Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Prosodic Trees for Boundary Detection in ASR in French

Natalia Segal, Philippe Martin, Katarina Bartkova

France Télécom R&D, Lannion, France
EA333, UFR Linguistique, University Paris-Diderot, France
{natalia.segal; katarina.bartkova}@orange-ftgroup.com;
philippe.martin@linguist.jussieu.fr

Abstract

Prosodic trees as a hierarchical representation of prosodic organization in French proved to be efficient for automatic processing of continuous speech. We applied this technique to the prosodic boundary detection on the output of a speech recognition application in order to test whether prosodic boundaries of different levels in tree confirm or not recognition hypotheses.
Two types of tree construction algorithms were tested: one using lexical information (word hypotheses), and another using only phonemic information (phoneme hypotheses). Both were successively used on the automatic alignment output ("perfect recognition" conditions) and on the ASR application output for the same spontaneous speech database so as to compare their applicability.

1. Introduction

Prosodic boundary detection can be useful for the task of ASR, as it can eliminate certain recognition hypotheses [3] and improve the performance of the system. Prosodic boundary detection demands either a theoretical linguistic description of phonetic and phonological specifics of a language intonation [2, 5, 8] or an approach based on machine learning techniques [4]. The problem of the first approach is that most of the intonation theories are based and tested on prepared speech and so its applicability to spontaneous speech processed by ASR needs a confirmation. The problem of purely probabilistic approaches is their limited application.

Our representation of French spontaneous speech intonation makes use of its theoretical description in the form of a prosodic tree conceived for prepared speech [5]. We used this theory as a base (with some necessary adjustments due to automatic data processing) and verified its applicability to spontaneous speech and the possibility of its application to improve the ASR system performance.
Two approaches to the prosodic tree construction have been tested in the present study: the first one uses assumptions made on word category (word-based) and the second one uses only the segmental transcription of the speech signal (phoneme-based). The first approach gives a significant advantage for the correct detection of prosodic group boundaries because it allows using some lexical constraints. Still, it can be observed that in case of multiple errors made in word assumptions, as is often the case in ASR output, lexical constraint becomes inconvenient. Since our principal research interest lies in the ASR domain, we decided to elaborate an alternative algorithm for prosodic tree construction using no lexical knowledge for prosodic group detection and then to compare these two approaches on the speech recognition data.

2. Prosodic groups and trees

2.1. Intonation theory for prepared speech in French

The intonation theory used in the implementation relies on the existence of a prosodic structure organizing prosodic groups (stress groups) hierarchically [5]. The F0 characteristics of these groups' stressed syllables indicate the prosodic structure through the use of a contrast of melodic slope. The prosodic structure is a priori independent and associated to the syntactic structure, each structure having its own set of constraints.

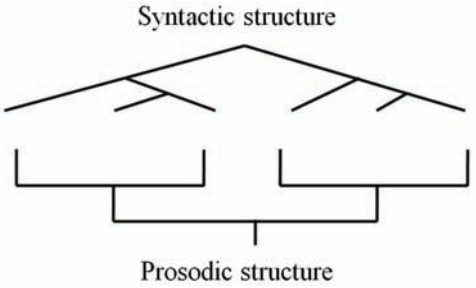

Figure 1: *Association of the prosodic and syntactic structures.*

2.1.1. Prosodic words

There is a general agreement to look on or around the accented (stressed) syllable for prosodic phenomena. Minimal prosodic units – prosodic words - contain one (lexical) stress and one optional initial stress. A minimum prosodic unit contains one or more content word (open class word), and optional grammatical words, constrained by a 7 unstressed syllable rule.

2.1.2. Prosodic structure

The prosodic structure organizes hierarchically the prosodic words and is not level limited. Prosodic words have no pre-established standard pattern, as their melodic characteristics depend on the application of 2 rules:

IMS: Inversion of Melodic Slope rule
AMV: Amplitude of Melodic Variation rule

Initial (secondary) accents do not play a role in the marking of the prosodic structure, and are therefore normally described with a melodic rise. Their role is only to ensure the presence of at least one stress in sequences of 7 consecutive syllables.

10.21437/SpeechProsody.2008-24

2.1.3. Intonation syntax association

In a phonosyntactic approach, prosodic structure (PS) is assumed to be independent but associated to the syntactic structure (SS). In general, more than one PS can be associated to a given SS, the final choice being governed either by syntactic congruence or eurhythmy (balanced of the number of syllables at each level of the prosodic structure).

Prosodic structure specifics:

- The prosodic structure organizes hierarchically minimal prosodic words (stress groups).
- Prosodic markers indicate the prosodic structure of the sentence.
- Grammars of prosodic markers are specific to every language.
- Specific realizations of prosodic markers characterize various dialects.

The association between the syntactic and the prosodic structures is not straightforward, even in prepared speech. The constraints of this association can be summarized as follows:

- Planarity (no tangled structures).
- Connexity (no floating segments).
- Stress clash (no consecutive stressed syllables if the implied syntactic units are dominated by the same syntactic node).
- Syntactic clash (no prosodic words grouped in the PS which are not themselves grouped in the SS by the same node, so at the lowest level in the structure)
- Stress group maximum number of syllables (a sequence of 7 syllables has at least one stress – either emphatic (narrow focus) or lexical, the number 7 depending on speech rate)
- Eurhythmy (balancing the number of syllables in the prosodic structure, generally at the expense of congruence with syntax)
- Neutralization (phonological features that are not necessary to encode a given PS are not necessarily realized).

2.2. Implementation for spontaneous speech

The main points of the construction of a prosodic tree for spontaneous speech had been presented in [9].

One of the crucial points in the algorithm is the detection of prosodic groups (the smallest prosodic units used as the leaves in the prosodic tree construction). In [9], only one solution for prosodic group detection had been proposed, based on word category hypotheses (*word-based* or *lexical* approach).

As we discussed above, prosodic group (also called rhythmic group or prosodic word) is the smallest prosodic unit in French. The division into prosodic groups is not completely voluntary, there being lexical and rhythmic constraints on the stress emplacement [2, 5]. Since we try to develop an approach for prosodic group detection without any lexical constraint, the only information we can make use of is prosodic parameters on phonetic segments and rhythmic constraints (*phoneme-based or phonemic approach*).

The principal feature of a prosodic group in French is its final stressed syllable, which is mainly marked by vowel lengthening and usually also by a pitch movement (rising or falling). In order to find prosodic group boundaries we must compare for each vowel both its length and F0 variation. One of the possibilities of taking them into account simultaneously is the use of a glissando threshold [6, 7, 10] for every vowel pitch movement.

The rhythmic constraint is usually expressed by imposing a maximum for the number of syllables per prosodic group. This maximum is reported to be either 7 or 8 syllables [2]. We have chosen the 8 syllables constraint as it produced better results on our data (most of the greater prosodic groups usually resulting from some errors of processing) [9].

In the following section we will describe the prosodic group detection algorithm in details.

2.2.1. Prosodic group detection

To detect stressed syllables, both vowel length and F0 amplitude (expressed in semitones, ST) were used.

For the detection of a perceptible pitch variation an auditory threshold known as the *glissando threshold* (*G*) was used, traditionally expressed in semitones per second (ST/s) [6, 7, 10]. This threshold changes with duration as the perception of F0 variation amplitude depends on its length. Minimal perceptible amplitude decreases with increasing duration of the stimulus. The glissando threshold measured in psychoacoustic experiments using isolated short stimuli, either pure tones or speech-like signals, was reported as $G = 0.16/T^2$ (*T* being the duration of the variation in seconds) [10]. This threshold is considered to be even higher for continuous speech where a prosodic variation needs to be more prominent to be perceived: $G = 0.32/T^2$ [6]. We adopted the later value for our further experiments, as it worked better on our continuous speech data.

In order to determine which of the vowels are stressed, all the vowels between two pauses were ranged according to their lengths. The vowels were then divided into two groups: stressed and unstressed. To do so, we calculated the glissando threshold for each vowel starting from the shortest. Once a vowel was found for which its F0 variation was greater than the glissando threshold, we considered it and all the longer vowels as stressed.

When calculating glissando we also used correction coefficients for vowel length according to vowel type in order to take into account the intrinsic differences. The coefficients were used as reported to be in [1], though with a somewhat rougher division providing smaller number of different vowel classes.

Moreover, two *rhythmic constraints* due to the principle of eurhythmy were also used. The first constraint limited the number of successive unstressed syllables to 7 (8 syllables being the maximum for a prosodic group). The second constraint limited the number of successive stressed syllables: the second consequent stressed syllable was, where possible, attached either to the previous or to the next prosodic group.

Excessively long prosodic groups were divided into smaller groups according to vowel length and glissando threshold, but also to the position of the vowel in the group. Thus, we chose as additional stressed vowel either the longest or the closest to the glissando threshold, the one which was closer to the middle of the group (to adhere to the eurhythmy principle).

As for isolated stressed syllables, they were added either to the previous or to the following prosodic group where it

was possible. Most commonly, isolated stressed syllable was stuck to the previous prosodic group as in fact the syllable continued the previous group's final F0 movement. When it was added to the following group, it was mostly due to the expressive stress which is often placed on the first syllable of a prosodic group in French. When neither of those options was possible, the syllable was considered as being an independent prosodic group.

2.2.2. Prosodic trees

Prosodic trees were constructed based on detected prosodic groups, in the same way as was described in [9], following the rules of Amplitude of Melodic Variation (**AMV**) and Inversion of Melodic Slope (**IMS**).

Thus, the prosodic structure of a phrase was represented in a hierarchical form: as a prosodic tree of different length and depth, with prosodic groups as its leaves. Particular values of prosodic parameters on the final syllables of prosodic groups reflect this structure, such as pitch amplitude and direction and segmental duration.

3. Speech database and evaluation

3.1. Speech database

The speech database used for the method evaluation contains the results of a customer satisfaction survey and is constituted of more than 1080 telephone messages in French. Every message is considered to be pronounced by a different user which gives an approximate number of speakers (male and female). The length of messages varies considerably, with an average of 54 words per message.

3.2. Evaluation for continuous speech

In the first place, word-based and phoneme-based approaches for prosodic boundary detection were opposed using the data of automatic alignment. The database used was manually transcribed in orthographic form with the annotation of non-speech noises as well as interrupted words and filled pauses. The orthographic representation was then phonetically transcribed and aligned with the speech signal. Though this alignment still includes some errors, nevertheless they are fewer than errors of speech recognition output; therefore, forced alignment technique can be used to compare the performance of the two approaches in what can be called "perfect recognition" conditions.

Firstly, we compared, for both approaches, the error rates for prosodic word detection. This part of the prosodic trees construction is particularly important because it provides initial prosodic boundary emplacements (used for the further tree construction).

About 100 files were chosen among the sufficiently long ones (well representing continuous unprepared speech). These files were parsed manually into prosodic words by experts for the evaluation of the two approaches (with and without any lexical knowledge). Even if the manual prosodic parsing remains expert dependent and can vary between different experts, it is still the most reliable prosodic group detection we can get. The comparison between the error rates for the two methods is reported in Table 1.

Table 1: *Error rate in prosodic word detection for the automatic alignment.*

	Lexical approach	Phonemic approach
Detected	2365	2992
Inserted	261	618
Omitted	657	387
Recall (%)	76	86
Precision (%)	89	79
F_1-measure	82	82

As expected, since the second approach makes use of no lexical constraint, it shows slightly less accurate performance for the precision (more borders inserted). Though precision is considered to be a priority for us, so as not to sort out good recognition hypotheses, accepting its degradation for the second approach allows the drop of lexical constraint, which is very convenient for the speech recognition. General accuracy, however, remains stable, compensated by better recall for phonemic approach and the performance of the system doesn't degrade dramatically.

A further speculative analysis of errors for the phonemic approach showed that among the wrongly placed boundaries (that is inserted) 160 (26%) were located in the middle of a lexical word. Another 458 (74%) were put at the end of a lexical word that wasn't actually stressed.

The second test applied to the two algorithms was an attempt to detect interdependence between the levels of prosodic boundaries in the prosodic tree and their error rates. The original supposition was that boundaries of higher levels in the tree would tend to be better detected than the less prominent boundaries of lower levels. The aim is to give every prosodic boundary a weight expressing a degree of confidence which could be useful for the speech recognition.

Thus, we checked the distinct error rates for different levels in the tree. The results presented on Figure 1 show a correlation between the level of the boundary and its error rate (percentage of inserted boundaries) for the 3 highest levels for both approaches. This difference becomes insignificant for the level 4 and deeper.

It is worth mentioning that the difference between level error rates is linear for the approach using word hypotheses, whereas for the phonemic approach the difference is quite pronounced between first and second levels and less so between second and third levels.

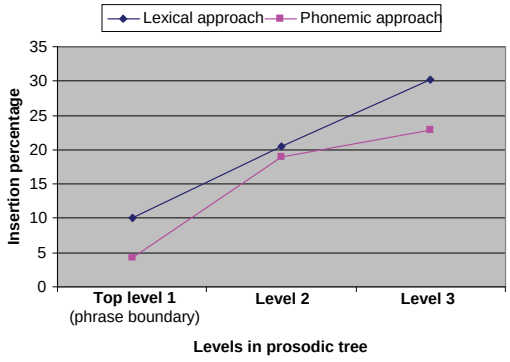

Figure 1: *Error rates for different prosodic levels.*

These preliminary tests applied to the automatically aligned database showed the applicability of both methods to spontaneous speech in the condition of forced alignment. This proved that our theoretic basis and hypotheses were appropriate for this kind of data.

We proceeded then with the actual recognition data so as to see how it will affect the boundary detection performance.

3.3. Evaluation for ASR data

To verify the performance of the two prosodic trees algorithms for a possible improvement of the speech recognition system, we applied them to the same database of spontaneous speech but using this time phonemic transcription and word hypotheses provided by an ASR system. We also used the indications for the recognition evaluation stating for every word whether it has been correctly recognised, substituted or inserted.

For the evaluation we used the same 100 files with manual prosodic word segmentation. A boundary provided by one of the automatic algorithms was considered as correct if it was placed on the same syllable as a manual boundary (even if the word was not correctly recognised). This way it was possible to compare how the ASR output affected the boundary detection performance compared to the automatic alignment.

In the Table 2 are given the error rates for both prosodic tree construction algorithms applied to speech recognition results.

Table 2: *Error rate in prosodic word detection for ASR output.*

	Lexical approach	Phonemic approach
Detected	2765	3016
Inserted	639	724
Omitted	635	469
Recall (%)	77	83
Precision (%)	77	76
F_1-measure	77	79

It can be seen that precision is considerably lowered for the lexical approach as compared to alignment data, whereas the precision of phonemic approach is less affected. This makes the general performance for the second approach slightly better than for the first one.

In order to test whether prosodic words detection could actually improve recognition results, we also compared, for the phonemic approach, the percentage of correct boundaries placed in the middle of wrongly recognised words to the percentage of boundaries wrongly placed in the middle of well-recognised words (Table 3).

Table 3: *Distribution of boundaries in the middle of the words.*

	Correct boundaries in the middle of a wrong word	Wrong boundaries in the middle of a correct word
Phonemic approach	139	47

Only prosodic boundaries placed in the middle of a correct word can possibly damage the recognition performance, and these are particularly few (only 47 of the total of 3016 boundaries).

There also has been established a correlation between prosodic boundary level and its error rate, similar to that for forced alignment (v. Figure 1).

4. Conclusions

The present study confirmed the possibility of applying a theoretical description of French intonation system to the development of an automatic intonation model detecting prosodic boundaries and prosodic structures in spontaneous speech.

Two different approaches for automatic prosodic boundaries detection – one using lexical structure and another using only phonemic structure – were tested, and both proved to be sufficiently well-adapted to spontaneous speech processing.

Although the application of the two algorithms to recognition data somewhat deteriorates their performance, both can eventually be used to improve the recognition results, as errors in prosodic boundary detection affect less well-recognized words than they do wrongly recognized ones.

There also has been established a correlation between the levels of prosodic tree nodes and the boundary detection accuracy. Thus, it is possible to improve the precision of boundary detection by attributing a degree of confidence to the boundary according to its level in prosodic tree.

In general, the results of this study open various possibilities of improving the performance of ASR system by using prosodic clues in post-processing of recognition output.

5. References

[1] Di Cristo, A., 1980. La durée intrinsèque des voyelles du français. In *Travaux de l'Institut de Phonétique d'Aix*, vol. 7, 211-23.

[2] Fónagy, I., 1979. L'accent en français : accent probabilitaire. In *L'accent en français contemporain*. I. Fónagy; P. Léon (ed.). Paris: Didier, 123-233.

[3] Kompe, R., Prosody in speech understanding systems. J. Siekmann, J. G. Carbonell (ed.). Springer-Verlag New York, Inc.

[4] Langlais, P., 1997. Estimating prosodic weights in a syntactic-rhythmical prediction system. *Proceedings of Eurospeech'97*. Rhodes, 1467-1470.

[5] Martin, Ph., 1987. Prosodic and rhythmic structures in French. *Linguistics* 25, 925-949.

[6] Mertens, P., 2004. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. *Proceedings of Speech Prosody 2004*. Nara (Japan), 23-26.

[7] Rossi, M., 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica* 23, 1–33.

[8] Rossi, M., 1999. *L'intonation. Le système du français : description et modélisation*. Gap: Ophrys.

[9] Segal, N.; Bartkova, K., 2007. Prosodic structure representation for boundary detection in spontaneous French. *Proceedings of ICPhS'2007*. Saarbrücken, 1197-1200.

[10] 't Hart, J., Collier, R., and Cohen, A. 1990. *A Perceptual Study of Intonation.* Cambridge U.P., London.