INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Combining energy and cross-entropy analysis for nuclear segments detection

Antonio Origlia[1], *Francesco Cutugno*[1]

[1]PRISCA-Lab, Federico II University, Naples, Italy
{antonio.origlia,cutugno}@unina.it

Abstract

Features related to rhythmic patterns are involved in the representation of the intonational content for spoken language analysis. Among others, speech rate is one of the most used measures extracted by systems using prosodic analysis and is typically measured in syllables per second. Automatic approaches designed to estimate this measure in absence of manual annotations usually mark the position of syllable nuclei as a single point in time. Approaches extracting duration features using automatic segmentation in units shorter than words but larger than phones tend to detect syllables. To represent the prosodic contents of an utterance, especially from the rhythmic point of view, automatic positioning of nuclear boundaries may, however, be more informative than syllable boundaries. In this paper we present a method combining the analysis of the energy envelope and of the cross-entropy profile to obtain a segmentation into nuclear and inter-nuclear segments, showing that the proposed method can be used to obtain a reliable estimate of speech rate and that accuracy in nuclear boundary positioning allows the extraction of segmental features useful for automatic prosodic analysis.
Index Terms: Syllable nuclei detection, speech segmentation, speech rate

1. Introduction

Speech segmentation is a very important step for every automatic approach for voice analysis. Segments can be found on different levels: automatic speech recognition needs to segment the input stream into regions of voice activity while different kinds of prosodic analysis may need to access different levels of detail ranging from intonational phrases down to single phones. Vowels, in particular, are important areas of the speech signal as rich spectral data describing *how* a human is speaking can be found there. Vowel detection has been studied in the past with a number of different approaches. In [1], the Reduced Energy Cumulation (REC) function was proposed to characterize the spectrum of vowels by detecting their formantic structure. This method was used in [2, 3] to detect vowel regions inside speech segments detected by applying the Divergence Forward Backward (DFB) presented in [4] and extract features for acted emotion recognition. More recently, a method to segment the speech signal into vowel-like regions was proposed in [5] to correct the position of vowel onset and offset detected by an HMM. The same approach was used in [6] to refine syllable boundaries in Assamese (a dialect of India).

While vowel detection may be used to estimate speech rate, as in [1], approaches designed for this task usually mark syllable nuclei as single points in time typically obtained by peak counting on some base function. In [7], this function was obtained by spectral subband correlation including temporal correlation, prominent spectral subbands, and pitch information. In [8], the smoothed energy profile is extracted and prominent voiced peaks are taken as syllable nuclei. In [9], rhythmicity features derived from a modified version of the short-time energy envelope are considered and, in [10], the Low Frequency Modulated Energy profile is taken as base function for peak counting. Since a complete view of the extension of nuclear and inter-nuclear segments is important for prosodic analysis, estimating nuclei boundaries is relevant to the task. As vowels, being the most sonorant class of vocal sounds, typically form the nucleus of the syllable, it is straightforward to explicitly look for them in the speech signal to perform relevant features extraction and to compute measures like speech rate and rhythmic descriptors. In some cases, however, nasal sounds can constitute the syllable nucleus. Therefore, while approaches that explicitly look for vowels are appropriate for acoustic features extraction, missing nuclei formed by this subset of vocal sounds risks introducing erroneous estimates for speech rate and for segment lengths evaluation, which are often significant in speaking style analysis tasks. Applications for nuclear segments detection range from intelligibility measures in pathological disorders [11] to automatic language identification [12] and continuous emotion tracking [13]. In this paper, we describe an approach combining an energy based algorithm for speech rate evaluation with the analysis of the cross-entropy profile to detect syllable nuclei without explicitly assuming they will be constituted by a vowel. The algorithm is easy to implement and it extends a similarly simple approach to provide a fast, efficient solution to automatic nuclei detection.

2. Material

In English, it is not rare to observe specific kinds of nasal sounds acting as syllable nuclei. For this reason, the training set of the TIMIT corpus is used in this work. The material consists of 4620 phonetically segmented utterances for a total of almost 4 hours of spoken material. The reference segmentation into nuclear and inter-nuclear segments is obtained automatically from the time-aligned phonetic transcription using the following rules (examples from the TIMIT documentation are also reported):

- All vowels are marked as a syllable nuclei;
- the glide *el* (bottle: bcl b aa tcl t EL), the nasal sounds *em* (bottom: b aa tcl t EM), *en* (button: b ah q EN) and *eng* (washington: w aa sh ENG tcl t ax n) are marked as nuclei if they are not adjacent to a vowel;
- the semivowels *w* (way: W ey), *y* (yacht: Y aa tcl t) become part of the nucleus if adjacent to a vowel. If they are found between two vowels, however, they form an internuclear segment on their own;
- adjacent segments marked as nuclei are merged into a single one.

http://dx.doi.org/10.21437/Interspeech.2016-1345

The resulting number of nuclei in the reference segmentation is 54879. Of these, 10% is represented by non-vowel sounds. While limited with respect to vowels, the role of non-vowels sounds acting as syllable nuclei cannot be neglected.

3. Automatic segmentation

In this section, the two approaches to syllable nuclei detection are presented together with the rule-based approach to merge the separate results.

3.1. Energy based nuclei detection

The energy based nuclei detection phase consists of the algorithm presented in [8] and we summarize it here. We will refer to this approach as the energy peaks marking (EPM) phase. The algorithm is composed of five steps as follows:

- the intensity profile is extracted, smoothing is applied over a time window of 64ms;
- all peaks above a certain threshold in intensity to be potential syllables. The threshold is set as the .99 quantile minus 25 dB to avoid potential energy bursts (this is different from the original algorithm, where the median was taken as reference. In our experiments, we used the updated version of the algorithm from 2010);
- only peaks with a preceding dip of at least 2 dB are kept as syllable nuclei candidates
- the pitch contour is extracted and unvoiced peaks are excluded. The remaining peaks are marked as syllable nuclei.

At the end of this procedure, syllable nuclei are marked as points in time rather than intervals.

3.2. Cross-entropy analysis for nuclei boundaries detection

Cross-entropy analysis is used to isolate portions of the speech utterance that are characterized by a rapid transition from areas containing less rich spectra to areas where the presence of peaks makes the spectrum richer and viceversa. The first type of transition marks Nuclear Onset Points (NOP) while the other marks Nuclear Ending Points (NEP). Between NOPs and NEPs, voicing activity must be found. Cross-entropy is often used to measure the difference between a reference probability distribution $p(x)$ and another probability distribution $q(x)$. It is formally defined as

$$H(p,q) = H(p) + D_{KL}(p||q) \tag{1}$$

where $H(p)$ indicates the entropy of the distribution p and $D_{KL}(p||q)$ is the Kullbak-Leibler divergence of q from p. In the discrete case, this is reformulated as

$$H(p,q) = -\sum_{x} p(x)\, log\, q(x) \tag{2}$$

Minimizing this measure is found, in the literature, to be the goal of machine learning algorithms and it is commonly used in the language processing community to evaluate language models.

The first step of our approach consists in filtering the signal with a low-pass filter (cutoff frequency set at 3000Hz). This is to avoid the noise caused by high frequency energy associated with other sounds than the ones we are interested in. At each step, we compute $LTAS_n(x)$, the Long-Term Average Spectrum (LTAS) of the $n-th$ frame, and $LTAS_{n-1}(x)$ using PRAAT's built-in function. Bins are 100Hz wide and, since we are interested in capturing details about the spectral configuration, in this step we consider the narrow-band spectrogram with overlapping analysis windows. In the presented experiments, these are 20ms wide and the time step is 10ms. The difference between consecutive frames is computed by rewriting Equation 2 as

$$H(LTAS_n(x), LTAS_{n-1}(x)) = -\sum_{f=1}^{N_b} LTAS_n^f(x)\, log\, LTAS_{n-1}^f(x) \tag{3}$$

where N_b is the number of bins in the generic $LTAS(x)$. Next, the first derivative of the cross-entropy profile is computed and normalized in the interval $[-1, 1]$. A multiple-pass moving average filter (5 steps in the presented experiments) is applied to the cross-entropy derivative profile to obtain a smooth curve. NOPs candidates are positioned in correspondence of the profile's valleys while NEPs candidates are positioned in correspondence of peaks. Only segments delimited by a NOP and a NEP containing pitch are retained. We will refer to this phase as the Cross-Entropy Derivative (CED) step.

3.3. Merging the approaches

When considered independently, the following problems may affect the two annotations:

1. An EPM peak is detected outside the actual nucleus because of a spurious energy peak preceding the actual one;
2. an EPM peak is not detected because of pitch tracking errors causing misalignment between the energy peak and the voiced region;
3. a CED interval is detected because of pitch tracking errors occasionally marking short unvoiced regions as voiced;
4. a CED interval contains more than one EPM peak.

Problem 1 is caused by a spurious peak detected just outside the actual nucleus. This causes the peak that should be associated to the nucleus to be discarded. In this situation, CED intervals correctly mark the nucleus and peak realignment is sufficient to address the issue.

Problem 2 is caused by EPM peaks being detected only if they fall inside a voiced region. Pitch tracking errors cause this strategy to miss nuclei if the starting point of the voiced region occurs even one time step later than the energy peak. CED intervals are less sensitive to this kind of error as pitch must be detected in a time interval rather than at a specific time instant. On the other hand, CED intervals can be oversensitive when cross-entropy boundaries contain erroneous pitch detection, leading to Problem 3. For this reason, CED intervals not containing an EPM peak are kept only if cross-entropy movements indicating nuclear boundaries are strong enough to suggest that disagreement was caused by Problem 2. If the difference between a segment's NOP and its NEP is weak, however, disagreement is considered to be likely caused by Problem 3 and the CED interval is not kept. In our experiments, movement strength is measured on the normalized scale [-1, 1] and the difference threshold between NOP and NEP is set to 0.5.

Problem 4 is caused by the moving average filter removing a cross-entropy movement because of a rapid transition from

nuclear to inter-nuclear to nuclear segments again. If the corresponding energy movement causes two EPM peaks to be detected, it is necessary to recover the missing boundaries. First of all, although the smoothed profile may not show a valley and its corresponding peak, an inflexion point may still indicate where the boundaries should have been. In this case, boundaries can be set by considering the second derivative of the cross-entropy profile. This is done by taking the valley and the peak enclosing the energy minimum occurring between the two EPM peaks as reference points for the missing NEP and NOP, respectively. If it is not possible to detect such movements on the second derivative, an inter-nuclear segment of fixed length centered on the energy minimum between the two EPM peaks is inserted.

The final rules set to combine the EPM and CED approaches is summarized as follows:

1. CED intervals containing an EPM peak are kept as nuclei;
2. EPM peaks that are not found inside a CED interval are moved in correspondence of the energy maximum of the nearest CED interval if this does not contain any other EPM peak.
3. EPM peaks that cannot be associated with CED intervals are discarded
4. CED intervals that do not contain an EPM peak are discarded unless the cross-entropy derivative difference value between the nucleus onset and the nucleus offset is higher than 0.5 on the normalized scale;
5. CED intervals containing more than one EPM peak are splitted by taking the second derivative of the cross-entropy profile and putting NEP and NOP markers in correspondence of the local maximum and minimum enclosing the EPM peak. If these cannot be found inside the CED interval, the algorithm assumes that the change was too fast to be captured by the CED step and introduces a 30ms long internuclear segment centered on the energy minimum between the two EPM peaks.

An example of the segmentation obtained with this procedure is shown in Figure 1.

4. Results

First of all, we check the impact the combined approach has on speech rate estimate. Speech rate is obtained by dividing the number of syllables in the reference and in the automatic annotation by the file length. The measures of evaluation are the Pearson's correlation coefficient (COR) and the Root Mean Square Error (RMSE). The first measure evaluates how strong is the linear relationship between the two variables while the second one provides an estimate of the error committed by the automatic approaches. This is summarized in Table 1.

	EPM	**EPM + CED**
COR	0,70	0,71
RMSE	0,64	0,55

Table 1: Broad evaluation for speech rate estimate.

Considering that, in [1], Pearson's correlation is reported to be $0,81$ for human annotators, the obtained result can be considered satisfactory. Also, the drop in RMSE provides an estimate of the magnitude in error reduction with the combined approach. This is mainly caused by an increased capability of detecting syllable nuclei. The difference between the two approaches is evaluated with a paired t-test on the reported speech rates and was found to be significant ($p < 0.001$).

In order to evaluate the accuracy of boundaries positioning, we use the SCLITE tool, included in the NIST Scoring Toolkit (SCTK), and enable the option to compute time-mediated alignment. In this mode, the tool applies the standard Dynamic Programming alignment algorithm used for scoring automatic transcriptions versus reference ones but it substitutes the fixed distance weights with measures based on segments' starting and ending times. In detail, the weights are computed as follows:

$$\begin{aligned} D(correct) &= |T1(ref) - T1(hyp)| + \\ &\quad |T2(ref) - T2(hyp)| \\ D(insertion) &= T2(hyp) - T1(hyp) \\ D(deletion) &= T2(ref) - T1(ref) \\ D(substitution) &= |T1(ref) - T1(hyp)| + \\ &\quad |T2(ref) - T2(hyp)| + 0.001 \end{aligned} \tag{4}$$

Where hyp indicates hypothesized segments and ref indicates reference segments. $T1$ and $T2$ indicate, respectively, the starting and ending times of the segments. To provide a basic segmentation for the baseline, we use a fast, empirical approach and delimit the nucleus boundaries by considering the -2dB band of the each EPM peak. We chose this threshold because the default value for the minimum energy dip in the baseline algorithm is set to -2dB, too. Results obtained with this procedure are reported in Table 2.

	EPM	**EPM + CED**
Correct	0,83	0,87
Substitutions	0,03	0,03
Deletions	0,13	0,11
Insertions	0,05	0,05
Accuracy	0,79	0,82

Table 2: Results obtained with SCLITE time-mediated alignment.

As in the preceding test, the combined approach shows an advantage with respect to the baseline that is mainly caused by a higher capability to detect nuclei, indicated by the reduction in the percentage of deletions. The number of substitutions and insertions is unchanged. Analysis with SCLITE provides data on nuclear boundaries positioning and, being a standard tool, favors future comparability. To check the significance of the difference between the baseline and the combined approach, we used the SCSTATS scoring tool, also included in SCTK, and performed a Matched Pairs Sentence Segment Word Error (MAPSSWE) test. The test indicated the difference to be significant ($p < 0.001$).

5. Discussion

Spectral changes in the speech signal are often evaluated with divergence operators. In [4], the DFB algorithm makes use of the cross-entropy operator, as we do in this work, but it is designed to isolate stationary regions from transitory ones. The comparison is made between the autoregressive models obtained from a long-term analysis window of fixed length and a second window growing inside the long-term analysis window.

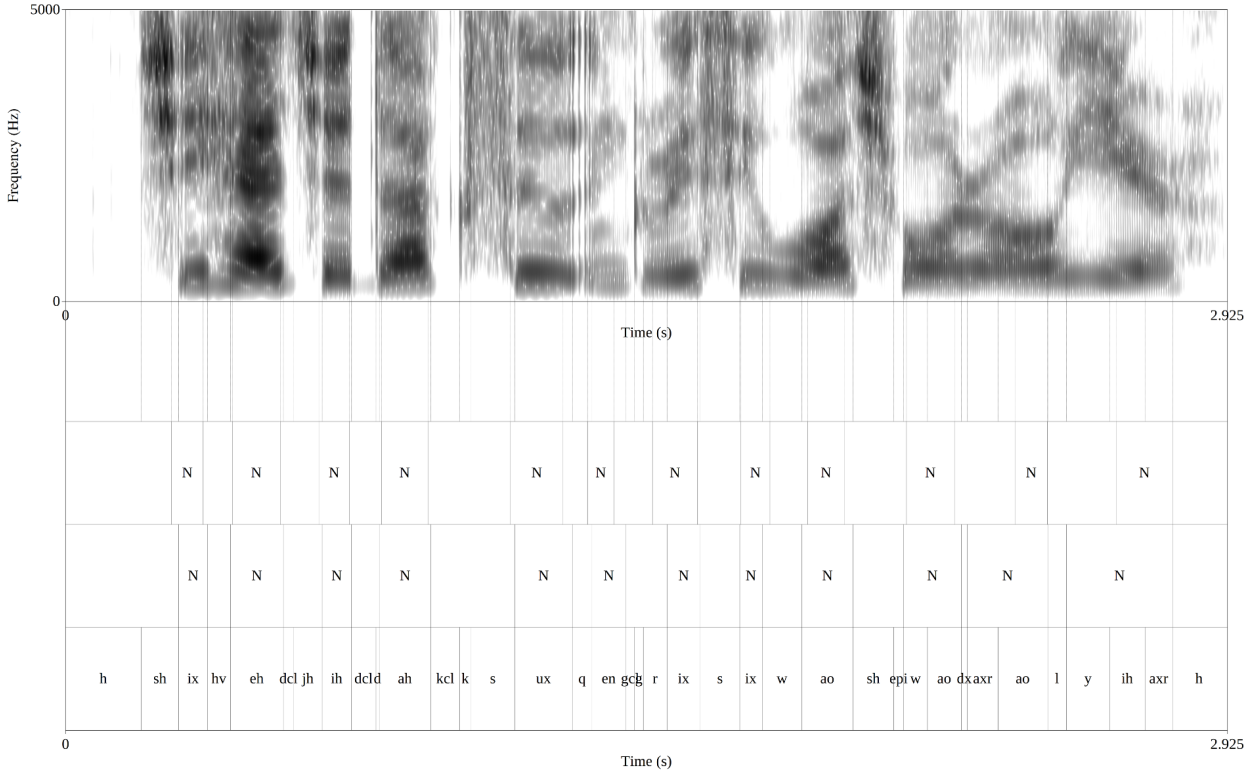

Figure 1: Automatic segmentation (first layer) of the TIMIT utterance *She had your dark suit in greasy wash water all year* compared with the reference segmentation (second layer). The phonetic transcription (third layer) is also reported.

In [2, 3], a second step based on the REC function is necessary to isolate vowels from which to extract acoustic features for acted emotion recognition. In this paper, we compute the cross-entropy profile of a given utterance using a single sliding window and then analyse its first derivative to directly detect the boundaries of a syllable nucleus. The approach is simple, unsupervised and easy to deploy, and it was possible to implement it as a PRAAT script [14].

It is difficult to compare our approach with other alternatives. The ones we reported vary in goals, use different performance measures or adopt different corpora. We report here, as reference, the results obtained in [2] as they were obtained on the full TIMIT dataset, among others, whereas other studies, like [15] use a limited subset of the same corpus. In the reference work, the authors reported that the best performance was obtained on TIMIT. In their evaluation of the system in terms of vowel spotting capability, the authors reported a detection rate of $87,56\%$, an insertion rate of $7,07\%$ and a Vowel Error Rate (VER) of $19,50\%$. VER is computed as

$$\frac{N_{del} + N_{ins}}{N_{tot}} \tag{5}$$

We can use the same formula with the data provided by SCLITE to obtain a Segment Error Rate (SER), considering segments deletions (12481) and insertions (5441) over the full number of reference segments (114310) to obtain a SER of $15,67\%$. This, of course, is not directly comparable to the reference approach but, together with the other measures, suggests that the proposed algorithm, although relatively simple, provides good performance.

6. Conclusions

We have presented a method to add the capability to estimate nuclear boundaries in a simple, yet reliable, algorithm to automatically compute speech rate using the energy envelope. The original algorithm is integrated with spectral analysis using the cross-entropy profile to find nuclear onset and ending times. We performed our tests on the full TIMIT training set and found the estimate of speech rate to be more stable than the baseline. Also, the results obtained when considering time-mediated alignment show that a statistically significant improvement is obtained on the accuracy of the segmentation with respect to the heuristical one. The obtained approach, although easy to implement, appears to be close to the performance of other approaches found in the literature, although direct comparison is difficult. This makes the proposed system appealing as it represents a fast solution to syllable nuclei detection. The positioning of nuclear boundaries also appears to be accurate enough to allow the extraction of features for automatic prosodic analysis. Future work will concentrate on evaluating the impact this approach has on tasks like continuous emotion tracking and prominence detection to check if the obtained improvement brings benefit to applications based on nuclei detection. Also, the algorithm will be included in the Prosomarker tool [16].

7. Acknowledgements

Antonio Origlia's work is supported by the Italian PAC project *Cultural Heritage Emotional Experience See-Through Eyewear* (CHEESE).

8. References

[1] F. Pellegrino, J. Farinas, and J. Rouas, "Automatic estimation of speaking rate in multilingual spontaneous speech," in *Speech Prosody 2004, International Conference*, 2004.

[2] F. Ringeval and M. Chetouani, "Exploiting a vowel based approach for acted emotion recognition," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, 2008, pp. 243–254.

[3] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *Proc. of Interspeech*, 2008, pp. 2763–2766.

[4] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 29–40, 1988.

[5] G. Pradhan and S. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 854–867, 2013.

[6] B. D. Sarma, B. Sharma, S. A. Shanmugam, S. Prasanna, and H. A. Murthy, "Exploration of vowel onset and offset points for hybrid speech segmentation," in *IEEE TENCON Conference*. IEEE, 2015, pp. 1–6.

[7] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[8] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[9] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters." in *Proc. of Interspeech*, 2011, pp. 1873–1876.

[10] T. Dekens, H. Martens, G. V. Nuffelen, M. D. Bodt, and W. Verhelst, "Speech rate determination by vowel detection on the modulated energy envelope," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, 2014, pp. 1252–1256.

[11] K. M. Yorkston, V. L. Hammen, D. R. Beukelman, and C. D. Traynor, "The effect of rate control on the intelligibility and naturalness of dysarthric speech," *Journal of Speech and Hearing Disorders*, vol. 55, no. 3, pp. 550–560, 1990.

[12] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, 2005.

[13] A. Origlia, F. Cutugno, and V. Galatà, "Continuous emotion recognition with phonetic syllables," *Speech Communication*, vol. 57, pp. 155–169, 2014.

[14] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[15] B. D. Sarma, S. S. Prajwal, and S. Mahadeva Prasanna, "Improved vowel onset and offset points detection using Bessel features," in *Signal Processing and Communications (SPCOM), 2014 International Conference on*. IEEE, 2014, pp. 1–6.

[16] A. Origlia and I. Alfano, "Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabi fication." in *Proc. of LREC*, 2012, pp. 997–1002.