PROSOTRAN: a tool to annotate prosodically non-standard data

Katarina Bartkova [1]*, Elisabeth Delais-Roussarie* [2], *Fabian Santiago-Vargas*[2]

[1] ATILF-UMR 7118, Nancy Université, France
[2] LLF-UMR 7110, Université Paris-Diderot, France

katarina.bartkova@atilf.fr, elisabeth.roussarie@wanadoo.fr, rotinet@hotmail.com

Abstract

Assigning a prosodic transcription that encompasses all prosodic phenomena (intonation, accentuation and phrasing) is difficult mainly because: (i) encoding all the prosodic phenomena usually supposes a knowledge of the language to transcribe; and (ii) a representation of the various phenomena cannot be achieved without taking into account the three prosodic parameters. In this paper, we present a tool, PROSOTRAN, which automatically assigns to each utterance a multi-tiered transcription that symbolically represents how the three prosodic parameters (F_0, duration & energy) do vary over time. Assigning labels to each syllable avoids segmenting the signal into linguistic units that are difficult to define when the language to transcribe is not known.

Index Terms: prosodic annotation systems, automatic annotation tools, phonetic implementation and phonological analysis

1. Introduction and problematic

Any prosodic transcription system, be it semi-automatic or manual, aims at providing a symbolic representation of various prosodic events that occur in the speech signal. As such, it is almost compulsory to carry research in prosody as it allows comparing, searching and quantifying what occurs in the data. So, transcription systems and annotation tools are of great importance as they facilitate the analyses of the data and can help in the development of new phonological modeling.

Among all the prosodic events occurring in the speech signal, a transcription should pay specific attention to the events that may be linguistically relevant. This latter condition explains partly why the transcription task, which consists in assigning a label to an interval (or a point in time), is particularly difficult when the data to transcribe are produced in a language or a dialect not described before (i.e L2 learners' productions, etc.). Therefore, the choice of the segmentation units as well as the choice of the labels should be done in such a way as to allow determining from the speech signal the phonological units and principles at play in the language/ dialect.

At present, several prosodic transcription systems are used, but they usually display some limitations (see [1] for a detailed description of the various systems and their limitations):

- Almost all the existing systems do not allow representing different types of prosodic events (phrasing, intonation and accentuation). In fact, most of them focus on intonation (and sometimes on the relation between tonal patterns and phrasing), relying mostly on an analysis of the melodic variations at the phonetic level. INTSINT, for instance, provides a transcription of tonal variations that occur in an utterance, by assigning directly labels to turning points that have been determined after a stylization of the pitch track (see [2]). As such, this system only gives a symbolic transcription of changes in pitch, without taking into account duration and intensity, which contribute to the categorization of some other prosodic events. As for manual systems, ToBi (*Tone and Break Indexes*), which has been originally developed for American English (see [3]), has mostly been designed to account for intonational patterns and phrasing, leaving aside all prosodic events related to the metrical structure.
- Even if it is not always explicit, many transcription systems assume that the phonological system of the language to transcribe is known. Consider, for instance, the IPA (or SAMPROSA), which has the advantage of encoding a wide array of prosodic phenomena (stress, accentuation, intonation, and phrasing). Firstly, the assignment of the labels presupposes a segmentation of the speech signal into units that are clearly phonological in nature (*tone groups* and *tone units*). In addition, the labels related to accentuation and stress cannot be assigned if one doesn't know which syllable is primarily or secondary stressed.
- Many transcription systems are developed in a specific theoretical paradigm (the British school of Intonation of the IPA, the metrical-autosegmental framework for ToBi, etc.). As a consequence, the segmentation units and the labels are more or less defined on the bases of the underlying paradigm. In the various systems developed in the metrical-autosegmental framework (ToBi or IViE), labels consist on tonal forms and are only assigned to stressed syllables and to prosodic phrase boundaries.
- Many transcription systems, and more particularly manual ones, do not allow achieving great agreements between transcribers. On the one hand, the units to which labels are assigned are not always defined in a rigorous way, which leads to great differences among transcribers (see [1] for concrete examples). On the other hand, the choice of the labels relies crucially on the level of transcription taken into consideration. Note, however, that some systems (i.e. ToBi) are not very clear regarding the level at which transcription is done. On that matter, IViE, which has been primarily developed to analyze intonational variations in English dialects (see [4] among others), provides a clear distinction between a phonetic and a phonological level of transcription: the tonal transcription is based on the auditory interpretation of the signal, instead of F_0 alone; and phonological interpretation follows auditory interpretation, so the intonational/prosodic system does not need to be fully understood for the transcription to be possible. These characteristics of the system come from the fact that it has been developed to analyze dialects of unknown phonological systems.

As just shown, most of the existing systems have some limitations that make them difficult to use (i) to encode all the prosodic events that occur in speech, and (ii) to annotate data uttered in a language/ dialect that has not been phonologically described before. These limitations are even more acute when the systems are used to transcribe non-standard data such as L2 learner speech or pathological speech.

An attempt is made here to develop a new annotation tool called PROSOTRAN. It should allow overcoming some of the limitations we observed in the existing systems. In this paper, our aim is twofold:

- to present PROSOTRAN and the symbolic labeling it generates, the latter being directly linked to the way the three prosodic parameters (duration, energy and F_0) are changing over time;
- to give an idea of how such a tool can be used to analyze non-standard data, and more precisely L2 learner productions, which are usually difficult to annotate with the most commonly used transcription systems. In the case of L2 learner data the difficulties are due to the fact that the underlying phonological system is unstable and unknown.

The paper is organized as follows: In the second section, PROSOTRAN is presented. We explain how the symbolic labels used to represent changes in duration, pitch and energy are computed and assigned. In the third section, concrete examples and uses of the tool are presented.

2. PROSOTRAN: description and output

PROSOTRAN is an annotation tool that has been designed in such a way as to overcome some of the limitations mentioned in section 1. It provides a multi-tiered transcription, in which each tier is associated with a single prosodic parameter (duration, intensity, F_0, etc), and represents how this parameter varies over time. In the next section, the main characteristics of the system will be presented. Then, we will explain the way the labels are computed from the signal.

2.1. Main characteristics

A prosodic transcription can be seen as a discrete symbolic representation of the linguistically relevant prosodic events occurring in the signal. Three major difficulties have to be faced while developing such a transcription system: problems related to the segmentation of the speech signal, problems related to the choice of the labels, and problems related to what is linguistically relevant in the data and how to represent it. As shown in section 1, many existing systems have not succeed in overcoming these problems, as they have chosen segmentation units and labels that either presuppose that the language to transcribe is known, or rely on some strong theoretical assumptions. In order to solve some of these problems, we have made some specific choices in developing PROSOTRAN.

In transcriptions generated by PROSOTRAN, the units of segmentation to which labels are assigned consist in vocalic nuclei (i.e. in syllables). By taking these units as basic unit, the system chooses a prosodic unit that can be easily defined and is universally recognized (see, among other, [5]). This departs from what is done in many existing systems. In many cases, the units taken as segmentation units for speech description are generally subject to controversy, as they do not allow (i) representing the prosodic events in all their complexity (these events can participate to phrasing, to accentuation or to intonation); nor (ii) indicating how the various prosodic parameters are changing over time. Units such as tone groups, accentual phrases or intonational phrases, which are related to prosodic and linguistic structures, can be good candidates to account for phrasing and intonation, but their extension is either theory-dependant, or language dependant. Turning points are also good candidates to represent variations in pitch, and thus to provide an intonational transcription; in addition, they are less language or theory-dependent; but they cannot be used to account for changes in duration, etc. Among all the possible units, the syllable is thus the only one that is quite neutral theoretically and that can be used to represent changes in duration, pitch and intensity.

Concerning the labels, there are roughly two ways to determine which symbols or labels to use: they can be associated to a function (as for the distinction between stressed and unstressed syllables) or to a phonetic event. If labels or symbols are directly related to a phonological function, they are difficult to use to annotate productions in languages/ dialects for which the prosodic system is not known (see for instance the distinction between primary and secondary stress in the IPA, which cannot be used accurately if the metrical system is not known). On the contrary, if symbols are associated to a form or a phonetic event, they will mostly use a single parameter (variation in pitch) and will not be sufficient to account for phonological phenomena such as phrasing, which is realized by a wide range of phonetic events (syllable lengthening, rise in pitch, change in pitch direction, etc.). In order to overcome these problems, PROSOTRAN generates for each utterance a multi-tiered transcription, in which each tier is associated to a specific prosodic parameter (duration, energy, etc.) and contains symbolic labels that account for the variation of the given parameter in every syllabic nucleus. In addition, the labels are determined by the acoustico-phonetic representation associated with the signal, and also by psycho-acoustic knowledge such as glissando threshold (see [6]). The advantages of such an approach are twofold:

- By being pluri-parametric, the generated transcriptions can help distinguishing prosodic phenomena that are realized by durational cues as well as by melodic variation at the acoustico-phonetic level;
- By taking into account only psycho-acoustic knowledge and acoustico-phonetic information, the system can be used to annotate data uttered in a language that has not been described before.

To summarize, we can say that the only assumption made in developing PROSOTRAN is the idea that any prosodic event that has a phonological status in a language is realized at the phonetic level by changes in pitch, in duration, or in intensity. As such, the system has the advantage of providing a symbolic representation of the various phonetic parameters without making too strong theoretical or phonological assumptions. In consequence, it can be used to describe languages whose phonological systems are unknown. Moreover, it allows distinguishing phonological errors from differences in phonetic implementation.

2.2. Data Processing

PROSOTRAN has been used to annotate three comparable passages in French and Spanish extracted from EUROM 1. The Spanish extracts have been read aloud by 10 native speakers, the French extracts by 6 native speakers, and by 6 Spanish learners of French (their level in French ranging from A2 to B1 according to the CEFRL).

At the acoustico-phonetic level, acoustic parameters such as F_0 in semi-tones and log energy are calculated from the speech signal every 10 ms using the acoustic analysis Aurora (see [7]). At the phonetic-linguistic level, the orthographic transcription associated with the speech signal is converted into phonemes (either automatically (French data) or manually (Spanish data)).

The speech signal is then segmented into phoneme units using the CMU sphinx speech recognition toolkit [8]. The forced alignment between the speech signal and its phonetic transcription provides phone duration as well as the duration of the pauses. The automatic segmentation has been verified by expert phoneticians for all the data of our corpus, in both French and Spanish.

Synchronization between the phoneme units and their acoustic parameters (F_0 and log energy values) is carried out and prosodic parameters are calculated for every relevant syllabic nucleus. Note that, during synchronization, the pitch track has been modified when zero values occurred in isolation. The latter have been replaced by values obtained by interpolation between the values of the previous and subsequent point.

2.3. Automatic analysis of the prosodic parameters and label assignments

From the forced alignment between the speech signal and its phonetic transcription and the calculation of the prosodic values (F_0 and energy) associated to several points in the signal (see section 2.2 for more details), PROSOTRAN generates for any utterance a multi-tiered transcription that can be read with PRAAT. In this transcription file, each tier contains a symbolic representation that account for the way the various prosodic parameters vary from one nucleus to the following.

2.3.1. Duration

The temporal axis of the speech signal is represented by vowel durations. Using solely vowel durations allows avoiding the problem due to the variability in syllabic structure, as vowel duration can be considered as more homogeneous and therefore more representative of the speech rate variation than syllabic duration. The vowel duration is compared to the mean duration and associated standard deviation of the vowels occurring in non-final positions (i.e. not at the end of a word or before a pause). This way of doing should hopefully discard vowel durations lengthened on prosodic boundaries. Each vowel duration is then compared to the mean vowel duration and associated standard deviation calculated on the speech signal between 2 pauses. When the speech signal between 2 pauses does not contain a sufficient number of vowels at non boundary positions to estimate the speech rate (current minimum threshold is set to 5 vowels), than the vowel durations are compared to the mean duration and the standard deviation calculated on the whole speech signal.

A label is assigned to each nucleus (or syllable), and it accounts for its lengthening rate in comparison to other nuclei. When the duration found for a vowel is equivalent to the mean vowel duration, the latter is considered as neither lengthened nor reduced, as a result no label is assigned. By contrast, when its duration is longer from the mean duration and once the standard deviation, it is seen as long (and encoded as *[long]*), from mean duration plus twice the standard deviation, it is considered as very long (encoded *[+ Long]*), and from mean duration and three time the standard deviation, it is encoded as extra-long (*[+ XLong]*). By contrast, if the value observed is reduced from once, twice or three time the standard deviation, it is considered respectively as reduced, very reduced and extra-reduced (*[reduced], [+ reduced]* and *[+Xreduced]*).

2.3.2. F_0 Height and F_0 Slope

Regarding the changes in pitch, three different types of information are provided for each syllable: the height and the direction of the pitch movement (rising, falling, etc.) and the importance of the slope.

To encode F_0 height, a melodic range is calculated between the maximum and the minimum values of F_0 in semi-tones observed in the speech signal for an entire speech file. The obtained range is then divided into 6 zones (from L1 to L6) after the calculation of the median value of the distribution of the F_0 values . For each vowel, the height is calculated on three distinct points that correspond to the beginning, the middle and the end of the vowel. Encoding the height in three distinct points has the advantage of showing exactly the temporal alignment of any pitch movement. From the information relative to the F_0 height, it is very easy to annotate for every syllable the direction of local pitch movements, by comparing the height on a given syllable with the height obtained in the previous and the following syllables. The pitch movements are encoded with the symbols H (for rising), L (for falling) and HL or LH (for a complex movement occurring on a single syllable).

A second label is related to the steepness of the F_0 slope itself. It is computed from the F_0 slope associated with a given vowel and with the previous vowel. The variation in F_0 obtained is compared to the glissando threshold ($0.16/T^2$) (see [6]) and is annotated symbolically between $\text{Vowel}_{\text{Slope}}$++++ (very strong slope, rising or falling) and $\text{Vowel}_{\text{Slope}}$- (flat).

2.3.3. Energy

The energy value of the sound is calculated as the mean value of the log energy of the frames (10 ms shift) of the phoneme. The mean energy is calculated only for vowels as the speech intensity is carried mainly by vowel segments. Differences in intrinsic vowel energy exist between [+ high] and [+ low] vowels (see [9]), for this reason the vowel energy calculation contains a normalization of the energy values. The energy value of every vowel is compared to the mean energy and associated standard deviation calculated on all the vowels of the speech signal comprised between two pauses and the difference between a current vowel energy value and this mean value is coded in a symbolic way. Thus the energy annotation is situated on a continuum comprised between VowEner---- (very low vowel energy) and VowEner++++ (very high vowel energy).

2.3.4. Perspectives

Although it is not implemented in the present study, a post-processing module, applied on the symbolic annotation, can be trained to decide whether a certain association of symbolic annotations will occur at prosodic boundaries or not or whether a certain association of parameters will occur in stressed or rather in unstressed syllables.

3. Concrete example: Prosotran and L2 learner's production

The transcription obtained in interrogative and declarative sentences are studied here to see which regularities emerge.

3.1. Annotation and tonal patterns observed in declarative questions

The prosodic annotation obtained in declarative questions produced by Spanish learners of French displays some differences when compared to the realizations of native speakers (see (1a), (1b) and (1c)). Pitch level starts usually higher in the productions of the Spanish learners, see (1a) and (1b). Moreover, phrasing boundaries at the end of the word "*reservation"* is not clearly marked in (1a). In this case, the realization is comparable to what is expected in Spanish declarative sentence where pitch level is usually high at the beginning, followed by a high plateau stretching over three to six syllables, then F_0 is declining steadily on the rest of the utterance reaching its lowest point on one of the last three syllables, before rising again on the last syllable (see for a complete description [10] and [11]). By contrast, the realization annotated in (1b) has already some characteristics observed in French, in particular in the marking of a prosodic phrase boundary at the end of "*reservation".*

(1) *Vous prenez les réservations par téléphone ?*

a. produced by a learner of French (level A2)

		vu	pr@	ne	lE	re	sEr	va	sjON	par	te	le	fon
F_0	Pitch height	L3	L4	L3	L3	L3	L3	L3	L3	L2	L2	L2	L4
	Slope		H	L						L			H

b. produced by a learner of French (level B1)

		vu	pr@	ne	lE	re	sEr	va	sjON	par	te	le	fon
F_0	Pitch height	L3	L4	L4	L4	L4	L3	L3	L4	L4	L4	L4	L5
	Slope		H				L		H				H

c. produced by a native speaker of French

		vu	pr@	ne	lE	re	sEr	va	sjON	par	te	le	fon
F_0	Pitch height	L2	L3	L4 (3)	L3	L3	L3	L3	L5	L4	L3	L3	L5
	slope			M	D				M	D	D		M

The annotations generated by PROSOTRAN allow comparing productions of learners and of native speakers. It can then be used to clearly describe the patterns used by the various speakers.

3.2. Annotation, phrasing and prosodic cues

The transcriptions obtained for all the sentences of one passage are studied for French and Spanish. This allows determining which syllables are realized as prominent, or at least differently than the surrounding syllables. In French, almost all syllables located at the end of a lexical word (verb, noun, adjective and adverb) are realized at a higher pitch level than the following syllable. This change in pitch height is accompanied by syllable lengthening in many cases, and sometime by a clear glissando. These prosodic events occur almost only in word final positions, and can be interpreted as cues for phrasing boundaries. When duration and pitch are both involved, the degree of the prosodic break is more important. These observations correspond to what is usually said in studies dedicated to phrasing in French ([12]).

In Spanish, by contrast, durational cues do not seem as important. The prosodic realizations observed in the French sentences produced by Spanish learners do confirm this hypothesis. Syllabic lengthening does not occur as regularly as in the productions of native speakers.

4. Conclusions

In this paper, we have presented PROSOTRAN, a tool that generates multi-tiered transcription that account for the way the three prosodic parameters vary over time in a given utterance. In developing this tool, we have made some choices to avoid some drawbacks observed in many transcription systems such as their tendency to focus on a specific category of prosodic events or parameters (intonational events or F_0), their use of linguistic knowledge in segmenting and labeling the speech. Further work is now carried out in order to evaluate precisely the advantages and limitations of PROSOTRAN in comparison to the existing systems. It is done by annotating various types of data, and by modifying the way the various labels are computed and assigned.

5. References

[1] Delais-Roussarie, E. and Post, B., "Corpus annotation: methodology, systems and reliability", in J. Durand, U. Gut and G. Kristoffersen [eds], *Handbook of Corpus Phonology*, Oxford university Press, to appear.

[2] Hirst, D.J., Di Cristo, A. and Espesser, R., "Levels of representation and levels of analysis for intonation", in M. Horne [ed], *Prosody: Theory and Experiment*, 51-87, Kluwer Academic Publishers, 2000.

[3] Beckman, M. E., Hirschberg, J. and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in S.-A. Jun [ed.], *Prosodic Typology: The Phonology of Intonation and Phrasing*, 9-54, Oxford University Press, 2005.

[4] Grabe, E.; B. Post & F. Nolan (2001). "Modelling intonational variation in English: The IViE system", in S. Puppel & G. Demenko (eds.) *Prosody*, 51-58, Adam Mickiewicz University Press, 2000.

[5] Segui, J., "The syllable: A basic Perceptual Unit in Speech Processing?", in H. Bouma and D.G Bouwhuis [eds], *Attention and Performance: Control of Language Processes*, 165-181, Lawrence Erlbaum Associates, 1984.

[6] Rossi, M., "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole ", Phonetica 23, 1-33, 1981.

[7] ETSI ES 202 212 V1.1.1, STQ. "Distributed speech recognition; Extended advanced front-end feature extraction". 2005

[8] Mesbahi, L., Jouvet, D., Bonneau, A., Fohr, D., Illina, I. and Laprie, Y., "Reliability of non-native speech automatic segmentation for prosodic feedback", Proceedings of SLaTE 2011, 2011.

[9] Di Cristo, A., De la microprosodie à l'intonosyntase, Thèse d'état, Université de Provence, 1985.

[10] Sosa, J.M., La entonación del español, Cátedra, 1999.

[11] De la Mota, C., Butragueño, P. M. and Prieto, P., "Mexican Spanish intonation", in P. Prieto and P. Roseano [eds], Transcription of Intonation of the Spanish Language, Lincom Europa, 2010.

[12] Di Cristo, A., A propos des intonations de base du français. Unpublished ms, 2010.