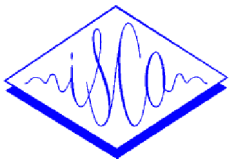

Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

Reliable Prominence Identification in English Spontaneous Speech

Fabio Tamburini

Dipartimento di Studi Linguistici e Orientali
University of Bologna, Italy
f.tamburini@cilta.unibo.it

Abstract

This paper presents a follow up of a study on the automatic detection of prosodic prominence in spontaneous speech. Prosodic prominence involves two different prosodic features, pitch accent and stress, that are typically based on four acoustic parameters: fundamental frequency (F0) movements, overall syllable energy, syllable nuclei duration and mid-to-high-frequency emphasis. A careful measurement of these acoustic 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 even when tested on spontaneous speech.

1. Introduction

This paper presents the preliminary results of a project on the use of automatic prosodic prominence identification methods in spontaneous speech. In [13, 14] we presented an automatic prominence detection system tested on continuous read speech. The techniques and the system presented in this paper are an evolution of the previous system and are able to handle more complex classification tasks such as those involved in spontaneous speech.

Automatic and reliable prominence identification would be very useful in various fields of speech processing. Automatic Speech Recognition systems can take advantage of software modules devoted to prosody management enhancing the global classification performances, as well as can do Automatic Speech Understanding systems. Prosodic modules can enhance the fluency and adequacy of automatic speech-generation systems and prominence is extremely useful for solving ambiguities in natural language parsing.

The review presented by Jun [7] proposed a model of prosodic typology that considered two different aspects of variation: the prominence and the rhythmic pattern of an utterance. She analysed in detail various languages in this perspective by considering the studies performed by some leading scholars using the Autosegmental Metrical model of intonational phonology, and proposed a complete taxonomy applied to the classification of 21 different languages by elaborating the various parameters of the two main lines of classification. The first of these two dimensions, namely prominence, has been studied in detail, following Jun's perspective, using early versions of our automatic prominence identification algorithm, exhibiting interesting results when applied to different languages [15].

Starting from the widely accepted definition of prosodic prominence given by Terken [18], "a word, or part of a word, made prominent is perceived as standing out from its environment", we designed an automatic method to identify prominent sections of an utterance based on the definition of a general prominence function that combines some acoustic parameters directly derived from speech waveforms. It does not require any additional resource such as speech transcriptions (either aligned or not) or any other source of linguistic data to perform the classification process.

In section 2 we present the techniques for computing the acoustic and prosodic parameters which support the prominence phenomenon. In section 3 we propose a definition of the continuous prominence function used in our system as well as a prominence detection criterion and its evaluation. Section 4 draw some conclusions.

2. Acoustic and prosodic parameters

In English, the other Germanic languages, and, more generally, all the stress-accented languages, it is widely accepted that syllables perceived as prominent either contain a pitch accent, a stress, or both [2, 3, 11]. 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. These prosodic parameters can be derived directly from combinations of four acoustic features: syllable duration, spectral emphasis, pitch movements and overall intensity [11].

The linguistic models of prosodic prominence mentioned above agree in considering syllable duration as one of the fundamental acoustic parameters for detecting syllable stress, certainly in 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 speech is often misleading.

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 [10, 19]. 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 [6, 20], which can be automatically obtained with a higher accuracy level.

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. Table 1 outlines the hierarchy of parameters as considered throughout this work, with respect to the different phenomena types.

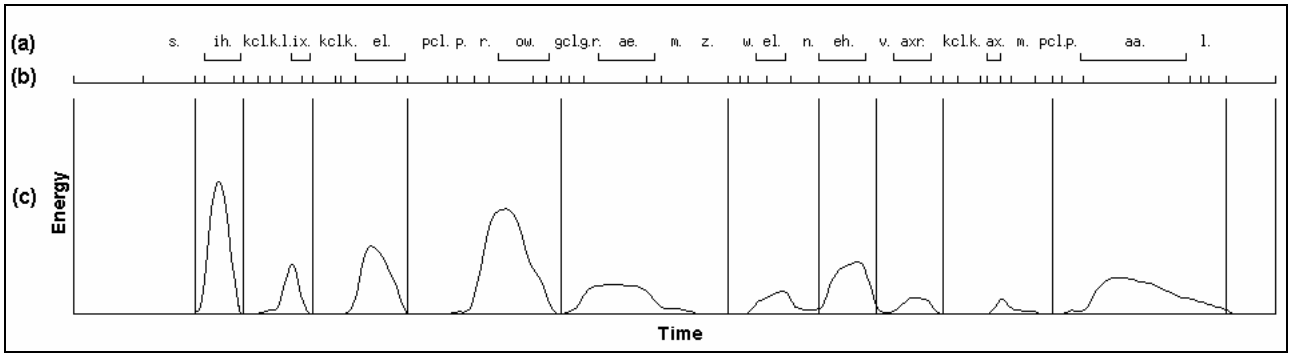

Figure 1: *The second step of the nuclei identification process. (a) The correct nuclei segmentation. (b) SQS regions for the considered utterance. (c) The energy profile and the nuclei isolation hypothesis proposed by the convex-hull based algorithm.*

Perceptual	Prominence			
Prosodic	Stress		Pitch accent	
Acoustic	Nucleus Duration	Spectral emphasis	Pitch movements	Overall intensity

Table 1: *The hierarchy of parameters involved in this study with respect to the phenomena type.*

2.1. Syllable Nuclei identification

To identify the syllable nuclei in the utterance and measure their duration to obtain the acoustic parameter needed for subsequent computations, we applied a three-step algorithm:

a) The first step identifies regions of spectrally quasi-stationary (SQS) speech in the utterance using the algorithm proposed by Andre-Obrecht [1]. Because of coarticulation phenomena between phones, such kind of algorithms tend to identify regions at sub-phonetic level [4], thus, typically, syllabic nuclei, as well as the other phones in the utterance, are split into two or more stationary regions. This method exhibit a good agreement with manually tagged phone boundaries (more than 93% of correctly identified transition points), but generate a lot of false alarms (more than 60% of insertion errors). A technique for grouping such intervals in a proper way must be introduced in the following steps.

b) The second step is based on a modified version of a convex-hull algorithm [8] applied to the utterance energy profile to isolate every syllable nucleus in a separate time region (tr_k, $k=1..m$), containing a peak in the energy profile. The latter was computed by multiplying the contributions of two frequency bands, 800-2000 and 2000-3000 Hz, to filter out energy information not belonging to the vowel unit which forms the syllable nucleus. These two band should contain the vowel F2 and F3 formants: multiplying the energy contribution of these bands is equivalent to a logical conjunction, and represent the request of having two strong formants in the selected spectral band. The segmentation points involved in the convex-hull algorithm were restricted to the ones derived from the first step. This reduces the need for a careful set-up of a large set of thresholds (see figure 1).

c) The last step examines each region *tr* containing a nucleus for determining the nucleus boundaries (see Fig. 2a). As in the previous step the allowed cutting points are only the one proposed by the Andre-Obrecht algorithm (Fig 2b). The energy profile is recomputed using the SQS regions as base for the integration process and it is normalized by considering its maximum value taken in one of the involved SPS regions (see Fig 2c). Let sqs_j, $j=1..n$ be the SQS regions contained into the examined time region *tr* and e_j the normalised energy associated with the SQS region sqs_j. Let us define $Area^+$ as the area built by multiplying the energy of a SQS region by its time duration and $Area^-$ the area of the rectangle obtained considering the remaining part ($Area_j^+ = e_j * time(sqs_j)$, $Area_j^- = (1-e_j) * time(sqs_j)$, where $time(x)$ gives the time span of a SQS region). Let us consider all possible binary partitions $P = (P^+,P^-)$ that we can obtain from the SQS region set *sqs* by identifying a subset of it composed by contiguous SQS. Let us call it P^+ (Fig. 2d). Fig. 3 outlines some of the various possibilities for partitioning the *sqs* set of the example in Fig. 2. By considering the score function

$$Score(P) = \sum_{sps_j \in P^+} Area_j^+ + \sum_{sps_j \in P^-} Area_j^-$$

we can select the partition P' which better approximate the maximum shape as

$$P' = \arg\max_P (Score(P)).$$

The composition of the SQS regions forming the P^+ subset of P' identifies the syllable nucleus.

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

2.2. Acoustic Parameters

Table 2 outlines the acoustic parameters used in the prominence identification algorithm. Previous works [13, 14] describe in detail the procedures for computing these acoustic parameters.

2.3. Prosodic parameters

The main correlates of syllable stress reported in the literature are syllable duration and energy [2, 12]. On this topic Sluijter & van Heuven [11] have introduced a further refinement, confirmed also in a later study [5], claiming that mid-to-high frequency emphasis is a useful parameter in

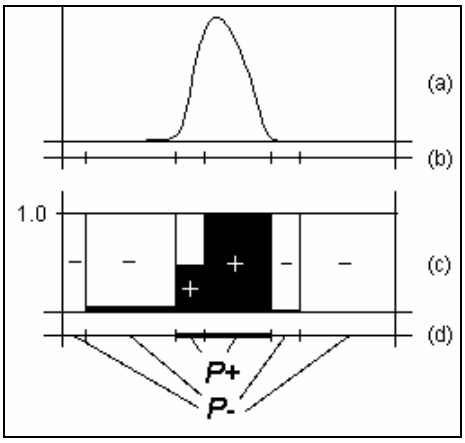

Figure 2: *The nucleus borders identification algorithm. (a) The energy profile inside a time region (tr). (b) the SQS regions forming the time region tr. (c) The energy profile considering the SQS regions as integration intervals. The '+' mark what we called Area$^+$, while the symbols '-' mark the Area$^-$. (d) An example of partition P = (P^+,P^-).*

Acoustic Parameter	Description
Nucleus Duration	Time duration of the syllable nucleus normalised by considering the mean and variance duration of the syllable nuclei in the utterance.
Spectral emphasis	RMS energy computed in the frequency band 500-4000 Hz normalised to the mean and variance of spectral emphasis inside the utterance.
Pitch movements	TILT model [17] representation of pitch movements derived from a pitch contour computed using the ESPS get_f0 program [16].
Overall intensity	RMS energy computed in the frequency band 50-5000Hz normalised to the mean and variance of intensity inside the utterance.

Table 2: *Acoustic parameters used in the prominence identification algorithm.*

determining stressed syllables when replacing the overall energy. Our previous work showed that there is 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 [13, 14].

Sluijter and van Heuven also suggested that pitch accents can be reliably detected by using overall syllable energy and some measure of pitch variation. As far as pitch variation is concerned, the intonational event amplitude, which is one of the TILT model parameters [17], 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. Qualitatively, a clear correlation emerges among overall syllable nucleus energy and the product of the event parameters when identifying prominent syllables [13, 14].

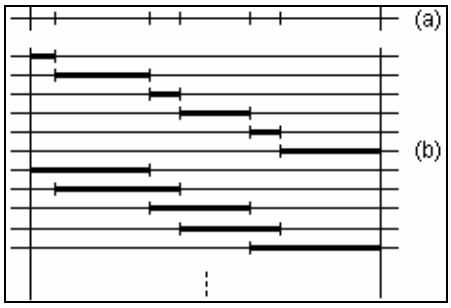

Figure 3: *Some of the possible partitions (b) of the SQS region set in (a).*

3. Prominence identification and evaluation

Bearing in mind the qualitative relationships among the acoustic and prosodic 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:

$$Prom^{i} = en^{i}_{500-4000} \cdot dur^{i} + en^{i}_{ov} \cdot (A^{i}_{event} \cdot D^{i}_{event})$$

where $en_{500\text{-}4000}$ is the energy in the 500-4000 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. It is slightly different from the one we used in our previous work, but the global recognition results were enhanced by using such a modified function. Although the *Prom* function is somewhat arbitrary and tentative, as all of the empirical functions, it has a rationale: normalised stress and normalised pitch accent values, the two arguments of the '+' operator, are summed together because both contribute to support prominence in a reinforcement fashion.

Considering the syntagmatic nature of prominence definition, identifying prominent syllables implies a search for the local maxima of the *Prom* function defined above. Therefore, in our classifier the prominence value of each syllable nucleus is compared with the two neighbours and, if it represents a maximum, then the corresponding syllable is considered prominent.

The model was extensively tested using three different corpora based on different speech type. Table 3 describes the subcorpora used in the tests and table 4 outlines the classification performances of the proposed system when applied to these corpora.

A plot of prominence function and the results of the detection algorithm for a sentence taken from the TIMIT corpus are shown in figure 4.

4. Conclusions

In this paper we presented work in progress for the automatic identification of prosodic prominence in spontaneous speech.

It is widely accepted in the literature that inter-human agreement, when manually tagging prominence in English continuous speech, is around 80-90% according to the

different number of prominence classes chosen for the annotation [6, 9]. The prominence detector presented here exhibits an overall agreement of about 80% with the data manually tagged, without exploiting any information apart from acoustic parameters derived directly from the utterance waveform even for spontaneous speech. It can be seen as a possible alternative to manual tagging for building large resources of speech annotated with prominence information or speech processing systems able to manage such prosodic information. Some preliminary experiments on other stress-accented languages, using read speech, produced similar encouraging results [15].

5. References

[1] Andre-Obrecht, R., 1988, A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals, *IEEE Trans. on ASSP*, 36(1), 29-40.

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

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

[4] Glass, J., Zue, V., 1988, Multi-level acoustic segmentation of continuous speech, In *Proc. of ICASSP'88*, New York, 429-432.

[5] Heldner, M., 2001, Spectral Emphasis as an Additional Source of Information in Accent Detection, In *Proc. of ISCA Tut. and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, 57-60.

[6] Jenkin, K.L. & Scordilis M.S., 1996, Development and comparison of three syllable stress classifiers. In *Proc. of ICSLP '96*, Philadelphia, 733-736.

[7] Jun, S., 2005, Prosodic Typology, In S. Jun (ed.), *Prosodic models and transcription: Towards prosodic typology*, Oxford University Press, 430-458.

[8] Mermelstein, P., 1975, Automatic segmentation of speech into syllabic units, *J. Acoust. Soc. Amer*., 58(4), 880-883.

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

[10] Silipo, R. & Greenberg, S., 1999, Automatic transcription of prosodic stress for spontaneous English discourse. In *Proc. of ICPhS '99*, San Francisco, 2351-2354.

[11] Sluijter, A., van Heuven, V., 1996, Acoustic correlates of linguistic stress and accent in Dutch and American English, In Proc. of *ICSLP' 96*, Philadelphia, 630-633.

[12] 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 Proc. of *Eurospeech '99*, Budapest, 551-554.

[13] Tamburini, F., Caini, C., 2004, Automatic Annotation of Speech Corpora for Prosodic Prominence, In *Proc. LREC-CPSLC workshop*, Lisbon, 53-58.

[14] Tamburini, F., Caini, C., 2005, An automatic system for detecting prosodic prominence in American English continuous speech, *Int. J. of Speech Tech.*, 8(1), 33-44.

[15] Tamburini F., 2005, Automatic Prominence Identification and Prosodic Typology. In *Proc. InterSpeech 2005*, Lisbon, 1813-1816.

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

[17] Taylor, P.A., 2000, Analysis and Synthesis of Intonation using the Tilt Model, *J. Acoust. Soc. Amer*., 107(3), 1697-1714.

[18] Terken, J., 1991, Fundamental frequency and perceived prominence, *J. Acoust. Soc. Amer*., 89(4):1768-1776.

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

[20] Waterson, N. (1987). *Prosodic phonology: The theory and its application to language acquisition and speech processing*. Grevatt and Grevatt: Great Britain.

Corpus	Speech Type	#Utts	#Sylls	#Speakers
TIMIT	read speech	382	4780	51 (31m, 20f)
AixMarsec	radio news	43	704	3 (2m, 1f)
HCRC maptask	spont. speech	62	901	10 (5m, 5f)

Table 3: *The subcorpora used to measure the system performances.*

Corpus	Error rate	Insertions	Deletions
TIMIT	18.64%	9.52%	9.12%
AixMarsec	18.89%	10.37%	8.52%
HCRC maptask	20.75%	8.99%	11.76%

Table 4: *System performances in prominence identification.*

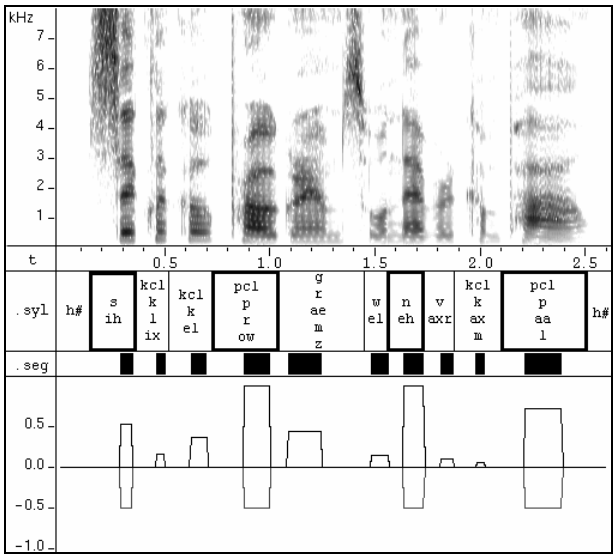

Figure 4: Prosodic prominence function values for the utterance "Cyclical programs will never compile". Proceeding from the top, we have: the spectrogram plot, the syllable segmentation (only for comparison purposes), the syllable nuclei as detected by the system, and finally the prominence values for every nucleus identified by the segmentation procedure (above the axis). The prominent nuclei, as identified by the automatic system, are marked below the axis, while prominent syllables, as classified by a human listener, are indicated by a thick box in the syllable segmentation tier ("syl").