

# Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System

Fabio Tamburini

CILTA/DEIS - University of Bologna  
Piazza S. Giovanni in Monte, 4, I-40124, Bologna, Italy  
f.tamburini@cilta.unibo.it

## Abstract

This paper presents work in progress on the automatic detection of prosodic prominence in continuous speech. Prosodic prominence involves two different phonetic features: pitch accents, connected with fundamental frequency (F0) movements and syllable overall energy, and stress, which exhibits a strong correlation with syllable nuclei duration and mid-to-high-frequency emphasis. By measuring these acoustic parameters it is possible to build an automatic system capable of correctly identifying prominent syllables with an agreement, with human-tagged data, comparable with the inter-human agreement reported in the literature. This system does not require any training phase, additional information or annotation, it is not tailored to a specific set of data and can be easily adapted to different languages.

## 1. Introduction

The study of prosodic phenomena in speech is a central topic in language investigation. Speakers tend to focus the listener's attention on the most important parts of the message, marking them by means of such phenomena. As outlined in Beckman & Venditti [4], a precise identification of such phenomena helps to disambiguate the meaning of some utterances. It is also a fundamental step for the automatic recognition of spontaneous speech [9], and enhances the fluency and adequacy of automatic speech-generation systems. Moreover the construction of large annotated language resources, such as prosodically tagged speech corpora, is of increasing interest both for research purposes and for language teaching.

One of the most important prosodic features is prominence: a word or part of a word made prominent is perceived as standing out from its environment [24]. A better understanding of how prominence is physically accomplished is a basic step in the construction of tools capable of automatically identifying such phenomena. This paper presents work in progress on the construction of a system for the automatic detection of prosodic prominence in speech using only acoustic parameters and cues.

Following Beckman's [3] phonological view, further developed by Bagshaw [1, 2], syllables that are perceived as prominent either contain a pitch accent or are somehow "stressed". On the acoustic/phonetic side, the accomplishment of such features has to be strictly correlated with acoustic parameters. As well as the works already cited, there are many studies [16, 17, 18], suggesting that some of the main acoustic correlates of prominence are pitch movements, strictly connected with fundamental frequency (F0), overall syllable energy, syllable duration and spectral emphasis.

The work presented here is operationally divided into two separate steps: the first step involves the automatic identification of syllable-nuclei boundaries to reliably measure the duration feature, while the second one concerns the identification of prominent syllables by means of acoustic measurements.

The data set used in these experiments is a subset of the DARPA/TIMIT acoustic-phonetic continuous speech corpus enriched with manually added prominence annotations.

Several studies have been conducted in this field for building automatic systems capable of reliably identifying either one acoustic correlate of prominence [6, 8, 11] or a complete set of prosodic parameters [2, 7, 25]. These latter studies, involved in the construction of a complete prosody identification system, rely on additional phonetic information such as phone labelling and/or utterance transcriptions. Such systems, based on Hidden Markov models, neural networks or similar models, require a training phase in order to work properly on new, unseen data. This way of processing data requires as an additional resource an adequately segmented and labelled speech corpus; this resource might not be available, would certainly be very expensive to build, and, moreover, permanently binds the system to one specific language. The aim of this study is to derive some methods for the reliable tagging of prominence, avoiding any training phase and the use of additional resources.

Despite the quantity and quality of studies on this topic, it seems that the automatic and reliable detection of prosodic prominence is still an open question.

## 2. The acoustic parameters

In the following subsections, each acoustic parameter involved in this study is considered. All acoustic parameters must be normalised to some extent to avoid the natural variations among different speakers. The specific normalisation procedures applied to each parameter will be described.

### 2.1. Duration

The studies of prosodic prominence listed above tend to consider syllable duration as one of the fundamental acoustic parameters for detecting syllable stress. Unfortunately the automatic segmentation of the utterance into syllables is a complex task; in [10] we can find a survey of syllable segmentation algorithms. None of these methods seem to perform well when applied to continuous speech. For these reasons, an alternative duration measure for prosodic prominence detection should be introduced.

One possible measure seems to be the duration of syllable nucleus. Considering some utterances taken from the TIMIT corpus and comparing the duration of the syllable nucleus

with the duration of the entire syllable, with respect to prominence, and approximating the logarithm of these measures with a gaussian distribution, it is possible to obtain the distributions in figure 1. The two sets of distributions look qualitatively very similar and the separation between the two classes remains almost the same using the two measures. Moreover, building two gaussian discriminators using the distributions in figure 1 and classifying a set of test syllables with them, with respect to prominence, we obtain almost the same ratio of correct classifications. The exact classification performance is not important in this context as this duration measure is only one parameter useful to build the prominence detector. The relevant conclusion, interesting for this study, is that we can reliably substitute the syllable duration measure, rather difficult to obtain with automatic procedures, with the measure of syllable nucleus duration, that can be automatically obtained more easily.

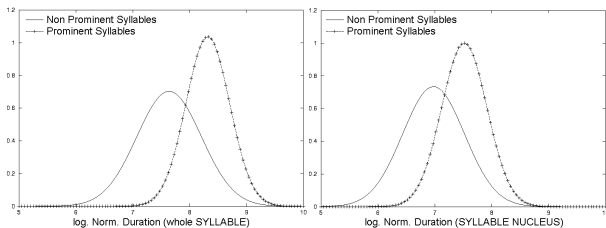


Figure 1: Gaussian approximation of duration measures: whole syllable (left), syllable nucleus (right).

Using a modified version of the convex-hull algorithm presented in [12], and applying it on the utterance energy profile in the band 300-900 Hz as suggested in [10], it is possible to reliably identify the syllable nuclei in the utterance and measure their duration to obtain the acoustic parameter needed for subsequent computations and to identify the nucleus positions to correctly assign prominence values (see the following sections).

This duration parameter is normalised, considering the mean duration of the syllable nuclei in the utterance. This is a standard technique for ROS (Rate-Of-Speech) normalisation, as described in [13].

## 2.2. Energy

The second acoustic parameter is overall syllable nucleus energy. It can be computed in various ways. Here I refer to RMS energy. Overall nucleus energy is normalised dividing it by the mean energy over the utterance syllable nuclei. This reduces the energy variation across different utterances and different speakers.

## 2.3. Fundamental frequency (F0) contour

The extraction of F0 contour, or pitch contour, is typically a complex task. Bagshaw [2] carried out an accurate comparison of different algorithms for fundamental frequency estimation. Most of the complexity of this process resides in post-processing optimisation of the contour. Stops and glitches often tend to distort the contour, introducing spurious changes in the profile and artificial maxima or minima. A post-processing procedure to smooth out such variations is often required in order to obtain reliable results. The Praat speech package [5] contains useful routines for fundamental

frequency determination as an effective set of post-processing functions. Removing octave jumps, smoothing, pitch lowering compensation at the end of the utterance and interpolation are common post-processing operations that can be successfully applied using the Praat package, also through its scripting additions.

## 2.4. Spectral emphasis

It has been shown, especially by the influential work of Sluijter & van Heuven [16], that mid-frequency emphasis is one useful parameter in determining stressed syllables. Each nucleus segment has been bandpass-filtered through FIR filters dividing it into three bands: from 0 to 500 Hz, from 500 to 2000 Hz and from 2000 to 4000 Hz. The RMS energy of each segment/band pair was computed. Examining the distributions of prominent and non-prominent syllable energies in the frequency bands considered, we find that the two bands 0-500 Hz and 2000-4000 Hz show a clear overlapping between prominent and non-prominent syllables, while the central band from 500 to 2000 Hz exhibits a clear separation between the two syllable categories. These results confirm a strict dependence of syllable prominence to vowel mid-frequency emphasis.

## 3. PROSODIC PARAMETERS

This section examines the prosodic quantities that are the object of the study: stress, pitch accent and prominence.

### 3.1. Stress

The main correlates of syllable stress indicated in the literature are syllable duration and energy [1, 2, 17, 18]. These works were further refined by Sluijter & van Heuven [16], casting some light on the exact correlation between the different acoustic parameters. Their studies pointed out that the most reliable correlates of syllable stress are duration and mid-frequency emphasis. The presence of a high quantity of energy in the mid-to-high band of vowel spectra, where the main formants reside, is one of the parameters indicating a strong possibility for syllable stress. From our experiments there is strong evidence supporting Sluijter & van Heuven's ideas: stressed syllables exhibit a longer duration and greater energy in the vowel mid-to-high-frequency band [19].

### 3.2. Pitch accent detector

There is a long tradition of studies dealing with intonation profiles and accents [6, 15]. The influential work of Pierrehumbert introduced a two-level categorisation of pitch profiles enriched by a wide combination of symbols and diacritics to represent all possible intonation contours and pitch accents. Unfortunately such a categorisation, as well as the famous ToBI labelling scheme, appears to be difficult to encode in an automatic system capable of reliably identifying such categories and combinations. Taylor [20, 21, 22, 23] proposed a different view of intonation events. Starting from a rise/fall/connection (RFC) model, he defined a set of parameters capable of uniquely describing pitch accent shapes and boundary tones, called the TILT parameter set.

Following the model proposed by Taylor, the Praat-produced F0 contour was first converted into an RFC model. The contour was divided into frames 0.025 seconds long, and the data in each frame was linearly interpolated using a Least

Median Squares method to obtain robust regression and deletion of outliers. Then every frame line was classified as rise, fall or connection depending on its gradient; subsequent frames with the same classification were merged into one interval and the duration and amplitude of the rise or fall section was measured.

Having obtained a compact RFC representation, it is possible to identify every intonational event in the F0 contour. The view adopted here is to identify every possible event candidate to be a pitch accent, and evaluate the best combination, among the acoustic and TILT parameters, for identifying the actual pitch accents in the utterances. As described by Taylor [23], an intonational event that can be considered a candidate for pitch accent exhibits a rise followed by a fall profile. The actual pitch accents can be found by examining the event amplitude and if necessary some others parameters.

Sluijter & van Heuven suggested that the pitch accent can be reliably detected by using the overall syllable energy and some measure of pitch variation. The event amplitude, normalised using the mean event amplitudes across the utterances, that is part of the TILT parameter set, can be considered a measure of this variation, being the sum of the absolute amplitude of the rise and fall sections of a generic intonational event.

### 3.3. Prominence detector

According to Taylor [23], all the prosodic parameters involved in prominence study should be considered as continuous quantities, avoiding any kind of categorisation. This view is not so common in linguistics, where there is a tendency to deal with categorical/discrete representations of the examined phenomenon and avoid any kind of continuous function. On the other hand, for testing the reliability of an automatic system hand-tagged data have to be used: the manual tagging of utterances for prosodic phenomena is a highly complex task for humans and the introduction of categories seems unavoidable. For these reasons every prosodic quantity presented here is described and managed as a continuous quantity, then some provisional categorisations are proposed, to compare the behaviour and the performance of the automatic process with the hand-tagged data.

A preliminary paper about this study [19] examined the dependencies of prominence on the acoustic parameters described in the previous section. As suggested in the literature, confirmed by our earlier experiments, prosodic stress strictly depends on syllable nuclei duration and energy in a specific spectral band: the longer the duration and the higher the energy in the syllable nucleus, the greater the stress perception. In the same way, high overall nucleus energy and wide pitch movement produce the strongest pitch accent. Bearing in mind these relationships and considering the four-dimensional space generated by these parameters, it is possible to combine them properly to build a prominence function able to assign a value of prominence parameter for each syllable nucleus entirely derived from acoustic features. One proposal for such a function could be:

$$Prom(i) = \max[en_{500-2000}(i) \times dur(i), en_{ov}(i) \times ev_{amp}(i)]$$

where  $en_{500-2000}$  is the energy in the 500-2000 Hz frequency band,  $dur$  is the nucleus duration,  $en_{ov}$  is the overall energy in the nucleus and  $ev_{amp}$  is the TILT event amplitude (if an event is present in the nucleus, zero otherwise), all referred to a generic syllable nucleus  $i$ . The *Prom* function is built in such a way as to express, mathematically, the fact that a prominent syllable is usually stressed or pitch accented or both. A plot of prominence function for the sentence “For girls the overprotection is far more pervasive” taken from the TIMIT corpus is shown in figure 2.

As pointed out before, to evaluate the system, comparing it with hand-tagged data, it is necessary to introduce some kind of categorisation in prominence. Following Terken, a word or part of a word made prominent is perceived as standing out from its environment. Starting from this perspective, identifying prominent syllables is a matter of finding the maxima of the *Prom* function defined above. The prominence value of every syllable nucleus is compared with the two neighbours and if it represents a maximum the corresponding syllable nuclei, and then the connected syllable, are considered as prominent.

However, it is neither impossible nor rare, to have subsequent syllables that are both prominent, for example if they represent two monosyllabic words that are both prominent. The peak (maximum) picking algorithm will fail in this case, not recognising one of the two prominent syllables. To partially overcome this problem, in the case of two subsequent syllables that differ only by 15% of their prominence value, the peak picking algorithm is modified and, for each syllable, the test is performed by ignoring the neighbours with the similar prominence value. Moreover, syllables that have high prominence value, greater than 70% of the maximum peak in the utterance, are also considered as prominent.

By using the *Prom* function and the peak picking method described above, it was possible to produce a reliable prominence detector. The whole system was tested using a subset of TIMIT utterances, composed of 7328 syllables taken from 485 utterances spoken by 51 different speakers. The prominence detector correctly classified 80.2% of the syllables as either prominent or non-prominent, with an insertion rate of 5.6% (false alarms) and a deletion rate of 14.2% (missed detections). As pointed out before this is an unsupervised system, thus there is no need for any training phase.

## 4. Conclusions

It is widely accepted in the literature that inter-human agreement, when manually tagging prominence in continuous speech, is around 80% [11, 14]. The unsupervised prominence detector presented here exhibits an overall agreement of 80.2% with the data manually tagged by a native speaker; this performance is obtained without using any information apart from acoustic parameters derived directly from the utterance waveform. The results are comparable with those obtained by human taggers, so the presented prominence detector can be seen as a valid alternative to manual tagging for building large resources useful for language research and teaching.

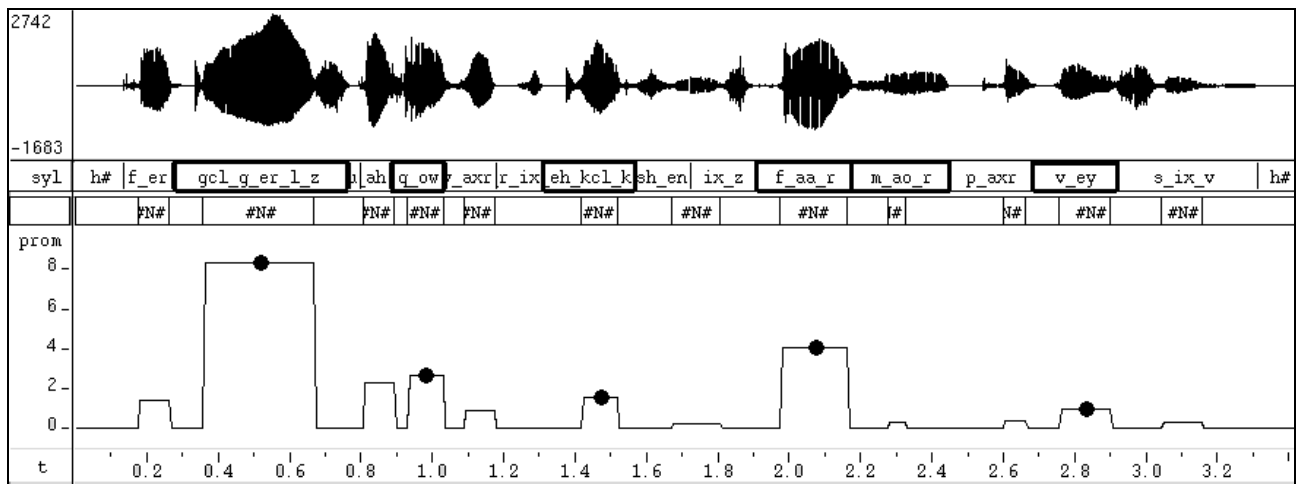


Figure 2: Prosodic prominence function values for the utterance “For girls the overprotection is far more pervasive”. From the top, the waveform plot, the syllable segmentation (shown only for the purposes of comparison as it is not used by the system presented here), the syllable nuclei as detected by the system and finally the prominence values for every nucleus identified by the segmentation procedure. Prominent nuclei, as identified by the automatic system, are marked by a dot on the function profile, while prominent syllables, as classified by a human listener, are indicated by a thick box in the syllable segmentation track (“syl”).

## 5. References

- [1] Bagshaw, P.C., “An investigation of acoustic events related to sentential stress and pitch accents, in English.”, *Speech Comm.*, 13, pp. 333-342, 1993.
- [2] Bagshaw, P.C., *Automatic prosodic analysis for computer-aided pronunciation teaching*. PhD thesis, University of Edinburgh, 1994.
- [3] Beckman, M.E., *Stress and non-stress accent*. Foris Publications, Dordrecht, Holland, 1986.
- [4] Beckman, M.E. and Venditti, J.J., “Tagging prosody and discourse structure in elicited spontaneous speech.” In Proc. *Science and Technology Agency Priority Program Symp. on Spontaneous Speech*, Tokyo, pp. 87-98, 2000.
- [5] Boersma, P. and Weenik, D., “Praat, a system for doing phonetics by the computer.” *Report 132*, Institute of Phonetic Sciences, University of Amsterdam, 1996.
- [6] Campione, E. and Veronis, J., “A multilingual prosodic database”, In Proc. *ICSLP98*, Sydney, 1998.
- [7] Delmonte, R., “SLIM prosodic automatic tools for self-learning instruction”, *Speech Comm.*, 30, pp. 145-166, 2000.
- [8] Fach, M. and Wokurek, W., “Pitch Accent Classification of Fundamental Frequency Contours by HMM”. In Proc. *Eurospeech '95*, Madrid, pp. 2047-2050, 1995.
- [9] Hieronymus, J.L., McKelvie, D. and McInnes, F.R. “Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition”, In Proc. *ICASSP-92*, San Francisco, pp. 225-227, 1992.
- [10] Howitt, A.W., *Automatic Syllable Detection for Vowel Landmarks*, PhD Thesis, MIT, 2000.
- [11] Jenkin, K. and Scordilis, M., “Development and Comparison of Three Syllable Stress Classifiers”, In Proc. *ICSLP96*, 1996.
- [12] Mermelstein, P., “Automatic segmentation of speech into syllabic units.”, *JASA*, 58 (4), pp. 880-883, 1975.
- [13] Neumeyer, L. *et al.*, “Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech.”, In Proc. *ICSLP96*, Philadelphia, pp. 1457-1460, 1996.
- [14] Pickering, B., Williams, B. & Knowles, G., “Analysis of transcriber differences in SEC”, In Knowles G., Wichmann, A. & Alderson, P. (eds), *Working with speech*, London: Longman, pp. 61-86, 1996.
- [15] Pierrehumbert, J.B., *The phonetics and phonology of English intonation.*, PhD thesis, MIT, 1980.
- [16] Sluijter, A. and van Heuven, V., “Acoustic correlates of linguistic stress and accent in Dutch and American English.”, In Proc. *ICSLP96*, Philadelphia, pp. 630-633, 1996.
- [17] Streefkerk, B.M., “Acoustical correlates of prominence: a design for research.”, In Proc. *Inst. of Phon. Sciences*, Vol. 20, University of Amsterdam, pp. 131-142, 1997.
- [18] Streefkerk, B.M. *et al.*, “Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's.”, In Proc. *Eurospeech '99*, Budapest, pp. 551-554, 1999.
- [19] Tamburini, F., “Automatic detection of prosodic prominence in continuous speech”, In Proc. *LREC2002*, Las Palmas, Spain, 301-306, 2002.
- [20] Taylor, P.A., *A phonetic model of English intonation*, PhD thesis, University of Edinburgh, 1992.
- [21] Taylor, P.A., “Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model”, In Proc. *Eurospeech '93*, Berlin, 1993.
- [22] Taylor, P.A., “The rise/fall/connection model of intonation.”, *Speech Comm.*, 15, pp. 169-186, 1995.
- [23] Taylor, P.A., “Analysis and Synthesis of Intonation using the Tilt Model”, *JASA*, 107 (3), pp. 1697-1714, 2000.
- [24] Terken, J., “Fundamental frequency and perceived prominence.”, *JASA*, 89 (4), pp.1768-1776, 1991.
- [25] Wightman, C.W. & Ostendorf, M., “Automatic labelling of prosodic patterns.”, *IEEE Transaction on Speech and Audio Processing*, 2 (4), pp. 469-481, 1994.