Building *OMProDat*: an open multilingual prosodic database

Daniel Hirst[1,2], *Brigitte Bigi*[1], *Hyongsil Cho*[3,4], *Hongwei Ding*[2], *Sophie Herment*[1], *Ting Wang*[2]

[1]Laboratoire Parole et Langage, UMR 7309 CNRS & Aix-Marseille University, France
[2]School of Foreign Languages, Tongji University, Shanghai, China
[3]Microsoft Language Development Center, Lisbon, Portugal
[4]ADETTI – 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 mutlilingual 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.

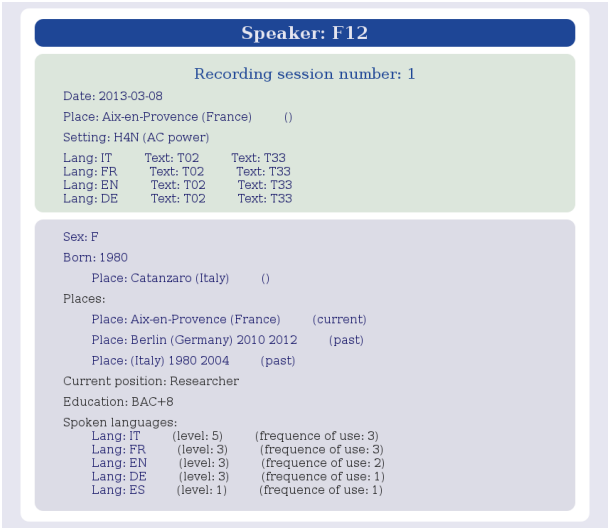

Figure 1: *Available participant information sheet*.

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. Acknoweldgements

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.; Mouropoulos, J.; Senia, F.; Trancoso, I.; 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., Jaèn, 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, Saarbrucken, 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.; Kiriyama, 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.