INTERSPEECH 2009 BRIGHTON

Determining intonational boundaries from the acoustic signal

Lourdes Aguilar[1], *Antonio Bonafonte*[2], *Francisco Campillo*[3], *David Escudero*[4]

[1]Universitat Autònoma de Barcelona, [2]Universitat Politècnica de Catalunya,
[3]Universidad de Vigo, [4]Universidad de Valladolid, Spain

Abstract

This article has two-fold aims: it reports firstly the improvement of a speech database in Catalan for speech synthesis (Festcat) with the information about prosodic boundaries using the break index labels proposed in the ToBI system; and secondly, it presents the experiments undergone to determine the acoustic markers that can differentiate among the break-indexes. Several experiments using different classification techniques were performed in order to compare the relative merit of different attributes to characterize breaks. Results show that the prosodic phrase breaks are correlated with: presence of a pause, lengthening of the pre-break syllable and the F0 contour of the span between the stressed syllable and the following post-stressed, if there are, immediately preceding the break.
Index Terms: ToBI Breaks, prosodic analysis

1. Introduction

It is generally acknowledged that so as to make real progress in understanding prosodic mechanisms that dominate the spoken communication between humans one needs large, prosodically labelled corpora. Nevertheless, prosodic labelling of corpus is a time consuming process which requires a high level of expertise. This makes it difficult to generate massive corpora which are required to use prosody in speech technology. Even speech synthesis systems, which have traditionally modelled prosody, tend to rely on implicit mapping between text and acoustics without analyzing the *meaning* of each particular prosody instance.

One of the goals of the authors is to develop reliable tools to automatically annotate prosody. These tools could be widely used. The first step in this research line was to label a subset of the Festcat database ([1]) using the Cat_TOBI proposal ([2]). But due to the fact that a ToBI-based system is a fine-grained labelling which requires native well-trained transcribers and the supervision of experts to achieve reliable and consistent databases, we decided to start by the so-called break-tier, marking the placement of prosodic boundaries and their degree of perceived strength in the framework of metrical and intonational phonology ([3]). Since evidence for prosodic phrasing can come from either tonal or segmental phenomena, or from both, we do not start from a purely phonetic determination of constituent boundaries, but we adopt the perception-based approach in determining boundary location proposed in the ToBI system ([3], http://www.ling.ohio-state.edu/ tobi/). Catalan, as does every language that has been described in the ToBI framework, has tones that are anchored to phrase edges, and these phrases are an important component of the metrical structure in the language ([2]). However, any phonological model of intonation is not complete without an accompanying description of phonetic realization: a means to map contrastive events onto the phonetic space in time and in frequency. Although some studies have examined some phonetic properties of intonational boundaries in Romance languages ([4]), the contrastive levels of prosodic grouping have not been fully described.

This article reports firstly the improvement of a speech database in Catalan for speech synthesis (Festcat) with the information about prosodic boundaries using the break index labels proposed in the ToBI system; and secondly, the experiments undergone to determine the acoustic markers that can differentiate among the break-indexes and to compare the relative merits of different information in a hierarchical model.

2. Experimental procedure

2.1. Corpus

The corpus is part of the one developed to provide with two Catalan voices the speech synthesis system Festival ([1]). Two speakers have recorded 10 hours of speech each one. The database was designed with exhaustive combinations of Catalan segments in different types of texts (news, transcriptions of the parliament, etc.) to have the highest phonetic coverage. All the utterances were segmented in words, syllables and allophones. The process of segmentation, as well as the extraction of F0 curve, were done automatically.

In this study, 1397 sentences (27866 words) were extracted from the following domains: transcriptions of the parliament (150 sentences), monologues from literary plays (318 sentences), novels (446 sentences) and phonetically enriched sentences (483 sentences). All sentences are declarative read at a normal rate with a natural pronunciation. The total time labelled is of 3h30m per each speaker.

2.2. Labelling procedure

The Festcat sentences have been annotated by means of the ToBI-based annotation conventions to signal break indexes: as in other ToBI systems, the procedure is perceptually-based, although the labeller has visual information of the signal. Without making strong claims about the existence of theoretical units such as prosodic word, phonological phrase or intermediate phrase (differently to the Cat_ToBI proposal in [2]) we have decided to use all the levels in Table 1 to better capture the relationship among prosodic constituents. In order to mark the absolute end of the elocution, at the end of the file, the level 5 proposed in [5] has been added. This decision has two advantages: first, in declaratives, it serves as the minimum F0 value of the declination baseline; second, it prevents the processing of the silences in this position, without any linguistic content.

All the speech data (1397 sentences $\times$ 2 speakers) was labelled manually by a graduate in linguistics, with no prior training in prosodic labelling. She received a two-hour session on the basics of the AM theory, and some written information about the core concepts (prominence, prosodic structure, units of phras-

Table 1: *Break descriptions*

Break	Description
0	Any clear example of cohesion between orto-graphic forms, such as vowel contacts
1	Any inter-word juncture (provided as default at every word boundary)
2	End of groups with some sense of disjuncture with respect to the following speech chunk
3	End of minor prosodic group
4	End of major prosodic group

ing). The transcriber was looking at a computer screen with a display of the signal (F0 curve and waveform) together with the phonetic marks corresponding to words, syllables and silences. Nevertheless, she was encouraged to attend preferable to perception. To ensure the consistency of the data, only one transcriber was recruited and one of the authors (L.A.) reviewed the corpus. The annotation has not been considered definitive until the transcriber and the reviewer arrived to a consensus in the labels.

Table 2: *Distributional analysis of the boundary types*

BI0	BI1	BI2	BI3	BI4	BI5
510	21,900	556	2,727	1,548	1,393

Table 2 shows the distributional analysis of the boundary types for the male speaker. For the analysis, we consider a three-level hierarchy and therefore map breaks 0-1-2 to *non break*, 3 to a *minor break*, and 4 to a *major break* (excluding absolute final breaks). Roughly, 10% of the words are followed by a minor break and 10% of the words are followed by a major break (BI4 and BI5) To control possible interspeaker variation, due to limited space, only data in time and F0 corresponding to the male speaker will be presented.

3. Determining break indexes using acoustic features

Several experiments using different classification techniques were performed in order to compare the relative merit of different attributes to characterize breaks. In this section, first we define the considered acoustic features and then we address the problems of detecting breaks and discriminating BI3 type versus BI4 type. The analysis is limited to the boundaries after content words, as a preliminary analysis has shown that there are no boundaries after function words. Furthermore, the corpus data was filtered to avoid some problems caused by the pitch detection algorithm.

3.1. Features considered

In the experiments reported here, the features selected include only information that can be extracted from the signal and the sentence, that is, corresponding to the time and frequency domains. This way, we took into account features such as the duration of different speech chunks (silence after the boundary, syllables around the accent, etc), several F0 values (maximum values, nucleus of the accented syllable, beginning and end of the accent group, etc), To model the intonation contour at the phrase edges, values concerning the last syllable of the word have also been considered.
Table 3 depicts the features that were finally chosen after the experiments that are described in section 3.2.

Table 3: *Features description*

Name	Description
D_PAUSE	Duration of the silence after the break
D_PPSYL	Duration of the last syllable before the break
D_PPSYLNORM	D_PPSYL normalised by the intrinsic duration of the phonemes in the syllable
F0E	Last F0 value found in the last sonorant before the break
F0_DIFF	Difference between the mean F0 value at the nucleus of the accented syllable and F0E
F0max_nextWRD	Maximum F0 value of the word following the break
F0max_lastSYL	Maximum F0 value found in the last syllable before the break
F0_RESET	Difference between F0max_lastSYL and the maximum F0 value in the following words before the next accented syllable
F0_DECLIN	Difference between maximum f0 value in the sentence and F0max_lastSYL

3.2. Detection of breaks

Following a similar approach to the one described in [6], we first select the acoustic features that are more relevant for discriminating the breaks and second we test the performance of a decision tree for classifying the breaks in terms of the observed acoustic features.

Table 4: *Relevance of the attributes to discriminate break versus no-break*

Attribute	Gain Ratio	Info Gain
D_PAUSE	0.62296	0.5891
D_PPSYLNORM	0.10843	0.2652
D_PPSYL	0.06212	0.1801
F0_DIFF	0.02838	0.0646
F0max_nextWRD	0.0235	0.0399
F0_RESET	0.02032	0.032
F0max_lastSYL	0.01707	0.0414
F0_DECLIN	0.01117	0.0133
F0E	0.00768	0.0108

Table 4 shows the ranking of attributes for the break versus no-break problem within the data from the male speaker. The metrics *InfoGain* and *and GainRatio* are computed to rate the attributes. $InfoGain$ evaluates the worth of an attribute by measuring the relation $H(Class) - H(Class|Attribute)$. $GainRatio$ relates $InfoGain$ with respect to $H(Attribute)$. $H(Attribute)$ measures the entropy of the entities classified in terms of $Attribute$. We used WEKA data mining tools[1] to compute the metrics.
The features D_PAUSE, D_PPSYLNORM, F0_DIFF were used as questions in a decision tree using a CART algorithm. With

[1] http://www.cs.waikato.ac.nz/ml/weka/

only three decisions, the tree illustrated in Figure 1 correctly classifies 7006 out of 7793 potential boundary sites, an overall success rate of 89.9%. The tree was obtained after balancing the data, decimating the non-break type values.
The first node indicates that there is a break after a pause (silences < 8ms. are not considered pauses from a perceptual point of view, since the listener cannot perceive them; if they are encountered in these data is due to the appearance of short occlusions), whereas the second node identifies a boundary when the F0 contour significantly falls (for reference, the mean F0 value of the speaker is 181 Hz, $\sigma = 32$ Hz). In the other cases, there is a break if the last syllable lengthens its intrinsic duration. The condition about extreme lengthening (> 1.59) applies mostly to segmentation errors.

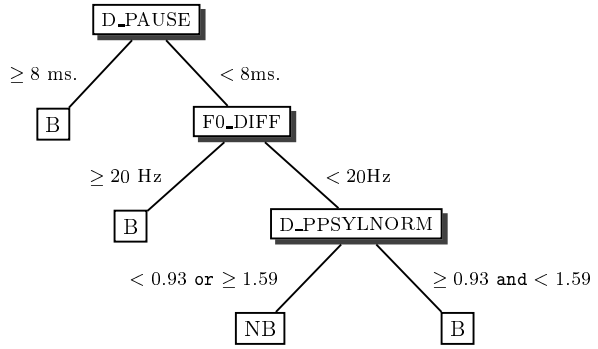

Figure 1: CART tree for classifying breaks (B: BI3 and BI4) versus no-break (NB)

3.3. Classification of major and minor breaks

For the BI3 versus BI4 problem we repeated the experiment obtaining Table 5. A CART algorithm was trained with the four best ranked attributes F0E, D_PAUSE, F0_DIFF and D_PPSYLNORM, giving an overall rate success of 79.2% (2869 correctly classified instances, 753 incorrectly classified instances). Nevertheless, a closer look at the confusion matrix indicates that BI4 are correctly identified in 556 sites, but incorrectly classified in 522 sites. This can be explained to some extent by the fact that in the data set, the number of instances of BI3 is significantly higher than the number of instances of BI4: 2544 BI3 versus 1078 BI4. Alternatively, we can decimate the data set, as we have done with the problem of discriminating break-type versus non-break type. In this case, the overall success results are worst: 1771 correctly classified instances (75.4%), 579 incorrectly classified instances (24.6%), but the results of predicted BI4 improve (789 out of 1078 are correctly identified), while the results of predicted BI3 remains (982 out of 1272 are correctly identified). We hypothesize that the subdivision of BI3 and BI4 breaks into L or H types will improve the confusion observed.

4. Analysis of acoustic features

Once we have identified the most relevant acoustic features, the role of each attribute is discussed in the next sections.

4.1. Time domain

According to the results of the decision trees performed in previous sections, the presence of a pause and the lengthening of the final syllable seem to be the most relevant features to determine if a break occurs after a word, independently of the degree of strength of this break. In the normal junctures between words (BI1) a pause cannot appear, and the duration of the syllables perform as if they were not before a BI (the factors affecting syllabic duration are well described in other studies [7, 8]).
Moreover, it has been shown that the duration of the pause plays a significant role to discriminate between BI4 type and BI3 type. Firstly, the pause is almost always associated to BI4 (99.55%) while only in 70.30% of the cases the BI3 precedes a pause. Secondly, when appears, the pause associated to the BI3 is shorter than the pause associated to the BI4: a mean of 256 ms. (133) versus 451 (211).
With respect to the duration of the syllable preceding the break-index, syllables become noticeably longer immediately prior to a prosodic break and that the degree of final lengthening can differentiate break-index 3 from break-index 4. The lengthening of a syllable located before a BI3 is of 45% ($\mu = 1.45$, $\sigma = 0.42$), while the lengthening of a syllable located before a BI4 is of 62% ($\mu = 1.62$, $\sigma = 0.39$).
By itself, the duration of the pause can account for the majority of the potential boundary site classifications for BI3 and BI4. However, when the duration of the pause is shorter than a given threshold (in our data, 452 ms), another temporal factor is combined to give the listener the perception of a strong boundary: if the degree of lengthening is relevant (the duration of the syllable is greater than 366 ms in this data set), a BI4 is identified. On the contrary, when the temporal cues are not evident (with a pause duration shorter than 452 ms, and a syllable duration shorter than 366 ms), the most prominent acoustic cue is the scaling of the final contour: if the last F0 value found immediately before the BI (F0E) is low (< 79 Hz), the break index is classified as BI4; in the rest of the cases, more acoustic cues are needed, referred to the duration of the pause and the duration of the syllable preceding the break index.

Table 5: *BI3 versus BI4 gain ratio and information gain*

Attribute	Gain Ratio	Info Gain
D_PAUSE	0.1356	0.3288
F0_DIFF	0.0446	0.0884
F0E	0.0425	0.0561
D_PPSYLNORM	0.0338	0.0511
D_PPSYL	0.0522	0.0404
F0_DECLIN	0.0499	0.0196
F0_RESET	0.0149	0.0186
F0max_nextWRD	0.0238	0.0175
F0max_lastSYL	0.0273	0.016

4.2. Frequency domain

We will now turn to the frequency data set. The decision tree makes use of the F0 contour of the span between the stressed syllable and the following until the break, using the variables F0E and F0_DIFF.
The F0 data concerning to the non-break type cases are not observed, because it is assumed that there are other factors different from the boundary classification that determines their F0 values, and that it is necessary a complete modelling of the F0 in the sentence. On the contrary, the scaling of the F0 value immediately preceding the break-index, that is, F0E, can discriminate between BI3 and BI4, with means of 93 Hz ($\sigma = 16$)

and 73 Hz ($\sigma = 17$), respectively. This means that in the great majority of cases, the melodic curve ends higher when a BI3 follows the word than when a BI4 does. This could be interpreted as the presence of a rise, or a continuation rise versus the presence of a fall. Nevertheless, if we observe the distribution of F0E preceding a BI3, or BI4, it is inferred that values are distributed into two Gaussian distributions. This can be due to the fact that BI3 and BI4 can both use either ascending or descending patterns (the types L- and H- and L% and H% respectively, in ToBI terminology).

Another feature refers to F0_DIFF, defined as the difference in Hz between the maximum F0 value in the stressed syllable and the F0 value immediately prior to the break index (F0E). The following descriptives are obtained: a mean of -2 Hz ($\sigma = 13$) for the data set corresponding to BI3, and a mean of -14 Hz ($\sigma = 18$) for the data set corresponding to BI4. The differences are statistically significant as revealed by a paired t-test. These results are interpreted as a tendency to an ascending F0 trajectory before a BI3 and a descending trajectory before a BI4. Figure 2 shows more clearly this pattern.

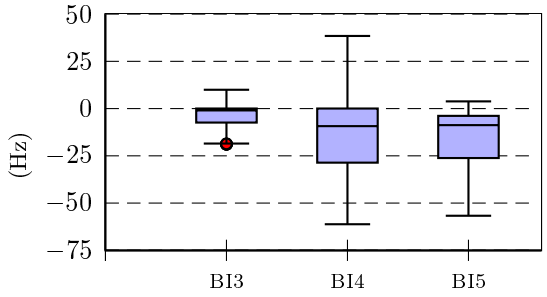

Figure 2: Box-Plot for F0_DIFF

5. Discussion and conclusions

From the data, it is concluded that in Catalan, prosodic phrase breaks are discrete events that are associated with the following acoustic cues: presence of a pause, lengthening of the pre-break syllable, F0 contour of the span between the stressed syllable and the following until the break (measured in terms of the variables F0E and F0_DIFF). These acoustic correlates are looked to predict from the signal if there is a boundary site, because they are present only when a prosodic break occurs in the sentence. The data then confirm the three-level hierarchy of *non break*, *minor break*, and *major break*.

Prosodic phrase breaks differ from one another in the phonetic realisation of the acoustic correlates identified: the duration of the pause (longer if it associated to a major group, that is, BI4), the degree of lengthening of the pre-break syllable (greater if associated to a major group) and the intonational tune type (continuation rises, or ascending tones, mostly associated to minor groups, that is, BI3). In other words, it is not the selection of different acoustic correlates but their phonetic manifestation which differentiates the prosodic strength between units of speech. A good illustration of this is the pause: we can say that the appearance of a pause is obligatory to mark a major prosodic group, but optional to mark a minor prosodic group. Consistently, when appears, the pause associated to BI3 is always shorter than the pause associated to BI4.

Furthermore, in the time domain, it is important to note that the well-known process of final lengthening before a pause also applies before a break that is not correlated with a pause.

The main drawback of the approach presented in this article is that the series of high and low tones (and more complex tones) occurring at the boundaries are not considered. We have attempted an approach directly mapping the F0 scaling and dynamics to the break-index types, but for a complete description we will require the labelling of boundary tones, to which include distinctions between rises, falls and continuation rises. Another promising area is to characterize the shape of the stress groups around the potential boundaries.

Moreover, attending to classification rates, the algorithms described above should be complemented with linguistic features such as part-of-speech tags, position in the utterance, punctuation signs and distance to previous breaks.

Future research will concentrate on the prediction of prosodic boundaries from text both for corpus labelling and for prosody generation. The organization in prosodic groups, or chunks, helps the listener to better understand the synthetic message, and at the same time, the appropriate location of the boundaries is needed to achieve natural-sounding TTS outputs. This is not a trivial problem: the melodic movements give clues about where is the most relevant information in the utterance (signalling focus, for example) and the location of silences gives the listener enough time to process speech.

6. Acknowlegments

Partially founded by the Ministerio de Ciencia e Innovacion, Spanish Government AVIVAVOZ project TEC2006-13694-C03, Glissando project FFI2008-04982-C003-02, Xunta de Galicia "Isidro Parga Pondal" research programme and PGIDIT05TIC32202-PR programme. The work done by L. Aguilar was possible thanks to the visiting position at the Universitat Politècnica de Catalunya during the academic year 2008-09.

7. References

[1] Bonafonte, A., Adell, J., Esquerra, I., Gallego, S., Moreno, A. and Pérez, J (2008), "Corpus and Voices for Catalan Speech Synthesis", Proc. of the Sixth International Language Resources and Evaluation (LREC' 08).

[2] Prieto, P., Aguilar, L., Mascaró, I., Torres-Tamarit, F. and Vanrell, M.M. (2008). "L'etiquetatge prosòdic Cat_ToBI". Estudios de Fonética Experimental XVIII, pp. 287-309.

[3] Beckman, M., Hirschberg, J. and Shattuck-Hufnagel, S., (2005) "The original ToBI system and the evolution of the ToBI framework", Prosodic Typology and Transcription: A Unified Approach, ed. by Sun-Ah Jun. Oxford University Press, Oxford.

[4] D'Imperio, M., Elordieta, G., Frota, S., Prieto, P. , and Vigário, M. (2005) "Intonational phrasing in Romance: the role of syntactic and prosodic structure", Prosodies, ed. by S. Frota, M. Vigário & M. J. Frietas, Berlin & New York, Mouton de Gruyter, 59-97.

[5] Price, P., M. Ostendorf, Shattuck-Hufnagel, S. and C. Fong, C. (1991). "The use of prosody in syntactic disambiguation", Journal of the Acoustical Society of America, 90: 2956-70.

[6] C. V. Wightman and M. Ostendorf (1994). "Automatic Labeling of Prosodic Patterns", IEEE Transactions on Speech and Audio Processing, Vol. 2 No.4

[7] Van Santen, J. (1994), "Assignment of segmental duration in text-to-speech synthesis", Computer Speech & Language, Vol. 8, No.2

[8] Febrer, A., Padrell J. and Bonafonte A. "Modeling Phone Duration: Application to Catalan TTS", Proceedings of the Third ISCA Speech Synthesis Workshop, Jenolan Caves Australia, Nov. 1998.