SONORITY BASED SYLLABLE SEGMENTATION

Bogdan Ludusan, Serena Soldo
Department of Physical Sciences, 'Federico II' University, Naples
ludusan@na.infn.it, soldo@na.infn.it

1. ABSTRACT

This paper proposes a new method for detecting syllable boundaries. It is based on the sonority and it uses the so-called 'Sonority Sequencing Principle' for the boundary detection. As acoustic correlate of the phonological concept of sonority we use the regularities present in the spectrogram of the signal. By finding the maxima of the sonority function we will be finding the syllable nuclei, while the syllable boundaries are to be found at the minima of the sonority function. Due to the fact that it uses only the information contained in the speech signal it could be implemented, with small modifications, for almost any language.

2. INTRODUCTION

Automatic speech segmentation is a topic of great interest in nowadays speech related literature due to its multiple use. One of its most important application areas is Automatic Speech Recognition (ASR), in which speech segmentation techniques are applied for obtaining the units used for recognition. In the recent ASR literature, the syllable is a frequent choice for such a unit because it offers a good representation of the variability present in the speech signal while retaining a also good trainability. This is the reason behind our proposal for an algorithm for automatic syllable segmentation.

Although the syllable is intuitively recognized by most of the people, there is not yet a universally agreed definition of the syllable. For example, from an acoustic point of view it was observed that energy temporal patterns play a fundamental role (Jespersen, 1904), syllable nuclei being usually found in correspondence with energy maxima, while syllable boundaries correlating with energy minima. In contrast, in phonology, the most widely used syllable definition is based on the sonority scale.

The sonority is a concept present in the phonological theory from the nineteenth century. The opinions on whether the sonority has or not a phonetic basis are divided some suggesting that it is correlated in some way with audibility (Sievers, 1881), some that it can be defined in terms of the loudness of a sound, which is related to its acoustic energy relative to other sounds having the same length, stress and pitch (Ladefoged, 1993), while others do not even recognize it as a phonological concept (Harris, 2006). Taking on a different stance, Clements (1990) argues that the absence of a physical basis for characterizing sonority in language-independent terms would make it impossible to explain the nearly identical nature of sonority constraints across languages.

Based on the measure of sonority, several relative rankings of the sonority of sounds were developed, among which, I recall the one presented in (Ladefoged, 1993): low vowels > mid vowels > high vowels > liquids > nasals > obstruents. The Sonority Sequencing Principle (SSP) is used as principle for syllabification stating that the sounds inside a syllable increase in sonority from the onset to the nucleus, with a maximum value corresponding to the nucleus and decrease in sonority from the nucleus to the coda.

The sonority was used previously as feature for segmentation in speech processing, but it was either used to detect only syllable nuclei (Kawai & van Santen, 2002) or to detect syllable boundaries, but combined with other features and in conjunction with statistics from previous segmentations (Mayora-Ibarra & Curatelli, 2002).

In (Kawai & van Santen, 2002) multiple linear regression is used in order to obtain, what the authors call, the instantaneous sonority. As predictor variables for the regression they use bandpass-filtered acoustic energy from the central part of each phone. The authors argue that the five frequency bands chosen can efficiently locate boundaries between different phone classes. They report accuracies of over 60% for syllable nuclei detection and over 80% for speech rate recognition for a corpus of read news.

Mayora-Ibarra & Curatelli (2002) obtain their segmentation by using time-domain signal processing followed by a refinement of the results based on a fuzzy-logic approach. As time domain feature they use the zero-crossing rate in the intervals of sonority decrease, which, they state, it is related to the attenuation of the acoustic intensity of speech that occurs between the transition of adjacent syllables. The second step represents a refinement of these results and it is implemented using statistics from previous segmentation tests together with fuzzy logic rules. The accuracies reported on a corpus of isolated Italian digits are of 87% after the first phase and 95% after the refinement of the results.

Recent work (Galves *et al.*, 2002) has proved the usefulness of the sonority in other areas, like rhythmic class discrimination. In their paper, the authors propose a formulation for the sonority function, defined on the interval [0,1]. The proposed function has values close to 1 for sounds displaying regular patterns, characteristic of sonorant portions of the signal and close to 0 for regions characterized by obstruency.

In Cassandro *et al.* (2002) the authors refine the previously proposed function using an exponential. Subsequently, the sonority is defined as a decreasing function of the values of the relative entropies between neighbouring columns of the spectrogram of the speech signal:

$$S(t)=\exp(-\beta\sum_{1}^{3} h(p_t|p_{t-i})) \qquad (1)$$

where h denotes the relative entropy between two probability measures, p_t is the power spectrum renormalized in order to become a probability measure and β is a free parameter assuming positive real values.

Among the methods used in the literature for syllable boundary detection there are many algorithm using only the information extracted from the speech signal, without any linguistics or phonetic knowledge (Petrillo & Cutugno, 2003; Nagarajan *et al.*, 2003). The first approach (Petrillo & Cutugno, 2003) is based on the energy of the signal and it searches the syllable boundaries at the minima of the energy envelope. In Nagarajan *et al.* (2003), the authors obtain the syllable segmentation based on a minimum phase group delay approach. To our knowledge, the results presented in the previous paper are the best segmentation results on an English corpus of conversational speech.

3. METHODS

3.1 Algorithm

Based on the previous formulation of the sonority function (1), we propose an algorithm for the detection of syllable boundaries. The algorithm uses exclusively speech processing techniques (both frequency and time domain), having no knowledge about the phonetic content of the signal, in order to obtain the syllable boundaries from the continuous speech signal. Figure 1 presents the block scheme of the algorithm.

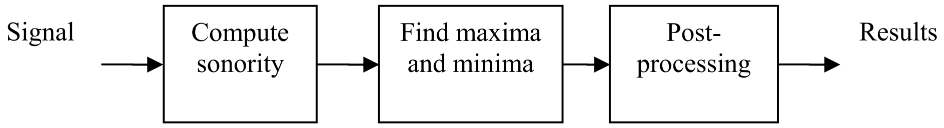

Figure 1: Block scheme of the algorithm

In a first step, for each of the utterances, the sonority function is computed in a similar manner to the one described in Cassandro *et al.* (2002). The major difference consists in the use of a different distance function for the computation of the sonority – the normalized Euclidean distance instead of the relative entropy between the columns of the spectrogram. The following steps are executed in order to obtain the sonority of the signal:

1. computing the spectrogram of the signal for the frequency band below 1000 Hz, using a 25 ms window;
2. normalization of the power spectrum;
3. computing the normalized Euclidean distance of five consecutive columns; the formula of the normalized Euclidean distance between two vectors (here the vectors representing the columns of the spectrogram) is listed in (2); the distance function used has the same properties as the relative entropy – gives low values for vowels, nasals and voiced stops (because of the regularity introduced by voicing) and high values for voiceless stops, fricatives and flaps (Garcia *et al.*, 2002):

   $$d(\vec{x},\vec{y})=\sqrt{\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{\sigma_i^2}} \qquad (2)$$

   where σ_i is the standard deviation of x_i over the sample set;
4. applying relation (1) for the normalized Euclidean distance we will obtain the value of the sonority function; by multiplying in the exponential function with -1, we will obtain in the sonority function high values for vowels, nasals and voiced stops and low values for all the other sounds.

As an example, the sonority profile of the word ‘cinquecentoventunomiladuecentouno’ along with its syllable segmentation is presented in Figure 2.

The frequency band for which the spectrogram of the signal is computed (0-1000 Hz) was established empirically. The bandwidth was determined through testing of various bandwidths between 400-500 Hz (in order to catch at least the F0 of the voiced segments) and 1500-2000 Hz (to avoid a high computation time for the sonority function).

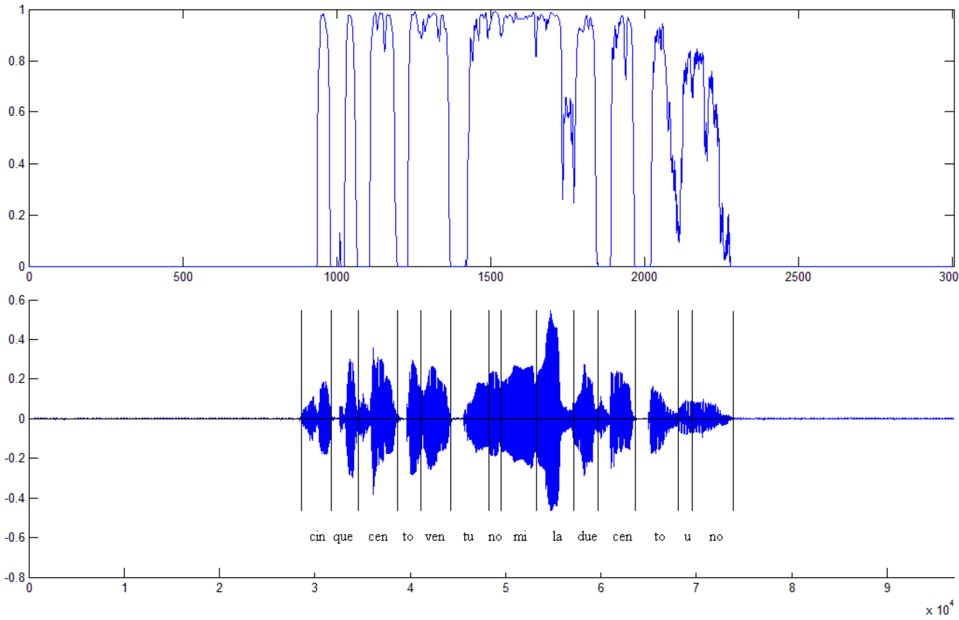

Figure 2: Sonority profile and syllable boundaries for the word ‘cinquecentoventunomiladuecentouno’

As the distance metric used is the most important factor in computing the sonority function, we needed a robust function for it. By using the relative entropy (Garcia *et al.*, 2002) relatively high sonority values for the silence periods and for the fricative segments were obtained (due to the quasi-periodicity of the noise).

The normalized Euclidean distance was chosen because it has the same characteristics of the relative entropy, while eliminating its drawbacks: giving low sonority values for the silence periods and better behaviour for the fricative segments. A problem with this metric was the fact that it returned high differences between the sonority of voiced regions and that of the unvoiced regions (by several orders of magnitude) which posed problems when computing the inverse of the function. This issue was solved by introducing a variable normalizing factor β, that reduces the distances between the values for the voiced and unvoiced regions while still keeping a significant difference between them.

At the beginning of the second step the envelope of the sonority is computed in order to find the sonority maxima. The envelope is computed by low-pass filtering the sonority function, thus obtaining only the long-term variations of the signal. Because the normalized Euclidean distance has a much smoother form than the relative entropy, it needs also a less sophisticated method for computing the envelope. The syllable boundaries will be placed in accordance with the SSP. As it states that the peaks in the sonority function correspond to the syllable nuclei, by finding the maxima of the function, we will be finding the syllable nuclei. This is done first by imposing a minimum threshold on the sonority maxima and then searching for all local maxima in the signal. Having found the syllable nuclei and knowing that the syllable boundaries correspond to the minima in the sonority function, the next step consists in finding the minima between each two consecutive maxima.

The post-processing step tries to correct some of the errors that might appear in the segmentation process. In a first stage, a voice activity detection procedure is used to determine the beginning and the end of the speech region and all syllables boundaries found outside this interval are eliminated. One of the most important errors is the existence of spurious maxima close to the syllable nuclei, due to the semi-vowels, nasals or liquids that are in the vicinity of the vowel. Because the minima between two such maxima has very high values, these types of errors can be corrected by comparing each minima with their neighbouring maxima and eliminating the lower maximum in case of an insertion.

Another type of error might appear due to segments violating the SSP, in which a more sonorous segment is found further away from the nucleus than a less sonorous segment. In this case, our system tends to consider this segment as a unique syllable. But, due to the fact that these segments are quite short (usually under 75 ms), by setting a minimum threshold to the syllable length and assigning these isolated segments to one of the neighbouring syllables. For this value of the threshold the speech rate is not an issue as a speech rate of 14 syllables/second (corresponding to syllables 75 ms long) is difficult, if not impossible to be reached.

The erroneously found syllables corresponding to nasal segments are corrected by taking into consideration one of the most important acoustic characteristics of nasal consonants, i.e. the existence of a high intensity F1 around 300 Hz. In the case of boundaries with relatively high sonority and one of the syllables it confines short enough, a comparison between the F1 of this short segment and its neighbouring segment is done. If its mean F1 value is lower than the one of its neighbouring segment and also lower than 400 Hz the boundary will be deleted.

4. RESULTS

The corpora on which our system was tested are the Italian part of the SPEECON corpus (Siemund *et al.*, 2000) and the Switchboard corpus (Godfrey *et al.*, 1992). The Italian part of the SPEECON corpus contains numbers from 0 to 999,999 pronounced by male speakers. There are a total of 1906 recordings made by approximately 400 speakers. The Switchboard corpus instead is a corpus of English conversational speech recorded over the telephone line. It contains 2500 conversations collected from 500 American English speakers (both males and females).

The evaluation of the segmentation was performed using the algorithm presented in (Petek *et al.*, 1996). The algorithm defines for each of the manually annotated syllable boundaries a search region in which a corresponding automatically found syllable boundary will be searched for. The search interval spans from the middle of the interval between the previous and the current syllable boundary to the middle of the interval between the current and the next syllable. If no automatic boundary is found in the search interval a deletion is considered. If a boundary is found, depending on the distance to the closest manual boundary, it is considered either a correct boundary or a substitution. If more automatic boundaries are found in the search interval, the closest one is considered the correct one and all the others are considered insertions.

In Table 1 we present a summary of the accuracies obtained using several algorithms for automatic syllable segmentation, while in Table 2 we show a comparison of the errors obtained between our system and the approaches presented in (Nagarajan *et al.*, 2003) and (Petrillo & Cutugno, 2003).

In each cell of Table 2 we have the substitutions, insertions and deletions that occurred during the segmentation process. Substitutions were considered if the distance between the found boundary and the manual one exceeds 40 ms, unless stated otherwise.

Corpus→ ↓Approach	Switchboard [%]	SPEECON [%]	Other [%]
Our algorithm	54.11	77.70	-
Mayora-Ibarra & Curatelli	-	-	SPK-IRST (Italian digits) 95[1]
Kawai & van Santen	-	-	Read news 62[2]
Nagarajan *et al.*	74.84	-	-
Petrillo & Cutugno	57.25	82.43	-

Table 1: Accuracies obtained using different segmentation algorithms

Corpus→ ↓Approach	Switchboard sub/ins/del [%]	SPEECON sub/ins/del [%]
Our algorithm	15.33 / 14.97 / 15.58	10.31 / 7.84 / 4.15
Nagarajan *et al.*	12.79 / 5.25 / 7.1	-
Petrillo & Cutugno	13.67 / 8.88 / 20.2	8.74 / 4.29 / 4.55

Table 2: Errors obtained using different segmentation algorithms

[1] The error interval was set at 15 ms

[2] For syllable nuclei

Although our approach gives accuracies values far from the state of the art system presented in Nagarajan *et al.* (2003), we obtain closer values to the systems using less complex speech processing techniques. The accuracies, both on the SPEECON corpus as on the Switchboard corpus, close to the modified system (Petrillo & Cutugno, 2003) as well as a much better accuracy (of the syllable boundaries) with respect to the Kawai & van Santen (2002) approach are an encouraging result.

5. CONCLUSIONS AND FUTURE WORK

A syllable segmentation algorithm based on the sonority of the speech signal was presented. The algorithm, which is based entirely on the signal, without any linguistic or phonetic knowledge, uses the Sonority Sequencing Principle for finding the syllable boundaries, corresponding to the minima in the sonority. The results obtained are encouraging, having obtained better accuracies than previous systems based on the sonority and accuracies similar to those of systems based on the energy of the signal.

In order to increase the segmentation accuracy, several other features could be used together with the sonority function. An example of such features are the acoustic properties of the signal that help us characterizing the manner of articulation of the speech segments included in the utterance. These can be in particular useful in the case of SSP violation – for example the presence of /s/ before /t/ in the onset of the syllable. By finding an isolated strident segment between two syllables and the second syllable beginning with a very low sonority segment (a stop consonant) we could consider it as a onset /s/ and assign it to the second syllable.

Another alternative would be the combination of our system with the one presented in Petrillo & Cutugno (2003). Initial tests on the errors that the two systems do showed that most of the errors are quite complementary and a combination of the two systems will increase the segmentation accuracy.

ACKNOWLEDGEMENTS

This work was supported by the EU FP6 Marie Curie Research Training Network 'Sound to Sense'.

6. REFERENCES

Cassandro, M., Collet, P., Duarte, D., Galves, A. & Garcia, J. (2002), *An universal linear relation among acoustic correlates of rhythm*, Retrieved from http://www.ime.usp.br/~tycho/participants/a_galves/linearity.pdf.

Clements, G. (1990), The role of the sonority cycle in core syllabification, in *Papers in laboratory phonology 1: between the grammar and physics of speech* (J. Kingston & M. Beckman, editors), Cambridge: Cambridge University Press, 283-333.

Galves, A., Garcia, J., Duarte, D. & Galves, C. (2002), Sonority as a basis for rhythmic class discrimination, in *Proceedings of the Speech Prosody 2002*, Aix en Provence, France, 323-326.

Garcia, E., Gut, U.B. & Galves, A. (2002), Vocale – a semi-automatic annotation tool for prosodic research, in *Proceedings of the Speech Prosody 2002*, Aix en Provence, France, 327-330.

Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992), SWITCHBOARD: Telephone speech corpus for research and development, in *Proceedings of IEEE ICASSP 1992*, 517-520.

Harris, J. (2006), The phonology of being understood: further arguments against sonority, *Lingua*, 116, 1483-1494.

Jespersen, O. (1904), *Lehrbuch der Phonetik*, Leipzig: B.G. Teubner.

Kawai, G. & van Santen, J. (2002), Automatic detection of syllabic nuclei using acoustic measures, in *2002 IEEE Workshop on Speech Synthesis*, Santa Monica, California, 39-42.

Ladefoged, P. (1993), *A Course in Phonetics*, 3rd edition (International Edition), Orlando: Harcourt Brace & Company.

Mayora-Ibarra, O. & Curatelli, F. (2002), Time-Domain Segmentation and Labelling of Speech with Fuzzy-Logic Post-Correction Rules, in *Proceedings of the Second Mexican International Conference on Artificial intelligence: Advances in Artificial intelligence*, 1-14.

Nagarajan, T., Murthy, H.A. & Hegde, R. M. (2003), Segmentation of speech into syllable-like units, in *EUROSPEECH-2003*, Geneva, Switzerland, 2893-2896.

Petek, B., Andersen, O. & Dalsgaard, P. (1996), On the robust automatic segmentation of spontaneous speech, in *ICSLP-1996*, Philadelphia, USA, 913-916.

Petrillo, M. & Cutugno, F. (2003), A syllable segmentation algorithm for English and Italian, in *EUROSPEECH-2003*, Geneva, Switzerland, 2913-2916.

Siemund, R., Höge, H., Kunzmann, S. & Marasek, K. (2000), *SPEECON* - Speech Data for Consumer Devices, in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000, vol. 2, 883-886.

Sievers, E. (1881), *Grundzüge der Phonetik,* Leipzig: Breitkopf und Härtel.