Automatic Prominence Identification and Prosodic Typology

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 continuous 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, 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. This automatic system has been used to cast light on the actual correlation among the acoustic parameters and the prominence phenomenon from an typological point of view, by examining data derived from some stress-accented languages.

1. Introduction

This paper presents the preliminary results of a project on the use of automatic prosodic prominence identification methods in continuous speech for investigating the contribution of the prominence phenomenon to prosodic language typology.

The review presented by Jun [10] proposed a model of prosodic typology that considered two different aspects of variation: the prominence and the rhythmic pattern of an utterance (this view is supported by other scholars, for example [6]). 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 presented in her book [11], 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 by many scholars (see for example [3, 25]) and the results, from a language typology point of view, seem to be fairly uncontroversial among researchers. Languages can be typologically broadly classified into four categories:

- stress-accented,
- lexical pitch-accented,
- non stress-accented and non lexical pitch-accented,
- tonal,

though most scholars suggest that the most promising view of these classes is a continuum, instead of considering a rigid classification [8, 14].

On the other hand, the rhythmic dimension, the traditional classification in stress-timed, syllable-timed, and mora-timed languages is less accepted and the basic concept of isochrony is often seen as problematic. There are studies, based on experimental data, that tend to criticise this traditional view [16, 27], while others provide data and classification methods that tend to support it [13, 17].

In this paper we concentrate our attention on the prominence dimension: starting from the widely accepted definition of prosodic prominence given by Terken [25], "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.

Using this algorithm, described briefly in section 2, we investigated the correlations between acoustic parameters and prominence in some stress-accented languages (section 3). Section 4 discusses the preliminary results obtained and examines some future issues.

2. Automatic Prominence Identification

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, 19]. 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: nucleus duration, spectral emphasis, pitch movements and overall intensity [19]. 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 outline the hierarchy of parameters as considered throughout this work, with respect to the different phenomena types.

The automatic prominence identification method proposed in this section is described in detail in our previous work [21, 22] and it is mainly referred to American English language. Some enhancements of the basic method were introduced to achieve better recognition performances.

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. Speech segmentation

To identify the syllable nuclei in the utterance and measure their duration to obtain the acoustic parameter needed for subsequent computations, we applied a modified version of the convex-hull algorithm [15] to the utterance energy profile. This was computed by multiplying the contributions of two frequency bands (640-2800 and 2000-3000 Hz [5]), to filter out energy information not belonging to vowel units which forms the syllable nucleus. The segmentation points were restricted to the ones derived from the algorithm proposed by Andre-Obrecht [1] that detects regions of spectrally quasi-stationary speech in the utterance.

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 [21, 22] describe in detail the procedures for computing these acoustic parameters.

Acoustic Parameter	Description
Nucleus Duration	Time duration of the syllable nucleus normalised by considering the mean duration of the syllable nuclei in the utterance.
Spectral emphasis	RMS energy computed in the frequency band 500-4000 Hz normalised to the maximum spectral emphasis inside the utterance.
Pitch movements	TILT model [24] representation of pitch movements derived from a pitch contour computed using the ESPS get_f0 program [23].
Overall intensity	RMS energy computed in the frequency band 50-5000Hz normalised to the maximum intensity inside the utterance.

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

2.3. Prosodic parameters

The main correlates of syllable stress reported in the literature are syllable duration and energy [2, 20]. On this topic Sluijter & van Heuven [19] have introduced a further refinement, confirmed also in a later study [9], claiming that mid-to-high frequency emphasis is a useful parameter in 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.

Sluijter and van Heuven also 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 intonational event amplitude, which is one of the TILT model parameters [24], 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.

2.4. Prominence identification

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:

$$Prom^{i} = en^{i}_{500-4000} \cdot dur^{i} + en^{i}_{ov} \cdot (A^{i}_{event} \cdot D^{i}_{event}) \qquad (1)$$

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 on American English were enhanced by using such a modified function.

Considering the syntagmatic nature of prominence definition, identifying prominent syllables implies the 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 tested using a subset of TIMIT utterances, composed of 4780 syllables taken from 382 utterances spoken by 51 different speakers of American English. The prominence detector correctly classified 81.05% of the syllables as either prominent or non-prominent, with an insertion rate of 7.28% (false alarms) and a deletion rate of 11.67% (missed detections).

A plot of prominence function and the results of the detection algorithm for a sentence taken from the TIMIT corpus are shown in figure 1.

3. Prominence and Prosodic Typology

By considering the potentialities of the method presented, it will be interesting, in an interlinguistic perspective, to introduce a variation in the *Prom* function definition. By modifying its structure as follows

$$Prom^{\mathrm{i}} = \Phi(en^{i}_{500-4000},\alpha)\cdot\Phi(dur^{i},\beta)+ \Phi(en^{i}_{ov},\gamma)\cdot\Phi(A^{i}_{event}\cdot D^{i}_{event},\delta) \qquad (2)$$

it is possible to hypothesise that the *Prom* function plays the role of a cross-language component, while the vector of parameters **w** = (α, β, γ, δ) adapts the method behaviour to a specific language, enhancing or reducing the contribution of each acoustic parameter to prominence identification through the application of a weighting function Φ. For this study we chose $\Phi(x, y) = (x + 1)^y - 1$ for its mathematical properties.

On this basis we designed a completely different set of experiments testing this idea on some stress-accented languages. Using the correct segmentation of the utterances provided with the respective corpora, we computed the acoustic and prosodic parameters described in section 2 and applied the prominence function (2) to three different languages: American English, Dutch and Italian. A parametric scanning in the search space of each **w** component allowed us to determine the optimal combination to obtain the maximum agreement with manually annotated data for prominence identification. The same process was repeated for each of the languages considered.

For a better adaptation of the *Prom* function to the different characteristics of the various languages we introduced a new parameter representing the alignment type (*at*) between the intonational events of the TILT model and the syllable segments composing the utterance. In English the rise part of the intonational event has been recognised as the most relevant section to align the event with the corresponding syllable, but it is not necessarily true for other languages. Figure 2 shows the different intonational event timing interval chosen as reference to determine the syllable associated with this event.

The data used for these experiments are taken from the TIMIT corpus for American English, from IFA corpus for Dutch and from a small corpus built by the author for Italian. The latter is composed of utterances extracted from radio news. Table 3 describes briefly the composition of the corpora in the study.

Language	#Utteran.	#Syllables	#Speakers	Ref.
AmEnglish TIMIT	382	4780	51 (20F, 31M)	[7]
Dutch IFA	103	2006	7 (4F, 3M)	[26]
Italian RADIO	29	801	9 (4F, 5M)	-

Table 3: Composition of the data sets used for the experiments.

Table 4 shows the results obtained from the languages analysed.

In accordance with previous studies, in American English [12, 24] and Dutch [18] the most relevant anchor point appears to be the rise section of the intonational event, while Italian tends to prefer the event maximum [4]. With regard to the other acoustic parameters, English presents a configuration in which every parameter is considered roughly equally important, whereas Dutch tends to disfavour spectral emphasis and in Italian energy measures seems to be slightly less important than duration and pitch movements.

	at	α	β	γ	δ
AmEnglish	1	0.9	1.0	0.9	0.8
	Acc. = 82.5% Rec. = 75.6% Prec. = 77.9%				
Dutch	1	0.3	1.0	0.8	0.9
	Acc. = 81.7% Rec. = 79.1% Prec. = 66.3%				
Italian	3	0.7	1.0	0.5	1.0
	Acc. = 81.6% Rec. = 77.6% Prec. = 61.9%				

Table 4: Results on prominence identification with respect to the considered parameters. Accuracy, Recall and Precision of the identification process are outlined after the parameter combination that leads to the best results.

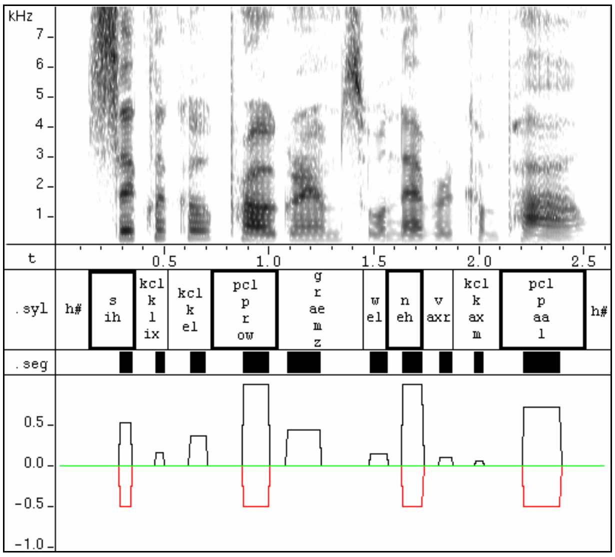

Figure 1: Prosodic prominence function values for the utterance “Cyclical programs will never compile”. Proceeding from the top, we have: the spectrogram plot, the syllable segmentation (only for comparison purposes), the syllable nuclei as detected by the system, and finally the prominence values for 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”).

4. Conclusions

In this paper we presented work in progress for the automatic identification of prosodic prominence in continuous speech. The prominence detector presented here exhibits an overall agreement of more than 81% with the data manually tagged by an English native speaker, without exploiting any information apart from acoustic parameters derived directly from the utterance waveform.

A different set of experiments was performed using some parts of this detector trying to cast light on the actual correlation among the acoustic parameters and the prominence phenomenon from an interlinguistic point of view.

It must be pointed out that these are preliminary results and it is necessary to investigate in greater depth the relationships between these parameters, supplementing these results with more data and more languages.

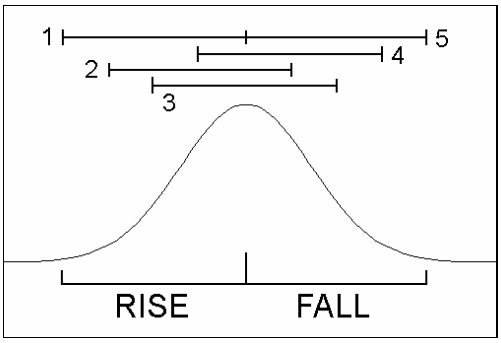

Figure 2: Alignment parameter definition. The numbers near the intervals represent the *at* parameter values.

5. References

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

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

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

[4] D'Imperio, M., "Italian intonation: An overview and some questions", *Probus*, 14:37-69, 2002.

[5] Espy-Wilson, C.Y., "A feature-based semivowel recognition system", *JASA*, 96:65-72, 1994.

[6] Fitzpatrik, J., "On intonational typology", In P. Siemund (ed.) *Methodological Issues in Language Typology. Sprachtypologie und Universalienforshung*, 53:88-96, 2000.

[7] Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S., and Dahlgren N.L., *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*, NIST order number PB91-100354, 1993.

[8] Grabe, E., "Variation Adds to Prosodic Typology", In *Proc of Speech Prosody 2002*, Aix-en-Provence, 127-132, 2002.

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

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

[11] Jun, S. (ed.), *Prosodic models and transcription: Towards prosodic typology*, Oxford University Press, 2005.

[12] Ladd, D.R., Faulkner, D., Faulkner, H. and Schepman, A., "Constant 'segmental anchoring' of F0 movements under changes in speech rate", *J. Acoust. Soc. Amer.*, 106(3): 1543-1554, 1999.

[13] Low, E.L., Grabe, E., Nolan, F., "Quantitative characterisation of speech rhythm: Syllable-timing in Singapore English", *Language and Speech*, 43(4),377-401, 2001

[14] McCawley, J.D., "What is a tone language?", In V. Fromkin (ed.), *Tone: A Linguistic Survey*, Orlando: Academic Press, 113-131, 1978.

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

[16] Pamies Bertran, A. "Prosodic Typology: on the Dychotomy between Stress. Timed and Syllable-Timed Languages", *Language Design*, 2:103-130, 1999.

[17] Ramus, F., Nespor, M., Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition*, 73:265-292, 1999.

[18] Rietveld, T., Kerkhoff, J., "The temporal alignment of L*H Accents", In *Proc of Speech Prosody 2002*, Aix-en-Provence.

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

[20] Streefkerk, B M., Pols L.C.W. , ten Bosch L.F.M., "Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's", In Proc. of *Eurospeech '99*, Budapest, 551-554, 1999.

[21] Tamburini, F., Caini, C. "Automatic Annotation of Speech Corpora for Prosodic Prominence", In *Proc. Compiling and Processing Spoken Language Corpora workshop - LREC-CPSLC*, Lisbon, 53-58, 2004.

[22] Tamburini, F., Caini, C. "An automatic system for detecting prosodic prominence in American English continuous speech", *International Journal of Speech Technology*, 8(1):33-44, 2005.

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

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

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

[26] van Son, R.J.J.H., Binnenpoorte, D., van den Heuvel, H., Pols, L.C.W., "The IFA Corpus: a Phonemically Segmented Dutch 'Open Source' Speech Database", In Eurospeech 2001, Aalborg, 2051-2054, 2001.

[27] Warner, N., Arai, T. "Japanese Mora-Timing: A Review", *Phonetica* 58(1-2):1-25, 2001.