Speech Prosody 2004
Nara, Japan
March 23-26, 2004

The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model

Piet Mertens

Department of Linguistics
University of Leuven, Belgium
Piet.Mertens@arts.kuleuven.ac.be

Abstract

This paper describes a system for semi-automatic transcription of prosody based on a stylization of the fundamental frequency data (contour) for vocalic (or syllabic) nuclei. The stylization is a simulation of tonal perception of human listeners. The system requires a time-aligned phonetic annotation. The transcription has been applied to several speech corpora.

1. Introduction

Speech corpora providing a transcription of prosody are essential to the study of prosody, to the development of intonation models, text-to-speech systems, and speech recognition systems. To this date, such corpora are fairly rare. One of the reasons is that there is little agreement about the desired nature of the transcription.

Which prosodic transcription should be used? One can distinguish three major types of prosodic transcription, each of which corresponds to a different stage in the processing of prosodic information in spoken language communication.

Acoustic analysis of the speech signal provides parameters such as fundamental frequency, intensity, voicing. Auditory analysis, i.e. manual transcription of pitch contours, involves introspection about short-term auditory memory. Finally, symbolic transcription, using a finite set of discrete symbols represents aspects of prosody that are assumed to be relevant for language communication.

a. *Symbolic transcriptions* use a small number of discrete symbols, indicating pitch levels, pitch movements (e.g. IPO standard pitch movements), tones (e.g. ToBI), boundaries, etc. These symbols are selected to note aspects of prosody that are assumed to be relevant for language communication; the transcription is reductionist. The number and nature of the symbols depends upon the model, resulting in a transcription which is model-specific. To our knowledge, to this date no system is able to obtain the prosodic transcription of a speech signal in a fully automatic way.

Since these transcriptions are usually made by hand or interactively using a computer program (e.g. starting from acoustic data), they require human intervention and therefore imply a certain amount of subjectivity.

b. In order to obtain an objective and quantified representation of prosody, it is possible to use *acoustic* information, in particular fundamental frequence (F_0). However, the interpretation of F_0 data is not straightforward. In order to interpret the F_0 curve in relation to the speech chain (i.e. the sequence of sounds, words...) its alignment with segmental data (spectral information, amplitude information, sound identity) is needed, something requiring phonetic expertise.

c. Acoustic analysis of prosodic phenomena became available only around 1950, with the introduction of the spectrograph and of specialized analog measurement instruments. Earlier linguistic studies on prosody were based on observations by ear by phoneticians and linguists. Obviously, *auditory transcription* is prohibitively time-consuming and requires capabilities that are rather uncommon among phoneticians and linguists (only musicians are trained to transcribe pitch). However, such transcriptions are based on auditory perception.

Given the various types of transcriptions, it is important to define the *objectives* for a prosodic transcription system.

1. The representation of prosody should be objective, robust, and easy to interpret. This transcription should represent perceived intonation: it should distinguish audible F_0 variations from inaudible ones, for individual syllables as for sequences of syllables.

2. The transcription should display pitch evolution over longer stretches of speech, in order to identify phenomena such as declination, onset pitch, register and register change.

3. The displayed pitch should be quantified to enable the estimation of melodical intervals at each level (local or global).

4. The temporal organisation of the speech signal should be preserved to identify and evaluate pauses and hesitations, to determine speech rate and to study rhythm (e.g. accelerations and decelerations).

5. Transcription should be automatic or semi-automatic.

6. The transcription should be neutral, i.e. model-independent, in order to enable its use by people with different theoretic backgrounds.

7. The transcription provides time-aligned phonetic transcription and text, for readability and text-based search.

8. The quantified representation of pitch and time autorizes manipulations, e.g. in resynthesis (PSOLA) and in synthesis, allowing the evaluation of the obtained transcription.

This article describes a prosody transcription system which meets these goals to a large extent. It is based on a stylization method proposed earlier [1]. Its particularity resides in the fact that it is based on a simulation of pitch perception and that it takes the syllabic nucleus as a basic unit. The proposed transcription will be called a *prosogram*, by analogy with the oscillogram and the spectrogram, which represent the evolution in time of the waveform and the spectrum, respectively.

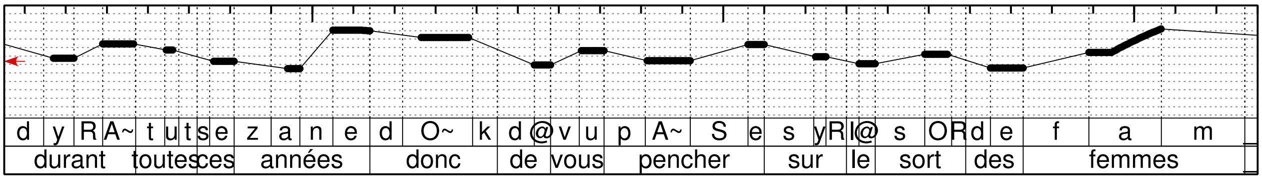

Figure 1. *Compact plain prosogram (using glissando threshold 0.32/T^2, cf. infra)*

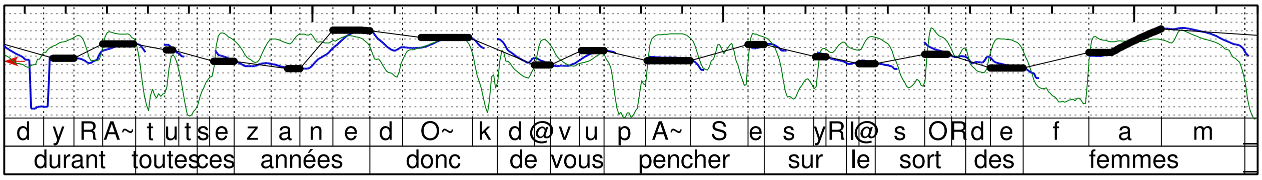

Figure 2. *Compact rich prosogram (using glissando threshold 0.32/T^2, cf. infra)*

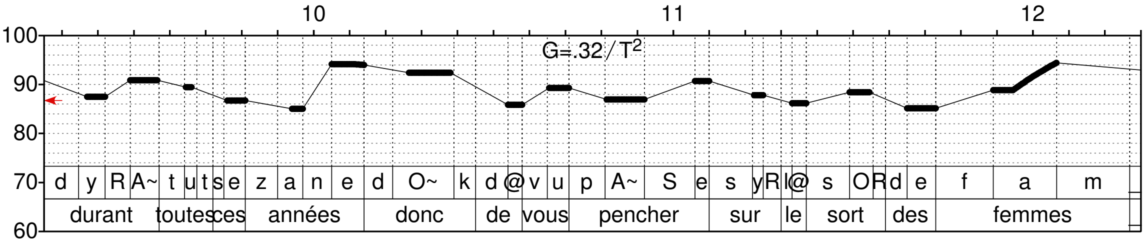

Figure 3. *Wide plain prosogram (using glissando threshold 0.32/T^2, cf. infra)*

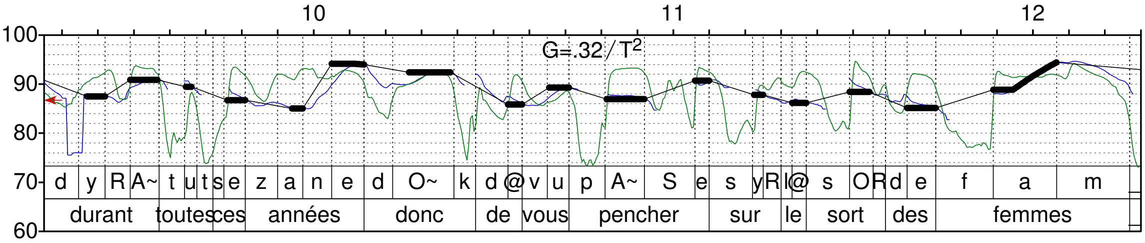

Figure 4. *Wide rich prosogram (using glissando threshold 0.32/T^2, cf. infra)*

Figures 1 to 4 illustrate the basic concept of the transcription: it's an estimation of the pitch contour perceived by the average listener, computed as a stylization of selected F_0 data. A prosogram consists of two parts: (1) the pitch contour and (2) one or more time-aligned annotations. The phonetic transcription (IPA, SAMPA or other formats) is required for the computation of the stylization (although it need not be shown necessarily). Other types of annotation (text, tones, stress type, prosodic units…) are optional.

Several variants are available for the part containing the pitch data. Whereas the *basic* prosogram shows only perceived pitch (thick line), the *rich* one adds F_0 (thin black line) and intensity (thin grey line). This information is presented in two formats: the *compact* format, intended for corpus transcription, and the *wide* format, which includes a calibration of time (in s) and pitch (in semi-tones). This results in four variants. The rich transcriptions provide more detailed information, including F_0 and intensity. The former (plotted on the same ST scale) is used to validate the stylization.

The parallel dotted lines are separated from one another by 2 semitones (ST), similarly to a musical score. These lines are used for interpreting pitch movements and melodic intervals either between syllabes, or within vowels.

For instance, a rising interval of 4 ST separates the first syllable of the word “pencher” from the next. Since both vowels (or syllables) are shown as a level line, they are perceived without internal pitch variation (according to the tonal perception model). On the other hand, for the syllabe “femmes”, which is shown as a rising line, the estimated size of the internal rise is about 8 ST.

The displayed pitch range (in ST, relative to 1 Hz) may be set to match that of the speaker. The small arrow on the left border indicates a reference value of 150 Hz (or 86.75 ST, relative to 1 Hz); it provides a key for the interpretation of the absolute pitch in the compact format. This value was chosen because it is likely to appear in the register of most speakers, both men and women.

2. Pitch perception

In spoken language communication, pitch contours are processed by human listeners, not by machines. The processing by the auditory system differs from the spectral analysis or pitch detection used in digital signal processing.

Psychoacoustics studies the relation between properties of acoustic stimuli and their effect in auditory perception. Several phenomena have been observed for pitch perception in speech. For details, see [1].

1. In order to be audible, a fundamental frequency (F_0) variation should have a minimal size; this size varies as a function of start frequency and stimulus duration (it decreases with duration). This *glissando threshold* has been measured for linear frequency variations in pure tones, synthetic speech, or resynthesized speech (to obtain a linear variation). ’t Hart [2] formulates a unified definition of the glissando threshold, where pitch change is expressed as an melodic interval in semitones, to eliminate start frequency. The “standard” threshold determined in psychoacoustic experiments for isolated vowels is $G = 0.16/T^2$ (ST/s), where T is the duration of the variation, cf. also [3]. (The semitone is a musical scale, in which an octave is divided in 12 equal intervals on a logarithmic scale.)

2. Of course, frequency variations in natural speech are rarely linear. So how does one perceive variations with slope changes? This question may be rephrased as follows: which slope changes are audible? To answer this, [1] introduce the notion of *differential glissando threshold* (DG). Slope changes are compared to this DG threshold; when the slope change is subliminal, the movement of both parts is replaced by a single linear variation starting at the beginning of the first part to the end of the second. (There has been little research on this threshold. The value used here is DG = g2 - g1 = 20 ST/s, where g1 and g2 indicate the slopes of the parts of the variation (in ST/s) on both sides of the slope change.)

There has been little research on the perception of complex variations: rise-fall, rise-level... (however, cf. [4, 5]). It is assumed here that if each of the parts is audible, the complex variation is perceived as the sequence of the constituting simple variations.

3. Up to now, we have only mentioned pitch variations in isolated sounds (typically vowels). However, in the speech chain the concatenation of sounds entails changes in intensity and voicing, as well as important spectral changes. *In most cases,* the alternation of vowels and consonants (or clusters) gives rise to an intensity and sonorance peak during the vowel, characterized by relative spectral stability. The vowel constitutes the syllabic nucleus then. The consonants, on the other hand, occur at the intensity dips and may give rise to spectral changes which are relatively rapid and important. The contrast between vowels and consonants is most clear for plosives and fricatives, whereas liquids, nasals and glides are similar to vowels. Major acoustic changes are thus located at syllable boundaries.

In fig. 1, the vowels [u] and [e] in “toutes ces” constitute energy peaks relative to the contiguous consonants. The intensity difference between [t] and [u], in “toutes ces années”, is larger than that between [e] and [z]. However, [a] and [m] in “femmes” have comparable intensity and the [m] has its own intensity peak.

Work by House [6] on perception of pitch change in speech shows that a same F0 variation is perceived differently depending on its location relative to syllable boundaries. If it appears within the vowel, the change is audible provided is exceeds the glissando threshold. If located partly on the transition at the syllabic boundary, only the part of the transition on the vowel will be perceptually integrated. This seems to indicate that simultaneous changes of intensity, spectral energy distribution and voicing impede the perceptual integration of pitch changes. This phenomenon becomes more important as acoustic variation increases. It results in a *segmentation of the pitch continuum* into elements corresponding to syllabic nuclei.

4. In continuous speech, sounds and syllables follow each other at a high rate and pitch information has to be processed in real time before it is masked by new sounds. Tonal perception is optimal for isolated vowels, it is worst for continuous speech at high rate of speech: the glissando threshold is higher in continuous speech. House [7] shows that pitch variations are perceived with more accuracy when followed by a *pause*. In other words, the presence of a pause after the variation lowers the glissando threshold.

3. Pitch stylization based on tonal perception

The term *stylisation* indicates a simplified pitch curve that is supposed to preserve pertinent or audible elements. The idea goes bach to the work of J. ’t Hart at IPO (Eindhoven).

Stylizations come in many flavors; we are interested here in (semi-)automatic procedures. The Momel system [8, 9] models the pitch curve by a quadratic spline function, as a sequence of parabolic segments. Other systems use linear regression to determine inflection points. Stylisation methods often are based on mathematical or statistical properties of the pitch curve.

[1] propose an approach based on a *simulation of tonal perception*. The parameters of the model are psycho-acoustic thresholds. The glissando threshold is applied to fragments of the pitch curve that are simple variations (rise, fall, or level, below differential glissando threshold).

An important issue is the selection of the unit used. The stylisation procedure may be applied to any voiced portion of the speech signal. By proper selection of the portion to be analysed, the segmentation effect described above is taken into account. The optimal unit seems to be the syllabic nucleus. It allows for (1) the localisaton and elimination microprosodic variations at vowel onset, (2) an adequate treatment of sonorant coda consonants (liquids, nasals, glides, voiced fricatives sonores).

4. Application to prosody transcription

The stylisation procedure can be used for transcription of prosody, in particular as a representation of perceived prosody.

Given the lack of an automatic segmentation of speech into syllabic nuclei based on perception, we adopt a pragmatic solution that consists in modelling pitch for *vowels* only. The information on vowel identity is provided by the phonetic annotation, which for each sound in the signal indicates its identity and the start and end times.

For each vowel the *vocalic nucleus* is determined. It is defined as the voiced part around the local intensity peak, and delimited by the points located at -3 dB (left) and -9 dB (right) from the peak. The value for the left boundary (-3 dB) eliminates most microprosody perturbations at syllable onset as well as microprosodic phenomena for voiced consonants at syllable boundaries; the value for the right boundary (-9 dB) preserves late pitch variations in stressed vowels. Obviously the results depend to large extent upon the quality of the phonetic alignement available.

In order to select the glissando threshold appropriate for continuous speech, stylisations obtained for different thresholds were compared to the manual transcription of a test corpus obtained earlier by two trained transcribers. Using $G = 0.16/T^2$, the stylization retains many intrasyllabic pitch glides not present iin the manual transcription. In other words, the stylization with the standard threshold overestimates the capabilities of the average listener. Using $G = 0.32/T^2$, i.e. twice the standard thresold, the stylisation is very near to the manual notation, as far as the distinction between glissando and level tone is concerned. As for the global pitch changes, extending over several seconds, the semi-automatic transcription appears to be more accurate than the manual one.

In order to validate the system, several speech corpora for which manual transcriptions by expert transcribers were available were analyzed (about 25 min).

The confrontation of the manual and automatic transcriptions allows for a study of their agreement. One can see to what extent the semi-automatic transcription is representative for the manual (or vice versa) and to what extent the prosogram reproduces the auditory image.

5. Conclusion

The approach for prosody transcription described in this paper has the following properties. It's a stylization of the F_0 curve of vowel nuclei, aiming at the reconstruction of the perceived pitch contour, based on a psycho-acoustic model of tonal perception. The transcription preserves temporal structure of the acoustic signal, includes annotations of text and phonetic transcription. The latter is used in the identification of vocalic nuclei. Pitch variations are shown on a pitch scale in semitones, for readability. The temporal alignement enables duration measurements of sounds and syllables, the localisation of pauses, and the study of speech rate and rhythm.

By comparison with approaches aiming at symbolic (e.g. tone) transcription, that retain a small inventory of units, the tonal stylisation is more detailed and avoids to take a stand about the nature and inventory of these abstract units (contour, tone, group, etc.).

Two transcription formats are used. The basic format retains only the pitch stylization aligned with the phonetic and text annotations. The rich format adds the F_0 curve (plotted on a musical scale) and the intensity data, for validation. Both formats can also be plotted in a compact size, e.g. for corpus transcription.

The tool has been used to transcribe 3 corpora (5 speakers; 2 male, 3 female). The results demonstrate the robustness of the transcription, its similarity to manual transcription, and the similarity of the contours with those used in text-to-speech synthesis [10]. The stylisation has been validated in an informal way using resynthesis.

Several improvements and extensions are considered, such as a graphical representation of rate of speech, or the automatic adjustement of the pitch range on the basis of the distribution of F_0 values. The vocalic nucleus should be replaced by the syllabic nucleus as the basic unit for stylization. The latter could be computed on the basis of acoustic properties or using a syllabification starting from the phonetic transcription.

For more information, see
http://bach.arts.kuleuven.ac.be/pmertens/prosogram .

6. References

[1] d'Alessandro, C.; Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9(3), 257-288.

[2] Hart, J. 't., 1976, Psychoacoustic backgrounds of pitch contour stylisation. *I.P.O. Annual Progress Report* 11, 11-19.

[3] d'Alessandro, C.; Rosset, S. ; Rossi, J.-P., 1998. The pitch of short-duration fundamental frequency glissandos. *J. Acoust. Soc. Am.* 104(4), Oct 1998, 2339-2348.

[4] Rossi, M., 1978a. La perception des glissandos descendants dans les contours prosodiques. *Phonetica* 35(1), 11-40

[5] Rossi, M., 1978c. Interactions of intensity glides and frequency glissandos. *Language & Speech* 21, 384-396

[6] House, D., 1990. *Tonal Perception in Speech*. Lund: Lund University Press.

[7] House, D., 1995. The influence of silence on perceiving the preceding tonal contour. *Proc. Int. Congr. Phonetic Sciences* 13, vol. 1, 122-125.

[8] Hirst, D.; Espesser, R., 1993. Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15, 75-85.

[9] Campione, E.; Hirst, D.; Véronis, J., 2000. Stylisation and symbolic coding of F0: comparison of five models. In *Intonation: Analysis, Modelling and Technology*, Botinis, A. (ed.). Kluwer Academic Publishing, 185-208.

[10] Mertens, P. ; Goldman, J-P.; Wehrli, E.; Gaudinat, A., 2001. La synthèse de l'intonation à partir de structures syntaxiques riches. *Traitement Automatique des Langues* 42(1), 145-192.