Automatic Annotation of Speech Corpora for Prosodic Prominence

F. Tamburini and C. Caini

University of Bologna
f.tamburini@cilta.unibo.it ccaini@deis.unibo.it

Abstract

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. Perceptual prosodic prominence is supported by two different prosodic features: pitch accent and stress. 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 phenomena, makes it possible to build automatic systems capable of identifying prominent syllables in utterances with performance comparable with the inter-human agreement reported in the literature without using any kind of information apart the acoustic parameters derived directly from speech waveforms.

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 performances 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 signal its correct interpretation by means of intonation, pauses, prominences, ...

Automatic Speech Recognition systems can take advantage of software modules devoted to prosody management enhancing the global classification performances (Hastie, *et al.* 2001; Hieronymous, *et al.* 1992; Shriberg & Stolcke, 2001), as well as can do Automatic Speech Understanding systems (Beckman & Venditti, 2000; Nöth, *et al.* 2000; Shriberg, *et al.* 1998). Prosodic modules can enhance the fluency and adequacy of automatic speech-generation systems (Bulyko, *et al.* 1999; Portele & Heuft, 1997; Wightman, *et al.* 2000) and it may be extremely useful for solving ambiguities in natural language parsing (Hirschberg & Avesani, 2000; Warren, 1996).

One of the most interesting applications of automatic techniques for handling prosodic phenomena is that of language resource annotation, such as prosodically tagged speech corpora, both for research purposes and for language teaching (Beckman & Venditti, 2000; Campione & Veronis, 1998; Hirst, 2001). In this field the request of prosodically annotated resources is increasing and the difficulty and the prohibitively high costs for a manual production often limit, and have limited, the design of such resources.

One of the most important prosodic features is prominence. "Prominence is the property by which linguistic units are perceived as standing out from their 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 are stressed. On the acoustic/phonetic side, the accomplishment of such features is strictly correlated with particular behaviour of acoustic parameters, either considered as single features or, more commonly, as combinations of them. As well as the works already cited, there are many other studies suggesting that some of the main acoustic correlates of prominence are pitch movements, overall syllable energy, syllable duration and spectral emphasis (Sluijter & van Heuven, 1996; Streefkerk, 1997; Taylor, 2000).

This paper presents a study on the relationships between prosodic prominence and acoustic features with the aim of designing a system for the automatic detection of prominence in speech using only acoustic/phonetic parameters and cues. Our work has been developed restricting the information sources to the utterance waveform, avoiding any other resource that might not be always available, it would certainly be expensive to build, and permanently bound the system to one specific language. The method we will present do not rely on additional phonetic information, such as phone labelling and/or utterance transcriptions as well as the use of complex techniques requiring heavy training phases on manually annotated data such as hidden Markov models, neural networks or similar methods. This study performs an analysis of the correlation between prominence and a set of acoustic features to identify the best acoustic correlates of prosodic prominence and use such information to build a system capable of identifying prosodic prominence in continuous speech.

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.

The data set used in our 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. In this study, the TIMIT annotations are used only for testing and measuring system performance, not as additional information for the prominence detection algorithms.

The rest of the paper is organised as follows. Section 2 presents syllable nuclei identification for duration measures. Section 3 outlines the computation procedures for the other parameters involved in prominence detection. Section 4 presents a study about the combination and relationships of these acoustic features to identify prosodic features such as pitch accents, stress and prominence, and section 5 discusses the automatic detector of prosodic prominence presented in this study. Section 6 draws the conclusions of the work, comparing and discussing the results obtained with the literature examined.

Syllable nuclei identification and duration measures

The linguistic theories of prosodic prominence mentioned above 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; even defining the syllable concept in continuous speech is often misleading. Resyllabification phenomena and ambisyllabic units contribute to making syllables an entity with fuzzy boundaries. Moreover a lot of studies have made clear that the main contribution of prominence to syllable lengthening is concentrated in the vocalic part of it, mainly increasing the syllable nucleus duration (Greenberg, *et al.* 2003; Silipo & Greenberg 1999; van Bergem, 1993; van Kuijk & Boves, 1999). The relevant conclusion, interesting for the present prominence study, 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 as in (Jenkin & Scordilis, 1996; Waterson, 1987), which can be automatically obtained with a higher accuracy level.

To reliably identify the syllable nuclei in the utterance and measure their duration we applied a modified version of the convex-hull algorithm proposed by 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 units which forms the syllable nucleus. The segmentation points were restricted the to the ones derived from the algorithm proposed by Andre-Obrecht (1988) that detects regions of spectrally quasi-stationary speech in the utterance. 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 normalisation, described, for example, in (Neumeyer, 1996).

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

Other acoustic parameters

Fundamental frequency (F0) contour

The extraction of F0 contour, or pitch contour, is another demanding task. Most of the complexity of pitch extraction process resides in candidate selection and post-processing optimisations. 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. 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, based on the algorithm presented in (Talkin, 1995), that is considered in literature the best pitch-tracking method. 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.

Energy measures

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:

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

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.

In the recent literature, and in particular in the influential work of Sluijter & van Heuven (1996), it has 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 the two bands 0-300 Hz and 2200-4000 Hz show a clear overlapping between prominent and non-prominent syllable distributions, while the central band from 300 to 2200 Hz exhibits a clear separation between the two classes. These quantitative results confirm the 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 & van Heuven, with a view to identifying stressed syllables, we will consider that the spectral emphasis is measured by the energy of this specific frequency band.

Prosodic parameters

This section examines the prosodic quantities, stress, 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 and pitch accent, both sufficient to identify a prominent syllable, but none of them necessary to mark a syllable as prominent.

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 phenomena, namely stress and pitch accent, 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 final detector. Unfortunately it is very difficult for humans to distinguish between stress and pitch accents when listening to an utterance. It is only possible to reliably perceive if a syllable is prominent or not with respect to the surrounding context. This lead to a certain degree of overlapping in the study of the involved phenomena.

Stress

The main correlates of syllable stress reported in literature are syllable duration and energy (Bagshaw, 1993; 1994; Streefkerk, 1997; 1999). On this topic Sluijter & van Heuven (1996) have introduced a further refinement, confirmed also in later studies (Heldner, 2001), casting some light on the exact correlation between the different acoustic parameters. Their studies clearly divided the two phenomena involved in supporting prominence perception, pointing out that the most reliable correlates of syllable stress are syllable duration and mid-to-high frequency emphasis.

In figure 1 two sets of prominent and non-prominent syllables are depicted as a function of both log-normalised nucleus duration and log-normalised RMS energy in the 300 to 2200 Hz band. There is a clear evidence supporting Sluijter & van Heuven's ideas: prominent syllables exhibit a longer duration and greater energy in the vowel mid-to-high-frequency band.

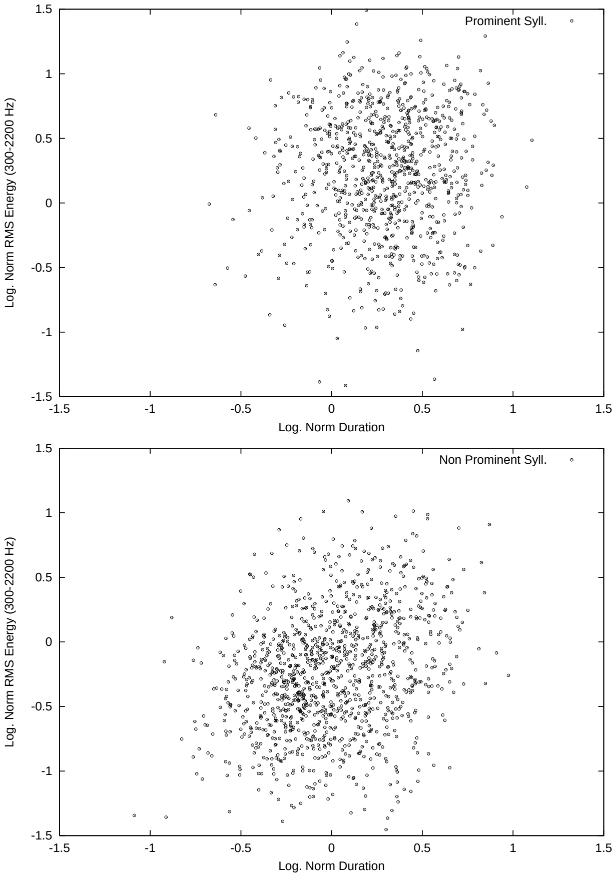

Figure 1: Prominent and non-prominent syllables as a function of log-normalised nucleus duration and log-normalised nucleus energy in the spectral band from 300 to 2200 Hz.

Pitch accent

There is a long tradition of studies dealing with intonation profiles and pitch accents (Pierrehumbert, 1980; Beckman, 1996; Campione & Veronis, 1998). Unfortunately, the categorisations introduced in these studies, as well as the famous ToBI labelling scheme (Pitrelli, *et al.* 1994), appears to be difficult to implement in an automatic system. Taylor (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. This set, called TILT, consists of five parameters defined as:

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

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

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

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.

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 and the data in each frame are linearly interpolated using a Least Median Squares method. Then every frame interpolating line is classified as rise, fall or connection, depending on its gradient, as suggested in (Taylor, 1993), and subsequent frames with the same classification are merged into one interval. The duration and amplitudes of the rise and fall sections are measured to finally derive the TILT parameter set and assign them to the intonational events in the F0 contour extracted from the RFC representation. 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}) to reduce the significance of spike errors. Figure 2 shows a plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and the product of the event parameters on a log scale. Qualitatively, a clear correlation emerges among these parameters when identifying prominent syllables.

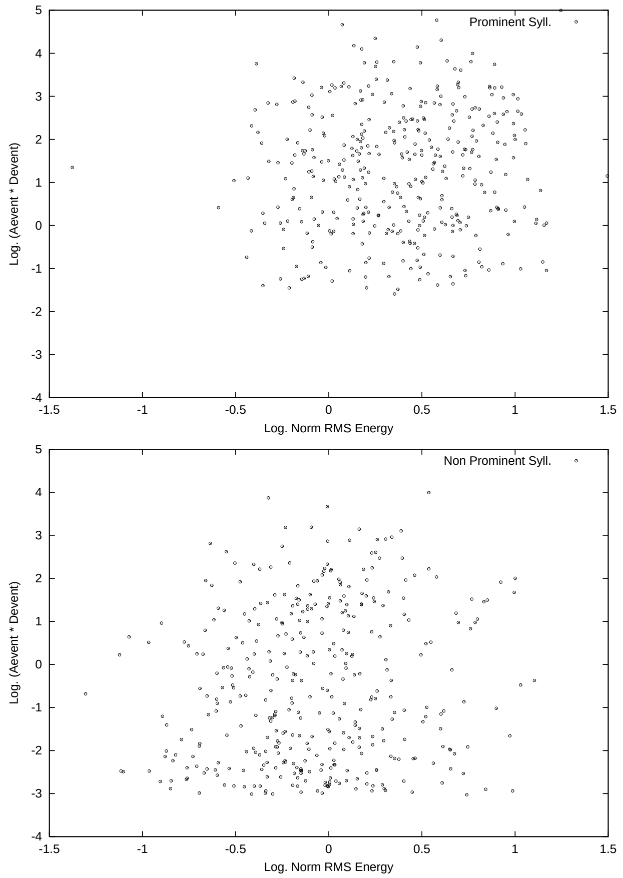

Figure 2: A plot of prominent and non-prominent syllables as a function of overall syllable nucleus energy and intonational event parameters.

Prominence

We have established some qualitative relationships between acoustic parameters and some prosodic quantities, in particular stress and pitch accent. As suggested in the literature and confirmed by our experiments, metrical 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.

The two phonological concept, namely stress and pitch accent, considered in this study as in intermediate level, will help us to relate the acoustic/phonetic parameters with prominence. As we will see in the next section, the relationships between these phenomena and the qualitative observations described before will be useful in defining the behaviour of the prominence detector.

Prominence detector

According to Taylor (2000), all the prosodic parameters involved in prominence study should be considered as continuous quantities, avoiding any kind of categorisation. On the other hand, for testing the reliability of an automatic system, hand-tagged categorical data have to be used. For these reasons we chose to describe and manage the prosodic parameters presented above as continuous values, and successively introduce some provisional categorisations to compare the behaviour and performance of the automatic process with the hand-tagged data.

Bearing in mind the qualitative relationships among the acoustic parameters outlined above, 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:

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

where $en_{300\text{-}2200}$ is the energy in the 300-2200 Hz frequency band, *dur* is the nucleus duration, en_{ov} is the overall energy in the nucleus and A_{event} and D_{event} are the parameters derived from the TILT model. 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, 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. 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, by considering the prominence level of the syllable compared with its neighbours. Following Terken 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, then it is considered prominent. However, it is neither impossible nor rare for consecutive syllables to be prominent, for example whenever two successive monosyllabic words are both emphasised. 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. To partially overcome this problem, the peak picking algorithm was enhanced to tackle this relatively 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 a sentence taken from the TIMIT corpus is shown in figure 3.

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 model was tested using a subset of TIMIT utterances, composed of 5708 syllables taken from 384 utterances spoken by 51 different speakers of American English. The prominence detector correctly classified 80.80% of the syllables as either prominent or non-prominent, with an insertion rate of 8.22% (false alarms) and a deletion rate of 10.98% (missed detections).

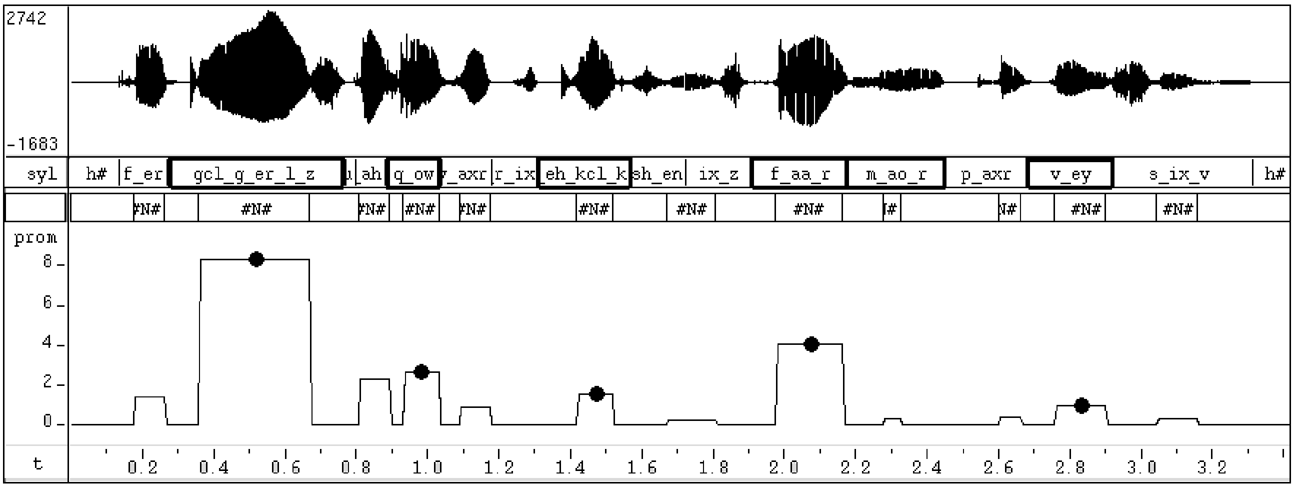

Figure 3: Prosodic prominence function values for the utterance “For girls the overprotection is far more pervasive”. Proceeding from the top, we have: the waveform plot, the syllable segmentation (only for comparison purposes), the syllable nuclei as detected by the system (marked by #N#), and finally 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”).

As pointed out before, this is an unsupervised system, thus there is no need for any training phase.

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 80-90% according to the different number of prominence classes chosen for the annotation (Pickering, *et al.* 1996; Jenkin & Scordilis, 1996). The prominence detector presented here exhibits an overall agreement of 80.80% 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 in the same range of those obtained by human taggers, the prominence detector can be seen as a possible alternative to manual tagging for building large resources of speech annotated with prominence information.

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 obtaining 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 training phases. The most performant is based on neural networks and achieved a correct classification on 81-84% of cases. All systems presented in their study require a complex training phase and additional tagged data to do it. Similar considerations can be made about the results obtained by Wightman & 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 85-86% of correct classification when applied to prominence detection. All these methods make an heavy use of additional information such as phonetic and orthographic transcriptions, segmentation information or ASR systems.

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

Andre-Obrecht, R. (1988). A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals. IEEE Transactions on Acoustics, Speech and Signal processing, 36(1), 29-40.

Bagshaw, P.C. (1993). An investigation of acoustic events related to sentential stress and pitch accents, in English. Speech Communication, 13(3-4), 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. & Venditti, J.J. (2000). Tagging prosody and discourse structure in elicited spontaneous speech. In Proc. of Science and Technology Agency Priority Program Symposium on Spontaneous Speech (pp. 87-98), Tokyo.

Bulyko, I., Ostendorf M. & Price P. (1999). On the Relative Importance of Different Prosodic Factors in Improving Speech Synthesis. In Proc. of ICPhS ‘99 (pp. 81-84), San Francisco.

Campione, E. & Veronis, J. (1998). A multilingual prosodic database. In Proc. of ICSLP ’98 (pp. 3163-3166), Sydney.

Greenberg, S., Carvey, H., Hitchcock, L. & Chang S. (2003). The Phonetic Patterning of Spontaneous

American English Discourse. In Proc. of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo.

Hastie, H.W., Poesio, M. & Isard, S. (2001). Automatically predicting dialog structure using prosodic features. Speech Communication, 36(1-2), 63-79.

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

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

Hirshberg, J. & Avesani, C. (2000). Prosodic disambiguation in English and Italian. In A. Botinis (Ed.), Intonation (pp. 87-95), Kluwer Academic Publisher.

Hirst, D.J. (2001). Automatic analysis of prosody for multilingual speech corpora. In E. Keller, G. Bailly, J. Terken & 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. & Scordilis M.S. (1996). Development and comparison of three syllable stress classifiers. In Proc. of ICSLP '96 (pp. 733-736). Philadelphia.

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. & Price, P. (1996). Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech. In Proc. of ICSLP '96 (pp. 1457-1460). Philadelphia.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (2000). VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System. IEEE Transactions on Speech and Audio Processing, 8(5), 519-532.

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, MIT.

Pitrelli J., Beckman M. & Hirschberg J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In Proc. of ICSLP '94 (pp. 123-126). Yokohama, Japan.

Portele, T. & Heuft, B. (1997). Towards a prominence-based syntesis system. Speech Communication, 21(1-2), 61-72.

Shriberg, E., Baker, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M. & van Ess-Dykema, C. (1998). Can Prosody Aid the Automatic Classification of Dialog Act in Conversational Speech? Language and Speech, 41(3-4), 439-487.

Shriberg, E. & Stolcke, A. (2001). Prosody modeling for automatic speech recognition and understanding. In Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding (pp. 13-16), Red Bank.

Silipo, R. & Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In Proc. of ICPhS '99 (pp. 2351-2354), San Francisco.

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

Streefkerk, B.M. (1997). Acoustical correlates of prominence: a design for research. In Proc. of Inst. of Phon. Sciences, University of Amsterdam, 20, 131-142.

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 Eurospeech '99 (pp. 551-554), Budapest.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W.B. Kleijn & K.K. Paliwal (Eds.), Speech coding and synthesis (pp. 495-518). New York: Elsevier.

Taylor, P.A. (1993). Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model. In Proc. of Eurospeech '93 (pp. 789-792), Berlin.

Taylor, P.A. (1995). The rise/fall/connection model of intonation. Speech Communication, 15(1-2):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.

van Bergem, D. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels. Speech Communication, 12(1), 1-23.

van Kuijk, D. & Boves L. (1999). Acoustic characteristic of lexical stress in continuous telephone speech. Speech Communication, 27(2), 95-111.

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 language acquisition and speech processing. Grevatt and Grevatt: Great Britain.

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

Wightman, C.W., Syrdal, A.K., Stemmer, G. & Conkie, A. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In Proc. of ICSLP 2000 (pp. 71-74), Beijing.