EVALUATION OF IMPLICIT BROAD PHONEMIC SEGMENTATION OF SPEECH SIGNALS USING PITCHMARKS

Iosif Mporas, Panagiotis Zervas, and Nikos Fakotakis

Wire Communication Laboratory, Department of Electrical & Computer Engineering, University of Patras
Rion Patras, 261 10, Patras, Greece
phone: + 30 2610 991855, fax: + 30 2610 991855, email: imporas@wcl.ee.upatras.gr
web: http://www.wcl.ee.upatras.gr/ai

ABSTRACT

In this paper, we evaluate an implicit approach for the automatic detection of broad phonemic class boundaries of continuous speech signals. The reported method is consisted of the prior segmentation of speech signal into pitch-synchronous segments, using pitchmarks location, for the computation of adjacent broad phonemic class boundaries. The approach's validity was tested on a phonetically rich speech corpus of Greek language as well as on the DARPA-TIMIT American-English language corpus. Our framework's results were very promising since by this method we achieved 25 msec accuracy of 76% and 74,9% respectively, without presenting over-segmentation on the speech signal.

1. INTRODUCTION

The segmentation of continuous speech into linguistically defined segments, such as narrow or broad phonemic segments is considered as a key issue in several speech processing areas. Speech signals that are annotated on phoneme, diphone or syllable-like level are essential for tasks such as the training of a speech recognizer [18], the building of a language identification model [20], the construction of databases, or even in various speech synthesis techniques such as formant and unit selection approaches [4]. Due to the fact that, segmentation is a time-consuming and tedious task which can be carried out only by expert phoneticians, several automated procedures have been proposed. Those approaches are roughly divided into two major categories depending on whether we possess or not knowledge of the uttered message. Those categories are known as explicit and implicit segmentation methods [15], respectively. In explicit approaches, the speech waveform is aligned with the corresponding phonetic transcription. On the other hand, in implicit approaches the phoneme boundary locations are detected without any textual knowledge of the uttered message. Although explicit approaches achieve better accuracy than implicit, the requirement of prior phoneme sequence knowledge makes them inappropriate for real life applications, such as language identification tasks.

Over the past years, several procedures for automatic segmentation of speech have proposed in the literature. In [2], Aversano et al., proposed a segmentation method which was based on the critical-band perceptual analysis of preprocessed speech that fed a decision function and reported an accuracy of 73,58% within a range of ±20 msec on DARPA-TIMIT [7]. Suh and Lee [12], proposed a structure, based on multi-layer perceptron and reported a 15msec phoneme segmentation performance of 87% with 3,4% insertion rate in speaker dependent mode. Svendsen and Kvale [13], proposed a two-stage boundary detection approach consisted of an acoustic segmentation of speech followed by an HMM based phonemic segmentation, and reported an accuracy of 80-85% for four languages and a range of 20 msec. Svendsen and Soong [14] presented an accuracy of 73% within three frames, based on a constrained-clustering vector quantization approach. Grayden and Scordilis [6], proposed a Bayesian decision surface for dividing speech into distinct obstruent and sonorant regions and applied to each of them specific rules; an 80% of accuracy was reported with an insertion rate of 12%. In conclusion, an approach similar to our method was proposed in [5], which was taking advantage of the visual clues at each pitch period for the detection of the voiced phoneme boundaries.

In this work we evaluate an implicit method for the automatic detection of boundaries of broad phonemic classes, using *pitchmarks* [11] locations. In particular, we segment the speech signal into voiced phoneme segments and unvoiced intervals. With regard to voiced segments, they were chunked pitch-synchronously from the pitchmark locations into fragments. Subsequently, we compare the frame contours using the well established, dynamic time warping (DTW) [3] algorithm to compute the distance path between adjacent frames. Finally, the local maximums of the resulted distance path contour correspond to broad phoneme class boundaries. In contrast to [5] explicit approach, our implicit method does not use geometrical features to detect signal changes related to co-articulation. Contrarily, we utilize DTW on a smoothed transformation of the speech waveform for the task of extracting broad phonemic class sequences.

2. METHOD DESCRIPTION

Our method depends on the hypothesis that the voiced parts of a speech signal are composed of periodic fragments produced by the glottis during vocal-fold vibration [11]. By this hypothesis and as the articulation characteristics are almost constant in the middle of a voiced segment, each of these fragments will differ from its adjacent ones at the co-articulation regions, where the candidate voiced-phoneme boundaries reside. The above contemplation leads to segmentation of the speech waveform to voiced phoneme segments and unvoiced interval segments. Apparently, since an unvoiced phoneme is between two voiced its boundaries are detected. Otherwise, the algorithm detects the boundaries of the interval that contains an unvoiced phoneme sequence.

For this purpose, we initially segment the speech signal into fragments determined by the pitchmarks location. Subsequently, a moving average smoothing is applied to each fragment for the task of abrupt local irregularities reduction.

Ultimately, we utilize an evaluation algorithm for the measurement of the distance between adjacent smoothed fragments. In that way we were able to detect the co-articulation points, which correspond to the voiced phoneme boundaries. A general diagram of the method outline is illustrated in figure 1.

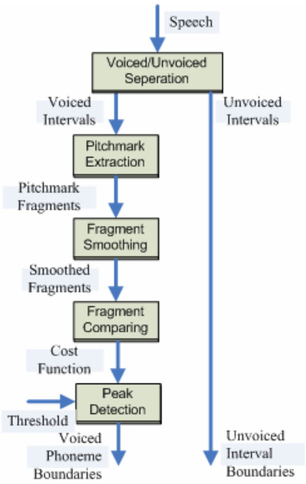

Figure 1 - Block diagram of the proposed procedure.

2.1 Pitchmark extraction algorithm

For the extraction of pitchmarks we used the point process algorithm of Praat [9]. The voiced intervals are determined on the basis of the voiced/unvoiced decisions extracted from the corresponding F_0 contour. For every voiced interval, a number of points (glottal pulses) are found. The first point, t_1, is the absolute extremum of the amplitude of the sound

$$t_1 = \max\left[t_{mid} - T_0/2,\ t_{mid} + T_0/2\right] \qquad (1)$$

where t_{mid} is the midpoint of the interval, and T_0 is the period at t_{mid}, as can be interpolated from the pitch contour. Starting from time instant t_1, we recursively search for points t_i to left until we reach the left edge of the interval. These points must be located between $t_{i-1} - 1.2T_0(t_i -1)$ and $t_{i-1}-0.8T_0(t_i-1)$, and the cross-correlation of the amplitude of the environment of the existing point t_{i-1} must be maximal. Between the samples of the correlation function parabolic interpolation has been applied. The same procedure is followed and for the right of t_1 part of the particular voiced segment.

Though the voiced/unvoiced decision is initially taken by the pitch contour, points are removed if the correlation value is less than 0.3. Furthermore, one extra point may be added at the edge of the voiced interval if its correlation value is greater than 0.7. An example of the detection of the first two pitchmarks t_1 and t_2 of a voiced speech interval is illustrated in figure 2.

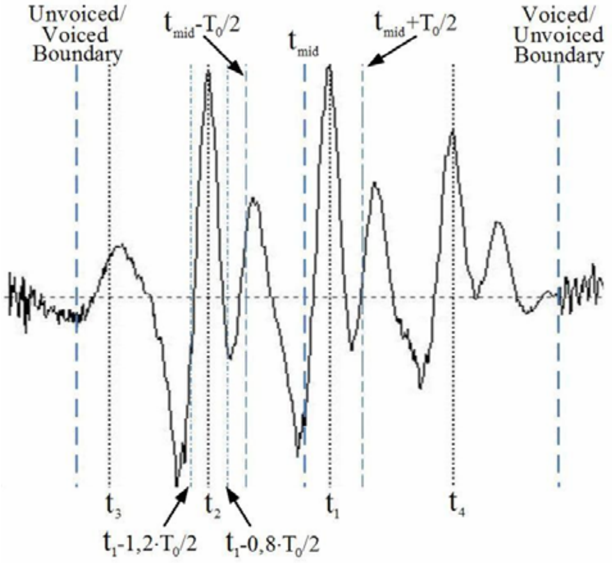

Figure 2 - Pitch-mark extraction from speech signal

2.2 Broad Phoneme Class Boundary Detection

Voiced phoneme boundaries are observed into speech regions that are marked with heavy co-articulation phenomena. Since the manner of articulation is almost constant during each specific phoneme, fragments lying in the same phoneme and away from the co-articulation regions have similar amplitude evolution. In contrast to that, frames that are located in such regions will have different contours while the articulation behaviour changes.

For the task of calculating the difference between the amplitude contour of each fragment and its adjacent ones, we employed the dynamic time warping (DTW) algorithm [3], which computes the distance path between each pair of following fragments of speech that are determined by the pitchmarks. As a consequence a cost function is computed for each pair of adjacent fragments.

$$CostFunction(i) = DTW(\,fragment(i),\ fragment(i+1)) \qquad (2)$$

Consequently, equation 2 provides a measure of similarity between adjacent fragments of the speech waveform. In other words, the local maxima of the function are equivalent to the phoneme boundaries of the utterance, since the warping path between the adjacent fragments is longer. A typical contour of the computed cost function is illustrated in figure 3.

As a final step, peaks of the cost function are detected. To decide which of the peaks correspond to candidate segment boundaries a threshold operational parameter, *Thr,* is introduced. For each peak we calculate the magnitude distances from its side local minimums. The minimum of the two resulted magnitude distances is compared to *Thr*. For values higher to *Thr* the corresponding fragment is considered to contain a detected boundary. For values lower than *Thr* the related peak is ignored. Finally, each detected boundary is assumed to be located on the middle sample of the prior chosen fragment.

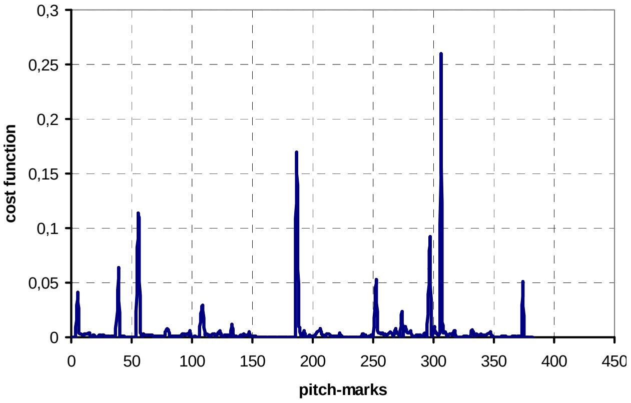

Figure 3 - Cost function for the identification of the boundaries

3. SPEECH CORPORA

The validation of the proposed technique for implicit voiced-phoneme segmentation was carried out with the exploitation of two databases: DARPA-TIMIT and WCL-1 [19].

As regards DARPA-TIMIT, it is considered as an acoustic-phonetic continuous speech corpus that contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. It includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. The DARPA-TIMIT corpus transcriptions have been hand verified. Test and training subsets are balanced for phonetic and dialectal coverage.

Concerning the WCL-1, is regarded as a phonetically and prosodically balanced corpus of Greek speech annotated on phonemic level. It is consisted of 5.500 words, distributed in 500 paragraphs, each one of which may be a single word utterance, a short sentence, a long sentence, or a sequence of sentences. Newspaper articles, paragraphs of literature and sentences were used, in order to cover most of the contextual segmental variants. Furthermore, the data was processed so as targeted phenomena could be obtained in a real communicative way and appropriate text was composed by linguist where was need. The database was phonetically annotated by expert phonetician. For the high quality narrow phonetic transcriptions, the SAMPA alphabet adapted for Greek was employed.

4. EVALUATION

For the task of evaluating our broad phonemic class segmentation framework we have conducted experiments with both databases practicing different thresholds. A detected segmentation point is defined as *correctly-detected* only if its distance from the true segmentation point is less than *t* msec. In order to compute the performance of the method we introduce accuracy and over-segmentation. Accuracy is defined as the percentage of the number of the *correctly-detected* segmentation points P_c to the total number of the *real-boundary* points P_t,

$$Accuracy = \frac{P_c}{P_t} \cdot 100\% \qquad (3)$$

where the *real boundary* points are the boundaries of the voiced phonemes and the boundaries of the unvoiced intervals.

As regards explicit approaches, the number of detected segmentation points is equal to the number of the true segmentation points. In contrast, regarding implicit approaches, where our method falls, detected segmentation points are not equal to the true ones. An effective way of measuring the reliability of a segmentation method regarding the estimated and actual number of boundary location is over-segmentation measure. Over-segmentation is defined as the ratio of the number of the detected segmentation points P_d to the total number of the true segmentation points P_t,

$$Over-Segmentation = \frac{P_d}{P_t} \qquad (4)$$

It is clear from equation 4 that over-segmentation near to one denotes that the number of the estimated boundaries is close to the actual number of boundaries.

4.1 Results

For the evaluation of the proposed technique several experiments were carried out. Our main purpose was the accuracy improvement while keeping the over-segmentation factor close to the value of one. As a result, a vast variety of threshold values were tested for several smoothing factors. In addition, we investigated the accuracy of our procedure for *t*=25msec. Results regarding the achieved accuracy for the DARPA-TIMIT American-English corpus are illustrated in figure 4.

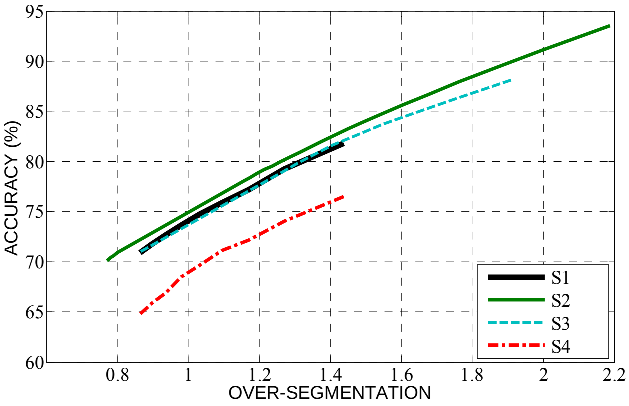

Figure 4 – Broad phonemic segmentation accuracy with respect to over-segmentation for different smoothing factors S (S1=70, S2=100, S3=140, S4=1) on DARPA-TIMIT.

Furthermore, figure 4 presents an empirical way for selecting practically optimal values for free parameters such as smoothing factor and threshold. In that way, the accuracy of the method could be optimized.

The best result obtained through the optimization procedure was 74,9%, without presenting over-segmentation, for a smoothing factor equal to 100 and *Thr*=$1{,}25 \cdot 10^{-6}$, (Over-Segmentation<1,05). Accuracy of the method could be further elevated if higher values of over-segmentation are accepted. In previous research [10] has been demonstrated that over-segmentation control is a tedious task with values higher than 1. For over-segmentation of 1,6 our method achieved about 85% accuracy, as shown in figure 4.

Regarding WCL-1 Greek speech corpus, figure 5 depicts the segmentation performance that was achieved. In particular, the best obtained accuracy was 76% with over-segmentation less than 1,05, a smoothing factor equal to 80 and a *Thr* value of $2{,}5 \cdot 10^{-4}$. For an over-segmentation equal to 1,6 the method achieved an accuracy of more than 90%, as illustrated in figure 5.

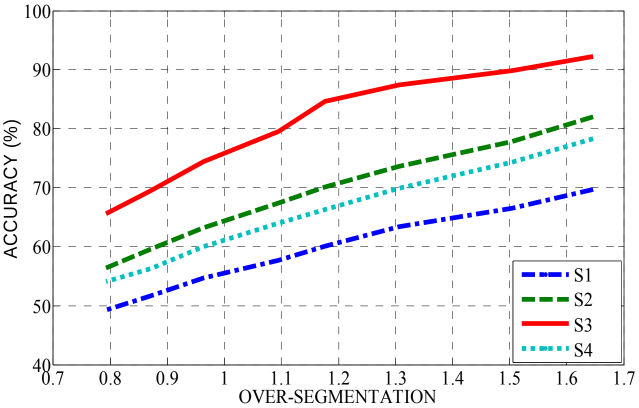

Figure 5 – Broad phonemic segmentation accuracy with respect to over-segmentation for different smoothing factors S (S1=1, S2=50, S3=80, S4=130) on WCL-1.

5. CONCLUSIONS

In this work, we have implemented and evaluated a speaker independent method for automatic broad phoneme class segmentation of speech signals using the knowledge of pitch-mark locations. For the approach's validity, experiments were conducted on a phonetically rich speech corpus of Greek language as well as on DARPA-TIMIT American-English database. Specifically, segmentation experiments on DARPA-TIMIT showed an accuracy of 74,9%. On the other hand, WCL-1 database accuracy was 76%. Due to the fact that the textual message of the speech utterance in not necessary for the extraction of the boundary locations makes it appropriate for applications that require automatic broad annotation of speech.

Future research will focus on the utilization of the described approach for online segmentation of continuous speech for tasks such as language identification, emotion recognition and the rapid development of resources speech synthesis tasks.

REFERENCES

[1] Adami A., Hermansky H., "Segmentation of speech for speaker and language recognition", In Proceedings of EUROSPEECH 2003, pp. 841-844, Geneva 2003.

[2] Aversano G., Esposito A., Esposito A., Marinaro M., "A new text-independent method for phoneme segmentation", In Proceedings of the 44th IEEE Midwest Symp. Circuits and Systems, vol. 2, pp.516-519, 2001.

[3] Deller J., Proakis J., Hansen J., "Discrete-time processing of speech signals", MacMillan Series for Prentice-Hall Publishers, New York, 1993.

[4] Dutoit, T., "An Introduction to Text-To-Speech Synthesis", vol. 3, Text, Speech and Language Technology. Kluwer Academic Publishers, 1997.

[5] Essa O., "Using prosody in automatic segmentation of speech", In Proceedings of 36th ACM Southeast Regional Conference, 1998.

[6] Grayden D., Scordilis M., "Phonemic segmentation of fluent speech", In Proceedings of ICASSP 1994, pp.73-76, 1994.

[7] NIST Speech Disc CD1-1.1.

[8] Paul Boersma & David Weenink (2005): Praat: doing phonetics by computer. Retrieved from http://www.praat.org/.

[9] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", In Proceedings of IFA, 17: 97-110, 1993.

[10] Petek B., Andersen O., Dalsgaard P., "On the robust automatic segmentation of spontaneous speech", In Proceedings of ICSLP '96, pp. 913-916, 1996.

[11] Reddy, D., R., "Pitch Period Determination of Speech Sounds", Communication of the ACM, vol. 10, pp. 343-348.

[12] Suh Y., Lee Y., "Phoneme segmentation of continuous speech using multi-layer perceptron", In Proceedings of ICSLP '96, pp. 1297–1300, 1996.

[13] Svendsen T., Kvale K., "Automatic alignment of phonemic labels in continuous speech", In Proceedings of ICSLP '90, Kobe, Japan, 1990.

[14] Svendsent T., Soong F. K., "On the automatic segmentation of speech signals", In Proceedings of ICASSP '87, pp.77-80, Dallas, April 1987.

[15] van Hemert J., "Automatic Segmentation of Speech", IEEE Transactions on Signal Processing, vol. 39, no. 4, April 1991.

[16] Wang D., Lu L., Zhang H.-J., "Speech segmentation without speech recognition", In Proceedings of IEEE ICASSP '03, Vol. I, pp. 468-471, Hong Kong, April 4-10, 2003.

[17] Wokurek W., "Entropy rate-based stationary/ non-stationary segmentation of speech", PHONUS 5, pp. 59-71. Saarbrócken: Institute of Phonetics, University of the Saarland, 2000.

[18] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., P. Woodland P., "The HTK Book", Revised for HTK Version 3.0, July 2000.

[19] Zervas P., Fakotakis N., Kokkinakis G., "Development of a prosodic database for greek speech synthesis", In Proceedings of SPECOM 2005,pp. 603-606, Patras, Greece, 2005.

[20] Zissman M., "Comparison of four Approaches to Automatic Language Identification of Telephone Speech", IEEE Trans. Speech and Audio Proc., SAP-4, pp.31-44, Jan.96.