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.)

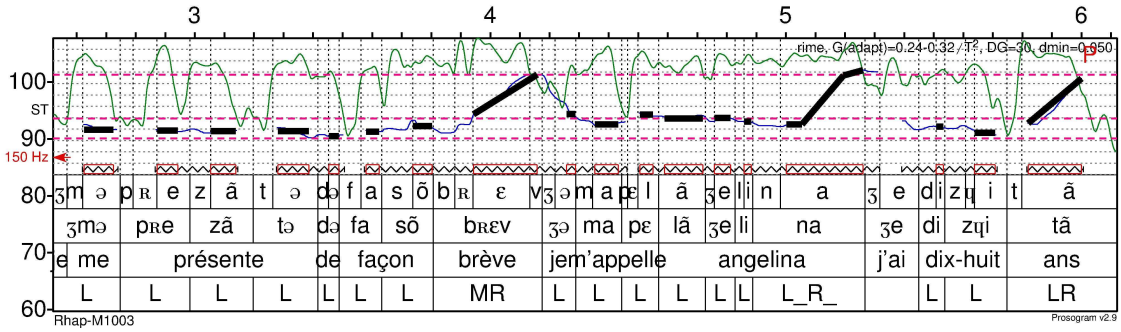

Figure 1. Automatic tonal annotation for the French utterances "*Je me présente de façon brève. Je m'appelle Angelina. J'ai dix-huit ans.*" [Rhap-M1003] ("*I briefly introduce myself. My name is Angelina. I'm 18 years old.*"), by a female speaker. The automatic prosodic labelling is shown in the lower 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*" (creak) indicates a syllable with creak (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).

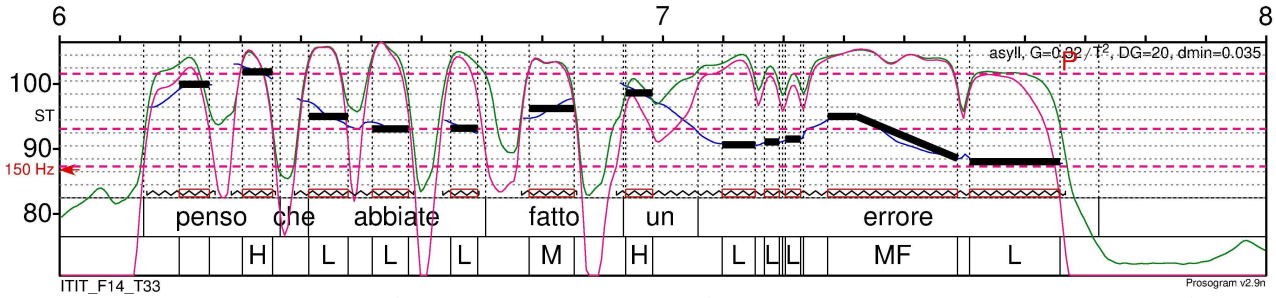

Figure 2. Automatic prosodic labelling of the utterance "*penso che abbiate fatto un errore*" ("*I think you made a mistake.*"), by a female speaker [ITIT_F14_T33] (OpenProDat corpus, [22]). The automatic labelling is shown in the lower tier. Acoustic parameters and pitch range are shown in the same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