THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.24, NO.4E 2005, 12 pp. 128-131

A New Method for Segmenting Speech Signal by Frame Averaging Algorithm

Byambajav.D*, Chul-Ho Kang*
*Department of Electronics and Communications Engineering, Kwangwoon University
(Received September 16 2005; rivised October 27; accepted December 26 2005)

Abstract

A new algorithm for speech signal segmentation is proposed. This algorithm is based on finding successive similar frames belonging to a segment and represents it by an average spectrum. The speech signal is a slowly time varying signal in the sense that, when examined over a sufficiently short period of time (between 10 and 100 ms), its characteristics are fairly stationary. Generally this approach is based on finding these fairly stationary periods. Advantages of the algorithm are accurate border decision of segments and simple computation.

The automatic segmentations using frame averaging show as much as 82.20% coincided with manually verified segmentation of CMU ARCTIC corpus within time range 16 ms. More than 90% segment boundaries are coincided within a range of 32 ms. Also it can be combined with many types of automatic segmentations (HMM based, acoustic cues or feature based etc.).

Keywords: *Boundary Refining, Frame Averaging, Speech Signal Segmentation*

I. Introduction

Speech technology development is strongly related to corpus-based methodologies and to quality and availability of good speech corpora. In order a corpus to be very desirable and useful, it should contain information about speech contents or, speech signals should be phonetically segmented and labeled. The very precise way to obtain this information is manual segmentation. However, manual segmentation and labeling (especially segmentation) are very costly, time consuming and required much effort. For preparing a large inventory of subword units or phonemes, an automatic segmentation is more desirable to manual segmentation as it substantially reduces the work. Researchers try to use many different techniques and features for automatic phonetic segmentation[1-3]. The most frequent approach for automatic phonetic segmentation is to modify an HMM based phonetic recognizer to adapt it to the task of automatic phonetic segmentation.

HMMs and other techniques automatically produce segmentation, but it is still less precise than manual segmentation. Manual segmentation has two major drawbacks, first the process is both laborious and tedious, requiring extensive listening and spectrogram interpretation. Second, due to the subjective nature of a manual process, there will be inconsistencies from trial to trial, even for segmenting the same utterance and even this process requires linguistic experiences[4, 6].

Researchers always try to develop accurate automatic segmentation techniques and methodologies, and evaluate the automatic approaches by comparing the segmentations with manual segmentation and by computing some figures of merit. Moreover most of automatic segmentation techniques start from manually segmented speech databases for training or testing purposes to make the reference template or pattern[1].

During the last few years the need has raised the interest in segmentation techniques to develop new voices and spoken languages quickly and also the maximum quality.

The question how to estimate the quality of the available

Corresponding author: Chul-Ho Kang (chkang5136@kw.ac.kr)
Kwangwoon University, 447-1 Wolgye-Dong, Nowon-Ku, 139-701 Seoul, Korea

segmentations becomes even more important issue for the automatic segmentation systems to satisfy the accuracy near to manual segmentation, and then the comparison between manual segmentation with automatic one has no meaning.

In order to alleviate these problems for both of manual and automatic segmentations, a new algorithm for segmenting speech into sub-word units is proposed in this paper.

II. Frame Averaging Algorithm

This algorithm is based on finding successive similar frames belonging to a segment and each of segments may be described in terms of length and average spectrum. This representation of segments is reported in[5].

Assume that we have observed a sequence of N speech frames {x1,x2,...,xN} with corresponding spectral representations {X1,X2,...,XN}. We wish to segment the utterance into m consecutive segments where each segment corresponds to a subword unit or relatively stationary period.

The number of segments m is assumed to be known. If we denote ending frame of segment i is bi, the ith segment starts in frame bi-1+1 and ends in frame bi. Our objective is to find boundaries {b0,b1,...,bm}, obviously b0=0 and bm=N. To find the boundaries or to divide into m segments a number of segmenting iteration is repeated until n=m based on frame averaging algorithm, where n is number of subsegments after segmenting iteration or segment number in the middle of processing. Obviously before first iteration subsegment number n is equal to total frame number (n=N) and after finish segmentation n=m.

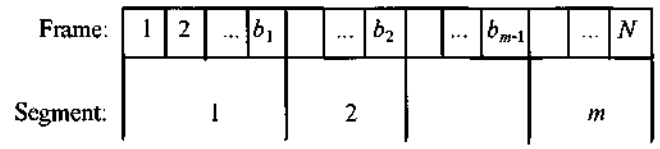

Figure 1. Segmentation of frames into m segments.

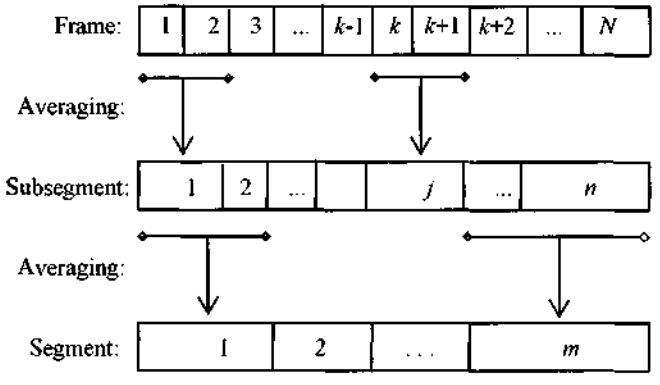

Figure 2. Frame averaging segmentation of frames into m segments.

Two main approaches are developed for segmentation. First algorithm is to find shortest spectral distance from all of pair neighbor frames and to average the most similar frame spectrums then to modify spectrogram by replacing similar frame spectrums with averaged frame spectrum, then repeat the iteration. Second algorithm is to average more similar neighbor frames of successive three frames and to modify spectrogram by replacing averaged frame spectrum, then continue it for next three frames. Before starting procedures all frames are assumed as subsegment or n = N.

2.1. Averaging Most Similar Neighbor Frames

The first approach of averaging most similar neighbor frames can be described in the following way:

1. Calculate all log spectral distances of neighbor subsegments.

2. Find shortest log spectral distance and average the corresponding subsegment spectrums. Actually subsegment spectrum is averaged spectrum of frames due to previous iteration, thus the averaging must be weighted with number of frames in subsegment.

$$X'(\omega, j) = \frac{w(k+1)\log X(\omega, k+1) + w(k)\log X(\omega, k)}{w(k+1) + w(k)} \quad (1)$$

where w is weight or number of averaged frames in the subsegment (duration of the subsegments).

3. Modify spectrogram by averaged spectrum.

Most similar successive subsegment's spectrums are replaced by their averaged spectrum. Total number of subsegments is reduced by one and new subsegment will be generated with duration w'(j) = w(k+1) + w(k).

4. Repeat 1, 2 and 3 until subsegments number is reduced to the required number or n=m.

2.2. Forward Averaging

Difference of this approach from the previous one is no search to the most similar frames. It is based on finding more similar frames from successive three frames and/or subsegments.

1. Calculate all log spectral distances of neighbor subsegments. Middle subsegment is 2nd subsegment.

2. Measure distances of first and middle subsegments, and middle and last subsegments for selected three subsegments.

3. If first and middle subsegments are more similar, average them and increase index by 2 for middle subsegment.

If middle and last subsegments are more similar, average them and increase the index by 3 for middle subsegment.

(Averaging is same as the previous approach)

4. Repeat 2 and 3 until reach to last subsegment.

5. Modify spectrogram and subsegments. It is same as the previous approach.

6. Repeat 1-4 until reach to desired number of subsegments.

III. Implementation of Frame Averaging Algorithms

The two algorithms developed have advantages and disadvantages relatively to each other. Advantage of averaging most similar frames is that the iteration can be finished with a desired number of subsegment or segment. Drawback of this approach is that some frame may remain without belonging to a segment due to noise and non-stationary sounds, because its spectrum is sometimes very different from neighbor frame spectrums. Forward averaging algorithm eliminates this problem. Two of three frames must be averaged for current selected frames. Disadvantage of this approach is the number of generated subsegments not predictable.

Combination of these approaches can solve the problems. In this case forward averaging must be implemented before averaging most similar frames. The frame averaging algorithm in many ways can be applied to segmenting speech signals.

For manual segmentation the averaging algorithm can improve accuracy or eliminate discrepancies between two manual segmentations, and reduce time consuming of the process. It is helpful to obtain same border locations for trial-to-trial procedures, because a number of border locations can be generated before segmentation and a man only can make decisions which border locations are correct.

The applications of the frame averaging algorithm for automatic segmentations have followed one of two basic approaches to the problem. The first approach is to utilize the explicit information that is known a priori, such as the correct number of phonemes, phonetic transcriptions and/or reference templates. The second approach does not require any explicit information, but should utilize combination with other automatic segmentation techniques such as HMM.

Table 1. Percentage of segmentation difference smaller than several tolerances (8,16,32,64ms) between segmentation of CMU ARCTIC corpus and proposed segmentation with different frame steps.

	Frame step (sample)	< 8 ms	< 16 ms	< 32 ms	< 64 ms
Database Set1	128	55.26	74.18	95.73	99.40
	256	51.18	73.20	94.99	98.98
	512	42.70	80.53	92.86	99.21
Database Set2	128	50.81	71.44	94.85	98.79
	256	55.35	73.39	94.81	99.35
	512	41.72	82.20	92.44	99.54

IV. Experiment

Experiment is performed on data taken from CMU ARCTIC speech corpus designed for the purpose of speech synthesis research or more precisely segmented. The CMU ARCTIC corpus consists of four primary sets of recordings (3 males, 1 female) with automatically segmented phonetic labels and hand pruning is performed for examination. The database is consisted of nearly 1150 phonetically balanced English utterances (approximately 39000 phonemes). The speech was recorded with a microphone at 16 kHz in clean environment.

A measure of the discrepancies between two different segmentations can be interpreted as a measure of the performance of new segmentation technique.

The frame averaging segmentation is evaluated for given number of segments by measuring boundary locations with deviation between segmentation of CMU ARCTIC corpus and

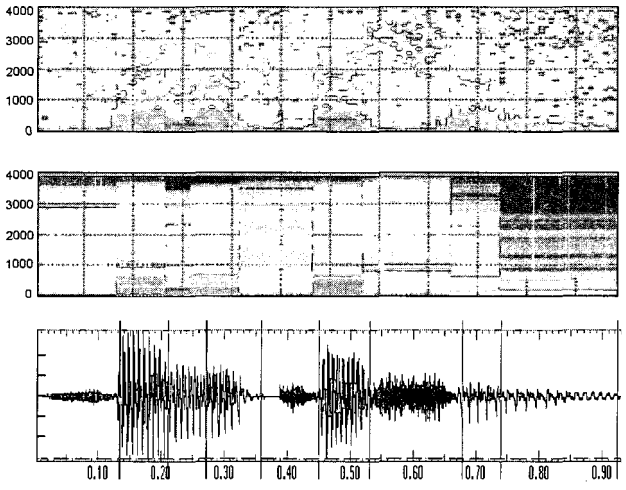

Figure 3. Result of segmentation for 'phonetician' signal.

proposed segmentation with tolerances 8, 16, 32 and 64 ms. Table 1 shows the data of percentage of segmentation difference smaller than several tolerance (8, 16, 32 and 64ms) between segmentation of the corpus and proposed segmentation with different number of frame steps for two primary sets of recordings. The two sets are speech databases for same utterances and different speakers. For example, meaning of 82.20% is percentage of boundary coincidence segmentation of the corpus and proposed segmentation within time range 16 ms.

Figure 3 shows an example, which is a result of segmentations for signal phonetician'. The bottom figure indicates manual segmentation, middle figure indicates frame averaging segmentation and average spectrum representations of segments, and upper figure indicates spectrogram of the signal.

V. Conclusions and Future Work

This study shows the possibility of using frame averaging technique for speech signal segmentations and boundary refinement. Numerical evaluations for manual segmentations are hard because there is no estimation of how much time consuming for the segmentation. But we just observed that making decisions is very easy for selecting the correct boundary of phonetic from subsegment's boundaries.

The automatic segmentations using frame averaging show as much as 82.20% coincided with segmentation of CMU ARCTIC corpus within time range 16 ms. More than 90% segment boundaries is coincided within a range of 32 ms. These are remarkable results obtained by the frame averaging segmentation technique, which can be implemented for speech signal segmentation.

This figure of merit commonly reported in several research works reveals that the good results (around 90% of segment boundaries coincided within a range of 20 ms) have been achieved with HMMs, DTW and other methods[1-2]. Advantages of the proposed algorithms are that same performance as HMMs and other methods is achieved by simple computation and low cost, and can be implemented by automatic or manual segmentation. Also it can be combined with many types of automatic segmentations (HMM based, acoustic cues or feature based etc.), in which case frame averaging technique is implemented as a preprocessing phase.

As a future work, some specific applications of this technique and combinations with other techniques can be applied to improve the performance of speech segmentations.

Acknowledgment

The present research has been conducted by the research grant of Kwangwoon University in 2004.

References

1. Toledano, D.T., Gomez, L.A.H. and Grande, L.V., "Automatic Phonetic Segmentation" , Speech and Audio Processing, IEEE Transactions, 11, Issue 6, Nov. 2003, 617~625
2. Wesenick, M.-B and Kipp, A, "Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals" , Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference 1, 3-6 Oct. 1996, 129~132 vol.1
3. Kris Demuynck and Tom Laureys, A Comparison of Different Approaches to Automatic Speech Segmentation" , Text, Speech and Dialogue, 5th International Conference, TSD 2002, 277~284
4. Milone, D.H., Merelo, J.J. and Rufiner, H.L., "Evolutionary Algorithm for Speech Segmentation" , Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress, 2, 12-17 May 2002, 1115 ~1120
5. Bridle, J. and Sedgwick, N., "A Method for Segmenting Acoustic Patterns, with Applications to Automatic Speech Recognition" , Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77, 2, May 1977, 656~659
6. Svendsen, T. and Soong, F., "On the Automatic Segmentation of Speech Signals" , Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87, 12, Apr 1987, 77~80

[Profile]

•Byambajav.D

Byambajav.D received the B.S. degree in Physics from National University of Mongolia (NUM), in 1989 and the M.S. degree in Telecommunications from Asian Institute of Technology, Thailand, in 1997. From 1989 to 2002, he worked as a research engineer and lecturer at the NUM. Since 2003 he has been a doctoral student at the School of Electronic Engineering in the Kwangwoon University, Seoul.

•Chul-Ho Kang

Chul-Ho Kang received the B.S degree in electronics engineering from Hanyang Univ. in 1975. And he also received the M.S and Ph.D. degree in electronics engineering from Seoul National Univ. in 1979 and 1988, respectively. He worked at ADD as a research engineer for five years from 1977 to 1982. He is currently working as a professor since 1983 at the Department of Electronics and Communications Engineering, Kwangwoon University. His research interests include digital signal processing with application to the speech/speaker recognition and communication systems.