APA: towards an Automatic Tool for Prosodic Analysis

F. Cutugno, L. D'Anna*, M. Petrillo*§, E. Zovato°*

***CIRASS - Università di Napoli "Federico II" Italy**
§Dipartimento di Informatica e Sistemistica - Università di Napoli "Federico II" Italy
°Loquendo S.p.A. Torino Italy

{ cutugno,danna,petrillo} @cirass.unina.it, enrico.zovato@loquendo.com

Abstract

In this paper a tool for the speech signal prosodic analysis is described. The system APA (*Automatic Prosodic Analysis*) is based on a tool for speech segmentation into syllabic units and on their description in terms of pitch, energy and duration. A particular linear stylization of the fundamental frequency function is proposed, which helps in describing efficiently intonation movements at phrase level. Finally the energy information, together with the f0 information, are used to find intonation boundary markers in order to segment speech into tone units. A general description of the tool is provided and recent results are reported. Future works in this field are also announced.

1. Introduction

Prosodic analysis plays a fundamental role in the understanding of speech structures: many models have been proposed to describe macroscopic features that characterize the speech in different contexts. In particular, the application of prosodic rules in text to speech systems strongly increases the pleasantness and naturalness of the speech production. On the other hand prosodic features could improve automatic speech recognition if added to the traditional acoustic features[1]. Furthermore, in order to synthetically reproduce prosodic features, a sort of phonetic description of the same features is necessary. The system here proposed is based on procedures that extract information like pitch, energy and syllable duration directly from the speech signal.

As far as the syllable segmentation is concerned [2], the proposed algorithm works in the time domain and calculates the boundaries analyzing local peaks in the energy temporal patterns, selecting them and considering the neighborhood to detect the valleys that constitute syllables edges.

The syllable segmentation is then used for further analysis: f0 movements and energy dynamic are considered in order to detect macro prosodic boundaries (i.e. tone units).

In the next paragraph, the algorithm to segment speech into syllables is described in detail. Paragraph 3 refers to the stylization of f0 using linear regression and in paragraph 4 the tone units segmentation and other kinds of analysis are considered. Results and discussion are reported in the last section.

2. Segmentation into syllabic units

As the basic unit for this prosodic analysis we have considered the syllable. The reason for this choice lays on the fact that syllables are the most natural speech units [3]; moreover co-articulation phenomena are confined inside a syllable and this discourages the segmentation at phoneme level that can be both more complicated and affected by many errors.

The algorithm used to segment speech signals into syllabic units in based on the energy calculation. A previous work [4] based the algorithm for syllable segmentation of the loudness patterns. Energy frames are calculated every 11.6 ms without overlap between contiguous windows.

Among the relative maxima present in the energy contour, we will select those related to syllabic nuclei. The energy frame length is small enough not to loose any of these maxima, but, of course, many of them will be discarded. Our intention here is to loose none of them at this phase, a finest selection will be executed subsequently. In order to select only the most evident of them, during the execution of this step, we use a segmentation window (the size of this window as well as all other parameters used in APA, are set up automatically in a way that will be explained in §6). As it can be seen in Fig.1, this window has been centered on each of the peaks in the energy pattern previously found. If these peaks result to be absolute maxima in the window interval, they are marked as syllabic nuclei and the corresponding minima as syllabic boundaries (a similar approach can be found in [5]).

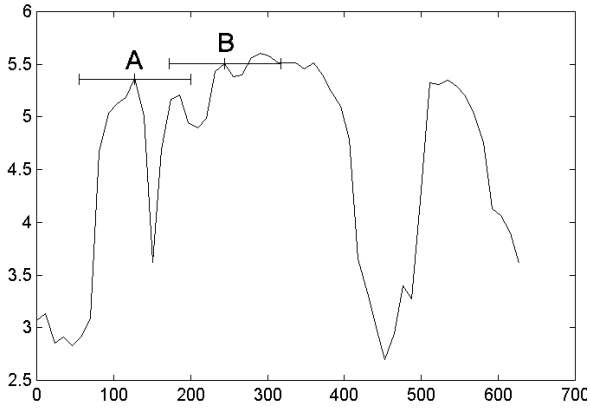

Figure 1: Segmentation window. The maximum in A corresponds, at this stage, to a syllable nucleus because it is an absolute maximum in the segmentation window. The maximum in B does not correspond to a syllable nucleus.

After this phase still some insertion or deletion errors may occur in syllable determination and, consequently, some refinement modules have been introduced. The first of them (*split*) tries to individuate syllabic boundaries that were erroneously discarded in the previous analysis. For each previously found syllable all the minima inside the marked

interval are taken into account in order to find further significant intensity variations around these points. This is achieved looking for the peaks that fall inside a fixed length window centered in the minimum. If the ratio between the lowest peak and the minimum exceeds a particular threshold a new syllable marker is inserted.

One of the drawbacks deriving from the initial segmentation into syllables is that some produced segments contain only fricative parts of speech. The following stage (*fricative grouping*) is necessary to manage this error. In practice they have to be assigned to one of the adjacent syllables. In order to assign these segments, a low pass digital filter has been used. The cut-off frequency has been set to 1100 Hz and the residual energy has been calculated starting from the filtered signal. In this way, segments containing only fricatives present a significant difference between full band energy and residual energy. In these cases the movement of the residual energy has been considered: if it shows a decreasing trend, then the fricative segment is assigned to the left syllable otherwise it is assigned to the following syllable.

The third stage (*stressed vowels grouping*) is necessary to detect insertion errors due to long stressed vowels. In these cases, in fact, some slight energy falls may occur, particularly when the vowel lasts more than 300 ms causing syllable insertion errors. In general, however, these local minima are limited.

The minimum marker is deleted if one of the following conditions, regarding the inverse of the ratio between the valley and its lowest adjacent peak, is true:

- the ratio is less than **R.**
- the ratio is less than **R_med** and the duration of the shortest obtained segment is less than **l_high** ms.
- the ratio is less than **R_med** and the duration of the shortest obtained.

(where R, R_med, l_high are further parameters, see §6)

The last stage is used to move the markers positions when they fall in the middle of a phone. This phenomenon is frequent when a fricative consonant is present at the edge of a syllable. Following the phonologic rules, fricatives should lay at the beginning of the syllable except when they are followed by an occlusive consonant, but in these cases the syllable boundaries are easily detected due to the presence of very low energy intervals. The wrong syllable markers are shifted left bound until the ratio between residual energy and full energy reaches values below a calculated threshold.

3. F0 Stylization

In the course of the automatic prosodic analysis we were also interested in the principal intonation movements at phrase level. For this reason an algorithm for f0 curve stylization is necessary in order:

1. To smooth the unavoidable errors due to the f0 extraction algorithm,
2. To represent in a simple but realistic way the pitch movements and trends, excluding micro-oscillations present in the pitch pattern.

Initially, the f0 function is calculated analyzing speech signals with fixed length windows (25 ms) shifted by 10 ms using an autocorrelation function. According to 't Hart [6], a first order approximation is sufficient to perceptually describe relevant pitch movements. For this reason, a linear approximation has been applied to render the pitch function.

First of all, for each pitch contour, a set of target points has been selected. These points are local minima and maxima. A first group of three points belonging to this set, is used to calculate the first interpolating linear tract. The interpolation is based on the minimum squares technique that is simple and efficient. The following target points are assigned to the current line if their distance is less than 2σ. If this condition is not satisfied, a break point is found and a new interpolating line has to be calculated. The new line starts from the last point of the preceding line and passes through the point whose deviation was too large to belong to the first line. In this way the next target point is considered in order to decide if it is aligned to the new line or if, on the contrary, a new break point is found (see fig.2).

3.1. Energy analysis, speech pauses.

In order to skip all the non-speech parts, i.e. pauses made by the speaker, a further energy analysis has been carried out. Also in this case, windows 25 ms long shifted by 10 ms have been used. Silence frames have been detected by means of an adaptive energy threshold and this parameter has been calculated every 4 seconds in order to consider the speech energy dynamic variations. Signal intervals longer than *max_d* and having energy below the energy thresholds have been considered as speech pauses. Once again all the variables cited in this paragraph are set according to what explained in §6.

4. Segmentation into Tone Units

It is known that when we speak we group words into chunks in order to subdivide the whole information into smaller meaningful portions. The speaker performs these "subdivisions" in different ways. The most natural one is the introduction of a speech pause, often necessary to breath, but sometimes other more complex artifacts occur. Intonation rapid change is one of these.

In literature [7] [8], 4 tone unit boundary markers have been defined:

- Speech pauses
- F0 resets
- Energy resets
- Prepausal syllable lengthening

The second one in particular is intended as the gap of pitch values that occur when, in coincidence of a tone unit boundary, the left part of the pitch contour has a local maximum or minimum, while the right part starts with values in the highest part of the speaker's pitch values range. As far as the energy reset concerns, it just derives from a natural fall that occur when the speaker tends to produce words decreasing the speech intensity (this is due to physiological reasons) and raising again the intensity after a sort of re-starting (breath groups).

At this point we have to decide which syllable boundaries are the edge of a tone unit. The two features that help us in finding these boundaries are the pitch curve and the syllabic normalized energy curve (i.e. the average energy calculated

for each syllable interval). Syllables with lower intensity values are strong tone unit boundaries candidates.

The algorithm applied to the pitch contour initially detects the unvoiced parts of speech and selects only those intervals longer than a fixed value. This parameter has been set experimentally in order not to consider smaller intervals due to occlusive phones. For each unvoiced interval, the f0 gap between the left and right voiced values is calculated in semitones. If the gap of the fundamental frequency is significant (i.e. higher than a fixed threshold) the syllable energy is also taken into account. If this parameter also presents a significant minimum at phrase level and if it is not in the proximity of a speech pause, a tone unit boundary marker is inserted. Figure 2 shows an example of stylization with integration of the silent pauses and of the energy–f0 resets and consequently the subdivision of the signal into tone units.

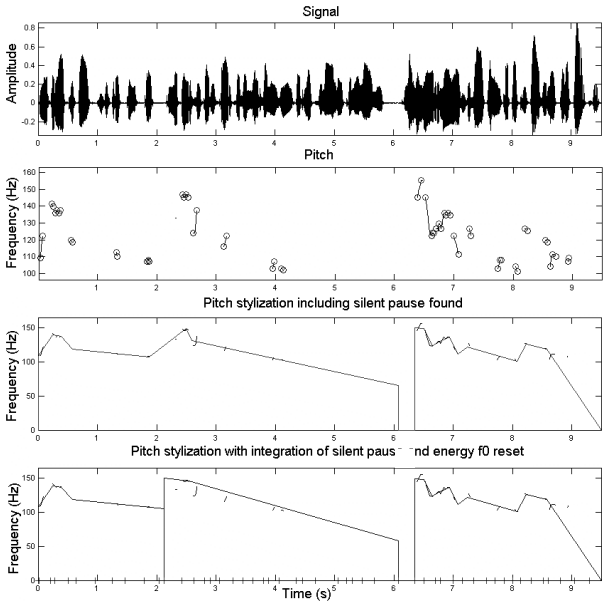

Figure 2: Example of automatic prosodic analysis: (a) signal, (b) pitch target points, (c) pitch stylization including silent pauses, (d) TU boundaries. Syllable markers are represented by crosses on the x-axis.

5. Stress discrimination

Syllables segments have also been described in terms of stress. Turning back to the problem of the prepausal lengthening determination, we should describe the rhythmical structure over syllables segments. To achieve this result we used a simple function *F* defined as the product between the syllable duration and the energy [9]. The function *F* was applied to all the individuated syllables and when a maximum was found then the relative syllable was marked as stressed. This algorithm simply decides if a given syllable is stressed or not but does not performs a discrimination among different types of stress (primary, secondary and pitch stress). In this way the deaccentation present over certain syllables is individuated indirectly when the algorithm marks this syllable as unstressed.

6. Parameters setting

The algorithms described above are based on a set of parameters that have to be tuned in an accurate manner in order to work properly. The algorithmic nature of the system allows us to know the physical meaning of each parameter, so that it is possible to set an interval of meaningful values for each parameter. For every analysis algorithm included in APA, a procedure was developed to set a suitable combination of parameters values. In this paper we will present only the parameter optimisation of the syllable detection via these *setting procedures*, but the same approach can be used for other modules.

Two manually labelled corpora, training and a test data, are required to set parameters. The former is used, as in other paradigms, to effectively choose the best parameters, the latter to verify if the procedure works correctly if applied to signals outside the training corpus. An evaluation function is also required to state the accordance between automatic analysis and corpora labels obtained manually. The smaller is its value on a parameters set, the better is the accordance with the manual analysis. In practice it is necessary to find the set of parameters values that minimizes the evaluation function value.

Three strategies were used. The first one assigns to each parameter a set of possible values, equally spaced, and tries all possible combinations of parameters values. The strategy is suitable only if the number of parameters is very small (e.g. if a 20 parameters procedure is to be tuned, and if only five different values is proposed for each parameter, we get 5^{20} $\sim 10^{14}$ combination to be tried, 3 million years are required even if only one second, for each combination, is needed to test the tuning procedure on the entire training corpus).

However, due to the physical meaning of each parameter, it is often possible to restrict the interval of possible values of each parameter, so that a random procedure can get a suitable result in a smaller computing time.

Better results, in terms of computing time, seem to be obtained via a semi-random procedure: starting from a default value for each parameter, values are then slightly changed at each step. If the evaluation function decreases, then the new set of parameters will be used as default in the next steps.

7. Results and Discussion

The syllable segmentation was trained and tested using two subsets of AVIP corpus (spontaneous speech)[10]. The former was composed of 667 syllables and the latter 645 syllables. In Table 1 its performances are shown.

The system was tested on two further corpora, TV news and Loquendo internal corpus. Performances improve due to the formal speech style of these corpora (see Table 1 and 2).

An estimate on boundary agreement was made. Fig. 3 shows the distance distribution between each manual and automatic syllable marker over the Loquendo corpus. This distribution can be used as a measure of the agreement (the mean is, obviously consistent with zero), its standard deviation is 36 ms. The worst values are obtained on the left marker in stops where a minimum of energy may be found all over the silent regions, or near pauses (by the way, the greatest difference found, occurring only once on 1154 syllables is about 200 ms.). Values different from zero, (but not greater than previous ones) occur in the proximity of nasal phones.

	Training	%	Test	%	TV News	%
Syllables	667		645		1351	
Missed or inserted	66	9.9%	70	10.9%	105	7.8%
Wrong markers	26	3.9%	18	2.8%	31	2.3%

Table 1: performance of syllable segmenter on training, test and TV news corpora.

Syllables	1154	%
Insertion	51	4,40%
Deletion	37	3,20%
Markers	35	3,00%

Table 2: performance of syllable segmenter on Loquendo corpus.

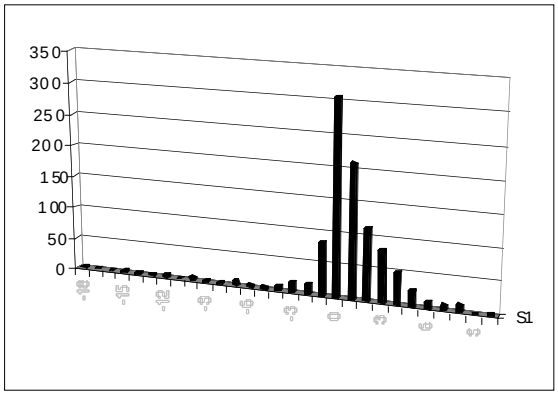

Figure 3: Distribution of differences between manually and automatically detected margins. The unit on the x axis corresponds to the length of the energy frame (11.6ms).

Regarding the stress discrimination, the algorithm performance was evaluated by processing 50 tone units of the AVIP corpus (equal to 365 syllables) manually labeled by expert phoneticians.

Type	Marked manually	Marked automatically	Agreement (%)
Stressed	44	28	64
Unstressed	297	222	75
Deaccented	24	14	58

Table 3: Performances of algorithm for stress discrimination.

In table 3 results are shown. The algorithm achieves the best performance when discriminates unstressed syllables (75%), then the stressed syllables follows with 64% and the worst performance occurs in case of deaccented syllables (58%).

8. Conclusions

A tool to analyze speech prosodic features has been described. The results are encouraging but they could be probably improved if we will use more information regarding f0. Actually we are working to a new function that takes into account the syllabic f0 mean and range variations.

9. Acknowledgments

Authors strictly appear in alphabetic order. The present work started within a national research project (API, *Archivio di Italiano Parlato* – Archive of Spoken Italian – MURST/MIUR- cofin/99 coordinated by Prof. Federico Albano Leoni).

Starting from June 2001 Loquendo s.p.a. became a research partner in the APA project.

All software products will be released under GNU General Public License [11].

10. References

[1] Wu, S.L.; Shire, M.; Greenberg, S.; Morgan, N., 1997. Integrating syllable boundary information into speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing.* Munich, 987-990.

[2] Petrillo, M., 2000. Sillabificazione dei segnali vocali: un approccio procedurale. *Atti Del XXVIII Convegno Nazionale dell'Associazione Italiana di Acustica*, Trani, 303-306.

[3] Greenberg, S., 1998. Speaking in shorthand- a syllable-centric perspective for understanding pronunciation variations. *Proceedings of ESCA Workshop on Modeling Pronunciation variation for Automatic Speech recognition*, Kekrade, 47-56.

[4] Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units, *Journal of Acoustical Society of America*, 58 (4), 880-883.

[5] Pfitzinger, H.R.; Burger, S.; Heid, S., 1996. Syllable detection in read and spontaneous speech. *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)* Philadephia, 1261-1264.

[6] 'T Hart, J., 1990. F_0 stylization in speech: straight lines versus parabolas. *Journal of the Acoustical Society of America*, 90 (6), 3368-3370.

[7] Caputo, M. R., 1992. Aspetti prosodici del processo di segmentazione nel parlato spontaneo. *Atti del XX Convegno Nazionale dell'AIA*, Roma 361-366.

[8] Vaissière, J., 1983. Language-Independent Prosodic Features. In *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd (eds), Berlin: Springer-Verlag, 53-66.

[9] Silipo, R.; Greenberg, S., 1999. Automatic transcription of prosodic stress for spontaneous English discourse. *Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, 3:2351-2355.

[10] Bertinetto, P.M. (ed.), 2001. *Archivio delle Varietà di Italiano Parlato*, Scuola Normale Superiore, Pisa, 2001.

[11] GNU General Public License – http://www.gnu.org/licenses/gpl.html.