Broad Phonemic Class Segmentation of Speech Signals in Noise Environments

Iosif Mporas, Panagiotis Zervas, Nikos Fakotakis

Artificial Intelligence Group, Wire Communication Laboratory
Department of Electrical and Computer Engineering, University of Patras
261 10, Rion Patras, Greece, phone: + 30 2610 991855, fax: + 30 2610 991855
{imporas, pzervas, fakotaki}@wcl.ee.upatras.gr

Abstract

In this paper, we evaluate the performance of an implicit approach for the automatic detection of broad phonemic class boundaries from continuous speech signals in different additive noise environments. We exploit the prior knowledge of glottal pulse locations for the estimation of adjacent broad phonemic class boundaries. The approach's validity was tested on the DARPA-TIMIT American-English language corpus and NOISEX-92 database. Our framework's results were very promising since by this method we achieved 25 msec accuracy of 74,9% for un-noisy environment, while the performance reduced about 5% for wideband distortion noise.

1. Introduction

Speech signals that are annotated on phoneme, diphone or syllable-like level are essential for tasks such as, speech recognition [1], construction of language identification models [2], prosodic database annotation, and in speech synthesis assignments such as formant and unit selection techniques [3]. Since, segmentation of speech signals is a time-consuming and tedious task which can be carried out only by expert phoneticians, several automated procedures have been proposed. Speech segmentation methodologies can be classified into two major categories depending on whether we possess or not knowledge of the uttered message. These categories are known as explicit and implicit segmentation methods [4], respectively.

Regarding 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.

In the area of automatic speech segmentation extensive research has been conducted. Aversano et al. [5], proposed a segmentation method 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 [6]. Suh and Lee [7], 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 [8], 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 [9] presented an accuracy of 73% within three frames, based on a constrained-clustering vector quantization approach. Grayden and Scordilis [10], 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%. An approach similar to our method was proposed in [11], which was taking advantage of the visual clues at each pitch period for the detection of the voiced phoneme boundaries. In conclusion Pellom and Hansen [12] evaluate an HMM based explicit segmentation approach in a variety of additive noise environments. Since most real life applications operate in noise environments we focus in the evaluation of our implicit, pitch-synchronous method of detecting broad phonemic class boundaries from speech signals with additive noise.

Initially, segmentation of the speech signal into voiced phoneme segments and unvoiced intervals is carried out. Subsequently voiced segments, were chunked pitch-synchronously according to pitchmark locations into fragments followed by the comparison of the frame contours using the well established, dynamic time warping (DTW) [14] algorithm for the computation of the distance path between adjacent frames. Finally, the local maximums of the resulted distance path contour correspond to broad phoneme class boundaries.

The outline of the paper is as follows. Section 2, describes the proposed method. In section 3 the utilized speech corpora is presented and in section 4 we discuss the results.

2. Segmentation Methodology

Our method builds on the theory that, *voiced parts of a speech signal are composed of periodic fragments produced by the glottis during vocal-fold vibration* [13]. Furthermore, since the articulation characteristics of voiced phonemes are almost constant in the middle of their region, co-articulation regions will be possible places for a possible phoneme boundary to reside. By following this observation we are led to segmentation of speech waveform to broad phonemic classes consisting of voiced phoneme segments and unvoiced intervals.

As regards the unvoiced phoneme sequence, it could be recognized with the utilization of a language model, fed with the neighboring recognized voiced phoneme sequences. In the case which an unvoiced interval consists of one phoneme, its boundaries are detected from the adjacent voiced phoneme boundaries.

In the framework of our approach we initially segment the speech signal into voiced and unvoiced intervals, using

Boersma's algorithm [15]. This method uses the short-term autocorrelation function r_x of the speech signal:

$$r_x(\tau) \equiv \int x(t)x(t+\tau)dt \quad (1)$$

The pitch is determined as the inverted value of τ corresponding to the highest of r_x. Threshold values for silence, voiced and unvoiced detection are introduced in order to extract the corresponding intervals.

After distinguishing voiced and unvoiced regions, voiced speech is segmented to fragments determined by the pitchmarks location. Subsequently, a moving average smoothing is applied to each fragment for the task of abrupt local irregularities reduction.

Finally, we utilize an evaluation algorithm for the measurement of the distance between adjacent smoothed fragments. In that way we detect the co-articulation points, which correspond to the voiced phoneme boundaries.

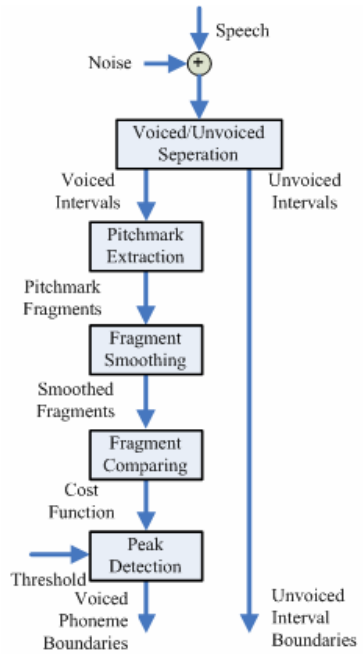

Figure 1: Block diagram of the proposed procedure.

The above figure illustrates a general diagram of the proposed methodology. It clearly shows that our approach results boundaries for the voiced phonemes and for adjacent unvoiced phoneme sequences.

2.1. Pitchmark extraction algorithm

For the extraction of pitchmarks we have used the point process algorithm of Praat [16]. The voiced intervals are determined on the basis of the voiced/unvoiced decision 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

$$t_1 = max[t_{mid} - T_0/2, t_{mid} + t_0/2] \quad (2)$$

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.

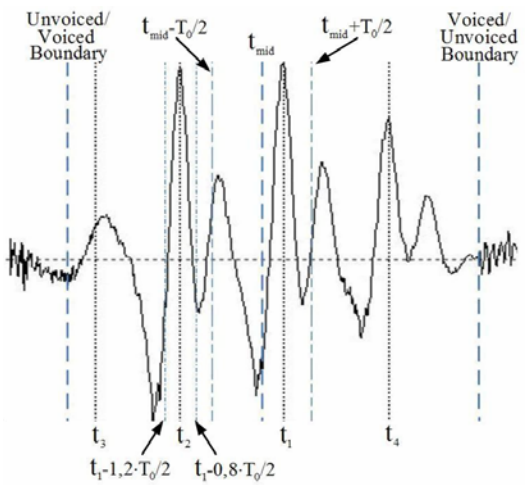

Figure 2: Pitch-mark extraction from speech signal

2.2. Voiced phoneme boundary detection

As discussed in a previous section, heavy co-articulation phenomena mark the transition from on phoneme to the next. Fragments lying in the same phoneme and away from the co-articulation regions have similar amplitude contours. On the other hand, fragments that are located in co-articulation regions will be rendered with variations in their amplitude contours, resulted from the changes in the articulation.

In calculating the difference between the amplitude contour of each fragment and its adjacent one, we have employed the dynamic time warping (DTW) [14] algorithm. DTW calculates the distance path between each pair of successive fragments of speech that are determined by the pitchmarks. As a consequence the outcome of a cost function is computed for each pair of adjacent fragments.

$$Cost\ Function\ (i) = DTW(fragment(i), fragment(i+1)) \quad (3)$$

In other words, equation 3 could be described as a measure of similarity between adjacent fragments of a speech waveform. 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. An example of a typical contour of the computed cost function is illustrated in figure 3.

As a final step in our approach is the detection of peaks in the cost function. In order 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 possible boundary. A peak related to values that is lower to *Thr,* is ignored. Finally, each detected boundary is assumed to be located on the middle sample of the prior chosen fragment.

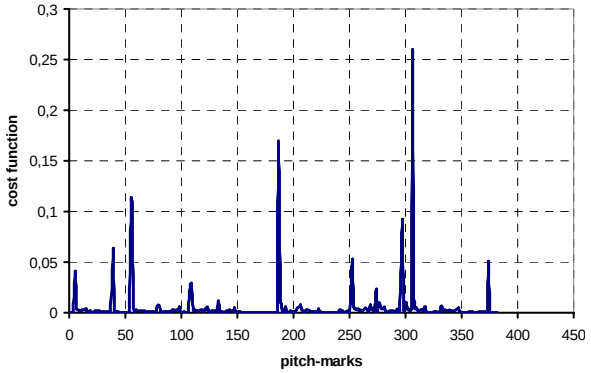

Figure 3: Cost function for the identification of the boundaries

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 NOISEX-92 [17].

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 NOISEX-92 database, it contains recordings of various noises such as voice babble, factory noise, high-frequency radio channel noise, pink noise, white noise. In addition, there are provided various military noises, as fighter jets (Buccaneer, F16), destroyer noises (engine room, operations room), tank noise (Leopard, M109) and machine gun noise. Finally, car noise (Volvo 340) is provided.

4. Performance Evaluation

For the task of evaluating our broad phonemic class segmentation framework, we conducted experiments in both un-noisy and noisy environments practicing different thresholds.

A segmentation point is defined as *correctly-detected* only if its distance from the actual annotation point is less than t msec. In order to measure the performance of our method we introduce accuracy metric 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,

$$Accuracy=P_c/P_t \cdot 100\% \quad (4)$$

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

Regarding 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,

$$Over\text{-}Segmentation=P_d/P_t \quad (5)$$

It is clear from equation 5 that over-segmentation near to one means that the number of the estimated boundaries is close to the actual number of boundaries.

4.1. Results

In this section we present and discuss the results of the carried out experiments. We have focused on improving accuracy while keeping the over-segmentation factor close to the value of one. As a result, a vast variety of threshold values was tested for several smoothing factors. Additionally, we investigated the accuracy of our procedure for *t*=25msec. Results that present the achieved accuracy for the DARPA-TIMIT American-English corpus in un-noisy environment are illustrated in figure 4.

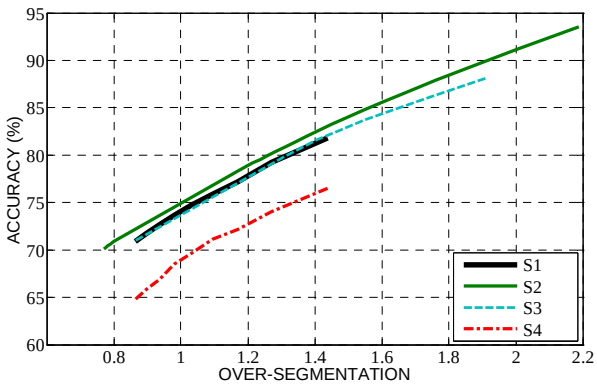

Figure 4: Broad phonemic segmentation accuracy with respect to over-segmentation for different smoothing factors S (S1=70, S2=100, S3=140, S4=1) on DARPA-TIMIT.

Figure 4 presents an empirical way for selecting practically optimal values for the free parameters such as the smoothing factor and the threshold. In that way, we were able to optimize the accuracy of our method.

The best results were obtained through the optimization procedure was 74,9%, without presenting over-segmentation, for a smoothing factor equal to 100 and *Thr*=1,25·10-6, (Over-Segmentation<1,05). Accuracy of the method could be further elevated if higher values of over-segmentation are accepted. In previous research [18] 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.

The next step to our experimental procedure was the evaluation of our segmentation schema in several additive noise environments. For this task we selected *white noise, Gaussian white noise, voice babble, noise in pilot cockpit, tank noise, HF radio channel noise, car noise, pink noise* and

machine gun noise.

Sentences from the TIMIT database were corrupted by the various noise conditions at a global SNR of 10 dB. The method's accuracy under noise conditions was tested using the practically optimal values of the parameters as were obtained from the un-noisy environment experiments (S=100, Thr=1,25·10-6). The achieved results are tabulated in Table 1.

Table 1: Speech segmentation accuracy using standard scoring method for additive noise environments (SNR=10dB)

ADDITIVE NOISE	ACCURACY
No noise	74,90%
White noise	70,21%
Gaussian white noise	70,01%
Speech in background (voice babble)	69,53%
Pink noise	69,50%
Noise in pilot cockpit (F-16)	69,36%
Tank noise (Leopard, M109)	69,24%
HF radio channel noise	69,05%
Car noise (Volvo 340)	68,11%
Machine gun	57,41%

Table 1 clearly presents that our method performs equally well in noise environments with high frequency, as well as with wideband distortion characteristics, like HF radio channel noise or Gaussian white noise respectively. Method's accuracy reduces significantly in the case of machine gun noise since it is described by high colored energy distribution characteristics. It causes such a distortion to the speech waveform contour that smoothing process effort to maintain the articulated glottal pulse contour, of the compared fragments, performs poorly.

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 pitchmark locations in un-noisy and noisy environments. For the approach's validity, experiments were conducted on DARPA-TIMIT American-English and NOISEX-92 databases. Segmentation experiments without noise showed an accuracy of 74,9%. On the other hand, the method demonstrated robustness for wideband distortion noise characteristics. Given the fact that the textual message of the speech utterance in not necessary for the extraction of the boundary locations as well as its robustness to noisy environments, makes it appropriate for applications that require automatic broad annotation of speech in real environment conditions.

6. References

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

[2] 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.

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

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

[5] 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.

[6] Garofolo J., "Getting started with the DARPA-TIMIT CD-ROM: An acoustic phonetic continuous speech database", National Institute of Standards and Technology (NIST), Gaithersburgh, MD, 1988.

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

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

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

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

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

[12] Pellom B., Hansen J., "Automatic segmentation of speech recorded in unknown noisy channel characteristics", Speech Communication, 25, pp. 97-116, 1998.

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

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

[15] Boersma, P., "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.

[16] Boersma, P., Weenink, D., Praat: doing phonetics by computer, (2005). Retrieved from *http://www.praat.org/*

[17] Varga, A., Steenneken, H., J., M., Tomlinson, M., and Jones, D., The NOISEX-92 study on the effect of additive noise on automatic speech recognition, 1992. Documentation included in the NOISEX-92 CD-ROMs.

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