PROSODIC PROMINENCE DETECTION IN SPEECH

Fabio Tamburini
CILTA/DEIS - University of Bologna - Italy
f.tamburini@cilta.unibo.it

ABSTRACT

This paper presents work in progress on the automatic detection of prosodic prominence in continuous speech. Prosodic prominence involves two different phonetic features: pitch accents, connected with fundamental frequency (F0) movements and syllable overall energy, and stress, which exhibits a strong correlation with syllable nuclei duration and high-frequency emphasis. By measuring these acoustic parameters it is possible to build an automatic system capable of correctly identifying prominent syllables with an agreement with human-tagged data comparable with the inter-human agreement reported in the literature. These results were achieved without using any information apart from acoustic parameters.

1 INTRODUCTION

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

One of the most important prosodic features is prominence: a word or part of a word made prominent is perceived as standing out from its environment [23]. A better understanding of how prominence is physically accomplished is a basic step in the construction of tools capable of automatically identifying such phenomena.

This paper presents work in progress on the construction of a system for the automatic detection of prosodic prominence features in speech using only acoustic/phonetic parameters and cues.

Following Beckman's [3] phonological view, further developed by Bagshaw [1, 2], syllables that are perceived as prominent either contain a pitch accent or are somehow "stressed". On the acoustic/phonetic side, the accomplishment of such features has to be strictly correlated with acoustic parameters. As well as the works already cited, there are many studies [15, 16, 17], suggesting that some of the main acoustic correlates of prominence are pitch movements (strictly connected with fundamental frequency - F0), overall syllable energy, syllable duration and spectral emphasis.

The work presented here is divided into two separate steps: the first step involves the automatic identification of syllable-nuclei boundaries to reliably measure the duration feature, while the second one concerns the identification of prominent syllables by means of acoustic measurements. This paper will report on the first experiments conducted on the whole system.

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. In this study the TIMIT annotations are used only for measuring the system performances, not for prominence detection.

Several studies have been conducted in this field for building automatic systems capable of reliably identifying either one acoustic correlate of prominence [5, 7] or a complete set of prosodic parameters [2, 6, 24]. 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.

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, is still an open question.

2 THE ACOUSTIC PARAMETERS

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

2.1 Duration

The linguistic theories of prosodic prominence listed above tend to consider syllable duration as one of the fundamental acoustic parameters for detecting syllable stress. Unfortunately the automatic segmentation of the utterance into syllables is a complex task; in [9] we can find a survey of syllable segmentation algorithms. None of these methods seem to perform well when applied to continuous speech. For these reasons, an alternative duration measure for prosodic prominence detection should be introduced.

One possible measure seems to be the duration of syllable nucleus. Considering some utterances taken from the TIMIT corpus and comparing the duration of the syllable nucleus with the duration of the entire syllable, with respect to prominence, and approximating the

0-7803-7946-2/03/$17.00 ©2003 IEEE.

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

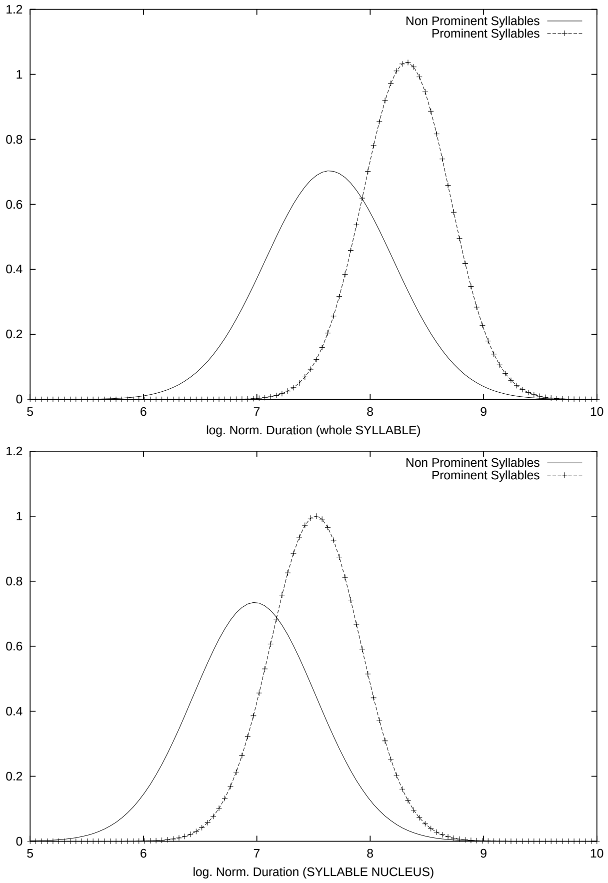

Figure 1: Gaussian approximation of duration measures: whole syllable (top) and syllable nucleus (bottom).

Using a modified version of the convex-hull algorithm [10] applied on the utterance energy profile in the band 300-900Hz as suggested in [9], it is possible to reliably identify the syllable nuclei in the utterance and measure their duration to obtain the acoustic parameter needed for subsequent computations. This duration parameter is normalised, considering the mean duration of the syllable nuclei in the utterance. This is a standard technique for ROS (Rate-Of-Speech) normalisation, as described in [11].

2.2 Energy

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

2.3 Fundamental frequency (F0) contour

The extraction of F0 contour, or pitch contour, is typically a complex task. Bagshaw [2] carried out an accurate comparison of 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 out such variations is often required in order to obtain reliable results. To extract pitch contour we used the ESPS get_f0 program derived from the algorithm presented in [18]. The post-processing phase involves octave-jump removers and profile smoothers, as proposed in [2], applied at different levels and a final interpolation between voiced regions to obtain a continuous profile.

2.4 Spectral emphasis

It has been shown, especially by the influential work of Sluijter & van Heuven [15], that mid-frequency emphasis is one useful parameter in determining stressed syllables. Each nucleus segment has been bandpass-filtered through FIR filters dividing it into three bands: from 0 to 500 Hz, from 500 to 2000 Hz and from 2000 to 4000 Hz. The RMS energy of each segment/band pair was computed. Examining the distributions of prominent and non-prominent syllable energies in the frequency bands considered, we find that the two bands 0-500 Hz and 2000-4000 Hz show a clear overlapping between prominent and non-prominent syllables, while the central band from 500 to 2000 Hz exhibits a clear separation between the two syllable categories. These results confirm a strict dependence of syllable prominence to vowel mid-frequency emphasis.

3 PROSODIC PARAMETERS

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

3.1 Stress detector

The main correlates of syllable stress indicated in the literature are syllable duration and energy [1, 2, 16, 17]. These works were further refined by Sluijter & van Heuven, casting some light on the exact correlation between the different acoustic parameters. Their studies pointed out that the most reliable correlates of syllable stress are duration and mid-frequency emphasis. The presence of a high quantity of energy in the mid-to-high band of vowel spectra, where the main formants reside, is one of the parameters indicating a strong possibility for syllable stress. Figure 2 shows prominent and non-prominent syllables as a function of log. syllable-normalised duration and log. RMS energy in the band from 500 to 2000 Hz. There is strong evidence supporting Sluijter & van Heuven's ideas: stressed syllables exhibit a longer duration and greater energy in the vowel mid-to-high-frequency band. 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 contributing to prominence, the pitch accent.

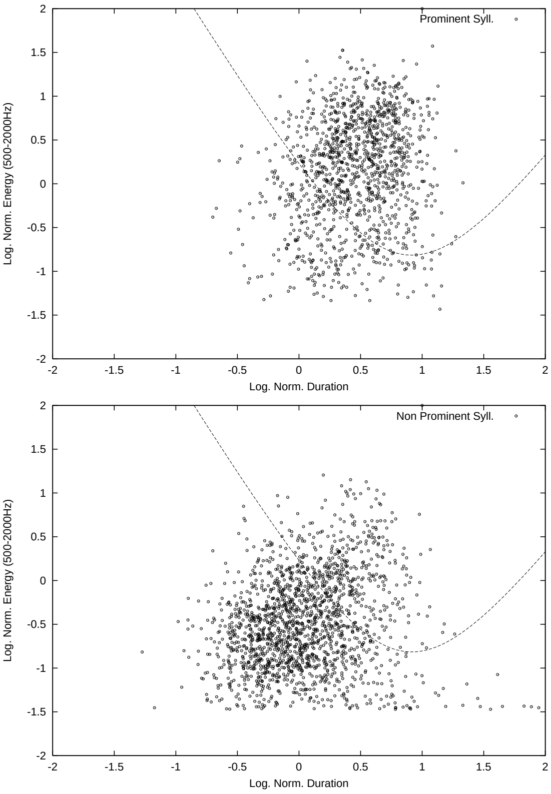

Figure 2: Prominent and non-prominent syllables as a function of log-normalised duration and log-spectral energy in the band from 500 to 2000 Hz.

Considering this general picture, it is possible to represent each set with a multivariate gaussian distribution using the centroid of the set and the sample covariance matrix as parameters of the distribution. In this way a discriminant function can be built and used for classifying general vectors. A similar procedure for designing a multivariate gaussian discriminator is described, for example, in [8]. The dashed line in figure 2 represents the decision threshold between the two sets. We took the log of the two acoustic parameters considered in figure 2 to adapt them to achieve a better fit with a gaussian distribution.

3.2 Pitch accent detector

There is a long tradition of studies dealing with intonation profiles and pitch accents [5, 13]. 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 [19, 20, 21, 22] proposed a different view of intonation events. Starting from a rise/fall/connection (RFC) model, he defined a set of parameters capable of uniquely describing pitch accent shapes and boundary tones, called the TILT parameter set.

Following the model proposed by Taylor, the F0 contour was first converted into an RFC model. The contour was divided into frames 0.025 seconds long, and the data in each frame was linearly interpolated using a Least Median Squares method to obtain robust regression and deletion of outliers [14]. Then every frame line was classified as rise, fall or connection depending on its gradient; subsequent frames with the same classification were merged into one interval and the duration and amplitude of the rise or fall section was measured.

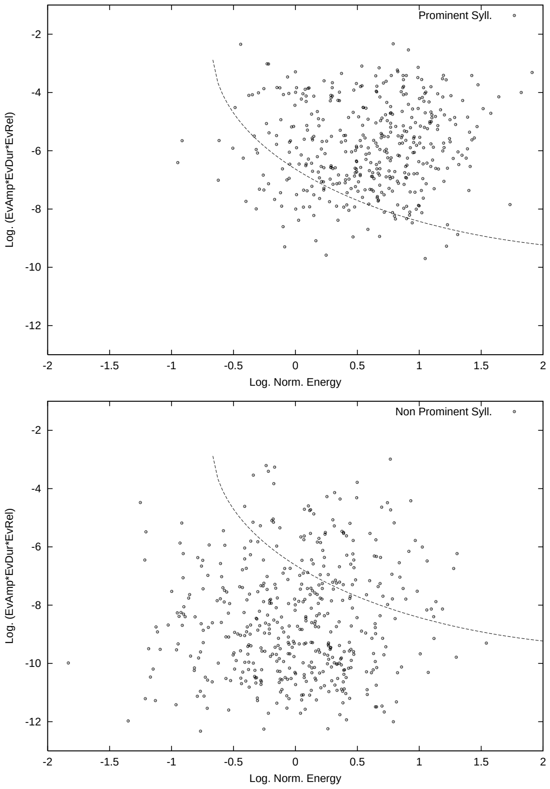

Figure 3: A plot of prominent and non-prominent syllables as a function of overall syllable energy and intonational event parameters (the prominent set contains only syllable with *EvAmp* > 5Hz and *EvDur* > 25ms).

Having obtained a compact RFC representation, it is possible to identify every intonational event in the F0 contour. The view adopted here is to identify every possible event candidate to be a pitch accent, and evaluate the best combination, among the acoustic and TILT parameters, for identifying the actual pitch accents in the utterances. As described by Taylor [22], an intonational event that can be considered a candidate for pitch accent exhibits a rise followed by a fall profile. There are different degrees of such profiles and, in general, rise sections are more relevant for prominence. The actual pitch accents can be found by examining the event amplitude and if necessary some others parameters.

Sluijter & van Heuven suggested that the pitch accent can be reliably detected by using the overall syllable energy and some measure of pitch variation. The event amplitude, 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. Better results can be obtained by multiplying the event amplitude (*EvAmp*) by its duration (*EvDur*) and a further factor that expresses the relevance of the event along the utterance (*EvRel*). Figure 3 shows a plot of prominent and non-prominent syllables as a function of overall syllable energy and the product of event parameters on a log scale. Quite a clear correlation emerges among these parameters when identifying prominent syllables. As in the previous section, the dashed curve represents the threshold for discriminating between the two sets, computed, again, using a multivariate gaussian discriminator.

3.3 Prominence detector

By combining the two detectors described, on the basis of the methodological issues presented above, it should be 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 detectors were estimated using a subset of TIMIT utterances, composed of 3637 syllables, spoken by 25 different speakers. Table 1 shows the results of the prominence detector when applied to a test set extracted from 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 are different from the 25 used for parameter estimation.

	Stressed	Pitch Accented	Stressed+ Pitch Acc.	None
Prominent	650	53	280	271
Non-Prom.	314	41	50	1984

Table 1: The results obtained by applying the prominence detector to the TIMIT test set considered in this study.

The prominence detector correctly classified 81.44% of the syllables as either prominent or non-prominent, with an insertion rate of 11.12% (false alarms) and a deletion rate of 7.44% (missed detections).

4 CONCLUSIONS

It is widely accepted in the literature that inter-human agreement, when manually tagging prominence in continuous speech, is around 80% [12]. The prominence detector presented here exhibits an overall agreement of 81.44% with the data manually tagged by a native speaker; this performance is obtained without using any information apart from acoustic parameters derived directly from the utterance waveform. The results are comparable with those obtained by human taggers, so the presented prominence detector can be seen as a valid alternative to manual tagging for building large resources useful for language research and teaching.

5 REFERENCES

[1] Bagshaw, P.C., “An investigation of acoustic events related to sentential stress and pitch accents, in English.”, *Speech Comm.*, 13, pp. 333-342, 1993.

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

[3] Beckman, M.E., *Stress and non-stress accent*. Foris Publications, Dordrecht, Holland , 1986.

[4] Beckman, M.E. and Venditti, J.J., “Tagging prosody and discourse structure in elicited spontaneous speech.” In Proc. *Science and Technology Agency Priority Program Symp. on Spontaneous Speech*, Tokyo, pp. 87-98, 2000.

[5] Campione, E. and Veronis, J., “A multilingual prosodic database”, In Proc. *ICSLP98*, Sydney, 1998.

[6] Delmonte, R., “SLIM prosodic automatic tools for self-learning instruction”, *Speech Comm.*, 30, pp. 145-166, 2000.

[7] Fach, M. and Wokurek, W., “Pitch Accent Classification of Fundamental Frequency Contours by HMM”. In Proc. *Eurospeech '95*, Madrid, pp. 2047-2050, 1995.

[8] Harrington, J. and Cassidy, S., *Techniques in speech acoustics*, Dordrecht, Holland: Kluwer, 1999.

[9] Howitt, A.W., *Automatic Syllable Detection for Vowel Landmarks*, PhD Thesis, MIT, 2000.

[10] Mermelstein, P., “Automatic segmentation of speech into syllabic units.”, *JASA*, 58 (4), pp. 880-883, 1975.

[11] Neumeyer, L. *et al.*, “Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech.”, In Proc. *ICSLP96*, Philadelphia, pp. 1457-1460, 1996.

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

[13] Pierrehumbert, J.B., *The phonetics and phonology of English intonation*., PhD thesis, MIT, 1980.

[14] Rousseeuw, P.J., *Robust regression and outlier detection*, New York: Wiley, 1987.

[15] Sluijter, A. and van Heuven, V., “Acoustic correlates of linguistic stress and accent in Dutch and American English.”, In Proc. *ICSLP96*, Philadelphia, pp. 630-633, 1996.

[16] Streefkerk, B.M., “Acoustical correlates of prominence: a design for research.”, In Proc. *Inst. of Phon. Sciences,* Vol. 20, University of Amsterdam, pp. 131-142, 1997.

[17] Streefkerk, B M. *et al.*, “Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's.”, In Proc. *Eurospeech '99*, Budapest, pp. 551-554, 1999.

[18] Talkin, D., “A robust algorithm for pitch tracking (RAPT)”, In Kleijn W.B. & Paliwal K.K. (eds.), *Speech coding and synthesis*, New York: Elsevier, 1995.

[19] Taylor, P.A., *A phonetic model of English intonation*, PhD thesis, University of Edimburgh, 1992.

[20] Taylor, P.A., “Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model”, In Proc. *Eurospeech '93*, Berlin, 1993.

[21] Taylor, P.A., “The rise/fall/connection model of intonation.”, *Speech Comm.*, 15, pp. 169-186, 1995.

[22] Taylor, P.A., “Analysis and Synthesis of Intonation using the Tilt Model”, *JASA,* 107 (3), pp. 1697-1714, 2000.

[23] Terken, J., “Fundamental frequency and perceived prominence.”, *JASA*, 89 (4), pp.1768-1776, 1991.

[24] Wightman, C.W. & Ostendorf, M., “Automatic labelling of prosodic patterns.”, *IEEE Transaction on Speech and Audio Processing*, 2 (4), pp. 469-481, 1994.