Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Pitch behavior detection for automatic prominence recognition

*Giovanni Abete[1], Francesco Cutugno[2], Bogdan Ludusan[2], Antonio Origlia[2] **

[1] Friedrich-Schiller-Universität, Jena, Germany
[2] LUSI-lab, Department of Physical Sciences, "Federico II" University, Naples, Italy
giovanni.abete@libero.it, {cutugno, ludusan}@na.infn.it, antori@gmail.com

Abstract

In this paper a non-supervised approach for automatic syllable prominence recognition is presented. Previous research in this field showed that syllable nuclei energy and duration are the main cues for prominence detection. The role of the fundamental frequency has also been investigated in the past but was considered secondary or irrelevant for this task. The proposed system uses the energy and the duration of the nucleus while taking into account also the pitch behavior. The algorithm was tested by comparing its results with the annotations of two human experts and a 5.6% accuracy increase with respect to the system not using the pitch behavior was found.

Index terms: prominence detection, pitch behavior, syllable

1. Introduction

An important role in speech processing and speech understanding is played by the systematic, prosodically driven phonetic variation [1, 2, 3]. Prominence is one of the main sources of this type of variation: prominent syllables exhibit several kinds of articulatory expansion, such as vowel lengthening, strengthening of consonants, or hyper-articulation [4, 5, 6, 7, 8].

Formalizing an appropriate model for the prominence effects is therefore important for automatic speech recognition and speech synthesis [9, 10, 11]. A first step in these types of applications consists in elaborating efficient algorithms for automatic prominence annotation [e.g. 12, 13, 14, 15]. Developing such algorithms has both theoretical and practical implications because it sheds light on the complex relation between prominence, which is essentially a perceptual reality [16], and its multilayer acoustic correlates [17, 18, 7].

This paper presents an approach employing classic features for automatic prominence detection along with an analysis of the fundamental frequency. This analysis goes beyond simple measurements and considers, instead, the occurrence of specific prosodic events to improve the classification accuracy with respect to the human experts' annotations.

Acoustic correlates of prominence in Italian have not been thoroughly inspected: many researches were limited to lexical stress [e.g. 19] while only a few study the prominence correlates at levels higher than the word [e.g. 20, 21, 22, 7]. This literature outlines the role of duration and intensity as main correlates of prominence, while F0 is documented to have a secondary importance.

In this work we will show that pitch behavior, while being nearly irrelevant in many cases, acquires a greater importance than classical features in specific, localized situations.

2. The corpus

The data we worked with are a subset of the SPEECON [23] corpus, a collection of speech data recorded in more European languages with the goal of developing voice-driven interfaces for consumer applications. It was collected to ensure "wide range of speaking styles, voice qualities and regional influences". The subset contains 288 natural numbers with at least five syllables (15 on average; tot. 4265), read by over 100 male speakers from several Italian regions.

In Figure 1 we show the length distribution of sentences in the considered subset.

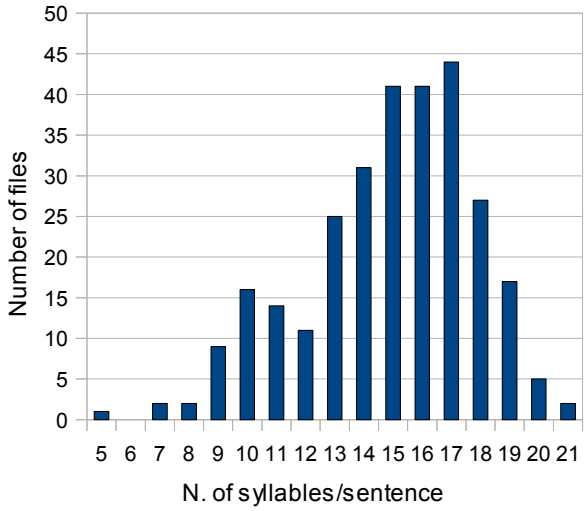

Figure 1. *Length distribution for sentences in the employed SPEECON subset*

The choice of the corpora was done taking into account that the information obtained from the prominence detection stage will be further used for a speech recognition task. Attempts to implement this kind of approaches for speech recognition will be made easier by the use of a language model partly based on a regular grammar. Moreover, choosing to work with read numbers helped the annotators not to be influenced by semantic and pragmatic factors. As it has been shown in [27, 28] these elements can significantly affect the manual annotation procedure, thus hindering the comparison with the algorithm performance.

* Authors appear in strict alphabetical order.

Even if we are dealing with read numbers, the material exhibits many features belonging to connected speech, like coarticulation, hesitations, deaccentuations, etc. Moreover, the speakers display strong regional inflections, thus showing many different rhythmic patterns. The prosodic typology among regional varieties of Italian is deeply diversified as shown in [24]. The linguistic features we found in the considered SPEECON subset are therefore sufficiently descriptive for phenomena commonly found in spontaneous speech. Because of this, we expect that our approach will not suffer major drawbacks when applied to natural speech.

3. Manual annotation

In this work, we consider the prominence as the subjective salience of a syllable within a prosodic unit [c.f. 16, 22]. Different levels of prominence can be identified for each language. For example, four levels of prominence were found in British English [25], while Caputo [20] defined for Italian an annotation system based on 4 prominence levels (from 0 to 3), even if the highest level is not well documented in her corpus.

In our case, in order to simplify the match between manual and automatic annotation, a binary opposition [+ prominent] [- prominent] was preferred. Non prominent syllables in our approach coincide with the 0 level of Caputo's system (i.e. unaccentuation), while prominent syllables coincide with levels 1 and 2 (i.e. word stress and sentence stress). The level 3 (emphasized stress) is not pertinent in our corpus.

The manual annotation was carried out by two human experts, according to a list of basic operational criteria. The procedure was totally perceptual, without any reference to the position of lexical stress in the citation form of words produced in isolation.

The rhythmic structures in the corpus are quite variable, as briefly shown in Figure 2, and typical prosodic phenomena of connected speech often occur, e.g. deaccentuation of lexical stressed syllables and accentuation of unstressed ones [26].

a.	* * no	* ve	* * tSen	* to
b.	* no	* ve	* * tSen	* to

Figure 2: *Variability of prominence structure. Parts of longer sequences found in the corpus*

The agreement rate between the two human experts is 91.51%, which is in line with other researches displaying an agreement index approximately between 80% and 90% [e.g. 27, 28]. When comparing these agreement rates, however, it is necessary to be careful because of differences in the chosen material and in the annotation methodologies.

4. The algorithm

In this paper we present an approach for detecting prominent syllables by taking into account mean energy, duration of syllable nuclei and pitch behavior while crossing syllable nucleus. It has been shown [13] that high performances in prominence detection can be obtained by means of a so called "evidence" variable *Ev*, computed with the formula:

$$Ev = \Delta A \, \Delta D \qquad (1)$$

where ΔA and ΔD are the amplitude and the duration of the syllable nucleus. Also, the same study it concluded that pitch was less significant than the other two features for prominence detection. The same remark has been pointed out in [22] where F0 is indicated as the less significant cue to detect prominence. In [19: 392], however, while referring to lexical stress, it was stated that "when certain conditions are met, the combined effect of intensity and F0 may in fact exceed the weight of duration". In this work we assumed this observation to be valid even in a prosodic context wider than the word. In the proposed system, we decided to consider that these conditions are met when the syllable nucleus is crossed by a rising pitch. This particular situation was described in a number of works on tonal alignment, for example [29], and revealed itself to be useful for the automatic detection of prominent syllables.

By taking into account the pitch behavior, equation (1) can be rewritten as:

$$Ev = m \, \Delta A \, \Delta D \qquad (2)$$

where ΔA and ΔD are the normalized mean amplitude and the normalized duration of the syllable nucleus, while *m* is a parameter describing the contribution of the pitch for the prominence detection process and it is computed according to the following rule:

$$m = \begin{cases} \dfrac{p_{max} - p_{min}}{max(p_{max} - p_{min})} & if \; t_{max} > t_{min} \\ \dfrac{1}{max(p_{max} - p_{min})} & otherwise \end{cases} \qquad (3)$$

where p_{max} and p_{min} are the maximum and the minimum pitch values inside the syllabic nucleus and t_{max} and t_{min} are the time instants at which p_{max} and p_{min} occur.

As defined in equation (3), the parameter *m* has a significant contributions to the definition of *Ev* only if the pitch variation inside the nucleus has a rising pattern. The more steep the rise through the nucleus is, the more important the weight of the *m* parameter will be. Thus, syllables having their nucleus crossed by a rising pitch will generally be preferred by the algorithm when computing their prominence value.

For the implementation of our system, we used the speech analysis software PRAAT [30]. Scripts for automatically extracting syllable nuclei amplitudes and durations and for pitch behavior evaluation were used in order to implement the explained strategy. The pseudocode description of our algorithm is presented in the following paragraph.

1. Extract pitch curve and smooth it
2. Emphasize the signal to make energy peaks point out
3. Extract energy profile and smooth it
4. For each manually marked syllable
   - calculate nucleus duration as the 5 db bandwidth of the energy profile
   - calculate mean energy inside the nucleus
   - evaluate pitch behavior over the nucleus
   - calculate E_v value as described by (2)
5. Mark syllables containing local E_v maxima as prominent

In Figure 3, we show the energy profile and the pitch curve of a speech signal along with its syllable level segmentation, manual prominence annotation and automatic prominence annotation. The case of the syllables [mi – la] is particularly interesting: using equation (1), [la] was seen as prominent while both our experts agreed on marking [mi] as prominent. As it is clear from the figure, the characteristic [mi] exhibits among other syllables, even other than [la], is the rising pitch movement crossing the nucleus.

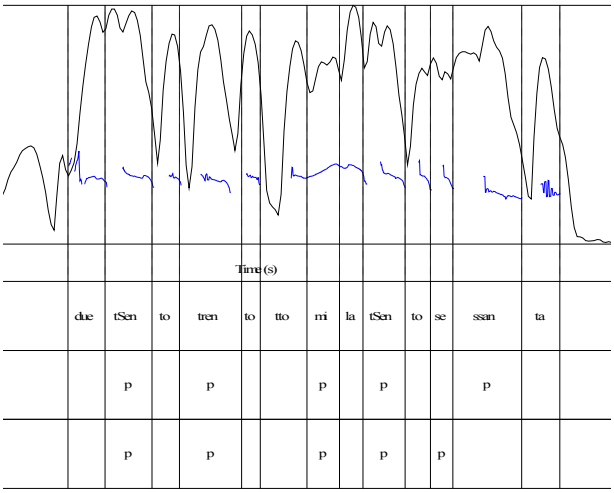

Figure 3. *Intensity profile of a speech signal along with its pitch curve and segmentation data. On the second text tier manual prominence annotations are found. On the third tier automatic ones are found.*

5. Results

We tested our algorithm on the chosen dataset and we compared the results obtained against the annotations provided by human experts. The results are shown in Table 1. Test 1 represents the results against the first human expert, while Test 2 represents the results against the second expert.

Table 1. *Prominence detection results against the two human experts annotations*

	Accuracy	Precision	Recall	F-Measure
Test 1	80,04%	79,59%	70,78%	74,93%
Test 2	78,54%	72,52%	70,88%	71,69%
Average	79,29%	76,05%	70,83%	73,31%

We also compared our results with the ones presented in [14]. In this work, a non-supervised algorithm was proposed to detect syllabic prominences in Italian language. Compared to the corpus used in [14], the SPEECON subset we used here is considerably larger. The results of this comparison are shown in Table 2.

Table 2. *Comparison between proposed system and results presented in the reference work*

	Accuracy	Precision	Recall	F-Measure
Reference system	80,32%	58,10%	70%	63,49%
Proposed system	79,29%	76,05%	70,83%	73,31%

While the values of accuracy and recall are very similar, the precision of our algorithm is significantly higher, indicating that our results are much purer than the ones obtained by the reference algorithm. Even though we understand that the comparison between the two systems is not immediate, because of the difference in dataset types used (the study in [14] used natural connected speech), we considered the obtained results to be indicative of the usefulness of our approach.

We wanted to check how much of the final result depended on the introduction of the parameter concerning the pitch behaviour. We therefore ran our algorithm, deactivating pitch behavior detection, thus employing (1). In Table 3 we show the obtained results.

Table 3. *Performance obtained by deactivating pitch behavior detection*

	Accuracy	Precision	Recall	F-Measure
Test 1	73,41%	71,39%	61,97%	66,35%
Test 2	73,57%	66,21%	63,42%	64,79%
Average	73,49%	68,8%	62,69%	65,57%

After seeing the results in the previous table, it is clear that pitch behavior detection has played a crucial role in determining the results presented in Table. Performance drop was caused by a reduced capability to detect prominent syllables as well as an increased trend of marking as prominent syllables that were not prominent.

6. Discussion

Automated systems for prominence annotation are important not only for the support they can offer to more complex technological applications but also, as standalone software, for basic phonetic research. Phoneticians interested in studying acoustic correlates of prominence could take great advantage of the use of an automatic annotation system like the one we propose here, manual work being reduced to only correcting the errors the algorithm makes. This would certainly be a much easier task than doing the whole annotation from scratch. This kind of procedure, other than being much faster than the manual one, would make it possible to avoid many inconsistencies to which the perceptual annotation is subject to [27, 28].

Automatic prominence detection plays also an important role in speech recognition systems development, especially for those based on automatic syllable segmentation performed on speech chains before recognition. Prominent versus non-prominent opposition is known to cause a high number of systematic phonetic variations, not just on the suprasegmental layer but also on the segmental layer. Studies highlighted how prominent vowels show a tendency to be less centralized and coarticulation resistant [e.g. 4, 7] than non prominent vowels. Furthermore, prominent syllables show a greater displacement of articulators and are more subject to lengthening phenomena [4, 6, 7, 17]. Research on automatic speech recognition should focus on exploiting these systematic variations in order to improve the systems' performance.

A possible approach in this direction could be the use of prominent syllables as anchor points to perform speech recognition by considering them to be probabilistically better recognized than non prominent ones. Supervised approaches, in particular, could benefit of this opposition by generating two different models for the same syllable class describing

both the prominent and the non prominent manifestation of the instances it should contain. This kind of training should lead to a better recognition capability both for the prominent and for the non prominent syllables, thus improving the overall performance of the syllable recognition task and, consequentially, of the whole speech recognition system.

This work is part of a more articulated and complex research involving syllable based speech recognition systems and we are going to test in the near future how prominence can aid in performing these tasks.

7. Conclusions

We proposed an automatic prominence detection algorithm which takes into account classic features, syllable nucleus energy and duration, while also considering pitch behavior, in accordance to results shown in tonal alignment studies.

From the results we obtained it can be observed that this feature helps solving a number of cases that were misclassified by the approach employing energy and duration only. We agree with the idea of F0 being generally less important than energy and duration when performing prominence detection but we also observe that localized pitch behavior analysis is significant to detect prominences. In particular, we observed that when pitch signals a prominence, its weight tends to be more important than the one coming from the nucleus energy and the nucleus duration.

This finding is consistent with many prosodic studies where localized pitch abnormal behaviors are considered rather than raw F0 measurements.

8. References

[1] Wightman, C. W., Shattuk-Hufnagel, S., Ostendorf, M. & Price, P. J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91. 1707-1717.

[2] Keating, P., Cho, T., Fougeron, C. & Hsu, C. 2003. Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, R. Temple (eds.), *Phonetic interpretation. Papers in Laboratory Phonology 6.* Cambridge: Cambridge University Press. 143-161.

[3] Cho, T., McQueen, J. M. & Cox, E. A. 2007. Prosodically driven phonetic detail in speech processing: the case of domain-initial strengthening in English. *Journal of Phonetics* 35. 210-243.

[4] De Jong, K. 1995. The supraglottal articulation of prominence in English. *Journal of the Acoustical Society of America* 97. 491-504.

[5] Erickson, D. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59. 134-149.

[6] Cho, T. 2006. Manifestation of prosodic structure in articulation: Evidence from lip kinematics in English. In L. M. Goldstein, D. H. Whalen, & C. T. Best (eds.), *Varieties of phonological competence. Papers in Laboratory Phonology 8*. Berlin/NewYork: Mouton de Gruyter. 519 - 548.

[7] Avesani, C., Vayra, M., & Zmarich, C. 2007. On the articulatory bases of prominence in Italian. In *Proceedings of the 16th International Congress of Phonetic Science*. Saarbrücken. 981-984.

[8] Cho, T. & Keating, P. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics* 37. 466-485.

[9] Portele, Th. & Heuft, B. 1997. Towards a prominence-based synthesis system. *Speech Communication* 21 (1-2). 61-72.

[10] Batliner, A., Möbius, B., Möler, G., Schweitzer, A. & Nöth, E. 2001. Prosodic models, automatic speech understanding and speech synthesis: toward the common ground. In *Proceedings of Eurospeech 2001.* Aalborg. 2285-2288.

[11] Shriberg, E. & Stolcke, A. 2001. Prosody modeling for automatic speech recognition and understanding. In *Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding.* Red Bank, NJ. 13-16.

[12] Vereecken, H., Martens, J., Grover, C., Fackrell J. & Van Coile, B. 1998. Automatic prosodic labeling of 6 languages. In *Proceedings of the 5th International Conference on Spoken Language Processing.* Sydney. 1399-1402.

[13] Silipo, R. & Greenberg, S. 1999. Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco. 2351-2354.

[14] Tamburini, F. 2005. Identificazione automatica della prominenza frasale nel linguaggio parlato. In *Atti del 1° Convegno Nazionale AISV 2004* [Associazione Italiana di Scienze della Voce]. Padova. 725-754.

[15] Tamburini, F. 2006. Reliable prominence identification in English spontaneous speech. In *Proceedings of Speech Prosody 2006.* Dresden. PS1-9-19.

[16] Terken, J. 1991. Fundamental frequency and perceived prominence. *Journal of the Acoustical Society of America 90 (4)*. 1768-1776.

[17] Beckman, M. E. & Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In P. A. Keating (ed.), *Phonological structure and phonetic form. Papers in Laboratory Phonology 3.* Cambridge: Cambridge University Press. 7-33.

[18] Eriksson, A., Thunberg, G. C. & Traunmüller, H. 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Proceedings of the 7th European Conference on Speech Communication and Technology*. Aalborg. Vol. 1. 399-402.

[19] Bertinetto, P. M. 1980. The perception of stress by Italian speakers. *Journal of Phonetics 8*. 385-395.

[20] Caputo, M. R. 1993. Gradi accentuali nell'italiano parlato spontaneo. In A. Perretti & F. Ferrero (eds.), *Atti del XXI Convegno Nazionale AIA* [Associazione Italiana di Acustica]. Padova. 81-86.

[21] Farnetani, E. & Zmarich, C. 1997. Prominence patterns in Italian: An analysis of F0 and duration. In *Proceedings of the ESCA workshop on intonation*. Athens. 115-118.

[22] D'Imperio, M. 2000. Acoustic-perceptual correlates of sentence prominence in Italian. *Working Papers in Linguistics 52.* Ohio State University. 59-77.

[23] Siemund, R., Höge, H., Kunzmann, S. & Marasek, K., 2000. SPEECON-Speech Data for Consumer Devices. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*. Athens. Vol. 2. 883-886.

[24] Canepari, L. 1980. *Italiano standard e pronunce regionali*. Padova: CLEUP.

[25] Jensen, C. 2003. Perception of prominence in Standard British English. In *Proceedings of the 15th International Congress of Phonetic Sciences.* Barcellona. 1815-1819.

[26] Nespor, M. 1993. *Fonologia*. Bologna: Il Mulino.

[27] Buhmann, J., Caspers, J., Heuven, van V. J., Hoekstra, H., Martens, J. P. & Swerts, M. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).* Las Palmas. 779-785.

[28] Eriksson, A., Grabe, E. & Traunmüller, H. 2002. Perception of syllable prominence by listeners with and without competence in the tested language. In *Proceedings of Speech Prosody 2002.* Aix-en-Provence.

[29] D'Imperio, M. & House, D., Perception of questions and statements in Neapolitan Italian. In *Proceedings of EUROSPEECH 1997*. Rhodes. 251-254.

[30] Boersma, P. & Weenink, D. 2009. Praat: doing phonetics by computer (Version 5.1.20) [Computer program]. Retrieved October 31, 2009, from http://www.praat.org/