INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 8, 33–44, 2005

An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech

F. TAMBURINI
Centro Interfacoltà di Linguistica Teorica e Applicata, University of Bologna, Italy and Dip. di Elettronica, Informatica e Sistemistica, University of Bologna, Italy
f.tamburini@cilta.unibo.it

C. CAINI
Dip. di Elettronica, Informatica e Sistemistica, University of Bologna, Italy
ccaini@deis.unibo.it

Abstract. A precise identification of prosodic phenomena and the construction of tools able to properly manage such phenomena are essential steps to disambiguate the meaning of certain utterances. In particular they are useful for a wide variety of tasks: automatic recognition of spontaneous speech, automatic enhancement of speech-generation systems, solving ambiguities in natural language interpretation, the construction of large annotated language resources, such as prosodically tagged speech corpora, and teaching languages to foreign students using Computer Aided Language Learning (CALL) systems. This paper presents a study on the automatic detection of prosodic prominence in continuous speech, with particular reference to American English, but with good prospects of application to other languages. Prosodic prominence involves two different prosodic features: pitch accent and stress accent. Pitch accent is acoustically connected with fundamental frequency (F0) movements and overall syllable energy, whereas stress exhibits a strong correlation with syllable nuclei duration and mid-to-high-frequency emphasis. This paper shows that a careful measurement of these acoustic parameters, as well as the identification of their connection to prosodic parameters, makes it possible to build an automatic system capable of identifying prominent syllables in utterances with performance comparable with the inter-human agreement reported in the literature. Two different prominence detectors were studied and developed: the first uses a training corpus to set up thresholds properly, while the second uses a pure unsupervised method. In both cases, it is worth stressing that only acoustic parameters derived directly from speech waveforms are exploited.

Keywords: prosody, automatic feature extraction, prominence, stress accent, pitch accent

1. Introduction

The study of prosodic phenomena in speech is a central topic in language investigation and it is generally agreed that it represents one of the main streams for improving the performance of speech processing systems. Speakers tend to focus the listener's attention on the most important parts of the message by means of prosodic markers and, as outlined in Beckman and Venditti (2000), a precise identification of such phenomena is often essential in order to solve possible ambiguities in the meaning of some utterances. Automatic prosody analysis or synthesis is a fundamental step in a variety of speech processing applications. For example, it can improve performance in spontaneous speech automatic recognition (Hieronymus et al., 1992), it enhances the fluency and adequacy of automatic speech-generation systems (Bulyko, 1999) and it may be useful for solving ambiguities in natural language parsing (Warren, 1996).

Moreover, the construction of large annotated language resources, such as prosodically tagged speech corpora, shows an increasing interest both for research purposes and for language teaching (Hirst, 2001). Also teaching languages using CALL systems with software modules capable of properly managing prosodical information (Auberg et al., 1998; Nouza, 1999; Delmonte, 2000) seems to be an interesting option, especially in second language acquisition.

There are many studies that examine the general aspects of correlation between prosodic phenomena, such as rhythm (Ramus, 2002), boundary tones (Yang and Wang, 2002), prominence (Bagshaw, 1994), and acoustic parameters. Other studies are focused on the building of automatic systems capable of reliably identifying either one acoustic correlate of prominence (e.g. Fach and Wokurek, 1995; Campione and Veronis, 1998) or a complete set of prosodic parameters, such as prominence, intonation, rhythm, etc., and their acoustic correlates (Bagshaw, 1993, 1994; Wightman and Ostendorf, 1994; Jenkins and Scordilis, 1996; Delmonte, 2000). These studies typically rely on additional phonetic information, such as phone labelling and/or utterance transcriptions, for deriving prosodic phenomena. Such systems, based on Hidden Markov models, neural networks or similar methods, often require a training phase in order to work properly on test data. Although powerful, they present the drawback of requiring an adequately segmented and prosodically labelled speech corpus for carrying out the training phase. This resource might not be always available and it would certainly be expensive to build. Moreover, the system would be permanently bound to one specific language.

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). Following Beckman's (1986) phonological view, further developed by other scholars, for example Bagshaw (1993, 1994), syllables that are perceived as prominent either contain a pitch accent or a stress accent. On the acoustic/phonetic side, the accomplishment of such features has to be strictly correlated with particular behaviour of acoustic parameters, either considered as single features or, more likely, as combinations of them. As well as the works already cited, there are many other studies (Sluijter and van Heuven, 1996; Streefkerk, 1997; Streefkerk et al., 1999), 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. These studies perform an in-depth analysis of the correlation between prominence and a set of acoustic features, using various methods and techniques, to identify the best acoustic correlates of prosodic prominence. Although providing an accurate analysis of the correlations between single acoustic parameters and prominence, they do not assess quantitatively the combinations of such features for the best prominence identification. They suggest, from a qualitative point of view, possible combinations without exploring them using specific data. These suggestions formed the basis of our study, but their claims were further analysed and verified using quantitative data.

Despite the quantity and quality of studies on this topic, it seems that the automatic and reliable detection of prosodic prominence, without providing phonetic information, such as utterance transcription or training corpora composed of segmented utterances, is still an open question.

This paper presents a study on the relationships between prosodic prominence and acoustic features with the aim of designing a prototype system for the automatic detection of prominence features in speech using only acoustic/phonetic parameters and cues. First, the problems related to the automatic derivation of basic acoustic parameters, such as duration, energy, pitch contour and spectral emphasis, from speech waveforms, are addressed, proposing, whenever necessary, novel methods to the specific acoustic feature extraction. Then, the relationships between these acoustic features and the wanted prosodic features, such as pitch accents, stress accents and prominence, are studied, casting some light on the possible design of automatic prominence detectors. Finally, two alternative solutions are presented in the paper. The first is grounded on a Gaussian mixture discriminator model. It offers good performance but it relies on a training corpus, manually tagged inserting prominence information, to properly set up the parameters involved in the decision procedure. The second, more innovative, was developed to explore the possibility to obtain basically the same satisfactory performance without making use of training procedures or other additional resources. This new method is based on the definition of a general prominence function that combines some acoustic parameters directly derived from speech waveforms. As such it does not require any additional resource.

The data set used in these experiments is a subset of the DARPA/TIMIT acoustic-phonetic continuous speech corpus, consisting of thousands of transcribed, phone-segmented and aligned sentences of American English (Garofolo et al., 1993). In this study, the TIMIT annotations are used only for testing and measuring system performance, rather than additional information for additional the prominence detection algorithms.

The rest of the paper is organised as follows. Section 2 describes the automatic derivation of acoustic parameters from speech waveforms. Section 3 presents a study about the combination and relationships of these acoustic features to identify prosodic features such as pitch accents, stress accents and prominence, casting some light on the construction of automatic detectors of such prosodic features. Section 4 discusses the two different detectors of prosodic prominence presented in this study. Section 5 draws the conclusions of the work, comparing and discussing the results obtained with the literature considered.

2. The Acoustic Parameters

The methods for prosodic feature extraction described in the next section are based on the computation of the following acoustic parameters: syllable duration, fundamental frequency (F0) contour, overall energy, and spectral emphasis. Before examining them in detail, it is necessary to state in advance that all the acoustic parameters considered here must be normalised to some extent to avoid the natural variations among different speakers and different utterances. Thus, all graphs and measurements presented here refer to normalised parameters. This is why units of measurements are not always indicated in the diagrams. In the following sections, we describe the specific normalisation procedures applied to each parameter.

2.1. Duration

The linguistic theories of prosodic prominence mentioned in the introduction agree in considering syllable duration as one of the fundamental acoustic parameters for detecting syllable stress, certainly in American English, but also in many other languages. Unfortunately, the automatic segmentation of the utterance into syllables is a challenging task. In (Howitt, 2000) a survey of syllable segmentation algorithms is presented, but none of the methods analysed in this study seems to perform well when applied to continuous speech. To overcome this problem an alternative duration measure for prosodic prominence detection, capable of offering good performance even in this case, should be considered.

Instead of measuring the whole syllable length, heavily affected by consonant durations, it would be preferable, from an automatic implementation point of view, to measure the syllable nucleus duration, as done, for example, in Jenkin, Scordilis (1996). However, even if there are scholars claiming that durational variations due to the presence of stress mainly affects vowel durations in the syllable (Waterson, 1987), it is necessary to show that considering syllable nuclei instead of whole syllables does not lead to any information loss, as far as the ability to discriminate between prominent and non prominent syllables is concerned. To explore such an interesting possibility, we considered some utterances taken from the TIMIT corpus, containing syllables classified as prominent and non- prominent, and compared the duration of the syllable nucleus with the duration of the entire syllable. Taking the logarithm of these measures and adopting a Gaussian approximation (see Section 4.1), we obtained the distributions presented in Fig. 1. The two sets of Gaussians look qualitatively very similar and the separation between the two classes (prominent and non-prominent syllables) remains almost the same using the two different measures. As a further test, we built two Gaussian discriminators on the basis of the distributions presented in Fig. 1 and classified a set of test syllables, with respect to prominence, obtaining almost the same ratio of correct classifications. Note that the exact classification performance is not important in this context, as the syllable duration measure is only one of the parameters that can be jointly applied for prominence detection. The relevant conclusion, interesting for the present and future prominence studies, is that we can reliably replace the syllable duration measure, necessarily affected by large measurement error whenever obtained by automatic procedures, with the measure of syllable nucleus duration, which can be automatically obtained with a higher accuracy level. To the best of the authors' knowledge, this important consideration has been qualitatively discussed in various works, but a quantitative study based on a large amount of data, as the one reported here, has not been presented in the literature before.

To reliably identify the syllable nuclei in the utterance and measure their duration to obtain the acoustic parameter needed for subsequent computations, we

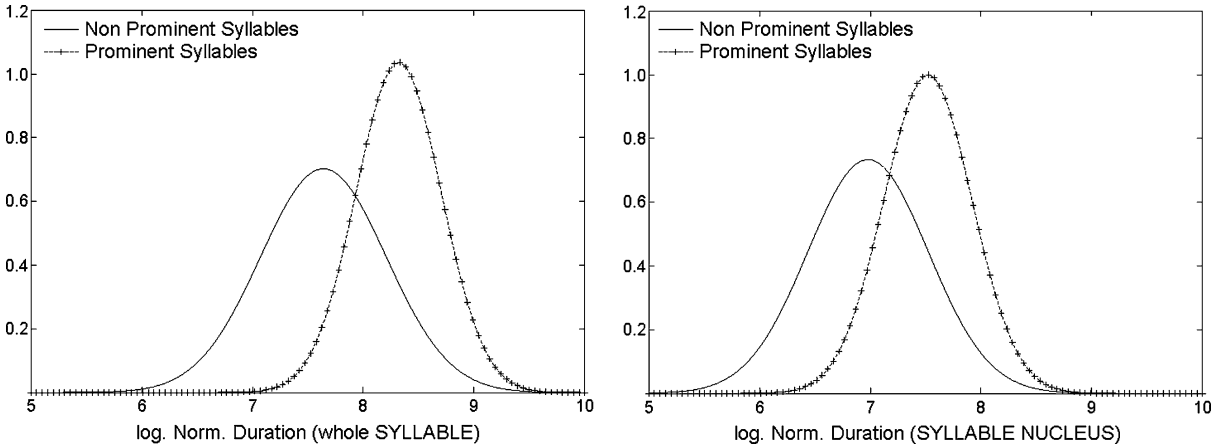

Figure 1. Gaussian approximation of duration measures: whole syllable (left) and syllable nucleus (right).

applied a modified version of the convex-hull algorithm (Mermelstein, 1975) to the utterance energy profile. This was computed after band-pass filtering (300–900 Hz) the speech-samples, as suggested in Howitt (2000), to filter out energy information not belonging to vowel phones, which forms the syllable nucleus. The duration parameter is then normalised by considering the mean duration of the syllable nuclei in the utterance. This is a standard technique for Rate-Of-Speech (ROS) normalisation, described, for example, in Neumeyer (1996) and Venkata Ramana (2000).

All the subsequent measurements of acoustic parameters will be referred to the syllable-nucleus intervals computed using the method described in this section.

2.2. *Fundamental Frequency (F0) Contour*

The extraction of F0 contour, or pitch contour, is another demanding task. Bagshaw (1994) carried out an accurate comparison of different algorithms for fundamental frequency estimation. Most of the complexity of pitch extraction process resides in the post-processing optimisation. Stops and glitches often tend to distort the contour, introducing spurious changes in the profile. Other typical problems in obtaining a correct pitch profile derive from octave jumps, where the pitch frequency computed by the algorithm, in a specific speech frame, is found to be double (or half) the correct pitch frequency. These F0 computation errors led to spurious sharp rises or falls in the pitch contour. A post-processing procedure to smooth out such variations is often required in order to obtain more reliable results.

To extract pitch contour we used the ESPS get_f0 program, derived from the algorithm presented in Talkin (1995), and, in particular, the version included in the wavesurfer speech package (Sjölander and Beskow, 2000). To obtain a continuous profile, the post-processing phase involves octave-jump removers and profile smoothers, derived from the ones proposed in Bagshaw (1994), applied both at voiced interval and sentence level, and a final interpolation between voiced regions.

2.3. *Energy*

Differently from the parameters presented in the previous subsections, the third acoustic parameter considered here, namely the syllable nucleus energy (or intensity), can be automatically computed in various ways without any particular difficulty. Here we refer to RMS energy, defined as:

$$E_j^{\mathrm{RMS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_{ji}^2}$$

where N is the number of samples per frame and $a_{1..N}$ are the speech samples in the j-th frame. The nucleus energy is successively normalised to the mean nucleus energy over the utterance. This reduces the energy variation across different utterances and different speakers.

2.4. *Spectral Emphasis*

In the recent literature, and in particular in the influential work of Sluijter and van Heuven (1996), it has

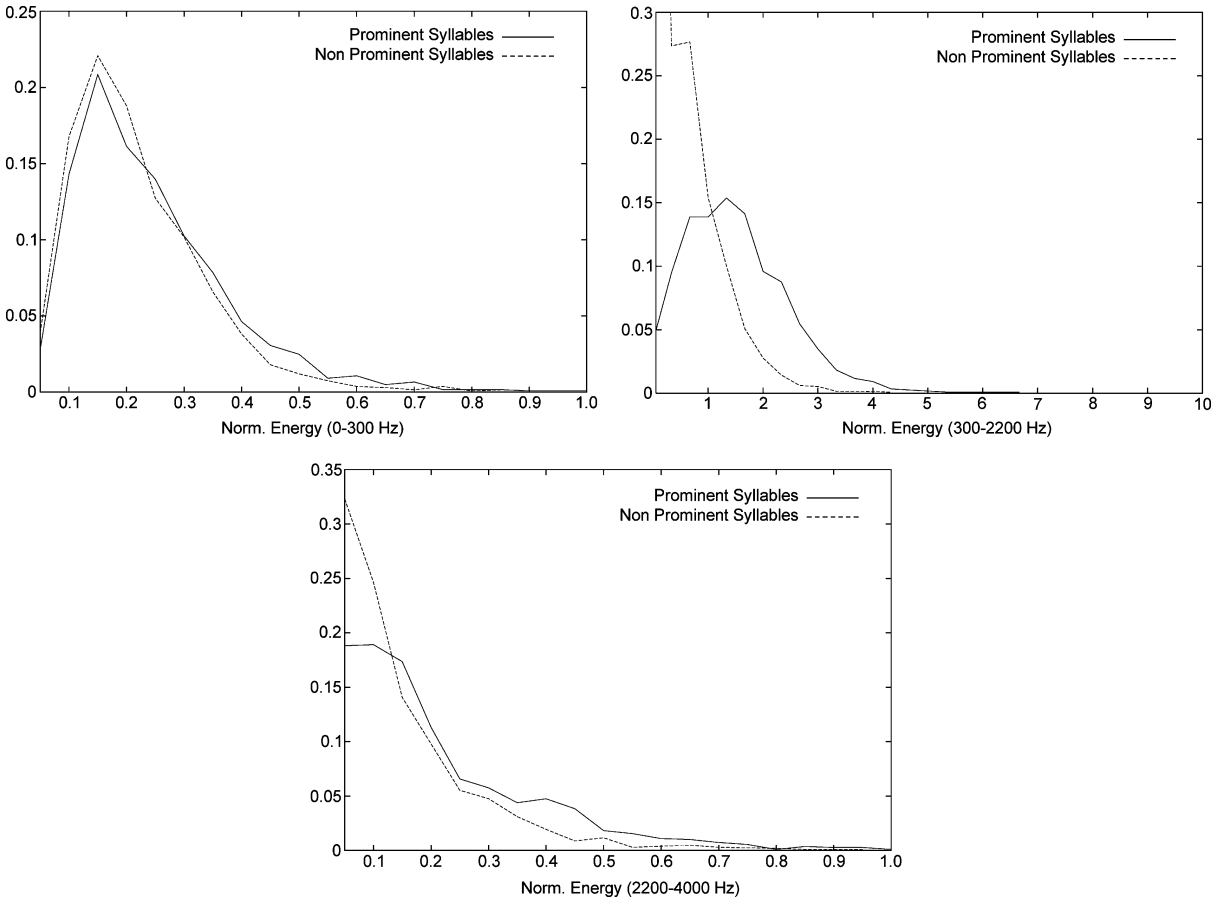

Figure 2. Distributions of prominent and non-prominent syllable nucleus energies in the frequency bands 0–300 Hz (top left), 300–2200 Hz (top right) and 2200–4000 Hz (bottom).

been claimed, that mid-to-high frequency emphasis is a useful parameter in determining stressed syllables. To verify this hypothesis, each nucleus segment spectrum was divided into three bands, making use of band-pass FIR filters, namely from 0 to 300 Hz, from 300 to 2200 Hz and from 2200 to 4000 Hz. The RMS energy of each segment/band pair was successively computed. By examining the distributions of prominent and non-prominent syllable energies in the frequency bands considered (see Fig. 2), the two bands 0–300 Hz and 2200–4000 Hz show a clear overlapping between prominent and non-prominent syllables, while the central band from 300 to 2200 Hz exhibits a clear separation between the two classes. These quantitative results confirm a strict dependence of syllable prominence to vowel mid-to-high frequency emphasis, the frequency band where the main vowel formants reside. Thus, agreeing with the hypothesis suggested by Sluijter and van Heuven, with a view to identifying stress accents (see Section 3.1), we will consider that the spectral emphasis is measured by the energy of this specific frequency band.

3. Prosodic Parameters

This section examines the prosodic quantities, stress accent, pitch accent and prominence, that are the object of the study, and their acoustic correlates. As already mentioned in the introduction, syllables that are perceived as prominent either contain a pitch accent or a stress accent, or both. Thus, prominence can be described by relying on two different prosodic parameters, stress accent and pitch accent, both sufficient to identify a prominent syllable, but none of them necessary to mark a syllable as prominent. These prosodic

parameters can be derived directly from combinations of the four acoustic features described above. The relationships between the prosodic and acoustic parameters define a hierarchy of parameters in which the higher levels are defined and built over the lower ones.

The data used in the following sections are derived from the TIMIT corpus and every syllable was manually classified as prominent or non-prominent. It emerges quite clearly in the following subsections that being able to classify these syllables with respect to the two different accents, instead of classifying them with respect to prominence, would have been preferable, for both the qualitative analysis that we will carry out in this section and the design of the detectors described in Section 4. Unfortunately it is very difficult for humans to distinguish between stress accents and pitch accents when listening to an utterance. It is only possible to reliably perceive, also with considerable difficulty, if a syllable is prominent or not with respect to the surrounding context. That is why we had to perform this study with data classified only with respect to prominence.

3.1. Stress Accent

The main correlates of syllable stress reported in literature are syllable duration and energy (Bagshaw, 1993, 1994; Streefkerk, 1997, 1999). On this topic Sluijter and van Heuven (1996) have introduced a further refinement, confirmed also in a later study (Heldner, 2001), casting some light on the exact correlation between the different acoustic parameters. "Previous research on American English was generally hampered by covariation of stress and (pitch) accent" they claim. Their studies clearly divided the two phenomena, pointing out that the most reliable correlates of syllable stress are syllable duration and mid-to-high frequency emphasis.

In Fig. 3 two sets of prominent and non-prominent syllables are depicted as a function of both log-normalised nucleus duration and log-normalised energy in the 300 to 2200 Hz band. There is a clear evidence supporting Sluijter and van Heuven's ideas: prominent syllables exhibit a longer duration and greater energy in the vowel mid-to-high-frequency band. Although an overlapping region emerges quite clearly from the diagrams, it should be considered that in the model presented here stress accent is only one of the parameters contributing to prominence. In other words, the prominent syllables that cannot be selected on the basis of the process presented in this section can still be identified correctly exploiting the other parameter contributing to prominence, the pitch accent.

3.2. Pitch Accent

There is a long tradition of studies dealing with intonation profiles and pitch accents (Pierrehumbert, 1980; Beckman, 1996; Campione and 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

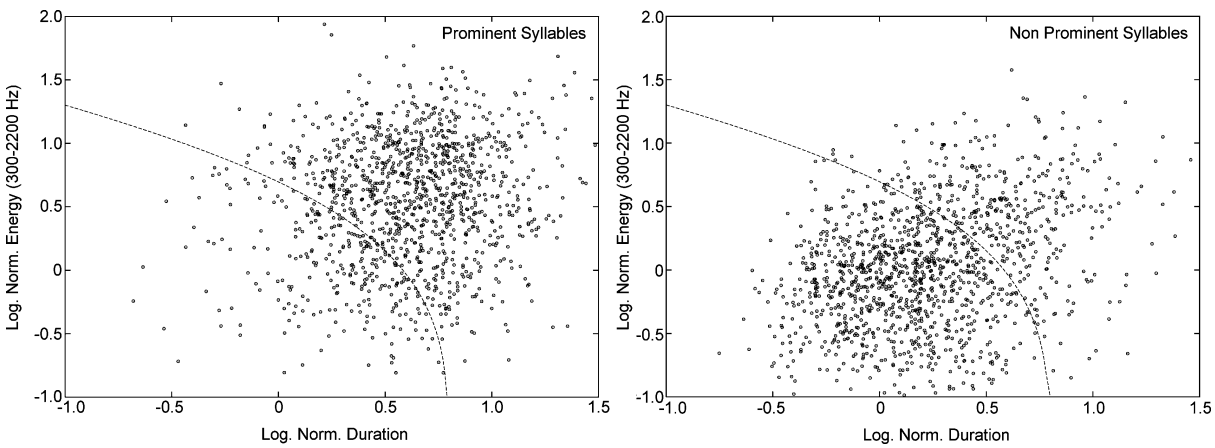

Figure 3. Prominent and non-prominent syllables as a function of log-normalised nucleus duration and log-normalised nucleus energy in the spectral band from 300 to 2200 Hz. The dashed line represents the decision threshold between the two sets computed using the Gaussian discriminator (see Section 4.1).

all possible intonation contours and pitch accents. Unfortunately, such a categorisation, as well as the famous ToBI labelling scheme (Pitrelli et al., 1994), appears to be difficult to implement in an automatic system. Taylor (1992, 1993, 1995, 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 events in the pitch contour (pitch accent shapes and boundary tones). This set, called TILT, consists of five parameters defined as:

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

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

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

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.

Our implementation for the extraction of the pitch shape follows Taylor's proposal. The F0 contour is first converted into an intermediate RFC model. To do that the contour is segmented into frames 0.025 second long; next, the data in each frame are linearly interpolated using a Least Median Squares method to obtain robust regression and deletion of outliers (Rousseeuw, 1987); then every frame interpolating line is classified as rise, fall or connection, depending on its gradient, as suggested in Hieronymous (1989) and Taylor (1993). After that, subsequent frames with the same classification are successively merged into one interval and the duration and amplitudes of the rise and fall sections are measured to finally derive the TILT parameter set.

Having obtained a compact RFC representation of the utterance pitch profile, it is possible to exploit it in order to extract intonational events in the F0 contour. Each of these events is then assigned to the nearest nucleus. The aim is to determine the best combination, among the acoustic and TILT parameters, for identifying the actual pitch accents in the utterance. As described by Taylor (2000), an intonational event that can be considered as a good candidate for pitch accent exhibits a rise followed by a fall in the pitch profile. There are different degrees of such profiles and, in general, rise sections appear to be more relevant for prominence.

Sluijter and van Heuven suggested that the pitch accent can be reliably detected by using overall syllable energy and some measure of pitch variation. As far as pitch variation is concerned, the event amplitude, which is one of the TILT parameters, can be considered as a proper measure, being the sum of the absolute amplitude of the rise and fall sections of a generic intonational event. However, a further refinement can be obtained by multiplying the event amplitude (A_{event}) by its duration (D_{event}) and normalising the product by means of a weighting factor that expresses the relevance of the event along the utterance (R_{event}). This factor is computed by dividing the event amplitude by the maximum pitch variation and the maximum pitch absolute value across the utterance. Figure 4 shows a plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and the product of

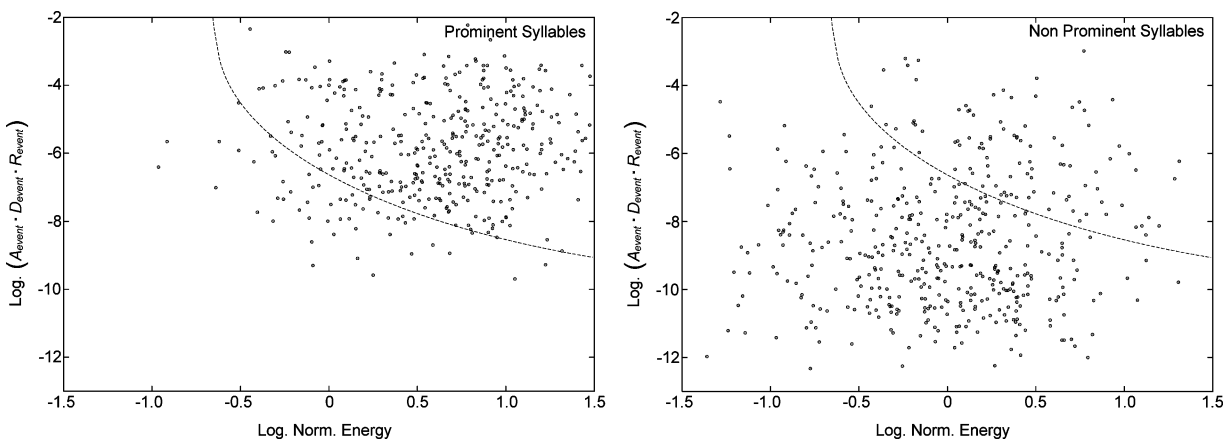

Figure 4. A plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and intonational event parameters (the prominent set contains only syllable with $A_{\text{event}} > 5$ Hz and $D_{\text{event}} > 25$ ms). The dashed line represents the decision threshold between the two sets computed using the Gaussian discriminator (see Section 4.1).

the event parameters, just described, on a log scale. Quite a clear correlation emerges among these parameters when identifying prominent syllables.

3.3. Prominence

In the previous subsection we established some qualitative relationships between acoustic parameters and some prosodic quantities, in particular stress accent and pitch accent. As outlined before, prominence can be defined, from a theoretical point of view, as a combination of these two prosodic features. In particular the presence of either one accent or the other inside a syllable is sufficient for classifying it as prominent in the utterance considered.

Remaining in a qualitative perspective, we have already seen that the acoustic parameters considered in the previous section and the relationships between them presented in the previous subsections should allow us to build a reliable prominence detector. In general we can say, especially by looking at Figs. 3 and 4, that the higher the acoustic parameters, the stronger the accent perception, and thus also the prominence perception.

In the next section two possible detectors of prosodic prominence, based on these qualitative observations, will be introduced.

4. Prominence Detectors

This paper presents two different detectors of prosodic prominence, based on different assumptions and on different theoretical models. The first is grounded on multivariate Gaussian discriminators, so it needs to be tuned with annotated data in order to properly set up the model parameters, while the second is based on the definition of a continuous prominence function and it does not need any training phase.

4.1. The Supervised System

The first prominence detector stems from the observation that the raw data representing acoustic parameters are approximately distributed in a lognormal way. This property can be qualitatively verified by inspection of Figs. 3 and 4, which showed the considered parameter on a log scale. It can be observed that the logarithm of the four acoustic parameters follows, at least approximately, a normal distribution. This attractive property suggests the adoption of an interesting theoretical approach. Modelling the data as lognormal random variable allows us to successfully apply Multivariate Gaussian Discriminator techniques to build a prominence detector with a thorough theoretical basis. Although these techniques are well known in the literature, to the best of the authors' knowledge they have not yet been applied to this problem, probably because it was not recognized that the acustic parameters actually follow a lognormal distribution, at least approximately.

The prominence detector is based on two separate detectors, one for the stress accent and one for the pitch accent. They are defined using the acoustic parameters described above and a syllable is consider as prominent if at least one of the two detectors finds an accent in it.

Considering this framework, it is possible to represent each prominence class by a multivariate Gaussian distribution with the centroid

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

and the sample covariance matrix

$$\mathbf{W} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

as parameters of the distribution, where $\mathbf{x}_i$ is the vector of acoustic parameters of syllable i (log-normalised duration and log-normalised energy—300–2200 Hz—in the case of stress accent detector and log-normalised overall nucleus energy and log(Aevent·Devent·Revent) in case of pitch accent detector), and N is the number of vectors composing the set we want to model. In this way a discriminant function

$$g_{nj}(\mathbf{v}) = -\log|\mathbf{W}_j| - (\mathbf{v} - \boldsymbol{\mu}_j)^{\mathrm{T}}\mathbf{W}_j^{-1}(\mathbf{v} - \boldsymbol{\mu}_j)$$

can be built and used for classifying vectors $\mathbf{v}$ ($|\mathbf{W}_j|$ and $\mathbf{W}_j^{-1}$ are respectively the determinant and the inverse of the sample covariance matrix). If, for example, $g_{n1}(\mathbf{v}_i) > g_{n2}(\mathbf{v}_i)$, then the corresponding syllable i will be classified as belonging to the set represented by the Gaussian $n1$. A similar procedure for designing a multivariate Gaussian discriminator is described, for example, in Gish and Schmidt (1994) and Harrington and Cassidy (1999) applying it to different kind of problems.

The dashed lines in Figs. 3 and 4 represent the decision thresholds between the two sets computed using this method and obtained, in particular, by imposing $g_{n1}(\mathbf{v}) = g_{n2}(\mathbf{v})$.

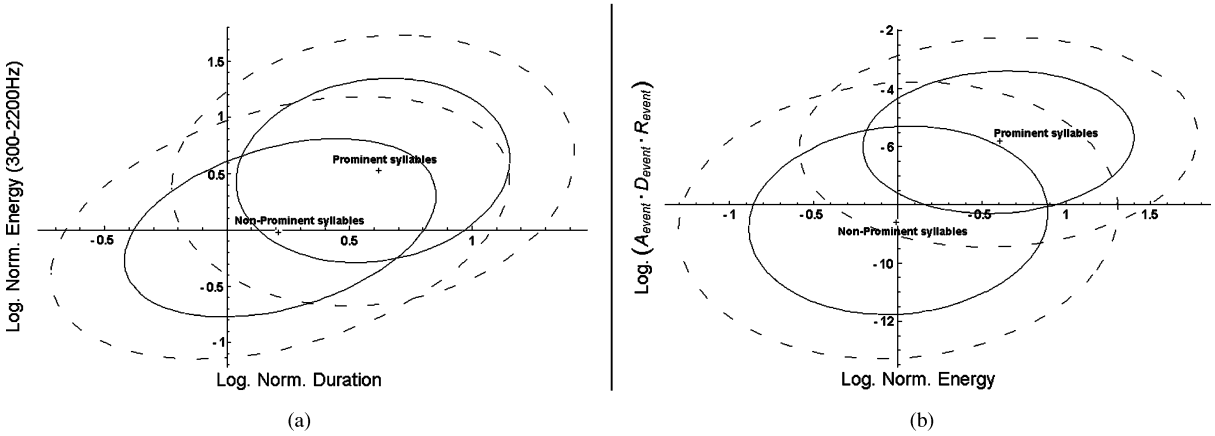

Figure 5. Gaussian mixture that approximates the data shown in Fig. 3 concerning the stress accent (a) and Fig. 4 regarding the pitch accent (b). The ellipses represent the area in which are contained, respectively, 95% (dashed line) and 75% (solid line) of the data for both the classes of prominent and non-prominent syllables.

Figure 5 shows, in a bidimensional picture, the tridimensional Gaussian mixtures that approximate respectively the data of Fig. 3 regarding the stress accent (Fig. 5(a)) and Fig. 4 regarding the pitch accent (Fig. 5(b)). The ellipses represent the area containing, respectively, 95% and 75% of the data for both the classes of prominent and non-prominent syllables.

By combining the two detectors, one for detecting stress accents and the other for detecting pitch accents, on the basis of the methodological considerations presented above, it is possible to produce a reliable prominence detector. Prominent syllables can thus be identified either as pitch accented or stressed syllables.

The parameters involved in the multivariate-Gaussian discriminators ($\boldsymbol{\mu}$ and $\mathbf{W}$) were estimated using a subset of TIMIT utterances, composed of 3634 syllables, spoken by 25 different speakers. The prominence detector was applied to a test set extracted, again, from the TIMIT corpus. The test set consisted of 3643 syllables, uttered by 26 different speakers of American English. The 26 speakers used to test the system were different from the 25 used for parameter estimation, as well as the utterances. Note that thanks to the normalisation processes applied during the computation of the considered acoustic parameters, neither the variations across the utterances of the same speaker nor the variations introduced by different speakers needed to be considered. This prominence detector correctly classified 80.73% of the syllables as either prominent or non-prominent (54.24% are correct rejection and 26.49% correct detection), with an insertion rate of 11.34% (false alarms) and a deletion rate of 7.93% (missed detections). As pointed out before, these results were obtained by training the system on manually classified data with regard to prominence.

4.2. *The Unsupervised System*

According to Taylor (2000), all the prosodic parameters involved in prominence study (namely, prominence, stress and pitch accent) should be considered as continuous quantities, avoiding any kind of categorisation. This view is not usually adopted in linguistics, where there is a tendency to deal with categorical/discrete representations of the examined phenomenon. On the other hand, for testing the reliability of an automatic system, hand-tagged data have to be used, and as manual tagging of utterances for prosodic phenomena is a highly complex task for humans, the introduction of categories seems unavoidable. For these reasons we chose, following Taylor, to describe and manage the prosodic parameters presented in this section as continuous values, to successively introduce some provisional categorisations, following the linguistic point of view, to compare the behaviour and performance of the automatic process with the hand-tagged data.

As suggested in the literature and 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 among the

acoustic parameters, it seems possible to combine them properly to build a "prominence function" able to derive a continuous value of prominence directly from the acoustic features of every syllable nucleus. Our proposal for such a function is:

$$Prom^i = \max\left\{en^i_{300-2200} \cdot dur^i, en^i_{\text{ov}} \cdot \left(A^i_{\text{event}} \cdot D^i_{\text{event}} \cdot R^i_{\text{event}}\right)\right\}$$

where $en_{300-2200}$ is the energy in the 300–2200 Hz frequency band, *dur* is the nucleus duration, en_{ov} is the overall energy in the nucleus, while A_{event}, D_{event} and R_{event} are the same as before (note that if an event is not present in the nucleus, all these values are set to zero). All the parameters refer to a generic i-th syllable nucleus in the utterance examined. Although the *Prom* function definition is somewhat arbitrary and tentative, as all of the empirical functions, it has a rationale, as it was derived in such a way as to mathematically express the fact that a prominent syllable is usually stressed or pitch accented or both and that these prosody parameters can be successfully derived from the acoustic parameters that appear in the formula. However, by contrast with the Gaussian model prominence detector, the syllables are not classified into stress accented or pitch accented before being considered for prominence. Moreover, even the prominence of a syllable nucleus is not two-level quantised into prominent or non-prominent, at least on a syllable-by-syllable basis. This continuous approach is fully justified by considering that the classification into prominent or not prominent cannot be carried out, at least in an optimal way, if the context of the neighbouring syllables is neglected.

As pointed out before, to evaluate the system by comparing it with hand-tagged data, it is necessary to introduce some kind of categorisation in prominence, by considering the prominence level of the syllable compared with its neighbours. 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 implies the search for the local maxima of the *Prom* function defined above. Therefore, in our classifier the prominence value of every syllable nucleus is compared with the two neighbours and, if it represents a maximum, the corresponding syllable nuclei (and the whole syllable) is considered prominent.

However, it is neither impossible nor rare, in American English, for consecutive syllables to be prominent, for example whenever two successive monosyllabic words are both prominent. The two syllables would certainly present a different "level" of prominence, but, in a dichotomic-classification perspective (prominent or non-prominent), levels of prominence cannot be taken into account. The previous peak (maximum) picking algorithm would fail in this case, not recognising one of the two prominent syllables. To partially overcome this problem, the peak picking algorithm was enhanced to tackle this relatively

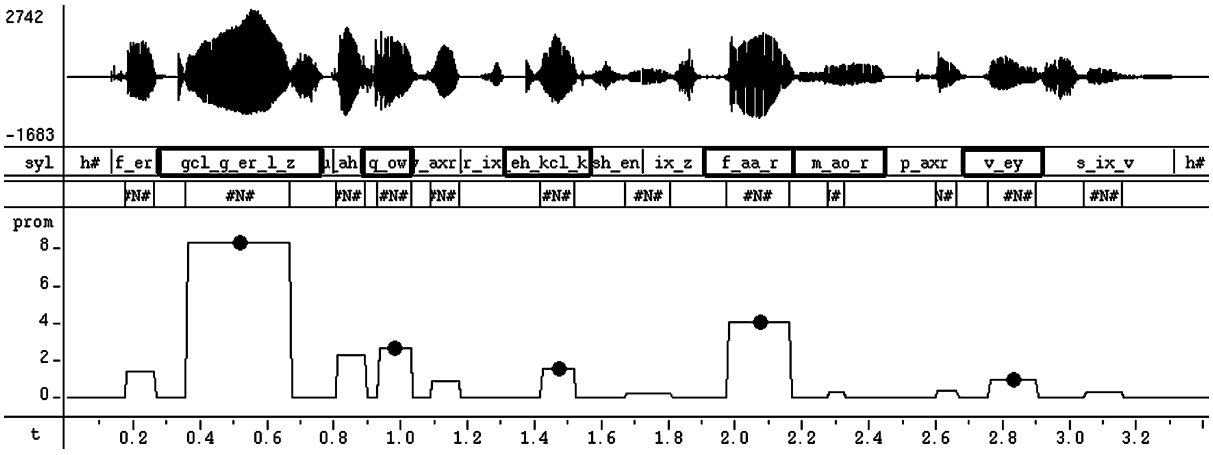

Figure 6. Prosodic prominence function values for the utterance "For girls the overprotection is far more pervasive". Proceeding from the top, we have: the waveform plot, the syllable segmentation (only for comparison purposes), the syllable nuclei as detected by the system (marked by #N#), and finally the prominence values for every nucleus identified by the segmentation procedure. The 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").

frequent case. Whenever two subsequent syllables differ only by 15% of their prominence value, the test is performed by ignoring the neighbours with similar prominence and by considering instead the next syllable nuclei. Moreover, syllables that have a very high prominence value, greater than 70% of the maximum peak in the utterance, are also considered as prominent, independently of the context.

A plot of prominence function for the sentence "For girls the overprotection is far more pervasive" taken from the TIMIT corpus is shown in Fig. 6.

Numerical results show that by making use of the *Prom* function and the enhanced peak picking method described above, it is possible to design a reliable prominence detector. The continuous model system was tested using a subset of TIMIT utterances, composed of 7327 syllables taken from 485 utterances spoken by 51 different speakers of American English. The prominence detector correctly classified 80.61% of the syllables as either prominent or non-prominent (58.65% are correct rejection and 21.96% correct detection), with an insertion rate of 7.22% (false alarms) and a deletion rate of 12.17% (missed detections). As pointed out before, this is an unsupervised system, thus there is no need for any training phase.

5. Conclusions and Discussion

It is widely accepted in the literature that inter-human agreement, when manually tagging prominence in American English continuous speech, is around 80–82% (Pickering et al., 1996; Jenkins and Scordilis, 1996). Both the prominence detectors presented here exhibit an overall agreement of about 80.7% with the data manually tagged by a native speaker, without exploiting any information apart from acoustic parameters derived directly from the utterance waveform. As these results are comparable with those obtained by human taggers, both the prominence detectors can be seen as a valid alternative to manual tagging for building large resources of speech annotated with prominence information useful for language research and teaching.

Although the supervised detector is built on a thorough theoretical basis, it does not outperform the unsupervised detector and tends to make more insertion errors, recognising as prominent syllables that are actually not prominent. It is likely that its overall performance is penalised by the syllable-by-syllable detection, which does not take into account the prominence context. By contrast the unsupervised detector, derived from less theoretical considerations, is context aware, and providing the same level of accuracy without the need of any training information for system tuning, it could be considered preferable, at least from this point of view.

As outlined in the introduction, previous studies tend to use different approaches. Bagshaw (1993) built a prominence detection system for computer aided pronunciation teaching, thus using the utterance transcription to guide the segmentation and the detection process. He obtained a 61.6% of agreement with human-tagged data, that is much less than the one obtained by the systems presented in our work. Jenkin, Scordilis (1996) implemented and compared three different system for prominence detection, all based on theoretical models that require a training phase. The first and best performing system is based on neural networks and achieved a correct classification on 81–84% of cases. The second system uses hidden Markov models and correctly classified syllables with respect to prominence with a precision ranging from 78 to 80%, while the third, and worst performing, rule-based system achieved a score between 67–75%. The best two systems presented in this study reach a level of correct classification close to the one achieved by the unsupervised systems presented here, but all of them require a complex training phase and additional tagged data to do it. Similar considerations can be made about the results obtained by Wightman and Ostendorf (1994) with their system, based on a model that uses decision trees similar to a discrete HMM and an Automatic Speech Recognition module. The model is trained using maximum likelihood estimation and achieves 83% of correct classification when applied to prominence detection.

It would be interesting to test the validity of our approach with different languages. Theoretically, different languages involve different combinations of acoustic parameters or different weightings among them, but the methods presented here should be easily adapted to cope with these inter-language variations. A study in this direction is presently under way considering the Italian language.

References

Auberg, S., Correa, N., Rothenberg, M., and Shanahan, M. (1998). Vowel and intonation training in an English pronunciation tutor. *ESCA-StiLL '98 Proceedings*. Marholmen, Sweden, pp. 69–72.

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 Edinburgh.

Beckman, M.E. (1986). *Stress and Non-Stress Accent*. Dordrecht, Holland: Foris Publications.

Beckman, M.E. and Venditti, J.J. (2000). Tagging prosody and discourse structure in elicited spontaneous speech. *Science and Technology Agency Priority Program Symp. on Spontaneous Speech Proceedings*. Tokyo, pp. 87–98.

Bulyko, I., Ostendorf, M., and Price, P. (1999). On the relative importance of different prosodic factors in improving speech synthesis. *ICPhS '99 Proceedings*, pp. 81–84.

Campione, E. and Veronis, J. (1998). A multilingual prosodic database. *ICSLP '98 Proceedings*. Sydney, pp. 3163–3166.

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

Fach, M. and Wokurek, W. (1995). Pitch accent classification of fundamental frequency contours by HMM. *Eurospeech '95 Proceedings*. Madrid, pp. 2047–2050.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., and Dahlgren, N.L. (1993). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*, (printed documentation for NIST Speech Disc 1-1.1), NTIS order number PB91–100354, 1993.

Gish, H. and Schmidt, M. (1994). Text-independent speaker identification. *IEEE Signal Processing Magazine*, *11*(4):18–32.

Harrington, J. and Cassidy, S. (1999). *Techniques in Speech Acoustics*. Dordrecht, Holland: Kluwer.

Heldner, M. (2001). Spectral Emphasis as an Additional Source of Information in Accent Detection. *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, pp. 57–60.

Hironymous, J.L. (1989). Automatic sentential stress labelling. *Eurospeech '89 Proceedings*. Paris, pp. 226–229.

Hironymous, J.L., McKelvie, D., and McInnes, F.R. (1992). Use of acoustic sentence level and lexical stress in HSMM speech recognition. *ICASSP '92 Proceedings*. San Francisco, California, pp. 225–227.

Hirst, D.J. (2001). Automatic analysis of prosody for multilingual speech corpora. In E. Keller, G. Bailly, J. Terken, and M. Huckvale (Eds.), *Improvements in Speech Synthesis.* Chichester, UK: Wiley.

Howitt, A.W. (2000). *Automatic Syllable Detection for Vowel Landmarks*. PhD Thesis, MIT.

Jenkin, K.L. and Scordilis M.S. (1996). Development and comparison of three syllable stress classifiers. *ICSLP '96 Proceedings*. Philadelphia, USA, pp. 733–736.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal Acoustical Society of America*, *58*(4):880–883.

Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. *ICSLP '96 Proceedings*. Philadelphia, pp. 1457–1460.

Nouza J. (1999). Computer-aided spoken-language training with enhanced visual and auditory feedback. *Eurospeech '99 Proceedings*. Budapest, pp. 183–186.

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

Pierrehumbert, J.B. (1980). *The Phonetics and Phonology of English Intonation*. PhD thesis, MIT.

Pitrelli J., Beckman M., and Hirschberg J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. *ICSLP '94 Proceedings*. Yokohama, Japan, pp. 123–126.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. *Speech Prosody 2002 Proceedings*. Aix-en-Provence.

Rousseeuw, P.J. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Sjölander, K. and Beskow, J. (2000). WaveSurfer—an Open Source Speech Tool. *ICSLP '2000 Proceedings*. Beijing, China.

Sluijter, A. and van Heuven, V. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. *ICSLP' 96 Proceedings*, Philadelphia, pp. 630–633.

Streefkerk, B.M. (1997). Acoustical correlates of prominence: A design for research. *Inst. of Phon. Sciences Proceedings.* University of Amsterdam, vol. 20, pp. 131–142.

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

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*. New York: Elsevier, pp. 495–518.

Taylor, P.A. (1992). *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh.

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

Taylor, P.A. (1995). 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 Acoustical Society of America, 107*(3):1697–1714.

Terken, J. (1991). Fundamental frequency and perceived prominence. *Journal Acoustical Society of America*, *89*(4):1768–1776.

Venkata Ramana, R.G. (2000). Modeling Word Duration for better speech recognition. *Speech Transcription Workshop Proceedings*, University of Maryland, MD.

Warren, P. (1996). Prosody and Parsing: An introduction. *Language and Cognitive Processes*, *11*(1/2):1–16.

Waterson, N. (1987). *Prosodic phonology: The Theory and its Application to Aanguage Acquisition and Speech Processing*. Grevatt and Grevatt: Great Britain.

Wightman, C.W. and Ostendorf, M. (1994). Automatic labelling of prosodic patterns. *IEEE Transaction on Speech and Audio Processing*, *2*(4):469–481.

Yang, Y. and Wang, B. (2002). Acoustic correlates of hierarchical prosodic boundary in mandarin. *Speech Prosody 2002 Proceedings*, Aix-en-Provence.