Automatic detection of prosodic prominence in continuous speech

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 duration and high-frequency emphasis. By deriving a set of acoustic parameters it is possible to build syllable-stress detectors as well as pitch-accent detectors and combine them to build an automatic system devoted to prominence detection. Starting from a syllable-segmented utterance, the system presented here is capable of correctly identify prominent syllables with an agreement, with human-tagged data, comparable with the inter-human agreement reported in the literature.

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 (2000), a precise identification of such phenomena helps to disambiguate the meaning of some utterances and is a fundamental step for the automatic recognition and synthesis of spontaneous speech. Moreover the construction of large annotated language resources, such as prosodically tagged speech corpora, is of increasing interest both for research purposes, in the phonetic/phonological field, and for teaching languages and their correct pronunciation.

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 (Terken, 1991). A better understanding of how prominence is physically accomplished is a basic step in the construction of tools capable of automatically identifying such phenomena. These tools will be extremely useful in speech recognition to distinguish between different meanings, and in speech production to enhance the fluency and adequacy of automatic speech-generation systems.

This paper presents work in progress on the construction of a system for the automatic detection of prosodic features in speech using only acoustic/phonetic parameters and cues. In particular, the paper presents a study analysing the connections between the acoustic parameters of speech and the perception of prosodic prominence. Building an acoustic model of perceived prosodic prominence, and casting light on the mathematical correlations between acoustic measures and prominence, allows for the construction of tools that will be useful in building speech language resources, such as automatic taggers of prosodic features in speech corpora and automatic training systems for the self-access study of pronunciation.

Following Beckman's (1986) phonological view, further developed by Bagshaw (1993, 1994), syllables that are perceived as prominent either contain a pitch accent or are somehow "stressed". Prominent syllables containing a pitch accent are called *accented syllables*, while prominent syllables without a corresponding pitch accent are called *stressed syllables*. On the acoustic/phonetic side, the accomplishment of such features has to be strictly correlated with acoustic parameters. Beyond the works already cited, there are many studies (Heldner, 1996; Streefkerk *et al.* 1996, 1997, 1999; Sluijiter & Heuven, 1996) 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 main goal of the project presented here is to build an automatic system capable of reliably identifying prominence in speech, using only cues derived from acoustic measurements. The project is divided into two separate steps: the first step involves the automatic identification of syllable boundaries, while the second one concerns the identification of prominent syllables by means of acoustic measures. This paper will concentrate on the second step, proposing a possible combination of acoustic parameters to solve such problems, basing the processing on reliable syllable segmentation.

The data set used in these experiments is a subset of the DARPA/TIMIT acoustic-phonetic continuous speech corpus, that consists of thousands of transcribed, phone-segmented and aligned sentences of American English. Starting from the phone transcription of the utterances, a native speaker manually tagged all the syllables she perceived as prominent, as well as grouping the phones into syllables, obtaining a new utterance segmentation containing syllable boundaries and prosodic prominence labels. This new segmentation was the starting point for all the measurements presented in this paper.

Several studies have been conducted in this field for building automatic systems capable of reliably identifying either one acoustical correlate of prominence (Taylor, 1995a; Fach & Wokurek, 1995; Campione & Veronis, 1998) or a complete set of prosodic parameters (Wightman & Ostendorf, 1994; Bagshaw, 1994; Delmonte, 2000). 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 algorithms for the reliable tagging of prosodic features, in particular prominence, avoiding the training phase and the use of additional resources. The subset of the TIMIT corpus referred to in this study is used only in the test phase to outline how the different acoustic parameters behave on prominent and non-prominent syllables.

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.

Section 2 describes the basic acoustic parameters involved in prominence detection; section 3 outlines the computation of the prosodic parameters (pitch accent, stress and prominence), while section 4 discusses the experimental results and draws some provisional conclusions.

2. The computation of acoustic parameters

Before examining in detail each acoustic parameter involved in this study, it is necessary to consider the normalisation of each measurement presented here. All acoustic parameters must be normalised to some extent to avoid the natural variations among different speakers. Thus, in the following sections, I present the specific normalisation procedures applied to each parameter; it is important to bear in mind that all graphs and measurements presented here refer to normalised parameters. This is the reason why units of measurements are not always indicated in the diagrams.

2.1. Energy

The first acoustic parameter involved in this study is overall syllable energy. It can be computed in various ways. Here I refer to RMS energy, computed as follows:

$$E_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i^2}$$

where a_i is the i-th syllable sample.

Overall syllable energy is normalised by dividing it by mean syllable energy over the utterance. This reduces the energy variation across different utterances and different speakers.

2.2. Duration

Every utterance considered in the data is segmented and every syllable is clearly defined over time. Computing syllable duration is therefore straightforward. The duration parameter is also normalised, considering the mean duration of the syllables in the utterance. This is a standard technique for ROS (Rate-Of-Speech) normalisation, as described in Neumeyer *et al.* (1996) and Venkata Ramana Rao Gadde (2000).

2.3. Fundamental Frequency (F0) contour

The extraction of F0 contour, or pitch contour, is a complex task. Bagshaw (1994) carried out an accurate comparison of the different algorithm 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. A post-processing procedure to smooth such variations is often required in order to obtain reliable results. The Praat speech package (Boersma, 1993, 1996) 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. Figure 1 shows an example of a raw F0 contour as computed by F0 detector in Praat and the corresponding post-processed contour used for subsequent computations.

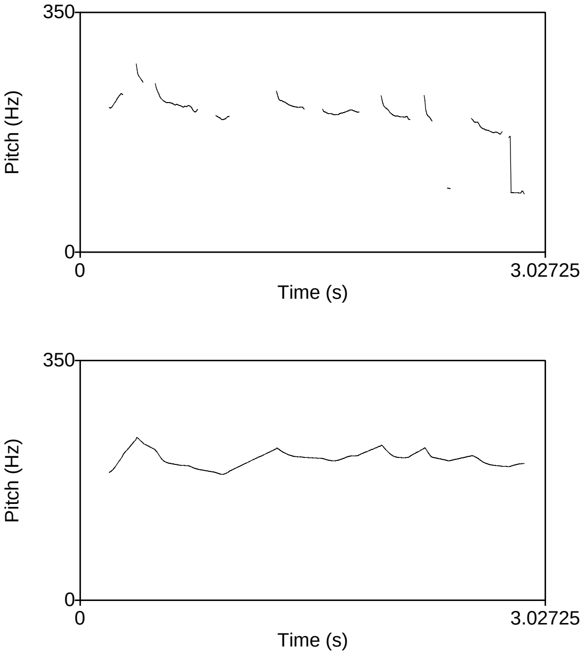

Figure 1: An F0 contour before and after post-processing.

2.4. Spectral emphasis

It has been shown, especially by the influential work of Sluijter & van Heuven (1996), that high-frequency emphasis is one useful parameter in determining stressed syllables. Each syllable 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 and used as the parameter for the subsequent computations. Figure 2 shows the distributions of prominent and non-prominent syllable energies in the frequency bands considered. 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 reveal a strict dependence of syllable prominence to vowel high frequency emphasis.

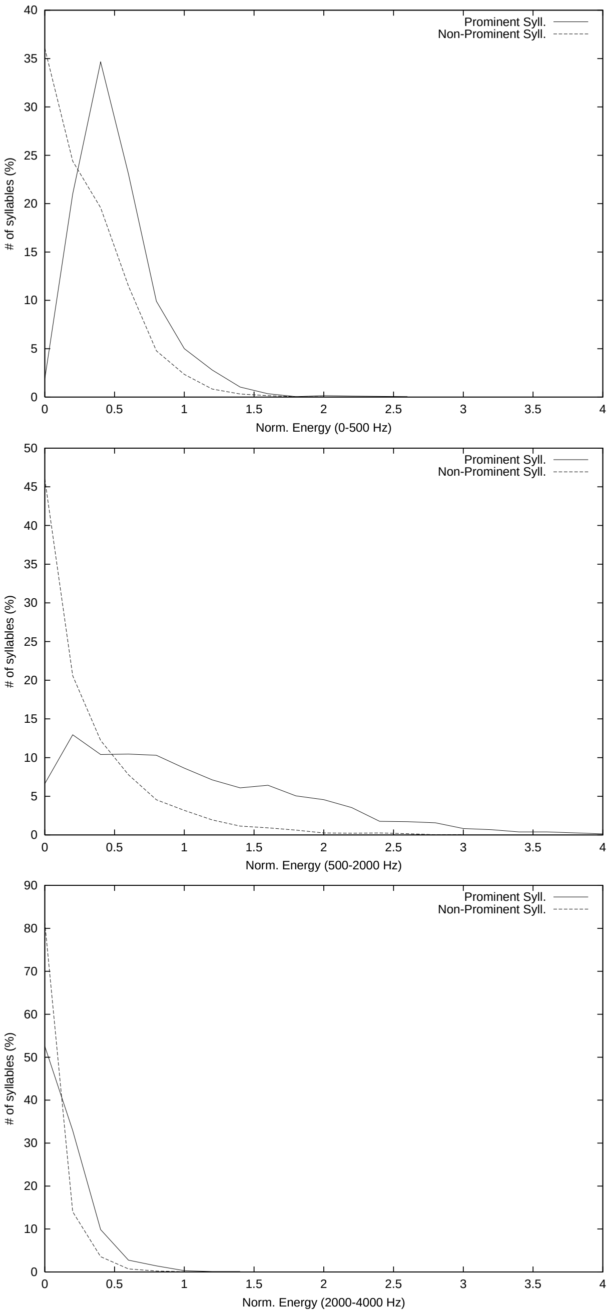

Figure 2: Distributions of prominent and non-prominent syllable energies in the considered frequency bands

3. Prosodic parameters

This section examines the prosodic quantities that are the object of the study: stress, pitch accent and prominence. According to Taylor (2000), all these parameters should be considered as continuous quantities, avoiding any kind of categorisation. It is common practice in linguistics to deal with categorical/discrete representations of the examined phenomenon and totally avoid any kind of continuous function. However, for testing the reliability of an automatic system one can rely only on hand-tagged data: the manual tagging of utterances is a highly complex task for humans and the introduction of some categories seems unavoidable. For these reasons every prosodic quantity presented here is first briefly described as a continuous quantity, then some provisional categorisations are proposed, often as threshold values or functions, to compare the behaviour of the automatic process with the hand-tagged data.

3.1. Stress

The main correlates of syllable stress indicated in the literature are syllable duration and energy (Bagshaw 1993,1994; Streefkerk *et al* 1996, 1997, 1999). However the work of Sluijter & van Heuven casts some light on the exact correlation among the different acoustic parameters. "Previous research on American English was generally hampered by covariation of stress and accent" they claim. Their studies clearly divided the two phenomena, pointing out that the most reliable correlates of syllable stress are duration and high-frequency emphasis. The presence of a high quantity of energy in the high band of vowel spectra, where the main formants reside, is one of the parameters indicating a strong possibility for syllable stress.

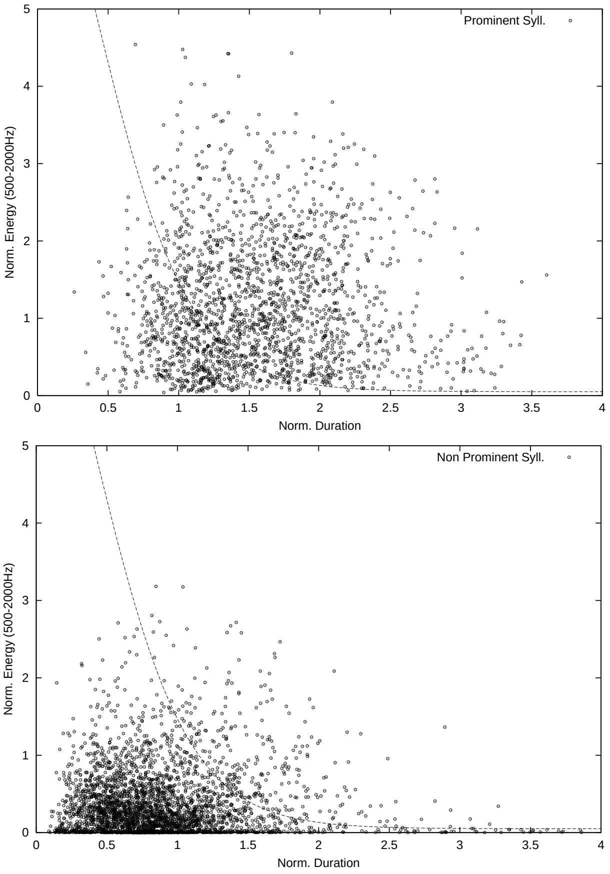

Figure 3: Prominent and non-prominent syllables as a function of normalised duration and spectral energy in the band from 500 to 2000 Hz.

Figure 3 shows prominent and non-prominent syllables as a function of syllable-normalised duration and RMS energy in the band from 500 to 2000 Hz. There is strong evidence supporting Sluijter & van Heuven's ideas: a stressed syllable exhibits a longer duration and a remarkable energy in the vowel high frequency band. The dashed curve in the diagrams is the proposed threshold, experimentally determined, to distinguish stressed from unstressed syllables. Figure 3 shows a clear separation between the cluster of prominent syllables and the cluster of non-prominent ones. Nevertheless, a small overlapping region emerges quite clearly from the diagrams. Ideally it could be perfectly correct, because in the model presented here, stress is only one of the parameters contributing to prominence, so the prominent syllables that are not captured by the process presented in this section may be identified correctly by the other parameter, pitch accent.

3.2. Pitch Accent

There is a long tradition of studies dealing with intonation profiles and accents (Pierrehumbert 1980; Taylor, 1992; Campione & Veronis, 1998). 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 (1992, 1993, 1995a, 1995b, 2000) 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. This set consists of five parameters defined as follows:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \qquad tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}}$$

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2 \cdot (|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2 \cdot (D_{rise} + D_{fall})}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \qquad D_{event} = D_{rise} + D_{fall}$$

where A_{rise}, A_{fall}, D_{rise}, D_{fall} are respectively the amplitude and the duration of the rise and fall segments of the intonation event.

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 (Rousseeou, 1987) 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 the fall section was measured.

Having obtained a compact RFC representation, it is possible to identify every intonational event in the F0 contour. Taylor used a system based on neural networks (1995a) and Fach *et al.* (1995) a system based on HMM for event identification; such methods require a training phase, which I would like to avoid for the reasons mentioned above. My work is mainly concerned with pitch accent detection, so the events I am looking for are accent shapes. 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 (1992, 2000) an intonational event that can be considered a candidate for pitch accent exhibits a rise profile followed by a fall profile. There are different degrees of such profiles leading to the degenerate cases in which only a rise or fall section exists. All the events exhibiting these shapes are possible candidates for pitch accents. The actual pitch accents can be found by examining the event amplitude, as outlined by Taylor, and eventually some others parameters.

Sluijter & van Heuven proposed that the pitch accent can be reliably detected by using the overall syllable energy and some measure of pitch variation. The event amplitude (A_{event}), 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. Figure 4 shows a plot of prominent and non-prominent syllables as a function of overall syllable energy and event amplitude. There is an evident correlation among these parameters when identifying prominent syllables. Again the dashed curve represents the proposed threshold.

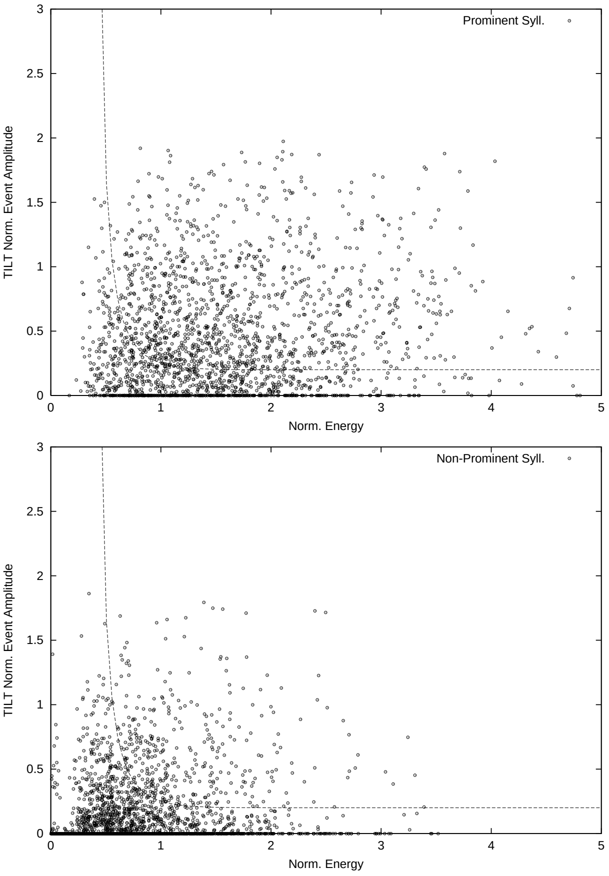

Figure 4: A plot of prominent and non-prominent syllables as a function of overall syllable energy and event amplitude (A_{event}).

On the basis of the proposed threshold, we can plot the syllables carrying a pitch accent in the same way as plotted for stressed syllables in figure 3 (see figure 5). As supposed before, the prominent syllables that are not identified by the stress parameter are captured using pitch accent. The overlapping region between prominent and non-prominent syllables can be resolved by means of a pitch accent detector. In fact figure 5 shows a number of identified prominent syllables in the "messy" overlapping region outlined above.

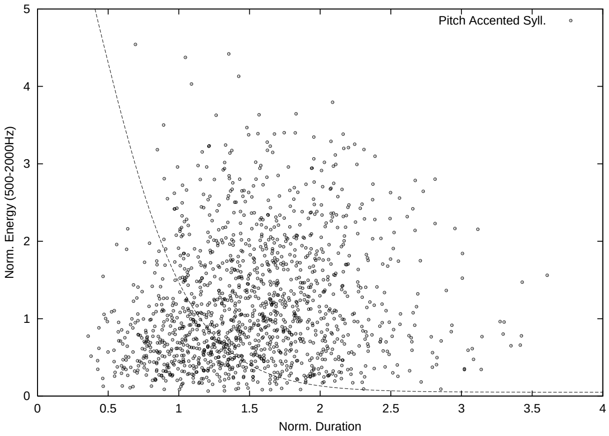

Figure 5: Pitch accented syllables.

3.3. Prominence detector

As described in the previous section, the pitch accent detector is able of capturing most of the prominent syllables not identified by the stress detection method. By combining the two detectors, on the basis of the methodological issues presented above, one should be able to produce a reliable prominence detector. As described before, prominent syllables can be identified either as pitch accented or stressed syllables. Table 1 shows the results of the prominence detector when applied to the TIMIT subset considered here. The set consists of 367 utterances divided into 5531 syllables.

	Stressed	Pitch Accented	Stressed+ Pitch Acc.	None
Prominent	544	216	877	401
Non-Prom.	184	210	117	2982

Table 1: The results obtained by applying the prominence detector to the TIMIT subset considered here.

The prominence detector correctly classify 83.5% of the syllables as either prominent or non-prominent, with an insertion rate of 7.2% and a deletion rate of 9.2%.

For completeness, I tested the method on a different subset of the TIMIT corpus, tagged and segmented in the same way, consisting of 118 utterances and 1797 syllables. I obtained the same general figure: 80.4% of correct classifications with an insertion rate of 13.5% and a deletion rate of 6%.

4. Conclusions

It is widely accepted in literature that inter-human agreement, when manually tagging prominence in continuous speech, is around 80%. For example in the study conducted by Pickering (1996) on the Spoken English Corpus, such agreement has been estimated around 83%.

The prominence detector presented here exhibits an overall agreement of 82% with the manually-tagged data by a native speaker. The results are comparable with those obtained by humans 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.

5. Acknowledgements

I would like to thank especially Carlo Caini for his helpful support and suggestions at every stage of this project, and the teams of both Departments where I work who sustained me during the development process. Special thanks are due to Maureen Lister, who manually tagged all the utterances used in this paper with respect to perceived prosodic prominence.

6. References

Bagshaw, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, 13: 333-342.

Bagshaw, P.C. (1994). *Automatic prosodic analysis for computer-aided pronunciation teaching*. PhD thesis, University of Edimburgh.

Beckman, M.E. (1986). *Stress and non-stress accent*. Dordrecht, Holland: Foris Publications.

Beckman, M.E. & Venditti, J.J. (2000). Tagging prosody and discourse structure in elicited spontaneous speech. In *Proceedings of the Science and Technology Agency Priority Program Symposium on Spontaneous Speech: Corpus and Processing Technology* (pp. 87-98). Tokyo.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound. In *Proceedings of the Institute of Phonetic Sciences* (pp. 97-100), Vol. 17, University of Amsterdam.

Boersma, P. & Weenik, D. (1996). Praat, a system for doing phonetics by the computer. *Report 132 of the Institute of Phonetic Sciences*, University of Amsterdam.

Campione, E. & Veronis, J. (1998). A multilingual prosodic database, In *Proceedings of the International Conference on Spoken Language Processing - ICSLP98*, Sydney.

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, 30: 145-166.

Fach, M. & Wokurek, W. (1995). Pitch Accent Classification of Fundamental Frequency Contours by Hidden Markov Models. In *Proceedings of the Eurospeech '95 Conferenc*e (pp. 2047-2050), Madrid.

Heldner, M. (1998). Is an F0-rise a necessary or a sufficient cue to perceived focus in Swedish? In *Nordic Prosody: Proceedings of the VII Conference* (pp. 109-125), Frankfurt am Main: Peter Lang.

Neumeyer, L., Franco, H., Weintraub, M. & Price P. (1996). Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech. In *Proceedings of the 4th International Conference on Spoken Language Processing - ICSLP96*, Philadelphia, PA.

Pickering, B., Williams, B. & Knowles, G. (1996). Analysis of transcriber differences in SEC. In Knowles G., Wichmann, A. & Alderson, P. (eds), *Working with speech* (pp. 61-86). London: Longman

Pierrehumbert, J.B. (1980). *The phonetics and phonology of English intonation*. PhD thesis, Massachusetts Institute of Technology.

Rousseeuw, P.J. (1987). *Robust regression and outlier detection*. New York: Wiley.

Sluijter, A. & van Heuven, V. (1996) Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings of the 4th International Conference on Spoken Language Processing - ICSLP96* (pp. 630-633), Philadelphia, PA.

Streefkerk, B M. & Pols, L.C.W. (1996). Prominent accents and pitch movements. In *Proceedings of the Institute of Phonetic Sciences* (pp. 111-119), Vol. 21, University of Amsterdam.

Streefkerk, B M. (1997). Acoustical correlates of prominence: a design for research. In *Proceedings of the Institute of Phonetic Sciences* (pp. 131-142), Vol. 20, University of Amsterdam.

Streefkerk, B M., Pols, L.C.W. & ten Bosch, L.F.M. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. In *Proceedings of the Eurospeech '99* (pp. 551-554), Budapest.

Taylor, P.A. (1992). *A phonetic model of English intonation*. PhD thesis. University of Edimburgh.

Taylor, P.A. (1993). Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model, In *Proceedings of the Eurospeech '93 conferenc*e, Berlin.

Taylor, P.A. (1995a). Using Neural Networks to Locate Pitch Accents, In *Proceedings of the Eurospeech '95 conferenc*e, Madrid.

Taylor, P.A. (1995b). The rise/fall/connection model of intonation. *Speech Communication*, 15: 169-186.

Taylor, P.A. (2000). Analysis and Synthesis of Intonation using the Tilt Model, *Journal of the Acoustical Society of America*. Vol. 107, 3: 1697-1714.

Terken, J. (1991). Fundamental frequency and perceived prominence. *J. Acoust. Soc. Am.*, Vol. 89, No. 4:1768-1776.

Venkata Ramana Rao Gadde (2000). Modeling Word Duration for better speech recognition, In *Proceedings of the Speech Transcription Workshop*, University of Maryland, MD.

Wightman, C.W. & Ostendorf, M. (1994). Automatic labelling of prosodic patterns. *IEEE Transaction on Speech and Audio Processing*, Vol. 2, No 4:469-481.