

TRASP 2013

Tools and Resources for the Analysis of Speech Prosody

An Interspeech 2013 satellite event

August 30, 2013

Laboratoire Parole et Langage

Aix-en-Provence, France

Proceedings

Editors: Brigitte Bigi, Daniel Hirst

Assistant Editors: Joëlle Lavaud, Claudia Pichon-Starke

(c) LPL – Laboratoire Parole et Langage. All rights reserved.

TRASP 2013 Proceedings

Distributed by:

Laboratoire Parole et Langage (LPL)

5 avenue Pasteur

13100 Aix-en-Provence

Phone: +33 (0)4 13 55 36 20

Email: communication@lpl-aix.fr

Website: <http://www.lpl-aix.fr>

ISBN 978-2-7466-6443-2

Welcome to the Aix-en-Provence TRASP 2013 Workshop

TRASP 2013 is the first international meeting on Tools and Resources for the Analysis of Speech Prosody. TRASP is an Interspeech 2013 satellite event.

TRASP is a unique occasion to bring together people involved in developing tools and resources for the analysis of speech prosody in order to evaluate the state of the art in this area, summarising what tools and resources are currently available and what other of tools and resources are in greatest need of development. It is also an opportunity to discuss ways in which work in this area might benefit from harmonisation and collaboration.

We welcome you to Aix-en-Provence, a city steeped in both culture and academia. We hope you will enjoy your stay and the workshop.

Organizers

Brigitte Bigi and Daniel Hirst

Local organising committee

Members of the committee : Carine André, Nadéra Bureau, Laurence Colombo, Stéphanie Desous, Sophie Herment, Joëlle Lavaud, Frédéric Lefèvre, Nadia Monségu, Catherine Perrot, Claudia Pichon-Starke. TRASP was organised with the support of CEP (Thierry Legou) and SLDR (Bernard Bel).

Scientific Committee

- Véronique Aubergé, CNRS and Université Joseph Fourier, France
- Cyril Auran, University of Lille, France
- Plinio A. Barbosa, Instituto de Estudos da Linguagem/Unicamp, Brazil
- Katarina Bartkova, CNRS and Université de Lorraine, France
- Brigitte Bigi, CNRS and Aix Marseille University, France
- Nick Campbell, Trinity College, Éire
- HyongSil Cho, Microsoft Language Development Center, Portugal
- Jennifer S. Cole, University of Illinois, USA
- Piero Cosi, ISTC CNR, Italy
- Franco Cutugno, Università Degli Studi di Napoli Federico II, Italy
- Elisabeth Delais-Roussarie, CNRS and Université Paris Diderot, France
- Céline De Looze, Trinity College, Dublin, Éire

- Volker Dellwo, University of Zurich, Switzerland
- Grazyna Demenko, Uniwersytet im. A. Mickiewicza w Poznaniu, Poland
- Hongwei Ding, Tongji University, China
- Juan-María Garrido, Universitat Pompeu Fabra, Spain
- Dafydd Gibbon, Universität Bielefeld, Germany
- Jean-Philippe Goldman, Université de Genève, Switzerland
- Mark Allan Hasegawa-Johnson, University of Illinois, USA
- Sophie Herment, CNRS and Aix Marseille Université, France
- Keikichi Hirose, University of Tokyo, Japan
- Daniel Hirst, CNRS and Aix Marseille Université, France
- Ruediger Hoffmann, Technische Universität Dresden, Germany
- Philippe Martin, CNRS and Université Paris-Diderot, France
- Piet Mertens, KU Leuven, Belgium
- Amandine Michelas, CNRS and Aix Marseille Université, France
- Hansjoerg Mixdorff, Beuth University Berlin, Germany
- Antonio Origlia, Università degli Studi di Napoli Federico II, Italy
- Andrew Rosenberg, Queens College (CUNY), USA
- Jianhua Tao, Chinese Academy of Sciences, China
- Shu-Chuan Tseng, Academia Sinica, Taipei, Taiwan
- Martti Vainio, Institute of Behavioural Sciences, Helsinki, Finland
- Petra Wagner, Universitat Bielefeld, Germany
- Yi Xu, University College, London, UK

Acknowledgements

We wish to thank all of our sponsors for their generous support:



and the following institutions for their support:



We express our gratitude to all the external reviewers for having helped us put together a wonderful program.

Last, but certainly not least, we thank all of the conference members for their participation.

Program

8:00 - 8:45 Registration

9:00 - 9:30 Opening Session

9:30 - 10:30 [O1-1] Oral Session

Digital curation: the SLDR experience

Bernard Bel, Frédérique Bénard

ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis

Yi Xu

10:30 - 11:00 Coffee break

11:00 - 12:30 [P1-1] Poster Session – Ressources

Building OMProDat: an open multilingual prosodic database.

Daniel Hirst, Brigitte Bigi, Hyongsil Cho, Hongwei Ding, Sophie Herment, Ting Wang

Aix MapTask: A (rather) new French resource for prosodic and discourse studies

Ellen Gurman Bard, Corine Astésano, Alice Turk, Mariapaola D'Imperio, Noel Nguyen, Laurent Prévot, Brigitte Bigi

A Taiwan Southern Min spontaneous speech corpus for discourse prosody

Sheng-Fu Wang, Janice Fon

A heuristic corpus for English word prosody: disyllabic nonce words

Sophie Herment, Gabor Turcsan

C-PROM-Task. A New Annotated Dataset for the Study of French Speech Prosody

Mathieu Avanzi, Lucie Rousier-Vercruyssen, Sandra Schwab, Sylvia Gonzalez, Marion Fossard

12:30 - 14:30 Lunch

14:30 - 15:45 [P2-1] Poster Session – Pitch

Rapid and Smooth Pitch Contour Manipulation

Michele Gubian, Yuki Asano, Salomi Asaridou, Francesco Cangemi

Anonymising long sounds for prosodic research

Daniel Hirst

ModProso: A Praat-Based Tool for F0 Prediction and Modification

Juan-María Garrido

Automatic labelling of pitch levels and pitch movements in speech corpora

Piet Mertens

14:30 - 15:45 [P2-2] Poster Session – Prosody in Interface

An integrated tool for (macro)syntax-intonation correlation analysis

Philippe Martin

Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features

Katarzyna Klessa, Maciej Karpiński, Agnieszka Wagner

ProsoReportDialog: a tool for temporal variables description in dialogues

Jean-Philippe Goldman

Prosodic phrasing evaluation: measures and tools

Klim Peshkov, Laurent Prévot, Roxane Bertrand

15:45 - 16:15 Coffee break

16:15 - 17:30 [P3-1] Poster Session – Automatic annotation tools

What's new in SPPAS 1.5?

Brigitte Bigi, Daniel Hirst

TGA: a web tool for Time Group Analysis

Dafydd Gibbon

Timing analysis with the help of SPPAS and TGA tools

Jue Yu

SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora

Juan-María Garrido

16:15 - 17:30

[P3-2] Poster Session – Prosody Analysis

Continuous wavelet transform for analysis of speech prosody

Martti Vainio, Antti Suni, Daniel Aalto

Modeling Speech Melody as Communicative Functions with PENTAtainer2

Santitham Prom-on, Yi Xu

Semi-automatic and automatic tools for generating prosodic descriptors for prosody research

Plinio Barbosa

Variability of fundamental frequency in speech under stress

Grażyna Demenko, Magdalena Oleśkowicz-Popiel, Krzysztof Izdebski, Yuling Yan

17:30 - 18:30

Discussion / Closing Session

Table of contents

Digital curation: the SLDR experience <i>Bernard Bel, Frédérique Bénard</i>	1
ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis <i>Yi Xu</i>	7
Building OMProDat: an open multilingual prosodic database. <i>Daniel Hirst, Brigitte Bigi et al.</i>	11
Aix MapTask: A (rather) new French resource for prosodic and discourse studies <i>Ellen Gurman Bard, Corine Astésano et al.</i>	15
A Taiwan Southern Min spontaneous speech corpus for discourse prosody <i>Sheng-Fu Wang, Janice Fon</i>	20
A heuristic corpus for English word prosody: disyllabic nonce words <i>Sophie Herment, Gabor Turcsan</i>	24
C-PROM-Task. A New Annotated Dataset for the Study of French Speech Prosody <i>Mathieu Avanzi, Lucie Rousier-Vercruyssen et al.</i>	27
Rapid and Smooth Pitch Contour Manipulation <i>Michele Gubian, Yuki Asano et al.</i>	31
Anonymising long sounds for prosodic research <i>Daniel Hirst</i>	36
ModProso: A Praat-Based Tool for F0 Prediction and Modification <i>Juan-María Garrido</i>	38
Automatic labelling of pitch levels and pitch movements in speech corpora <i>Piet Mertens</i>	42
An integrated tool for (macro)syntax-intonation correlation analysis <i>Philippe Martin</i>	47
Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features <i>Katarzyna Klessa, Maciej Karpiński et al.</i>	51
ProsoReportDialog: a tool for temporal variables description in dialogues <i>Jean-Philippe Goldman</i>	55

Prosodic phrasing evaluation: measures and tools	59
<i>Klim Peshkov, Laurent Prévot et al.</i>	
What's new in SPPAS 1.5?	62
<i>Brigitte Bigi, Daniel Hirst</i>	
TGA: a web tool for Time Group Analysis	66
<i>Dafydd Gibbon</i>	
Timing analysis with the help of SPPAS and TGA tools	70
<i>Jue Yu</i>	
SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora	74
<i>Juan-María Garrido</i>	
Continuous wavelet transform for analysis of speech prosody	78
<i>Martti Vainio, Antti Suni et al.</i>	
Modeling Speech Melody as Communicative Functions with PENTA-trainer2	82
<i>Santitham Prom-on, Yi Xu</i>	
Semi-automatic and automatic tools for generating prosodic descriptors for prosody research	86
<i>Plinio Barbosa</i>	
Variability of fundamental frequency in speech under stress	90
<i>Grażyna Demenko, Magdalena Oleśkiewicz-Popiel et al.</i>	
Author Index	97
List of Authors	99

Digital curation: the SLDR experience

Frédérique Bénard, Bernard Bel

Laboratoire Parole et Langage (LPL)

CNRS – Aix-Marseille University

B9 80975, 5 avenue Pasteur

13604 Aix-en-Provence, France

frederique.benard@lpl-aix.fr, bernard.bel@lpl-aix.fr

Abstract

This paper deals with the description, packaging, and preservation of digital objects submitted to the Speech & Language Data Repository (www.sldr.org). SLDR is a Trusted Digital Repository offering the sharing oral/linguistic data and its submission for medium-term and long-term preservation via an institutional archive. Its work environment offers a flexible integrated management of access rights at all phases of a project. Data include all signals associated with oral production, documents created or collected during an experiment or a field enquiry, material derived from primary data with their associated resources and tools designed for data processing in the domain. Currently, information packages are distributed via the Adonis/Huma-Num grid hosted by *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3) and preserved on the platform of *Centre informatique de l'enseignement supérieur* (CINES), a site beneficiary of the Data Seal of Approval.

Index Terms: digital curation, data repository, resource sharing, OAIS, archive, Digital Humanities

1. A change of practice with respect to archives

In recent years, funding agencies supportive of linguistic research projects have been putting pressure on scholars to include long-term preservation and sharing of data in their project agenda. Initiating the archiving process at the very onset of the project is a radical change of practice because of associated technical and legal constraints. It is made possible by tools and procedures compliant with the life-cycle of present-day research projects. These will be introduced in this paper.

Compliance of procedures is achieved by responding favourably to the requests of data producers. Among these, we give brief answers to the most challenging ones:

- *Should I wait for the completion of my research work to submit only final versions of data and results?* Answer: A digital repository offers the options of upgrading stored material either via the submission of new versions or correcting mutable data.
- *Should I wait for the availability of informed consent by all participants to share recording in open access?* Answer: Open access is one among many options. Access to sensitive data may be restricted (in compliance with the legal framework) and an integrated management of access rights facilitates their gradual modification in more or less restrictive directions.

2. A repository for sharing oral/linguistic resources

In 2006, the LPL laboratory was commissioned by the Social Science and Humanities department of the French *Centre national de la recherche scientifique* (CNRS, www.cnrs.fr) to set up a resource centre for speech research.

This initiative was driven by a growing concern with the existence of scattered oral resources in non-persistent formats and locations, many of which could not be reused nor shared due to access restrictions.

Another incentive was to promote the self-archiving of linguistic resources in a manner similar to that of scientific publications submitted to *Centre pour la communication scientifique directe* (CCSD, www.ccsd.cnrs.fr). In those days, the dissemination of speech corpora was mostly carried out by corporate agencies (ELDA, www.elda.org, and the LDC, www ldc.upenn.edu). Admittedly, CNRS' initiative could initially be perceived as creating a public facility to replace a business framework unfit for academic work in small research units. Later on, this feeling of competition vanished thanks to a fair combination of private and public models (see *infra* §5.3).

At LPL, a generic (multidisciplinary and multilingual) digital repository was implemented from scratch after comparing existing similar initiatives [1]. During the same period, CNRS supported the creation of another site mostly replicating the design of LACITO's archive of rare languages [5].

In 2008-2011, LPL and LACITO were enrolled in a pilot project coordinated by TGE Adonis (www.tge-adonis.fr) for digital resource pooling in social sciences and the humanities. In this context, both repositories became submission sites for long-term preservation. After the completion of this project, LPL's resource centre was renamed *Speech & Language Data Repository* (SLDR) and LACITO's *Collections de Corpus Oraux Numériques* (COCOON).

A new phase of development started at the end of 2012. Six leading institutions joined efforts in the ORTOLANG project (www.ortolang.fr/english) with the aim of building a French sub-network of CLARIN centres (www.clarin.eu) involving SLDR for speech and CNRTL (www.cnrtl.fr) for text linguistics. Current focus is on interoperability with respect to descriptive metadata, persistent identifiers and controlled vocabularies.

3. From secure backup to long-term preservation

When initiating a research project, scholars should be aware of archival limitations with respect to file formats, and proceed to a comprehensive separation of primary and secondary data

with respective cycles of versioning. It is also advisable that files contained in a given folder share identical access restrictions.

These constraints are easier to comply with when secure backups are replaced with a process carrying out a technical validation of information package content and a set-up of controlled access to documents. In other words, medium-term preservation should follow the same procedures as long-term preservation.

As claimed on the CINES website (www.cines.fr), preserving digital resources is neither a backup service nor 'the ultimate step of storing data before oblivion or permanent loss.' In a long-term preservation scheme, data should be eligible for reuse after an unspecified period of time, typically more than 30 years. This calls for reliance on an institutional archive (CINES) rather than a consortium of computing centres whose policy might vary (because of fund scarcity) at a time data producers are no longer able to negotiate an extension of its preservation.

The commitment of CINES is threefold: (1) preserving data and its associated metadata; (2) preserving access-right information; (3) ensuring the reusability of data, which is achieved by migrating file formats (without loss of information) once these are becoming obsolete.

Digital items stored at the CINES archive are processed as generic 'information packages' regardless of their origin. Subject-specific information is stored in descriptive metadata encapsulated in XML files. Nonetheless, this implies technical limitations with respect to the packaging of Submission Information Packages (SIP) and a restricted set of open formats eligible for long-term preservation and logical data migration (www.sldr.org/wiki/Formats).

The issue of file formats is problematic when it comes to sound/video material. CINES accepts sound recordings in WAVE and AIFF with PCM encoding, or in compressed AAC — all non-proprietary formats. The popular MP3 format is not suitable because of its commercial restriction. Thus, if a corpus contains MP3 files, these will be stored in medium-term preservation whereas their replications in WAVE or AIFF format can be submitted for long-term preservation.

Preserving video recordings requires a trade-off between accuracy (e.g. high-resolution, 3D or multiple cameras), reasonable storage space and eligible formats (currently MP4/AVC/AAC, OGG/Theora/Vorbis and MKV/AVC/Flac). Even though SLDR may accept items featuring thousands of files and sizes beyond 100 Gigabytes, it should be kept in mind that the current annual cost of long-term preservation is roughly 5,000 euros per Terabyte (CINES estimation).

4. Digital curation

The term 'digital curation' is a combination of 'digital preservation' and 'curation', the latter in the sense of 'activities that add value and knowledge to the collections' (Tammaro & Madrid, cited in [6, p. 2]). Combining these words reflects an evolution of archival practice made possible by the use of digitized documents as research material eligible for long-term preservation, dissemination and reuse outside their original production environment. This creates new opportunities for enriching data provided that the issue of its portability has been properly addressed [2].

As put by Hedstrom [4, p. 2], digital curation calls for expertise (and training) 'across a spectrum from curation-centric needs to discipline or application specific requirements'. The challenge is to fill the gap of expertise

between producers (scholars, participants, informants) and archive curators in charge of the preservation of research/documentation material. On the side of archive curators, records management is no longer restricted to the preservation of 'semi-current' or 'inactive' records. It now includes 'curation at the source': assistance with the elaboration of descriptive metadata even before starting the collection of primary data.

Digital curation requires collaborative work and a proper coordination of initiatives to enrich archived material. To this effect, curation tasks accomplished by data depositors or administrators are traced by SLDR and displayed on the curation page (www.sldr.org/curation).

4.1. Packaging research data

The Huma-Num framework for long-term preservation of digital resources is based on the Open Archival Information System (OAIS) [3]. It comprises the two submission sites (SLDR and COCOON) connected with CINES (www.cines.fr) for long-term preservation and CC-IN2P3 (cc.in2p3.fr) for data dissemination, as shown Figure 1.

Items stored at SLDR are generic: any tree-structure of computer files with no limitation on size or (UTF8-encoded) file names. File formats incompatible with long-term preservation are automatically redirected to the dissemination site; this is the case for instance of sound files in MP3 format or video files in FLV format used for sound/video streaming, or ZIP files making it easy to download subsets of the tree.

SLDR deposit policy is the same as CCSD with respect to scientific papers (<http://hal.archives-ouvertes.fr>): documents, resources, corpora are not reviewed in terms of their presumed scientific or cultural-heritage value. The same permissiveness applies to the assessment of acoustic quality, accuracy of transcriptions, relevance of annotations etc., all of which should be taken care of by data producers. For this reason we incite producers to specify institutions responsible for data production and verification, as well as the funding bodies associated with the project (see Figure 2). Digital curation only cares for the proper technical packaging of data.

In a near future, ORTOLANG will develop facilities for the analysis and proper reuse of archived material. At this stage, the issue of data quality will be raised and guidelines produced to optimize interoperability and facilitate automated linguistic analytical processes.

4.2. Descriptive metadata

Every item has a specific space for 'documentary files' which may include documents describing its content, experimental protocols, associated material (diagrams etc.), transcriptions, translations and annotations of sound/video recordings etc. If the item is stored in medium-term or long-term preservation, documentary files will be modified without resubmitting a new version of the same item.

Documentary files include descriptive metadata, i.e. structured sets of information describing the contents of an item and its associated documents. Metadata are stored as XML files (at least one for each item) in the archive. In addition, the same are available in a database for quick access and reformatting.

4.2.1. Dublin Core OLAC

This metadata format is a qualification of Dublin Core applicable to the description of linguistic resources. See for

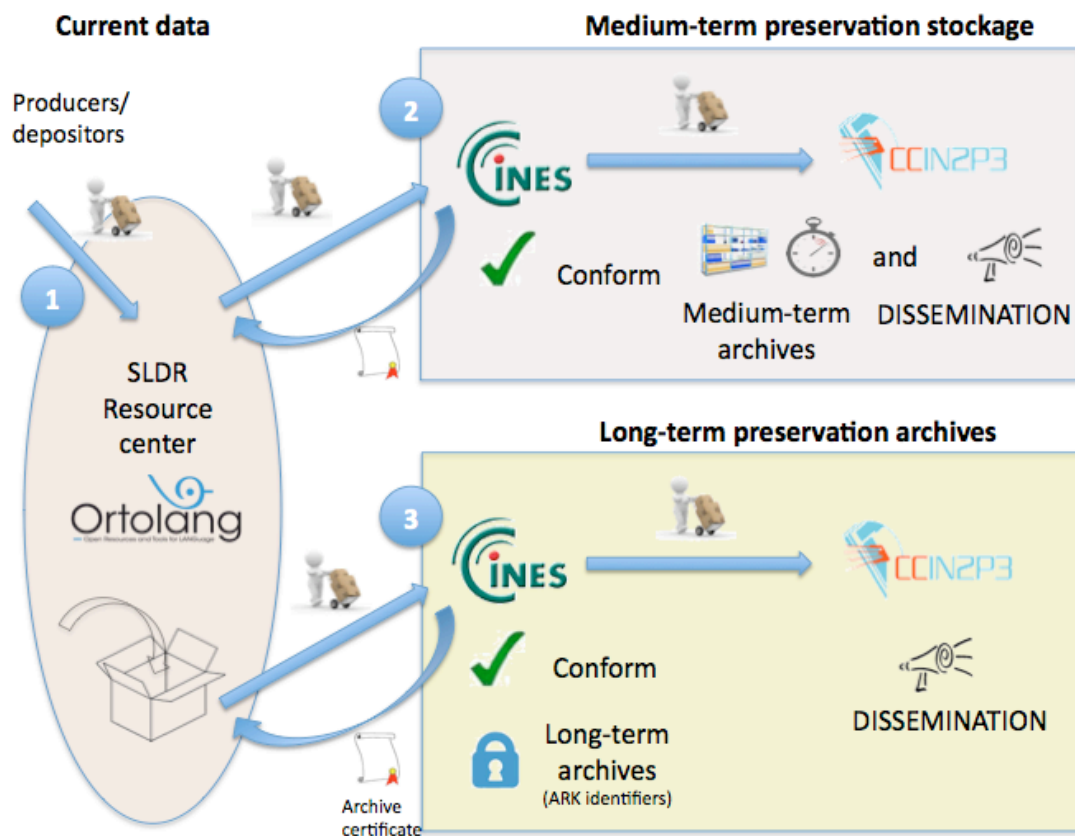



Figure 1. A descriptive outline of the data flow for medium or long-term preservation between SLDR, CINES and CC-IN2P3.

Speech & Language Data Repository

Speech & Language Data Repository (SLDR) <http://sldr.org>





Open archives ([OAI-PMH](#))

[\[Sign up\]](#) / [\[Login\]](#)

/ 中文 / English / español / français /

The Open ANC (OANC)

Nancy Ide, Randi Reppen, Keith Suderman
Department of Computer Science, Vassar College (New York US)

OAI: [oai:sldr.org/sldr000770](http://oai.sldr.org/sldr000770) ([olac](#) - [oai_dc](#) - [VLO](#) - [language-archives](#))
Persistent Identifier: hdl:11041/sldr000770
SLDR id: <http://sldr.org/sldr000770>
ARK: ark:/87895/1.4-183691
ARK: ark:/87895/1.4-183706
ARK: ark:/87895/1.4-183705
ARK: ark:/87895/1.4-183707
ARK: ark:/87895/1.4-183709
ARK: ark:/87895/1.4-183708
ARK: ark:/87895/1.4-183710

Sponsored by :

- National Science Foundation (BCS-98009, KDI, SBE)
- TalkBank project

Figure 2. A descriptive page of The Open ANC (62003 files). Note the careful mention of institutional support, funding agencies, identifiers according to OAI, Handle and SLDR schemes, and Archival Resource Keys associated with the 7 segments.

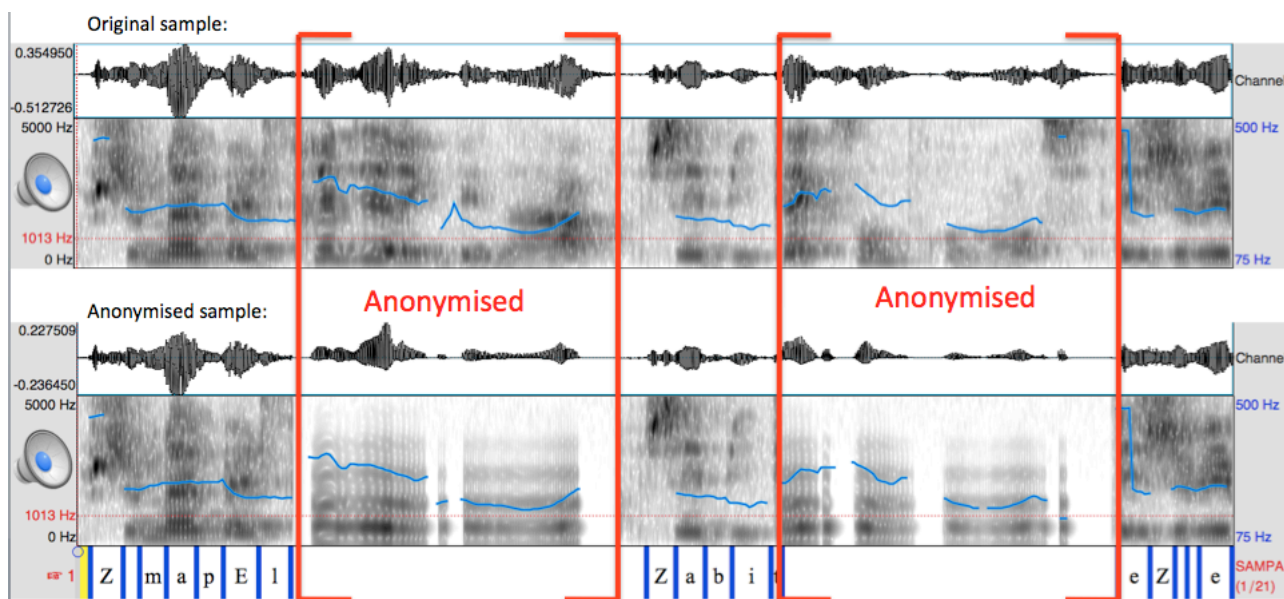


Figure 3. An illustration of the impact of the anonymisation PRAAT script by Daniel Hirst on a sentence before and after anonymisation. The pitch line is preserved.






Downloaded (58) primary data (corpus) Videos of CID - hdl:11041/sldr000027			
First name and last name	Institution	Field of research	Date of download
M Pascal NOCERA Contact	 Centre d'enseignement et de recherche en Informatique - EA 4128 (LIA, Avignon FR)	Traitement Automatique de la Parole	2008-10-17 licence #1
M Jean-Claude MARTIN Contact	 Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur - UPR 3251 (Limsi, Orsay FR)	communication multimodale	2008-10-20 licence #1
M Paul ISAMBERT Contact	 Langues, textes, traitements informatiques, cognition - UMR 8094 (LaTTiCe, Paris FR)	Structure du discours	2008-11-04 licence #1
M Olivier ROUCHON Contact	CINES 950 Rue de Saint Priest 34097 Montpellier Cedex 5  http://www.cines.fr/	Archivage pérenne de documents électroniques	2008-11-20 licence #1
Mme Sophie JAOUŁ Contact	 Formes et représentations en linguistique et littérature - EA 3816 (FoReLL, Poitiers FR)	didactique	2008-12-11 licence #1

Figure 4. A excerpt from the users' community for CID videos (hdl.handle.net/11041/sldr000027)

instance sldr.org/sldr000027/metadata/olac for a description of the Corpus of Interactional Data (hdl.handle.net/11041/sldr000027). Elements such as links and table of contents are automatically computed to describe the content.

4.2.2. CMDI

In addition to Dublin Core OLAC, SLDR and its partners in the ORTOLANG project will support Component Metadata (CMDI, www.clarin.eu/cmdi). A minimum set will use the information encapsulated in DC OLAC. Later, data producers will be given the option to select one among CMDI profiles for which web forms will be available.

4.2.3. EAD, METS

Structural metadata describing the tree-structure of an item will technical details derived from file analysis will be automatically incorporated into metadata files in the EAD and/or METS format. Access to this information will be required for processing queries over large sets of data.

4.2.4. Importing metadata formats

In the long term, the repository will be able to import metadata in various formats. Import from Dublin Core, CMDI and IMDI (www.clarin.eu/imdi) is easy to figure out. Other formats (such as DDI used by social scientists) will require the mapping of elements following schemes created by the research community.

4.3. Anonymisation

Today, the strong demand for sharing linguistic data results in an increasing need for anonymising audio files for both legal and ethical reasons. This curation task requires technical support. A PRAAT script for anonymising speech recordings is distributed on SLDR (hdl.handle.net/11041/sldr000526) with particular interest for speech prosodists. It replaces segments labeled with a key word on the accompanying TextGrid with a hum sound with the same prosodic characteristics as the original sound (Figure 3).

Producing the TextGrid from a simple text file is further facilitated by the Table2TextGrid script (hdl.handle.net/11041/sldr000811).

4.4. Persistent identifiers

A necessary feature for the proper reuse of research data is the assignment of links to documents that do not depend on their location in the repository. This location is bound to vary during the life cycle of an item. First it is available as 'source data' on the submission site (SLDR), but later it is transferred to the dissemination site (CC-IN2P3) as a 'secure backup'. Once the item has become stable, producers may decide to submit it for long-term preservation, which results in yet another location.

End users and web designers expect that the URLs pointing at files or 'datastreams' remain unchanged despite these changes of location. This is accomplished by the assignment of a persistent identifier (PID) to each individual document (www.sldr.org/wiki/Handle). Referring to PIDs makes it easy for outsiders to feed their web pages or blog articles with material directly extracted from the SLDR repository.

The SLDR algorithm for assigning PIDs relies on the assumption that a document shall retain its PID across changes of its location as well as new versions of the item which it belongs to. To this effect, the identity of a file is assessed by checking its digital signature (MD5): as long as both file name and digital signature remain unchanged, the same PID is assigned. In this way, every document deposited on SLDR is accessible to automated queries regardless of its archival status.

5. Accessing research data

There is a genuine interest for the interoperability of repositories for Digital Humanities, here meaning the possibility of launching analytical processes over sets of data stored in multiple repositories. Nonetheless, most current 'showcases' only work with open-accessible material.

5.1. Controlled access

Research scholars submitting data to a digital repository strongly insist on keeping control over its dissemination and access protocols. They often find it difficult to formalize access rights in limited technical frameworks set up by engineers. Admittedly, classical solutions fail to comply with the details of regulations for the protection of private data and intellectual property. This calls for an integrated management of access rights covering the broadest diversity of cases.

France takes advantage from a significant advance on archive law, namely its *Code du patrimoine* (the Heritage Code) clarifying the notion of 'public archive' with a set of formal rules regulating access to archived documents (Act of 15 July 2008, articles L213 1-5). This framework prompted a radical change of policy as any public archive is expected to be open-accessible, with the exception of 24 derogations applicable to certain categories of documents (www.sldr.org/wiki/table_derogations_en). Each derogation has been assigned a code facilitating a systematic management. A frequent derogation case is the protection of privacy (50 years, code AR048, art. L213-2, I, 3). For a recorded audio/video corpus, this derogation may be invoked to restrict access until authorisations have been signed by all participants.

In SLDR, users are assigned categories according to profiles defined by institutional producers. If no profile is available, the default SLDR profile (sldr.org/wiki/Groupees) is applied which makes a distinction between 'academics' — teachers, students and research scholars working on subjects related with linguistics — and other users, including the ones affiliated with the speech industry. This status is carefully verified at the time of signing up. SLDR default categories and procedures are similar to the 'URCS protocol roles' used by ELAR (Nathan 2013: 6).

Users granted access to restricted material are requested to check the SLDR licence (sldr.org/wiki/Licences_en) for a licit use of the resource. Peer-to-peer exchange is only allowed for items bearing a Creative Commons licence. Data producers may optionally impose additional clauses in a specific licence.

SLDR keeps records of all controlled-access downloadings. Thus, any user of a resource may consult the list of other users, check their credentials or contact them to seek information about their planned usage of the material (see Figure 4).

5.2. Shared licences

Current development of access rights management at SLDR is aiming at a social networking approach inspired by ELAR's 'protocol' approach, here meaning 'the concepts and processes that apply to the formulation and implementation of language speakers' rights and sensitivities, and the consequent controlled access to materials.' [7, p. 4].

Transactions with groups of users are facilitated by 'shared licences' granted to sets of archived items, persons or institutions. This technique applies to individuals or groups belonging to a particular community of research participants.

An example of non-commercial licence is the Buckeye Corpus of Conversational Speech distributed by Ohio State University (hdl.handle.net/11041/sldr000776). This corpus is under a licence shared by all members of a laboratory. Thus, access is granted to persons whose affiliation with the licenced institution has been authenticated by SLDR.

Shared licences may also be used for disseminating material purchased by a group of laboratories, as is the case with the Treebank collection acquired by CNRS (hdl.handle.net/11041/ldc000828).

5.3. 'Commercial' versus 'academic'

The distinction between 'commercial' and 'non-commercial' is not a rigid one. Agencies distributing language resources (such as ELDA and the LDC) adopt a pragmatic approach with respect to financial participation: scholars and public laboratories are granted access to resources at rates significantly lower than the speech industry; the resource may even be given free on request.

SLDR has provision for links with ELRA and LDC resources of this type. See for example two instances of the EUROM collection (hdl.handle.net/11041/sldr000034, hdl.handle.net/11041/sldr000035) and the Open ANC (hdl.handle.net/11041/sldr000770).

This pragmatism is not resented as discriminative by corporations because they prefer to pay a high fee for the service which implies a contractual protection against litigation. Engineers feel reluctant to use resources labelled 'public domain' because of the trouble their company might face if this free licencing is challenged due to their use in a commercial product.

5.4. Individual permission

Resource producers sometimes need to retain full control of the sharing of their material. This may be the case with scholars using SLDR for a restricted sharing of their speech corpus until the completion of their research work. In this case, owners of the resource receive mail queries sending them to a web form for granting or denying access to applicants during a given period of time.

6. New perspectives

SLDR is a promising work environment for archive curators in charge of linguistic/oral/multimodal research material. Current focus is on the automation of curation tasks such as the production of accurate metadata and the packaging of generic items following the OAIS model. In the context of ORTOLANG, new features are being integrated such as:

- Tools for automatic phonetic annotations and alignment of transcriptions (SPPAS, www.sldr.org/sldr000800);

- A systematic pre-processing of sound files including their segmentation to intonation units and MOMEL/INTSINT labeling; these will be applicable to multitrack recordings (more than 2 microphones);
- Automatic conversion of descriptive metadata and the production of structural metadata.

This list is not exhaustive. We expect that dealing with a great diversity of linguistic material from a wide range of disciplines will encourage the development of tools and procedures coping with the requirements of high-quality research.

7. References

- [1] Bel, B. and Blache, P., "Le Centre de Ressources pour la Description de l'Oral (CRDO)", Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA), 25:13-18, 2006. Online: <http://hal.archives-ouvertes.fr/hal-00142931> accessed on 19 Jun 2013.
- [2] Bird, S. and Simons, G., "Seven Dimensions of Portability for Language Documentation and Description", Language, 29:557-582, 2003. Online: [arXiv:cs/0204020](http://arxiv.org/abs/cs/0204020) accessed on 19 Jun 2013.
- [3] CCSDS, "Reference Model for an Open Archival Information System (OAIS)", Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1, August 2009.
- [4] Hedstrom, M., "Digital Data Curation – Workforce demand and educational needs for digital data curators", Proceedings of conference Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnals, Florence, 11-12 December 2012 (in press). Online: http://www.rinascimento-digitale.it/conference2012/paper_ic_2012/hedstrom_paper.pdf accessed on 19 Jun 2013.
- [5] Michailovsky, B., Michaud, A. and Guillaume, S., "A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive", International Conference on Speech Database and Assessments (Oriental COCOSA), Hsinchu: Taiwan, 2011. Online: <http://halshs.archives-ouvertes.fr/halshs-00620893> accessed on 19 Jun 2013.
- [6] Moulaison, H.L. and Corrado, E.M., "LAM education for digital curation: A North American perspective", Proceedings of conference Cultural Heritage on Line, Trusted Digital Repositories & Trusted Professionnals, Florence, 11-12 December 2012 (in press). Online: http://www.rinascimento-digitale.it/conference2012/paper_ic_2012/moulaison_paper.pdf accessed on 19 Jun 2013.
- [7] Nathan, D., "Digital archiving", in P.K. Austin and J. Sallabank, [Eds], The Cambridge Handbook of Endangered Languages, 255-273, Cambridge University Press, 2001.

ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis

Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK

yi.xu@ucl.ac.uk

Abstract

This paper introduces ProsodyPro — A software tool for facilitating large-scale analysis of speech prosody, especially for experimental data. The program allows users to perform systematic analysis of large amounts of data and generates a rich set of output, including both continuous data like time-normalized F_0 contours and F_0 velocity profiles suitable for graphical analysis, and discrete measurements suitable for statistical analysis. It maximizes efficiency by automating tasks that do not require human judgment, and saving analysis output in formats that are ready for further graphical and statistical analysis.

Index Terms: ProsodyPro, time-normalization, annotation, F_0 velocity

1. Introduction

The need for stringent experimental control in prosody research is increasingly recognized [1, 4, 7, 8]. The analysis of experimental prosody data, however, is not always a straightforward matter. There is often a dilemma between systematic comparison of discrete measurements [8] and detailed analysis of continuous prosody [3, 5]. In most cases, details are sacrificed in favor of straightforward comparisons. In the intonation literature, for example, typically continuous F_0 contours of a few utterances are shown as illustrations, and then subsequent analyses are done on only a limited set of measurements, with little or no further examination of continuous contours. This leaves many details in continuous prosody unknown. A likely reason for such compromise is the lack of tools that can simultaneously facilitate both types of analysis.

2. ProsodyPro — An integrated tool

This paper introduces ProsodyPro, a Praat script that allows users to combine systematic comparison with detailed analysis of continuous prosody. The key design is to use time-normalization to facilitate direct comparison of continuous F_0 contours, while at the same time generate multiple measurements from non-time-normalized data suitable for statistical analysis.

2.1. Time-normalization

Time-normalization is the key method of ProsodyPro to facilitate close scrutiny of continuous F_0 contours over multiple tokens. When an experiment has recorded many sentences, especially when each unique sentence is repeated several times by each speaker and by multiple speakers, it becomes difficult to analyze and virtually impossible to report

all the data. For example, Figure 1a shows F_0 contours of the Mandarin tone sequence HLFHH produced by four male speakers. These F_0 tracks are generated by the “Export visible pitch” function in Praat [2]. Visually we can see some similarities across the speakers despite the differences. But how can we capture the similarities? One way is to take a measurement in the middle of all syllables and average them across the repetitions as well as the speakers, as shown in Figure 1b, from which we can see that the greatest differences occur on syllable 2, which carries four alternative tones. However, while statistics may show a significant difference across the four tones with this kind of measurement, many finer differences are lost. In Figures 1c and 1d, as two or three measurements are taken from each syllable, more details start to emerge. But it is not until Figure 1e, where eight measurements are taken from each syllable, does the continuous nature of the F_0 contours clearly emerge.

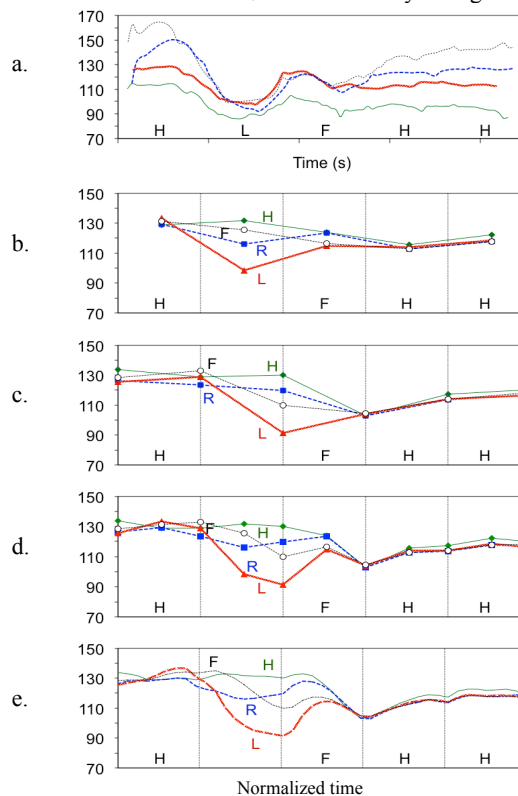


Figure 1: a. Raw F_0 tracks of HLFHH by 4 male speakers, generated by Praat. b-e. Mean time-normalized F_0 of a five-syllable Mandarin sentence sampled at 1, 2, 3 and 8 samples per syllable.

Thus time-normalization allows averaging across repetitions as well as speakers, a process that also smoothes out random variations unintended by the speaker, as well as individual differences, leaving only consistent variations due to tone and contextual tonal variations. From Figure 1 we can also see that time-normalization is only a further extension of the coarser sampling as in Figures 1b-1d, which are in fact also time-normalized. But the finer sampling allows us to see much more details, leaving little to guesswork. This can be seen in the two examples in Figure 2, where similarities in focus realization can be seen between Mandarin and English. Note that the English F_0 contours in Figure 2b are displayed in real rather than normalized time. This is achieved by taking time values together with F_0 values, and averaging both across repetitions and speakers. The averaged real time can then be used as the time axis for plots like Figure 2b.

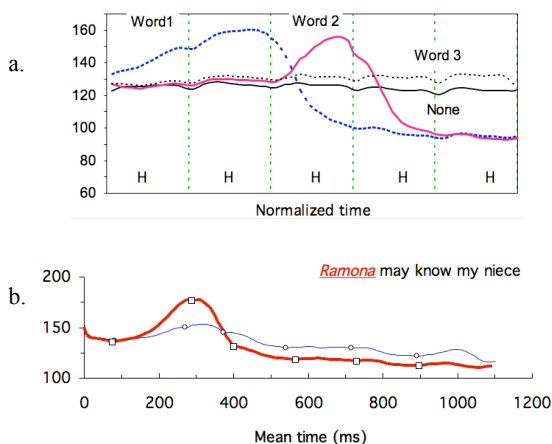


Figure 2: Mean F_0 contours of Mandarin (a) and English (b) sentences in different focus conditions. The plots are adapted from studies [6, 11] using precursor versions of ProsodyPro.

Time-normalization, however, requires users to define the temporal domain of normalization. In ProsodyPro this is done by inserting interval boundaries in the TextGrid of an utterance. Technically ProsodyPro allows the use of any units as the temporal domain of normalization, e.g., syllable, word, or even phrase. But it is important that there are good reasons to believe that the F_0 contours of the unit are consistently produced. Our recommendation is to use the syllable (or rhyme) whenever possible. This is based on evidence that speakers produce syllable-sized contours consistently [9, 10].

2.2. Other continuous prosody output

In addition to time-normalized F_0 contours, ProsodyPro also generates a number of non-time-normalized continuous prosody outputs. The following is a list of all the continuous prosody files, with a variety of time scales:

- X.rawf0 — Real-time F_0 (Hz) converted directly from vocal periods ($F_0 = 1/T$, where T is vocal period in second) marked in X.pulse
- X.f0 — Real-time F_0 (Hz) smoothed with a trimming algorithm (Xu, 1999)
- X.smoothf0 — Real-time F_0 (Hz), smoothed from X.f0 with a triangular filter

- X.normtimef0 — Time-normalized F_0 (Hz), with default sampling rate of 10 points/interval
- X.actutimenormf0 — Time-normalized F_0 (Hz) with real-time x-axis values
- X.samplef0 — F_0 (Hz) sampled at at fixed time intervals specified by “f0 sample rate”
- X.semitonef0 — Semitone version of X.samplef0
- X.f0velocity — Continuous F_0 velocity (instantaneous rates of F_0 change) in semitone/s sampled at time intervals specified by “f0 sample rate”
- X.normtime_f0velocity — Time-normalized continuous F_0 velocity

All these files are automatically generated after an utterance is annotated. These outputs allow users to examine continuous prosody of each utterance. However, as explained above, only the time-normalized ones can be averaged into mean contours across repetitions and speakers.

2.3. Discrete measurements

Time-normalization, despite its many advantages as mentioned above, is meant only for graphical comparisons. For statistical analysis, it is impractical to compare every time-normalized point. Nonetheless, time-normalization allows us to see the locations and manners of maximum differences when mean continuous contours from different experimental conditions are plotted in overlaid graphs, like those in Figure 1e and Figure 2. This can help identify optimal measurements that best reflect the key differences between experimental conditions, and, just as importantly, avoid pitfalls. Figure 2a, for example, shows that the F_0 at the onset of the first post-focus syllable “-na” is higher than in the same syllable in the neutral focus contour. This means that maxf0 of that syllable is not the best measurement for showing the lowered F_0 characteristic of all the other post-focus syllables.

The following measurements are automatically generated by ProsodyPro for each non-blank interval in the TextGrid and saved in the file X.means, where X stands for the name of the sound file being analyzed:

1. maxf0 — maximum F_0 in Hz
2. minf0 — minimum F_0 in Hz
3. excursion_size — difference between maximum F_0 and minimum F_0 in semitones
4. meanf0 — average F_0 in Hz
5. max_velocity — maximum F_0 velocity in semitones/s
6. finalf0 — F_0 near the interval offset in Hz
7. final_velocity — F_0 velocity near the interval offset in semitones/s
8. duration — interval duration in ms
9. mean intensity — mean intensity in dB

2.4. Ensemble files

To facilitate both graphical and numerical analysis, ProsodyPro has a function to pool the outputs of all individual sounds in a folder together into a large set of ensemble files. The following is a list of ensemble files generated by the current version of ProsodyPro (some are optional):

- 1) normf0.txt
- 2) normtime_semitonef0.txt
- 3) normtime_f0velocity.txt
- 4) normtimeIntensity.txt

- 5) normactutime.txt
- 6) samplef0.txt
- 7) f0velocity.txt
- 8) maxf0.txt
- 9) minf0.txt
- 10) excursionsize.txt
- 11) meanf0.txt
- 12) duration.txt
- 13) maxvelocity.txt
- 14) finalvelocity.txt
- 15) finalf0.txt
- 16) meanintensity.txt
- 17) mean_normf0.txt
- 18) mean_normtime_semitonef0.txt
- 19) mean_normtime_f0velocity.txt
- 20) mean_normtimeIntensity.txt
- 21) mean_normactutime.txt
- 22) mean_maxf0.txt
- 23) mean_minf0.txt
- 24) mean_excursionsize.txt
- 25) mean_meanf0.txt
- 26) mean_duration.txt
- 27) mean_maxvelocity.txt
- 28) mean_finalvelocity.txt
- 29) mean_finalf0.txt
- 30) mean_meanintensity.txt
- 31) mean_normf0_cross_speaker.txt

These are all text files that can be opened by spreadsheet, graphing or statistical programs. The first seven files contain continuous prosody of all the annotated sounds in the folder in normalized, real or sampled time. Files 8-16 each contains values of a specific measurements from all the sounds. Files 17-30 contain values averaged across the repetitions of unique sentences. And file 31 contains time-normalized F_0 values averaged across multiple speakers. Values in files 17-21 are ready for graphing mean F_0 , F_0 velocity or intensity contours of individual speakers. Values in file 31, which contains across-speaker means, can generate plots like those in Figures 1e and 2. The discrete mean measurements in files 22-30 are ready for statistical analysis such as anova and t-test, for which each subject needs to contribute only a single mean value for each of the factors being tested.

3. Workflow and time-saving features

To analyze large amounts of experimental data, much labor is needed simply to process the data files, taking and recording the measurements, and readying the output data for graphical and statistical analysis. ProsodyPro has incorporated many labor-saving features specifically designed for experimental research.

- First, ProsodyPro is written as a Praat script, so that it is executable on all major computer platforms.
- Second, the entire program consists of a single script file to be run in the same folder as the sound files being analyzed, thus reducing installation effort to a minimum.
- Third, ProsodyPro automates tasks that require little human judgment, including, in particular, finding and opening sound files, and saving analysis results.
- Finally, ProsodyPro arranges output data in formats that are nearly ready to be processed further by spreadsheet, graphing and statistical programs.

The following is a brief sketch of the workflow of ProsodyPro, with graphic illustrations shown in Figure 3.

1. Put ProsodyPro.praat in the folder containing the sound files to be analyzed.
2. Launch ProsodyPro either from “Open Praat Script...” command from Praat menu, or by double-clicking the ProsodyPro icon if it is recognizable as a Praat script by the operating system.
3. When the script window opens in Praat, select “Run” from the Run menu (or use the shortcut command-r or control-r).
4. In the startup window (Figure 3a), check or uncheck the boxes when necessary, and set appropriate values in the text fields or simply use the default ones. Select the task “Interactive labeling” by ticking the first radio button.
5. Click Ok and three windows will appear. The PointProcess window (Figure 3b) displays waveform together with vocal cycle marks (vertical lines) generated by Praat. Here one can manually add missing marks (e.g., in the last vocalic sound in the figure) and delete redundant ones. This needs to be done only for the named intervals, as explained next.
6. The TextGrid window (Figure 3c) displays waveform and spectrogram of the current sound. (Note that the pitch track in the spectrogram panel is not used by ProsodyPro.)
7. At the bottom of this window are the annotation tiers, where users can insert interval boundaries (Tier 1) and comments (Tier 2). Any blank intervals (or those labeled “sil”) in Tier 1 are ignored, which allows easy exclusion of temporal regions from analysis. The interval labels can be as simple as a, b, c... or 1, 2, 3..., since ProsodyPro ignores label content.
8. The Pause window (Figure 3d) controls the workflow. To bring up the next sound to be analyzed, change the number (or leave it as is) in the current_file box and press “Continue”. The number indicates the order in the String object “list” in the Object window (and in the file “FileList.txt” automatically saved to the current folder). The next sound will be 1 + current_file (So, entering 0 opens sound 1).
9. To end the current analysis session, press “Finish” in the Pause window, and the order number of the last sound analyzed is shown in the Praat Info window. That number can be used as a starting point in the next analysis session.
10. After processing individual files, ensemble files can be generated by running ProsodyPro again with the “Get ensemble files” radio button checked.
11. The analysis parameters in the startup window can be modified after annotating all the sound files. To do so, start a new session with the radio button “Process all sounds without pause” checked. ProsodyPro will exhaustively run through all files nonstop.
12. To generate the means files (files 17-30 listed in Section 2.4), set the value of “Nrepetitions” in the startup window to the number of repetitions in the dataset when running ProsodyPro with the “Get ensemble files” button checked. Note that the number of labeled intervals must be identical across repetitions.

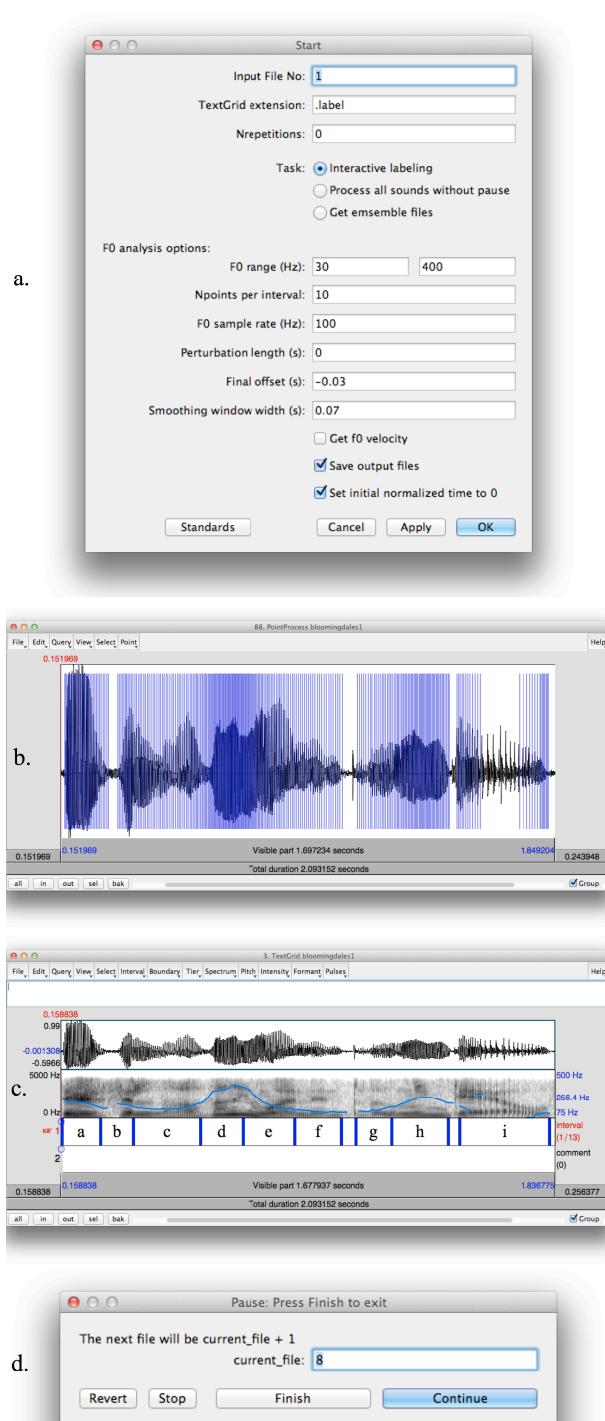


Figure 3: *ProsodyPro* workflow. *a.* Startup window for setting control and analysis parameters, *b.* PointProcess window for manually rectifying vocal pulse markings, *c.* TextGrid window for segmentation and annotation, *d.* Pause window for controlling workflow.

13. To generate mean F_0 contours averaged across speakers, the following steps can be followed:

- Create a text file (speaker_folders.txt) containing the speaker folder names arranged in a single column.

- Run ProsodyPro with the 4th task (Average across speakers) checked. The script will read mean_normf0.txt from all the speaker folders, average the f_0 values on a logarithmic scale, and then convert them back to Hz.
- The grand averages are saved in "mean_normf0_cross_speaker.txt".

4. Limitations

Despite its many time-saving features, the use of ProsodyPro is still often labor-intensive. This is first because typical experimental datasets are large, having multiple speakers and multiple repetitions. Secondly, the rectification of vocal pulse markings can be time-consuming, which is especially the case when the expectation for accuracy of F_0 tracking is raised once human intervention is involved. While this may not be a real disadvantage given that the amount of labor is proportional to the increase in accuracy of prosody analysis, it is desirable to develop methods in future versions that can accelerate the labeling and vocal pulse marking process.

5. Conclusions

As an integrated prosody analysis tool, ProsodyPro resolves the dilemma between detailed analysis of continuous prosody and systematic comparison of discrete measurements. It also minimizes labor by automating tasks that do not require human judgment, and facilitates human intervention of processes that are prone to error, thus delivering high accuracy and reliability in prosody analysis.

6. References

- [1] Bruce, G., and Touati, P., "On the analysis of prosody in spontaneous speech with exemplification from Swedish and French", *Speech Communication*, 11:453-458, 1992.
- [2] Boersma, P., "Praat, a system for doing phonetics by computer", *Glott International*, 5:9/10:341-345, 2001.
- [3] Hawkins, S., "Roles and representations of systematic fine phonetic detail in speech understanding", *Journal of Phonetics*, 31:373-405, 2003.
- [4] Nakai, S., Turk, A. E., Suomi, K. et al., "Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish", *Journal of Phonetics*. 40(6):796-807, 2012.
- [5] Post, B., D'Imperio, M., and Gussenhoven, C., "Fine phonetic detail and intonational meaning", *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrücken, 191-196, 2007.
- [6] Xu, Y., "Effects of tone and focus on the formation and alignment of F_0 contours", *Journal of Phonetics*, 27:55-105, 1999.
- [7] Xu, Y., "In defense of lab speech", *Journal of Phonetics*, 38: 329-336, 2010.
- [8] Xu, Y., "Speech prosody: A methodological review", *Journal of Speech Sciences*, 1:85-115, 2011.
- [9] Xu, Y., and Liu, F., "Tonal alignment, syllable structure and coarticulation: Toward an integrated model", *Italian Journal of Linguistics*, 18:125-159, 2006.
- [10] Xu, Y., and Liu, F., "Intrinsic coherence of prosody and segments", *Understanding Prosody – The Role of Context, Function, and Communication*, O. Niebuhr, ed., 1-26: Walter de Gruyter, 2012.
- [11] Xu, Y., and Xu, C. X., "Phonetic realization of focus in English declarative intonation", *Journal of Phonetics*, 33:159-197, 2005.

Building *OMProDat*: an open multilingual prosodic database

Daniel Hirst^{1,2}, Brigitte Bigi¹, Hyongsil Cho^{3,4}, Hongwei Ding², Sophie Herment¹, Ting Wang²

¹Laboratoire Parole et Langage, UMR 7309 CNRS & Aix-Marseille University, France

²School of Foreign Languages, Tongji University, Shanghai, China

³Microsoft Language Development Center, Lisbon, Portugal

⁴ADETTI – ISCTE, IUL, Lisbon, Portugal

daniel.hirst@lpl-aix.fr, brigitte.bigi@lpl-aix.fr, t-hych@microsoft.com,

hongwei.ding@tongji.edu.cn, sophie.herment@univ-amu.fr, 2011ting_wang@tongji.edu.cn

Abstract

Current research on speech prosody generally makes use of large quantities of recorded data. In order to provide an open multi-lingual basis for the comparative study of speech prosody, the *Laboratoire Parole et Langage* has begun the creation of an open database *OMProDat* containing recordings of 40 five sentence passages, originally taken from the European SAM project, each read by 5 male and 5 female speakers of each language. The database will contain both primary data, the recordings, and secondary data in the form of different annotation files. Currently the database contains recordings and annotations for five languages: Korean, English, French and Chinese plus a smaller subset for several languages which will be used for the TRASP workshop. All the data will be freely available on the *Speech and Language Data Repository*.

Index Terms: resources, database, speech prosody, multilingual, open source

1. Introduction

In the last two decades, there has been an increased awareness of the need to establish prosodic descriptions on the basis of large quantities of empirical data. Comparing the prosody of different languages, in particular, requires the analysis of comparable data from several speakers for each language.

1.1. The EUROM1 corpus

One of the first systematic attempts to provide a multi-lingual resource for speech technology was the Eurom1 corpus, [04], created as a deliverable of the European Esprit project 2589 SAM (Speech Assessments and Methodology) and its follow-up project SAM-A. *Eurom1* contained, in particular, a series of 40 continuous and thematically connected five-sentence passages, intended to represent a *clean* version of the various types of speech which speech technology might be expected to deal with. The passages were based on identical themes for the different languages, freely translated and adapted from the original English texts for the different languages.

Two sample passages from the Eurom1-EN database are:

[T02] I have a problem with my water softener.
The water-level is too high and the overflow keeps dripping.
Could you arrange to send an engineer on Tuesday morning please?

It's the only day I can manage this week.

I'd be grateful if you could confirm the arrangement in writing.

[T33] Hello, is that the telephone-order service?

There seems to have been some mistake.

I ordered a teddy bear from the catalogue and was billed for an electric lawnmower.

And I don't even have a garden.

Would you put me through to the complaints department, please?

The passages were originally recorded in the 1980's for eleven European languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish.

The recordings of the passages were separated into two corpora: the many talkers corpus (MANY) and the few talkers corpus (FEW). For the MANY corpus, 3 passages were read by 30 male and 30 female speakers. For the FEW corpus, 5 male and 5 female speakers each read only a limited number of the 40 passages, typically 15 passages per speaker for most of the languages but 20 passages per speaker for German and only 10 passages per speaker for French. The result of this is that in the FEW corpus there are only 2 or 3 recordings of each passage for most languages.

1.2. The Babel corpus

A compatible speech database for East European languages was later recorded during the Copernicus project 1304, *Babel*, with similar recordings for Bulgarian, Estonian, Hungarian, Polish, and Romanian [18].

1.3. The MULTEXT Prosodic Database

The continuous passages from the FEW corpus in five languages, (English, French, Italian, German, and Spanish) were re-used during the Esprit project Multext. The recordings were provided with manually created annotation files for word labels and with automatically stylised f0 patterns using the Momel algorithm [08, 10]. The database was published as the *MULTEXT Prosodic Database* [03].

A compatible version of the database for East European languages was produced as Multext-East [06].

1.4. Other recordings

A Japanese version of the corpus with 3 male and 3 female speakers reading all 40 passages in two different speaking styles [19, 20], included recordings and stylized F0 curves using Momel. It also contains the time-aligned labels of phonemes, phrases, and J-ToBI annotation as well as the native speakers' judgment of lexical accents and data from EGG electrodes. This was followed by a Chinese version [17] with 5 male and 5 female speakers. One speaker read all 40 passages, and each of the other 9 speakers read 15 passages. Each passage was read by 4 or 5 speakers.

1.5. Availability

The original Eurom1 recordings were protected by copyright assigned to the different laboratories that produced the recordings. For details see <http://www.phon.ucl.ac.uk/shop/eurom1.php>. The database contained on 30 CDs is available for sale from the same address for £100.

The Babel corpora are available from ELRA (http://catalog.elra.info/product_info.php) at 600€ per language for researchers. ELRA members get a 50% discount and ELRA membership costs 750€ for non-profit-making organisations.

The Multext Database is available from the same address for 100 € for academic researchers. The Multext-East recordings are freely available from <http://nl.ijs.si/ME/>.

The Japanese version of the corpus is distributed by the Faculty of Information, Shizuoka University via the author Shigeyoshi Kitazawa <kitazawa@cs.inf.shizuoka.ac.jp>, after signing a licence agreement which prohibits redistribution of the corpus. The Chinese version of the corpus is available from *Speech Resources Consortium* (NII-SRC) free of charge apart from a minimal sum to cover shipping. (<http://research.nii.ac.jp/src/en/MULTEXT-C.html>).

2. Building OMProDat

In order to provide a more solid basis for the analysis of prosodic metrics, we decided to build an open multilingual prosodic database **OMProDat**, to be archived and distributed by the recently created *Speech and Language Data Repository* (SLDR) (<http://sldr.org>) under an open database license. The database will be available at:

<http://sldr.org/sldr000725>

The aim of this database is to collect, archive and distribute recordings and annotations of directly comparable data from a representative sample of different languages representing different prosodic typological characteristics.

As mentioned above, the passages of the different versions of the original Eurom1 corpus were typically read by only two or three speakers each. This makes the corpus of limited use for the study of speaker variability.

We consequently decided to make new recordings of the corpus, with all 40 passages read by 10 speakers each.

2.1. Korean

The first language recorded under these conditions was Korean [16]. The original English version of the Eurom1 text was translated into Korean. The texts in Korean alphabet were Romanized and also transcribed in SAMPA and IPA. 10 Seoul speakers

(5 male and 5 female) took part in the recording session, all were Korean native speakers in their twenties, either undergraduate or graduate students of Seoul National University. Each speaker read all 40 passages.

For prosodic annotation, the Momel algorithm was used [10] and the pitch targets obtained were manually corrected. The prosodic events were annotated in two ways: first, with the automatic annotation algorithm, INTSINT [10] and second, with manual labelling of prosodic units using just two tone labels (H and L).

2.2. English and French

This was followed by new recordings for English and French read by native speakers, as well as for English read by native speakers of French and for French read by native speakers of English [07]. The speakers were all 20-30 years old. All speakers were from monolingual families. The English speakers were recorded in Oxford and spoke Southern British English; the French speakers were recorded in Aix-en-Provence and spoke either a Southern or a standard variety of French, or something between the two.

The originality of this corpus is that it provides recordings for both natives and non-native speakers, so as to allow comparative studies on L1 and L2 productions.

Three groups of learners were recorded for each language, one group of native speakers on the one hand and two groups of non-native speakers, corresponding to the levels of the Common European Framework of Reference for Languages, (CEFR): classified respectively as *independent users* (level B1/B2) and *proficient users* (level C1/C2).

The recordings are accompanied by TextGrid annotation files obtained semi-automatically from the sound and the orthographic transcription using the SPPAS alignment software [01] using manual correction when necessary.

Prosodic annotation was also obtained using the Momel and INTSINT automatic annotation algorithms [10].

2.3. Chinese

Most recently we have added recordings for Standard Chinese [05]. The speakers were 10 Chinese native speakers: 5 female and 5 male. Their ages ranged from 21 to 31 years old, and they were all postgraduate students and speakers of standard Chinese. Before recording started, they were asked familiarise themselves with the texts and were given some practices at reading them at a normal speaking rate and with a natural intonation. During the recording, the speakers were asked to repeat the whole passage whenever a word was produced wrongly. Each speaker read all 40 passages. The annotation of the recordings using SPPAS and Momel/INTSINT is currently in progress.

2.4. The OpenProDat multilingual sample

In the context of the TRASP workshop, we collected and distributed a more limited set of data from a larger number of languages. In order to increase comparability for the different tools, we asked TRASP participants to apply their tools to this corpus. Since it is expected that tools may concern several different languages, a multilingual corpus was necessary.

We choose the two paragraphs from the English Eurom1 corpus given as examples in section (1.1) and we translated and recorded the texts. This shared corpus is hosted by the SLDR

forge (repository number 805) under the name *OpenProDat* as a part of the more general *OMProDat* database described in this paper.

These texts were transcribed in: Dutch, French, German, Italian, Arabic, Spanish, Finnish, Hungarian, Japanese and Thai. Each participant read both paragraphs, first in their mother tongue and then in each language that they felt able to read.

By April, 2013, this corpus included data described in Table 1, recorded by 24 speakers (14 female, 9 male and 1 child).

Language	L1	L2
English	5	18
French	5	21
German	4	1
Italian	4	4
Dutch	1	1
Arabic	2	0
Spanish	1	4
Finnish	1	0
Hungarian	1	0
Japanese	1	0
Thai	1	0

Table 1: *OpenProDat*: number of speakers.

The participants information sheets were saved as an XML file (see figure 1). This information is attached to recordings.

Speaker: F12

Recording session number: 1

Date: 2013-03-08
 Place: Aix-en-Provence (France) ()
 Setting: H4N (AC power)

Lang: IT Text: T02 Text: T33
 Lang: FR Text: T02 Text: T33
 Lang: EN Text: T02 Text: T33
 Lang: DE Text: T02 Text: T33

Sex: F
 Born: 1980
 Place: Catanzaro (Italy) ()

Places:
 Place: Aix-en-Provence (France) (current)
 Place: Berlin (Germany) 2010 2012 (past)
 Place: (Italy) 1980 2004 (past)

Current position: Researcher
 Education: BAC+8

Spoken languages:
 Lang: IT (level: 5) (frequency of use: 3)
 Lang: FR (level: 3) (frequency of use: 3)
 Lang: EN (level: 3) (frequency of use: 2)
 Lang: DE (level: 3) (frequency of use: 1)
 Lang: ES (level: 1) (frequency of use: 1)

Figure 1: Available participant information sheet.

Moreover, some files were manually transcribed and annotated with SPPAS [01]. These annotations are also freely available in the SLDR repository.

We intend to continue to record new participants. Any new contribution is welcome in the form of:

- new recordings (existing languages or new ones);
- transcriptions;
- annotations.

3. Using the OMProDat database

The tone patterns obtained from the Momel/INTSINT coding of the Korean version of the corpus [16] were compared to those defined in K-ToBI [15], which is regarded as a standard intonation model of Korean. The same corpus was used to evaluate two versions of the Momel algorithm [14], comparing the original version [08] to the improved version described in [10]. The second version of Momel was shown to be qualitatively and quantitatively superior to the earlier version for all 10 speakers and for 38 of the 40 passages analysed. [14]

In [07], a pilot study is described applying the multi-tiered annotation files of the Aix-Ox corpus to compare the intonation of questions in L1 and L2 for English and for French. A number of other applications described in the paper are also currently in progress.

The Chinese version of OMProDat has been used for a preliminary investigation of the third tone Sandhi in standard Chinese [05]. The results tend to support the argument that it is prominence rather than reduction that is one of the factors for the formation of 3rd tone sandhi. The data also support lend support to the idea of a binary foot-like sandhi domain.

The English, French, Chinese native-speaker recordings from the database were used in a cross-language study of the [12, 13]. The examination of a set of pitch-normalised melody metrics for English, French and Chinese, revealed a significant difference between Chinese on the one hand and English and French on the other. In Chinese, pitch movements were found to be larger (mean interval, fall and rise), with greater variability (standard deviation of interval, fall and rise) and are faster (mean slope, rise-slope, fall-slope) than in English and French. For the two European languages there was also a significant gender difference which was not observed for Chinese: female speakers making larger and faster pitch movements than male speakers in English and French.

It was suggested that this effect could be the result of pressure from the lexical tone system of Chinese which restricts the use of pitch for non-lexical functions such as gender distinctions.

4. Perspectives

It is intended that all the corpora included in the database shall be annotated using our automatic annotation tools, and that all the recordings and annotations will be made freely available under an open-database licence as part of *OMProDat*: the open multilingual speech-prosody database.

Linguists and engineers are welcome to download and use the corpora freely. They are kindly requested, in return, to make any additional annotations which they may carry out on the primary data publicly available on *OMProDat*.

5. Acknowledgements

The Korean data was recorded with the support of the *Korean-French Science and Technology Amicable Relationship* (STAR) project, funded by *EGIDE* (a partner of the French Ministry of Foreign Affairs) and the *Korean Foundation for International Cooperation of Science and Technology*.

The English and French data was recorded with the support of an *ALLIANCE PHC* (Partenariat Hubert Curien) project, funded by the *British Council* and *EGIDE*.

The Chinese data was recorded with the support of the *Innovation Program of Shanghai Municipal Education Commission* (12ZS030) and with the funding to the first author for the 985 project of the *School of Foreign Languages of Tongji University*, Shanghai.

Our thanks to Minhwa Chung, Sunhee Kim, Greg Kochanski, So-Young Lee, Anastassia Loukina, Qiuwu Ma, Anne Tortel, Hyunji Yu for their help with these different projects.

6. References

- [01] Bigi, B. and Hirst, D. J. "Speech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody". In *Proceedings of the 6th International Conference on Speech Prosody*, May 2012.
- [02] Boersma, P. and Weenink, D. "Praat, a system for doing phonetics by computer". <http://www.praat.org> [version 5.3.41, February 2013], 1992 (2013).
- [03] E. Campione and J. Véronis "A multilingual prosodic database". In *Proceedings of ICSLP'98*, Sidney, Australia. 1998.
- [04] Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouroupoulos, J.; Senia, F.; Trancoso, L.; Veld, C. and Zeiliger, J. "Eurom - a spoken language resource for the EU". In *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, 1, 867–870, Madrid., 18-21 September 1995.
- [05] Ding, D. and Hirst, D.J. "A preliminary investigation of third-tone sandhi in Standard Chinese with a prosodic corpus". *8th International Symposium on Chinese Spoken Language Processing*, Hong Kong 2012.
- [06] Erjavec, T. "MULTEXT-East Version 3: Multilingual morphosyntactic specifications, lexicons and corpora". *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal: 1535-1538. [available at <http://nl.ijs.si/ME/>] 2004
- [07] Herment, S., Tortel, A., Bigi, B. Hirst, D., and Loukina, A. "AixOx: A multi-layered learners corpus: automatic annotation". *4th International Conference on CorpusLinguistics*, Jaën, Spain, (forthcoming in Díaz Pérez, J. and Díaz Negrillo, A. (eds.) *Specialisation and variation in language corpora*, Peter Lang.) 2012.
- [08] Hirst, D.J. and Espesser, R. "Automatic modelling of fundamental frequency using a quadratic spline function". *Travaux de l'Institut de Phonétique d'Aix*, 15: 75–85, 1993.
- [09] Hirst, D.J., "Pitch parameters for prosodic typology. A preliminary comparison of English and French". In *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelona, 2003.
- [10] Hirst, D.J. "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation". In *Proceedings of the XVIth International Conference of Phonetic Sciences*: 1233–1236, Saarbrücken, 2007.
- [11] Hirst, D.J. "The analysis by synthesis of speech melody: from data to models". *Journal of Speech Sciences*, 1(1): 55–83, 2011.
- [12] Hirst, D.J. "The automatic analysis by synthesis of Speech Prosody with preliminary results on Mandarin Chinese". *8th International Symposium on Chinese Spoken Language Processing*, Hong Kong, [Invited keynote lecture]. 2012.
- [13] Hirst, D.J. "Melody metrics for prosodic typology: comparing English, French and Chinese". *Proceedings Interspeech*, Lyon August. 2013 (submitted)
- [14] Hirst, D.J.; Cho, H.; Kim, S. and Yu, H. "Evaluating two versions of the Momel pitch modeling algorithm on a corpus of read speech in Korean". In *Proceedings of Interspeech VIII*, Antwerp, Belgium: 1649–1652, 2007.
- [15] Jun, S.-A. "K-ToBI (Korean ToBI) labeling conventions: Version 3.1". *UCLA Working Papers in Phonetics* 99. pp.149-173. 2000.
- [16] Kim, S.-H.; Hirst, D.J.; Cho, H.-S.; Lee, H.-Y. and Chung, M.-H. "Korean Multext: A Korean prosody corpus". In *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil., 2008.
- [17] Komatsu, M. "Chinese MULTEXT: recordings for a prosodic corpus". *Sophia Linguistica*, 57:359–369, 2009.
- [18] Roach, P.; Arnfield, S. and Hallum, E. "BABEL: A multi-language speech database". In *Proceedings of SST-96: Speech and Science Technology Conference*, Adelaide: 351–4, 1996.
- [19] Shigeyoshi, K.; Tatsuya, K.; Kazuya, M. and Toshihiko, I. "Preliminary study of japanese MULTEXT: a prosodic corpus". In *Proceedings of ICSLP 2001*, 2001.
- [20] Shigeyoshi, K.; Kiriya, S.; Toshihiko, I. and Campbell, N. "Japanese MULTEXT: a prosodic corpus". *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal: 2167-2170. 2004.
- [21] Véronis, J.; Hirst, D.J. and Ide, N. "NL and speech in the MULTEXT project". In *Proceedings of AAAI Workshop on Integration of Natural Language and Speech*, Seattle, USA: 72–78, 1994.

Aix MapTask: A new French resource for prosodic and discourse studies

*Ellen Gurman Bard¹, Corine Astésano^{2,3}, Mariapaola D'Imperio³, Alice Turk¹,
Noël Nguyen³, Laurent Prévot³, Brigitte Bigi³*

¹The University of Edinburgh, English and Language Studies &
Human Communication Research Center, Edinburgh, United Kingdom

²Toulouse Université and CNRS, Octogone, Toulouse, France

³LPL, CNRS, Aix-Marseille Université, Aix-en-Provence, France

ellen@ling.ed.ac.uk, corine.astesano@univ-tlse2.fr,
mariapaola.dimperio@lpl-aix.fr, turk@ling.ed.ac.uk, noel.nguyen@lpl-aix.fr,
laurent.prevot@lpl-aix.fr, brigitte.bigi@lpl-aix.fr

Abstract

This paper introduces the Aix MapTask corpus. This corpus was modelled after the original HCRC Maptask. Lexical material selection has been carefully crafted for speech and prosodic analysis [1]. We present the design of the lexical material, the protocol and basic quantitative facts about the existing corpus. We also describe an additional face-to-face condition now being collected. Finally, we explain how the material has been transcribed and processed.

Index Terms: corpus, maptask, French

1. Introduction

Leading on from pioneering work on communicative skills [2], the Map Task protocol had been designed in Edinburgh with the HCRC Map Task corpus [3]. The usefulness of the data produced with this protocol has led many teams to create their own Map Task corpora on various languages including Italian (different varieties), Japanese or Occitan. However, until now no Map Task Corpus was available for French.

Map Task corpora are interesting in particular because they can be simultaneously well controlled (in terms of lexical material, difficulty of the task, participant pairings....), while allowing genuine spontaneous speech exhibiting all the phenomena of speech production (pauses, disfluencies, etc.). The lack of Map Task for the French language is therefore, at a general level, a missing element for approaching speech and discourse in French and comparing certain phenomena across languages. Moreover, some of the authors wanted to investigate the findings from previous work [1] on spontaneous speech and the Map Task protocol was the perfect one for achieving this goal.

Thanks to EU Marie-Curie funding, the corpus was recorded and transcribed in 2002. With additional funding from ANR projects PhonIACog¹ and CoFee² [4] it has been developed for further use. We have gathered data and metadata and archived it in the Ortolang speech and language repository³. We are now working on a new set of recordings in a face-to-face condition.

¹<http://aune.lpl.univ-aix.fr/~phoniacog/>

²<http://cofee.hypotheses.org/>

³The corpus itself is archived at <http://sldr.org/sldr000732>

The present paper sets out the experimental design of the corpus (Section 2), explains how it has been processed (Section 3) and provides some quantitative information (Section 4) before introducing ongoing work and planned research (Section 5).

2. Lexical Material and Design

2.1. Lexical Material

The critical lexical material used for the Aix Map Task is a subset of the material used in [1]. That corpus consists of syntactically ambiguous sentences which prosodic cues (namely boundaries, Final Accent –FA– and Initial Accent –IA–) help to disambiguate. Syntactic ambiguity is created by manipulating adjective scope as in ‘les gants et les bas lisses’, where the adjective (A) ‘lisses’ either qualifies:

1. the second noun ‘bas’ (N2) only: [les gants][et les bas lisses], with an intermediate phrase (ip) boundary (B2) between N1 and N2, and a word (w) boundary (B5) between N2 and A (hereafter Case 1 or C1);
2. or the two nouns ‘gants et bas’ (N1 and N2): ([les gants et les bas][lisses], with an ip boundary (B5) between N2 and A, and an accentual phrase (ap) boundary (B2) between N1 and N2 (hereafter Case 2 or C2).

The manipulation of adjective scope thus yields 4 sites of interest (C1-B2 ; C1-B5 ; C2-B2 ; C2-B5) for observing indications of prosodic boundaries via FA and IA (see Figure 1).

The prosodic structure is also manipulated with regard to constituents’ length, nouns and adjectives ranging from one to four syllables, in all possible combinations (eg. ‘les gants et les bas lisses’ vs. ‘les bonimenteurs et les baratineurs fabulateurs’).

Results from [1] showed that IA was a consistent marker of structure. More than its ‘classic’ rhythmic role as a marker of long stretches of speech, IA was shown to preferentially be used as a marker of constituency over FA, especially at the minor-phrase (ap) level, thus clarifying its role and putting it at the centre of the prosodic description of French.

As a follow-up to this first study on controlled speech, we wanted to test whether IA’s role as structure marker would apply in more spontaneous speech. We ask whether IA will still

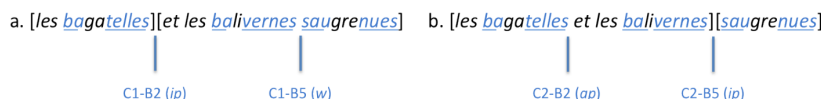


Figure 1: The 4 prosodic sites of interest. Underscored syllables are where FA and IA potentially can occur to mark prosodic structure.

be elicited as a structure marker in dialogue, for the same controlled target words and phrases as were used in the previous study. A subset of the corpus noun phrases was thus chosen to be represented within a Map Task design for semi-guided speech. Our goal is to compare IA occurrence in guided dialogues with our previous results on read speech. The target words and phrases chosen to appear on the maps are described below.

2.2. Material Design

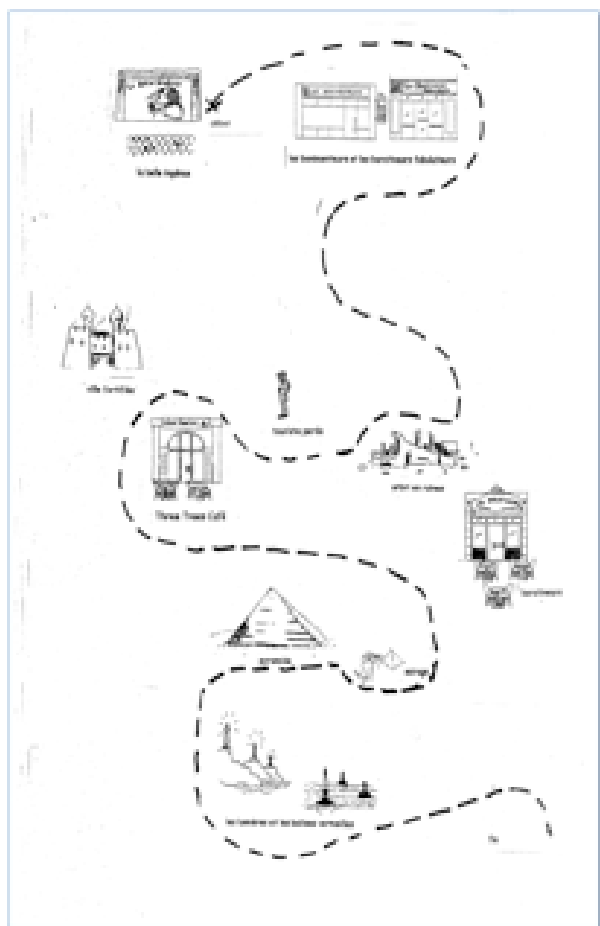


Figure 2: Instruction Giver's map includes a route.

To elicit spontaneous speech forms for comparison with the read speech examples used in our earlier work, we collected and transcribed a corpus of task-oriented dialogues, following the general method used in the HCRC Map Task Corpus [3]. In this task two players collaborate to reproduce on the map before one of them the route drawn on one of the player's maps (Figures 2 and 3). Neither can see the other's map. They know

that the maps describe the same features but that some details may differ. In fact, the maps differ in alternate route-critical landmarks, so that discussion of the mismatches is common. To hold prior experience constant, the maps are of imaginary places and the players proceed through a series of different maps in a way that balances their experience in Instruction Giver and Instruction Follower roles. Whatever their assigned roles, players are allowed to say anything necessary to accomplish their communicative goals. However, since the two speakers cannot see each other, gestures would be ineffective.

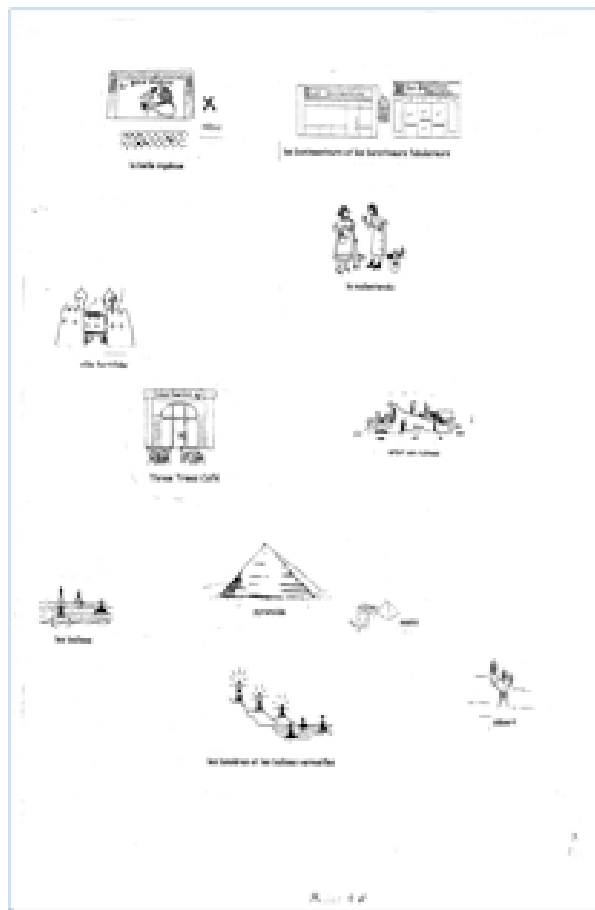


Figure 3: Instruction Follower's map does not include a route.

The maps were designed around labelled cartoon landmarks, the names of which gave us the freedom to elicit nominals of the desired structure. So as not to make the names suspiciously alike, we included materials for several experiments by colleagues – on /r/ placement and on final high vowels. The critical landmark names, however, were chosen from the conjoint noun phrases of our read materials as described above. Thus there was one set of lighthouses at the top of a cliff with tall

buoys tilting on a choppy sea (*les lumières et les balises vertigineuses*: broad adjective scope) and another sketch with light-houses at sea level and tall buoys (narrow adjective scope). The length of the second noun and of the following adjective were varied to see whether IA was encouraged by longer phonological words (N) or by longer phrases (N + Adj). In half the cases, the two players' versions of a landmark matched in adjective scope, but elsewhere they did not. To make the two Ns necessary to the naming process, single exemplars (a lone lighthouse) were also found on maps.

Sixteen pairs of speakers performed 8 map tasks each. Digital channel per-speaker stereo recordings were made in studio conditions and transcribed by native speakers of French.

The distribution of target words (and therefore landmarks) across the maps is set out in Table 3. In this table, each cell corresponds to a map. The condition $IG = IF$ means that the corresponding landmark has the same scope on follower and giver's maps ($IG \neq IF$ if not). More precisely, the conditions are Broad Broad, Narrow Narrow, Broad Narrow and Narrow Broad. Finally the four colors correspond to the four dyads of participants (each dyad had 8 maps to communicate and participants switched role after 4 maps).

3. Transcribing and Processing the data

When a speech corpus is transcribed into a written text, the transcriber is immediately confronted with the following question: How to reflect the reality of oral speech in a corpus? Conventions are thus designed to provide a set of rules for writing speech corpora. These conventions establish which phenomena have to be annotated and also how to annotate them.

The corpus was transcribed in standard French orthography, using Transcriber [5]. The transcription includes short pauses, truncated words and hesitations.

SPPAS is a tool to produce automatic segmentations from a recorded speech sound and its transcription [6]. The resulting segmentations are represented in a set of TextGrid files, the native file format of the Praat software [7].

SPPAS tools and resources are currently available under the GNU Public License, at the URL:

<http://www.lpl-aix.fr/~bigi/sppas/>

SPPAS generates separate TextGrid files for utterance, word, syllable, and phoneme segmentations. (i) utterance segmentation, (ii) word segmentation, (iii) syllable segmentation and (iv) phoneme segmentation. An example of SPPAS output is represented in Figure 4.

4. Quantitative aspects

The combination of the four dyads explaining each other 8 maps (alternating roles) provides 32 dialogues of an average duration of 6 minutes 52 seconds. The corpus includes about 50 000 tokens but with a vocabulary size of only 1500 different forms.

As can be seen in table 1, which lists the corpus frequencies the 30 most frequent words with their number of occurrences, other than function words, feedback items (*ouais*, *mh*, *d'accord*, *voilà*, *oui*, *non*) are well represented. Words related to space are also extremely frequent (*gauche* / left, *droite* / right, *vers* / towards, *vas* / go³, *sur* / on, *dessous* / under).

³Since the corpus has not been lemmatized yet, this table may not give other spatial verbs their true rank.

2389	tu	703	à	347	du
1146	la	643	les	337	vas
1077	ouais	507	un	332	ai
1032	euh	494	gauche	331	voilà
1017	de	446	d'accord	325	sur
927	et	438	donc	321	c'est
815	le	420	droite	317	oui
812	mh	414	as	314	j'
800	je	397	là	310	non
704	en	375	vers	293	dessous

Table 1: Most frequent forms with number of occurrences

word	occurrences	mean duration (sec)
balises	187	0.3538
lumières	120	0.3675
bonimenteurs	105	0.6777
baratineurs	82	0.5809
vertigineuses	81	0.6823
fameux	45	0.4388
vermeilles	41	0.4440
fabulateurs	23	0.7513

Table 2: Target words with number of occurrences and average duration

Table 2 displays the number of occurrences of target words and their average durations. Due to their presence on all the maps, the nouns are used more often than the adjectives. Noun frequencies are however difficult to interpret at this early stage. The spread of frequencies for the adjectives is bigger with *vertigineuses* occurring about four times more than *baratineurs*. This is however difficult to explain without analyzing the dialogues more deeply. For example, though all the design-critical landmarks were also route-critical, some landmarks may in the end have been less useful or easier to use than others, yielding less discussion and resulting in a lower frequency of occurrence.

Finally the target phrases occurrences were:

- les balises [euh] vermeilles 25
- les balises [pause] vertigineuses 42
- les baratineurs fameux 15
- les baratineurs [pause] fabulateurs 13

5. On-going work and perspectives

We are currently working in two directions on this corpus: (i) shallow natural language processing for extracting more linguistic information of the corpus, (ii) constitution of an additional condition with participants seeing one another as they work. Natural language processing includes POS tagging with probabilistic tagger trained on written and spoken data [8], lemmatizing and chunking. The additional condition is basically a replication of the existing corpus but allowing participants to see each other (but of course not each other's maps). The recordings are once more performed in an anechoic room with a high-quality headset and 3 cameras (1 for each participant and general one) following a technical setting already used in the lab [9].

At this stage, the main studies planned include the annotation of feedback items (which are extremely frequent as it can

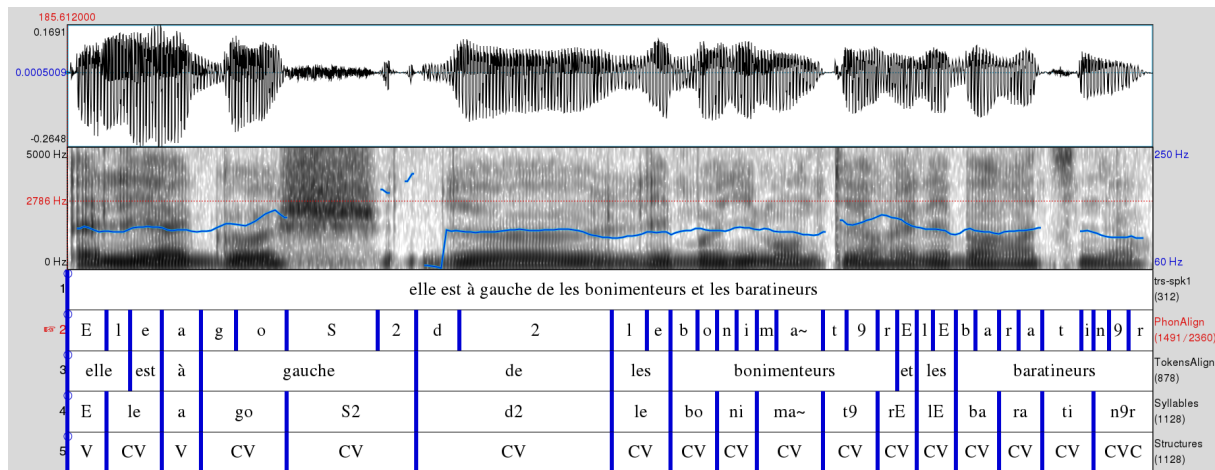


Figure 4: SPPAS output example.

be seen in table 1) in the context of a broader study on multi-modal feedback, which will focus on the different distributions of feedback between verbal and visual modalities with the absence or presence of a visual channel.

A second study planned consists in comparing compare IA occurrence on guided dialogues with previous results on controlled speech [1]. The number of occurrences (Table 2) of the target words are encouraging in this respect .

6. Acknowledgements

This work was funded by a European Community Marie Curie Fellowship award to Corine Astésano (HPMF-CT-2000-00623). We are grateful for the collaboration, assistance and support of colleagues the Laboratoire Parole et Langage, Bernard Teston, Cheryl Frenck- Mestre, Mariapaola d'Imperio, Robert Espesser, Louis Seimandi, Annelise Coquillon, Ludovic Jankowski and from the University of Edinburgh, Eddie Dubourg and Ziggy Campbell.

7. References

- [1] C. Astésano, E. Bard, and A. Turk, "Structural influences on initial accent placement in french," *Language and Speech*, vol. 50, no. 3, pp. 423–446, 2007.
- [2] A. Anderson, G. Brown, R. Shillcock, and G. Yule, *Teaching talk: Strategies for production and assessment*. Cambridge University Press New York, 1984.
- [3] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The HCRC Map Task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.
- [4] L. Prévot and R. Bertrand, "CoFee-toward a multidimensional analysis of conversational feedback, the case of French language," in *Proceedings of the Workshop on Feedback Behaviors*, Stevenson, 2012, (poster).
- [5] C. Barras, E. Geoffrois, Z. Wu, , and M. Liberman, "Transcriber: a free tool for segmenting, labeling and transcribing speech," in *Language Resources and Evaluation Conference*, Granada (Spain), 1998, pp. 1373–1376.
- [6] B. Bigi, "SPPAS: a tool for the phonetic segmentation of speech," in *Language Resource and Evaluation Conference*, Istanbul (Turkey), 2012, pp. 1748–1755, ISBN 978-2-9 517 408-7-7.
- [7] P. Boersma, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5, no. 9/10, pp. 341–345, 2001.
- [8] S. Rauzy and P. Blache, "Un point sur les outils du LPL pour l'analyse syntaxique du français," in *Actes du workshop ATALA 'Quels analyseurs syntaxiques pour le français ?'*, Paris, France, 2009, pp. 1–6.
- [9] J. Saubesty, "L'activité gestuelle en situation d'incompréhension," 2013, Master Thesis, Aix Marseille Université.

			noun 4-syllables + adjective 2-syllables <i>les bonimenteurs et les baratineurs fameux</i>				noun 4-syllables + adjective 4-syllables <i>les bonimenteurs et les baratineurs fabulateurs</i>			
			IG=IF		IG≠IF		IG=IF		IG≠IF	
			BB	NN	BN	NB	BB	NN	BN	NB
noun 2-syllables + adjective 2-syllables <i>les lumières et les balises vermeilles</i>	IG=IF	BB			Ai1	Aiii1			Bi6	Biii6
		NN			Aii1	Aiv1			Bii6	Biv6
	IG≠IF	BN	Av5	Avii5			Bv2	Bvii2		
		NB	Avi5	Aviii5			Bvi2	Bviii2		
noun 2-syllables + adjective 4-syllables <i>les lumières et les balises vertigineuses</i>	IG=IF	BB			Ci3	Ciii3			Di8	Diii8
		NN			Cii3	Civ3			Dii8	Div8
	IG≠IF	BN	Cv7	Cvii7			Dv4	Dvii4		
		NB	Cvi7	Cviii7			Dvi4	Dviii4		

Table 3: Map composition in terms of landmarks (target words)

A Taiwan Southern Min spontaneous speech corpus for discourse prosody

Sheng-Fu Wang, Janice Fon

Graduate Institute of Linguistics, National Taiwan University, Taiwan
10617 Le-xue Building, No.1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C.)
sftwang0416@gmail.com, jfon@ntu.edu.tw

Abstract

This paper presents a Taiwan Southern Min (Taiwanese) spontaneous speech corpus primarily constructed and annotated for studying discourse prosody. The corpus contains monologue-like speech elicited from interviews. Eight hours of speech contributed by sixteen interviewees, evenly split by gender and age, have been transcribed and annotated. Transcription and the recordings were aligned at the level of syllable with the aid of EasyAlign (Goldman, 2011). Discourse annotation was done by identifying one-verb clausal units and labeling the strength of unit transitions to show the hierarchical structure of discourse using Grosz and Sidner's model (1986). As for prosodic labeling, two major levels of prosodic breaks were identified, along with truncation and prolongation caused by disfluencies and hesitation. The present state of the corpus allows for research on the relationship between acoustic cues, prosodic structure, and discourse organization in unscripted speech.

Index Terms: discourse, spontaneous speech, Taiwanese, Southern Min, prosodic break

1. Introduction

Natural speech is inherently variable. The construction of spontaneous speech corpora is one way of approaching such fascinating variability. This kind of corpora that contain speech collected by a more natural setting provides a greater range of variability than a reading task or other kinds of simple but designed laboratory setting. Especially in studying the relationship between natural discourse and speech, spontaneous speech is able to reveal phenomena that would otherwise be unavailable to researchers.

The paper presents a spontaneous Taiwan Southern Min speech corpus constructed with the aforementioned objectives in mind. Another important object concerns the target language. Taiwan Southern Min, more commonly known as Taiwanese, is the native language of approximately 70% of the population in Taiwan [1]. Because of the policy that promoted Mandarin Chinese as the official language, Taiwan Southern Min was marginalized in the realm of education, media, and administration. A consequence in linguistic studies was that language resources such as annotated speech corpora in Taiwan Southern Min have been very scarce as compared with resources in Mandarin Chinese. The construction of a spontaneous speech corpus will certainly be an important step to further understandings on relevant theoretical issues on Taiwan Southern Min, as well as providing an important resource for applicational purposes such as speech synthesis and speech recognition.

2. Corpus Construction

The corpus is the continuation of a Mandarin-Min bilingual spontaneous speech corpus started in 2004 [2]. Currently in the annotated dataset, there are eight hours of monologues contributed by sixteen native speakers of Taiwan Southern Min. These speakers can be grouped according to gender and age. For the age of the speakers, the young group contains speakers born in the 1980s and the old group contains speakers born in the 1940s. All of the speakers were from the same region (Taichung, the mid-Taiwan metropolitan area) so that research claims based on the corpus would not be confounded by dialectal differences.

Speech was elicited in the form of an interview in which the interviewer asked the interviewee to talk about his or her personal experiences in childhood, in school, or at work. Marriage, health, and traveling experiences were also common topics. The aim of the interviewer was to elicit longer monologues rather than to engage in a conversation with the interviewee. Whenever the interview became loaded with short turn exchanges between the interviewer and the interviewee, the recording would be not be included in the present dataset.

The currently annotated dataset contains 110873 syllables, 10603 discourse boundaries, and 19433 prosodic breaks. The annotation conventions for discourse and prosodic break will be described in Section 3.

2.1. Transcription

The recordings were transcribed with the Taiwanese-Romanization convention mainly based on the online dictionary constructed by Iuⁿ [3]. The convention uses both Chinese characters and romanized transcription. Chinese characters are used when it is possible to identify the source characters for a given Southern Min expression. When the characters were not readily identifiable, romanization was used instead.

When the transcription was aligned with the recording in Praat [4], all Chinese characters were romanized. The particularities of the romanization includes using the *h* symbol for plosive aspiration (e.g., *pha* for /p^ha/), double *n* for nasalized vowels (e.g., *pinn* for /pi/), and *h* for glottal stops at the coda position (e.g., *peh* for /peʔ/).

Since Taiwan Southern Min is a tone language, information on lexical tones is provided in the annotation. In addition, since lexical tones in Taiwan Southern Min may be realized as a fixed low tone [5], and the occurrence of such a low tone is not predictable from the transcription, further effort was devoted in identifying syllables with the low tone.

2.2. Syllable alignment

Syllable segmentation and alignment were first automatically conducted with the EasyAlign plug-in [6] and were further manually checked by the first author. Since investigation on the relationship between syllable duration and other linguistic parameters was a primary goal for research based on this corpus, it was crucial to determine the criteria for deciding syllable boundaries, especially at the utterance-final positions. The major criterion taken by the labeler was to put the syllable boundary at the decay of formant structures as shown on Praat spectrograms.

A second trained phonetician labeled 10% of the data and a test of cross-labeler agreement was conducted. Results showed a mean difference of 12.4 ms in terms of the alignment of syllable boundaries, which amounts to 5.7% of mean syllable duration. In addition, 91% of the alignment differences are smaller than 20% of mean syllable duration. This degree of agreement is on par with, or even slightly better as compared with similar tests for the Switchboard Corpus [7], which reports a mean phone-alignment difference of 16.4 ms (19% of mean phone duration), and of the Buckeye corpus [8], which reports that 75% of the phone-alignment differences are smaller than 20% of phone duration.

3. Annotation

3.1. Discourse annotation

Discourse segmentation was done with the transcribed texts of the recordings. The annotation was based on Fon's [9] adaptation of Grosz and Sidner's [10] model on "Discourse Segment Purpose", which identifies the basic intentional units that structured in a hierarchical fashion. In practice, the texts were first segmented into basic discourse units, which are clauses, defined as units that contain one verb, according to the definition of "a simple clause" by a classical study on the functional grammar of Mandarin Chinese [11]. Next, the relationship between clauses was judged according to the level of discourse juncture. Four different levels of "Discourse Boundary Indices (DBI)" were distinguished in this study.

The first level was DBI0, which suggests that the two adjacent clauses describe the same entity or event, thus the boundary between these two is merely a clausal boundary. In the corpus, DBI0 is often used in the following situations: between a matrix and a subordinate clause (Examples 1 and 2), two clauses showing parallel syntax (Example 3), between a tag question and its preceding clause (Example 4), between clauses sharing an anaphora (Example 5), and the boundary between two simple discourse units having the relation of cause-effect or topic-comment (Example 6).

- (1) [伊講]_{DBI0} [“我嘛欲來”]
[he said]_{DBI0} [‘I also want to come’]
- (2) [這我感覺]_{DBI0} [足歡喜ê]
[it makes me feel]_{DBI0} [very happy]
- (3) [可能按呢彼個m著]_{DBI0} [啊彼個著按呢]
[perhaps that was right]_{DBI0} [still that was right]
- (4) [咱也無法度kah伊陪伴伊一世人啊]_{DBI0} [著無?]
[we can't be on her side for her whole life/DBI0 right?]
- (5) [伊其實本來著無心欲買]_{DBI0} [只是入來覷]
[she didn't really want to buy]/_{DBI0} [(she) just came in and looked around]

- (6) [對以前ê查某來講翁婿著是家己ê天]_{DBI0}
[所以一定愛結婚]
[“a husband was like the sky for women in the past”]_{DBI0}
[so getting married was a must]

DBI1 refers to the juncture around which the clauses describes about different subtopics or "scenes" within a theme or an episode in narration. In the corpus, DBI1 is often used to label the following cases: an anaphoric change or update (Example 7), a change of aspect or the introduction of a new time reference (Example 8), comments (Example 9), and the boundary between a complex discourse unit to a simple or another complex discourse unit where the units between the boundary have the relation of cause-effect or topic-comment (Example 10).

- (7) [因為阮哥哥他們攏讀到高職]_{DBI0}
[讀了專科讀了]_{DBI0}
[就攏去做工作啊]_{DBI1}
[阿本來阮阿公的意思是講]...

“[because my brothers they went to the vocational high school]_{DBI0}
[after graduating from the vocational high school]_{DBI0}
[(they) went to work]_{DBI1}
[and originally my grandpa's intention was]...”

- (8) [若講交朋友啦]_{DBI0}
[我其本上都是專門靠he lah]_{DBI1}
[基本上啊我讀二專ê時陣]...

“[when talking about making friends]_{DBI0}”
[basically I particularly depended on that]_{DBI1}
[basically, when I was studying at the two-year technological college]”

- (9) [假那鳥仔放出籠leh]_{DBI1} [足歡喜]
[just like birds out of their cage]_{DBI1} [very happy]
- (10) [但是我是帶ti農家]_{DBI1}
[所以對這leh採棉ê空課嘛lóng真真熟]_{DBI0}
[ah實在有落去做]

“[but I lived at the farm]_{DBI1}
[so I was quite familiar with the work concerning the silkworms]_{DBI0}
[I actually did it]”

DBI2 referred to situations where the boundary clearly differentiates two themes or episodes, yet these themes and episodes are still within a bigger general topic. Example 11 showed one such switch, in which the speaker was still narrating the experiences working in a hospital but changed from the sentiment of the fragility of human-beings to the description of how long he had worked there.

DBI3 was an additional label for handling radical shifts of themes which may be considered boundaries of totally different pieces of monologue or interview within the same recording. This label was often used when the interviewer directed the interviewee to another totally unrelated topic. Example 12 shows a rare case where the jump between topics were initiated by the speaker herself, as she switched from describing her son's working experience to her happy days as a young girl.

- (11) [彼陣都一種感覺啦]_{DBI0}
 [感覺講]_{DBI0}
 [人原仔是真<MAN 脆弱 MAN>按呢]_{DBI2}
 [啊彼陣佇遐做做做]_{DBI1}
 [我會記ê做一兩年吧]

“[at that time there was a feeling]_{DBI0}
 [(I) felt]_{DBI0}
 [that human beings are very fragile]_{DBI2}
 [at the time I worked here (for some time)]_{DBI1}
 [I remember it was a year or two]”

- (12) [有啊]_{DBI0}
 [逐工轉來啊]_{DBI1}
 [lóng 愛足暗才轉來eN]_{DBI3}
 [我會感覺哦]_{DBI0}
 [查某gín仔時代eh足快樂ê]

[yes]_{DBI0}
 [(he; the speaker's son) came back everyday]_{DBI1}
 [always came back until it's very late]_{DBI3}
 [I feel that]_{DBI0}
 [(I) was very happy when I was a young girl]

The labeling described above is believed to be able to reflect the hierarchical organization of discourse units. The resulting labeling on discourse structure was subsequently annotated and aligned with recordings using Praat [4]. The first author labeled all of the data. A second labeler labeled two of the sixteen transcription files, and the agreement rate was 85%. The discrepancies were discussed and the rest of the labeling were rechecked accordingly by the first author.

3.2. Prosodic annotation

Prosodic units was also annotated in with a ToBI-style system of prosodic breaks. Although there is an prosodic annotation framework of Taiwanese available [12], in their proposal, shown in Table 1, the level below the intonation phrase (IP) is the Tone Sandhi Group (TSG), whose occurrence is defined by rule-governed tonal alternations. This kind of definition is very different from what is commonly used for defining or describing a level of prosodic unit, such as the perception of a certain pitch movement or acoustic cues such as final lengthening and final pitch-lowering[13, 14, 15]. Also, the occurrence of TSG boundaries can almost regularly be predicted from syntax [16, 17], with a very low degree of freedom if the syntactic structure is to be held unaltered. It makes the inclusion of TSG as a level of prosodic phrasing doubtful.

Table 1: TW-ToBI break indices [12]

b4	intonation phrase boundary, either utterance-finally or -medially
b3	tone sandhi group (TSG) boundary
b3m	percept of TSG boundary without sandhi tone
b2m	base tone without percept of the TSG ending
b2	ordinary “word-internal” syllable boundary
b1	resyllabification
b0m	syllable fusion

Thus, a new labeling scheme was devised, as shown in Table 2. This new scheme focuses on two things: The first focus is the differentiation of the intonation phrase and a lower level

of prosodic unit. This differentiation makes sure that discourse boundaries corresponding to the same level of prosodic boundaries are investigated. The second focus is the annotation of hesitation, truncation, and abrupt stops.

Table 2: A proposed Taiwan Southern Min BI tagset

4	intonation phrase boundary, either utterance -finally or -medially
4-	utterance-final boundary without obvious IP boundary cues
3	a lower level of boundary, utterance-medially
3p	hesitation that gives the percept of a major prosodic boundary
2p	hesitation that does not give the percept of a major prosodic boundary
1p	truncation and abrupt stop

Figure 1 and Figure 2 show examples of a level 4 and level 3 break in the proposed break index framework. The difference between these two types of breaks can be clearly seen from Figure 2, where the syllable at level 3 break has lowered F0 not lengthened as a level 4 break does. Figure 3 shows a prolonged syllable perceived as hesitation, labeled with 3p. Figure 4 labeled two cases of 1p: the former showed an abrupt stop followed by an immediate repair, and the latter showed a segmental deletion.

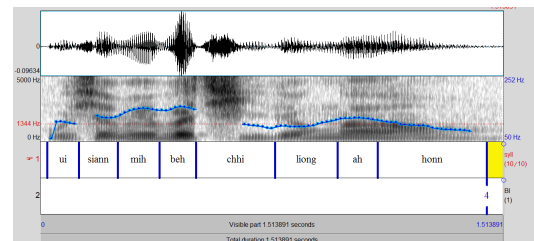


Figure 1: A BI4 (intonational phrase boundary) example; “the reason why I raise fish”

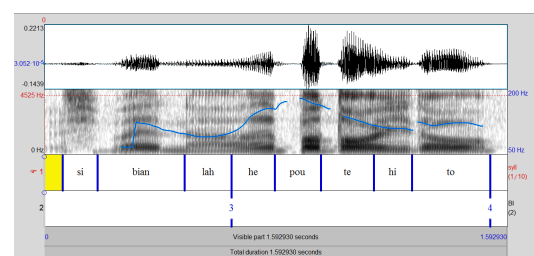


Figure 2: A BI3 boundry (tentative label for a prosodic break smaller than an intonational phrase boundary but bigger than word boundary) followed by a BI4 boundary; “No, the glove puppetry”

A second labeler annotated the prosodic breaks of 10% of the corpus and the agreement with the author was calculated with the kappa statistic [18]. The kappa statistic on boundary placement was 0.86. When both labelers agreed on placing a BI, the kappa statistic for BI category agreement was 0.6. These values are comparable with the report on the interlabeler agreement on ToBI labeling on the Switchboard corpus [19], which yielded a kappa statistic of 0.75 for pitch accent placement, 0.67

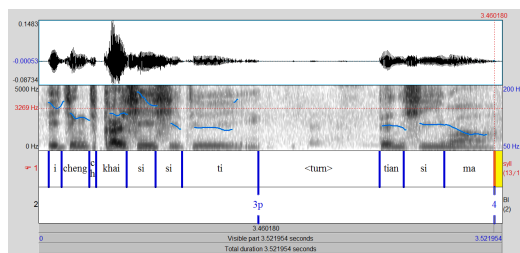


Figure 3: A BI3p boundary that labeled prolongation perceived as hesitation; "in the past it was on..... TV"

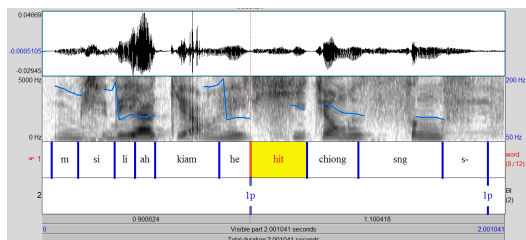


Figure 4: Two 1p boundaries that labeled abrupt stop and segmental deletion: "Not the salted plum; that kind of..."

for phrasal accent placement, 0.58 for boundary tone placement, and 0.51 for pitch accent choice, and 0.48 for phrasal accent choice. The only statistic that exceeds our current agreement rates was the kappa value for boundary tone choice (0.71). [20] also examined interlabeler agreement on Break Index in the ToBI framework, which yielded a kappa value of 0.75 for phrasal boundary placement and 0.67 for phrasal boundary size, slightly exceeding the value obtained in the current study. Table 3 presents the agreement matrix of break indices.

Table 3: Agreement matrix of break indices (Column headings indicate labels assigned by the author and row headings are labels assigned by the second labeler)

	1p	2p	3	3p	4	NO	Sum
1p	189	10	6	10	78	64	357
2p	0	18	10	10	5	23	66
3	5	10	61	1	49	81	207
3p	1	30	5	153	55	13	257
4	43	5	44	25	868	103	1088
NO	17	8	91	2	34	8226	8378
Sum	255	81	217	201	1089	8510	10353

4. Conclusions & Prospects

The present state of the corpus allows for research on the relationship between acoustic cues, prosodic structure, and syntactic/discourse structure. In addition to further annotation on larger amounts of data, more dimensions of annotation such as POS tagging and the identification of Tone Sandhi Group boundaries will also be implemented to make the corpus a more valuable resource for a wider varieties of research topics on Taiwan Southern Min.

5. References

- [1] Government Information Office, *The Republic of China Yearbook*

2012. Kwang Hwa Pub. Co., 2012.

- [2] J. Fon, "A preliminary construction of taiwan southern min spontaneous speech corpus," Tech. Rep. NSC-92-2411-H-003-050, National Science Council, Taiwan, 2004.
- [3] U.-G. Lu, "Taiwen huawen xianshangzidian jian-zi jishu ji shi-yongcixing tantao [Construction and utilization of Taiwanese-Chinese online dictionary]," in *Proceedings of the 3rd International Conference on Internet Chinese Education*, pp. 132–141, 2003.
- [4] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program], version 5.1. 44," 2010.
- [5] U.-J. Ang, *Taiwan Helaoyu shengdiao yanjiu (Research on Taiwanese Tones)*. Taipei: Zili wanbao, 1985.
- [6] M.-H. Chen, J.-P. Goldman, H.-h. Pan, and J. Fon, "Easyalign: an automatic phonetic alignment tool under praat," in *Proceedings of the Workshop on New Tools and Methods for Very-Large-Scale Phonetics Research*, vol. 2011, pp. 109–112, 2011.
- [7] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *International Conference on Spoken Language Processing*, pp. S32–S35, Citeseer, 1996.
- [8] W. Raymond, M. Pitt, K. Johnson, E. Hume, M. Makashay, R. Dauricourt, and C. Hilts, "An analysis of transcription consistency in spontaneous speech from the buckeye corpus," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 1125–1128, 2002.
- [9] Y.-J. J. Fon, *A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech*. PhD thesis, The Ohio State University, 2002.
- [10] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational linguistics*, vol. 12, no. 3, pp. 175–204, 1986.
- [11] C. Li and S. Thompson, *Mandarin Chinese: A functional reference grammar*. Univ of California Pr, 1981.
- [12] S. Peng and M. Beckman, "Annotation conventions and corpus design in the investigation of spontaneous speech prosody in taiwanese," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- [13] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," *Prosodic typology: The phonology of intonation and phrasing*, pp. 9–54, 2005.
- [14] S. Godjevac, "Transcribing serbo-croatian intonation," *Prosodic typology: The phonology of intonation and phrasing*, pp. 146–171, 2005.
- [15] J. J. Venditti, "The J-ToBI model of Japanese intonation," *Prosodic typology: The phonology of intonation and phrasing*, pp. 172–200, 2005.
- [16] R. L. Cheng, "Tone sandhi in taiwanese," *Linguistics*, vol. 6, no. 41, pp. 19–42, 1968.
- [17] M. Chen, "The syntax of xiamen tone sandhi," *Phonology Yearbook*, vol. 4, pp. 109–149, 1987.
- [18] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.
- [19] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa-Johnson, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 2729–2732, Nara Japan, 2004.
- [20] M. Breen, L. C. DiLley, and J. Kraemer, "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)," *Corpus Linguistics and Linguistic Theory*, vol. 8, no. 2, pp. 277–312, 2012.

A heuristic corpus for English word prosody: disyllabic nonce words

Sophie Herment & Gabor Turcsan

Aix-Marseille University, Laboratoire Parole et Langage, Aix-en-Provence, France

sophie.herment@univ-amu.fr, gabor.turcsan@univ-amu.fr

Abstract

It is generally admitted that disyllabic words in English are stressed according to their morphological make-up. While prefixed words show differential behaviour according to major grammatical category, non-derived nouns are allegedly trochaic and underived verbs are either iambic or trochaic following rules of quantity-sensitivity. This paper presents a database which was compiled in order to test native speakers' intuition about the stress of disyllables. 53 nonsense words were created displaying different phonological and morphological structures forced by the spelling. These words were embedded in sentences so that each form appears twice, once as a nominal and once as a verbal form. We recorded 20 speakers reading 106 sentences giving 2120 tokens. The construction of nonce words is the main issue at stake, the paper is therefore concerned with methodological questions regarding the design of a heuristic corpus. The data will be freely available on SLDR for the scientific community.

Index Terms: resources, database, phonology, prosody, disyllables, nonce words.

1. Introduction

It is generally admitted that disyllabic words in English are stressed according to i. grammatical category (noun/verb), ii. syllable weight (light/heavy) and iii. lexical properties (e.g. prefixation). While prefixed words show differential behaviour according to major grammatical category (Noun, Verb and Adjective), non-derived nouns are allegedly trochaic and non-derived verbs are either iambic or trochaic following rules of quantity-sensitivity (see [1] for an overview). Table 1 below illustrates the major word classes as far as stress is concerned:

	N	V	A
/10/	paper, fellow	offer, <i>comfort</i>	clever, <i>narrow</i>
/01/	<i>parade, debate, July, hotel</i>	neglect, cajole, <i>begin</i>	distinct, <i>extreme</i>

Table 1. Interaction of syllable weight (heavy H/light L), extrametricality, morphology, analogy and part of speech category: exceptional classes in italics.

Nouns are generally trochaic regardless of syllable weight (HL *paper*, LH *fellow*), as opposed to verbs, which are either trochaic (LL *offer*) or iambic (LH *cajole*) following syllable weight. Moreover, all verbs ending in a consonant cluster (*neglect*) are late stressed. Disyllabic adjectives behave like verbs. There are systematic exceptions to these patterns, like verbs having a prefix + root structure (*begin*), nouns derived from verbs (*debate*) and verbs derived from nouns (*comfort*), and nouns containing specific endings (*parade*). We can also find lexical exceptions like *hotel* or *July*.

In order to test native speakers' intuition about the stress of disyllables, an experiment was carried out involving reading tasks where nonce words were embedded. Similar experiments have been proposed for Spanish [2] or for Italian

[3], languages much alike English in that they also display stress patterns conditioned by either syllable weight (phonology) (see [4]) or language specific lexical properties (morphology) (see [5] for a review).

This paper presents the compilation of the corpus conceived as a heuristic corpus (see [6]), and in particular the making up of the nonsense words. Given that nonce words do not have lexical properties, it is impossible to come up with a comprehensive list reflecting all the above categories. We can test syllable weight and grammatical category relatively easily, but prefixation to some extent only and systematic lexical exceptions not at all.

2. Making up of nonce words

2.1. Syllable weight

We focused first on syllable weight. We made up words combining different syllable weights:

- Light/Light (LL): *begin*;
- Light/Heavy (LH): *recane*;
- Heavy/Light (HL): *furna*;
- Heavy/Heavy (HH): *hastelk*.

One of the major difficulties of having a balanced nonce word corpus is that English spelling allows various possible pronunciations and therefore different rhyme structures: *manem* could either have an LL structure [mə'nem] or an HL structure ['meməm]. *Calbain* could be pronounced ['kælbən] (HL) or [kæl'beɪn] (HH). *Capult* could be [kə'palt] (LH) or ['keɪpalt] (HH) and it is possible to imagine at least 4 different pronunciations for *divey*: ['dvi] (LL), ['dver] (LH), ['darvi] (HL) and [dar'veɪ] (HH). We had to ensure that all types would be sufficiently represented in the database.

We also paid attention to the different possible heavy syllable types: *furnoy* contains a heavy final syllable with a VV type ['fɜ:nɔɪ], while the final heavy syllable of *ducasp* is of the VC type [dju:kæsp].

2.2. Nature of the word final consonant

Following [7]'s claim that certain configurations in final unstressed syllables in English do not seem to exist in verb forms, we also took the nature of the word final consonant into consideration. According to [7], there are no constraints on noun forms, but in disyllabic trochaic verbs, no final cluster and no final schwa plus non coronal sequences can be found. For instance, verbal forms like *meluct* and *lanop* are expected to be stressed on the final syllable because, respectively, of the final consonant cluster and of the non-coronal coda consonant.

2.3. Grammatical category

In order to test the influence of the grammatical category of the word, the same nonce words were embedded in carrier sentences once in a nominal and once in a verbal position:

My Mum likes these

She often ... when she's tired.

So as to mask the task, the words were used in a plural or 3rd person singular form in the sentences.

We also embedded the words in a sentence where they were understood as a proper noun (a place name) so as to see if the speakers pronounced the common noun and the proper noun differently, as it is often the case in the lexicon (the proper noun was written with a final -s, like the common noun):

My Mum likes these... vs. My Mum lives in ...

2.4. Morphological structure

Our word list also contains items that may be associated to a prefix + root construction. Although we are fully aware that testing morphological structure without meaning is a delicate issue, we nevertheless thought it was worth a try. A few words were therefore made up with the common prefixes a-, ab-, ad-, be-, de-, di-, dis-, ex- and re-: *anem, abmone, adnop, bepult, debilk, dilact, disper, exbain, adnop, recane*.

3. Recordings

All in all, following the criteria described above, 53 words were created. We submitted the list to native speakers not participating in the experiment to exclude items that may call for analogical responses with existing items in the lexicon of English.

Each form appears twice, once as a nominal and once as a verbal form, randomly distributed in the test, so that the two word forms are never too close to each other:

My Mum lives in Ducasp.

She often galeafts when she's tired.

My Mum likes these furnoys.

She often calbens when she's tired.

We recorded 20 native speakers of English reading 106 sentences giving 2120 tokens embedded in two sentences.

The 20 speakers were between 20 and 30 years old. They all worked as language assistants at Aix-Marseille University at the time of the recording and they all had a university degree (B.A. or higher). Most of them found the task easy and did not figure out the aim of the experiment. They do not speak the same variety of English but to our best knowledge, while there may be slight differences in vowel reduction patterns, variation of stress placement in disyllables is non-existent.

The recordings took place at Aix-Marseille University, in a recording studio equipped with a Shure SM 58 microphone, a TASCAM M512 mixing desk, related to an iMac with a digidesign Mbox 2 sound card. The software Protocols LE 7 was used.

Questionnaires collecting data about the speakers were filled in by each speaker, along with a consent form. The data have been anonymized, each speaker being assigned a code.

4. Annotation

The results were analysed in an auditory way by two specialists: in the overwhelming majority of cases stress placement was a straightforward issue, accompanied by vowel reduction. The remaining dubious cases, mostly heavy – heavy structures were submitted for judgment to two other trained phoneticians. Figure 1 at the bottom of the page shows an extract of the file containing the results for 10 speakers, with two types of information: the phonetic transcription and the stress pattern. The final column shows the proportion of speakers choosing a trochaic versus iambic pattern.

5. Results and perspectives

The results of our experiment are beyond the scope of this paper and they are detailed in [8]. They confirm our claim that nonce word corpora contribute to our understanding of how languages work. While some of our results confirm generalisations based on random language samples [7] or on dictionary data [9], others refine our knowledge of grammar. Let us just give an example for each type.

- The robustness of the noun/verb dichotomy as far as stress placement is concerned (see table 1 above) is a pleasant surprise given that nonce words do not have meaning. Most of the nouns in our corpus are trochaic (76%) while both iambs and trochees are equally found for verbs (48% are trochees).
- Contrary to what we can see in dictionary data, in our corpus final consonant clusters do not necessarily attract stress for verbs: /01/ *bepult, capult, debilk, galeaft, meluct, nabbast, nabelk*: all LH structures; /10/ *dilact, finlact, foslaint, hastelk*: all HH structures (except *dilact* pronounced /dɪ/). This shows that the observation that a final consonant cluster attracts stress is not an active constraint: verbs displaying this type of final just happen to have a light initial syllable in the English lexicon. Thus a nonce word corpus allows us to separate active dynamic constraints from static lexical patterns.

adnop (n)	'ædnops	æd'nops	'ædnops	'ædnops	'ædnops	'ædnop	'ædnop	'ædnop	æd'nop	'ædnop	7_3
adnop (v)	'ædnops	æd'nops	'ædnops	'ædnops	'ædnops	æd'nops	æd'nops	æd'nops	æd'nops	'ædnops	3_7
befin (n)	'bi:finz	bi'finz	'bi:finz	be'finz	'befinz	'befin	'befin	'befin	'bi:fin	'bi:fin	8_2
befin (v)	bi'finz	bi'finz	bi'finz	be'finz	bə'finz	be'finz	bi'finz	be'finz	bə'finz	bə'finz	0_10
bepult (n)	'bi:pɒlts	be'pɒlts	'bi:pɒlts	be'pɒlts	bə'pɒlts	'belpɒt	be'pɒlt	'bepɒlt	'bepɒlt	(h)'bepɒlt	6_4
bepult (v)	bi'pɒlts	bi'pɒlts	bi'pɒlts	be'pɒlts	bə'pɒlts	bə'pɒlts	bi'pɒlts	bi'pɒlts	bə'pɒlts	be'pɒlts	0_10
dilact (n)	'di:lækt	dɪ'lækt	'di:lækt	dɪ'lækt	'di:lækt	'di:lækt	'di:lækt	dæ'lækt	də'lækt	'di:lækt	6_4
dilact (v)	'di:lækt	dai'lækt	'di:lækt	'di:lækt	'di:lækt	'dailækt	dai'lækt	dɪ'lækt	'dailækt	dɪ'lækt	6_4
gapel (n)	'geɪpəl	gə'pel	'gæpəl	'geɪpəl	'gæpəl	'gæpəl	'gæpəl	'gæpəl	'gæpəl	'gæpəl	8_2
gapel (v)	'geɪpəl	gə'pel	gə'pel	'geɪpəl	'gæpəl	'gæpəl	'gæpəl	gə'pel	'gæpəl	gə'pel	6_4

Figure 1: Results for a few words of the experiment (10 speakers)

In this paper we also want to insist on the idea that a heuristic corpus should not necessarily serve one purpose and fall into oblivion. As mentioned in the introduction, the list of nonce words was created to test native speakers' intuition about the stress of disyllables in English but it can be used for other prosodic purposes. Rhythm can be an interesting issue: one of the speakers almost always stresses the second syllable of the words and reduces the first syllable of the verbs, not of the nouns. This behaviour is mysterious and definitely worth investigating. Although we could not find any significant difference in stress placement according to the origin of the speakers, vowel reduction patterns might have something to do with varieties: some speakers make more reductions on unstressed syllables than others. Moreover, some syllable types display more vowel reductions than others and it would be interesting to try and understand why. Generally speaking, anyone interested in phonological strength relations and more specifically in the interaction of word prosody and the licensing of segmental features (see [10] and the references therein) will find valuable material in the annotated corpus.

These are only possible lines of research and we think that a nonce word corpus can be helpful for the scientific community. This is the reason why it will soon be made freely available on the Speech Language Data Repository (<http://www.sldr.org>).

6. References

- [1] Halle, M., "The Stress of English Words 1968-1998", *Linguistic Inquiry* 29, 539-568, 1997
- [2] Bárkányi, Zs., "A fresh look at quantity sensitivity in Spanish", *Linguistics* 40, 375-394, 2002.
- [3] Krämer, M., "Main stress in Italian nonce nouns", In D. Torck, and W. L. Wetzels [Eds], *Romance Languages and Linguistic Theory 2006*, Amsterdam and Philadelphia: John Benjamins, 127-141, 2009.
- [4] Hyman, L., "A Theory of Phonological Weight", Stanford: CSLI publications, 2003.
- [5] Hulst, H.G., van der, "Word accent", in H. van der Hulst [Ed], *Word prosodic systems in the languages of Europe*, Berlin & New York: Mouton de Gruyter, 3-116, 1999.
- [6] Scheer, T. "Le corpus heuristique : un outil qui montre mais ne démontre pas", *Corpus* [On line], 3 | 2004, <http://corpus.revues.org/210>.
- [7] Hammond, M., "English Phonology", Oxford: Oxford University Press, 1999.
- [8] Turcsan, G. & Herment, S., "Making sense of nonce word stress in English", *Proceedings on line of the 3rd international conference on English Pronunciation: Issues and Practices (EPIP3)*, Murcia, Spain, May 8-10, 2013. <https://sites.google.com/site/epip32013/home>
- [9] Fournier, J-M., "Manuel d'anglais oral", Paris: Ophrys, 2010.
- [10] Nasukawa, K. & Backley Ph. [Eds]. "Strength Relations in Phonology", Berlin and New York: Mouton de Gruyter, 2009.

C-PROM-Task: A New Annotated Dataset for the Study of French Speech Prosody

Mathieu Avanzi¹, Lucie Rousier-Vercruyssen¹, Sandra Schwab², Sylvia Gonzalez¹, Marion Fossard¹

¹ Chaire de logopédie, University of Neuchâtel, Ruelle Vaucher 22, Neuchâtel, 2000, Switzerland

² ELCF, University of Geneva, Rue Candolle 5, Geneva, 1211, Switzerland

mathieu.avanzi@unine.ch, lucie.rousier-vercruyssen@unine.ch, sandra.schwab@unige.ch,
sylvia.gonzalez@unine.ch, marion.fossard@unine.ch

Abstract

The aim of this paper is to describe C-PROM-Task, a dataset created and annotated in the same spirit as C-PROM [1]. C-PROM-Task was annotated following a perceptually-based and computer-assisted procedure for the study of syllabic prominences, syllabic disfluences and two ranges of prosodic units. All told, C-PROM-Task comprises recordings and annotated TextGrids of story-telling by 20 native French speakers from Switzerland. The entire dataset is 2 hours 20 minutes long. Some observations are also made regarding accentuation (prominence rate), disfluency rate and phrasing (length of prosodic units) in the corpus.

Index Terms: corpus, spoken French, prosodic annotation, prominence, phrasing.

1. Introduction

Until fairly recently, annotation of continuous French speech was either relatively rudimentary and approximate or the province of a group of specialists working within the framework of phonologic theories. On the one hand, specialists of spoken French ([2], [3] and [4]) transcribe prosodic events using a reduced set of symbols, which does not reflect the actual complexity of prosodic phenomena. On the other hand, phonologists who work in the Autosegmental-Metrical (AM) framework [5], such as [6] and [7], use or develop annotation systems which are not really applicable to spontaneous speech since the data they deal with mostly consist of laboratory speech, that is "light years ahead of the complexity of spontaneous speech" [8]. However, in the past few years, mostly thanks to automatic processing advances, the situation has been changing. Protocols and tools designed to annotate French prosodic structure (semi-)automatically are emerging (see [9] for an overview). An increasing number of projects aim to create available and public annotated corpora [10]. In this context, the purpose of this paper is not to make an inventory of existing systems and resources but (i) to present C-PROM-Task, a prosodically annotated corpus based on the same hypotheses and with the same aims as C-PROM [1]; (ii) to summarize the perceptually-based and computer-assisted procedure used to annotate accentuation (calculation of the position and strength of pitch accents within a given group of words) and phrasing (identification of the different prosodic groups in the prosodic hierarchy) in this corpus; and (iii) to briefly discuss what such annotations can teach us about French speech prosody.

2. Method

2.1. Participants and task

Twenty French-speakers from Switzerland (10 male and 10 female) took part in the study. Ten of the participants were from 19 to 27 years old (mean age: 22.6, SD: 2.2), and the other ten were between 71 and 82 years old (mean age: 75.5, SD: 3.1). We will refer to these two groups as the "young group" and the "older group". The participants in the two groups were strictly matched for gender. Speakers were recorded in a storytelling situation. They were asked to describe verbally 18 different story picture sequences with 3 increasing levels of complexity according to the number and gender of the characters involved in the story picture sequence, i.e., level 1 or easy level: one character; level 2 or medium level: 2 characters of different genders; level 3 or difficult level: 2 characters of the same gender. Half of the stories were presented with a logical order of events (logical condition) and the other half with a non-logical order (non-logical condition). For each of the three complexity levels and each of the two conditions (logical vs. non-logical order), speakers had to describe three different story picture sequences. Using a referential communication paradigm (see [11] and [12]), the storytelling in sequence test enables one to assess how a participant (the director of the interaction) plans his/her discourse and what type of verbally discriminating information he/she produces that will enable an addressee (the researcher) to identify and order the 6 pictures that constitute a story sequence. To avoid non-verbal communication, the participant and researcher were separated by an opaque screen. For each interaction, stories were presented in a pseudo-randomized order and verbal productions were recorded.

2.2. Selection of the files

We processed one story for each level of complexity for both orders, making six stories in all for each speaker. Thus the C-PROM-Task corpus contains 120 files (3 levels of difficulty*2 order conditions*20 speakers). The total duration of the corpus is 2 hours 20 minutes.

2.3. Annotations

2.3.1. Orthographic transcription and text-to-sound alignment

Each of the 120 files was first orthographically transcribed within Praat software [13]. Transcriptions were then semi-automatically aligned in phones, syllables and words with

EasyAlign [14] script. Alignments were manually checked and corrected when necessary by two of the authors (each author was in charge of half of the data). Silent pauses and non-transcribed segments (interventions from the researcher, overlapped speech, laughs, etc.) were transcribed with the symbol "_".

2.3.2. Annotation of syllabic prominences and disfluencies

Syllabic prominences and disfluencies were manually annotated in parallel by two of the authors, using the method described in [1]. To summarize, the annotators had to listen to small stretches of the signal (2-3 seconds on average), 3 times at most, and to code in a dedicated tier (an empty copy of the syllable tier) with "p" and "P" the syllables they perceived as weakly and strongly prominent. They were asked to annotate the syllables perceived as associated with a disfluency with "H" (false starts, breaks in the syntactic program, elongations due to a hesitation, "euh", etc.). To ensure the coding was performed on perceptual bases as far as possible, the researchers did not have visual access to acoustic information (f0 and intensity lines, spectral envelope). The Analor tool [15] was then used to obtain an automatic annotation of prominent syllables in a new tier. The algorithm calculates the relative height and duration of each syllable in a given stretch of speech by comparing the value of the analyzed syllable with the average of the six adjacent syllables (i.e. three preceding and three following ones); the pitch rise slope is then processed and the presence of a subsequent silent pause is considered. Thresholds to activate prominence were the ones trained for spontaneous speech indicated in [15], that is to say 1.5 for relative duration, 1.3 st for relative height and 2.5 st for melodic rise. To avoid false-alarm prominence detection, [16]'s algorithm was used to ensure that pitch path files were as clean as possible.

The inter-annotator agreement coding was statistically tested. Regarding prominence, "p" and "P" were merged and considered together as a single category (which contrasts with "0", see [1] for the justification). The total number of intervals considered was 18'604 (syllables associated with an "H" were not taken into account). First, Cohen's kappa [17] was used to assess reliability between a pair of annotators. It appeared that the agreement between the two human annotators was substantial ($\kappa = 0.68$), while it was fair between the first annotator and Analor ($\kappa = 0.56$) and between the second annotator and Analor ($\kappa = 0.48$). Fleiss' Kappa, a measure used to assess reliability between more than two annotators [19], indicated a fair agreement between the three annotators ($\kappa = 0.57$). Regarding disfluencies, Cohen's kappa revealed an almost perfect agreement ($\kappa = 0.82$) out of the 12'530 intervals taken into account (syllables annotated with the symbols "p" or "P" were excluded from the calculation). A syllable was considered prominent in the reference tier if it was marked in two of the three annotation tiers. The final status of a syllable hesitating between "H" and something else was decided after discussion between the two annotators.

2.3.3. AP segmentation

Next, a tier indicating the boundaries of minor prosodic units was obtained as follows: each final syllable of a lexical word or polysyllabic functional word that was coded as prominent generated the boundary of a minor prosodic

constituent, including every element without prominence on its left side. Following the AM theory [20], we will refer to these units as Accentual Phrases (henceforth APs), even though in many cases they are closer to Clitic Groups than to APs. When the last syllable's item was a schwa, the penultimate syllable was considered as carrying the final pitch accent, thus marking the right boundary of the AP. Syllables labeled "H" are either comprised in the AP of the surrounding valid syllables or form an AP on their own (such syllables are excluded from the calculations presented below).

2.3.4. IP segmentation

In spite of its importance for speech processing, the definition of the major prosodic units called Intonational Phrases (IP) in the AM framework is still an issue for scholars working on French [17]. It is either described in terms of syntactic/information structure (root clauses, embedded coordinated clauses, left- and right-peripheral constituents map onto IPs) or regarding their intonational realization: IPs are defined by the presence of a nuclear accent (syllable associated with a major pitch movement, a pre-boundary lengthening and/or followed by a silent pause). To identify IP in the C-PROM-Task database, the prominence degree detection function provided by the Analor tool [15] was used. On the basis of four automatically measured acoustic parameters (relative syllabic duration, relative f0 average, slope contour amplitude and presence of an adjacent silent pause), the software estimates a degree of strength for the last syllable of each AP on a scale from 0 to 10 (from the least to the most prominent). The calculations rely on two fundamental principles. The first is a quantity principle: the greater the number of acoustic parameters involved in the identification of a prominence and the distance from predetermined thresholds, the stronger the prominence is perceived. The second is a compensation principle, which stipulates that if one of the classic parameters involved in the perception of prominence in French presents a low value and another presents a high value, there will be the same feeling of prominence as if the two parameters involved both presented a medium score. We considered that the last syllable of an AP was associated with a nuclear contour, i.e. an IP boundary, if its strength reached a score of 4/10.

3. Analysis

We illustrate and briefly comment on some descriptive statistics regarding the distribution of the syllables in the corpus according to their status (+/- prominent, +/- disfluent). We then discuss the effects of different factors influencing accentuation and phrasing.

3.1. Descriptive Analysis

3.1.1. Distribution of the syllables according to their labels

Among the 21'161 syllabic intervals in our corpus, 7'546 were identified as prominent (35.65%), 2'464 as disfluent (11.64%) and 11'151 were left blank (52.69%), as seen in Figure 1:

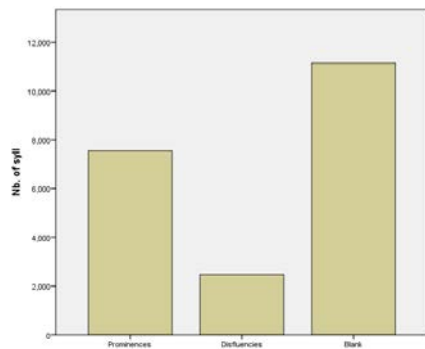


Figure 1. Distribution of the syllables in the corpus according to their type. From left to right: prominent syllables, disfluent syllables and blank syllables.

3.1.2. AP length

In total, the corpus contains 6'547 APs. The mean length of an AP is 2.85 syllables. On average, as seen in Figure 2 below, an AP mostly comprises between 2 and 4 syllables (80.3% of the data), rarely less or more.

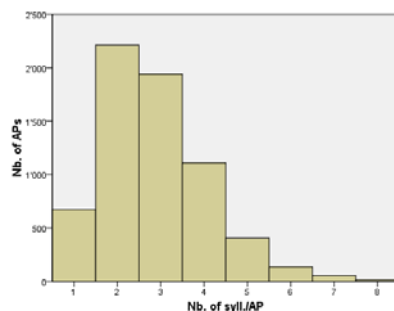


Figure 2. Number of APs according to their syllabic length.

3.1.3. IP length

In total, the corpus contains 3'895 IPs. The mean size of an IP is 4.79 syllables. As seen in Figure 3, most of the IPs that comprise our corpus are from 2 to 7 syllables. Beyond 10 syllables, the number of syll./IP decreases significantly:

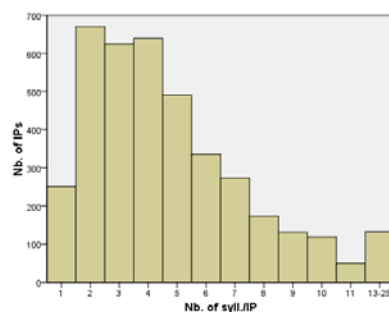


Figure 3. Number of IPs according to their syllabic length.

3.2. Statistical Analysis

We wanted to go further and provide a statistical analysis of the data presented in this paper. We report the results we obtained by testing the effects of age, articulation rate, order condition (logical vs. non-logical) and level of complexity

(easy, medium, difficult) on accentuation (prominence rate) and phrasing (length of prosodic units). Three Generalized Estimated Equations (with repeated measures) models were run, for the first with the rate of prominence as a dependent variable, for the second with AP length as a dependent variable, and for the third with IP length as a dependent variable. Age, local articulation rate (mean syllabic duration excluding the silent pause for each AP and each IP), order condition and level of complexity were entered as predictors.

First, results indicate that none of the 4 predictors mentioned above had any effect on prominence rate. In other words, young speakers did not behave differently from the older ones, and both groups did not produce less pitch accents when they increased the pace at which they talked, or when they told an easy story in the logical order compared with a difficult story in the same order or not.

However, the level of complexity of the task had a significant effect on AP length ($p < 0.01$). Post hoc tests reveal that APs were shorter for level 3 stories than for level 1 stories ($p < 0.01$) but that they did not show any differences for level 2 stories. In addition, it appears that articulation rate had an effect on AP length: the faster the speaker articulated, the longer his/her APs ($p < 0.001$). We should note that there was an interaction between articulation rate and age ($p < 0.001$). This effect was not the same for young speakers as for older ones. As seen in Figure 4, older speakers had a longer syllabic duration when the prosodic constituent was short compared with the young speakers. Note that this difference in length tended to weaken as the length of the prosodic unit increased:

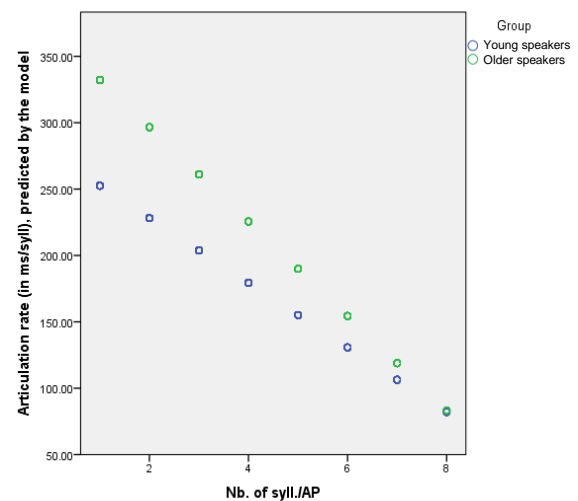


Figure 4. Number of syll./AP according to the age of the speakers and articulation rate, expressed in ms/syll., and predicted by the model.

Finally, the analysis revealed that IP length was not influenced by the condition or level of difficulty. However, it was influenced by articulation rate ($p < 0.001$), which interacted with age ($p < 0.001$). As was the case for the AP, long IPs presented a shorter mean duration than short IPs, which, on the other hand, presented a higher syllabic mean duration.

4. Discussion

The results presented in the preceding section are interesting with respect to our knowledge of French prosody. First,

regarding prominence and disfluency rate, the results found in this study can be compared with other corpus-based analyses of French. For example, out of the 21'161 syllabic intervals in C-PROM-Task, 35% were annotated as prominent while 11.6% were annotated as disfluent. Out of the 17'778 syllables in the C-PROM corpus [1], 4'570 syllables were annotated as prominent (26%) and 805 syllables (4.5%) were associated with a disfluency. In the Rhapsodie corpus [10], which contains 45'192 syllables, the prominence rate is 41% while the disfluency rate is slightly less than 8%.

Next, regarding phrasing, our results are in agreement with the ones obtained by [21], who found that the average length of the minor prosodic units that she calls "rhythmic groups" (and which correspond to the units we call APs) was 3.5 syllables in her corpus (2.85 syll./AP in our dataset). She also found that 80% of the APs in her data was composed of 2 to 4 syllables. Based on our data, we made exactly the same observation. The results obtained in our study are also in agreement with [6]'s description of French. Indeed, the authors found that an AP is composed of 3.5 to 3.9 syllables on average. Our data also confirm the idea that in French an AP cannot contain more than 8 syllables ([22]). Very little work is available for IP and it is therefore impossible to compare our results with other work.

Preliminary analyses examining the impact of some factors such as order condition, level of difficulty, articulation rate or age on accentuation and phrasing led to quite unexpected conclusions. On the one hand, it appeared that prominence rate was not influenced by one of the 4 factors, while AP and IP lengths varied according to articulation rate (the effect differing by gender). Additional analysis, not detailed here, revealed a significant effect of age on articulation rate: older speakers articulated much more slowly than younger speakers. These results are in agreement with previous studies on articulation rate in French [23].

5. Conclusions

The aim of this paper was to present C-PROM-Task, a new prosodically annotated dataset for the study of French prosody. The entire database is more than 2 hours long and contains speech by a group of young and a group of older native French-speakers from Switzerland. Recordings of controlled story-telling were first transcribed and aligned. They were then annotated for the study of prominent and disfluent syllables. Kappa measures were used to assess reliability between the annotators, which was judged to be fair to substantial. Recordings were also segmented in minor and major prosodic units, here called Accentual Phrases and Intonational Phrases. From a descriptive analysis of the data we were able to confirm previous findings. On average, one syllable carries a pitch accent every three or four syllables; 10% percent of a speaker's syllables is associated with a disfluency; AP length comprises between 3 and 4 syllables, and cannot exceed 8 syllables, while IP length is more sensitive to variation. Finally, the effects of many factors on articulation rate were tested, and it appeared that only age and constituent length had an effect on this prosodic variable.

6. Acknowledgments

This work was supported by the SNF under Grant No. 100012-140269, hosted at Neuchâtel University.

7. References

- [1] Avanzi, M., Simon, A. C., Goldman, J.-P. and Auchlin, A. "C-PROM. An annotated corpus for French prominence studies", Proc. of Prosodic Prominence, Speech Prosody Satellite Workshop, 2010.
- [2] Blanche-Benveniste, C., Bilger, M., Rouget, C. and van den Eynde, K. "Le français parlé. Etudes grammaticales", Paris, Éditions du CNRS, 1990.
- [3] Morel, M.-A. and Danon-Boileau, L. "Grammaire de l'intonation : l'exemple du français", Paris/Gap, Ophrys, 1998.
- [4] Groupe de Fribourg. "Grammaire de la période", Bern, Peter Lang, 2012.
- [5] Ladd, R. "Intonational Phonology", Cambridge University Press, 1996.
- [6] Jun, S. A., and Fougeron, C. "Realizations of Accentual Phrase in French intonation", *Probus*, 14: 147-172, 2002.
- [7] Delais-Roussarie, E. et al. "Developing a ToBI system for French", in Frota, S & Prieto, P. (eds). *Intonational Variation in Romance*, Oxford University Press, in press.
- [8] Lacheret, A. "La prosodie des circonstants", Paris/Leuven, Peeters, 2003.
- [9] Delais-Roussarie et al. "Outils d'aide à l'annotation prosodique de corpus". *Bulletin PFC*, 6: 7-26, 2006.
- [10] Lacheret A., Kahane, S. and Pietrandrea, P. "Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French". New York/Amsterdam, Benjamins, in press.
- [11] Champagne-Lavau, M., Fossard, M., Martel G., Chapdelaine, C., Blouin, G., Rodriguez, J.-P. and Stip, E. "Do patients with schizophrenia attribute mental states in a referential communication task?", *Cognitive Neuropsychiatry*, 14(3): 217-239, 2009.
- [12] Clark, H. H. and Wilkes-Gibbs, D. "Referring as a collaborative process", *Cognition*, 22, 1-39, 1986.
- [13] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer (Version 5.5)". www.praat.org, 2013.
- [14] Goldman, J.-Ph. "EasyAlign: an automatic phonetic alignment tool under Praat", Proc. of Interspeech, 3233-3236, 2011.
- [15] Avanzi, M., Obin, N., Lacheret-Dujour, A. and Victorri, B. "Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree", Proc. of Interspeech, 2011.
- [16] De Looze, C. and Hirst, D. "Detecting key and range for the automatic modelling and coding of intonation", Proc. of the Speech Prosody Conference, 135-138, 2008.
- [17] Delais-Roussarie, E. and Post, B. "Unités prosodiques et grammaire de l'intonation : vers une nouvelle approche", Actes des 27èmes JEP, Avignon, 2008.
- [18] Cohen, J. "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement*, 20(1): 37-46, 1969.
- [19] Fleiss, J. L. "Measuring Nominal Scale Agreement among Many Raters", *Psychological Bulletin*, 33: 613-619, 1973.
- [20] Avanzi, M. "Note de recherche sur l'accentuation et le phrase du français à la lumière des corpus", Tranel, 2013.
- [21] Delais-Roussarie, E. "Pour une approche parallèle de la structure prosodique. Étude de l'organisation prosodique et rythmique de la phrase française". PhD thesis, Toulouse-le Mirail University, 1995.
- [22] Martin, P. "Prosodic and Rhythmic Structures in French", *Linguistics*, 25: 925-949, 1987.
- [23] Schwab, S. "Les variables temporelles dans la production et la perception de la parole", PhD thesis, Geneva University, 2007.

Rapid and Smooth Pitch Contour Manipulation

Michele Gubian¹, Yuki Asano², Salomi Asaridou^{3,4}, Francesco Cangemi⁵

¹Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

²Department of Linguistics, University of Konstanz, Germany

³International Max Planck Research School, Nijmegen, The Netherlands

⁴Donders Institute for Brain, Cognition and Behavior, Nijmegen, The Netherlands

⁵University of Toulouse II, France and University of Cologne, Germany

m.gubian@let.ru.nl, yuki.asano@uni-konstanz.de,

s.asaridou@donders.ru.nl, fcangemi@uni-koeln.de

Abstract

Speech perception experiments are often based on stimuli that have been artificially manipulated, e.g. to create hybrids between two given prosodic categories. Tools like the widely used PSOLA re-synthesizer available in Praat provide the user full editing control on the shape of pitch and intensity contours, as well as on local relative speech rate of recorded utterances. However, high level experimental specifications, e.g. “generate a number of pitch contours whose shapes are gradual transitions between two reference contours”, are not easily translated into a sequence of low level manipulation operations. Often, the viable solution is the manual stylisation of contours, which drastically reduces the degrees of freedom, but at the same time can introduce unwarranted alterations in the stimuli. In this paper we introduce a method, implemented in a software tool interfaced with PSOLA, that automatically translates high level specifications into low level operations in a principled way. No manual editing in the form of contour stylization or otherwise is required. The tool enables the rapid generation of manipulated stimuli of desired properties, while it guarantees that acoustic feature alterations are always smooth. Three use cases demonstrate the efficacy of the tool in real experimental conditions.

Index Terms: stimuli manipulation, perception experiments, prosody

1. Introduction

The artificial manipulation of natural speech is common practice in the preparation of stimuli for perception experiments. A number of speech processing software tools, like the PSOLA (Pitch Synchronous Overlap Add Method) re-synthesis tool available in Praat [1], offer the possibility to modify the shape of f_0 and intensity contours extracted from recorded utterances, and to selectively alter segment duration. Speech scientists use these tools to investigate the perceptual effect of those features in isolation, e.g. by keeping the original f_0 contour and varying segment duration or vice versa. These tools have opened new possibilities for the experimental verification of phonological theories of prosody and intonation. Manipulated resynthesized stimuli are also being used in combination with brain imaging in the investigation of the mental processes involved in language acquisition and learning. However, changing prosodic parameters in isolation can give rise to stimuli that sound unnatural if

constraints on the co-variation of the parameters are violated. Therefore, there is a need for embedding manipulation tools in a work bench that makes it possible to vary parameters in combination and to quickly generate large numbers of stimuli that can then be assessed auditorily to remove unnaturally sounding tokens.

In this work we focus on two typical scenarios involving f_0 contour manipulation. The first one is the creation of f_0 contours whose shapes gradually vary between two reference contours. By analysing subjects’ responses (e.g. through reaction times) to stimuli whose f_0 contours present hybrid characteristics between two known categories, it is possible to make inferences on the perceptual space, which in general is not linearly mapped on the acoustic space. The second scenario involves the creation of stimuli where one or more speech features are imported, or ‘transplanted’, from other stimuli. An example is the creation of hybrids that do not exist in a natural language, like the creation of stimuli that combine spoken words in a non-tonal language with f_0 contours extracted from words spoken in a tonal language (cf. Section 4.2).

Although stimuli manipulation procedures are widely used, a closer analysis of the common practice reveals a number of operational limitations. In the case of gradual modification of f_0 contours between two reference shapes, the creation of intermediate shapes using default Praat tools requires two operations, namely (i) aligning corresponding segmental boundaries in time, and (ii) stylizing the reference f_0 contours using straight line segments. The manipulation is carried out by changing the position of the junction points (usually only one) in the stylized contours using a graphical editor or a script (e.g. both options available in Praat). For example, in [2] artificial mixtures of two Mandarin tones are generated from a three-points stylization of each tone and by gradually moving the point in the middle. Similarly, in [3] the shape of a pitch rise is changed from concave to convex by imposing a three-points stylization and by varying the middle point height, and in [4] a modulation between two more complex shapes (dip and hat) is obtained by moving more than one point at the same time. While segmental alignment is performed in order to preserve the anchoring of f_0 movements to the segmental material, f_0 contour stylization is not necessarily justified by theoretical or experimental reasons. On one hand, stylization may help in reducing the complexity of a prosodic model by isolating simple

shape features. On the other hand, there is always the risk of losing important detail, e.g. the type of curvature (concave or convex) of a rising gesture [5]. Stylization is often carried out manually, which entails empirical judgement and time consuming procedures. Semi-automatic stylization procedures exist, which are claimed to preserve all perceptual properties of the original f_0 contour (see [6] for an overview). However, those faithfulness guarantees refer to the so-called ‘close copy’ of a contour, thus they do not necessarily extend to manipulation.

While full f_0 contour grafting does not necessarily involve stylization, which is sometimes applied nonetheless (e.g. in [7]), segmental alignment remains a requirement. Ideally, the time warping involved in the alignment should alter the utterance structure as little as possible, in order to minimise the risk of introducing unwanted perception effects. Praat provides manual editing facilities for the selective manipulation of segment duration, and scripted procedures are also available, like [8] or the one used in [9]. To our knowledge, the available scripts apply duration changes locally on each segment, which may introduce noticeable discontinuity effects whenever alteration values of adjacent segment durations differ too much.

In this paper we present a contour manipulation method that eliminates the limitations described above. The proposed method (i) implements smooth deformation of f_0 and intensity contours in order to align them to given segmental boundaries, and (ii) provides a transparent way to combine two or more contours in desired proportions, e.g. for the creation of hybrids or for averaging. The procedures are automatic and controlled by the user through a few parameters, like the degree of ‘elasticity’ of contour deformation. This method improves the experimental procedures involved in stimuli manipulation in two ways. First it minimises the risk of introducing unwanted alterations in the original speech samples. Second, it provides for the automatic generation of stimuli in large quantities, enabling the rapid examination of several experimental conditions at design time.

The rest of the paper is structured as follows. In Section 2 the method is described in its main principles, which are adapted from techniques applied in Functional Data Analysis [10]. In Section 3 we give a brief description of the software tools that carry out all the necessary operations. In Section 4 we report three use cases where the method has been applied in real experimental conditions. Finally, in Section 5 we draw conclusions.

2. Method

2.1. Smoothing

We illustrate the basic principles of the proposed method by referring to the manipulation of f_0 contours; similar considerations apply for intensity contours. The first operation to be carried out on the input data is called *smoothing*, which transforms a sampled f_0 contour into a continuous curve represented by a mathematical function of time $f(t)$. This function is constructed by combining a set of so called basis functions such that the combination fits the sampled data. In the case of features like f_0 , whose contours in time can assume arbitrary shapes and do not present periodicity, it is customary to adopt B-splines as basis [11]. An example of smoothing is shown in Figure 1. The user can control the degree of smoothing through a number of parameters (see Section 3).

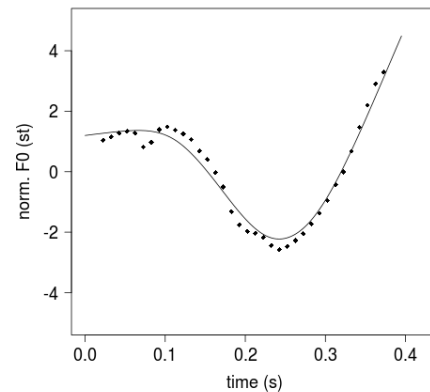


Figure 1: Example of a smoothed f_0 contour. Dots represent f_0 samples obtained from the pitch tracker available within Praat. The curve is a B-spline. This contour is extracted from a realisation of a three-syllabic word. The y-axis reports f_0 values in semitones after the global mean value was subtracted (corresponding to the zero level).

Once contours are represented by functions of time, expressing combinations of them becomes trivial. For example, to create hybrids between two base contours A and B, the arithmetic operation $(1 - \alpha) \cdot f_A(t) + \alpha \cdot f_B(t)$ produces the desired combinations in proportions controlled by the parameter α .

2.2. Landmark registration

Suppose A and B are realisations of the same three-syllabic word. The operation $(1 - \alpha) \cdot f_A(t) + \alpha \cdot f_B(t)$ defined above combines values of contours A and B at corresponding points in time. However, the inevitable segment duration differences between the two realisations would mix shape traits belonging to different syllables, which would blur the timing relation between f_0 movement and segmental content.

This problem is solved by a convenient transformation applied to the time axis that alters each contour in such a way that corresponding segmental boundaries get aligned in time. This operation, called *landmark registration*, is carried out automatically and it is based on the position of each boundary (landmark) on each of the input contours. The time warping carried out by landmark registration guarantees that the qualitative aspects of the curves are not altered. Moreover, the local speech rate alterations are spread gradually throughout the entire contour, which minimises discontinuity effects. Figure 2 shows the effect of landmark registration on an f_0 contour extracted from a three-syllabic word, where the syllable boundaries are shifted to a desired position.

2.3. Manipulation and re-synthesis

Here we illustrate how to combine smoothing and landmark registration in order to create a set of stimuli whose f_0 contours are combinations of two base contours A and B, extracted from two realisations of the same word or phrase in two different conditions (e.g. yes-no question vs. statement); analogous steps are required in other manipulation schemes.

First, f_0 contours are extracted from utterance A and B,

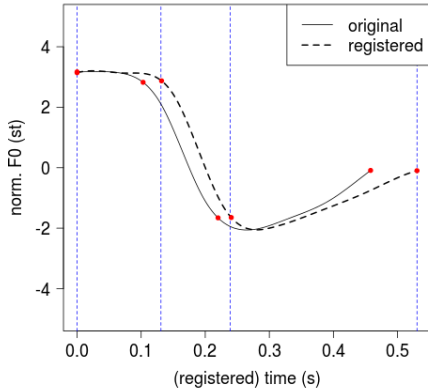


Figure 2: Example of landmark registration of a smoothed f_0 contour extracted from a realisation of a three-syllabic word. Dots show the position of syllable boundaries, vertical dashed lines the position where the boundaries are going to be shifted by landmark registration. The solid curve is the original f_0 contour, the dashed curve the contour after registration.

and all relevant landmarks, like syllable boundaries, are marked (e.g. on a Praat textgrid). Then f_0 contours are smoothed and turned into functions, $f_A(t)$ and $f_B(t)$, which have different duration and whose landmarks are not synchronised yet. Suppose we want to carry out re-synthesis on the recording of utterance A, let us call it the *base* utterance. Before mixing f_0 contours of A and B we have to synchronise utterance B on the landmarks of the base. Thus, landmark registration is carried out on the boundaries of utterance B by imposing a time warp that aligns its boundaries on those of the base A. This is internally represented by a warping function $h_{B \rightarrow A}(t)$. After this, function $h_{B \rightarrow A}(t)$ is applied on $f_B(t)$ to obtain a different function $f_B(t_A)$, which has (qualitatively) the shape of $f_B(t)$ but is aligned with the landmarks of the base A. At this point we can create a number of mixtures of the form $f_\alpha(t) = (1 - \alpha) \cdot f_A(t) + \alpha \cdot f_B(t_A)$, say for $\alpha = 0, 0.2, 0.4, \dots, 1.0$, where the value $\alpha = 0$ will produce a stimulus that should be identical to the original A and will be employed in the experiment in order to control for the re-synthesis effect, as well as being a useful sanity check for the re-synthesis. Finally, all the $f_\alpha(t)$ are converted into PitchTiers and used in Praat PSOLA re-synthesizer to modify the shape of the f_0 contour of the base utterance A.

3. Software

The software to carry out all the operations described above consists of a main R script [12] and a number of auxiliary R and Python scripts (www.python.org). The package is developed and maintained by the first author and is available for download from his website [13] (direct link [14]). The core functionalities are based on the `fda` library [15], with minor modifications. The software accepts Praat formats as input (e.g. TextGrids) and produces output also in Praat formats (e.g. PitchTiers) or wave files by calling Praat. The main script is not intended to be executed in a single call, because the procedure is composed by a cascade of operations, some of which require the user to

set a number of parameters, like those controlling smoothing, whose outcome has to be checked by plotting (cf. Figure 1). A simple expedient has been devised in order to alleviate the problem of having gaps in f_0 contours due to voiceless sounds. This is a hindrance when such a contour is transplanted on speech material where voiced sounds occupy the f_0 gap, as reported in [16]. The input contour is smoothly interpolated by (automatically) padding extra samples at a level computed by averaging neighbour sample values.

4. Use Cases

In this section we describe three perception experiments, conducted by the second, third and fourth author, respectively, where stimuli manipulation was carried out using the method and the software introduced above.

4.1. Perception of pitch and length cues in Japanese as a second language

The second author is conducting a study on the influence of first language (L1) in the discrimination of pitch and length in a second language (L2). A number of AX (same-different) discrimination tasks were designed based on German as L1 and Japanese as L2, where German participants were either learners of Japanese or not. In Japanese, which has both consonantal and vocalic length contrasts, pitch appears to be a secondary cue for length [17, 18]. For German participants, whose L1 does not have the consonantal length contrast, the discrimination of length is expected to be harder for consonants than for vowels [19], while the role of pitch in discrimination of length has not been studied yet.

Six non-sense disyllabic words were created that respect Japanese phonotactics and differ from each other in manner of articulation and voicing of the medial consonant. Each word is created in three versions, which differ either in the duration of the first vowel or in the duration of the second consonant, resulting in a singleton (CVCV), a geminate (CVCCV) and a long-vowel (CVVCV) as counterparts (e.g., /punu/, /pu:nu/, /pun:u/) [20].

All of the tokens were produced either with a lexical pitch accent on the first syllable (High-Low) or with no pitch accent at all (High-High). Each token was recorded six times by the same L1 speaker of Japanese, in order to present different tokens to the participants. Five segmental boundaries were considered: /C/N/C/N/, /C/N/CC/N/ or /C/NV/C/N/.

AX tasks were designed to measure discrimination in one cue, either pitch or duration, while keeping the other constant. In order to keep segment duration constant across stimuli, landmark registration was applied as illustrated in Section 2.2, where the 12 realisations of a given segmental pattern (e.g. six /pu:nu/ realisations for each of the two pitch patterns) were aligned on the average time location of those boundaries across realisations. Conversely, to keep pitch constant across stimuli in the duration contrast condition, a cascade of two landmark registrations was applied as follows. An average f_0 contour was created by first aligning the $N = 18$ contours $f_i(t)$, $i = 1, \dots, N$ (i.e. six realisations times three duration patterns) on the landmarks of one of them, say the first one, obtaining $f_i(t_1)$. This allowed to compute a time-aligned average contour $f_A(t_1) = \frac{1}{N} \sum_i f_i(t_1)$. This average shape was re-aligned on the segmental timing of each of the N realisations, obtaining

ing N pitch-normalised contours $f_A(t_i)$, which were eventually imposed on the recorded stimuli. The naturalness of the stimuli was highly satisfactory in both manipulations. Moreover, f_0 padding was successfully used in order to accommodate for gaps due to voiceless segments (cf. Section 3). The results of the AX tasks are being collected and analysed at the time of writing of this paper.

4.2. Dutchinese

A perception experiment is being conducted by the third author with the purpose of investigating neural and genetic correlates of sound learning performance. The study was performed using Dutch native speakers as subjects. We planned to use a phonetic contrast that would be unknown and difficult enough for Dutch natives while being ecologically valid. We chose the pitch contrasts used in the four Mandarin tones. Following the paradigm used by [21], we opted for Dutch-Chinese hybrid stimuli, i.e. words that respect Dutch phonotactic rules but with Mandarin tone contours superimposed on them. By using hybrid stimuli we could create minimal quadruplets where we manipulated f_0 whilst keeping all the other variables (e.g. word duration, intensity, vowel length, production rate etc.) constant. Twenty-four pseudowords with a consonant-vowel-consonant (CVC) structure were created (e.g. /ket/, /ba:f/, /nal/, /be:m/). We recorded eight Dutch native speakers reading aloud the list of those 24 CVC pseudowords. Similarly, we recorded eight native speakers of Chinese uttering the word /mi/ on four Mandarin tones.

Manipulation was applied for the grafting of f_0 contours from the word /mi/ to the pseudowords. The boundary between /m/ and /i/ was disregarded, since the duration of /m/ was always so short compared to /i/ that no difference could be noticed by considering the f_0 contour shape either as starting from the onset of /m/ or of /i/. On the other hand, we had to investigate how to align the Chinese tone contours on the CVC tokens. In /CVC/ there are four boundaries, we called them 1, 2, 3, 4. Using the software described in Section 3, the Mandarin tone contour from the word /mi/ was applied on each pseudoword in 4 different ways: from boundary 1 to 3, 1 to 4, 2 to 3, and 2 to 4. Landmark registration was applied, which in this case was equivalent to a linear time compression or expansion, because only two landmarks were present. Depending on the nature of the consonants (voiced or voiceless) and tones, some combinations were expected to be better than others. The full application of this scheme produced thousands of stimuli, one for each pseudoword, tone, Dutch speaker, Chinese speaker and alignment criterion. From those, a subsample of 384 tokens was extracted and used in a rating study, whose purpose was to find the combinations that would result in the highest tone identification as well as highest naturalness ranking when judged by native Mandarin speakers.

The naturalness of the sound was very satisfactory with the exception of three out of eight speakers whose voices sounded less natural in some of the tokens. The identifiability of tones by native speakers partially suffered from the experimental requirement of excluding all cues except for pitch, as different tones tend to have different durations in Mandarin. This problem mostly affected tone 3, which is the longest in duration. The general trend of the rating study revealed that overlap from boundaries 1-3 and 1-4 were preferred compared to 2-3 and 2-4 although that varied as a function of the Dutch-Mandarin speakers pitch agreement, the specific phonemes, and the tone

in question.

4.3. Tempo and sentence modality in Italian

In languages such as Italian and its regional varieties, sentence modality contrasts, e.g. the opposition between declaratives and yes-no questions, are conveyed through prosodic means alone. The intonational aspects, expressed phonetically by f_0 contours and their synchronization with segments and syllables, have been thoroughly studied in production and perception (e.g. [22] for Neapolitan Italian). Recent studies, however, point to the existence of consistently produced differences in segmental durations as well (e.g. [23, 24], and [25] for Neapolitan Italian), but no evaluation of their perceptual role had been provided yet. In order to study the perceptual role of these temporal aspects, it is crucial to manipulate both f_0 contours and durational patterns. If tempo is relevant in the perception of sentence modality contrasts, we expect listeners to react differently to stimuli featuring the same f_0 contour but different durational patterns. Moreover, we can also expect that the effect of temporal manipulations will be stronger if the intonational cues are made unavailable or ambiguous.

The fourth author tested these hypotheses by having 26 subjects participate in a forced-choice identification task based on 18 manipulated stimuli [26]. These were created by manipulating two base stimuli, namely the sentence *Danilo vola da Roma* ('Danilo takes the Rome flight') read as a Question (bQ) and as a Statement (bS) (notation coherent with [26]). For both stimuli, we extracted phone durations (dQ , dS) and f_0 contours (fQ , fS). Then we defined an acoustically Ambiguous durational pattern (dA) as the average of corresponding phone durations in dQ and dS , and an acoustically ambiguous f_0 contour (fA), as the average of fQ and fS . This conceptual scheme was operationalised by applying landmark registration and averaging as explained in Section 2. For example, to obtain an Ambiguous f_0 contour in the duration pattern of a Statement, first apply landmark registration on fQ to its base timing ($fQdQ$) to synchronise it on the phone pattern dS , obtaining contour $fQdS$. Then compute the average $(1 - \alpha) \cdot fQdS + \alpha \cdot fSdS$ with $\alpha = 0.5$ to obtain $fAdS$, i.e. an ambiguous f_0 contour in the timing of the Statement base utterance. Finally $fAdS$ can be directly applied to bS , or also to bQ , provided that it is first manipulated by applying the duration transform from dQ to dS . This scheme allowed for the generation of 18 stimuli, i.e. the combination of three levels (Q, S and A) on two factors (f_0 and duration) applied on both base utterances.

The transformation of a base stimulus into its opposite sentence modality was extremely successful in that subjects' responses to original questions ($bQfQdQ$) were not significantly different from responses to statements re-synthesized as questions ($bSfQdQ$), and the same happened for the $bSfSdS$ - $bQfSdS$ pair. On the other hand, the creation of stimuli with ambiguous intonation did not yield the expected results, since subjects exhibited a strong question-bias. This is because in the absence of established knowledge on the warping of perceptual space for utterance-long stimuli, function averaging was accomplished by combining f_0 contour with equal weights $\alpha = 0.5$. Apparently, acoustical ambiguity does not always result in perceptual ambiguity. Despite this limitation, we were able to conclude that listeners do not seem to use temporal information when categorizing stimuli as questions or statements.

5. Conclusions and future work

In this paper we have presented a method for the rapid and effective manipulation of f_0 and segmental duration values aimed at the re-synthesis of stimuli for speech perception experiments. The method provides an automation layer between the level of specification of segmental alignment constraints and contour linear combinations on one hand, and the lower level provided by state-of-the-art editors, like the one available in Praat (PSOLA). The effectiveness of the method was illustrated by three use cases, where it was successfully applied in real experimental conditions. The software that implements the method is available for download and use [13]. Several extensions of the software are possible, for example an explicit mechanism for expressing rigid contour shift, which is a way to create timing variants (e.g. [27]).

6. References

- [1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.3.42) [computer program]," *online*: <http://www.praat.org/>, 2013.
- [2] R.-X. Yang, "The phonation factor in the categorical perception of mandarin tones," in *in Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.
- [3] E. Dombrowski and O. Niebuhr, "Shaping phrase-final rising intonation in german," in *in Proceedings of the 5th International Conference on Speech Prosody, Chicago, Illinois, USA*, 2010.
- [4] G. I. Ambrazaitis and O. Niebuhr, "Dip and hat pattern: a phonological contrast of german?" in *in Proceedings of the 4th International Conference of Speech Prosody, Campinas, Brazil*, 2008.
- [5] E. Dombrowski and O. Niebuhr, "Acoustic patterns and communicative functions of phrase-final f_0 rises in german: Activating and restricting contours," *Phonetica*, no. 62, pp. 176–195, 2005.
- [6] D. Hermes, "Stylization of pitch contours," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, I. Augurzky, P. Mleinek, N. Richter, and J. Schliesser, Eds. Berlin, New York: De Gruyter (= Language, Context, and Cognition 3), 2006, pp. 29–62.
- [7] O. Niebuhr, "The signalling of german rising-falling intonation categories - the interplay of synchronization, shape, and height," *Phonetica*, no. 64, pp. 174–193, 2007.
- [8] H. Quené. (2011) Software tools - adjustdurpitch.praat. [Online]. Available: <http://www.let.uu.nl/Hugo.Quene/personal/tools>
- [9] P. B. de Mareüil and V. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, pp. 247–267, 2006.
- [10] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis - 2nd Ed.* Springer, 2005.
- [11] C. de Boor, *A Practical Guide to Splines, Revised Edition*. Springer, New York, 2001.
- [12] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: <http://www.R-project.org/>
- [13] M. Gubian. (2013) Functional data analysis for speech research. [Online]. Available: <http://lands.let.ru.nl/FDA>
- [14] —. (2013) Rapid and smooth pitch contour manipulation - software tool. [Online]. Available: http://lands.let.ru.nl/FDA/papers/rapid_smooth_manipulation.zip
- [15] J. O. Ramsay, G. Hookers, and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer, 2009.
- [16] S. Winters and M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of f_0 and duration," *Speech Communication*, vol. 55, 2013.
- [17] K. Kinoshita, D. M. Behne, and T. Arai, "Duration and f_0 as perceptual cues to japanese vowel quantity," in *Proc. of the International Conf. on Spoken Language Processing, Denver*, 2002, pp. 757–760.
- [18] H. Kubozono, H. Takeyasu, M. Giriko, and M. Hirayama, "Pitch cues to the perception of consonant length in japanese," in *Proc. of the 17th International Congress of Phonetic Sciences Hong Kong*, 2011, pp. 1150–1153.
- [19] H. Altmann, I. Berger, and B. Braun, "Asymmetries in the perception of non-native consonantal and vocalic length contrasts," *Second Language Research*, vol. 28, no. 4, pp. 387–413, 2012.
- [20] D. M. Hardison and M. Motohashi-Saigo, "Development of perception of second language japanese geminates: Role of duration, sonority, and segmentation strategy," *Applied Psycholinguistics*, vol. 31, no. 01, pp. 81–99, 2010.
- [21] B. Chandrasekaran, P. D. Sampath, and P. C. M. Wong, "Individual variability in cue-weighting and lexical tone learning," *Journal of Acoustical Society of America*, vol. 128, no. 1, pp. 456–465, 2010.
- [22] M. D'Imperio and D. House, "Perception of questions and statements in Neapolitan Italian," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., Rhodes, 1997, pp. 251–254.
- [23] J. Ryalls, G. Le Dorze, N. Lever, L. Ouellet, and C. Larfeuil, "The effects of age and sex on speech intonation and duration for matched statements and questions in French," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2274–2276, 1994.
- [24] V. van Heuven and E. van Zanten, "Speech rate as a secondary prosodic characteristic of polarity questions in three languages," *Speech Communication*, vol. 47, no. 1, pp. 87–99, 2005.
- [25] F. Cangemi and M. D'Imperio, "Local speech rate differences between questions and statements in italian," in *Proceedings of the 17th International Congress of Phonetic Sciences*, W. Lee and E. Zee, Eds. Hong Kong: City University of Hong Kong, 2011, pp. 392–395.
- [26] F. Cangemi and M. D'Imperio, "Tempo and the perception of sentence modality," *Laboratory Phonology*, no. 4(1), in press, 2013.
- [27] O. Niebuhr and H. R. Pfitzinger, "On pitch-accent identification – the role of syllable duration and intensity," in *in Proceedings of the 5th International Conference on Speech Prosody, Chicago, Illinois, USA*, 2010.

Anonymising long sounds for prosodic research.

Daniel Hirst^{1,2}

¹LPL, UMR 7309 CNRS, Aix-Marseille University, Aix-en-Provence, France

²School of Foreign Languages, Tongji University, Shanghai, China

daniel.hirst@lpl-aix.fr

Abstract

It is more and more standard practice, in speech research, to make publicly available the data used in the research, in particular the speech recordings. This can potentially raise the problem of how to respect the anonymity of the speakers, particular if the recordings consist of unmonitored conversations, which may contain references to people by name or other material which it may be thought preferable not to make public. This paper describes a simple procedure, originally proposed by Paul Boersma, which has been implemented as a Praat script. The script replaces portions of the original recording annotated with a key word by a *hum* sound, which reproduces the prosodic characteristics (fundamental frequency and intensity envelope) of the corresponding original speech signal. The script described is freely available from the *Speech and Language Data Repository* (<http://sldr.org/sldr000526/en>).

Index Terms: database, speech prosody, anonymisation.

1. Introduction

Research on speech in the last decades has shown a tendency to make use of larger and larger quantities of speech recordings for analyses. It has, furthermore, become standard practice to make the primary data, as far as possible, publicly available in order to make it possible for other researchers to check and replicate the analyses.

Making speech recordings publicly available can, in some cases, come into conflict with the need to respect the anonymity of the speakers. This is particularly true when the material recorded consists of unmonitored spontaneous conversation material in which the speakers may make explicit reference to themselves or to other people by name. Besides the question of preserving anonymity, there may be other reasons for not wishing to allow some parts of the recorded material to be in full public access, when, for example, one of the speakers gives some intimate information about someone else.

One radical solution to this problem is that which has been adopted for the oral part of the British National Corpus (BNC [2]), and which consists of replacing the portions of the recording which contain personal information by a silence.

The obvious disadvantage of this solution is that it makes it impossible to use those parts of the corpus for prosodic analysis since the prosodic information has been removed along with the speech signal. A preferable solution would be to modify the speech signal so that the lexical content of the signal is no longer recognisable but without removing the prosodic information.

2. Anonymising speech signals

In psycho-acoustic experiments, one solution which is often used is to ‘invert’ the spectrum of the speech signal ([3], chap 33 p. 619). This renders the speech signal unintelligible without removing the prosodic information. Unfortunately, a second inversion of the spectrum of the acoustic signal will restore the original signal and consequently the technique is not appropriate for the task we are describing.

A second technique, also used in psycho-acoustic experiments, renders the speech signal inaudible by applying a (usually low-pass) filter, which removes most of the spectral information without removing the prosodic information. The problem with this technique is that even quite severe filtering can sometimes still leave the original signal intelligible. In order to be sure that the signal is not intelligible, the filtering tends to make the original signal sound so muffled that it is difficult to use it for perceptual evaluation of prosodic information, for example.

A different technique was suggested by Paul Boersma on the Praat-users mailing list in 2006, (<http://uk.groups.yahoo.com/group/praat-users/message/2537>) in response to a query on how to anonymise speech data.

The algorithm proposed consisted of the following steps:

1. Select the word in the editor window
2. Extract the selected word as a Sound to the Objects list (File menu)
3. Choose Get Intensity (dB) from the Query menu
4. Choose To Pitch
5. Choose To Sound (hum)
6. Select the original extracted word in the Objects list
7. Choose To Intensity
8. Choose Down To IntensityTier
9. Select the hummed sound plus the IntensityTier
10. Choose Multiply
11. Select the sound and choose Scale Intensity from the Modify menu
12. in the Scale Intensity window, type the value from step 3 and click OK
13. In the window with the original sound, select the same times as in step 1.
14. Choose Cut (Edit menu; the word disappears).
15. Select the hummed sound and choose Edit.

16. In the sound window, select the whole sound (the hum) and choose Copy (Edit menu).
17. Go back to the window with the original sound and choose Paste (Edit menu).

Paul Boersma noted that since there may be some irregularities at the edges, some of the cutting and copying could be carried out on zero crossings.

3. A Praat script to anonymise long sounds

The algorithm described in the preceding section was implemented as a Praat script which can be used to anonymise a whole folder of speech sounds which are treated as long sounds.

Instead of using the editor as described in the algorithm, the script uses a TextGrid annotation object on which a specified tier contains target labels indicating the portions of the signal which need to be 'anonymised'. The original sound is then 'dissected' and the labeled segment of sound is treated as described in the algorithm before being splice back with the rest of the sound.

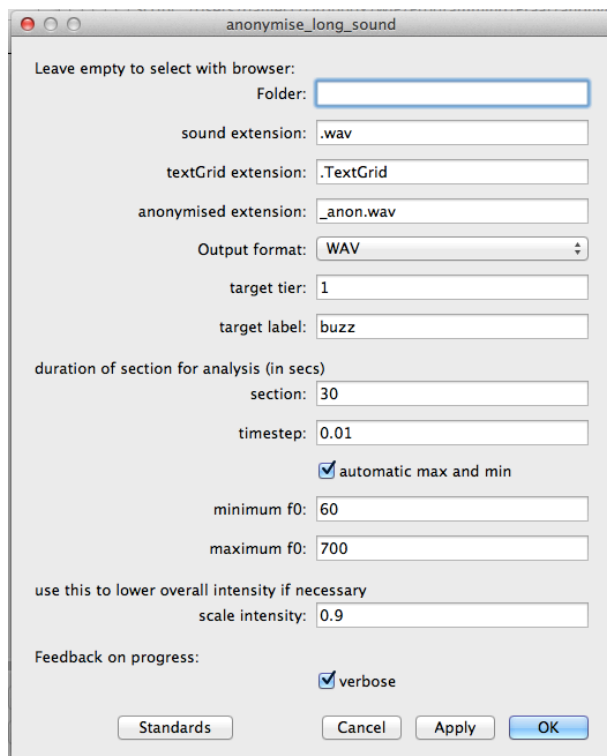


Figure 1: Script window for the Praat script `anonymise_long_sounds.praat`.

The path to the folder can either be specified directly in the script window, or selected with the browser (see Figure 1).

By default, the script expects to find Sounds and TextGrids in the folder with the extensions `.wav` and `.TextGrid`, respectively. The script can also read and write other sound formats.

Each sound is associated with a TextGrid containing at least one interval tier as target (the default target tier is 1). Intervals which are labelled with the *target label* (default is *buzz*) will be replaced by a *hum* sound which has the same pitch and intensity envelope as the original sound.

In order to be able to apply the script to sounds with a duration exceeding that which can be handled in the computer's RAM, the sounds are treated as LongSound objects, which means that instead of loading the whole sound into the computer memory, only one section of the sound (the duration of which is specified in seconds (default = 30) is treated at a time.

Figure 2 shows the sentence "Could you arrange to send an engineer on Tuesday morning please?" with the f0 display and intensity profile. The words "an engineer" have been marked with the label "buzz".

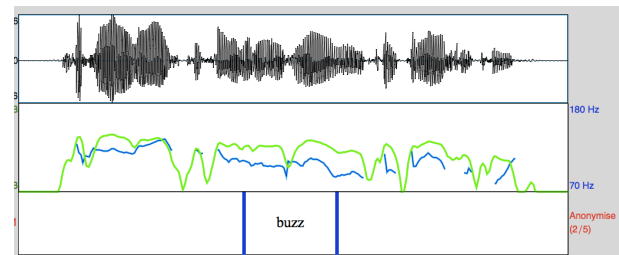


Figure 2: The sentence "Could you arrange to send an engineer on Tuesday morning please?"

In Figure 3, the words "an engineer" have been replaced by a buzz with the same pitch and intensity profile. The words are thus incomprehensible but the sentence can be used for the analysis of prosodic parameters without any appreciable loss of information.

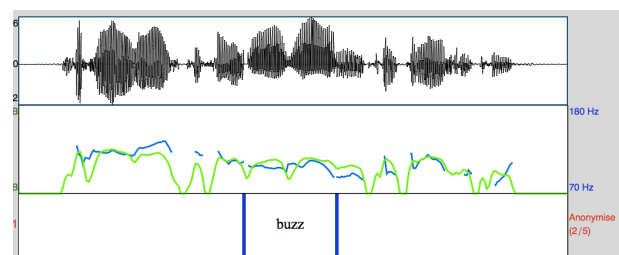


Figure 3: The same sentence as that in figure 2 with the words "an engineer" replaced by a buzz with the same pitch and intensity profile.

4. Conclusions

The script is freely downloadable from *SLDR*, the Speech and Language Data Repository, at the following address:

<http://sldr.org/sldr000526/en>

5. References

- [1] Boersma, P.; D. Weenink, D. Praat, a system for doing phonetics by computer. <http://www.praat.org> [version 5.3.41, February 2013], 1992 (2013).
- [2] Burnard, Lou (ed.) Reference Guide for the British National Corpus (XML Edition). British National Corpus Consortium, Research Technologies Service, Oxford University Computing Services, <http://www.phon.ox.ac.uk/SpokenBNC> 2007
- [3] Smith, Steven W. Digital Signal Processing: A Practical Guide for Engineers and Scientists Elsevier Science, Burlington, MA. 2003.

ModProso: A Praat-Based Tool for F0 Prediction and Modification

Juan María Garrido¹

¹Department of Translation and Language Sciences, Pompeu Fabra University,
Roc Boronat 138, 08018 Barcelona, Spain
juanmaria.garrido@upf.edu

Abstract

In this paper we describe ModProso, a Praat-based tool for prediction and modification of F0 contours in natural utterances. A general overview of the tool is given, and a brief description of the several steps carried out in the F0 contour generation are provided.

Index Terms: F0 contours, Analysis-by-synthesis, Speech Synthesis

1. Introduction

This paper presents ModProso, a Praat-based tool [1] for the perceptual evaluation of 'synthetic' F0 contours predicted from a chain of symbolic labels. It works in a similar way to other existing tools for F0 manipulation and prediction, as ProZed [2], in the sense that it replaces the original F0 contour of a natural utterance by the F0 contour predicted from the intonational labels given as input, but it accepts a different inventory of intonational labels (the ones predicted by the intonational model described in [3,4,5]). It was developed as a research tool to perceptually evaluate the output of the automatic F0 stylisation, annotation and modelling tool described in [5], but it has also been used to generate synthetic stimuli with modified F0 contours for several purposes.

ModProso was originally designed for its use with speech utterances in Spanish and Catalan, but current research in being carried out to adapt it to Brazilian Portuguese and Mandarin Chinese. Adaptation to other languages could be also done with a minimum effort.

2. Background

The tool assumes the model for intonation description proposed in [3,4]. This model conceives F0 contours as the result of the superposition of two types of F0 patterns, as shown in Figure 1

- **Local:** typical F0 shapes occurring at Stress Group (SG) level.
- **Global:** global evolution of an F0 contour along an Intonation Group (IG).

F0 contours can be viewed then as the sum of three types of local patterns, **initial**, **middle** and **final**, depending on its position within the IG, which are superimposed to a global pattern determining its relative height within the speaker F0 range.

Local patterns are modelled as sets of F0 turning points anchored to specific parts of the syllables that make up SG. Each pattern is identified with a label which includes information about:

- the level of the F0 points that make up the pattern: P (Peak), P+ (extra high peak), V (Valley) and V- (extra

low valley), depending on the relative height of each F0 point within the F0 range of its container IG;

- the syllable which contains the point: 0 (the stressed one), 1 (one after the stressed one), -1 (one before the stressed one), etc.
- the position of the point within the syllable: I ('initial', close to the beginning of the syllable nucleus), M ('middle', close to the centre of the nucleus), and F ('final', close to the end of the nucleus).

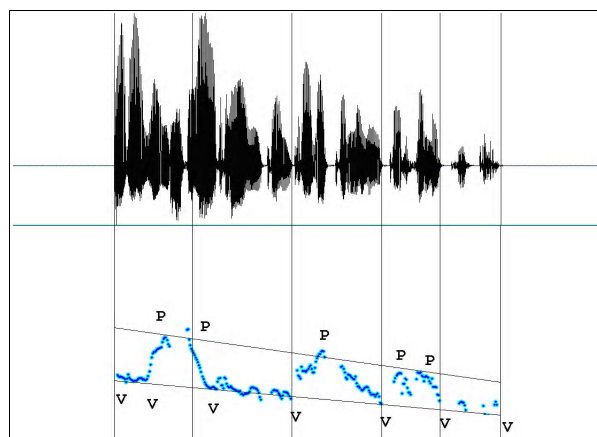


Figure 1: Waveform and F0 contour of the utterance "Aragón se ha reencontrado como motor del equipo", uttered by a Spanish female speaker. Vertical solid lines represent SG boundaries.

So for example, a pattern labelled as VI0_P M0_P I1, as the one shown in figure 2, is made up of three F0 inflection points: a V point located at the beginning of the stressed syllable of the container IG (I0); a P point in the middle of the stressed syllable (M0); and a P point at the beginning of the syllable after the stressed one (I1).

Global patterns are modelled as reference lines predicting F0 values as a function of time along the IG, as can be observed in figure 1. The model distinguishes several types of pattern lines (**initial**, **middle** and **final**), according to the position of the IG within its container sentence.

Both local and global patterns for a given utterance can be obtained automatically using MelAn, the modelling tool described in [5]. The tool stores the full listing of local patterns detected in the input utterance within a '.contour' file as the one shown in table 1.

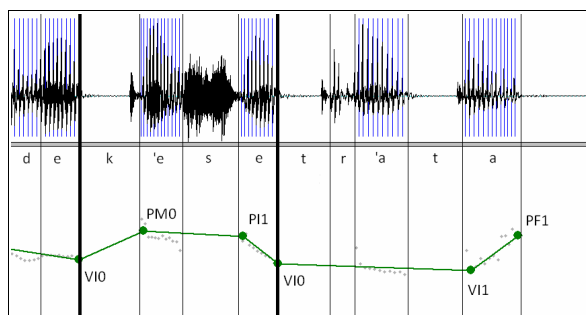


Figure 2: Notation example of two F0 patterns of the utterance '¿Quiere alguien explicarme de qué se trata?', uttered by a female speaker. Vertical solid lines represent SG boundaries.

```

VI-1_PI0_VI1_INICIAL_4.pattern
VI0_PM0_PM1_INTERIOR_4.pattern
VI0_VF0_FINAL_1.pattern
VF-3_PM-1_VF0_INICIAL_4.pattern
0_INTERIOR_3.pattern
VI0_PM0_INTERIOR_3.pattern
PI0_VM0_PI1_VM1_PI1_VF1_PI2_INTERIOR_3.pattern
VI0_VF1_INTERIOR_2.pattern
PM0_P+I1_PI2_INTERIOR_4.pattern
PI1_INTERIOR_3.pattern
VI2_INTERIOR_3.pattern
VF0_PI2_INTERIOR_3.pattern
VM0_PI1_INTERIOR_2.pattern
VF0_PI1_VI2_PF3_INTERIOR_4.pattern
PI0_VM0_VF0_FINAL_ENUNCIADO_1.pattern

```

Table 1. Example of 'contour' file containing the list of F0 patterns for the Brazilian Portuguese utterance 'Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.

Reference lines for global patterns are calculated as two regression lines approaching the P and V points respectively of any IG found in the input utterance. The result is stored in two separate files ('regression_P' and 'regression_V'), which contain the initial F0 of the calculated line, and its slope, in the format shown in table 2.

```

"x"
"(Intercept)" 257.848602365733
"Tiempo" -6.86524557236724

```

Table 2. Example of 'regression_P' file containing the initial F0 value and slope for P regression line of the Brazilian Portuguese utterance 'Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.

These three generated files ('contour', 'regression_P' and 'regression_V') can be used directly as input for ModProso to make 'analysis-by-synthesis' modification of F0 contours.

3. Description of the tool

3.1. General Overview

ModProso performs basically two tasks:

1. the prediction of a chain of F0 target values, in the form of a Praat-style stylised contour, using as input a list of local pattern labels contained in a 'contour' file, and two P and V reference lines, stored as regression lines in 'regression_P' and 'regression_V' files;
2. the substitution of the original F0 contour by the predicted one in the speech utterance provided as input.

To perform these two tasks, ModProso needs as input:

1. a wav file containing the utterance to be manipulated;
2. a Textgrid file containing the orthographic and phonetic transcription of the provided utterance, and its prosodic segmentation into syllables, SG, IG and breath groups (BG), as shown in Figure 3;
3. a 'contour' text file containing the list of pattern labels;
4. a couple of 'regression_P' and 'regression_V' files containing the values for the global F0 reference lines.

All three 'contour', 'regression_P' and 'regression_V' files required as input can be both files obtained from the automatic analysis of a given utterance using MelAn, or 'theoretical' files artificially built for research purposes.

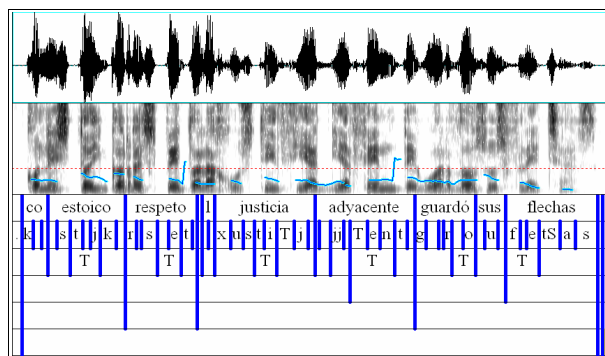


Figure 3: Speech waveform and TextGrid file for the Spanish utterance 'Con estoico respeto a la justicia adyacente guardó sus flechas', uttered by a male speaker. It contains the necessary tiers to be used as input by ModProso: orthographic transcription (tier 1), phonetic transcription (tier 2), syllable segmentation (tier 3), SG segmentation (tier 4), IG segmentation (tier 5) and BG segmentation (tier 6).

Figure 3 presents a workflow diagram representing the processing steps from an input wav file to a new audio file containing the original utterance modified with the predicted F0 contour.

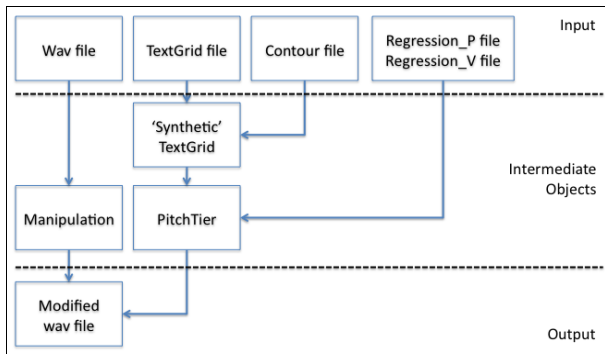


Figure 4: Workflow diagram showing the processing steps in the generation of a wav file with a predicted F0 contour using ModProso.

At the end of the process, an edition window showing the speech signal and the obtained stylised contour, as the one shown in figure 5, is displayed. Using this window, the user can obtain a synthesised version of input utterance using Overlap-Add or LPC techniques, which can be played directly or stored in an output wav file.

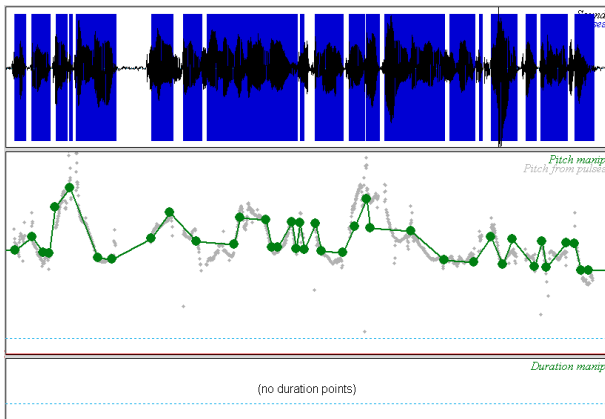


Figure 5: Praat edition window showing the speech signal and the predicted F0 stylised contour for the Brazilian Portuguese utterance 'Depois de tanto caminhar, amanheceu dia na luz da manhã descobriram trinta ou quarenta moinhos de vento que há no Campo de Montiel', spoken by a female speaker.

3.2. Prediction of the F0 chain

The generation of a 'synthetic' F0 contour is carried out in two steps:

- **Label alignment:** the labels contained in the 'contour' file are anchored to the predicted places in syllables within its corresponding SG of the input utterance.
- **F0 calculation:** F0 values for each inflection point predicted by the labels are calculated using the P and V regression lines.

In the label alignment phase, a new point tier is added to the input TextGrid showing the alignment of the labels with the signal, as shown in figure 6, to generate an intermediate TextGrid file ('TextGrid_generado'), to be used in the F0 calculation process.

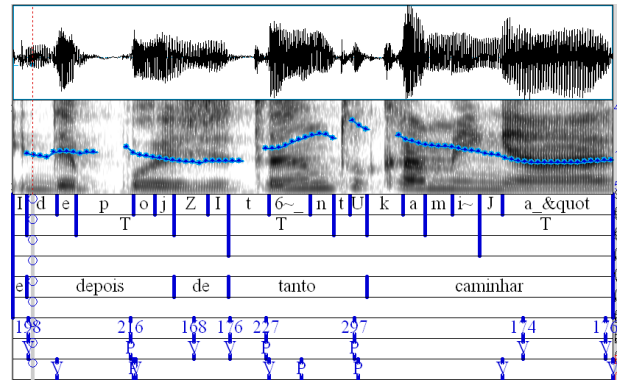


Figure 6: Speech waveform and TextGrid for the Brazilian Portuguese utterance 'e depois de tanto caminhar', uttered by a female speaker. Last tier shows the predicted alignment of the input labels; they can be compared with the ones obtained with MelAn, appearing in the previous tier

In this second phase, F0 values for each P and V value are calculated using their predicted time alignment value stored in the intermediate 'TextGrid generado' file and the regression lines provided as input. After this process, the obtained chain of F0 values is converted into a PitchTier object and then stored in a second intermediate file ('PitchTier_sintetico'), that will be used in the final F0 contour substitution process.

3.3. F0 contour substitution

Finally, the contour substitution process, which is carried out in two steps, takes advantage of the F0 manipulation facilities available in Praat:

1. A 'Manipulation' Praat object is created from the input wav file.
2. The original F0 contour in the obtained 'Manipulation' object is replaced by the PitchTier loaded from the intermediate 'PitchTier_sintetico' file. This modified 'Manipulation' object is the one which is presented to the user in the final edition window.

4. Applications

ModProso has shown to be useful to perceptually evaluate the symbolic representation of F0 contours automatically obtained with MelAn. The results of the experiments carried out with Spanish and Catalan speech corpora, presented in [5], showed that listeners evaluated the synthesised F0 contours as reasonably similar to the original ones, both in Spanish (mean rate 4.05 over a maximum of 5) and Catalan (mean rate 3.93). A similar experiment is being designed to carry out the same evaluation for Brazilian Portuguese.

ModProso has also been used for the generation of manipulated F0 stimuli in other perception experiments, such as the one described in [6], in which the perceptual interpretation of some final F0 patterns used in emotional speech in Spanish was evaluated.

5. References

- [1] Boersma, P. and Weenink, W., Praat: doing phonetics by computer [Computer program] <http://www.praat.org/>, 2012.
- [2] Hirst, D., ProZed: A speech prosody analysis-by-synthesis tool for linguists, Speech Prosody 2012. Online:

- http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_70.pdf, accessed on 24 Apr 2013.
- [3] Garrido, J. M., *Modelling Spanish Intonation for Text-to-Speech Applications*, Ph. D Thesis, Universitat Autònoma de Barcelona, 1996. Online: <http://www.tdx.cat/handle/10803/4885;jsessionid=376A9A0BED1D5E6DED7CDFD3880316F3.tdx1>, accessed on 24 Apr 2013.
- [4] Garrido, J. M., "La estructura de las curvas melódicas del español: propuesta de modelización", *Lingüística Española Actual*, XXIII/2, 173-209, 2001.
- [5] Garrido, J. M., "A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora", *Speech Prosody 2010*: 100041. Online: <http://speechprosody2010.illinois.edu/papers/100041.pdf>, accessed on 24 Apr 2013.
- [6] Garrido, J. M., Laplaza, Y. and Marquina, M., "On the use of melodic patterns as prosodic correlates of emotion in Spanish", *Speech Prosody 2012*, Shanghai, 2012. Online: http://www.speechprosody2012.org/uploadfiles/file/sp2012_submission_57.pdf, accessed on 24 Apr 2013.

Automatic labelling of pitch levels and pitch movements in speech corpora

Piet Mertens

Leuven University (KU Leuven), Linguistics Department, Belgium

Piet.Mertens@arts.kuleuven.be

Abstract

We describe a system for the automatic labelling of pitch levels and pitch movements in speech corpora.

Five pitch levels are defined: *Bottom* and *Top* of the speaker's pitch range, as well as *Low*, *Mid*, and *High*, which are determined on the basis of pitch changes in the local context. Five elementary pitch movements of individual syllables are distinguished on the basis of direction (rise, fall, level) and size (large and small melodic intervals, adjusted to the speaker's pitch range). Compound movements consist of a concatenation of simple ones.

The labelling system combines several processing steps: segmentation into syllabic nuclei, pause detection, pitch stylization, pitch range estimation, pitch movement classification, and pitch level assignment. Unlike commonly used supervised learning techniques the system does not require a labelled training corpus.

This approach results in an automatic, fine-grained and readable annotation, which is language-independent, speaker-independent and does not depend upon a particular phonological model of prosody.

Index Terms: speech prosody; transcription; annotation; automatic labelling; pitch range

1. Introduction

Prosodically annotated corpora, indicating prominence, stress, pitch levels, pitch movements, and prosodic units, enable systematic and quantified analyses of prosodic forms (tones, pitch contours) occurring in speech, of their distribution, their relation to syntax, their functions in discourse, and so on. In addition, they may be used in speech technology applications, such as text-to-speech synthesis and speech recognition. Large-scale prosodically annotated corpora are scarce, except for English. Manual annotation of prosody is so time-consuming that only automatic annotation is feasible. This paper describes a system for the automatic transcription of pitch-related aspects of prosody which is language-independent and may be applied to many languages.

Every system for automatic annotation of prosody faces the fundamental question about which aspects of prosody should be transcribed, and in what way.

A comparison of phonological intonation models for a given language immediately shows the lack of consensus, for each and every aspect of prosody: the nature of stress, the treatment of pitch variations (movements or targets, pitch levels, pitch range), the nature of prosodic units, and so forth. These differences reflect incompatible theoretical choices

about what is relevant (i.e. distinctive) in prosody and how it should be represented.

Most people may not be able to describe intonations analytically, but they are able to discriminate between intonations and to imitate a particular intonation, by repeating it or humming it. The lack of consensus therefore is likely to be due to theoretical preferences rather than to major perceptual differences between listeners.

To be useful to researchers from various approaches, and a fortiori to linguists and users without a background in intonation research, the annotation should not involve theoretical concepts such as prosodic units or contours and only indicate those pitch variations which may be heard by the average listener. This may be achieved by simulating tonal perception.

To avoid language-specific phonological choices, generic labelling schemes may be used. The suprasegmental diacritics of the IPA indicate pitch level, pitch movement, stress, boundaries, etc. The INTSINT notation [1, 2, 3] is based on the inventory of pitch contrasts found in published descriptions of intonation. It distinguishes absolute levels (Top, Mid, Bottom), relative levels (Higher, Same, Lower), and iterative relative levels (Up-stepped, Down-stepped). However, it does not provide symbols for pitch movements.

The system for automatic annotation described here first simulates tonal perception. In a later step, the perceived pitch event associated with a syllable in the speech signal is further categorized with respect to pitch level and pitch movement, taking into account the pitch range of the individual speaker. This categorization results in a label indicating the type of pitch movement (level, rise, fall, rise-fall, etc.) and the pitch level (low, mid, high, bottom, top) associated with each syllable in the speech signal.

Figure 1 illustrates the output obtained by the automatic annotation system. The upper part shows acoustic parameters, segmentations and the pitch stylization (cf. section 3.4). Four annotation tiers are shown: the phonetic alignment, the syllable alignment, the orthographic words, and the tonal label for each syllable. The first three tiers are provided by the speech corpus, the last is computed automatically. In this particular example, most syllables receive the label "L", indicating they are pronounced on a low pitch level (cf. section 3.7). The syllable "brève" carries the label "MR", indicating it is pronounced with a large rise ("R"), starting from the mid pitch level ("M"). A compound pitch movement is noted as a sequence of simple ones, as shown by the label for the syllable "na", which indicates that the syllable starts at a low level ("L") and contains a large rise ("R") preceded and followed by a level plateau ("."). (All three rises would be called "late pitch movements" in the IPO approach.)

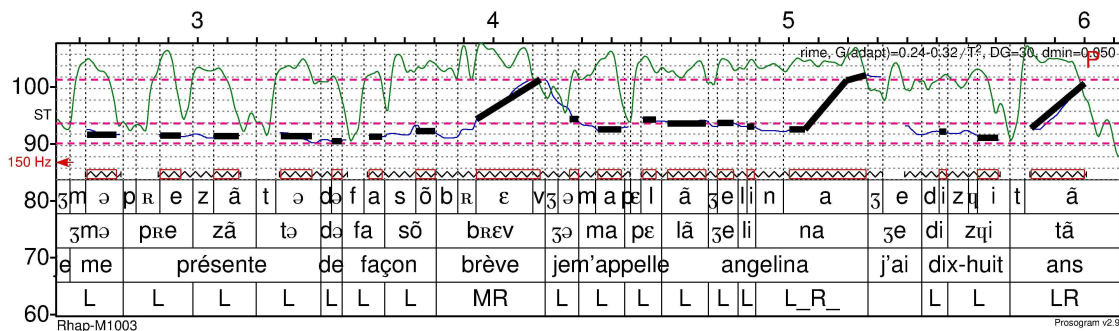


Figure 1. Automatic tonal annotation for the French utterances “*Je me présente de façon brève. Je m’appelle Angelina. J’ai dix-huit ans.*” [Rhap-M1003] (“*I briefly introduce myself. My name is Angelina. I’m 18 years old.*”), by a female speaker. The automatic prosodic labelling is shown in the lower tier. The upper part of the figure shows the acoustic parameters of intensity (continuous thin green line), voicing (saw tooth), fundamental frequency (thin blue line, mostly covered by the thick black line), as well the pitch stylization (thick black line). Pitch is plotted on a semitone (ST) scale (relative to 1Hz), with horizontal calibration lines (black dotted lines) at 2 ST steps. The three horizontal dashed lines in red indicate the pitch range of the speaker. The lower part shows various corpus annotation tiers: phonetic alignment, syllables, words, and tonal labels. The syllabic nuclei appear as red boxes on top of the voicing line (saw tooth). The “P” at 6 s indicates the start of a pause detected by the system.

2. The proposed labelling scheme

2.1. Pitch levels

In the proposed annotation, *pitch levels* are defined in two ways: *locally*, i.e. relative to the context, and *globally*, i.e. relative to the speaker's pitch range. The global interpretation results in the pitch levels *top* (T) and *bottom* (B). The local interpretation is based on pitch changes occurring between or within syllables in the near context and results in pitch levels *low* (L), *mid* (M) and *high* (H).

Two or more syllables at the same (local) pitch level and located at different points in the utterance, need not have the same fundamental frequency, but may differ considerably, provided there are local pitch changes motivating these differences. Since these pitch levels are based on *local* changes, they are compatible with the *declination line* phenomenon (see [4], p. 16).

2.2. Pitch intervals

An individual voice may be characterized by its *central pitch* and its *pitch span* [2, 4]. The central pitch (or *key*) opposes low pitched and high pitched voices. The pitch span, in contrast, indicates the interval between the lower and upper pitches used by the speaker in modal speech. The large variability in the pitch range of individual speakers calls for an interpretation of *pitch intervals* which is *relative to the individual speaker's range*.

The *number of pitch interval categories* used varies between models. Autosegmental models [4, 5, 6] typically postulate two pitch levels, and hence one size of pitch interval, while a specialized treatment is used for small size intervals (as found in “downstep” and “boundary tones”). Models such as the IPO model [7], INTSINT [1], or RaP (Rhythm and Pitch, [8]) distinguish large and small pitch intervals, where the latter typically occur in “down-stepping” or “up-stepping”.

The proposed annotation distinguishes *two sizes* of pitch intervals: *large* and *small* ones. Their size (in ST) is adjusted to the individual speaker’s pitch range, and such that small intervals exceed the size of micro-prosodic variations. This is in agreement with [9] who suggests that only differences exceeding 3 ST play a role in speech communication. The thresholds in table 1 were determined empirically by the author on the basis of data for 42 speakers.

Pitch range	Large interval	Small interval
> 8.5 ST	> 4.5 ST	3.0 - 4.5 ST
7.0 – 8.5 ST	> 3.5 ST	2.5 - 3.5 ST
< 7.0 ST	> 3.2 ST	2.5 - 3.2 ST

Table 1. *Thresholds for large and small pitch intervals used for pitch movement and pitch level determination, depending on the pitch range obtained for a given speaker.*

2.3. Symbols used in the labelling scheme

The notation used indicates (1) whether a given syllable presents an audible pitch variation or not, i.e. whether it is flat (level), rising or falling; (2) it distinguishes between large and small movements; (3) it allows for compound movements; (4) it indicates pitch level taking into account pitch range.

Pitch levels are indicated by “L” (low), “H” (high), “M” (mid), “T” (top of range) and “B” (bottom of range). *Pitch movements* will be represented by “R” (large rise), “F” (large fall), “r” (small rise), “f” (small fall) and “_” (flat). *Compound movements* use a sequence of these symbols: “RF” (rise-fall), “_R” (level-rise), “R_” (rise-level), and so on. Two additional symbols have a special status. First, “S” (sustain) indicates a syllable with a uniform level pitch and minimal duration of 250 ms, a marked contour which is fairly rare in French. Second, the symbol “C” (creek) indicates a syllable with creek (see section 3.1).

Although the annotation allows for compound intra-syllabic pitch movements of any complexity, such movements are fairly rare, even in spontaneous speech (less than 1% of the syllables in a 65 min. corpus of French).

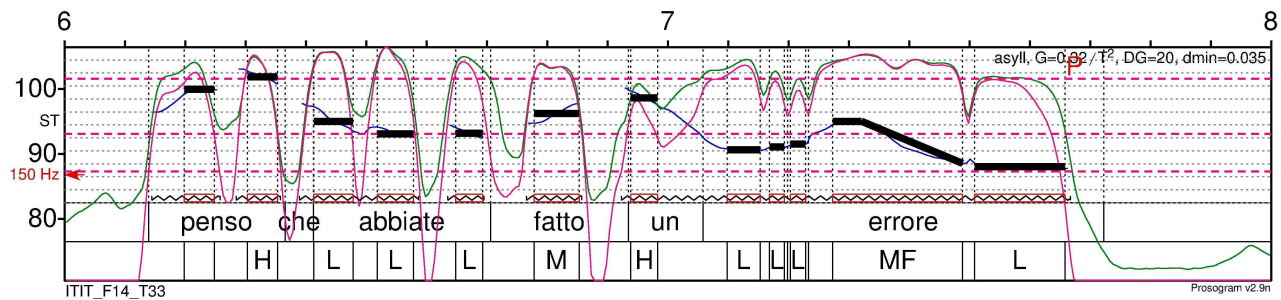


Figure 2. Automatic prosodic labelling of the utterance “penso che abbiate fatto un errore” (“I think you made a mistake.”), by a female speaker [ITIT_F14_T33] (OpenProDat corpus, [22]). The automatic labelling is shown in the lower tier. Acoustic parameters and pitch range are shown in the same way as in figure 1.

The pitch movement of a syllable is always identified, since it can be determined on the basis of F0 only. The *pitch level*, however, *may not be detected* for a given syllable (typically when the left context does not contain pitch changes). In such a case the pitch movement will be shown without the pitch level. When pitch movement is level and simple, the “_” is skipped, for conciseness: “H_” is simplified to “H”, whereas “H_R” and “HR_” are noted as such, in order to distinguish all three shapes. Moreover “_” (flat with missing pitch level) is skipped altogether.

The *pitch level reached at the end of the syllable* is indicated for “B” and “T”, when the syllable’s pitch contour starts at a different pitch level. For instance, “HF,B” indicates a high fall reaching the bottom level. This is justified by the fact that such cases combine the effect of “HF” and “B”.

3. The procedure for automatic annotation

For an overview of the approaches to automatic labelling of prosody, see [10].

The automatic annotation of pitch features includes several processing steps, stemming from the overall approach, which first simulates tonal perception and then categorizes the resulting pitch movements and levels, while taking into account the speaker’s pitch range, [10]. The processing steps are described below. The system is implemented as a script for the Praat speech analysis software [11].

3.1. Parameter extraction

Acoustic parameters are calculated using algorithms provided by Praat, with their default settings, except for the time step (frame rate), which is set to 5 ms. The voicing decision (V/UV) is derived from the F0 confidence (periodicity). Although the system does not include creak detection, it will use the creak annotation tier, when this is available.

3.2. Segmentation into syllabic nuclei

The *syllable* is a central unit for many aspects of prosody, including prominence, stress, syllable duration, pitch movements, speech rate and rhythm. Moreover, intensity changes and spectral changes within a syllable affect the perception of its pitch variation [12, 13, 14]. For this reason measurements are applied to the *syllabic nucleus*, which may be broadly characterized as the central part of the voiced area of a syllable rhyme (vowel and coda, as determined from the phonetic alignment), located around its local peak of intensity, for which the intensity only decreases to some amount, specified by a threshold (2 dB for left side, and for

right side relative to intensity dip at right boundary of the syllable). (Various segmentation types are supported: rhymes, syllables, vowels, or fully automatic.)

3.3. Detection of pauses

Silent pauses affect pitch perception, by lowering the glissando threshold [15]. In order to take this into account, speech pause detection is needed. When the gap between the end of a syllabic nucleus and the beginning of the next exceeds 350 ms, it is interpreted as a pause.

3.4. Pitch stylization

The next step applies a stylization to the F0 data, based on a model of *tonal perception* in speech [16, 17, 18, 19]. For each syllabic nucleus, the pitch contour is divided into one or more parts of uniform slope (“tonal segments”), on the basis of a perceptual threshold for slope change (the differential glissando threshold). For each part the pitch change is compared to the glissando threshold [20, 7] in order to determine whether the measured variation is perceived as a glissando or not. This model results in a representation of the audible pitch events in an utterance, as a sequence of forms, which is less complex than the acoustic data itself.

3.5. Automatic detection of the speaker’s pitch range

Information about pitch range will be used in three ways: (1) to discard pitch values outside the pitch range of the speaker; (2) to assign a pitch level to pitch values near both ends of the range; (3) to adjust pitch interval categories (small and large) to the pitch span of the speaker.

Unreliable values are discarded: syllables with octave jumps, creak, hesitations, outliers (≥ 18 ST from mean; which exceeds the average pitch range observed in a large corpus including many speakers, male and female). For each syllable pronounced by a given speaker, two pitch values are obtained: the minimum and maximum pitch inside the syllabic nucleus. The 2th and 98th percentiles of this set of data provide an estimate of the bottom and top of the global pitch range, respectively. In this way, outliers due to pitch detection errors and co-intrinsic pitch phenomena are mostly eliminated. Pitch range detection is based on all syllables for a given speaker in the corpus, rather than on individual utterances.

3.6. Intra-syllabic pitch movements

For each tonal segment (cf. section 3.4), the observed pitch variation is compared with the glissando threshold and

variations below the threshold are normalised to level pitch segments. The glissando threshold used by the stylization is set to $0.32/T^2$, except for syllables followed by a pause, where a threshold of $0.16/T^2$ is used (threshold for isolated stimuli in psychoacoustics, [20]). Next, pitch segments with an audible pitch variation are further categorized as large or small pitch intervals. This results in the elementary forms for intra-syllabic pitch movements used in the labelling scheme of section 2.

3.7. Pitch level detection

Various types of information are used: the speaker's pitch range, the pitch changes between successive syllables in the near context, and the intra-syllabic pitch movements. In addition, when pitch level cannot be determined directly, it may often be derived indirectly from the identified pitch level of neighbouring syllables. These cases are examined in the order indicated below, until the pitch level is detected. For some syllables, however, pitch level remains unidentified.

For syllables where F0 starts above the top or below the bottom of the estimated pitch span, the pitch level will be set to T (top) or B (bottom) respectively, provided the pitch range can be determined reliably (at least 200 syllables for this speaker) and the pitch span is sufficiently wide.

The pitch variation in the left context of the *target* syllable, i.e. the syllable to be labelled, may be used to *infer its pitch level*. For instance, when the start pitch is sufficiently higher than the lower pitch value in the left context, the target is high within that context. The local context consists of up to 3 syllables (of the same speaker) preceding the target syllable without an intervening pause and occurring within a window of 500ms. Syllables tagged as hesitations are discarded from the context, as well as syllables with a top or bottom pitch level. For instance, in figure 1, the left context of syllable "brève" (3.8s), has as its lower point syllable "de", and the pitch interval separating them results in level M for "brève".

When a syllable contains a large pitch variation, this variation also provides information about the pitch level at the start of that syllable. In this case the information about the position in the pitch range is also taken into account.

For syllables where pitch level remains unknown after the previous steps, detected pitch levels in the immediate context will be used as a *reference*, by measuring the pitch interval between a target syllable (with unknown pitch level, but known F0) and an adjacent or near syllable. The procedure is applied with increasing context size, looking first for an adjacent reference, then for a more distant one, but within a time window of 0.5s separating the target from the reference.

Pitch level detection relies mainly on pitch *changes*; it is not effective for sequences of level syllables pronounced at the same pitch level. Such plateaus receive a pitch level according to their position in the pitch range.

4. The resulting tonal annotation

Figure 2 illustrates the results obtained by the automatic annotation for an utterance in Italian, taken from the OpenProDat corpus [23]. Since no phonetic alignment was available, the automatic segmentation provided by Prosogram was used. The word annotation was added manually for the purpose of interpreting the results. The figure illustrates (1) the distinction between syllables with a glissando (such as the fall on the second syllable of "errore") and those with a steady

pitch (all other syllables), (2) the distinction between gradual changes (as for "abbiate fatto un") and abrupt pitch changes (as between "penso" and "che"), (3) the local interpretation of pitch level: syllables with the same pitch level may be at different frequencies, as is the case for the *H* levels in "penso", "un", and the *L* levels in "abbiate", and "errore".

A preliminary evaluation of the system for automatic transcription of pitch movements and levels is given in [24].

5. Conclusion

The proposed annotation system has several interesting properties. First, it provides a very *narrow transcription* of pitch movements (their direction and size), pitch level and pitch range (bottom, top).

Second, the approach allows for a *speaker-independent* annotation of tonal features. The system automatically adapts to the speaker, by calculating his pitch span and key and by adapting accordingly various thresholds used in the system.

Third, the tonal annotation system is *language-independent*. It does not refer to properties of particular languages. As a result, the system may be applied to many languages, to obtain a tonal annotation for existing speech corpora.

Fourth, the system uses little information other than the *acoustic* signal itself. In this study, the phonetic alignment was used to avoid segmentation errors having an impact on the tonal annotation. Many speech corpora already include an annotation of phonemes and syllables. Moreover, the system may also be applied using a fully automatic segmentation of the speech signal, resulting in an annotation tool which does not require any annotation whatsoever.

Fifth, the approach described in this paper does not require a *training corpus*. This constitutes a major advantage over common techniques for automatic classification by supervised learning, which all require such corpora. Since the validation of corpora (both training and reference corpora) is extremely time consuming [21, 22], the need for training corpora constitutes a major obstacle for the realization of automatic annotation systems for new prosodic transcriptions, for which such corpora are lacking. This obstacle does not apply to our system.

Finally, the transcription is not linked to a particular phonological model of prosody. Instead it is "*theory-friendly*" [2, 3], because it is compatible with a number of theoretical approaches to the representation of tonal aspects in speech. It would be fairly straightforward to map the obtained tonal annotation to other annotation schemes.

Our future research will focus on the detection of other aspects of prosody in continuous speech, including prominence, lengthening, stress and prosodic boundaries. The combination of these prosodic features with the tonal aspects will result in a more comprehensive transcription of prosody. However, since some of these aspects are language- or theory-dependent, the resulting transcription will follow a particular phonological model for a given language.

6. References

- [1] Hirst, D. J. and Di Cristo, A., "A survey of intonation systems", in Hirst, D. and Di Cristo, A. [Ed], *Intonation Systems. A Survey of Twenty Languages*, 1-44, Cambridge University Press, 1998.

- [2] Hirst, D. J., "Form and function in the representation of speech prosody", *Speech Communication*, 46: 334–347, 2005.
- [3] Hirst, D. J., "The Analysis by Synthesis of Speech Melody: From Data to Models", *Journal of Speech Science*, 1(1): 55-83, 2011.
- [4] Ladd, D. R., *Intonational Phonology*, Cambridge University Press. Second edition, 2008.
- [5] Grice, M., "Intonation", in Brown, K. [Ed], *Encyclopedia of Language and Linguistics*, 2nd Edition, Elsevier, vol. 5, 778-788, 2006.
- [6] Beckman, M.E., Hirschman, J. and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in Jun, S-A. [Ed.], *Prosodic Typology*, 9-54, Oxford University Press, 2005.
- [7] Hart, J. 't, Collier, R. and Cohen, A., *A perceptual study of intonation*, Cambridge University Press, 1990.
- [8] Dilley, L., Breen, M., Gibson, E., Bolivar, M., and Kraemer, J., "A comparison of inter-coder reliability for two systems of prosodic transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)", *Proc. of the Int. Conf. on Spoken Language Processing*, Pittsburgh, PA., 2006.
- [9] Hart, J. 't, "Differential sensitivity to pitch distance, particularly in speech", *J. of the Acoust. Soc. of Am.* 69 (3): 811-821, 1981.
- [10] Mertens, P., "From pitch stylization to automatic tonal annotation of speech corpora", in Lacheret, A., Kahane, S., and Pietrandrea, P. [Ed], *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, Benjamins, forthcoming.
- [11] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program]. Version 5.3.10, retrieved 12 March 2012 from <http://www.praat.org/>
- [12] Rossi, M., "Interactions of intensity glides and frequency glissandos", *Language and Speech*, 21: 384-396, 1978.
- [13] House, D., *Tonal Perception in Speech*, Lund University Press, 1990.
- [14] House, D., "Differential perception of tonal contours through the syllable", *Proc. of Int. Conf. of Spoken Language Processing*, 2048–2051. (Oct. 3-6, 1996. Philadelphia, PA, USA), 1996.
- [15] House, D., "The influence of silence on perceiving the preceding tonal contour", *Proc. Int. Congr. Phonetic Sciences* 13, vol. 1: 122-125, 1995.
- [16] Alessandro, C. d' and Mertens, P., "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech and Language*, 9(3): 257-288, 1995.
- [17] Mertens, P., Beaugendre, F. and Alessandro, Ch. d', "Comparing approaches to pitch contour stylization for speech synthesis", in Santen, J.P.H. van, Sproat, R. W., Olive, J. P., and Hirschberg, J. [Ed], *Progress in Speech Synthesis*, 347-363, Springer Verlag, 1997.
- [18] Mertens, P., "The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model", in Bel, B. & Marlien, I. [Ed], *Proceedings of Speech Prosody 2004*, Nara (Japan), 23-26 March 2004.
- [19] Mertens, P., "Un outil pour la transcription de la prosodie dans les corpus oraux", *Traitement Automatique des langues*, 45 (2): 109-130, 2004.
- [20] Hart, J. 't, "Psychoacoustic backgrounds of pitch contour stylisation", *IPO Annual Progress Report* 11: 11-19, 1976.
- [21] Tamburini, F. and Caini, C., "An automatic system for detecting prosodic prominence in American English", *International Journal of Speech Technology* 8(1): 33-44, 2005.
- [22] Jeon, J. H. and Liu, Y., "Automatic prosodic event detection using a novel labeling and selection method in co-training", *Speech Communication*, 54: 445-458, 2012.
- [23] OpenProDat - Italian (Brigitte Bigi, Daniel Hirst). Primary data (corpus). [Laboratoire parole et langage - UMR 7309 \(LPL, Aix-en-Provence FR\)](#). Created 2013-03-06. Speech & Language Data Repository. Identifier [hdl:11041/sldr000810](https://hdl.handle.net/11041/sldr000810)
- [24] Mertens, P., "Transcription of tonal aspects in speech and a system for automatic tonal annotation", *Advancing Prosodic*

Transcription Workshop at Laboratory Phonology 2012, Stuttgart, July 29, 2012.

1An integrated tool for (macro)syntax-intonation correlation analysis

Philippe Martin

UMR 7110, LLF, UFRL, Université Paris Diderot, ODG, rue Albert Einstein, 75013 Paris, France

philippe.martin@linguist.univ-paris-diderot.fr

Abstract

Many efforts are presently made to elaborate large corpora of spontaneous speech (PFC, CFPP2000, C-PROM, ESLO, PFC, Rhapsodie, ORFEO to quote a few), in order to offer the research community large databases that could be used in many aspects of linguistic research. I introduce here new functions of the WinPitch software addressing two aspects of oral corpus analysis:

1) Data mining functions involving specific speech unit (such as conjunctions, weak verbs, etc.) to retrieve rapidly and efficiently their occurrences and their context, displaying automatically the corresponding speech segments together with their acoustical analysis (F0 curve, spectrogram, etc.)

2) Tools enabling the correction of pitch curves resulting from adverse recording conditions, in order to obtain reliable F0 data for further processing (statistical analysis, automatic annotation of sentence intonation, etc.).

Index Terms: spontaneous speech, fundamental frequency, intonation transcription, concordancer.

1. Introduction

A lot of interest is presently devoted to the linguistic analysis of non-prepared speech, and in particular to the prosodic correlates of syntactic and macrosyntactic units. In this type of research, it is assumed that prosodic events help the listener to dynamically reconstruct the prosodic structure intended by the speaker, and eventually allow to infer the syntactic organization of the sentence with which the prosodic structure may be congruent or not.

To investigate this process, it seems at first that patient and meticulous examination of data would be required. Say for example that we want to know about the prosodic correlates of the occurrences of the conjunction “*parce que*” (because) in a set of spontaneous recordings of French.

Instead of listening to hours of recordings to retrieve pronounced occurrences of the key word, we would attempt to retrieve “*parce que*” in all the available text transcriptions of the recordings, and find in a second stage the corresponding speech segments in order to analyze their prosodic properties. This task would be further facilitated if the transcription is aligned, i.e. if bidirectional pointers between text segments and corresponding speech segments have been implemented. This would enable the easy retrieval of every occurrence of the appropriate speech segment from a text selection.

Still, most of the tools available today stop at this stage, even if concordancer of transcribed text items are readily available, listing all occurrences of the search item with together with its left and right contexts.

The new function implemented in WinPitch goes a little bit further by providing the following functions to allow the

user to efficiently and rapidly examine a large number of data with a minimum of manipulations:

1. Generation of a text transcription from alignment files in various largely used formats (Praat textgrid, Transcriber trs, C-Oral Rom xml, Necte xml, CRF alg, etc.);
2. Concordancer: generation of a list of occurrences of the search word, with its left and right contexts. This list is automatically created in Excel format;
3. Automatic retrieval of the search word occurrence in a context selected by the user on the Excel table generated in step 2, with a single mouse click.
4. Extraction of the corresponding speech segment from the proper sound file, played back with all relevant acoustical data displayed (spectrogram, fundamental frequency F0, intensity and duration curves).

2. Integrated concordancer

Figures 1 to 4 illustrate the details of the operations involved. In Fig. 1 The user enters the key word “*parce que*”, selects an appropriate alignment format (Transcriber trs in this example), and clicks on any of the file names stored in a common directory. This directory should contain all the alignment files of interest, together with their corresponding sound files. In the case of Praat textgrid files, the corresponding sound files must have the same name as their textgrid counterpart, as Praat textgrid files do not contain any reference to their corresponding speech file.

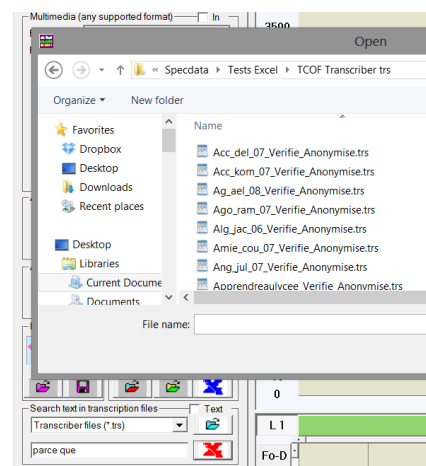


Figure 1. Entering the key word “*parce que*” and selecting a Transcriber files in a directory containing all files of interest.

An Excel table listing all found occurrences of the key word is immediately generated (Fig. 2). This operation is very fast, in the example of *parce que*, the completion takes less than one second to scan 104 files giving 1194 occurrences.

Figure 2. Fine Table generated automatically listing the occurrences of the entered keyword (“parce que”). The whole process takes less than 1 second for a list of 104 files and 1194 occurrences found.

3. Instant data access

When the user clicks on any line of the excel table, a specific occurrence of the keyword is selected together with its left and right contexts, with span values of about 256 characters (rounded to the next word limit). The corresponding text and speech segments are then automatically retrieved and displayed, as shown in Fig 3 and Fig. 4.

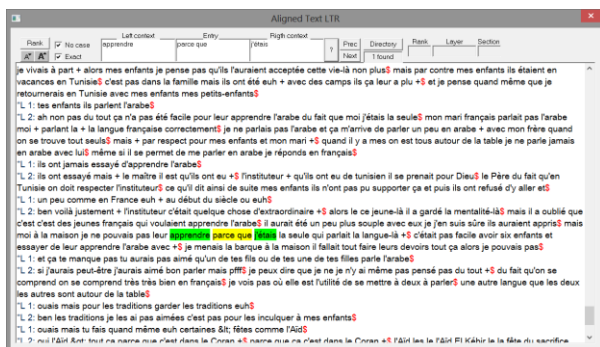


Figure 3. Automatic generation of text from alignment files and selection of the entered key word (“parce que”), highlighted with its immediate context.

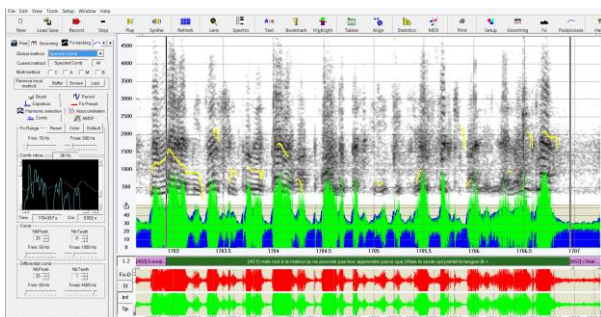


Figure 4. Resulting display of the spectrogram, intensity and pitch curves corresponding to the segment automatically retrieved from the Excel table.

Integrating this function in one single software package makes possible specific research topics on prosody that would have been seen as too time consuming previously.

4. F0 foes

Whereas the integrated concordancer described above can be a valuable time saver, the actual acoustic analysis of prosodic events can prove disappointing when researchers are confronted to obviously erroneous fundamental frequency curves, while this parameter is one of the most important in prosodic analysis.

Indeed spontaneous speech recordings are often performed in noisy conditions, in places where echo and other not so obvious sources of problems are present (an example is given in Fig. 4). More specifically, the measurement of fundamental frequency is particularly sensitive to recorded speech signal distortions due to:

- 1) Poor signal to noise ratio;
- 2) Filtering of low frequencies eliminating low harmonics for male voices;
- 3) Harmonic blur due to room echo in the recording places;
- 4) Encoding in formats such as mp3 or wma with excessive compression levels;
- 5) Presence of external sound sources (car engine, overlapping speech segments, etc.);
- 6) Presence of creaky segments where the fundamental frequency is not really defined.

The speech analysis software Praat [7] for instance, de facto standard in this domain, revealed itself unsatisfactory for F0 tracking for a large number of recordings of the Rhapsodie project [10]. This leads first to evaluate the most frequent causes of F0 errors, then to elaborate various solutions in order to obtain reliable pitch curves. Among causes identified as sources of reliable speech pitch curves, we have:

1. Use of microphones with a poor response in low frequencies, resulting in the absence of the first harmonics in the spectrum (especially for male voices);
2. The presence of an important echo in the signal linked to the recording room dimensions, producing harmonic blurs. An unvoiced consonant can for example appear voiced due to the falsely observed continuity on the first harmonic;
3. A recording level too low, often due to an excessive distance between the microphone and the speaker, resulting in a low signal to noise ratio;
4. Use of AVC (automatic volume control) in the recording process, which distorts the speech intensity curve and indirectly producing errors the evaluation of vowel spectra;
5. Presence of multiple sound sources, in particular generated by low frequency engines (presence of a fridge in the recording room, etc.), or speech overlapping;
6. Excessive compression of the speech signal (e.g. wma or mp3 with a high compression parameter), giving when converted into waveform shifted spectral peak frequency

values undesirable for spectrum based algorithms (Cepstrum, Spectral Comb,...);

To address these potential problems en to ensure the generation of reliable F0 data, WinPitch has a catalog of methods applicable independently on user-selected speech segments:

Frequency domain methods

1. Spectral comb [4], obtained by correlation of the signal spectrum with a spectral comb with variable teeth intervals. Harmonics frequency range retained in the computation are user selectable;

2. Spectral brush [5], obtained by aligning signal harmonics on a selectable time window followed by a spectral comb analysis;

3. Cepstrum [9], evaluation of the periodicity of the log spectrum;

4. Swipep, developed by IRCAM, derived from the Swipe algorithm [2] based on harmonic detection followed by a Viterbi smoothing process;

5. Harmonic selection followed by spectral comb, with the retained harmonics selected by the user from a visual inspection on a simultaneously displayed narrow band spectrogram;

Time domain methods

6. Autocorrelation, operating directly on the speech waveform, available in three flavors, standard, normed Praat [1] and Yin [3], with adjustable window duration;

7. AMDF: average magnitude difference function, with the window length and the clipping percentage user adjustable;

8. Period analysis: F0 values are obtained from period's measurements from pitch markers placed automatically in a first pass and later manually corrected by the user;

These various methods give globally comparable results on good quality recordings. However, for lower quality recordings, the main problems can occur.

By nature, spectral based methods (such as the Spectral Comb) evaluate the signal fundamental frequency from the harmonic structure (i.e. the harmonic spectral lines of voiced segments), obtained from a Fourier transform. This requires an analysis signal time window relatively long (in the order of 32 ms or 64 ms for male voices with F0 equals to 100 Hz), which in turn prevents a correct tracking for fast rising or falling F0 values. The autocorrelation-based methods such as Yin may also exhibit this limitation even if they are based on the time domain (The reason stems from the time window usually selected for the autocorrelation). Other problems may arise when the fundamental frequency is very weak or absent (due to some filtering in the recording process for example);

The presence of pseudo-harmonics due to the presence of echo in the recording room can adversely affect frequency-

based methods. The evaluation of the signal fundamental frequency of the Comb method for example is based on the detection of at least two consecutive harmonics. Echo produced by some harmonics, depending on the room dimensions, can generate trails of some harmonics long enough to make an unvoiced segment appear voiced (see Fig. 2) and confuse the algorithm detecting these components. It is quite difficult to differentiate automatically this spectral configuration from examples where a low frequency filtering would provoke spectral patterns similar to the ones generated by echo.

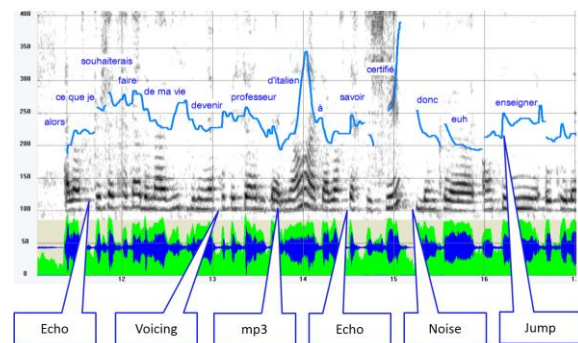


Figure 5. Most common sources of errors for F0 tracking (Rhap-D0003, PFC)

5. Cleaning F0 curves

To apply one of these methods, the user first selects a F0 tracking method in the command window (Fig. 6). Then a time window is selected on screen with the mouse guided by visual inspection of an underlying narrow band spectrogram. By releasing the mouse left button, the corresponding segment of the signal is automatically reanalyzed with the selected method, replacing F0 data with the new obtained values.

The new F0 curve segment is displayed in a color specific to the tracking method chosen, so that the user can identify visually on the overall F0 curve the tracking method pertaining to a specific time segment. Furthermore, by moving the cursor on screen, the corresponding command box corresponding to the F0 tracking method used for the wave segment defined by the cursor is displayed dynamically in the command box, together with all parameters values used for the chosen tracking method (Fig. 6).



Figure 6. Set of command boxes, for user selection of an alternate pitch-tracking algorithm applied locally on a speech segment.

A file containing all the information about corrections made can be saved in text format, as well as a .pitch file describing the corrected pitch curve to be exported to Praat.

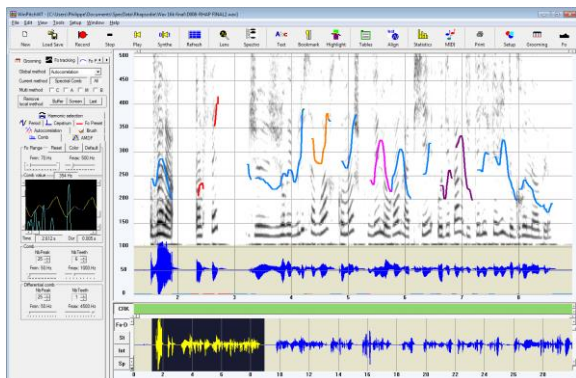


Figure 7. F0 curve sections are displayed in different colors according to the F0 tracking method used. The corresponding command box selected automatically on the left side (Rhap-D1001)

The two functions described above address the main concerns of researchers in the field of prosodic events analysis in their relationship with other structures of the sentence, syntactic and informational. The concordancer allows investigating a large number of occurrences of selected syntactic categories items, whereas the fundamental frequency “cleaning” gives reliable data in most cases retrieved by the concordancer, even in adverse recording conditions.

6. WinPitch as shareware

The software program is presently a shareware, whose installation code is free for the asking. WinPitch is downloadable from www.winpitch.com.

References

- [1] Boersma, Paul (1993) Accurate short time analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound, Proc. Institute of Phonetic Sciences, 17. Univ. Amsterdam, 97-110.
- [2] Camacho, Arturo (2007) Swipe: a sawtooth waveform inspired pitch estimator for speech and music, PhD thesis, University of Florida, 116 p.
- [3] de Cheveigné, Alain and Hideki Kawahara (2002) Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111(4).
- [4] Martin, Ph. (1981) Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne, Proc. 12e Journées d'Etude sur la Parole, SFA, Montréal, 1981.
- [5] Martin, Ph. (2008) Crosscorrelation of adjacent spectra enhances fundamental frequency estimation Proc. Interspeech, Brisbane, 22 – 26 September 2008.
- [6] Martin, Ph. (2012) Automatic detection of voice creak, Proc. Speech Prosody, Shanghai, September 26-28.
- [7] Praat, www.praat.org.
- [8] Transcriber, a tool for segmenting, labeling and transcribing speech, <http://trans.sourceforge.net/en/presentation.php>
- [9] Noll, A. Michael (1967) Cepstrum Pitch Determination, Journal of the Acoustical Society of America, Vol. 41, No. 2, (February 1967), 293-309.
- [10] Rhapsodie (2010) Corpus prosodique de référence en français parlé, <http://rhapsodie.risc.cnrs.fr/en/archives.html>
- [11] WinPitch, www.winpitch.com

Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features

Katarzyna Klessa, Maciej Karpiński, Agnieszka Wagner

Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland

{klessa, maciej.k, wagner}@amu.edu.pl

Abstract

This paper describes the design, development and preliminary verification of a new tool created for the purpose of annotation of spoken language recordings. The software extends the potential of a typical multi-layer annotation system with a new component based on the graphical representation of feature space that supports annotation of continuous and non-categorical features. Apart from the annotation options, the program provides a flexible perception experiment framework aimed especially at testing hypotheses related to continuous and non-categorical features.

The tool was initially tested and first applied for the annotation of a speech corpus composed of conversational and emotionally marked speech data within a larger project confessed to speaker characterisation and recognition.

Index Terms: annotation tools, perception based annotation, paralinguistic features, speech prosody

1. Introduction

We understand the process of speech annotation as assigning tags to selected portions of speech signal that may correspond to various units of analysis, from tiny phonetic segments to complex phrases or paratones. The tags usually come from an explicit, closed set like PoS. Still, there are situations that require more flexibility and where operation on fuzzy categories or gradable features is necessary. For example, in the annotation of emotional aspects of speech, there can be some intermediate affective states between extremes, e.g., between joy and sadness. Their number can be arbitrarily assumed or left for annotators to decide. Another problem pertains the fact that some of labels are two- or multidimensional, i.e., their values can be well represented in a multi-dimensional space. Therefore, they may consist of two or more values (tags) that can be placed on separate annotation layers. This somehow corresponds to the conceptual difference between tags and labels proposed by [1].

Many speech annotation programs (see, e.g., [2]) offer great potential but it is sometimes accompanied by less obvious user interface and complex operation. Some of them were primarily conceived for instrumental phonetic analysis and few of them offer direct support for non-categorical or complex-label annotation. Working on the annotation of spontaneous speech corpora on both linguistic and paralinguistic levels, the authors felt an increasing need for a more intuitive software that would support annotation on

multiple levels as well as various types of categorical and non-categorical data.

2. Software design and development

2.1. Assumptions and requirements

Paralinguistic features as well as other non-categorical features pose a challenge in the process of speech data annotation – both for software and for human annotators themselves. The way of defining the space for their annotations may strongly influence eventual results. The type of the scale used for a given dimension (linear, logarithmic, etc.) may be also of importance. Paralinguistic features often remain difficult to define in an unambiguous way, in clear and accessible terms. If “verbal” tags are used (e.g., the names of emotional categories, like “disturbed”, “angry”), their understanding by annotators may be strongly influenced by everyday usage of such words. Many of these and similar issues may be only partially solved or alleviated, and solutions will be most often context-dependent, designed or tuned for a particular kind of data and specific scientific aims.

The present program is intended for speech annotation for a range of applications, including those strictly technological (e.g., naturally-sounding speech synthesis, automatic speaker and speech recognition) as well as those focused on the psychology of interpersonal communication. Accordingly, the following functionalities and options were considered essential:

- simple and user-friendly interface, easy installation and configuration;
- multi-layer, synchronised annotation with precise boundary placement;
- various, adjustable annotation spaces and scales available as uploadable images;
- the option of using own spaces and scales represented as images;
- use of complex tags (e.g., for two-dimensional features);
- a slot for plugins that would extend the functionality of the program.

Well-organised software that supports annotation of paralinguistic features may also serve as an experimental tool in perception-based studies. While “top-down” approach, starting annotation with pre-defined categories, may keep annotators and researchers “blind” to new, undiscovered phenomena, leaving more flexibility to annotators and offering them non-categorical or continuous space may bring new

observations to the daylight or just allow for new categorisations to emerge.

2.2. Implementation and architecture

Annotation Pro was created using C# programming language and the Visual Studio programming environment and (in its current form) is designed for Windows operating system.

The main construction assumption was to create annotation software of general use, applicable for various types of projects involving both annotation of spoken and written (eg. morphological glossing) resources. On the other hand, the architecture was expected to be extensible and flexible in order to enable annotation according to user-specific needs which has been achieved thanks to plugin technology that enables the users to add their own functionality to the program top menu.

The structure of the system has been developed as a multilayer architecture, in which each application tier represents a specific functional layer. The layers of the program are shown in Figure 1.

n-tier architecture
Presentation
Logic
Database
Shared
Plugin

Table 1. *Programming tiers in Annotation Pro*

The **Database** layer is responsible for the process of writing and reading data on the most basic level. Currently, it concerns writing and reading of XML files and dealing with the software's annotation file format ANT which is in fact a ZIP archive containing the packed XML annotation file (see also 3.3 below). Such solution makes it possible to include various types of content inside the ANT file in future.

The **Logic** layer is an intermediate layer representing data in the form of C# objects that can be used by the programmer for operations on objects and collections of object in the application: Layer, Segment, Configuration.

The highest-level layer is the **Presentation** layer. This layer includes controls representing the elements of the software interface: Spectrogram, Waveform, Layer Collection, Input Device. All these components can co-operate automatically thus making it possible for the programmer to create any clone of the application based on *Annotation Pro*'s functionality. The controls of the Presentation layers are treated as components joined by a special control – Synchronizer. The Synchronizer is a special object which controls the state of variables whose synchronization is necessary for the consistent functioning of the application.

Plugin – a layer responsible for plugins. The plugin functionality makes it possible for the user to adapt the software to the individual needs of their own project. Including the plugin technology is a natural consequence of the main construction assumption: only general options that are required for most uses are built-in as fixed parts of the software while every functionality that is more project- or user-specific may be accessible via plugin menu. Any user familiar with C# programming language can easily create a plugin thus extending the software's functionality (e.g. speech

analysis options, automatic feature extraction from the speech signal or time-alignment procedures or any other desired by the user). While initialising, the program scans the *Plugins* folder located in the user's Documents/*Annotation Pro* folder and updates a list of plugins in the *Plugin* menu based on the contents of the folder. The plugins require a standard C# format (*.cs) with an appropriate structure, shown in an example plugin file, also available for the user in the *Annotation Pro* Plugins folder. The plugin file (*.cs) is compiled and launched at runtime. The user can access and use all controls of the interface, to annotation layers, and data.

Shared – a library including support classes for various layers.

3. User interface

3.1. Annotation interface

Apart from the “traditional” multi-layer annotation interface (accompanied by both spectrogram and waveform signal display), a universal graphic control was implemented in the program which enables using various graphical spaces as a basis for annotation (e.g. Fig. 2).

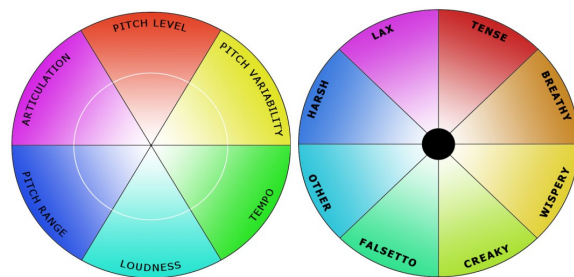


Figure 2: a) and b). Graphical representations used in the description of prosody (a, left) and voice quality (b, right).

Figure 3 (next page) shows the default program interface. The graphic control is visible in the right top corner of the program window. Instead of this particular picture representing two-dimensional space for annotation of emotional states, the user may select another picture (e.g. a *min-max* slider or a set of sliders for perception-based ratings using a continuous scale, etc.). It is also possible to create one's own picture representing any desired two-dimensional feature space. The space to which the graphic control picture is related is interpreted by the software as the Cartesian coordinate system. When the user clicks on the picture, the coordinates of the clicked points are stored and displayed both as dots in the picture and as numbers in the related typical annotation layer. While the user clicks on the picture while the sound is being played, the subsequent clicks result in the automatic insertion of segments in the annotation layer and the corresponding coordinates as annotation labels. The number of segments and their distribution over the layer's timeline is directly connected with the selections made by clicking the points in the graphic representation control. As a result, a collection of coordinates is obtained for which it is then possible to conduct a range of analyses, e.g. cluster analysis (compare also [3] for emotion analysis, and [4] for another examples of another graphic representations used in *Annotation Pro* for both corpus annotation and for conducting perception tests).

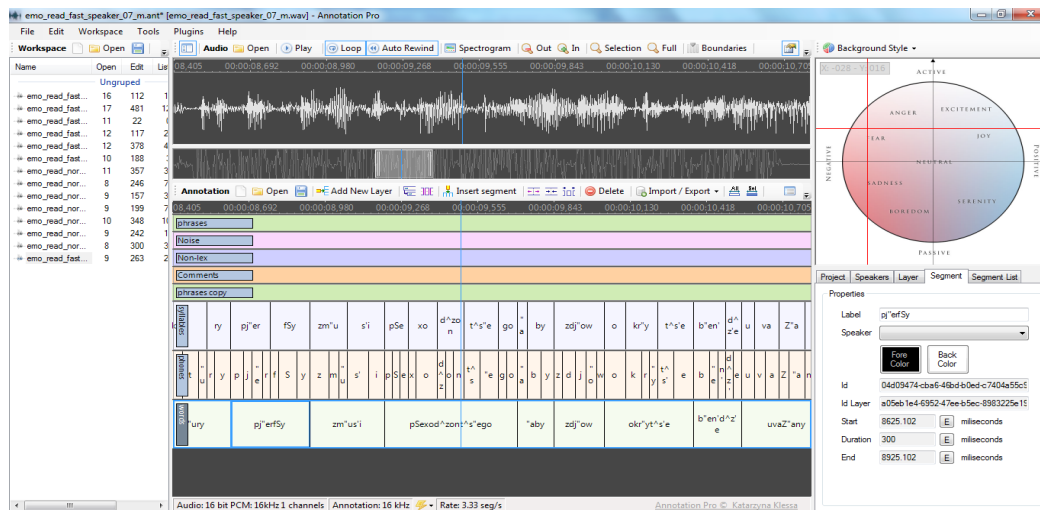


Figure 3: Annotation Pro user interface

3.1.1. Annotation of prosodic and paralinguistic features using graphical representation - example

The functionality of the software which enables annotation based on a graphic representation of the feature space was used in the preliminary annotation of perceived prosody and voice quality of emotion portrayals from *Paralingua* database [4] and in the perceptual recognition of speaker state in terms of emotion categories and dimensions [5]. Prosody was annotated in terms of perceived pitch level, pitch variability, tempo, loudness, pitch range and articulation.

The task of the labeler consisted in positioning the cursor in the regions of the circle corresponding to selected prosodic feature and specific intensity of the feature (Fig. 2 a). In the annotation of perceived voice quality the following labels were taken into account (based on [6]): *lax, tense, breathy, whispery, creaky, falsetto, harsh, modal* and *other*. These voice qualities were represented in a circle divided into eight parts with modal voice located in the center and intensity of a given voice quality (to distinguish different levels of e.g. creakiness or breathiness) increasing towards the edge of the circle (Fig. 2 b).

In order to investigate emotional speech production and perception two more graphical representations were created illustrating emotion dimensions of valence and activation (Fig. 4 a), and emotion labels (categories) describing speaker state (Fig. 4 b): *irritated/angry, interested/involved, proud/satisfied, joyful/happy, experiencing sensual pleasure, bored/weary, ashamed/embarrassed, in despair/sad, anxious/fearful* and *other*. The categorial and dimensional descriptions were based on [7, 8, 9, 10]. In the categorial representation, the twelve emotion labels used in the actor portrayals were collapsed to nine categories (plus *other*), because it was assumed that emotions belonging to the same family, of the same quality and valence, but of a different intensity should be represented together. In the perceptual annotation using the graphical representation (depicted in the Figure 4 b) these differences could be represented by the distance from the center of the circle which corresponded to greater or lesser intensity of the perceived emotion (i.e. intensity decreased from the center to the edge of the circle).

Perception-based annotation of prosody, voice quality and emotional state of the speaker consisted in placing the cursor in the appropriate area of the graphical representation. The resulting coordinates were automatically displayed on the associated annotation layer, saved in a text file and then exported to a spreadsheet.

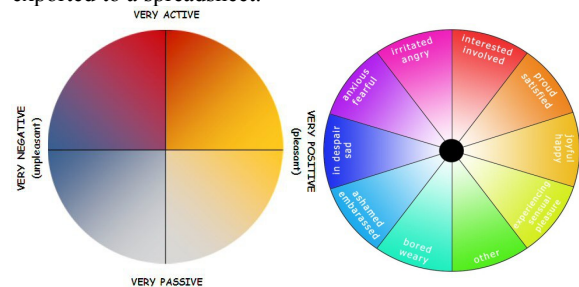


Figure 4 a) and b). Graphical representations used in the classification of emotional speech using valence/activation dimensions (a, left) and emotion labels (b, right).

3.2. Perception test session mode

The annotation interface can also serve as a tool for perception tests as it offers additional options in the *test session* mode. In this mode, the user can use the options for setting-up an experiment. First, it is possible to define options related to participants data management (participant's name or ID, age, gender, region of origin or other features). The perception test set-up is flexible and can be adjusted to particular needs. The experimenter can decide on the number of possible replays of each signal, the order of the signals, the possibility of returning to previous signals after marking the first answer/decision. The original file names can be either displayed or hidden during the test session. The results of the test are written to a CSV file where information about all the actions taken by the subject during the testing session (answers/decisions, opened files, number of listenings, etc.).

3.3. Annotation file format

Annotation Pro annotation (ANT) files are based on the XML format. The format was designed in a way to generalize the

annotation information. Any information narrowing the annotation information to a specific domain or project are introduced by the user via the user interface.

ANT files can store data for annotation using any desired number of annotation layers. The two crucial components of the XML file are **<Layer>** and **<Segment>**. The first one includes information about annotation layers and the second is a universal element that may contain various types of annotation labels (orthographic transcription, tagging of prosody, syntax, discourse markers, paralinguistic annotations, etc.) depending on the user's needs.

Any other relevant information related to the file, speaker, corpus etc. is stored using optional **<Configuration>** elements. This element is of dictionary type, and includes keys and values. The keys should be unique. A number of keys have been reserved for a set of standard properties related to e.g. date of creation, date of modification, version name, title of the project, characteristics of the recording environment, description of background noises characteristic to the project, the name of the collection including the project, the type of corpus, licence, etc.

Annotation Pro also reads XML files created with external tools provided that the format is compatible with ANT. Any information that has not been pre-defined in *Annotation Pro* should be included in the XML file using the **<Configuration>** elements. *Annotation Pro* will open such files, ignore the "foreign" information, but it will not be lost. Thanks to this solution, it is possible to make use of *Annotation Pro* on an intermediate, lossless basis.

Apart from the use of the default XML-based annotation files, *Annotation Pro* can import files from the following external formats: *Transcriber*'s TRS [11], and BLF [13], and also from TXT (each verse of the source text file will be imported to a separate segment in the selected annotation layer) and CSV (configurable import, including *Wavesurfer*'s LAB [12]) files.

4. Conclusions

Annotation Pro has already been employed for transcription and annotation of speech data in several research projects. Its applications included the analysis of perception and production of emotional speech. Presently, the technique of emotional speech analysis based on the functionalities provided by *Annotation Pro* is used in a larger-scale study on cross-linguistic perception of vocal communication of emotions. The tool is also used for transcription and annotation of corpora of lesser used languages requiring annotation with non-standard types of font family and morphological glossing. Annotators from both fields confirm that the software's principle is clear and user interface is easy to master while still retaining much flexibility. Although it is clear that all the issues mentioned in 1 and 2.1 cannot be solved by this piece of software, it offers some advantages over other available solutions.

Annotation Pro will be further tested and extended with new plugins. Import/export options will be elaborated in order to accept the data and metadata from other programs (e.g. *Praat* [14], *SPPAS* [15], *TGA* [16]). This is expected to allow to use *Annotation Pro* as a complementary tool for specific purposes and to exchange and integrate data with no information loss. In order to provide higher level of

interoperability of the software, it is currently considered to develop a new edition of *Annotation Pro* using *Mono* software platform, thus enabling the use of the program under operating systems other than MS Windows. As for the interface, among options under consideration, there is also video annotation and recording of other types of multiple-speaker data on separate layers.

Annotation Pro is freely available for research purposes from: annotationpro.org (contact e-mail: klessa@amu.edu.pl).

5. Acknowledgment

Annotation Pro is developed based on the experiences of the authors gained during the work on an earlier tool named *Annotation System* implemented within project no. **O R00 0170 12** supported from the financial resources for science in the years 2010–2012 as a development project.

6. References

- [1] Popescu-Belis, A., "Dialogue Acts: One or More Dimensions?" *ISSCO Working Paper*, 62 – 29th November 2005.
- [2] Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J., Traum, D.R., *Evaluation of Transcription and Annotation tools for a Multi-modal, Multi-party dialogue corpus*. Southern California, Inst. For Creative Technologies, 2004.
- [3] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. "FEELTRACE: An instrument for recording perceived emotion in real time". *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.
- [4] Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M. "Using *Paralingua* database for investigation of affective states and paralinguistic features". *Speech and Lang. Technology*. 14/15, to be published.
- [5] Wagner, A. "Emotional speech production and perception: A framework of analysis". *Speech and Lang. Technology*, vol. 14/15, to be published.
- [6] Laver, J. *The phonetic description of voice quality*. Cambridge Studies in Linguistics London, 31, 1-186, 1980.
- [7] Russell, J. A. "A circumplex model of affect". *Journal of personality and social psychology*, 39(6), 1161, 1980.
- [8] Banse, R., & Scherer, K. R. "Acoustic profiles in vocal emotion expression". *Journal of personality and social psychology*, 70(3), 614, 1996.
- [9] Laukka, P. *Vocal expression of emotion: discrete-emotions and dimensional accounts* Phd dissertation, Uppsala Univ., 2004.
- [10] Bänziger, T., Pirker, H., and Scherer, K. "GEMEP-Geneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions". *Proceedings of LREC*, Vol. 6, pp. 15-019, May, 2006.
- [11] Barras, C., Geoffrois, E., Wu, Z., Liberman, M. "Transcriber: Development and use of a tool for assisting speech corpora production". *Speech Communication*, 33(1–2), 5–22, 2001.
- [12] Sjölander, K., and Beskow, J. "WaveSurfer – an Open Source Speech Tool". *Proceedings of 6th ICSLP Conference 2000*, Vol. 4 (pp. 464–467). Beijing, 2000.
- [13] Breuer, S., & Hess, W. "The Bonn open synthesis system 3". *International Journal of Speech Technology*, 13(2), 75–84, 2010.
- [14] Paul Boersma & David Weenink (2009). "Praat: doing phonetics by computer (Version 5.1.05)" [Computer program]. Available: <http://www.praat.org/>
- [15] Bigi, B. "SPPAS: a tool for the phonetic segmentation of speech". *Language Resource and Evaluation Conference*, Istanbul, Turkey, 2012.
- [16] Gibbon, D. "TGA: a web tool for Time Group Analysis". *Proc. of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, August 2013, to be published.

ProsoReportDialog: a tool for temporal variables description in dialogues

Jean-Philippe Goldman

Linguistic Department, University of Geneva, Switzerland

jeanphilippegoldman@gmail.com

Abstract

Temporal variables were initiated by Grosjean in 1972 [1] who defined in details several dimensions in timing and rhythm in order to measure and compare these characteristics in different languages or various speaking context. Aside this approach dedicated to monologues, some studies applied notions from conversational domain (as in Sachs et al. 1974 [2]) to dialogues corpus in an automatic way.

Our contribution extends this work in both directions. (i) We suggest gathering these two approaches in an automatic tool for temporal variables description in dialogues. (ii) We compare these variables in a corpus of 35 dialogues to show their differences according their situational features.

Index Terms: tool, prosody, dialog, tools, resources, analysis, and speech prosody

1. Introduction

Temporal variables in speech have been studied extensively, in the domain of speech synthesis to improve the naturalness of the synthetic speech [3], in man-machine dialogue modelling [4], or in the field of conversation analysis, using a qualitative [5] or quantitative [6] approach within corpus linguistics. Finally, a number of studies have focused on temporal aspects in a descriptive approach [7][8][9], or in a contrastive approach [10] or even to study phenomena as hesitations [11].

These numerous studies deal with a wide range of languages and speaking situations, providing measures that are not often comparable because of the procedures used to extract and analyse them. To mention only few debated points: 1. Should the so-called “micro-pauses” (pauses smaller than a threshold) be considered or not (see [8] for a review of this bias) 2. Should the pause duration be log-transformed or not (see [4]). 3. Under which threshold of time difference should two turn boundaries (start or end point) be considered as simultaneous ?

After recalling some basics of temporal variables, we present an automatic tool to measure the temporal variables and apply it to a corpus of 35 various dialogs. We limit our current work to dialog (i.e. with exactly 2 speakers) leaving a third, a forth speaker for a later study.

2. Temporal variables

2.1. In monologues

More than forty years after its publication, the pioneering work of [1] remains a reference for defining the temporal variables of the oral language. The total (or speech) time is

composed of the articulation time (or phonation) and pause time, from which are derived the articulation and pause ratios (as a percentage of speech time) as well as the articulation rate (in syllables per second). Some other notions are taking into account like the number and the length of the speech sequences separated by pauses.

On the top of this, additional secondary variables are also considered such as filled pauses (hesitation and syllable lengthening), repetitions and false starts. These variables require manual annotation and thus are often ignored in studies on large corpora.

2.2. In dialogues

The analysis of conversations between two or more speakers makes the study of temporal variables more difficult; especially if overlapping speech occur. The notion of speech turn, which seems to be central, is extremely difficult to implement in an automatic analysis system, as this unit is the result of a dynamic analysis of how the speakers combine turn constructional units (TCU's) incrementally to produce what will be considered in context, as a turn of speech [12][13].

For automatic annotation, the notion of verbal production (VP) as a sequence of syllables assigned to a unique speaker should be preferred to speech turn ([14]). Verbal production can be long sequence syllables but in some cases a brief backchannel output, occurring within a pause or overlapping with the other speaker.

The silent pauses may occur within the VPs of a speaker (so-called within- or intra-speaker pause) or between the end of the verbal output of a speaker and the beginning of the next speaker (between- or inter-speaker pause or gaps). The former can simply be called “pauses” if the latter are referred as “gaps”.

The transition from a speaker to another can occur without any gap or overlap (the famous no-gap-no-overlap as in [2]), but often leads to a speech overlap of speech which, at most times, is not perceived as an interruption of the speaker being but as a slightly early transition [15].

The most complete list of turn change patterns is provided by [16], identifying 10 cases. This model has been often simplified (see [4][6]) because its implementation requires manual annotation of some phenomena, such as backchannels.

3. Procedure

To derive the temporal variables of a dialog, the main relevant information can be embedded in a single tier with speaker annotation, showing which speaker is speaking at every moment. Pauses, gaps and overlapping segment are also indicated.

The same information can possibly lie within several tiers (one tier per speaker). In this implementation, multiple tiers are firstly merged into a unique one.

As mentioned before, silent pauses are split into intra-speaker pause and inter-speaker gaps. If a silence is surrounded by two VPs of the same speaker, it is a pause. If a speaker transition occurs, then it is a gap. This pause-gap difference becomes a problem when an overlapping interval is adjacent to a pause. This happens when the two speakers start or stop speaking simultaneously.

[17] makes a systematic distinction, suggesting the *Instigator* and *Owner* status for each pause: "the *instigator* of a silence is the speaker who last spoke before the silence occurred (or who last spoke alone, in cases of a simultaneous end of speech); the *owner* of the silence is the speaker who breaks the silence (or the instigator, in cases of simultaneous start of speech); a *gap* is a silence with a different instigator and owner (aka *inter-speaker silence*); and a pause is a silence with the same instigator and owner (aka *intra-speaker silence*)"

Spk1	1				1		1	
Spk2		2		2				
Spk	1	gap	2	pause	2	overlap	gap	1

Figure 1. Example for separate tiers (above) and merge tier (below)

On the basis of a merged speaker tier where verbal productions, breaks, gaps and overlaps are distinguished, the tool offers the following measures, in order to depict as simply as possible the composition of the dialog:

- Recording Time
- Speech time (excluding side-breaks)
 - Articulation time
 - Exclusive articulation
 - Overlap (initiated, with/without transition)
 - Cumulated articulation
 - Silence time
 - pauses (intra)
 - gaps (inter)

These dimensions are shown with their duration (in seconds) and their count or frequency (number of pauses, number of verbal productions, number of overlap sequences). In addition, complex variables are derived as:

- ratios (as a percentage of the speech time)
- mean durations
- rates (per second or per minute)

For instance, the articulation ratio is identical to the so-called "rapport TA-TL" (so-called "rapport Temps Articulation-Temps de Locution") in [1]. These measures are also detailed for each speaker. It should also be mentioned that the notion of cumulated articulation can be greater than 100% of the speech time. In other words, the exact articulation time

is added for each speaker (i.e. overlap segments count more than once).

Moreover, the speaker initiating an overlapping interval (e.g. starting of VP while the other is currently speaking) can be clearly identified. Besides, this overlap segment leads to a turn change or not, it is counted with or without transition. In the second case (no transition), the overlap could be a backchannel VP or an aborted try of turn taking. The distinction of these latter cannot be done automatically at this moment.

In practice, the tool takes a TextGrid file with a speaker tier as input, and offers four different outputs. In the first two, all these above measures are displayed in a shortened version within the Info window of Praat (which can be saved in a text file). It is a simple overview of some measures. One is articulation-oriented as the other is speaker-oriented as can be seen below:

```
#####Processing TextGrid foot0...
Speech time                299
-articulation              214          (71.6%)
-overlap                   11          (3.6%)
-exclusive artic.         203          (67.9%)
-spk 1                     79          (26.4%)
-spk 2                    124 (41.5%)
-silence                   85          (28.4%)
-gap                       24          (8.0%)
-pause                     61          (20.4%)
-spk 1                     36          (12.0%)
-spk 2                     25          (8.4%)
```

```
#####Processing TextGrid foot0...
Speech time                299
-spk 1                    126          (42.1%)
-exclusive artic.         79          (26.4%)
-overlap                   11          (3.7%)
-pause                     36          (12.0%)
-spk 2                    160          (53.5%)
-exclusive artic.        124          (41.5%)
-overlap                   11          (3.7%)
-pause                     25          (8.4%)
-gap                       24          (8.0%)
```


The third output is an extended version as below:

```
#####Processing TextGrid foot0...
Tier speaker found in TextGrid : 4
Recording time          300.012
Speech time             299.134 (side.pauses
excl)
Silence time           85.042 (28.4%/speechime)
Pause (intra) time     61.402 (20.5% )
Gap (inter) time       23.640 (7.9 %)
Articulation time      214.092 (71.6%)
Cumulated articulation 224.725 (75.1%)
Overlap time           10.633 (3.6%/speech ime)
                        (5.0%/art.time)
Exclusive articulation 203.459 (68%/speech time)
                        (95%/art.time)

Nb of pauses           84 (dur:1.012;rate:16.8)
Nb of intra pauses     50 (dur:1.228;rate:10.0)
Nb of pauses inter     34 (dur:0.695;rate:6.8)
Nb of VP               120 (dur:1.784;rate:24.1)
                        (min:0.05;max:6.238)
Nb of overlaps         25 (dur:0.425;rate:5.0)
Nb initiated overlap   28 (w/o transition 13)
```

```
###Speaker #1###
Articulation time      90.023 (42.0%/art.time)
                        (40.1%/cumul.art)
Exclusive articulation 79.390 (26.5%/speech t.)
                        (37.1% /art.t.)
Pause (intra) time     36.386 (12.2%/speech t.)
Nb of VP               56 (dur:1.608;rate:24.1)
                        (min:0.05;max:5.813)
Nb overlaps            25 (rate: 5.0)
Nb initiated overlap   12 (w/o transition: 6)
Nb of intra pauses     21 (dur:1.733;rate:4.2)
```

```
###Speaker #2###
Articulation time      134.702 (62.9%/art.time)
                        (59.9%/cumul.art)
Exclusive articulation 124.069 (41.5%/speech t.)
                        (58.0%/art.t)
Pause (intra) time     25.016 (8.4% /speech.t)
Nb of VP               64 (dur:2.105;rate:24.1)
                        (min:0.05;max:6.238)
Nb of overlaps         25 (rate: 5.0)
Nb initiated overlap   16 (w/o transition:7)
Nb of intra pauses     29 (dur:0.863;rate:5.8)
###
```

Finally, the 4th option outputs all the measures of the full report in a table (tab-separated or csv) rather in a text window. In this case, many dialogs can be analysed at once and the results are represented in columns for all the speech files. The table format permits further analysis.

The tool has been developed as a Praat plugin and includes with some extra tools to manipulate intervals tiers such as:

- **Merging interval tiers:** if speaker tiers are separated, this tool produce a unique speaker interval tier
- **Report pauses from a syllabic tier to the speakers tier:** if a syllabic tier exists, pauses can be derived and added to the speaker tier.
- Merging similar consecutive intervals: if several pause intervals or several same-speaker intervals exist, they can be merge as one, avoiding future wrong measures.

4. Corpus

In this part, we apply the described methodology to a group of 35 extracts representing various speaking styles in different activities and situations. The total duration is 2 hours and 40 minutes. To cite a few examples, there are radio interviews, radio news dialogs as well as sports live report or map task dialogues.

We annotated each dialog according to a set of situational features [17][18] to allow further study and comparison. Our hypothesis is that these situational features may yield differences in the temporal variables. These features are the degree of interactivity (interactive, semi-interactive, non-interactive), the degree of preparation (spontaneous, semi-prepared, prepared) and the degree of media use as in the next Table.

	Interactive (total = 14)	semi-inter. (total =16)	non-inter. (total = 5)
Spontan. (total= 16)	D0009,D2008 foot0,foot1, foot31,intlib1, intlib2,intlib3	D0003,D0005 D1003,D2004	D0007 D0008 D0017 D0020
Semi- prepared (total= 16)	D0004,D0006 <u>D2001</u> <u>D2002 D2010</u> <u>D2012</u>	D1001,D1002 <u>D2009,infor1</u> infor2,intfor3 <u>intrad1,intrad2</u> <u>intrad3,intrad4</u>	
Prepared (total = 3)		<u>D2005 D2006</u>	D2013

Table 1. Distribution of the 35 speech samples according the 3 situational features. The media feature is indicated as non-media (total = 16), secondary media (total = 15), **media** (total = 4)).

5. Results

In the following table, some mean measures are represented for the 35 dialogs.

Variable	Mean	Std deviation
Speech time	275.0 (s)	110.3
Articulation ratio	79.2 (%)	7.9
Overlap ratio	4.2 (%)	4.7
VP duration	10.4 (s)	8.7
VP duration spk1	16.7 (s)	13.5
Speech ratio spk1	77.7 (%)	14.6
VP duration spk2	3.3 (s)	2.7
Speech ratio spk2	20.2 (%)	14.4
Gap duration	0.7 (s)	0.6
Gap ratio	5.4 (%)	3.3

Table 2. Vital statistics for the corpus of 35 dialogs

As a first result, we plotted the speech ratio of both speakers as in Figure 2. The 35 extracts are scattered along a diagonal as the sum of their speech ratio is supposed to be 100%. The items below the line have more gaps than overlaps.

The “interaction” feature is divided into three categories: non-interactive (red), semi-interactive (green) and interactive (blue). Each of these three categories gather in groups although interactive (blue) and non-interactive (red) ones seem to superimpose leaving aside the semi-interactive.

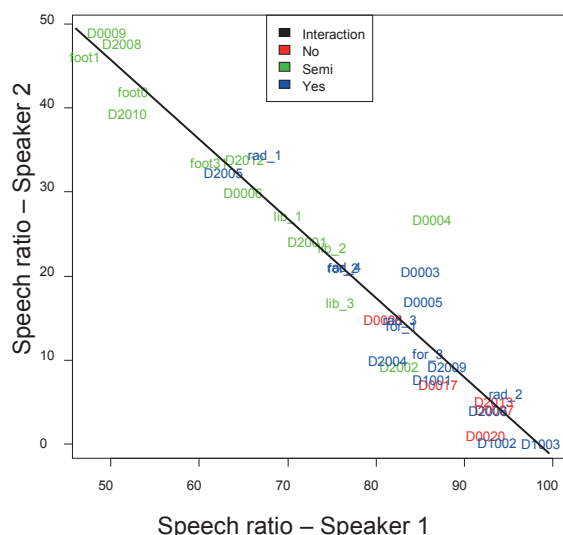


Figure 2. A corpus of 35 dialogues represented as speech ratio of speaker1 vs. speech ratio of speaker2

6. Discussion

This first attempt to automatically extract temporal variables in dialogs showed a high number of unexpected problems. Some questions still need further investigations. However, further developments are already in preparation like speech rate as well as an estimation of dynamic variation of the temporal variables along the total speech recording.

This tool is freely available and is distributed under this website:

<http://latlntic.unige.ch/phonetique>

7. Acknowledgements

This work is part of the Swiss-FNS project “Prosodic and linguistic characterization of speaking styles: semi-automatic approach and applications” (fund n°100012_134818).

8. References

[1] Grosjean, F. & A. Deschamps (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26, pp.129- 156.
 [2] Sacks, H., E. A. Schegloff & G. Jefferson. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4), pp. 696- 735. Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", *IEEE Trans. Speech and Audio Proc.*, 7(6):697-708, 1999.
 [3] Zellner, B., 1998. Caractérisation et prédiction du débit de parole en français Une étude de cas.

[4] Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38, 555-568. doi: 10.1016/j.wocn.2010.08.002
 [5] Auer, P., Couper-Kuhlen, E., & Müller, F. (1999). *Language in time: The rhythm and tempo of spoken interaction*. New York: Oxford University Press.
 [6] Ten Bosch, L., N. Oostdijk & L. Boves (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47:1-2, pp. 80- 86
 [7] Duez, D. (1987). Contribution à l'étude de la structuration temporelle de la parole en français, PhD thesis Université de Provence.
 [8] Campione, E. & Véronis, J. 2002. A large-scale multilingual study of silent pause duration. *SP-2002*, 199-202
 [9] Goldman, J. et al., 2010. Prominence perception and accent detection in French: a corpus-based account. In *Speech Prosody*. 2010, Chicago.
 [10] Grosjean, F. & Deschamps, A., 1975. Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, 31, pp.144-184. Candea 2000
 [11] Selting, M. (2005) Syntax and prosody as methods for the construction and identification of turn-constructual units in conversation. In: Hakulinen, Auli and Margret Selting (eds.), *Syntax and Lexis in Conversation: Studies on the use of linguistic resources in talk-in-interaction*. 2005. (pp. 17-44)
 [12] Mondada, L. (2008). L'interprétation online par les co-participants de la structuration du tour in fieri en TCUs: évidences multimodales. *Tranel* 48, pp.7- 38.
 [13] Groupe ICOR. (2006). Glossaire. Site CORINTE <http://icar.univ-lyon2.fr/projets/corinte/> [consulté le 7/5/2013]
 [14] Jefferson, G. (1983). Notes on some orderliness of overlap onset. *Tilburg Papers in Language and Literature* 28, Department of Linguistics, Tilburg University.
 [15] Weilhammer, K. & Rabold, S. (2003). Durational Aspects in Turn Taking, *ICPhS*, p. 931-934.
 [16] Edlund, J., Heldner, M. & Hirschberg, J., 2009. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*. Citeseer, pp. 2779-2782.
 [17] Koch, P., & Oesterreicher, W. 2001. *Langage parlé et langage écrit*. (G. Holtus, M. Metzeltin, & C. Schmitt, Éd.) *Lexikon der romanistischen Linguistik (LRL)*. Tübingen: Niemeyer.
 [18] Simon, A.C., A. Auchlin, M. Avanzi & J.-Ph. Goldman. (2009) Les phonostyles: une description prosodique des styles de parole en français. In: Abecassis, M. & G. Ledegen, *Les voix des Français. En parlant, en écrivant*, Berne: Peter Lang, 71-88.

Prosodic phrasing evaluation: measures and tools

Klim Peshkov, Laurent Prévot, Roxane Bertrand

Aix-Marseille Université
Laboratoire Parole et Langage
5 avenue Pasteur
Aix-en-Provence, France

klim.peshkov@lpl-aix.fr, laurent.prevot@lpl-aix.fr

Abstract

Over the recent years several transcription systems and tools have been created for marking prosodic phrasing. Although they correspond to different theoretical stances and objectives, it seems important to us to be able to compare the results of the tools and to study the reliability of the coding systems. However, only a few studies [0], [1] have focussed on reliability. We compare several segmentation evaluation metrics as well as intercoder reliability measures. About evaluation metrics, methodologies are coming mostly from clause or word segmentation: (i) precision and recall on boundaries ; (ii) WindowDiff and (iii) segmentation similarity. With regard to intercoder agreement, we discuss the standard measure (κ) and how it is applied to segmentation tasks. The poster consists in a practical application to two cases: (i) an evaluation of prosodic tools and (ii) a reliability evaluation of annotation campaign.

Index Terms: evaluation; intercoder reliability; speech prosody; prosodic phrasing detection

1. Introduction

Prosodic information is useful to answer linguistic questions and to create applications which deal with speech. Annotating prosody of large corpora by human means is costly and rarely possible. Automatic tools have been created to automate detection of prosodic events, but in order to use them, we would like to have a better idea of their performance. In order to evaluate tools in terms of human performance, one has to rely on reference segmentation made by human. However, using annotation made by only one person is risky, because a part of answers might be simple guesses. With multiple annotators, it is possible to create highly reliable “gold standard” [2]. First step in this direction is to obtain interannotator agreement measure.

The evaluations presented below are performed on the Corpus of Interactional Data (CID) [3]. This is a corpus made of 8 conversations of one hour involving two speakers. The protocol for obtaining this data was made in such a way that the interactions are highly natural featuring a lot of overlap and disfluencies.

Section 2 discusses precision/recall metrics and WindowDiff metrics in application to evaluation of prosodic phrasing tools evaluation. Section 3 presents estimation of interannotator agreement for prosodic phrasing annotation using κ statistics.

2. Evaluation of automatic segmentations

A number of tools for automating prosodic analyses have been proposed for French. We can cite Analor [6], Momel-Intsint [7], Prosogram [8]. Among these tools only Analor is directly concerned with prosodic phrasing. We also implemented an algorithm proposed by Simon et Degand (henceforth DS) [9], which is based on phonetic cues such as syllable length and fundamental frequency variation. Our baseline segmentation in Inter-Pausal Units (IPU), which assigns boundaries before and after pauses longer than 200 milliseconds.

In order to get a more precise idea about different tools for prosodic phrasing detection, we want to compare quantitatively the outputs of these tools with reference manual annotation and also to compare different outputs of the tools between them. In this section, we use an annotation of intonation phrases (IP) made by one expert linguist as reference segmentation.

2.1. Precision, recall and f -measure

Precision, recall and f -measure are conventional evaluation metrics from information retrieval. Applied to segmentation task, separate measures for left boundaries, right boundaries and the entire units. This method was used, for example, for the shared task of CoNLL-2001 (Conference on Computational Natural Language Learning) [4]. In our case, we do not work with text, but with aligned transcripts. Hence the alignment is not always perfect. We adopted a delta of 160 ms to tolerate near small mismatches. The value corresponds to the average length of syllables in our corpus.

Table 1 presents results of evaluation of tool's outputs using expert annotation. Low rates of detection, especially in case of the whole units, may be due to the fact, that the tools and the manual segmentation contain prosodic objects of different levels. The DS algorithm shows the best results in the detection of starts, ends and whole units. All the tools are better at detection of starts of the units than their ends.

2.2. WindowDiff

It should be noted that, when used for segmentation evaluation, information retrieval metrics present a serious drawback. They do not take in consideration the distance between the borders of the segmentations being compared. Near-miss errors are penalized as heavily as insertion or deletion of borders and using delta can result in a bias.

WindowDiff metrics was introduced to adress this problem

		spk1			spk2			Mean		
		Prec.	Recall	f	Prec.	Recall	f	Prec.	Recall	f
S	IPU	82.6	39.2	53.2	83.3	43.0	56.7	83.0	41.1	55.0
	DS	77.9	44.5	56.6	78.2	49.5	60.6	78.0	47.0	58.6
	An.	82.1	34.3	48.4	84.9	35.6	50.2	83.5	35.0	49.3
E	IPU	72.1	34.3	46.5	74.9	38.6	51.0	73.5	36.4	48.7
	DS	67.7	38.7	49.2	69.4	44.0	53.8	68.6	41.3	51.5
	An.	76.0	31.8	44.9	81.2	34.1	48.0	78.6	32.9	46.4
U	IPU	30.2	14.4	19.5	37.6	19.4	25.6	33.9	16.9	22.5
	DS	30.7	17.5	22.3	36.8	23.3	28.6	33.7	20.4	25.4
	An.	30.5	12.8	18.0	38.9	16.3	23.0	34.7	14.5	20.5

Table 1: Precision and recall. Evaluation of segmentations in terms of human preformance

[5]. The algorithm operates as follows. It consists in moving a fixed-length window along the two segmentations (cf. Figure 1¹), one unit at a time. On the scheme, the length of the window, which is represented by arrows, is 5 units. For each position, the algorithm compares the numbers of borders in both segmentations. If the number of borders is not equal, the difference of the numbers is added to the evaluated algorithm's penalty. The sum of penalties is then divided by the number of stops, yielding a score between 0 and 1. The score 0 means that the segmentations are identical. The length of the window is set to $1/2$ of the average length of a unit in the reference segmentation.

Initially, WindowDiff was created for text segmentation tasks. When applying it to prosodic units evaluation in time-aligned transcripts, we had to adapt it to our case by introducing a time-based step. If we had chosen to move the window by unit-based step, we would loose time dimension of our data. That's why we introduced a time-based step to move the window. Setting shorter step provides higher resolution of evaluation (but requires more computation time). Results shown here were obtained with a step of 20 milliseconds.

One of the problems of this relatively new metrics is that it is difficult to interpret results in absolute terms. In order to have the first picture of WindowDiff's behaviour, we tested it by perturbing identical segmentations. Figure 2 presents the evolution of WindowDiff score (y-axis) depending on the proportion of randomly moved boundaries (average distance of perturbation is 2.6 seconds and the minimal distance is set to 100 milliseconds). The score evolves in a linear fashion, but quite slowly. When 99% of the boundaries are moved, it reaches only 0.5. The score 0.3 can be interpreted as high divergence between segmentations, because it corresponds to 50% of moved boundaries.

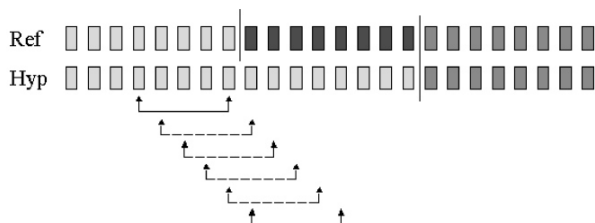


Figure 1: WindowDiff metrics

Table 2 presents WindowDiff metrics of tools' outputs in comparison to manual annotation. All the results indicate high divergence with the manual annotation. As in the case of precision and recall metrics, the DS algorithm's segmentation is

¹reproduced from [5].

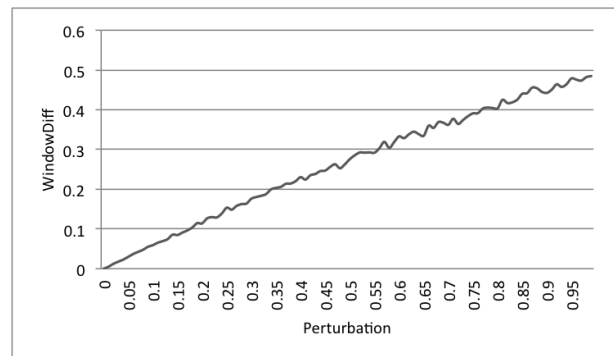


Figure 2: WindowDiff test by boundaries perturbation

	spk1	spk2
IPU	0.275	0.281
DS	0.263	0.265
An.	0.306	0.321

Table 2: WindowDiff. Evaluation against expert annotation

the closest to the reference. Although there is only a slight improvement over the baseline.

The next set of results (Table 3) is a comparison between automatic segmentations. The comparison was made in both directions, because depending on the choice of reference segmentation, the length of the window changes, producing different results. This is why, the results of IPU-Analor and Analor-IPU differ. It follows from the table, that DS algorithm is very close to IPU, and Analor's outputs differ a lot from both.

3. Interannotator agreement

During an annotation campaign of prosodic phrasing by naive annotators, the annotators were asked to assign a number between 0 and 4 to words' right boundaries, corresponding to 4 levels of prosodic break (similar to break indices in the ToBI system). 0 is the default boundary between two words without prosodic marking. Thus, each word's right boundary represents a decision point. All 8 dialogues of the CID corpus were annotated, each speaker was annotated by two judges.

In order to obtain a rough evaluation of the reliability of annotations we used a simple inter-annotator measure, the κ statistics. It is interpreted as "the proportion of joint judgments in which there is agreement, after chance agreement is excluded" [10]. The value of κ ranges between -1 and 1.

Table 4 shows interannotator reliability for several speakers of our corpus. First line takes in consideration all the four levels. The agreement is low, which means that the task was too difficult for the annotators. Second and third lines flatten levels to arrive to higher scores using just 2 classes instead of 4.

	Reference segmentation		
	IPU	An.	DS
IPU	–	0.311	0.077
An.	0.158	–	0.233
DS	0.089	0.390	–

Table 3: WindowDiff. Comparison between tools' segmentations

	spk1	spk2	spk3	spk4	spk5	spk6	Mean
0 1 2 3	0.38	0.28	0.27	0.48	0.16	0.36	0.32
(0 1) (2 3)	0.45	0.41	0.31	0.62	0.17	0.46	0.40
0 (1 2 3)	0.58	0.52	0.56	0.70	0.27	0.66	0.55

Table 4: Interannotator agreement

4. Conclusions and future work

Above, we presented such evaluation metrics as (i) precision and recall and (ii) WindowDiff with examples of their usage in the context of evaluation of tools for prosodic phrasing detection. The interannotator agreement of prosodic boundaries was also discussed with an example of results.

We continue to experiment with Analor tool by tweaking its parameters with the aim to obtain segmentations which would be more similar to the IPs.

In future, we would like to experiment with segmentation similarity metrics. It was proposed by [11] as an improvement of WindowDiff. This metrics relies on edit distance between the boundaries to compute penalties.

Acknowledgements

The author would like to thank Provence-Alpes-Cte d'Azur region which supported this work.

5. References

- [0] Lacheret, A. and Obin, N. and Avanzi, M. "Design and evaluation of shared prosodic annotation for spontaneous French speech: from expert knowledge to non-expert annotation" Proceedings of the Fourth Linguistic Annotation Workshop: 265-274, 2010
- [1] Breen, M. and Dilley, L.C. and Kraemer, J. and Gibson, E. "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)" In press, 2013
- [2] Beigman Klebanov, B. and Beigman, E. "From Annotator Agreement to Noise Models" Computational Linguistics, 35(4):495-503, 2009
- [3] Bertrand R. and Blache, P. and Espesser, R. and Ferr, G. and Meunier, C. and Priego-Valverde, B. and Rauzy, S. "Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle" Traitement Automatique des Langues 49(3):1-30, 2008
- [4] Tjong, E.F. and Sang, K. and Djean, H. "Introduction to the CoNLL-2001 shared task: clause identification", Proceedings of the 2001 workshop on Computational Natural Language Learning, 7:127-132, 2001
- [5] Pevzner, L. and Hearst, M. A. "A critique and improvement of an evaluation metric for text segmentation", Computational Linguistics, 28(1):19-36, 2002
- [6] Avanzi, M. and Lacheret-Dujour, A. and Victorri, B. "A corpus-based learning method for prominence detection in spontaneous speech" Vth International Conference Speech Prosody, 2010
- [7] Hirst, D. "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation" Proceedings of the XVth International Conference of Phonetic Sciences: 12331236, 2007
- [8] Mertens, P. "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model" Proceedings of Speech prosody, 2004
- [9] Simon, A. C. and Degand, L. "On identifying basic discourse units in speech: theoretical and empirical issues" Discours, 4, 2009
- [10] Cohen, J. "A coefficient of agreement for nominal scales" Educational and psychological measurement, 20(1):37-46, 1960
- [11] Fournier, C. and Inkpen, D. "Segmentation similarity and agreement" Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 152-161

What's new in SPPAS 1.5?

Brigitte Bigi¹, Daniel Hirst^{1,2}

¹LPL, CNRS, Aix-Marseille Université, Aix-en-Provence, France

²School of Foreign Languages, Tongji University, Shanghai, China

brigitte.bigi@lpl-aix.fr, daniel.hirst@lpl-aix.fr

Abstract

During Speech Prosody 2012, we presented *SPPAS*, *SP*eech *P*honetization *A*lignment and *S*yllabification, a tool to automatically produce annotations which include utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. *SPPAS* is open source software issued under the GNU Public License. *SPPAS* is multi-platform (Linux, MacOS and Windows) and it is specifically designed to be used directly by linguists in conjunction with other tools for the automatic analysis of speech prosody. This paper presents various improvements implemented since the previously described version.

Index Terms: phonetic, annotation, segmentation, intonation

1. Introduction

During Speech Prosody 2012, we presented version 1.3 of *SPPAS* (*SP*eech *P*honetization *A*lignment and *S*yllabification). *SPPAS* was presented as a tool to produce automatic annotations which include utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. The resulting alignments are a set of TextGrid files, the native file format of the Praat software [1] which has become the most popular tool for phoneticians today. *SPPAS* generates separate TextGrid files for 1/ utterance segmentation, 2/ word segmentation, 3/ syllable segmentation and 4/ phoneme segmentation.

An important point for a software which is intended to be widely distributed is its licensing conditions. *SPPAS* uses only resources and tools which can be distributed under the terms of the GNU Public License. *SPPAS* tools and resources are currently available at the URL:

<http://www.lpl-aix.fr/~bigi/sppas/>

Since the version presented in [2], we continued to improve the tool. Our improvements are related to the 4 following aspects:

1. Technical stuff: multi-platform, easy to install, UTF-8 support;
2. Graphical User Interface: improved ergonomics, documentation and help, some components added;
3. Annotations: Momel and INTSINT added; Tokenization added; IPU-segmentation improved;
4. Resources: acoustic model for Chinese changed, Taiwanese support, conversion to SAMPA.

The new *SPPAS* architecture can be summarized as:

- a set of automatic annotation tools,
- a set of components,

- two solutions to use them:

1. a Graphical User Interface (GUI) to use *SPPAS* which is as "user-friendly" as possible;
2. a set of tools, each one essentially independent of the others, that can be run on its own at the level of the shell.

2. Technical stuff

Since version 1.4, *SPPAS* is implemented with the programming language *python*. This allows the tool to work under Linux, Mac-OSX and Windows®. It is also much easier to install.

In the previous version, only TextGrid files were supported. The current version can import files from Transcriber [3] and Elan [4] softwares. We also fixed the encoding to UTF-8 only.

3. Graphical User Interface

The GUI consists of two main area, named the file list panel (FLP) and the automatic annotation panel (AAP).

The FLP displays a set of buttons and a tree-style list. The list contains Directories and Files which the user has added, but only files that *SPPAS* can handle (recognised by the file extension). The FLP makes it possible to exit the tool and to manage the list: add files, add directories, remove, delete, export.

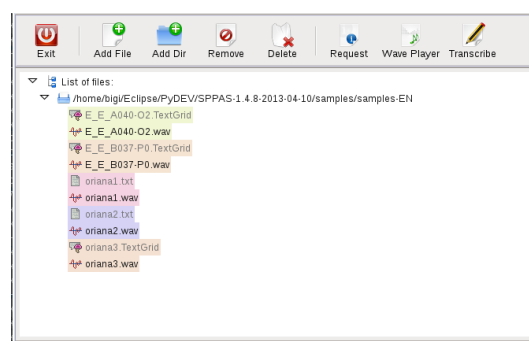


Figure 1: The file list panel.

The AAP consists of a list of buttons to check, the annotation name and buttons to fix the language of each annotation. A specific language can be selected for each annotation depending on the resources available in the package. This allows the users to add their own resources or to copy/modify existing resources.

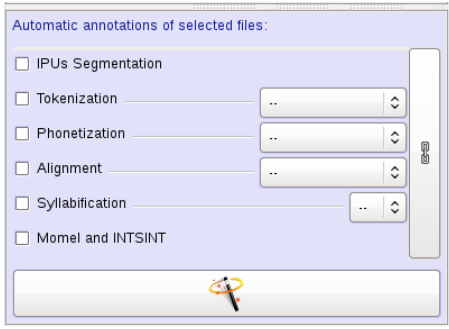


Figure 2: The automatic annotation panel.

As SPPAS is designed to be used directly by linguists, another important improvement is related to the Help and the Documentation. We paid particular attention to this. Finally, to facilitate the use of our tool, we decided to add some extra components. Currently, three components are available: 1/ **wav player** is a simple tool used to play sounds; 2/ **transcribe** is a tool dedicated to speech transcription; 3/ **requests** is a set of functionalities related to the annotation manipulation.

3.1. Transcribe

The key-point of this component is that it automatically performs a speech/silence segmentation. Then, only speech segments are displayed (see Figure 3). If more than one sound file has to be transcribed (as for a dialogue for example), speech segments are displayed interlaced to facilitate the transcription process.

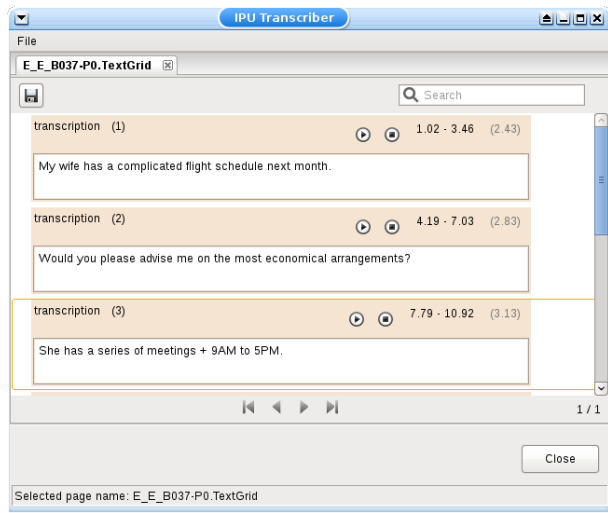


Figure 3: The transcription frame.

3.2. Requests

We added a component to get information, modify and request annotated files (see Figure 4). This allows the user to manage annotated files and the tiers of these files: rename, delete, cut, copy, paste duplicate, move up, move down, view. We also added a frame that prints elementary statistics (as in Figure 5).

	Label	Number of Occurrences	Total Duration	Average Duration
1	#	7	5.22	0.745714285714
2	m	14	0.9899	0.0707071428571
3	al	8	0.9	0.1125
4	w	2	0.17	0.085
5	f	6	0.59	0.0983333333333
6	h	4	0.41	0.1025
7	{	6	0.53	0.0883333333333
8	z	9	0.9477	0.1053
9	@	25	1.05	0.042
10	k	13	1.21	0.0930769230769
11	A	7	0.47	0.0671428571429
12	p	4	0.38	0.095
13	l	10	0.44	0.044
14	el	4	0.4	0.1
15	4	4	0.37	0.0925
16	d	8	0.44	0.055
17	t	11	0.89	0.0809090909091

Figure 5: The statistics of a tier.

Finally, we added an advanced filtering tool. In the following, X represents an interval, $L(X)$ the label of X , and $L(.)$ one label to find. We thus propose to select intervals depending on their label with the following capabilities:

- $L(X) = L(.)$, exact match: the labels must strictly correspond,
- $L(X) \in L(.)$, contains: the label of the tier contains the given label,
- $L(X) \sqsubset L(.)$, starts with: the label of the tier starts with the given label,
- $L(X) \sqsupset L(.)$, ends with: the label of the tier ends with the given label.

All these matches can be used in their negative form. To cope with specific needs, a multiple pattern selection has been implemented to search n patterns $L_1(.)$, $L_2(.)$, \dots , $L_n(.)$ as:

$$X : [L(X) \text{ op } (L_1(.) \vee L_2(.) \vee \dots \vee L_n(.))]$$

where op represents one of the relations $=$, \in , \sqsubset , \sqsupset . At last, the proposed filtering system makes it possible to fix duration constraints. Let $D_m(.)$ be a minimal duration, $D_M(.)$ be a maximal duration and $D(X)$ be the duration of interval X . Duration constraints are written as:

- $X : [D(X) > D_m(.)]$, to fix a minimal duration on X ,
- $X : [D(X) < D_M(.)]$, to fix a maximal duration on X .

For example, the request "Extract all words starting by "ch" with a duration of at least 100ms" is expressed as:

$$X : [L(X) \sqsubset L(ch)] \{eq\} [D(X) > D_m(100ms)]$$

These constraints can be applied to a whole tier or to just a part of the tier by fixing a start time and an end time.

4. Annotations

4.1. IPU segmentation

Inter-Pausal Units (IPUs) segmentation consists in aligning the macro-units of a document (based on their transcription) with

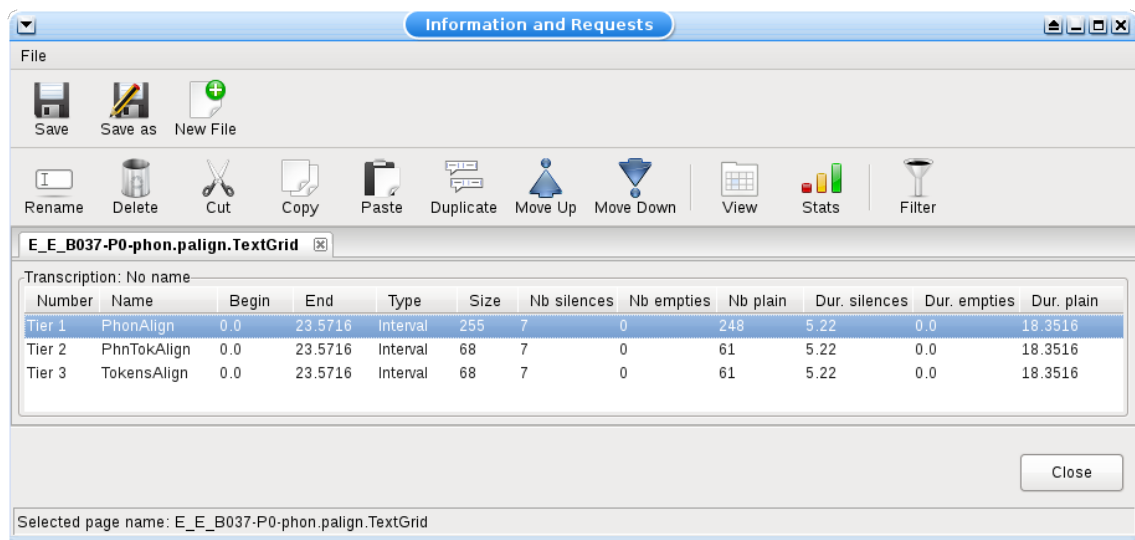


Figure 4: The frame to manipulate annotated files.

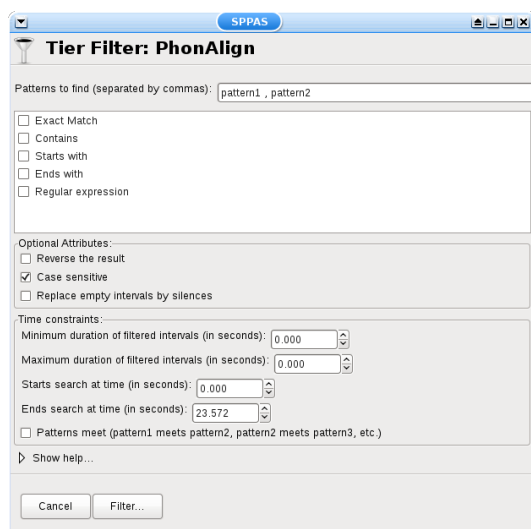


Figure 6: Filtering a tier.

the corresponding sound. A recorded speech file with the .wav extension should correspond to each .txt file. The segmentation provides a TextGrid file with one tier named "IPU". IPU Segmentation annotation performs a simple silence detection if no transcription is available (the volume is automatically adjusted). Current version allows to fix a shift value to speech boundaries.

4.2. Tokenization

Tokenization is the process of segmenting a text into tokens. In principle, any system that deals with unrestricted text needs the text to be normalised. Texts contain a variety of "non-standard" token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL's and e-mail addresses... Normalising or rewriting such texts using ordinary words is then an important issue.

SPPAS implements a generic approach for text normalisation, in view of developing a multi-purpose multi-lingual text corpus. This approach consists in splitting the text normalisation problem into a set of minor sub-problems each of which is as language-independent as possible. This approach is described in [5].

The Tokenization process takes as input a transcription that can be enriched by various phenomena, such as:

- truncated words, noted as a '-' at the end of the token string (an ex- example);
- liaisons, noted between '=' (an =n= example);
- noises, noted by a '*' (only for French and Italian);
- laughs, noted by a '@' (only for French);
- short pauses, noted by a '+' (a + example);
- elisions, mentioned in parenthesis;
- specific pronunciations with brackets [example,eczap];
- comments with braces or brackets {this} or [this];
- morphological variants with <like,lie ok>;
- proper name annotation, like \$John Doe\$.

4.3. Phonetisation

Phonetisation, also called grapheme-phoneme conversion, is the process of representing sounds with phonetic signs. The phonetisation is the equivalent of a sequence of dictionary look-ups. It is generally assumed that all words of the speech transcription are mentioned in the pronunciation dictionary. Otherwise, SPPAS implements a language-independent algorithm to phonetise unknown words. At this stage, it consists in exploring the unknown word from left to right and then finding the longest strings in the dictionary. Since this algorithm uses the dictionary, the quality of such a phonetisation will depend on this resource.

4.4. Alignment

Phonetic alignment consists in a time-matching between a given speech utterance and a phonetic representation of the utterance. For each utterance, the orthographic and phonetic transcriptions are used. SPPAS performs an alignment to identify the temporal boundaries of phones and words. Speech alignment requires an acoustic model in order to align speech. An acoustic model is a file that contains statistical representations of each of the distinct sounds of one language. The quality of such an alignment will depend on this resource.

4.5. Syllabification

The syllabification of phonemes is performed with a rule-based system previously described for French in [6]. A new set of rules was developed to deal with Italian.

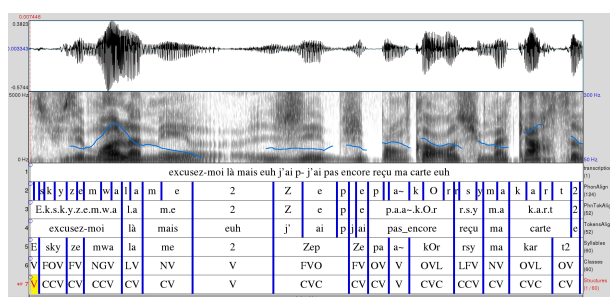


Figure 7: SPPAS output example on French spontaneous speech.

4.6. Momel and INTSINT

Momel (modelling melody) [7, 8] is an algorithm for the automatic modelling of fundamental frequency (F0) curves using a technique called asymmetric modal quadratic regression. This technique makes it possible by an appropriate choice of parameters to factor an F0 curve into two components:

1. a macroprosodic component represented by a quadratic spline function defined by a sequence of target points $\langle ms, Hz \rangle$.
2. a microprosodic component represented by the ratio of each point on the F0 curve to the corresponding point on the quadratic spline function.

Since several different techniques of F0 extraction are possible, Momel requires a file containing the F0 values detected from the signal.

Encoding of F0 target points using the "INTSINT" system [9] assumes that pitch patterns can be adequately described using a limited set of tonal symbols, T, M, B, H, S, L, U, D (standing for : Top, Mid, Bottom, Higher, Same, Lower, Upstepped, Downstepped respectively) each one of which characterises a point on the fundamental frequency curve.

The rationale behind the INTSINT system is that the F0 values of pitch targets are programmed in one of two ways: either as absolute tones T, M, B which are assumed to refer to the speaker's overall pitch range (within the current Intonation Unit), or as relative tones H, S, L, U, D assumed to refer only to the value of the preceding target point.

A distinction is made between non-iterative H, S, L and iterative U, D relative tones since in a number of descriptions it appears that iterative raising or lowering uses a smaller F0 interval than non-iterative raising or lowering.

5. Resources

Since the version we presented in [2], we continued to improve the resources as:

- Chinese: can deal with chinese characters or with pinyin, new Chinese acoustic model, and some minor changes in the dictionary;
- English acoustic models updated
- new Italian acoustic model;
- partially Taiwanese support.

A new French model is under construction. All models were converted to Sampa.

6. Conclusions

SPPAS is specifically designed with the aim of providing a tool for phoneticians rather than for computer-scientists, because no such a tool is currently available under a GPL license.

Current development is in progress to continue to improve the accessibility, to add new language, new annotations, and new components.

7. Acknowledgements

This work is supported by ORTOLANG: <http://www.ortolang.fr/>

8. References

- [1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer, <http://www.praat.org/>" 2009.
- [2] B. Bigi and D.-J. Hirst, "Speech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody," in *Proc. of Speech Prosody*, Tongji University Press, Ed., Shanghai (China), 2012.
- [3] TranscriberAG, "A tool for segmenting, labeling and transcribing speech. [computer software] paris: Dga," <http://transag.sourceforge.net/>, 2011.
- [4] H. Sloetjes, A. Russel, and A. Klassmann, "Elan: a free and open source multimedia annotation tool," 2010.
- [5] B. Bigi, "A multilingual text normalization approach," in *2nd Less-Resourced Languages workshop, 5th Language & Technology Conference*, Poznań (Poland), 2011.
- [6] B. Bigi, C. Meunier, I. Nesterenko, and R. Bertrand, "Automatic detection of syllable boundaries in spontaneous speech," in *Language Resource and Evaluation Conference*, La Valetta (Malta), 2010, pp. 3285–3292.
- [7] D.-J. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function," *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, pp. 75–85, 1993.
- [8] D.-J. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIIth International Conference of Phonetic Sciences*, Saarbrücken, 2007, pp. 1233–1236.
- [9] —, "The analysis by synthesis of speech melody: from data to models." *Journal of Speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

TGA: a web tool for Time Group Analysis

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft,
Universität Bielefeld, Germany

gibbon@uni-bielefeld.de

Abstract

Speech timing analysis in linguistic phonetics often relies on annotated data in *de facto* standard formats, such as Praat TextGrids, and much of the analysis is still done largely by hand, with spreadsheets, or with specialised scripting (e.g. Praat scripting), or relies on cooperation with programmers. The *TGA* (*Time Group Analyser*) tool provides efficient ubiquitous web-based computational support for those without such computational facilities. The input module extracts a specified tier (e.g. phone, syllable, foot) from inputs in common formats; user-defined settings permit selection of sub-sequences such as inter-pausal groups, and duration difference thresholds. Tabular outputs provide descriptive statistics (including modified deviation models like *PIM*, *PPD*, *nPVI*, *rPVI*), linear regression, and novel structural information about duration patterns, including difference *n-grams* and *Time Trees* (temporal parse trees).

Index Terms: web tools, speech timing, speech prosody, annotation processing, duration tokens, time trees

1. Background and requirements

Speech timing analysis in linguistic phonetics often relies on time-stamped annotated data in *de facto* standard formats, such as Praat TextGrids [1], Transcriber XML formats or tables with character separated fields (CSV tables). Typical applications are the analysis of speech rate, or measuring duration deviation and relative ‘fuzzy’ isochrony, either relative to the whole sequence, as with standard deviation, *Pairwise Irregularity Measure*, *PMI* [2], *Percentage Foot Deviation*, *PPD* [3], or relative to adjacent units (*raw* and *normalised Pairwise Variability Indices*, *rPVI*, *nPVI* [4]).

The literature reveals several methods for processing time-stamped data, in order of increasing sophistication:

1. copying into spreadsheets for semi-manual processing;
2. use of prefabricated Praat scripts for time-stamped annotations;
3. creation of Praat scripts for specific analysis tasks;
4. implementation of applications in general scripting languages such as *Perl*, *Tcl*, *Ruby* or *Python*, for TextGrid, *SAM*, *ESPS*, *WaveSurfer* etc., formats;
5. implementation in languages such as *C*, *C++* (mainly in speech technology applications), independently of time-stamping visualisation software.

The existence of many web applications and spreadsheet templates for manual calculation, sometimes with page user counts, documents the widespread use of (semi-)manual analysis methods. For those with programming abilities, libraries of analysis tools are available, e.g. those in the Aix-MARSEC repository [5], or parsing functions programmed in *Python*, such as the *Natural Language Took Kit*, *NLTK* [6], or

the *TextGrid tools* [7]. The web-based *Time Group Analyser* (*TGA*) tool, also implemented in *Python*, fills a gap between non-computational and computational users: a wide range of analyses is provided, with no need for *ad hoc* programming. TextGrid post-processing with the *TGA* is complementary to TextGrid generation with tools such as *SSPAS* [8].

The following account describes *TGA* input (Section 2), processing (Section 3), output (Section 4), and the *Python* implementation (Section 5). The term ‘annotation label’ is used for time-stamped triples $\langle \text{label}, \text{start}, \text{end} \rangle$, ‘annotation event’ is used for pairs $\langle \text{label}, \text{duration} \rangle$, and ‘Time Group’ refers to an event sequence with a well-defined boundary condition, such as an inter-pausal group or continuous deceleration or acceleration. The *TGA* functions analyse and visualise duration differences (Δdur) relative to thresholds: deceleration, rallentando, quasi-iambic (Δdur^+); acceleration, accelerando, quasi-trochaic (Δdur^-); equality, threshold-relative ‘fuzzy isochrony’ ($\Delta \text{dur}^=$).

2. Input and parameter setting

The *TGA* input module extracts a specified tier (e.g. phone, syllable, foot) from inputs in long or short TextGrid format or as CSV tables with any common separator. User-defined parameter settings currently include the following:

1. freely selected tier name, e.g. ‘Syllables’, and boundary symbol list (e.g. ‘_’, ‘p’, ‘sil’, ‘\$p’ for pauses);
2. *Time Group* division criterion (by *pauses*; or based on Δdur , i.e. changes in speech rate: decrease (*deceleration*) or increase (*acceleration*));
3. minimal *Time Group* length in duration intervals (where rhythm is concerned, at least 2 interval events (linking 3 point/boundary events) are needed to define a rhythm [9]);
4. global Δdur duration threshold range, e.g. 50...100 ms, 100 ... 200 ms, etc.;
5. local duration Δdur threshold, for local structure determination;
6. local Δdur tokens for visualising duration patterns, e.g. ‘\’, ‘/’, ‘=’ for ‘longer’, ‘shorter’, ‘equal’;
7. *Time Tree* type specification (decelerating, rallentando, ‘quasi-iambic’ vs. accelerating, accelerando, ‘quasi-trochaic’).

3. TGA modules

Currently there are three main *TGA* modules besides I/O and format conversion: (1) text extraction; (2) global basic descriptive statistics for all elements of the specified tier; (3) segmentation of the tier into *Time Groups* with statistics for individual *Time Groups*, and with three new visualisation techniques for Δdur duration patterns: duration difference tokens, duration column charts, and *Time Trees*.

3.1. Text extraction

Labels are extracted from annotation elements as running text, separated into sequences by the boundary criteria, e.g. pause, specified in the input. When the annotation has been made without prior transcription there may be a need for text extraction, as documented by a number of web pages providing this functionality, for various purposes such as discourse analysis, natural language processing, archive search, re-use as prompts in new recordings. No further computational linguistic analysis of the text output is undertaken by the TGA at present.

3.2. Global descriptive statistics

For calculating global descriptive statistics, three versions of the data are prepared: (1) with all annotation elements on the tier, including boundary elements (e.g. pauses); (2) with only non-boundary elements; (3) with only boundary elements. The following information is provided for each data version:

1. *n*, *len*: the number of elements in the input (for data versions with or without pauses, or only pauses), and the total duration Δt ;
2. *min*, *max*, *mean*, *median*, *range*: basic statistical properties;
3. *standard deviation*, *PIM*, *PFD*, *rPVI*, *nPVI*: ratio or difference measures of deviation Δdur , of an element from a reference value, e.g. mean or adjacent element;
4. *linear regression (intercept, and slope)*: slope indicates the average rate of duration change in the data (deceleration, acceleration, equality).

The *PIM*, *PFD* and *rPVI* metrics are distinguished partly for their popularity, rather than for significant differences: *PIM* and *PFD* relate closely to standard deviation, though the *PIM* uses global ratios rather than differences. The *nPVI* on the other hand factors out drifting speech rates, and may thus diverge very widely from the mean. The measures are claimed to be rhythm metrics, though they define only necessary, not sufficient conditions for rhythm: unsigned Δdur values ignore alternation in duration patterns, a necessary condition for rhythm models (cf. Section 3.3). The formulae for *PIM*, *PFD*, *rPVI* and *nPVI* are shown in Table 1.

Statistical ‘rule of thumb’ quality scores, such as *p*-value or confidence intervals, are not included at this time.

$PIM(I_{1...n})$	$= \sum_{i \neq j} \log \frac{I_i}{I_j} $
$PFD(foot_{1...n})$	$= \frac{100 \times \sum MFL - len(foot_i) }{len(foot_{1...n})}$ where $MFL = \frac{\sum_{i=1}^n len(foot_i)}{n}$
$rPVI(d_{1...m})$	$= \sum_{k=1}^{m-1} d_k - d_{k+1} / (m-1)$
$nPVI(d_{1...m})$	$= 100 \times \sum_{k=1}^{m-1} \left \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right / (m-1)$

Table 1: Definitions of *PIM*, *PFD*, *nPVI* measures.

3.3. Local Time Group statistics

Basic statistics and linear regression are calculated for each *Time Group* separately in the same way as for the global calculations. Minimal difference thresholds permit approximate (i.e. ‘fuzzy isochrony’) measures, rather than

strict time-stamp differences. Three novel structural Δdur pattern visualisations are defined:

1. tokenisation of duration differences Δdur into ‘longer’, ‘shorter’ and ‘equal’ duration difference tokens, represented by character symbols (cf. Figure 4 and Table 2), to support prediction of whether specific properties such as *rhythmic alternation* are likely to make sense;
2. top-suspended column chart illustrating the duration Δt of elements in the *Time Group* (Figure 4);
3. duration parse tree (*Time Tree*) for each *Time Group* (Figure 5), based on signed duration differences Δdur^+ and Δdur^- , [10], [11], to facilitate study of correspondences between duration hierarchies and grammatical hierarchies.

The *Time-Tree* induction algorithm follows a standard deterministic context-free bottom-up left-right shift-reduce parser schedule. The grammars use Δdur^+ and Δdur^- tests on annotation events in order to induce two types of *Time Tree*, with ‘quasi-iambic’ (decelerating, *rallentando*), and ‘quasi-trochaic’ (accelerating, *accelerando*) constituents:

Quasi-iambic:

$$TT_k \rightarrow TT_i TT_j$$

$$duration(TT_i) < duration(TT_j)$$

$$duration(TT_k) \text{ INHERITS } duration(TT_j)$$

Quasi-trochaic:

$$TT_k \rightarrow TT_i TT_j$$

$$duration(TT_i) > duration(TT_j)$$

$$duration(TT_k) \text{ INHERITS } duration(TT_i)$$

In each of these grammars, a right-hand side *TT* is a label-duration pair, and higher levels in the tree inherit durations recursively from the constituent annotation events.

Crucially, Δdur token patterns and *Time Trees*, (unlike *standard deviation*, *PIM*, *PFD*, *rPVI*, *nPVI*) use signed, not unsigned duration differences, and may therefore lay claim to representing true rhythm properties. In each case, the minimal local difference threshold setting applies.

4. Output

The output provides various list and table formats:

1. list of label text sequences within *Time Groups*, with any accompanying symbols for boundary events;
2. table of *Time Group* properties:
 1. statistical properties,
 2. tokenised Δdur^+ and Δdur^- deceleration-acceleration patterns,
 3. top-suspended column charts of durations,
 4. *Time Trees* built on the Δdur^+ or Δdur^- relations;
3. table with summary of basic statistics, linear regression, and correlations between different statistics, for the complete set of *Time Groups*;
4. list of Δdur duration difference token *n*-grams from all *Time Groups* (unigrams, digrams, trigrams, quadgrams and quingrams) to support analysis of rhythmically alternating patterns in the annotations;
5. various character-separated value tables of input and output for further analysis using other software.

5. Implementation

The architecture of the TGA tool implementation is visualised in Figure 1. The user inputs the annotation in a TextGrid or CSV format using an HTML form and selects the

required processing settings. The input is passed over the internet via the Common Gateway Interface (CGI) to the TGA server; processing is performed in Python, and the output returns to the user as an HTML page.

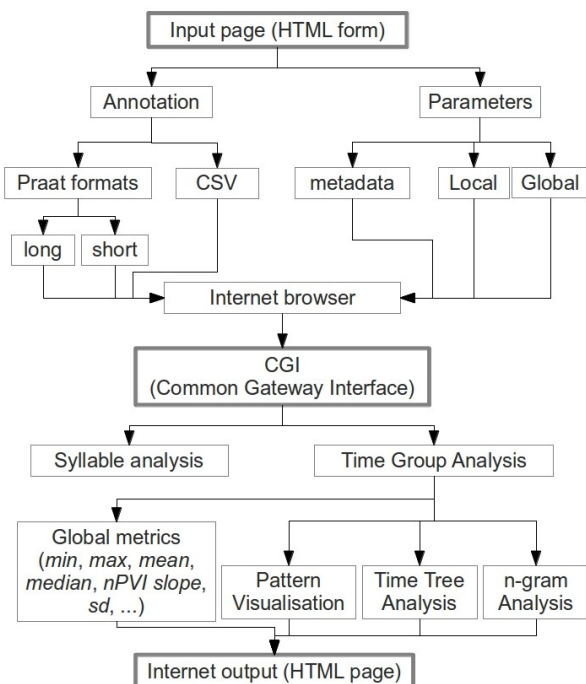


Figure 1: TGA architecture.

Currently accepted input formats are Praat short and long TextGrids, or CSV formats with various field separator options. The input screen layout is shown in Figure 2.

A number of output selection options are also provided on the input page: annotation text, global descriptive statistics, metrics for individual *Time Groups*, token patterns, *Time Trees* or selected output in CSV tables for further processing with spreadsheets, etc.

All the following illustrations are from TGA output for the syllable tier of Aix-MARSEC annotation A0101B.

The automatic label text output from the annotation elements in the *Time Groups* appears straightforwardly, as a list of *Time Group* text sequences:

'gUd 'mO:nIN

'mO: 'nju:z @'baUt D@ 'revr@n 'sVn 'mjVN 'mu:n

'faUnd@r @v D@,ju:nIfI'keISn 'tS3:tS

'hu:z 'kVr@ntl In 'dZell

The quantitative output types display as tables, both for individual *Time Groups* and for generalisations over individual *Time Groups*, as in Figure 3.

Figure 4 shows two aligned novel visualisation types: duration tokens and duration bars. The top sequence of symbols represents tokenisations of Δdur between adjacent intervals, in this case showing a possibly rhythmical shorter-longer alternation (Δdur tokenisation is controlled by adjusting the local Δdur threshold setting).

The the top-suspended column chart below the token sequence provides an iconic visualisation of durations, in width (to show time scaling) and in height. Top-suspension emphasizes the *rallentando* (deceleration, iambic,

downwards) and accelerando (acceleration, trochaic, upwards) tendencies, providing immediate visual sources of hypotheses about rhythmicity for perceptual testing and linguistic analysis.

TextGrid input control parameters (long or short TextGrid format accepted; only Interval Tiers, obviously)

Tier name: (max length 20; not needed for CSV formats)

Pause symbol: (max length 20; also needed for CSV formats)

More than one pause symbol permitted; separate with spaces. Delete any of the examples which might occur as an annotation label. If your pause symbol is not in the examples given, enter it

Time Group duration difference parameters:

TG criterion: ☒ *pausegroup* ☐ *deceleration* (increasing) ☐ *acceleration* (decreasing)

Local threshold: ms (try values less than common syllable lengths, e.g. 0 ... 300 ms)

Used for local pattern extraction and TimeTree parsing.

Local pattern symbols: **Longer:** (1 char) **Shorter:** (1 char) **Same:** (1 char)

Time Tree criterion: ☐ *(quasi-)iambic TTgr* ☐ *(quasi-)trochaic TTl* ☒ *show all TT*
☐ *(quasi-)iambic TTgre* ☐ *(quasi-)trochaic TTlre* ☐ *do not show TT*

Global TG threshold range: ... ms (minimal duration difference)

Ranges > 30 are not permitted because of possible server overload.

Global threshold is ignored with the 'pausegroup' criterion.

Experiment with values from 0 to 500 (negative values are permitted).

Equal range boundaries are adjusted to have range of 1, not null; if necessary values are switched to ensure 'low before high'.

Min TG length: > (generally >2, as 'minimal rhythm')

Time Group output control parameters:

Print text? ☒ *no* ☐ *yes* **n-grams?** ☒ *no* ☐ *yes* **All outputs:** ☐ *no* ☒ *yes*

TG element info? ☐ *no* ☒ *yes* **Time Trees?** ☐ *no* ☒ *yes*

TG detail? ☐ *no* ☒ *yes* **CSV output?** ☐ *no* ☒ *yes*

Figure 2: Screenshot of parameter input options.

Summary table of global and accumulated TG duration functions (some do make sense...)					
Time Group criterion: pausegroup , local threshold: 10 , Min valid TG length: 2					
Only inter-pause intervals measured; pauses not included					
Overall duration:	48504	Overall raw longer, ms:	15401	Overall raw shorter, ms:	14521
Overall min:	20.00	Overall max:	990.00	Overall range:	970.00
Valid Time Groups:	34	Overall rate/sec:	5.67		
Components: global tendencies					
Overall mean:	176.38	Overall median:	150.00	Overall SD:	113.58
Overall npvi:	62.00	Overall intercept:	156.12	Overall slope:	0.15
Mean of means:	182.18	Median of means:	176.70	SD of means:	34.75
Mean of medians:	168.68	Median of medians:	160.00	SD of medians:	40.88
Mean of SDs:	90.02	Median of SDs:	86.16	SD of SDs:	39.87
Mean of nPVIs:	60.00	Median of mnPVIs:	51.00	SD of nPVIs:	17.91
Mean of intercepts:	143.59	Median of intercepts:	130.80	SD of intercepts:	71.16
Mean of slopes:	10.65	Median of slopes:	11.86	SD of slopes:	41.10
Components: correlations					
mean::TGdur:	-0.190	median::TGdur:	-0.427	SD::TGdur:	0.230
nPVI::TGdur:	0.097	slope::TGdur:	0.061	intercept::TGdur:	-0.178
nPVI::mean:	0.128	slope::mean:	0.028	intercept::mean:	0.503
nPVI::median:	0.026	slope::median:	0.005	intercept::median:	0.310
nPVI::SD:	0.383	slope::SD:	0.051	intercept::SD:	0.229

Figure 3: Screenshot of summary of collated *Time Group* properties and correlations.

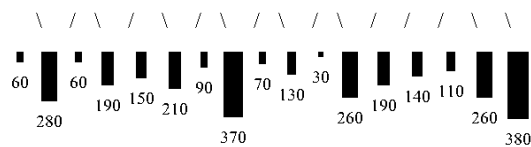


Figure 4: Duration difference token pattern (above) and top-suspended duration columns with duration represented by both width and length (below).

The *Adur* token digram analysis provides the following output format, showing rank and frequency of token digrams (see Table 2 for token frequencies above 10%).

Rank	Percent	Count	Token digram
1	22%	60	/\
2	20%	55	\/
3	11%	31	\ \

Table 2: *Adur* token rank and frequency analysis.

In this instance of ‘educated Southern British’ pronunciation, i.e. slightly modified Received Pronunciation (RP), alternations figure at the top two ranks, totalling 42% of the digrams, and therefore have potential for rhythm; deceleration patterns occupy rank 3.

Finally, perhaps the most interesting display format is the *Time Tree* visualisation, here shown as automatically generated nested parentheses. The example in Figure 5 illustrates this principle with the inter-pausal group ‘about Anglican ambivalence to the British Council of Churches’.

```
( ( (@ baUt)
  ( ( ({N gLI}
    (kn {m})
    (bI vl@ns)))
  ( ( ( (t@ D@)
    (brI tIS))
    ( ( kaUn
      ( sl
        (@v tS3:)))
      tSiz))
    PAUSE))
```

Figure 5: Automatic prettyprint of a quasi-iambic *Time Tree* in nested parenthesis notation.

The purpose of generating *Time Tree* output is to support study of the relation between temporal hierarchical structures and grammatical constituents in a systematic *a posteriori* manner, rather than postulating higher level units such as feet or other events types in an *a priori* prosodic hierarchy framework. This example shows a number of correspondences with grammatical units at different depths of embedding, e.g. ‘about’, ‘British’, ‘Anglican ambivalence’, ‘about Anglican ambivalence’, ‘Council of Churches’, ‘to the British Council of Churches’, including foot sequences of Jassem’s ‘Anacrusis + Narrow Rhythm Unit’ type [12].

6. Conclusion and outlook

The design and implementation of a web tool for support of linguistic phonetic analysis of speech timing, using time-stamped data, are described. Extensive basic statistical information is provided, including linear regression and correlations between different statistics. Three innovative visualisations are introduced: *Adur* duration difference tokens; top-suspension column charts for Δt and Δdur visualisation, and *Adur* based *Time Trees* automatically represented as nested parentheses. Informal evaluation by four trained phoneticians shows that the TGA tool reduces previous analysis times for time-stamped annotations by several orders of magnitude. Initial work on Mandarin Chinese is reported in [13] and [14]. An offline version of TGA for processing large annotation corpora rather than single files is undergoing testing.

The tokenisation and *Time Tree* techniques are very much research in progress. Ongoing work concerns extension of TGA functionality, particularly correspondences between *Time Groups* and grammatical, focus-based and rhetorical categories, coupled with the automatic discovery of inter-tier time relations based on temporal logics [15].

A recent version of the TGA, with the data illustrated in the present paper, can currently be accessed at the following URL:

<http://www.homes.uni-bielefeld.de/gibbon/TGA>

7. Acknowledgments

Special thanks are due to Jue Yu, Zhejiang University, Hangzhou, China, to Katarzyna Klessa and Jolanta Bachan, Adam Mickiewicz University, Poznań, Poland, for feedback on applications of the TGA, and to participants and reviewers of O-COCOSDA 2012, Macau, as well as reviewers of TRASP 2013 for very helpful critical discussion.

8. References

- [1] Boersma, P. “Praat, a system for doing phonetics by computer”, *Glott International* 5:9/10, 341-345, 2001.
- [2] Scott, D. R., Isard, S. D. and de Boysson-Bardies, B. “On the measurement of rhythmic irregularity: a reply to Benguerel”, *Journal of Phonetics* 14, 327–330, 1986.
- [3] Roach, P. “On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages”. In *Linguistic Controversies: Essays in Linguistic Theory and Practice*, D. Crystal, Ed., London: Edward Arnold, 73–79, 1982.
- [4] Low, E. L., Grabe, E. and Nolan, F. “Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English”. *Language and Speech* 43(4):377–401, 2000.
- [5] Hirst, D. Auran, C. and Bouzon, C. The Aix-MARSEC database. 2002-2004. Tech. Report, Equipe Prosodie et Représentation Formelle du Langage, Laboratoire CNRS UMR 6057 Parole et Langage, Université de Provence, Aix-en-Provence, 2009.
- [6] Bird, S., Klein, E. and Loper, E. *Natural Language Processing with Python*. Beijing, etc.: O’Reilly, 2009.
- [7] Buschmeier, H., and Włodarczak, M. “TextGridTools: A TextGrid Processing and Analysis Toolkit for Python”. *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013)*, Bielefeld, Germany, 152–15, 2013.
- [8] Bigi, B. SPPAS: a tool for the phonetic segmentations of speech, LREC 8, Istanbul, 2012.
- [9] Gibbon, D. “Computational modelling of rhythm as alternation, iteration and hierarchy,” in *Proceedings of ICPHS 15*, Barcelona, 2003.
- [10] Gibbon, D. “Corpus-based syntax-prosody tree matching”, in *Proceedings of Eurospeech 2003*, Geneva, 2003.
- [11] Gibbon, D. “Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data”. In: Sudhoff, S. et al. (2006). *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 281-209, 2006.
- [12] Gibbon, D., Hirst, D. and Campbell, N. 2012. *Rhythm, Melody and Harmony: Studies in Honour of Wiktor Jassem*. Poznań: Polish Phonetics Society (Speech & Language Technology 14/15).
- [13] Yu, J. and Gibbon, D. “Criteria for database and tool design for speech timing analysis with special reference to Mandarin”, *Proceedings of Oriental COCOSDA*, Macau, 2012.
- [14] Yu, J. 2013. *Timing analysis with the help of SPPAS and TGA tools*. TRASP 2013.
- [15] Carson-Berndsen, J. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*, Dordrecht: Kluwer Academic Publishers, 1998.

Timing analysis with the help of SPPAS and TGA tools

Jue Yu

School of Foreign Languages, Tongji University, Shanghai, China

erinyu@126.com

Abstract

This paper is trying to solve the big problems facing phoneticians and linguists in the study of duration, timing and speech rhythm, that is, heavy manual work during the annotating process, and how to generate more accurate and objective analysis results based on a large speech database. Two newly-developed speech tools are discussed: SPPAS, a tool for automatic phonetic segmentation of speech and TGA, a tool for automatic timing analysis. A case study was carried out to demonstrate that, with suitable models and tools for processing speech corpora, (1) the time required to transcribe speech data can be reduced with the help of SPPAS to about 33% of the manual annotation time, and (2) analysis of speech timing in annotations can be facilitated by using TGA.

Index Terms: SPPAS, TGA, Time Group Analysis, Time Trees, Chinese

1. Introduction

Speech timing is always a hot issue in phonetics, phonology, psychology and speech engineering. In order to study the relation between speech timing patterns and linguistic structures in Chinese dialects and in Chinese L2 speakers of English, a new approach is taken: Time Trees [1] are constructed from syllable annotations of speech recordings, and correspondences between their smallest constituents and language units are examined. This approach differs from more traditional studies in terms of “speech rhythm” or in terms of duration variation using models of duration difference averages, or in terms of different timing models, from the single-level duration models to multilevel modeling approaches and to studies of the multiple factors underlying durations.

So far, studies in speech timing show that an approach based on large corpora is necessary both for the study of speech production and for speech synthesis with reasonable quality. For example, Dellwo et al. [2] pointed out that rhythm studies really require the analysis of longer sequences of speech data, otherwise artifacts may appear in the results. Sagisaka et al. [3] state that fine control of segmental duration based on a large corpus has been proved to be essential in synthesizing speech with natural rhythm and tempo. Bigi & Hirst [4] came to the more general conclusion that today it is becoming more and more expected for linguists to take into account large quantities of empirical data, often including several hours of recorded speech.

Thanks to technological progress, a number of graphical software tools for creating annotated audio and/or video recordings of speech have become available such as Praat [5], Transcriber [6] and WaveSurfer [7], Anvil [8], Elan [9]. These tools are basically intended for manual annotation. But the

present problem is the production of large numbers of annotation and their analysis. Large quantities of data require many hours of manual work, which is time-consuming (and can be really frustrating) and therefore imposes a severe restriction on the amount of data which can be used. A better solution for this problem is to borrow methods from speech engineering and use an automatic time-aligned phonetic transcription tool [10] [11]. The second problem is how to analyze speech timing using the large quantities of annotated data. For this purpose a tool designed for automatic timing analysis [12] is available.

This present paper is concerned with these requirements imposed on speech database analysis by the study of duration, timing and speech rhythm, and with suitable models and tools for processing speech corpora, thus presenting a relatively efficient process for reducing the time required to transcribe speech data and for speech timing analysis.

2. Tools used in speech timing

2.1. SPPAS: automatic phonetic segmentation

SPPAS, Speech Phonetization Alignment and Syllabification, is a tool designed by Laboratoire Parole et Langage, Aix-en-Provence, France, to automatically produce annotations which include utterance, word, syllable and phoneme segmentations and their transcriptions from recorded speech. Currently it is implemented for four languages: French, English, Mandarin Chinese and Italian. It is said that adding other languages requires a very simple procedure [4] [11].

SPPAS has a) a phonetician-friendly interface; b) a high rate of correct alignment (correct phoneme alignment rate: 88%; correct word alignment rate: 97.6%) [4]; c) generation of files in the TextGrid format, which can be easily analyzed in detail with the widely used Praat software workbench [3]; d) free web-based software and e) constant improvement [4].

SPPAS generates six TextGrid outputs, four of which are relevant here: (i) utterance segmentation, (ii) word segmentation, (iii) syllable segmentation and (iv) phoneme segmentation. In Chinese, simple words are monosyllabic, so (ii) and (iii) are the same units, but (ii) is orthographic and (iii) is phonetic. The output is illustrated as a screenshot in Figure 1, with TextGrid files merged in Praat.

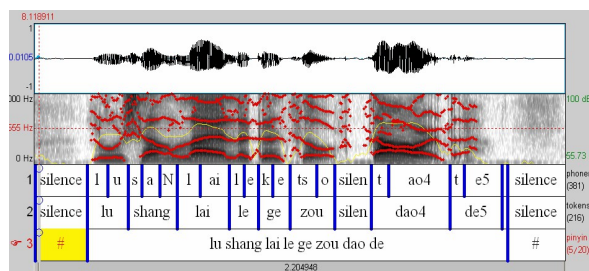


Figure 1: SPPAS output example for the Mandarin Chinese utterance “lu4 shang5 lai2 le5 ge4 zou3 dao4 de5” in Pinyin (On the street came a traveller).

2.2. TGA: automatic speech timing analysis

TGA, Time Group Analyzer [13] [14], is a tool for the automatic parsing of syllable sequences in speech annotations into Time Groups (TG), that is, inter-pausal groups, or into units based on deceleration models (consistent slowing down) or acceleration models (consistent speeding up). It captures not only the global timing patterns from the input speech annotation in TextGrid format, but also analyzes local patterns based on different duration thresholds (minimal duration difference between adjacent syllables). Most importantly for the present study, it also directly generates ‘Time Trees’ [1] using the local pattern, which are then available for analyzing the relationship between timing properties of the phonetic realizations and the underlying language categories.

TGA is also a web-based tool intended to facilitate the analysis work of phoneticians and linguists. TGA has been applied mainly to Mandarin and Hangzhou Chinese (a dialect, which is very different from Mandarin but shares the same typological structure) and English.

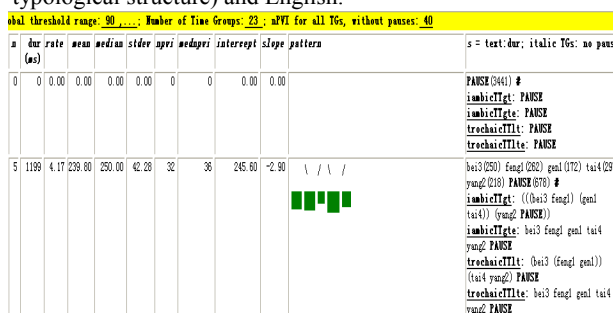


Figure 2: Local patterns of TGs for the Mandarin utterance “bei3 feng1 gen1 tai4 yang2” in Pinyin (the north wind and the sun).

Summary table of accumulated TG values				
Time Group count:	23	Overall syll rate/sec:	5.72	
Overall min:	51.00	Overall max:	342.00	Overall range:
Overall mean:	174.56	Overall median:	171.50	Overall SD:
Overall npvi:	40.00	Overall intercept:	174.91	Overall slope:
Mean of means:	177.48	Median of means:	174.53	SD of means:
Mean of medians:	172.77	Median of medians:	171.75	SD of medians:
Mean of SDs:	57.03	Median of SDs:	54.07	SD of SDs:
Mean nPVI:	42.00	Median nPVI:	38.00	SD nPVI:
Mean intercept:	160.15	Median intercept:	160.79	SD intercept:
Mean slope:	14.50	Median slope:	2.23	SD slope:
nPVI::TGdur:	-0.204	slope::TGdur:	-0.566	intercept::TGdur:
nPVI::mean:	-0.262	slope::mean:	-0.013	intercept::mean:
nPVI::median:	-0.157	slope::median:	-0.343	intercept::median:

Figure 3: Quantative information of TGA output example for the whole Mandarin IPA text “the north wind and the sun”.

The results generated by TGA are very informative, including (1) the corresponding text for each TG; (2) threshold

information, (3) global and local timing patterns; (4) many kinds of quantitative information for each TG and for all TGs, such as nPVI, speech rate, slope etc. An output sample of local timing patterns of TGs and relative quantitative information are illustrated as a Screenshot in Figures 2 and 3.

3. A case study in Hangzhou-accented Mandarin and Standard Mandarin

3.1. Objective: empirical evaluation

The objective is to apply these tools, SSPAS and TGA, together, in order to test whether the tools can facilitate the whole annotating process and whether the combination can generate more satisfactory results than manual annotation and analysis.

3.2. Data

For the study, recordings of six speakers reading the same material, a well-known coherent story (the classic Aesop fable and IPA standard text. ‘The North Wind and the Sun’), were used, in a Mandarin translation. 3 subjects are from Hangzhou and 3 are native Beijing Mandarin speakers. The data is from the CASS corpus [15].

3.3. Procedure

The sound file in WAV format is opened in SPPAS with a transcription of the prompt text, and an IPU (Inter-Pausal Unit) segmented TextGrid file is created. The file is then opened and the segmentation breaks in the input text file are corrected manually, if necessary. Then the prompt text is corrected if for example there are any repeated words or if the speaker does not read exactly the text as it was written. The text is corrected so that it corresponds to what the speaker actually says. Then, the other functions are applied, Phonetization and Alignment, and a merged TextGrid file is generated. Finally the output is checked manually.

In order to quantify the efficiency gained by applying the above procedure, 3 recordings of Hangzhou speakers were transcribed following the procedure described above, and the other 3 recordings of Mandarin speakers were totally manually worked without the aid of SSPAS. The result shows that the procedure using SPPAS reduces annotation time to 33% of the time required for wholly manual annotation.

Once the annotated TextGrid file is ready, timing analysis with the help of TGA is performed. This tool is designed to handle interval tiers. The procedure is: using the web interface, input the TextGrid file, choose the target interval tier, set analysis options. The analysis is performed automatically. There are options for global and local timing patterns.

For the local patterns, values less than common interval lengths can be tried, while for the global patterns, based on deceleration and acceleration models, a wider range of thresholds can be used. The variable threshold is introduced in order to define and traverse search space for possible TGs in terms of minimum differences between syllable durations: different thresholds are relevant for different sizes of TG. A previous study of deceleration and acceleration relations [15] has shown that there are conspicuous steps between certain thresholds, possibly indicating a ‘quantum leap’ between different sizes of linguistic units relating to timing unit sizes.

Finally, local Time Trees can be generated for the TGs, using local quasi-iambic (deceleration) or quasi-trochaic (acceleration) conditions (Figure 4).

An advantage of the TGA tool is that, rather than measuring timing properties of ‘a priori’ linguistic units, such as ‘foot’ or ‘phrase’, or focusing primarily on rhythm, the tool applies an inductive procedure for the automatic parsing of a hierarchy of interval sequences, and then the generalised results can be compared in an independent step with linguistic units. Further quantitative properties of these interval sequences are also available, in particular variation and ‘evenness’ of interval durations in the sequences.

Time Group duration difference parameters:
 TG criterion: ☒ pausegroup ☐ deceleration (increasing) ☐ acceleration (decreasing)

Local threshold: 10 ms (try values less than common syllable lengths, e.g. 0 ... 300 ms)
Used for local pattern extraction and TimeTree parsing

Local pattern symbols: Longer: \ (1 char) Shorter: / (1 char) Same: = (1 char)

Time Tree criterion:
☐ quasi-iambic TT gr ☐ quasi-trochaic TT gr ☐ show all TT
☐ quasi-iambic TT grs ☐ quasi-trochaic TT grs ☐ do not show TT

Global TG threshold range: 90 ... 120 ms (minimal duration difference)
Ranges > 30 are not permitted because of possible server overload.
 Global threshold is ignored with the pausegroup criterion.
 Experiment with values from 0 to 300 (negative values are permitted).
 Equal range boundaries are adjusted to have range of 1, not null, if necessary values are switched to ensure 'low before high'.

Min TG length: > 2 (generally >2, as 'minimal rhythm')

Figure 4: Option interface for TG duration difference parameters in TGA.

In the present study, the pause group condition was used. The global threshold is only required for determining deceleration and acceleration, and is ignored with the pause group criterion. In this study the local pattern and Time Trees are in the focus. Only the quasi-iambic Time Tree model is taken into account because initial inspection showed closer relationships for this model than for the quasi-trochaic model. Relations between Time Tree constituents and multisyllable words were investigated and the percentage of agreement in each TG was calculated. The following example shows a quasi-iambic Time Tree (using brackets, not a graph) of the Mandarin utterance “zhe4 shi1hou5, lu4 shang5 lai2 le5 ge4 zou3 daor4 de5” (at that time, on the street came a traveller), and a grammatical bracketing of the utterance:

quasi-iambic Time Tree:

((zhe4 (shi2 hou5))) (((lu4 shang5) (lai2 (le5 (ge4 zou3)))) daor4)) (de5 PAUSE))

Grammatical bracketing:

((zhe (shi hou)), (lu shang) ((lai) (le) (ge) (zou daor de)))

The groups (shi2 hou5) and (lu4 shang5) correspond to words; (ge4 zou3) is not a grammatical constituent. Also, factoring out the effect of the pause, (lai2 le5 ge4 zou3 daor4 de5) corresponds to a grammatical constituent.

3.4. Results on speech timing

The results illustrated in Figure 5 show:

1. Iambic groups with pause excluded are more meaningful than those with pause included.
2. The graphs in Figure 5 show that the results in the percentage of correlation of the Time Tree components with multisyllable words in each TG for all the speakers are very similar until about 50ms. This is approximately the length of the shortest syllables. It is only when the threshold gets longer that interesting results start to appear. It seems that

HZ-2 and HZ-3 speaker have better Mandarin-like timing, which corresponds to the evaluation results in Mandarin speech proficiency of the 3 Hangzhou speakers.

3. After 50ms, the percentage of correlation of the Time Tree components with multisyllable words increases rapidly and steadily until the increase stops at a certain threshold, varying with different speakers.
4. Quantitative properties show that the difference in speech rate between Hangzhou speakers and Mandarin speakers is not significant ($F(1, 5) = 0.04$, $p > 0.05$) and doesn't correlate with Mandarin speech proficiency ($R^2 = 0.028$, $p > 0.05$). Neither do the nPVIs of syllables durations between the two ($F(1, 5) = 0.444$, $p > 0.05$).

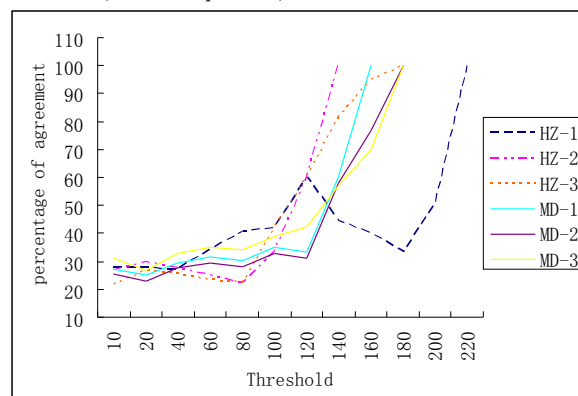


Figure 5: Comparison of local iambic patterns in pause groups between Hangzhou (HZ) and Mandarin (MD) speakers.

4. Conclusion

The study showed that SPPAS, an automatic annotation tool, applied in the procedures described in this paper, can reduce overall transcription time by about 33%. TGA, a tool for automatic timing parsing of interval sequences in speech annotations can be used, just as in the above case study, to investigate the relations between Time Tree constituents and multisyllable words, with comparison of quantitative properties; thus to distinguish native and non-native speech (though the data itself is not large enough). It can also facilitate research into areas such as cross-linguistic phonetic studies, into heuristics for using grammar-speech relations in speech technology, and into the provision of timing criteria for the evaluation of L2 speech proficiency. Both tools together allow phoneticians and linguists to work with larger corpora and to spend more of their time on analysis and less on manual tasks involved in transcription and calculation.

5. References

- [1] Gibbon, D., “Time Types and Time Trees: Prosodic mining and alignment of temporally annotated data”, S. Sudhoff et al., Methods in Empirical Prosody Research, Berlin: Walter de Gruyter. 281-209, 2006.
- [2] Dellwo, V. and Wagner, P., “Relations between language rhythm and speech rate”. In Proc. ICPHS XV. 471-474. Barcelona, 2003.
- [3] Sagisaka, Y., Kato, H., Tsuzaki, M. and Nakamura, S., “Speech timing and cross-linguistic studies towards computational human modeling”, In Proc. Oriental COCOSA 2009, 1-8, Beijing, 2009.

- [4] Bigi B., and Hirst D., "Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody", In Proc. of Speech Prosody 2012, Shanghai, 2012.
- [5] Boersma, P., "Praat, a system for doing phonetics by computer", *Glott International* 5:9/10, 341-345, 2001.
- [6] Barras, C., Geoffrois, E., Wu, Z. and Liberman. M. "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", First International Conference on Language Resources and Evaluation (LREC), 1373-1376, May 1998, and <http://transag.sourceforge.net/>, 2011.
- [7] Sjölander, K., Beskow, J., "Wavesurfer - an open source speech tool", in ICSLP/Interspeech, Beijing, China, October 16-20, 464-467, ISCA, 2000, <http://www.speech.kth.se/wavesurfer/>.
- [8] Kipp, M., "Anvil, DFKI, German Research Center for Artificial Intelligence", <http://www.anvil-software.de/>, 2011.
- [9] Sloetjes, H. and Wittenburg, P., "Annotation by category - ELAN and ISO DCR", LREC 6, 2008, <http://www.latmpi.eu/tools/elan/>.
- [10] Serridge, B., and Castro, L., "Faster time-aligned phonetic transcriptions through partial automation", In Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 189-192, Atenas, 2008.
- [11] Bigi, B., "SPPAS: a tool for the phonetic segmentations of speech", LREC 8, 1748-1754, Istanbul, 2012.
- [12] Yu, J. and Gibbon, D., "Criteria for database and tool design for speech timing analysis with special reference to Mandarin" In Proc. O-COCOSDA 2012, 41-46, Macau, 2012.
- [13] Gibbon, D., "TGA. A tool for automatic speech timing analysis", [Computer Software], Bielefeld: G. Dafydd, Universität Bielefeld, <http://www.homes.uni-bielefeld.de/gibbon/tga-3.0.html>, 2012.
- [14] Gibbon, D. "TGA: a web tool for Time Group Analysis". Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix-en-Provence, forthcoming, August 2013.
- [15] Li A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V. and Chen, X., "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech", in Proc. Interspeech 2000, 485-488, Beijing 2000.

SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora

Juan María Garrido¹

¹Department of Translation and Language Sciences, Pompeu Fabra University,
Roc Boronat 138, 08018 Barcelona, Spain

juanmaria.garrido@upf.edu

Abstract

In this paper we describe SegProso, a Praat-based tool for the automatic segmentation in prosodic units of speech corpora. It is made up of a set of Praat scripts that add several tiers, each one containing the segmentation of a different unit, to a previously existing TextGrid file including the phonetic segmentation of the associated wav file. It has been successfully used for the annotation of several corpora in Spanish and Catalan. The paper briefly describes the workflow of each detector, and presents the results of an evaluation of the performance of the tool in an automatic annotation task on two small Spanish and Catalan corpora.

Index Terms: Prosody, Speech Corpora, Automatic Annotation

1. Introduction

The annotation of prosodic boundaries in speech corpora is a time-consuming task if it is performed by manual means, especially in the case of the annotation of large corpora; and in many cases there can be a strong inter-annotator disagreement in the perceptual identification of some units. Automatic annotation is a promising alternative solution for this problem: even if the obtained output is not perfect, it allows to reduce significantly the time devoted by human experts to this task. In this paper we describe SegProso, a Praat-based tool [1] for the automatic segmentation of speech corpora into prosodic units. It is made up of a set of Praat scripts which add to a previously existing TextGrid file four tiers containing the segmentation into **syllables**, **stress groups (SG)**, **intonation groups (IG)** and **breath groups (BG)**. The tool uses a rule and knowledge-based approach to perform the boundary detection tasks. It was originally designed for the annotation of speech in Spanish and Catalan, but current research is being carried out to adapt the tool to Brazilian Portuguese, and languages could also be added with minimum or none adaptation of the scripts, if the corresponding phonetic transcription was provided. The tool is available for public download at <http://www.upf.edu/pdi/jmgarrido/recerca/projectes/segproso.zip>.

In the following pages, an overview of the tool is given, and the different scripts in charge of the detection of the each type of unit are described. Also, the results of an informal evaluation of the performance of the tool for the automatic annotation of a small speech corpus in Spanish and Catalan are provided.

2. Description of the tool

2.1. General Overview

SegProso is made up of a set of four Praat scripts, each one performing a different segmentation task:

- *Syllable boundaries detector*
- *SG boundaries detector*
- *IG boundaries detector*
- *BG boundaries detector*

These scripts can be run sequentially, to perform a full annotation task, or in isolation, to annotate a single level, if the necessary input for each script is provided: a wav file containing the speech signal of the utterance to be annotated, and a Praat TextGrid file containing the necessary tiers to perform the annotation task.

Full annotation using SegProso can be done by running another script which makes sequential calls to each individual detection script, in the necessary order to ensure that each one will find the necessary input information in the TextGrid file: syllable annotation is performed first; then IG annotation; next is SG annotation; and finally BG annotation.

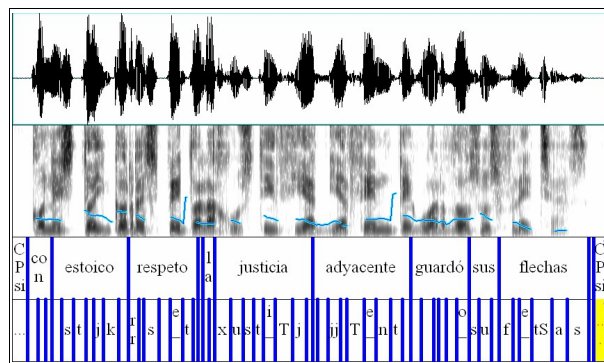


Figure 1: *Speech waveform and TextGrid containing the word segmentation (tier 1) and phonetic segmentation (tier 2) corresponding to the utterance 'con estoico respeto a la justicia adyacente guardó sus flechas', spoken by a male speaker.*

The TextGrid file provided as initial input to SegProso must contain at least two tiers: an interval tier the first one including the orthographic transcription of the utterance word-by-word; and a second interval tier with the phone segmentation in SAMPA format [2]. Figure 1 provides an

example of such an input. Pauses must be also marked and annotated with specific label in both tiers.

The tool provides as output the same input TextGrid file enriched with four new tiers, containing the segmentation of the four prosodic units mentioned above. Figure 2 offers an example of the appearance of such a file.

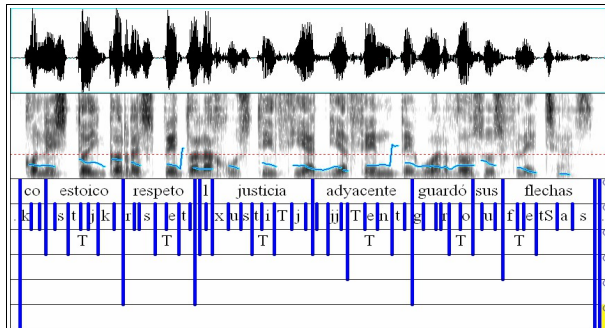


Figure 2: Speech waveform and TextGrid for the same utterance of Figure 1 containing the output tiers of SegProso: syllables (tier 3), SG (tier 4) IG (tier 5) and BG (tier 6).

2.2. Syllable boundaries detector

The syllable detection script creates a new tier in the input TextGrid with the syllable boundaries corresponding to the phone chain, in SAMPA transcription, provided as input in the same TextGrid. It also annotates the intervals corresponding to stressed syllables with a specific label.

To perform this task, the script implements a set of linguistic rules which predict the grouping of phones appearing in the input phonetic transcription tier. The general workflow of these rules is as follows:

- 1) The script first locates word boundaries in the orthographic tier: word boundaries are assumed to be a 'barrier' for phone grouping, so it is carried out separately within each word.
- 2) The script scans then the input phone chain of each word in search of phone symbols representing syllabic nuclei. The procedure in charge of this task basically checks if the input symbol appears in the implemented list of 'nuclear' phones (initially only Spanish and Catalan vowel symbols, recently enlarged with those of Brazilian Portuguese vowels; only vowels are allowed to be syllabic nuclei in those languages).
- 3) Once a nucleus has been detected, the script tries to define the boundaries of the corresponding syllable. To do this, it looks further in the phone chain, detects if there are non-nuclear phone combinations before the next nucleus and if so, tries to establish the placement of the final syllabic boundary by applying the corresponding syllabification rules. As already mentioned, a word boundary is considered always to be a syllable boundary as well; no re-syllabification procedures across words are applied.
- 4) If the nucleus contains a stressed vowel (it as to be transcribed as stressed in the phone tier to be identified), the syllable is labeled as 'stressed' in the corresponding interval of the output tier (label 'T', for 'Tonic').

This approach is different from other existing tools, such as APA [3], in which syllable boundary detection is attempted from the acoustic analysis of the speech signal. In this case, a 'theoretical' grouping of phones in the transcription tier is done based on phonological syllabification rules, not on the detection of acoustic cues for syllable boundaries.

The phone grouping rules implemented in the script are intended to be language-independent, in the sense that they try to represent phonological grouping principles valid at least for several Romance Languages. However, its current implementation is language-dependent, in the way that it uses language-dependent phone inventories in the nuclei detection and syllabification rules. The adaptation of the script to perform syllable segmentation in Brazilian Portuguese involved only the addition of new symbols to the inventory of possible syllable nuclei, with no extra syllabification rules, but probably the adaptation to other languages would not be so direct.

A similar approach has been described recently in the implementation of the syllable annotator of the SPPAS system [4], although in that case syllabification rules seem to be fully language-dependent.

2.3. Intonation group boundaries detector

IG is usually defined as the natural domain of a 'complete' intonation contour. A contour is considered to be complete if it is closed with a final (boundary) F0 pattern, a pause, or both. Other additional phonetic cues, such as declination resets or pre-boundary syllable lengthening, may also indicate the presence of an IG boundary. Some theoretical approaches make a distinction between major (usually ended with a pause) and minor IG (no pause, only boundary pattern at the end). Perceptual identification of IG boundaries is sometimes difficult by non-expert listeners. For this reason, manual segmentation is usually difficult, showing important inter-annotator disagreement.

The script in charge of the identification of IG boundaries in SegProso needs to have in the input TextGrid three tiers containing the word and syllable boundaries, and the phonetic transcription, as well as the corresponding wav file. It tries to detect boundaries at the end of stressed words, which are the candidate places, by looking for two types of F0 cues: the existence of specific F0 boundary patterns, on the one hand, and the existence of declination resets, on the other. The segmentation process is carried out by two sets of rules that look for specific differences between F0 values at specific syllables before and after word boundaries:

- Boundary pattern rules look for specific F0 risings just before word boundaries that could be perceptually interpreted as 'boundary' movements. Basically, these rules compare F0 values in the nucleus of the stressed syllable with F0 values in the last post-stressed syllable (if any), or at the end of the same stressed syllable, if it is the last syllable of the word. If the difference between these two values is beyond a fixed threshold (currently, 5% of the value at the center of the stressed syllable), a boundary is set at the end of the word. Figure 3 shows an example of it.
- F0 reset rules look for F0 jumps between the stressed syllables of two consecutive words. F0 values are taken in the middle of the nuclei of both stressed syllables. If the F0 of the second syllable is found to be significantly higher than the one at the first syllable (currently, at least 5% higher than the value at the center of the first

stressed syllable), an IG boundary is found in the word boundary between the two stressed syllables. Figure 4 shows an example.

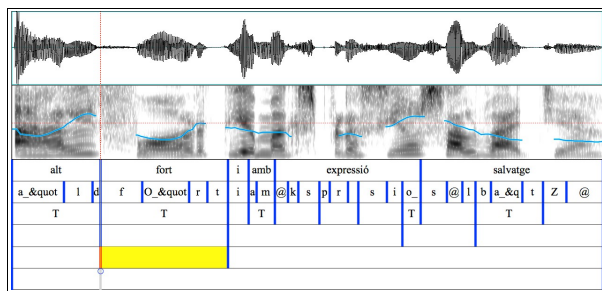


Figure 3: Speech waveform and prosodic segmentation of the Catalan utterance 'Alt, fort, i amb expressió salvatge', spoken by a female speaker. The selected boundary was inserted by the 'boundary pattern rules' of the IG detection script

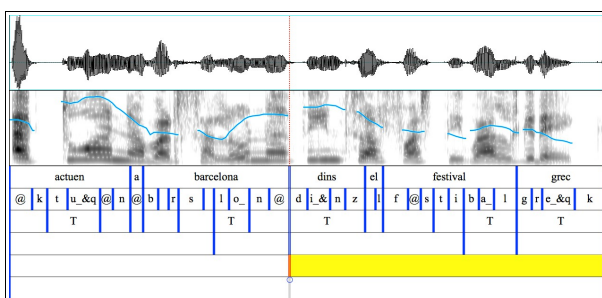


Figure 4: Speech waveform and prosodic segmentation of the Catalan utterance 'Actuen a Barcelona dins el festival grec', spoken by a female speaker. The selected boundary was inserted by the 'F0 reset rules' of the IG detection script

Other segmentation tools, such as APA [3] or ANALOR [5], make use of these F0 cues to detect prosodic breaks similar to IG, sometimes in conjunction with other non-tonal parameters (pauses, energy). However, the detection procedure in SegProso is slightly different, allowing, for example, the identification of F0 resets when no pause is present.

2.4. Stress group boundaries detector

Syllables and IG are two types of prosodic units widely accepted in the Prosodic Phonology literature independently on the theoretical approach. SG, however, is a more theory-dependent prosodic unit, proposed in Garrido [6, 7] and other intonation description frameworks, such as the one by Thorsen [8], for the description of intonation contours. It is defined as a segment of utterance starting at the beginning of a stressed syllable and ending at the beginning of the next stressed syllable, if any, or the end of the container IG. Unstressed syllables appearing at the beginning of an intonation group, before the first stressed one, are considered to be part of the first stress group. Then, for example, the Spanish sentence 'La universidad Pompeu Fabra está en Barcelona' would be segmented in the following SG:

[La universidad Pom] [peu] [Fa bra es] [tá en Barce] [lo na]

The script for the detection of SG needs the syllable and IG segmentation to be already available in the input TextGrid

file: it must be run, consequently, after applying the corresponding scripts for the detection of those units. Its workflow is very straightforward: basically, it looks for stressed syllables in the syllable chain (identified with a 'T' in the syllable tier), and places the beginning of each SG at the time marked for the beginning of the syllable in the syllable tier. If the stressed syllable is the first one in the IG, the initial mark of the SG is placed at the beginning of the IG. As far as final marks are concerned, they are placed at the beginning of next stressed syllable, or the end of the IG if the stressed syllable is the last one in the IG.

2.5. Breath group boundaries detector

BG is the last prosodic unit annotated by SegProso. BG are defined as portions of utterances between two silent pauses. They may include one or several IG, or even none (in interrupted utterances, for example).

BG detection script uses the information about the location of the pauses contained in the syllable tier to create a new tier with the segmentation in BG; it only needs then a syllable tier to be present in the input TextGrid file. Its processing workflow is also quite straightforward: the beginning of a new BG is set in the output tier when the end of a pause interval is detected in the syllable tier, and, accordingly, a BG end boundary is detected when the beginning of a new pause interval is found.

3. Evaluation

Two informal evaluation tests, one for Spanish and one for Catalan, were carried out to assess the performance of the tool. The goal of the evaluation was to check to what extent the tool is able to place correctly prosodic unit boundaries in a small automatic annotation task, assuming that the input (phonetic transcription, phone and pause alignment) is correct.

A set of 100 utterances for each language was selected as evaluation corpus. In the case of Spanish, the evaluation corpus was extracted from the Spanish subset of the INTERFACE corpus [9]: 50 utterances spoken by a male and 50 by a female speaker. For Catalan, the utterances were selected from two corpora recorded at Barcelona Media by two different female professional speakers for synthesis purposes [10, 11]. The utterances were rather short, and uttered with a neutral style. TextGrid files obtained automatically using an HMM segmentation tool were available for each file of the corpus. These utterances were processed using SegProso to obtain their automatic prosodic segmentation. The output was then manually checked, and compared with the automatic version.

The results of the evaluation show an excellent performance of the syllable and BG scripts for both languages: 100% of correct segmentations. This percentage does not include, however, errors in BG segmentation related to some wrong detection of pauses by the automatic segmentation tool. For SG and IG, the obtained rates are lower, although still high. In the case of SG tiers, the percentage of correct boundaries is very similar: 87.32% for Spanish and 86.46% for Catalan. It is important to notice that all detected errors were directly related to previous wrong boundary placements in the IG tier. The performance of the script is perfect when the IG boundaries are correctly detected. The script for IG identification is the one which obtained the poorest results: 82.46% of the boundaries inserted by the tool in the Spanish corpus were labelled as correct during the evaluation process,

and 77.40% in the case of the Catalan corpus. Some of the wrong boundaries inserted by the tool were moved to another boundary (21 cases, 6.81% of the automatic boundaries, in Spanish; 31 cases, 8.75%, in Catalan) and the rest was deleted (33 cases, 10.71% of the automatic boundaries, in Spanish; 49 cases, 13.84%, in Catalan). During the evaluation process, some boundaries not detected by the tool had to be added manually (22, 7.4% of the correct boundaries, in Spanish; 18, 5.57%, in Catalan).

4. Applications

SegProso has been successfully used for the automatic prosodic annotation of several corpora in Spanish and Catalan, such as Interface, I3Media or Glissando [12]. In all three cases, the obtained segmentation has been used as input for MelAn, the automatic F0 annotation and modeling tool described in [13]. A full inventory of the F0 patterns appearing in those corpora was successfully obtained using this tool. The prosodic segmentation provided by SegProso has also been used for the development of intonation models for TTS [14].

5. Conclusions and future work

SegProso has shown to be a useful tool for fast annotation of prosodic boundaries of large speech corpora in Spanish and Catalan. Although the performance of the IG detection script could be still improved, our experience has revealed that manual revision of the obtained output is much faster than manual annotation.

Future work will focus on the improvement of the IG annotation script: fine tune of the rules is expected to be done using the data of a detailed acoustic analysis of the F0 patterns appearing at those boundaries.

6. References

- [1] Boersma, P. and Weenink, W., Praat: doing phonetics by computer [Computer program] <http://www.praat.org/>, 2012.
- [2] Wells, J. C., "SAMPA computer readable phonetic alphabet", in D. Gibbon, R. Moore, R. Winski, [Eds.], Handbook of Standards and Resources for Spoken Language Systems, Part IV, section B, Mouton de Gruyter, Berlin and New York, 1997.
- [3] Cutugno, E., D'Anna, L., Petrillo, M. and Zovato, E., "APA: towards an automatic tool for prosodic analysis", Speech Prosody 2002, 231-234, 2002.
- [4] Bigi, B., "SPPAS: a tool for the phonetic segmentation of speech", LREC 2012 Proceedings, 1748-1755, 2012.
- [5] Avanzi, M., Lacheret-Dujour, A. and Victorri, B., "Analog: A tool for semi-automatic annotation of french prosodic structure", Speech Prosody 2008, 119-122, 2008.
- [6] Garrido, J. M., Modelling Spanish Intonation for Text-to-Speech Applications, Ph. D Thesis, Universitat Autònoma de Barcelona, 1996. Online: <http://www.tdx.cat/handle/10803/4885;jsessionid=376A9A0BED1D5E6DED7CDFS3880316F3.tdx1>, accessed on 24 Apr 2013..
- [7] Garrido, J. M., "La estructura de las curvas melódicas del español: propuesta de modelización", Lingüística Española Actual, XXIII/2, 173-209, 2001.
- [8] Thorsen, N., "An acoustical investigation of Danish intonation", ARIPUC, 10, 85-147, 1976.
- [9] Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A. and Nogueiras, A., "Interface databases: Design and collection of a multilingual emotional speech database", Proceedings of LREC'02, 2024-2028, 2002.
- [10] Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M. and Pidcock, Ch., "The Cerevoice speech synthesiser", Actas de las V Jornadas de Tecnología del Habla, 126-129, 2008.
- [11] Garrido, J. M., Laplaza, Y., Marquina, M., Pearman, A., Escalada, J. G., Rodríguez, M. A. and Armenta, A., "The I3Media speech database: a trilingual annotated corpus for the analysis and synthesis of emotional speech", LREC 2012 Proceedings, 1197-1202, 2012.
- [12] Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la-Mota, C., González, C., Rustullet, S., Larrea O., Laplaza, Y., Vizcaino, F., Cabrera, M. and Bonafonte, A., "Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan", Language Resources and Evaluation, DOI 10.1007/s10579-012-9213, 2013. Online: <http://link.springer.com/article/10.1007/s10579-012-9213-0>, accessed on 24 Apr 2013.
- [13] Garrido, J. M., "A Tool for Automatic F0 Stylisation, Annotation and Modelling of Large Corpora", Speech Prosody 2010: 100041. Online: <http://speechprosody2010.illinois.edu/papers/100041.pdf>, accessed on 24 Apr 2013.
- [14] Garrido, J. M., "GenProso: a parametric prosody prediction module for text-to-speech applications", IberSpeech 2012 Proceedings.

Continuous wavelet transform for analysis of speech prosody

Martti Vainio, Antti Suni, and Daniel Aalto

Institute of Behavioural Sciences (SigMe Group), University of Helsinki, Finland

`martti.vainio@helsinki.fi`, `antti.suni@helsinki.fi`, `daniel.aalto@helsinki.fi`

Abstract

Wavelet based time frequency representations of various signals are shown to reliably represent perceptually relevant patterns at various spatial and temporal scales in a noise robust way. Here we present a wavelet based visualization and analysis tool for prosodic patterns, in particular intonation. The suitability of the method is assessed by comparing its predictions for word prominences against manual labels in a corpus of 900 sentences. In addition, the method's potential for visualization is demonstrated by a few example sentences which are compared to more traditional visualization methods. Finally, some further applications are suggested and the limitations of the method are discussed.

Index Terms: continuous wavelet transform; speech prosody; intonation analysis; prominence

1. Introduction

The assumption that prosody is hierarchical is shared by phonologists and phoneticians alike. There are several accounts for hierarchical structure with respect to speech melody: In the tone sequence models which interpret the f_0 contour as a sequence of tonal landmarks of peaks and valleys (e.g. [15]) the hierarchy is mainly revealed at the edges or boundaries of units whereas in superpositional accounts (e.g., [13, 6]) it is seen as a superposition of different levels at each point of the contour. The problem with the tone sequence models stems from their phonological nature which requires a somewhat discretized view of the continuous phonetic phenomena. The superpositional accounts suffer, conversely, from the lack of signal based categories that would constrain the analysis in a meaningful way. Both models suffer from being disjointed from perception and require *a priori* assumptions about the utterances.

Wavelets emerged independently in physics, mathematics, and engineering, and are currently a widely used modern tool for analysis of complex signals including electrophysiological, visual, and acoustic signals [5]. In particular, the wavelets have found applications in several speech prosody related areas: The first steps of the signal processing by the auditory periphery are well described by models that rely on wavelets [23, 22, 17]; they are used in a robust speech enhancement in noisy signals with unknown or varying signal to noise ratio, in automatic speech segmentation, and in segregation along various dimensions of speech signal in a similar way as mel-cepstral coefficients [2, 1, 8, 9]; the multiscale structure of the wavelet transform has been taken advantage of in musical beat tracking [19]. The quantitative analysis of speech patterns through wavelets might also be relevant for understanding the cortical processing of speech (e.g. [3, 14, 7]).

In the present paper, we apply the wavelet methods to

recorded speech signals in order to extract prosodically important information automatically. Here, only the fundamental frequency of the speech signal is analyzed by wavelets although similar analysis could be performed to any prosodically relevant parameter contour (e.g., the intensity envelope contour or a speech rate contour) or even the raw speech signal itself.

The analysis of intonation by wavelets is not a new idea. Discrete wavelet analysis with Daubechies mother wavelets was the key component in automatically detecting the correct phrasal components of synthesized f_0 contours of the Fujisaki model further developed under the name general superpositional model for intonation proposed by van Santen et al. [21, 12]. Continuous wavelet transforms with Mexican hat mother wavelet have been used for Fujisaki accent command detection by Kruschke and Lenz [10]. Overall, previous work with wavelets and f_0 have been mainly concerned with utilizing wavelets as a part of model development or signal processing algorithm, instead of using the wavelet presentation itself.

In Finnish, the prosodic word is an important hierarchical level and the prominence at that level reveals much of the syntactically and semantically determined relations within the utterances. We have successfully used a four level word prominence in text-to-speech synthesis in both Finnish and English [20] and the automatic detection of word prominence is a prerequisite for building high quality speech synthesis. In relation to both a tone sequence and superpositional accounts the successful detection of word prominence would be related to distinguishing the accentedness of the unit as well as the magnitude of the accent.

Using an inherently hierarchical analysis we can do away with a fixed model and try to directly link acoustical features of an utterance to the perceived prominences within the utterance. In order to evaluate the wavelet analysis we calculated CTW based prominences for about 7600 separate words in 900 utterances previously annotated by human labelers and compared various wavelet and f_0 based features with each other. In this paper we first discuss the CWT and its application to f_0 and then show the quantitative evaluation followed by discussion and conclusion.

2. Continuous wavelet transform

The continuous wavelet transform (CWT) can be constructed for any one-dimensional or multidimensional signal of finite energy. In addition to the dimensions of the original signal, CWT has an additional dimension, scale, which describes the internal structure of the signal. This additional dimension is obtained by convolving the signal by a mother wavelet which is dilated to cover different frequency regions [5]. The CWT is similar to the windowed Fourier transform: the CWT describes the time-

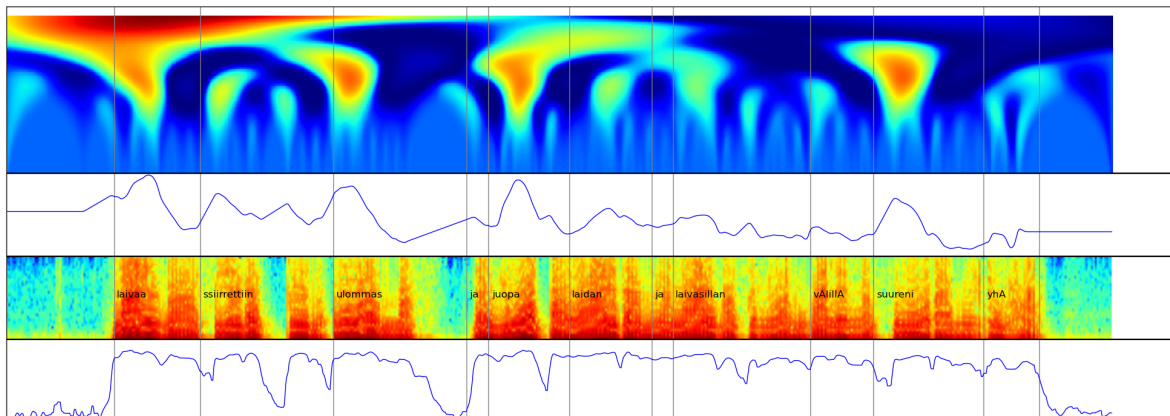


Figure 1: Different analyses aligned temporally. Top pane depicts the continuous wavelet transform with Mexican hat mother wavelet of f_0 , second pane shows the interpolated f_0 contour; third pane shows spectrogram of the speech signal; the bottom pane shows gain. The light gray vertical lines show the word boundaries. The text superposed to the third pane transcribes the uttered words (The ship was moved outwards and the gap between the board of the ship and the gangplank got wider, still.)

frequency behaviour of the signal and the signal can be reconstructed from the CWT by inverse wavelet transform. We use here a Mexican hat shaped mother wavelet which corresponds formally to the second derivative of the Gaussian, see pages 76–78 in [11]. In the Figure 1, the top pane shows the CWT of the f_0 contour shown in the second pane. The peaks in f_0 curve show up in the CWT as well, but the size of the peaks in the wavelet picture depends on the local context: the higher at the picture, or in other words, the coarser the scale, the slower the temporal variations and the larger the temporal integration window. Although several hierarchical levels emerge, the quantitative evaluation of the suitability of the CWT to prosodic analysis is only performed on word level. Note that in Finnish, content words have a fixed stress on the first syllable, clearly visible in the Figure 1. The third and fourth panes show the spectrogram and the intensity envelope of the same utterance. The time scales in the wavelet picture range from the 67 Hz as finest to less than 1 Hz as coarsest.

3. Quantitative evaluation

A visualization tool cannot be evaluated quantitatively as a whole. However, if the different temporal scales reflect perceptually relevant levels of prosodic hierarchy, the representation of f_0 at any scale should correlate with judgements of the relative prominence at that particular level. This hypothesis is tested at the level of prosodic word. Although word prominence is signaled by f_0 , it is, to large extent, signaled by other means as well including intensity, duration, word order, and morphological marking. Hence, the f_0 based prominence annotation is compared to a simple baseline f_0 prominence annotator and to the labels obtained from phonetically trained listeners.

3.1. Recorded speech data

The evaluation data consisted of 900 read sentences by a phonetically trained, native female speaker of Finnish. Linguisti-

cally, the sentences represented three different styles: modern standard scientific Finnish, standard Finnish prose, and phonetically rich sentences covering the Finnish phonemes. The sentences were recorded using high quality condenser microphone in a sound proof studio, digitized, and stored on a computer hard drive. The mean durations of the sentences had average durations of 6.1 s, 3.5 s, and 3.8 s. The total duration amounted to 1h 1 min. Acoustic features were extracted of the utterances with GlottHMM [16], and then the utterances were aligned with the text.

3.2. Fundamental frequency extraction

The fundamental frequency of the test utterances were extracted by GlottHMM speech analysis and synthesis software. In GlottHMM analysis, the signal is first separated to vocal tract and glottal source components using inverse filtering, and the f_0 is then extracted from the differentiated glottal signal using autocorrelation method. Parameters concerning voicing threshold and admissible range of f_0 values were tuned manually for the current speaker. While GlottHMM performs some post-processing on analyzed f_0 trajectories, deviations from perceived pitch remain, particularly in passages containing creaky voice. Thus, f_0 values were first transformed to logarithm scale and then all values lower than 2 standard deviations below the mean of log f_0 were removed.

The unvoiced segments of the speech and the silent intervals make the direct wavelet analysis impossible since f_0 is not well defined for these segments. Hence, the unvoiced gaps were filled using linear interpolation. Additionally, to alleviate edge artifacts, the continuous f_0 contour was extended over the silent beginning and end intervals by replacing the former by the mean f_0 value (logarithmically scaled) over the first half of the completed f_0 contour, and the latter by the mean over the second half. Then the f_0 curve was filtered by a moving average Hamming window of length 25 ms and finally normalized to zero mean and unity variance.

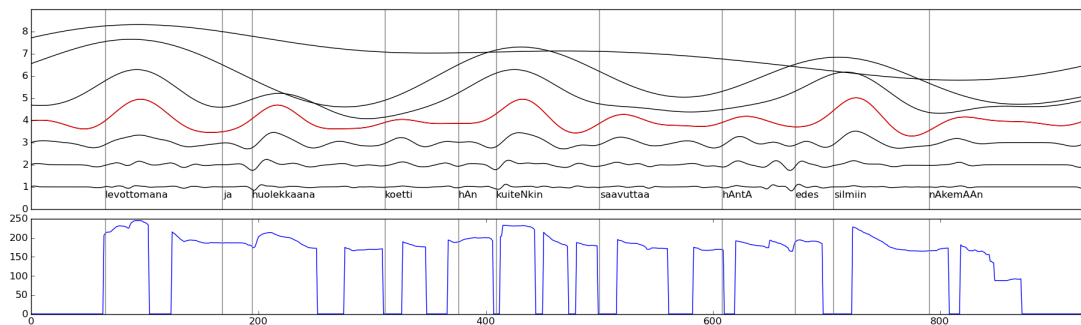


Figure 2: The word prosody scale is chosen from a discrete set of scales with ratio 2 between ascending scales as the one with the number of local maxima as close to the number of words in the corpus as possible. The upper pane shows the representations of f_0 at different scales. The word level (4.2 Hz; see text) is drawn in red. The lower pane shows the f_0 curve. The abscissa shows the frame count from the beginning of the utterance (5 ms frame duration).

3.3. Baseline annotation based on f_0 signal

For each word in the evaluation data, we extracted two common measurements from the preprocessed and normalized f_0 signal, the maximum value observed during word (BMax) and the maximum minus minimum (BRange). The measurements were not further processed, despite the scale differences compared to manual annotation, as only correlation was being tested.

3.4. CWT annotation based on f_0 signal

The CWT transform was first performed with one scale per octave, with finest scale being 3 frames or 15 ms. Then, the scale of interest for word prominence was selected as the one with positive peak count closest to the number of words (see Figure 2; the word scale corresponds to 4.2 Hz in the current data). This is intuitively suitable for Finnish, with relatively few unaccented function words. Three wavelet based measurements were then extracted for each word, height of the first local maximum (WPeak) as well as the same two measurements as in f_0 baseline (WMax, WRange). If the word contained no maxima, then the prominence of the word was set to zero. Note that the peak method is not applicable to raw F0, as the noisier contour contains many peaks. More complex measurements were experimented with, such as averaging over multiple scales, but with only moderate success.

3.5. Prominence labeling

Ten phonetically trained listeners participated in prominence labeling. The listeners were instructed to judge the prominence of each word in a categorical scale: 0 (unaccented, reduced); 1 (perceivably accented but no emphasis); 2 (accented with emphasis); 3 (contrastive accent). The listeners reported to have based their judgements mainly on listening and secondarily to the available Praat analyses of pitch, intensity, and spectrogram. Every listener labeled 270 sentences in such a way that every sentence was labeled by three listeners. The prominence of a word was set to the average of the three judgements.

3.6. Statistical analysis

The two baseline annotations and the three wavelet based annotations were compared to the listeners' judgements of word prominence by linear regression analysis. The amount of variance explained (R squared) by the regression model was used as an indicator for the goodness of the used measure.

3.7. Results

The baseline measure *BMax* has a strong correlation to the prominence judgements with 37 % of the variance explained. The other baseline measure *BRange* explained 36 % of the variance. The wavelet based measures fitted better to the data: *WMax* and *WRange* explained 47 % and 39 % of the variance, respectively. The more involved measures *WPeak* explained 53 % of the variance.

4. Discussion

The results of the evaluation show that it is fairly straightforward to extract prosodically relevant information from the CWT analysis. In this case it was at the level of prosodic word (which in Finnish corresponds well with the grammatical word). As can be seen in Figures 1 and 2, there are other levels both above and below the word that are relevant and if discretized, form a hierarchical tree which can be further exploited for instance in text-to-speech synthesis. However, such an analysis is not free of problems. For instance, the temporal scale corresponding to syllables becomes coarser (higher levels in the Figure 1) when the speech slows down, as is the case in e.g. pre-pausally.

What is important to notice here is that the CWT analysis – as applied to the pitch contour – takes into account both the f_0 level and its temporal properties as cues for prominence. Although we only used one level it is the analysis as a whole that we are interested in. As mentioned earlier, the wavelet analysis can be done on any prosodically relevant signal either alone or jointly – although multidimensional may no longer be easily visualizable.

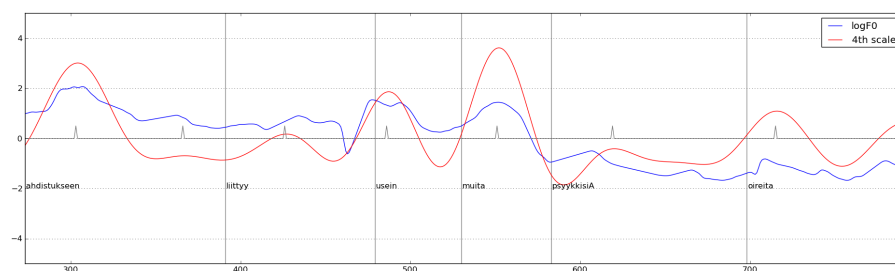


Figure 3: Comparison of selected word scale and original f_0 contour with detected peaks marked with gray triangles. Observe that the wavelet contour is free of noise and declination trend.

5. Conclusion

Continuous wavelet transform, a standard mathematical tool for simultaneous analysis and visualization of various temporal scales of a signal, is applied to f_0 signal of recorded speech. At the temporal scale corresponding to prosodic word, the local maxima correlate strongly with the listeners' judgements on the perceived word prominence. This is taken as evidence that the small and large scale contributions induced by segmental micro-prosody and phrasal intonation components are effectively removed by the analysis. Moreover, a hierarchical structure emerges which is easily visible and has similarities with the classical description of prosodic structure through a prosodic tree. Unlike other hierarchical models of prosody, the structure rises directly from the signal with no assumptions on the f_0 model.

Some interesting future directions could include building a 'spectrogram of prosody' -visualization tool combining spectrogram and prosody in the same picture, attempting to discretize the hierarchical structure for higher level applications, applying the decomposed prosodic features for TTS prosody models, studying other prosodic features such as energy by CWT, and, finally, exploring the relationship between the CWT analyses and human auditory processing.

6. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 287678 and the Academy of Finland (projects 135003 LASTU programme, 1128204, 128204, 125940). We would also like to thank Heini Kallio for collecting the prominence data.

7. References

- [1] Alani, A. and Deriche, M., "A novel approach to speech segmentation using the wavelet transform", 5th Int. Symposium on Signal Processing and its Applications, Brisbane, 1999.
- [2] Bahoura, M., "Wavelet speech enhancement based on the Teager energy operator", IEEE Signal Processing Letters, 8(1):10–12, 2001.
- [3] Bradley, A. P. and Wilson, W. J., "On wavelet analysis of auditory evoked potentials", Clinical neurophysiology, 115:1114–1128, 2004.
- [4] Chi, T., Ru, P., and Shamma, S. A., "Multiresolution spectrotemporal analysis of complex sounds", J. Acoust. Soc. Am. 118(2):887–906, 2005.
- [5] Daubechies, I., "Ten lectures on wavelets", Philadelphia, SIAM, 1992.
- [6] Fujisaki, H., Hirose, K., Halle, P., and Lei, H., "A generative model for the prosody of connected speech in Japanese", Ann. Rep. Eng. Reserach Institute 30: 75–80, 1971.
- [7] Giraud, A. and Poeppel, D., "Cortical oscillations and speech processing: emerging computational principles and operations", Nature Neuroscience, 15:511–517, 2012.
- [8] Hu, G. and Wang, D., "Segregation of unvoiced speech from non-speech interference", J. Acoust. Soc. Am., 124(2): 1306–1319, 2008.
- [9] Irino, T. and Patterson, R. D., "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform", Speech Communication, 36:181–203, 2002.
- [10] Kruschke, H. and Lenz, M., "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis", in Proc. Eurospeech'03, 4, pp. 2881–2884, Geneva, 2003.
- [11] Mallat, S., "A wavelet tour of signal processing", Academic Press, San Diego, 1998.
- [12] Mishra, T., van Santen, J., and Klabbers, E., "Decomposition of pitch curves in the general superpositional intonation model",
- [13] Öhman, S., "Word and sentence intonation: a quantitative model", STLQ progress status report, 2–3:20–54, 1967.
- [14] Petkov, C. I., O'Connor, K. N., and Sutter, M., L., "Encoding of illusory continuity in primary auditory cortex", Neuron, 54: 153–165, 2007.
- [15] Pierrehumbert, J., "The phonology and phonetics of English intonation", PhD Thesis, MIT, 1980.
- [16] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.
- [17] Reimann, H. M., "Signal processing in the cochlea: the structure equations", J. Mathematical Neuroscience, 1(5):1–50, 2011.
- [19] Smith, L. M. and Honing, H., "Time-Frequency representation of musical rhythm by continuous wavelets", J. Mathematics and Music, 2(2):81–97, 2008.
- [20] Suni, A., Raitio, T., Vainio, M., and Alku, P., "The GlottHMM entry for Blizzard Challenge 2012 – hybrid approach", in Blizzard Challenge 2012 Workshop, Portland, Oregon, 2012.
- [21] van Santen, J. P. H., Mishra, T., and Klabbers, E., "Estimating phrase curves in the general superpositional intonation model", Proc. 5th ISCA speech synthesis workshop, Pittsburgh, 2004.
- [22] Yang, X., Wang, K., and Shamma, S., "Auditory representation of acoustic signals", IEEE Trans. Information theory, 38:824–839, 1992.
- [23] Zweig, G., "Basilar membrane motion", Cold Spring Harbor Symposia on Quantitative Biology, 40:619–633, 1976.

Modeling Speech Melody as Communicative Functions with PENTAtainer2

Santitham Prom-on^{1,2}, Yi Xu²

¹Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand

²Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom

santitham@cpe.kmutt.ac.th, yi.xu@ucl.ac.uk

Abstract

This paper presents PENTAtainer2, a semi-automatic software package written as Praat plug-in integrated with Java programs, and its applications for analysis and synthesis of speech melody as communicative functions. Its core concepts are based on the Parallel Encoding and Target Approximation (PENTA) framework, the quantitative Target Approximation (qTA) model, and the simulated annealing optimization. This integration allows it to globally optimize for underlying pitch targets of specified communicative functions. PENTAtainer2 consists of three computational tools: Annotation tool for defining communicative functions as parallel layers, Learning tool for globally optimizing pitch target parameters, and Synthesis tool for generating speech melody according to the learned pitch targets. Being both theory-based and trainable, PENTAtainer2 can serve as an effective tool for basic research in speech prosody.

Index Terms: prosody modeling, parallel encoding, target approximation, communicative function, stochastic optimization

1. Introduction

Speech prosody conveys communicative meanings through the manipulation of fundamental frequency (F_0), duration, intensity, and voice quality. Of these cues, F_0 is one of the most important. Communicative function refers to the relation between a specific communicative meaning and how it is encoded in the structure of speech melody. Modeling communicative function is thus a key to achieve effective prosody analysis and synthesis.

This paper presents PENTAtainer2, a tool for prosody analysis and synthesis. It was created with an ultimate goal to assist speech researchers in prosody modeling studies. It provides users easy-to-use interfaces to perform three critical tasks in prosody modeling: data annotation, parameter estimation, and prosodic prediction. The program and the step-by-step tutorials in prosody analysis and synthesis can be freely downloaded from (<http://www.phon.ucl.ac.uk/home/yi/PENTAtainer2/>).

2. PENTAtainer2

PENTAtainer2 (pen-ta-train-ner-two) consists of a set of Praat scripts that facilitate the investigation of underlying representations of communicative functions in any language [1]. Its core concept is based on the Parallel Encoding and Target Approximation (PENTA) framework [2]. PENTAtainer2 encapsulates the quantitative Target Approximation (qTA) model, which represents dynamic F_0 control [3], and simulated annealing optimization [4], which is a stochastic learning algorithm used to globally optimize model parameters. Provided with annotated

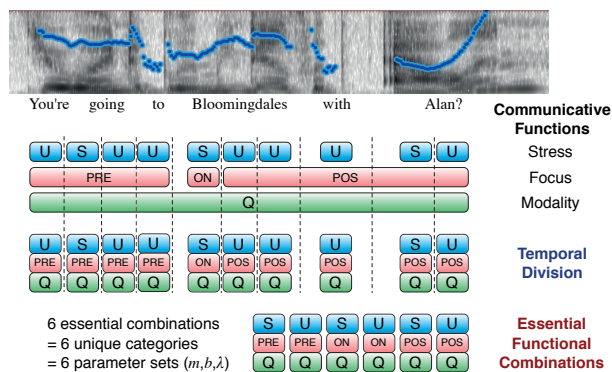


Figure 1: An illustration of the conversion from the parallel functional annotation to the essential functional combinations.

sound files, PENTAtainer2 automatically learns the optimal parameters of all possible functional combinations that users have annotated. After the optimization, the learned functional parameters can be used to synthesize F_0 contours according to any of the given communicative functions. Summaries of the modeling technique will be briefly discussed in the following sub-section.

2.1. Parallel annotation of communicative functions

PENTAtainer2 is a data-driven prosody modeling software. The specific values of the model parameters are optimized from the training speech material. The basic idea is to identify the number of functional layers and their corresponding prosodic categories that span across specified temporal units (e.g. tone localized with the syllable). It is critical for the system to know what to learn. Fig. 1 illustrates the annotation of three communicative functions of English intonation: Stress, Focus, and Modality. Each layer was annotated independently and the function-internal categories are defined manually by the investigator. Boundaries on each layer were marked according to the time span of that prosodic event, again defined by the investigator. For example, in Fig. 1, the "Stress" layer is associated with the syllable and can have two values: Stressed (S) and Unstressed (U). For a "Focus" layer, PRE, ON, POS denote pre-focus, on-focus, and post-focus regions respectively. For a "Modality" layer, Q denotes question and S denotes statement. Note that the names here carry no meaning to PENTAtainer2, as all it cares is which are the same categories and so should be given a common set of target parameters. This differs from annotation schemes in which the names are meaningful (e.g., ToBI [5], INTSINT [6]).

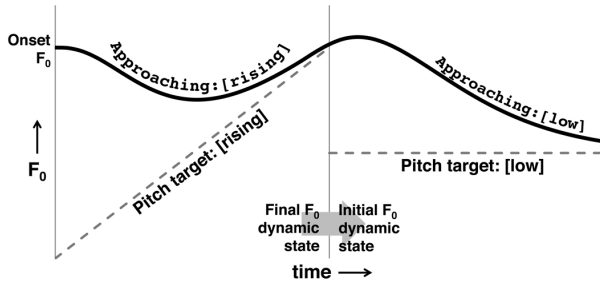


Figure 2: Illustration of target approximation process [3, 7].

2.2. Modeling F_0 movement with qTA model

To model F_0 movement, PENTAtainer2 uses the quantitative Target Approximation (qTA) Model [3], which is based on the theoretical target approximation model [7]. Fig. 2 is an illustration of the basic concept of target approximation. Surface F_0 contours (solid curve) are the responses of the target approximation process to the driving force of pitch targets (dashed lines). These targets represent the goals of the F_0 movement and are synchronized to the host syllable. Pitch targets are sequentially implemented syllable by syllable, starting from the beginning of the utterance. At the boundary of two syllables, the F_0 dynamic state at the end of the preceding syllable is transferred to the next syllable.

In qTA, a pitch target is defined as a forcing function that drives the F_0 movement. It is mathematically represented by a simple linear equation,

$$x(t) = mt + b \quad (1)$$

where m and b denote the slope and height of the pitch target, respectively. t is a relative time from the syllable onset. The F_0 control is implemented by a third-order critically damped linear system, in which the total response is

$$f_0(t) = x(t) + (c_1 + c_2t + c_3t^2)e^{-\lambda t} \quad (2)$$

where the first term $x(t)$ is the forced response of the system which is the pitch target and the second term is the natural response of the system. The transient coefficients c_1 , c_2 and c_3 are calculated based on the initial F_0 dynamic state and pitch target of the specified segment. The parameter λ represents the strength of the target approximation movement. The initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f'_0(0)$, and acceleration, $f''_0(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0 . The three transient coefficients are computed with the following formulae.

$$c_1 = f_0(0) - b \quad (3)$$

$$c_2 = f'_0(0) + c_1\lambda - m \quad (4)$$

$$c_3 = (f''_0(0) + 2c_2\lambda - c_1\lambda)/2 \quad (5)$$

qTA thus defines each pitch target with only three parameters, m , b , and λ . Of the three parameters, m and b are used to specify the form of the pitch target. For example, the Mandarin Rising and Falling tones, which differ mainly in target slope, have positive and negative m values, respectively; the Mandarin High and Low tones, which differ mainly in target height, have

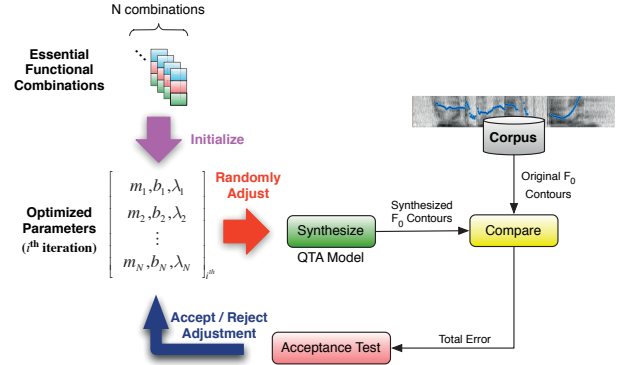


Figure 3: A diagram illustrating the application of the simulated annealing algorithm used for globally optimizing parameters of essential functional combinations.

relatively high and low b values, respectively [3, 8, 9]. Here the value of b is relative to a reference pitch, which can be either the speaker F_0 mean or the initial F_0 of an utterance. λ specifies how rapidly the target is approached, with a larger value indicating faster approximation. This approximation rate can define an additional property of a tone. For example, the Mandarin neutral tone is found to have a much smaller λ value than the full tones [9], which is consistent with the observation that the neutral tone may have a weak articulatory strength [10].

2.3. Parameter optimization

In PENTAtainer2, the parameter estimation is done via a stochastic global optimization that can directly estimate parameters of functional categories in a corpus. The general idea is illustrated as a block diagram in Fig. 3. At the initial stage, the algorithm randomly modifies parameters of all functional categories and tests whether or not such modification is acceptable by a probabilistic method. The number of initialized parameter sets is equal to the number of essential functional combinations obtained from the procedure to be discussed in the next section. These parameters are randomly adjusted and used in qTA to synthesize F_0 contours which are compared to the original data. The total sum of square error between original and synthesized F_0 contours calculated from the whole corpus is then used to determine whether the proposed adjustment is acceptable. The decision to accept or reject the proposed adjustment depends on the acceptance probability calculated from the change in error incurred from parameter adjustment and the annealing temperature,

$$p_{th} = \exp(-(E_{current} - E_{previous})/T) \quad (6)$$

where $E_{current}$ and $E_{previous}$ are the total sum of square errors calculated from the whole corpus. The difference between these two errors indicates the change in the total error incurred from the parameter adjustment. T is the annealing temperature which controls the degree at which a bad solution is allowed. In the decision process, a random testing probability p_{test} is generated and compared to p_{th} . If $p_{test} < p_{accept}$, the parameter adjustment is accepted; otherwise it is rejected. T is initially set to a high value and then gradually reduced as the procedure is repeated. This allows the solution to converge close to the global optimum over iterations.

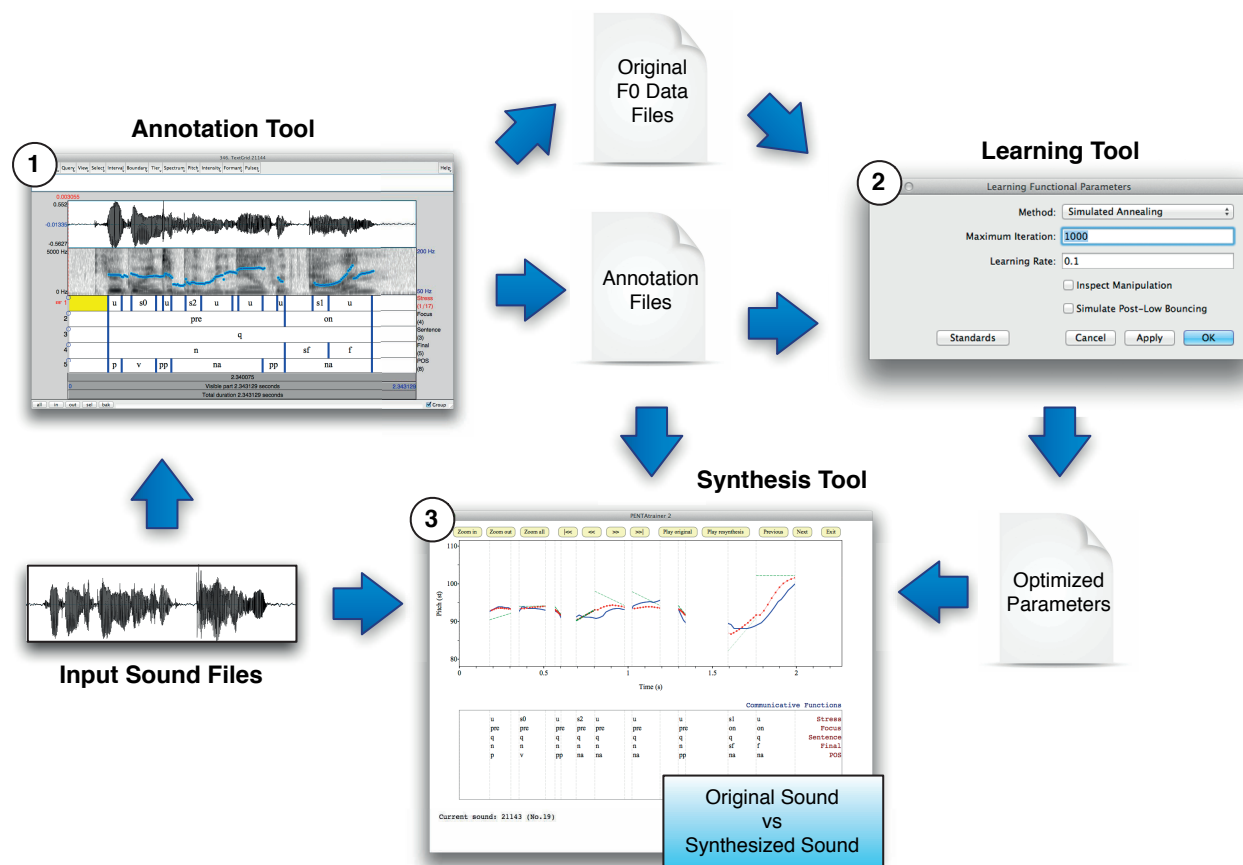


Figure 4: A workflow of analysis and synthesis of speech melody using PENTAtainer2.

2.4. Prosody modeling workflow

PENTAtainer2 composes of three main tools: Annotate, Learn, and Synthesize. Each of them can be accessed individually in the PENTAtainer2 plug-in menu. Each tool corresponds to a main task in the prosody modeling workflow shown in Fig. 4. First, the speech corpus is annotated using the Annotate tool. Communicative functions related to a corpus are annotated in separate tiers. Two tiers, tone and vowel length, are annotated for the Thai corpus in this project. Temporal boundaries in each tier are aligned consistently to the prosodic or segmental events of that tier. For the present study, because both tone and vowel length boundaries are synchronized to the syllable, the syllable boundaries are used as temporal markings for both tiers. The co-occurrences of events in the two tiers form functional combinations, which represent interactions between tiers. The annotation step is done iteratively for each sound file in the corpus. In this step, investigators can also inspect and manually rectify the vocal pulse marks used in F0 calculation. This step requires the most human effort. Next, after all the sound files are annotated, the second step is to estimate the pitch target parameters using the Learn tool. Investigators only need to provide the Learn tool with the optimization parameters, and it will then automatically estimate the optimal parameters of the functional combinations. The third and the last step allows investigators to synthesize or predict the F0 contours from the learned parameters (or humanly provided parameters if so desired) using

the Synthesize tool. The optimized parameters can be either speaker-dependent, i.e., learned from each individual speaker, or speaker-independent, i.e., derived by averaging the parameters of all the speakers.

3. Conclusion

This paper presents the technical detail of PENTAtainer2, and its workflow for prosody modeling. It provides analysis and synthesis functionalities to represent speech prosody as communicative functions. It has been found to be effective in capturing underlying representation of communicative functions in several languages and able to synthesize with high accuracy [1]. Being both theory-based and trainable, PENTAtainer2 can serve as a modeling tool for basic research in speech prosody.

4. Acknowledgement

The authors would like to thank for the financial supports the Royal Society (UK) and the Royal Academy of Engineering (UK) through the Newton International Fellowship Scheme, the Thai Research Fund through the TRF Grant for New Researcher, and the National Science Foundation.

5. References

- [1] Xu, Y. and Prom-on, S., "From variable surface contours to invariant underlying representations: Synthesizing speech melody via model-based stochastic learning", Manuscript submitted for publication, 2013.
- [2] Xu, Y., Speech melody as articulatory implemented communicative functions, *Speech Commun.*, 46(3-4): 220-251, 2005.
- [3] Prom-on, S., Xu, Y., and Thipakorn, B. Modeling tone and intonation in Mandarin and English as a process of target approximation, *J. Acoust. Soc. Am.*, 125: 405-424, 2009.
- [4] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P., Optimization by simulated annealing, *Science*, 220(4598): 671-680, 1983.
- [5] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., ToBI: A standard for labeling English prosody, In: *Proc. ICSLP 1992*, Banff, pp. 867-870, 1992.
- [6] Hirst, D. J., The analysis by synthesis of speech melody: From data to models, *J. Speech Science*, 1:55-83, 2011.
- [7] Xu, Y., and Wang, Q. E., Pitch targets and their realization: Evidence from Mandarin Chinese, *Speech Commun.*, 33: 319-337, 2001.
- [8] Prom-on, S., Liu, F., and Xu, Y., Functional modeling of tone, focus, and sentence type in Mandarin Chinese, In: *Proc. ICPhS XVII*, Hong Kong, pp. 1638-1641, 2011.
- [9] Prom-on, S., Liu, F., and Xu, Y., Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling *J. Acoust. Soc. Am.*, 132: 421-432, 2012.
- [10] Chen, Y., and Xu, Y., Production of weak elements in speech Evidence from f0 patterns of neutral tone in standard Chinese, *Phonetica*, 63:47-75, 2006.

Semi-automatic and automatic tools for generating prosodic descriptors for prosody research

Plínio A. Barbosa

Speech Prosody Studies Group, Dep. of Linguistics, State Univ. of Campinas, Brazil

pabarbosa.unicampbr@gmail.com

Abstract

This paper presents four Praat scripts which the author of this article developed for analysing speech rhythm and intonation, for helping scientific research on prosody modelling and for investigating the link between prosody production and perception. The BeatExtractor script does automatic, language-independent detection of vowel onsets; the SGdetector script does language-dependent, semi-automatic detection of syllable-sized normalised duration peaks for the study of prominence and boundary marking, the SaliencyDetector script does language-independent automatic detection of syllable-sized normalised duration peaks for the study of prominence and boundary, and the ProsodyDescriptor script allows to generate 12 prosodic parameters related to duration, F_0 and spectral emphasis to the study of rhythm and intonation. All scripts are freely available and were tested in previous research since 2006. The languages tested for the first three scripts were Brazilian Portuguese, European Portuguese, German, French, English and Swedish.

Index Terms: tools, Praat, duration, F_0 , speech prosody

1. Introduction

Due to volume of data, prosody research can enormously benefit from the automatization of procedures for describing prosodic functions such as prominence, boundary and discursive relations marking. Automatization is advantageous because the same procedures can be applied to analyze the entire corpus and that is useful for preparing data for the statistical analysis as well.

This paper presents four scripts running on Praat [1] which generate prosodic descriptors for prosody research. Assuming from early research that duration is a crucial parameter for signalling stress, prominence and boundary in languages such as English, Portuguese, French, German and Swedish, the scripts are able to detect prosodic boundaries and prominences, as well as to generate a 12-parameter vector of duration, intensity and F_0 -related descriptors for prosodic analysis. The usefulness of the scripts is discussed regarding the results obtained from their application in previous research.

2. Semi-automatic detection of acoustic salience via duration

The *SGdetector* script for Praat was implemented in 2004 and improved in 2009 and 2010 for allowing the semi-automatic detection of local peaks of smoothed, normalised syllable-sized durations. Languages that use duration to signal both stress

and prosodic boundary such as Brazilian Portuguese (henceforth BP) and Swedish [2], English [3, 4], German [5, 6] and French [7] are well suited to take advantage of such a tool. Although thoroughly tested since 2004 with BP (see, for instance, [8, 9]), the script was used to do analyses in the other languages cited here and has potential to be applied to other languages, at least to other genetically related languages.

The input files for running the script are a TextGrid file containing a phone-sized or syllable-sized segmentation and a broad phonetic transcription of the corresponding Sound file, as well as a TableOfReal file containing a table listing the means and standard-deviations in milliseconds of the phone durations of the language under study. This latter file is delivered as part of the script and is available for BP, European Portuguese, British English, German, Swedish and French. Manual and semi-automatic segmentations and transcriptions of audio files were repeatedly tested for BP along the years (see [9]), confirming the usefulness and correction of a method for detecting prominence and boundary based on syllable-sized duration.

Syllable-sized segmentation is meant as a first step to capture prosodic-relevant duration variation along the utterances [10] and is understood here as an interval between two consecutive vowel onsets. This unit constitutes a phonetic syllable called a VV unit. Besides the crucial importance of vowel onset detection for speech signal processing [11], a clear advantage of a segmentation based on vowel onsets is its potential for automatic detection [9] even under moderately noisy conditions. Automatic detection of vowel onsets can be carried out by using a Praat script developed in 2005 [12, 9], the *BeatExtractor* script, explained in more details in section 2.1.

In the *SGdetector* script, detection of peaks of prosodic-relevant VV duration is carried out by serially applying a technique of normalisation followed by a smoothing technique. For normalising VV duration, the script uses the z -score transformation given in equation 1, where dur is the VV duration in ms, the pair (μ_i, var_i) , the reference mean and variance in ms of the phones within the corresponding VV unit. These references are found in [12, p. 489] for BP. For the other languages, they can be freely obtained from the author.

$$z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}} \quad (1)$$

For smoothing, the script applies a 5-point moving average filtering technique given by equation 2 to the sequence of z -scores (z_i).

$$z_{smoothed}^i = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13} \quad (2)$$

The two-step procedure described here aims at minimising the effects of intrinsic duration and number of segments in the VV unit, as well as minimising the effect of the implementation of stress irrelevant for the prosodic functions of prominence and boundary marking. Local peaks of smoothed z - scores are detected by tracking the position of the VV unit for which the discrete first derivative of the corresponding smoothed z - score changes from a positive to a negative value.

At the output, the script generates two text files, a new TextGrid object and an optional trace of the syllable-sized smoothed/normalised duration along the time-course of the Sound file under analysis. The first text file is a 5-column table displaying the following values for each VV unit: (1) the given transcription recovered from the TextGrid itself, e.g., “eNs”, “at” (even for the case where the segmentation is made phonewise), (2) the raw duration in milliseconds, (3) the z - score of the raw duration, (4) the 5-point-smoothed z - score and (5) a binary value indicating if the position is a local peak of smoothed z - score (value 1) or not (value 0). The second text file is a 2-column table containing (1) the raw duration in milliseconds of duration-related stress groups, delimited by two consecutive peaks of smoothed z - scores and (2) the number of VV units in the corresponding stress group. This table was used a lot of times to evaluate the degree of stress-timing of a speech passage, for instance in [13].

The TextGrid generated by the script contains an interval tier delimiting the detected stress group boundaries, synchronised with the input TextGrid, which allows, when selected with the corresponding Sound file, to listen to the chunks that end with a duration-related salience. The optional feature, implemented when the option “DrawLines” is chosen in the input parameters windows, plots a trace of the smoothed z - scores synchronised with the VV unit sequence: each value of smoothed z - scores is plotted in the y-axis in the position of each vowel onset along the plotted original TextGrid. The advantage of this choice for integrating intonation and rhythm descriptions is discussed below.

The correspondence between smoothed z - scores peaks and perceived salience, which refers to both prominence and prosodic boundary, is striking. In [9], we demonstrated an accuracy varying from 69 to 82 % between perceived and produced salience, as shown in Table 1 for the semi-automatic algorithm described here.

Table 1: it Precision, recall, and accuracy in percentage (%) for semi-automatic detected salience against perceived salience for the Lobato corpus read by a female (F) and a male (M) speaker at slow (s), normal (n) and fast (f) rates.

Sp/rate	precision	recall	accuracy
F/n	90	74	82
F/f	73	57	69
M/s	88	67	73
M/f	61	70	70

Perceived salience was determined by asking two groups of ten listeners to evaluate two readings of a passage by two BP speakers (a male and a female at two distinct rates). The listeners in both groups were lay undergraduate students in Linguistics. They were free to listen to the four readings as many times as they wanted. In the first group, each listener was given

a handout with the ortographic transcription of the recording and was instructed to circle all the words s/he considered highlighted by the speaker. The second group was instructed to circle the words that preceded a boundary. In each group, the percentage of listeners that circled each word in the text for each reading was initially used to define three levels of salience, according to a one-tailed z-test of proportion. Since the smallest proportion significantly distinct from zero is about 28 % for $\alpha = 0.05$ and $N = 10$, words circled by less than 30 % of the listeners were considered non-salient. For $\alpha = 0.01$, the threshold for rejecting the null hypothesis is about 49 %. Thus, words circled by 50 % of the listeners or more were considered strongly salient. Words salient by between 30 and 50 % of the listeners were considered weakly salient. For the purpose of computing the performance measures in the table, weakly and strongly salient words were both considered as “salient”.

The relatively high correspondence between perceived and produced salience allowed us to evaluate the degree of stress-timing in two different speaking styles for two varieties of Portuguese [13]. This work revealed that the speech rhythm of Portuguese speakers differs remarkably from the rhythm of Brazilian speakers when both groups narrate but not when both groups of speakers read. This was possible to demonstrate through the linear correlation between interval durations delimited by smoothed z - score peaks and number of VV units in the same interval. These two series of values were recovered from one of the tables generated by the SGdetector script.

2.1. Making the script completely automatic

For helping detecting produced salience in large corpora, the SGdetector script was modified into a *SalienceDetector* script for which phone labelling and manual vowel onset marking was made unnecessary. For this we associated a script made some time ago, *BeatExtractor* script [12], with the SGdetector script described above.

The BeatExtractor script implements Cummins’ Beat Extractor [14] with some modifications. It generates a TextGrid containing intervals between consecutive vowel onsets. It runs according to five steps: (1) the speech signal is filtered by a default second-order Butterworth (or Hanning) filter; (2) the filtered signal is then rectified; (3) the rectified signal is low-pass filtered using 20 Hz (see step 4a) or 40 Hz (see step 4b) as cut-off frequencies. This signal is normalised by dividing all points by the maximum value. This normalised, band-specific amplitude envelope is called the beat wave, a technique also applied by [14, 15]; (4) a vowel onset is set either (a) at a point where the amplitude of the beat wave local rising is higher than a certain threshold, or (b) at a local maximum of the normalised first derivative of the beat wave, provided this maximum is higher than a certain threshold; (5) a Praat TextGrid is generated that contains all vowel onsets as interval boundaries. More details in [9].

After obtaining the vowel onset positions, the *SalienceDetector* script proceeds by computing duration z - scores by using fixed values for the reference mean ($Refmean = 193$ ms) and standard-deviation ($RefSD = 47$ ms) duration according to equation 3, where m estimates the actual number of VV units between each interval generated by the BeatExtractor algorithm, which may miss vowel onsets (up to 20 % from all vowels effectively present in the Sound file).

$$z = \frac{\frac{\sqrt{m}}{m} \cdot dur - \sqrt{m} \cdot Refmean}{RefSD} \quad (3)$$

Smoothed z – scores are determined in the same way as before, by using the 5-point moving average filter. The output files are the same of the semi-automatic SGdetector script. The performance of this algorithm is a little lesser than the semi-automatic algorithm, as it can be seen in Table 2, for which accuracy varies from 53 to 80 %.

Table 2: Precision, recall, and accuracy in percentage (%) for automatic detected salience against perceived salience for the Lobato corpus read by a female (F) and a male (M) speaker at slow (s), normal (n) and fast (f) rates.

Sp/rate	precision	recall	accuracy
F/n	80	69	74
F/f	61	53	61
M/s	75	57	62
M/f	78	67	79

Its performance can be enhanced by manually changing the input parameters or by using a gradient-descent technique to find the input parameters that achieve the better performance in a limited set of utterances of a particular language, since this script is not language-dependent. Its usefulness depends essentially on the relevance of syllable-sized duration to signal both boundary and prominence. As an additional feature, the SalienceDetector script also indicates the occurrence of silent pauses in the corresponding TextGrid interval.

3. Describing the relations between F0 trace and syllable-sized duration trace

The normalised syllable-sized duration trace obtained with the “DrawLines” option of the SGDetector script was conceived in such a way as to give the value of normalised duration along the vowel onsets of the utterance. This feature allows the possibility of plotting the F0 contour of the utterance against the evolution of normalised duration and examining the VV units for which pitch accents and boundary tones coincide with normalised duration peaks. This was presented in [8].

Table 3 presents results of such coincidences in terms of a priori and conditional probabilities for both read paragraphs (two male subjects) and spontaneous speech (a male and a female subject). A priori probabilities are the proportion of pitch accents, $p(F_0)$, and normalised duration peaks, $p(dur)$, considering the total number of phonological words. Conditional probabilities consider the co-occurrence between a duration peak with a pitch accent over the total number of duration peaks, $p(F_0/dur)$, or the total number of pitch accents, $p(dur/F_0)$. A significant difference, computed from a test of proportions with $\alpha = 0.02$, between a priori and conditional probabilities signals a dependence between pitch accent and duration peak.

The table shows that there is a dependence between duration peak and pitch accent for the female speaker in spontaneous speech, as well as for speaker AC in read speech: for the latter, a pitch accent implies 76 % of chance of a duration peak. For the female speaker both are inter-related. This inter-relation is confirmed when the analysis is restricted to major prosodic bound-

Table 3: A priori probability of pitch accent $p(F_0)$ and duration peak $p(dur)$ in percentage (%) of number of phonological words. Speaker and speaking style are indicated. Stars signal significant differences between a priori and conditional probabilities ($\alpha = 0.02$).

sp (sp.sty)	$p(F_0)$	$p(F_0/dur)$	$p(dur)$	$p(dur/F_0)$
F (spont.)	63 *	79 *	49 *	63 *
M (spont.)	73	80	48	56
AC (read)	54	66	56 *	76 *
AP (read)	70	83	65	74

aries in read speech (utterance boundaries, clause and subject-predicate boundaries): 98 % (speaker AP), and 100 % (AC) of the time, both pitch accent and duration peak occur in the same lexical item, usually in the stressed vowel for pitch accents, and in the stressed or pre-pausal VV unit for duration peaks. Fig. 1 illustrates how both traces can be visualised. This was possible with the use of the “DrawLines” option of the SGDetector script. In this figure, the labels “sg1” and “sg2” signal the first two stress groups. The first rising contour during “sg1” signals a prominence not accompanied by a duration peak. The two low boundary tones inside the stress groups ending in “ano” and “viver” occur during a VV unit with a duration peak.

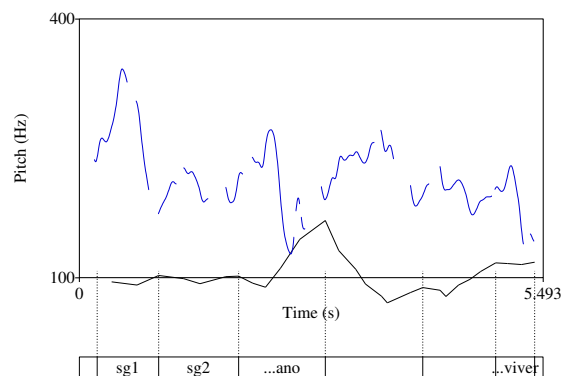


Figure 1: F0 contour superposed on the VV normalised duration contour of read utterance “Manuel tinha entrado para o mosteiro há quase um ano, mas ainda não se adaptara àquela maneira de viver.”

4. Semi-automatic extraction of global prosodic parameters

The *ProsodyExtractor* script delivers 12 prosodic descriptors for whole utterances or chunks of the same utterance in order to allow research on the link between prosody production and perception. This script has as input parameters the names of the Sound and corresponding TextGrid files. The TextGrid file must be composed of two interval tiers, one with the labelling and segmentation of the VV units (VV tier), and the other with the delimitation of the chunks of the audio file for analysis (Chunk tier). The number of intervals in the Chunk tier can vary from one to any number of units corresponding to any kind of phras-

ing needed for the intended analysis (e.g., syntactic phrases, prosodic constituents like stress groups, content-based chunks, among others). F_0 contour is also computed, thus, it is necessary, as for the Pitch buttons in Praat, to inform minimum and maximum pitch range.

For each chunk in the corresponding interval tier, the algorithm generates (a) 6 duration-related measures computed from the metadata obtained by using the algorithm of the previously described SGdetector script, (b) 5 descriptors obtained from the Pitch object computed by the script and (c) a measure of spectral emphasis as defined by [16]. The six duration-related measures computed in each chunk are: speech rate in VV units per second (sr), maximum of smoothed VV duration $z - score$, mean of smoothed VV duration $z - score$, standard-deviation of smoothed VV duration $z - score$, rate of smoothed VV duration $z - score$ local peaks (pr), and rate of non-salient VV units. The five F_0 descriptors are F_0 median, range, maximum, minimum, as well as F_0 peak rate. For computing the latter measure a smoothing function (with cut-off frequency of 1.5 Hz) followed by a quadratic interpolation function are applied before the F_0 peak rate computation.

The 12 measures generated per chunk can be used both to study the evolution of these prosodic parameters throughout a speech signal, as well as to correlate prosody production and perception. As regards the latter, we used the difference of these values between paired utterances as predictors of the degree of discrepancy between perceived manner of speaking [17]. The experimental design consisted in instructing 10 listeners to evaluate two subsets of 44 audio pairs combining 3 different speakers of BP and two speaking styles, storytelling and reading. The instruction was "Evaluate each pair of excerpts as to how they differ according to the manner of speaking given a scale from 1 (same manner of speaking) to 5 (very different manner of speaking)". After testing more than 50 models of multiple linear regression, results showed that the best model was the one which explained 71 % of the variance of the listeners responses (lr), as given in equation 4 with $p - value$ of at least 0.009 for all coefficients ($F_{3,11} = 12.4$, $p < 0.0008$).

$$lr = -1.5 + 10.4pr + 2.65sr - 10.75pr * sr \quad (4)$$

This reveals that the significant production parameters that explain the listeners' performance are speech rate in VV units/s and normalised duration peak rate, which can be associated with the syllable succession and salient syllable succession.

5. Summary and availability of the tools

The tools presented here were used to conduct research on speech rhythm analysis and modelling either in a single language or crosslinguistically, on the relation between intonation and rhythm both *stricto sensu*, as well as on the link between speech rhythm production and perception. They were tested in French, German, Brazilian and European Portuguese, Swedish and English, the latter two less systematically. All scripts are available freely from the author, including a Praat Script not considered in this article but which might be of interest to those who investigate speech expressivity, the ExpressionEvaluator script, which extracts five classes of acoustic parameters and four statistical descriptors, producing 12 acoustic parameters.

All scripts are available freely from the author, are licensed under the terms of the GNU General Public License as pub-

lished by the Free Software Foundation; version 2 of the License. They were tested in French, German, Brazilian and European Portuguese, Swedish and English, the latter two less systematically.

6. Acknowledgment

The author thanks a research grant from CNPq (301387/2011-7) and Sandra Madureira for revising the manuscript.

7. References

- [1] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" [Computer program], Online: <http://www.praat.org>.
- [2] Barbosa, P.A., Eriksson, A. and Åkesson, J., "Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese", in E.L. Asu and P. Lippus [Eds], Nordic prosody. Proceedings from the XIth conference, Tartu 2012 (pp. 97-106). Frankfurt am Main: Peter Lang, 2013.
- [3] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J., "Segmental durations in the vicinity of prosodic boundaries", Journal of the Acoustical Society of America, 91(3):1707-1717, 1992.
- [4] Fry, D. B., "Experiments in the perception of stress", Language and Speech, 1:126-152, 1958.
- [5] Dogil, G., "Phonetic correlates of word stress", in Van der Hulst, [Ed], Word Prosodic System of European Languages, 371-376, De Gruyter, Berlin, 1995.
- [6] Sluijter, A. M.C., "Phonetic Correlates of Stress and Accent", Ph.D. Thesis, Holland Institute of Generative Linguistics, Leiden, 1995.
- [7] Barbosa, P.A., "Caractérisation et génération automatique de la structuration rythmique du français". PhD thesis, ICP/Institut National Polytechnique de Grenoble, France, 1994.
- [8] Barbosa, P. A., "Prominence- and boundary-related acoustic correlations in Brazilian Portuguese read and spontaneous speech", Proc. Speech Prosody 2008, Campinas (pp. 257-260), 2008.
- [9] Barbosa, P. A., "Automatic duration-related salience detection in Brazilian Portuguese read and spontaneous speech", Proc. Speech Prosody 2010, Chicago (100067:1-4), 2010. Online: "<http://www.speechprosody2010.illinois.edu/papers/100067.pdf>."
- [10] Barbosa, P.A., "At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration", Proc. of the 1st ETRW on Speech Production Modeling, Autrans, (pp. 85-88), 1996.
- [11] Dogil, G., Braun, G., The PIVOT model of speech parsing, Verlag, Wien, 1988.
- [12] Barbosa, P. A., "Incursões em torno do ritmo da fala", Campinas: RG/Fapesp, 2006.
- [13] Barbosa, P. A., Viana, M. C. and Trancoso, I., "Cross-variety Rhythm Typology in Portuguese", Proc. of Interspeech 2009 - Speech and Intelligence. Brighton, UK (pp. 1011-1014). London: Causal Productions, 2009.
- [14] Cummins, F., Port, R., "Rhythmic constraints on stress timing in English", J. Phon., 26:145-171, 1998.
- [15] Tilsen, S., Johnson, K., "Low-frequency Fourier analysis of speech rhythm", JASA Express Letters, 124(2), EL34, 2008.
- [16] Traunmüller, H. and Eriksson, A., "The frequency range of the voice fundamental in the speech of male and female adults", Unpublished Manuscript. Online: <http://www.ling.su.se/staff/hartmut/aktupub.htm>.
- [17] Barbosa, P. A. and da Silva, W., "A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches", in H. Caseli et al. [Eds], PROPOR 2012, LNAI 7243 (pp. 329-337). Springer, Heidelberg, 2012.

Variability of voice fundamental frequency in speech under stress.

Grażyna Demenko^a, Magdalena Oleśkiewicz-Popiel^a,
Krzysztof Izdebski^{b,c}, Yuling Yan^c

^a Department of Phonetics, A. Mickiewicz University of Poznan, Poznan, (Poland)

^{b,c} Pacific Voice and Speech Foundation, San Francisco, CA (USA)

^c Santa Clara University, Santa Clara (USA)

lin@amu.edu.pl, mmj@amu.edu.pl
kizdebski@pvsf.org, yyan1@scu.edu

Abstract

By analyzing acoustic and phonetic structure of live recordings of 45 speakers from police 997 emergency call center in Poland, we demonstrated how stressful events are coded in the human voice. Statistical measurements of stressed and neutral speech samples showed relevance of the arousal dimension in stress processing. The MDVP analysis confirmed statistical significance of following parameters: fundamental frequency variation, noise-to-harmonic-ratio, sub-harmonics presence and voice quality irregularities. In highly stressful conditions a systematic over-one-octave shift in pitch was observed. Linear Discriminant Analysis based on nine acoustic features showed that it is possible to categorize speech samples into one of the following classes: male stressed or neutral, or female stressed or neutral.

Index Terms: call centers interfaces, detection of vocal stress, stress visualization, physiological correlates

1. Introduction

Recognition of whether a speaker is under stress is of crucial value in many civilian and military applications, hence automatic detections of vocal stress is becoming increasingly important. Applications of this technology can be found in multilingual communication, in security systems, in banking, in homeland security and in law enforcement [1, 2, 3, 4]. Automatic detection of voice under stress is specifically crucial in emergency call centers and in police departments, as all over the world these units of public safety are overloaded with different kinds of calls, only some of which represent a real danger and a need of an immediate response. Hence, to improve decision making process, response effectiveness, and to save lives, it is of pragmatic interest to detect automatically those speech signals that contain vocally mediated stress [2, 3, 4].

Several investigations [5, 6] showed direct evidence of emotion recognition to stress verification [5, 7, 8] by highlighting differences in acoustical features between the neutral and stressed speech signals brought by a variety of emotions [2, 9]. A number of these studies focused on the effects of emotions on stress because of a close relation between emotions and stress recognition, e.g. usage of similar acoustic features (F_0 , intensity, speech unit duration) and arousal dimension [10, 11, 12]. These studies demonstrate that emotional speech correlates are dependent on physiological constraints and do correspond to broad classes of basic emotions, but disagree on the specific differences between the acoustic correlates of particular classes of emotions [11, 13].

Certain emotional states can be correlated with physiological states, which in turn have predictable effects on speech and on its prosodic features. For instance, when a person is in a state of anger, fear or joy, the sympathetic nervous system is aroused and speech becomes louder, faster and enunciated with stronger high-frequency energy. When one is bored or sad, the parasympathetic nervous system is also aroused, which results in a slow, low-pitched speech with little high-frequency energy [10]. Apart from these differences, other studies showed an increase in intensity and in fundamental frequency, a stronger concentration of energy above 500 Hz and an increase in speech rate in cases of stressed speech [10].

A number of studies have considered analysis of speech under both simulated and actual stress condition, though the interpretation of speech characteristics is not unambiguous [10]. Research frequently reports on conflicting results, due to differences in experimental design, categorization of actual or simulated stress, and/or interpretation of results [10]. Studies using actors, simulated stress or emotions have the advantage of a controlled environment, but their major disadvantage is, however, artificial nature of these signals that can result in producing highly exaggerated misrepresentations of emotions in speech [10].

Few studies focused on analysis of authentic recordings coming from actual stressful situations [10]. There is usually no doubt as to the presence of stress in these situations; however there is a problem with categorization of the homogeneous classes of vocally embedded stress. Our study therefore focuses on the analysis of voice stress produced in response to real live situations, eliminating variables present in simulated stress studies. The research aims at extraction of those acoustic features which produce stressed in vocalization in a relatively homogenous group of actual threat stressors.

While some progress has been made in the area of stress definition and assessment from the acoustic or visual signals [10, 7, 14], visual correlates of vocal folds or supraglottic larynx contribution of these affected signals are essentially non-existing [10, 15], hence, we analyzed only the third order of stressor --the psychological ones-- which have their effect at the highest level of speech production [2]. External stimuli such as a threat are subject to individual cognitive evaluations and the emotional states they may bring about (i.e. fear, anger, irritation) to affect speech production at its highest levels. We hypothesized that models using live speech samples from both, stressful and neutral environment, will provide better determinants of acoustic stress indicators, and will help answering questions which of the prosodic derivatives are most valuable vocal stress indicators and which can be used in automatic stress detection.

2. Speech corpus construction and annotation

The 997 - Emergency Calls Database is a collection of spontaneous speech recordings that comprises crime/offence notifications and police intervention requests. All recordings are automatically grouped into sessions according to the phone number from which the call was made. In all over 8 000 sessions were available.

From this corpus, a six-levels preliminary manual phonetic annotation was performed: (1) background acoustics, (2) types of dialog, (3) suprasegmental features such as: (3.1) speech rate (fast, slow, rising, decreasing), (3.2) loudness (low voice or whisper, loud voice, decreasing or increasing voice loudness), (3.3) intonation (rising, falling or sudden break of melody and unusually flat intonation), (4) context (threat, complaint and/or depression), (5) time (passed, immediate and potential), (6) emotional coloring (up to three categorical labels and values for three dimensions: potency, valency, arousal; where potency is the level of control that a person has over the situation causing the emotion, valency states whether the emotion is positive or negative and arousal refers to the level of intensity of an emotion [10, 8, 16].

The annotation allowed for choosing a fairly uniform group of 45 speakers, both males and females, and voice stress detection was performed only on those speakers who manifested different arousal level in two or more dialogs.

3. Pitch characteristics of stress

3.1. Pitch register

A key issue of stress detection by machine is defining utterance segmentation that would result in clear units with respect to perceptual and acoustic homogeneity. Vocal registering, which divides voice region ranges into registers, is an important perceptual category. There are many approaches to define vocal register. For example vocal registration is perceptually a distinct region of vocal quality that can be maintained over some ranges of pitch and loudness over consecutive voice frequencies without a break [17]. However, vocal register definition and register classification terminology is one of the most controversial problems and that physiological register correlates are not defined [15, 18]. Therefore, as a solution to that problem, three cases have been presupposed: (1) different pitch position, same pitch range, (2) different pitch position different pitch range, (3) same pitch position, different pitch range, where pitch range is the difference between F_{max} and F_{min} . (values for F_{max} and F_{min} averaged in the region were maximum and minimum was detected, in order to avoid pitch detection errors).

3.2. Pitch ranges

3.2.1. Different pitch position same pitch range

Based on statistical analysis three pitch position settings were observed in the studies: (1) relative constant pitch position within the utterance and dynamic pitch position changes within the utterance: (2) pitch position shifted upward, (3) pitch position shifting up and down. These are shown in the following Figures 1-5

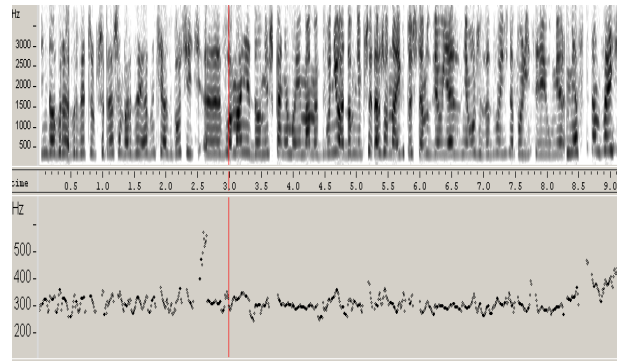


Figure 1a: F_0 contour of constant stress in the utterance: "Please, come over, there's a house-breaking. She's scared to death" ($F_{min}=240$ Hz, $F_{max}=352$ Hz).

1) Relative constant pitch position within the phrase.

Figure 1a is an acoustic representation of an utterance informing about a burglary and a life threat, whereas Figure 1b illustrates an utterance from the same person calling off the intervention (informing that the burglar has left the apartment), recorded one hour after the first call.

The follow-up call shows a downward F_0 shift in pitch position by approximately 40 Hz (Figure 1b), as compared to F_0 contour in utterance from Figure 1a.

The utterances in Figure 1a and 1b have similar pitch ranges but different pitch positions (we assume these were caused by stress).

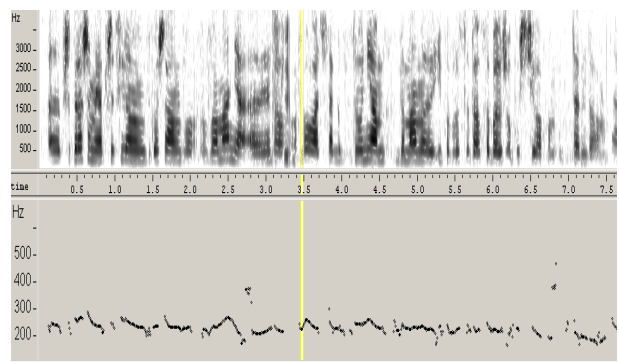


Figure 1b: F_0 contour of neutral speech in the utterance: "I called one hour ago, I want to call off the intervention" ($F_{min}=167$ Hz, $F_{max}=264$ Hz).

2) Dynamic change of pitch position within the utterance. Pitch position shifted upward.

In cases of high stress levels, F_0 can reach extreme values. For example female voices may be elevated up to 700 Hz. Figure 2a illustrates an utterance of a female speaking with extreme stress as she reports to the police, "a masked person has entered my apartment". Vocal stress decreases only slightly at the end of the recording, after hearing a dispatcher prompt asking her to calm down. As the stress of the speaker increases the following is noted: 1) an upward shift in the voice pitch, 2) as well as a prominence of the higher

frequencies in the spectrum, 3) an increase in the signal's energy and 4) rate changes.

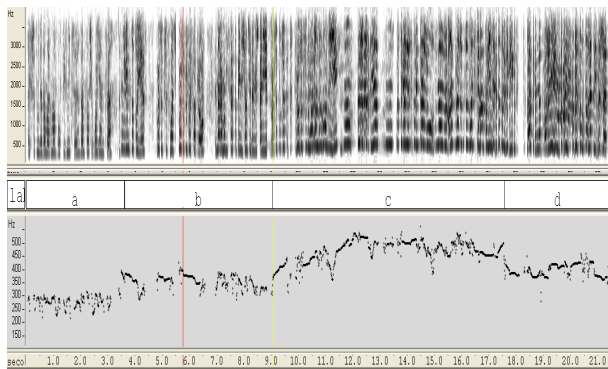


Figure 2a: A gradual stress increase in the utterances: a) "Someone is entering the apartment" ($F_{min}=220\text{ Hz}$), b) "He's masked" ($F_{min}=260\text{ Hz}$), c) "he is somewhere [here]" - direct threat ($F_{min}=320\text{ Hz}$), d) "Please come to Kwiatowa Street" - the answer after being asked by a police officer to calm down and tell him the address ($F_{min}=280\text{ Hz}$).

In cases of high levels of stress F_0 values can reach extreme values (even up to 750 Hz). Figure 2b illustrates an utterance marked by extreme stress increase that ended with a scream and an exceeding lengthening of some syllables. In this case F_0 changes are located in the range of 220 Hz - 750 Hz. As the stress of the speaker increases the following is observed: 1) an upward shift in the voice pitch, 2) as well as a prominence of the higher frequencies in the spectrum, 3) an increase in the signal's energy and 4) rate changes.

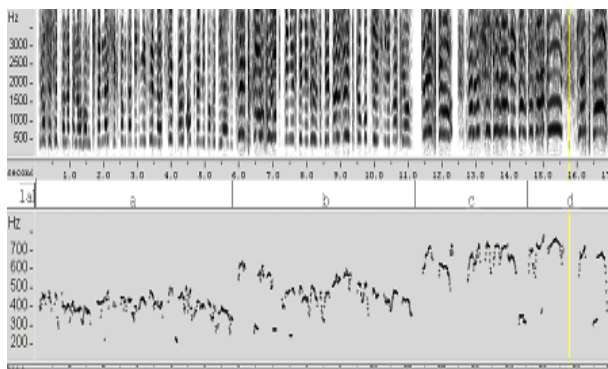


Figure 2b: A gradual increase in stress in the utterances: (a) "Please, [come] quickly to Kanatowa [street] 18, they want to kill my son, they've broken the window" ($F_{min}=289\text{ Hz}$), (b) "M E (name of the caller withheld), quickly, the mobsters have come" ($F_{min}=345\text{ Hz}$), (c) "Quickly. It's happening, they want to kill him" ($F_{min}=495\text{ Hz}$), (d) "Quickly, S... is killing him (scream)" ($F_{min}=495\text{ Hz}$, $F_{max}=748\text{ Hz}$).

3) Dynamic change of pitch position within the utterance. Pitch register shifted upward and downward.

The shaded part in the Figure 3 shows an utterance by male voice characterized by a significant, over 50Hz, upward shift of F_0 position.

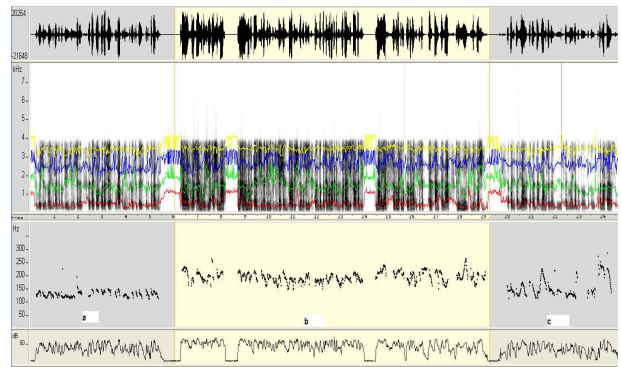


Figure 3: a) "I keep trying to get through..." ($F_{min}=121\text{ Hz}$), b) "I've reported it so many times already..." - clearly audible irritation ($F_{min}=173\text{ Hz}$) c) "... so I don't know anything anymore..." - the answer after being asked by a police officer to calm down ($F_{min}=115\text{ Hz}$).

3.2.2. Different pitch position different pitch range

In cases of anger and mixed emotions significant changes of both pitch position and pitch range were observed. Figure 4 illustrates F_0 contour for an utterance in a female voice, where first part (the end of which has been marked by the cursor) has been classified as the voice of indignation. The speaker can easily control her emotional state so that her message is clearly perceived by the listener. Each syllable that is lexically permissible is clearly stressed.

By comparison, in the final part of the recording (beginning of which has been marked by the cursor), as a result of the discourse, the female speaker softens and calms down her manner of speaking, so the recording has a different F_{min} and pitch range width than its first part.

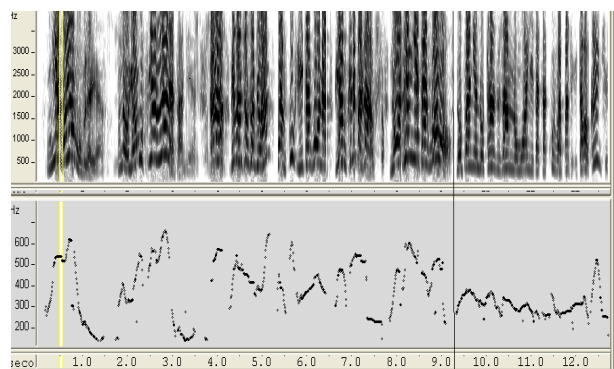


Figure 4: F_0 contour for an expressive utterance (indignation): "I've got here such a drunkard, he's maltreating me, I am going to trash him..." ($F_{max}=675\text{ Hz}$, $F_{min}=139\text{ Hz}$, first part of the utterance), "But what can I do..." ($F_{max}=275\text{ Hz}$, $F_{min}=206\text{ Hz}$, second part of the utterance).

3.2.3. Same pitch position different pitch range

Figure 5a and 5b illustrate utterances of the same male speaker, in neutral state and in anger respectively. Both utterances have similar F_{min} , their ranges of F_0 fluctuations however differ significantly.

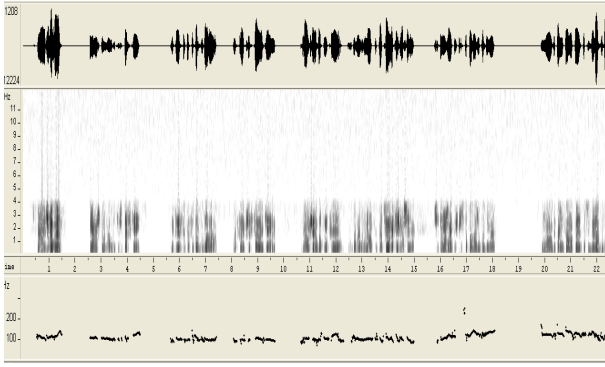


Figure 5a: F_0 contour for a neutral utterance: “Hi, I live on XXX street...” ($F_{max}=137\text{Hz}$, $F_{min}=92\text{Hz}$).

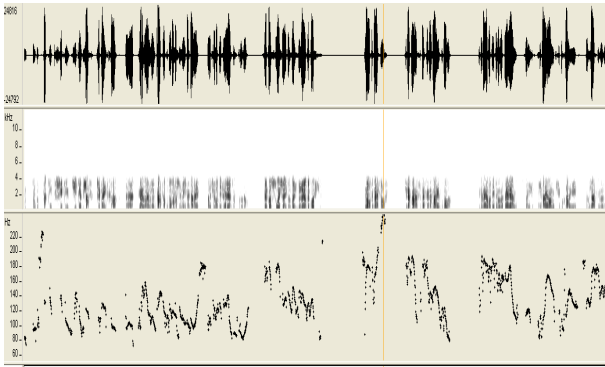


Figure 5b: F_0 contour for an expressive utterance of indignation: “I hear some shouting and name-calling... him...” ($F_{max}=252\text{Hz}$, $F_{min}=86\text{Hz}$).

4. Stress classification

The material was divided into four groups: G1: male – stress, G2: male – neutral/mild irritation, G3: female – stress, G4: female – neutral. Although acoustic analysis of MDVP allows for 32 features [19], only 9 have been used and correlated to LDA Linear Discriminant Analysis. The features used were: Average (F_0), Highest (F_{hi}) and Lowest Fundamental Frequency (F_{lo}), Fundamental frequency variation (vF_0 / %), Jitter (Jitt), Amplitude perturbation Quotient (sAPQ) / %, Degree of Subharmonic Segments (DSH) / %, Noise to Harmonic Ratio (NHR), Degree of voiceless DUV (%).

The LDA analysis of nine parameters enabled the classification of four groups with the average 80% accuracy, for two groups (neutral and stressed speech, males and female together) the accuracy was a bit higher, 84%. The results showed that extreme stress can be clearly identified by using only the amplitude information with mean and minimum F_0 values.

Figure 6 shows z-normalized F_{min} (F_{lo}) values for four groups: G1, G2, G3, G4. Highest pitch position (F_{min}) values are demonstrated by groups G1 and G3 (speech under stress), whereas F_{min} values for groups G2 and G4 are statistically substantially lower.

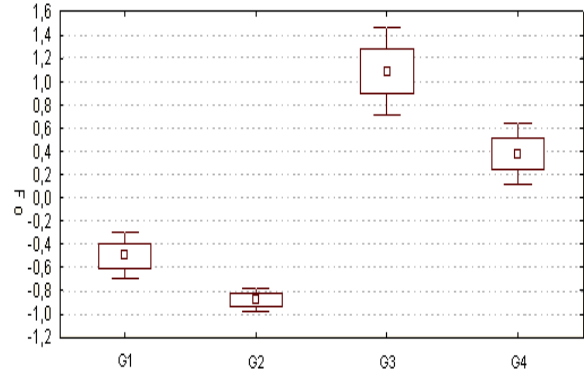


Figure 6: Z-normalized values F_{min} for G1, G2, G3, G4.

Table 1 shows classification results. Utterances by male voices affected by stress (G1) obtained better results than those of female voices affected by stress (G3).

	% correct	G_1:1 p=,23	G_2:2 p=,27	G_3:3 p=,23	G_4:4 p=,26
G_1:1	80,00	20	3	2	0
G_2:2	86,20	3	25	0	1
G_3:3	76,00	1	0	19	5
G_4:4	78,57	0	4	2	22
Total	80	21	35	21	30

Table 1: Classification matrix: rows – classification observed, columns – classification expected.

5. Stress visualization

An approach to characterize vocal folds (VF) vibrations from HSDI recordings using Nyquist plot was pioneered in Yan et al., [20, 21], while automatic and robust procedures to generate the glottal area waveform (GAW) from HSDI images were also provided by Yan et al. [22].

The principles underlying this approach are summarized below and illustrated in Figure 7. The HSDI-derived GAW is normalized for all of our analyses to a range of 0~1 with 0 corresponding to complete closure and 1 corresponding to maximum opening. This operation allows for standardized dynamic measurements of VF vibration. The Nyquist plot and associated analyses are used to represent the instantaneous property of the VF vibration, rather than a time averaged one. This property is revealed by the amplitude and phase of the complex analytic signal (e.g. in the form of Nyquist plot) that we generate from the Hilbert transform of the GAW as illustrated in Figure 7 (A, B, C). This operation is applied to as many as 200 glottal cycles taken from 4000-frames of a 2-second HSDI recording (i.e. at a 2000 Hz acquisition rate). Nyquist plots can be also derived from the acoustic signals. Here we submitted to Nyquist analysis the acoustic signals from neutral and stressed segments derived from our samples and to depict the differences in voice stress levels from the same speakers.

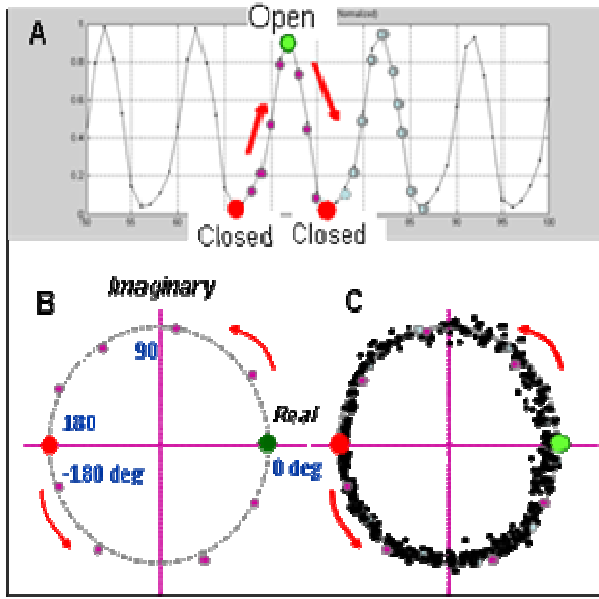


Figure 7: Concept of the Nyquist plot approach to characterize vocal fold vibrations.

- A)** a normalized GAW, representing 50 sequential frames (5 vibratory cycles) from a 2000 f/s HSDI recording; the open (0°) and closed (90°) glottal cycles are determined from automatic tracing of HSDI images (Yan et al, 2006b).
- B)** One vibratory cycle is mapped onto the complex plane, where the magnitude-phase of the analytic signal is graphed; the complex analytic signal is constructed from the Hilbert transform of the GAW.
- C)** Overlays of subsequent vibratory cycles generate a Nyquist plot - deviation of the points from the circle (scatter and shape distortion) reflects the effects of shimmer, jitter and nonlinearity.

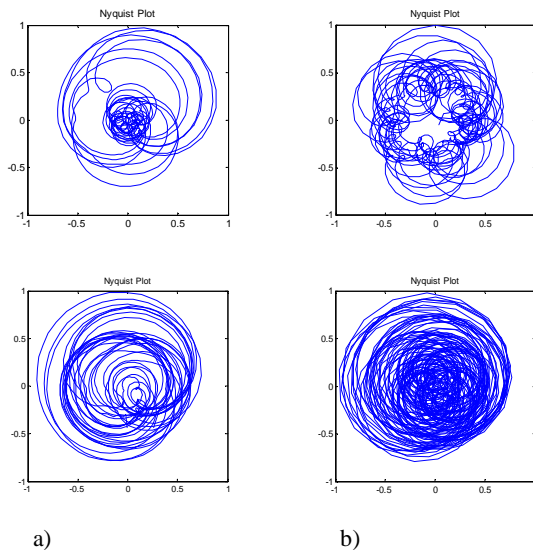


Figure 8: Nyquist plots for vowel "a" (Fig.8a) and "i" (Fig.8b) in neutral speech (upper plots) and speech under stress (bottom plots)

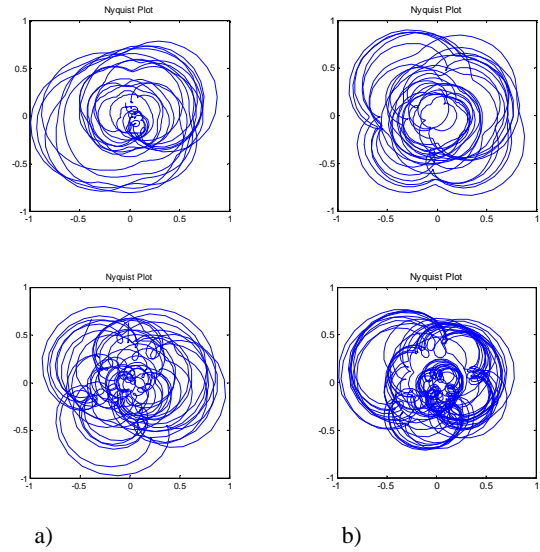


Figure 9: Nyquist plots for vowel "o" (Fig.9a) and vowel "o" from different phonetic context (Fig.9b), in neutral speech (upper plots) and speech under stress (bottom plots)

Figures 8a and 8b show Nyquist plots for vowel "a" and "i" in neutral speech (upper plots) and speech under stress (bottom plots).

Figures 9a and 9b show Nyquist plots for vowel "o" from two different contexts both in neutral speech (upper plots) and speech under stress (bottom plots). All vowels were extracted from continuous speech.

The differences in Nyquist plots for vowels in neutral speech (upper plots) and speech under stress (bottom plots) are obvious and very distinctive. Overall, more structured Nyquist patterns (for vowels "a", "i", "o") are observed in speech under stress in comparison to those in neutral speech. Yet, it should be noted that for an objective analysis it is necessary to use a standardized set of speech samples (mainly vowels or sonorants) that will enable evaluation of statistical significance.

6. Conclusion

Despite restricting the study to 45 speakers, a clear tendency in acoustic characterization of speech under stress was observed.

The results of this study confirm the crucial role of the F_0 parameter for investigating stress. Our results agree with literature [2, 5, 10, 23] and point that F_{\max} (averaged from several values within the region where maximum was detected, in order to avoid pitch detection errors) must be considered a particularly important parameter in the emotional stress detection. However, this and our previous work [24] showed that a shift in the F_0 contour is also a crucial stress indicator, thus an increase in F_{\max} in stressed speech results from a shift in the F_0 register. This holds specifically for vocalizations caused by fear. A systematic increase in the range of F_0 variability for the stress related to anger and to irritation was observed. The results also confirmed the need of including

shift of pitch position and change in pitch register width into prosodic structures segmentation.

We are now preparing to correlate these findings with visual (optical) observations of vocal fold activity using HSDI during production of various emotional vocal components. This will, in our opinion, enable improved explanation of the factors that influence pitch register changes in utterances diversified linguistically and in terms of situational context.

7. Acknowledgements

This project is supported by The Polish Ministry of Sciences and Higher Education (project no O R00 0170 12) and in parts by PVSF funding. We are grateful to Ms. Emma Marriott and Ms. Clara Lewis, for editing of this text.

8. References

- [1] Eisenberg, A., "Software that listens for lies," The New York Times, Sunday December 4, 2011.
- [2] Hansen, J., et al., "The Impact of Speech Under 'Stress' on Military Speech Technology," NATO report. Online: http://www.gth.die.upm.es/research/documentation/referencias/Hansen_The_Impact.pdf, 2007.
- [3] Lefter, J., Rothkrantz, L., Leeuwen, D., Wiggers, P., "Automatic stress detection in emergency (telephone) calls," International Journal of Intelligent Defence Support Systems 4(2), 148-168 (21), 2011.
- [4] Vidrascu, L., Devillers, L., "Detection of real-life emotions in call centers," Proc. of Interspeech, 1841-1844, 2005.
- [5] Shipp, T., Izdebski, K., "Current evidence for the existence of laryngeal macrotremor and microtremor," J. Forensic Sciences, 26, 501-505, 1981.
- [6] Cowie, R., Cornelius, R.R., "Describing the emotional states that are expressed in speech," Speech Communication, 40, 5-32, 2003.
- [7] Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., Friederici, A. D., "Affective encoding in the speech signal and in event-related brain potential," Speech Communication, 40 (1-2), 61-70, 2003.
- [8] Oudeyer, P.-Y., "The production and recognition of emotions in speech: features and algorithms," Int. J. of Human-Computer Studies 59 (1-2), 157-183 (2003).
- [9] Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann H., "Recognition of emotion in a realistic dialogue scenario," Proc. of the Int. Conf. on Spoken Language Processing Beijing, China, 665- 668, 2000.
- [10] Izdebski, K. (ed.), "Emotions in the Human Voice", [Vol. 1-3], Plural Publishing, San Diego, CA, 2008-2009.
- [11] Ekman, P., "An argument for basic emotions," Cognition and Emotion 6, 169-200, 1992.
- [12] Scherer, K.R., "What are emotions? And how can they be measured?", Social Science Information 44 (4), 695-729, 2005.
- [13] Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., "Desperately seeking emotions or: Actors, wizards, and human beings," Speech Emotion-2000, 195-200, 2000.
- [14] Izdebski, K., Yan Y., "Preliminary observations of vocal fold vibratory cycle with HSDI a function of emotional load," In progress (ePhonscope, 2013).
- [15] Izdebski, K., "SFCM 202 Voice Physiology Manual". In press e-Q&A-p, San Francisco, CA, 2013.
- [16] Fontaine, R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C., "The World of Emotions is not Two-Dimensional," Psychological Science 18 (12), 1050-1057, 2007.
- [17] Frič, M., Šram, F., Švec, J.G., "Voice registers, vocal folds vibration patterns and their presentation in videokymography," Proc. of ACOUSTICS High Tatras 06. 33rd International Acoustical Conference - EAA Symposium, Štrbské Pleso, Slovakia, October 4th - 6th, 2006. ISBN 80-228-1672-8, 42-45, 2006.
- [18] Shriberg, E., Ladd, D.R., Terken, J., Stolcke, A., "Modeling pitch range variation within and across speakers: predicting F0 targets when 'speaking up'," Proc. Of the Int. Conf. on Spoken Language Processing (Addendum, 1-4), Philadelphia, PA, 1996.
- [19] Deliyski, D., "Acoustic model and evaluation of pathological voice production," Proc. Eurospeech'93, 1969-1972, 1993.
- [20] Yan Y, Ahmad K, Kunduk M, Bless D, "Analysis of vocal fold vibrations from high-speed laryngeal images using a Hilbert transform based methodology", Journal of Voice 19(2), 161-175, 2005.
- [21] Yan Y, Edward Damrose, Diane Bless, "Functional Analysis of Voice Using Simultaneous High-Speed Imaging and Acoustic Recordings", Journal of Voice 21(5), 604-616, 2007.
- [22] Yan Y, Chen X, Bless D, "Automatic tracing of the vocal fold motion from high speed digital images", IEEE Trans Biomed Eng 53(7), 1394-1400, 2006.
- [23] Protopapas A., Lieberman P., "Fundamental frequency of phonation and perceived emotional stress," J. Acoust. Soc. Am. 101 (4), 2268-2277, 1997.
- [24] Demenko, G., "Voice Stress Extraction," Proc. of Speech Prosody Conference. May 6-9, 2008, Campinas, Brasil, 53-56, 2008.

Index of authors

Aalto, D.	78
Asano, Y.	31
Asaridou, S.	31
Astésano, C.	15
Avanzi, M.	27
Barbosa, P.	86
Bel, B.	1
Bertrand, R.	59
Bigi, B.	15, 11, 62
Bénard, F.	1
Cangemi, F.	31
Cho, H.	11
D'Imperio, M.	15
Demenko, G.	90
Ding, H.	11
Fon, J.	20
Fossard, M.	27
Garrido, J.M.	74, 38
Gibbon, D.	66
Goldman, J.P.	55
Gonzalez, S.	27
Gubian, M.	31
Gurman Bard, E.	15
Herment, S.	11, 24
Hirst, D.	11, 36, 62
Izdebski, K.	90
Karpiński, M.	51
Klessa, K.	51
Martin, P.	47
Mertens, P.	42
Nguyen, N.	15
Oleśkiewicz-Popiel, M.	90
Peshkov, K.	59
Prom-on, S.	82
Prévot, L.	59, 15
Rousier-Vercruyssen, L.	27
Schwab, S.	27
Suni, A.	78
Turcsan, G.	24
Turk, A.	15
Vainio, M.	78
Wagner, A.	51
Wang, S.F.	20
Wang, T.	11

Xu, Y.	7, 82
Yan, Y.	90
Yu, J.	70

List of authors

Aalto Daniel	daniel.aalto@helsinki.fi
Asano Yuki	yuki.asano@uni-konstanz.de
Asaridou Salomi	s.asaridou@donders.ru.nl
Astésano Corine	corine.astesano@univ-tlse2.fr
Avanzi Mathieu	mathieu.avanzi@unine.ch
Barbosa Plinio	pabarbosa.unicampbr@gmail.com
Bel Bernard	bernard.bel@lpl-aix.fr
Bertrand Roxane	roxane.bertrand@lpl-aix.fr
Bigi Brigitte	brigitte.bigi@lpl-aix.fr
Bénard Frédérique	frederique.benard@lpl-aix.fr
Cangemi Francesco	fcangemi@uni-koeln.de
Cho Hyongsil	t-hych@microsoft.com
D'Imperio Mariapaola	mariapaola.dimperio@lpl-aix.fr
Demenko Grażyna	lin@amu.edu.pl
Ding Hongwei	hongwei.ding@tongji.edu.cn
Fon Janice	jfon@ntu.edu.tw
Fossard Marion	marion.fossard@unine.ch
Garrido Juan-María	juanmaria.garrido@upf.edu
Gibbon Dafydd	gibbon@uni-bielefeld.de
Goldman Jean-Philippe	jeanphilpegoldman@gmail.com
Gonzalez Sylvia	sylvia.gonzalez@unine.ch
Gubian Michele	m.gubian@let.ru.nl
Gurman Bard Ellen	ellen@ling.ed.ac.uk
Herment Sophie	sophie.herment@univ-amu.fr
Hirst Daniel	daniel.hirst@lpl-aix.fr
Izdebski Krzysztof	kizdebskif@pvsf.org
Karpiński Maciej	maciejk@amu.edu.pl
Klessa Katarzyna	klessa@amu.edu.pl
Martin Philippe	philippe.martin@utoronto.ca
Mertens Piet	Piet.Mertens@arts.kuleuven.be
Nguyen Noel	noel.nguyen@lpl-aix.fr
Oleśkiewicz-Popiel Magdalena	mmj@amu.edu.pl
Peshkov Klim	klim.peshkov@lpl-aix.fr
Prom-on Santitham	santitham@cpe.kmutt.ac.th
Prévot Laurent	laurent.prevot@lpl-aix.fr
Rousier-Verduyssen Lucie	lucie.rousier-verduyssen@unine.ch
Schwab Sandra	sandra.schwab@unige.ch
Suni Antti	antti.suni@helsinki.fi
Turcsan Gabor	gabor.turcsan@univ-amu.fr
Turk Alice	turk@ling.ed.ac.uk
Vainio Martti	martti.vainio@helsinki.fi
Wagner Agnieszka	wagner@amu.edu.pl
Wang Sheng-Fu	sftwang0416@gmail.com
Wang Ting	sweetwangting@gmail.com

Xu Yi
Yan Yuling
Yu Jue

yi.xu@ucl.ac.uk
yyan@scu.edu
erinyu@126.com