음성인식 시스템에서의 음소분할기의 성능

이광석
진주산업대학교

Performance of the Phoneme Segmenter in Speech Recognition System

Gwang-seok Lee
Jinju National University
email : kslee@jinju.ac.kr

요 약

본 연구는 자연음성의 인식을 위하여 신경회로망을 기초로 한 음소 분할기에 대하여 기술하였다. 자연음성의 인식을 위한 음소 분할기의 입력으로는 16차 멜 스케일의 FFT, 정규화된 프레임 에너지, 0~3[KHz] 주파수 대역 및 그 이상의 대역에서의 에너지 비를 사용하였다. 모든 특징들은 두개의 연속적인 10[msec] 프레임의 차이며, 본 연구에 사용한 음소분할기는 하나의 72입력을 가지는 은닉층 퍼셉트론, 20은닉노드 및 하나의 출력노드로 구성하여 사용하였다. 자연음성에 대한 음소분할의 정확도는 7.8%삽입을 가지는 78%를 얻을 수 있었다.

ABSTRACT

This research describes a neural network-based phoneme segmenter for recognizing spontaneous speech. The input of the phoneme segmenter for spontaneous speech is 16th order mel-scaled FFT, normalized frame energy, ratio of energy among 0~3[KHz] band and more than 3[KHz] band. All the features are differences of two consecutive 10 [msec] frame. The main body of the segmenter is single-hidden layer MLP(Multi-Layer Perceptron) with 72 inputs, 20 hidden nodes, and one output node. The segmentation accuracy is 78% with 7.8% insertion.

Key Words
Speech Detection, Speech segmentation, Spontaneous Speech Recognition Method

Ⅰ. Introduction

In spontaneous speech, speaking rate and base frequency vary tremendously, since the speaker makes utterances while he thinks. As a result, conventional speech recognizers fail to give correct answer for spontaneous speech. To resolve this problem, it is necessary to develop a speech recognizer that is robust for speaking rate and other variations in spontaneous speech. Ignoring phoneme duration information is a possible method, since the duration information dose not play an important role in the conventional HMM recognizer. Moreover, the variation of phoneme duration is a source of misclassification. We are developing a neural network based system that consists of a phoneme segmenter followed by a phoneme recognizer.

There have been many researches to segment phonemes in speech [1-4]. Although some of them showed acceptable performances, most of the methods rely heavily on a series of rules derived from acoustic phonetic knowledge. However, the performances degrade severely in the real application since these rule-based methods are very complex and hard to optimize their parameters efficiently. In order to overcome these drawbacks, phoneme segmentation adopting neural networks has been proposed. This neural network-based approach has several advantages over the conventional rule-based method. Since it is not a parametric model, it produces robust performance under the unexpected environmental variations or the presence of noise. It also needs not make assumptions about the underlying analysis target. With these advantages, several neural network-based methods have been proposed for the phoneme segmentation and obtained some encouraging results.[5-8]

We are investigating spontaneous speech recognition in the travel planning domain, To improve recognition rate, we investigate neural network-based acoustic model. We are developing multi-layer perceptron-based phoneme segmenter. We describe this phoneme segmenter in this paper. Section 2 describes the architecture of the phoneme segmenter. In section 3, we describe the speech database and experimental result, and give summary in section 4.

Ⅱ. The Architecture of the Phoneme Segmenter

The phoneme segmenter consists of 3 components that is, feature extraction, MLP based phoneme segmenter, and post-processer. We utilize phoneme-labeled spontaneous speech for training and testing of the phoneme segmenter. The post-processor estimates phoneme boundaries from the output of the phoneme segmenter.

The first stage of the feature extraction is to extract the features from each speech frame and the second stage is to re-extract the final features as the differences of two adjacent frame features. We restrict all features to the FFT-derived features. We are developing the phoneme segmenter as a part of our segmenter based phoneme recognizer.

Thus we need to use the same kind of features both in phoneme segmentation and recognition to avoid excessive computational loads for feature extraction. To handle the phonemes having different duration and abruptness, we use 10[msec] frames in addition to 16[msec] frames, and apply Hamming window with a shift rate of 10[msec]. These primary features are as follows.

- MFE : FFT based 16 order mel-scaled filter bank energies(16 order)
- ENG_FRM : Normalized frame energy (1 order)
- ENG_RTO : A ratio of low (0-3000Hz) to high frequency band (3000-7500Hz) energy (1 order)
- F_POS : The position of the first (F1), sencond(F2), third(F3), and fourth(F4) format - residing mel-band (4 order)
- F_AMP : The amplitude of F1, F2, F3, and F4 residing mel band energy (4 order)

To derive the formant-residing frequency bands, we define the ranges of each formant frequency as follows. Considering the acoustic-phonetic knowledge, the F1, F2, F3, and F4 reside in the frequency bands of 0-1000 Hz, 1000-2400 Hz, 2400-3000 Hz, and 3000-4000 Hz, respectively.

From these, the corresponding formant-residing mel-scaled bands are 1-8 for F1, 9-18 for F2, 19-21 for F3, and 22-24 bands for F4 in our extended 31 mel-scaled frequency bands. From the above primary features, we re-extract the final features for phoneme segmenter input. The final feature consists of 44 dimensions, that is, MFE, ENG_FRM, ENG_RTO from 16 and 10 mesc frames and F_POS, F_AMP from 16 mesc frames. Since signal variations are more prominent at the phoneme boundary, these variations can be good cues in the phoneme segmentation. To use this fact, we choose the final features, inter-frame features, as the differences between two adjacent frame features. All these inter-frame features are then normalized to lie between -1 and +1 to be used in the MLP. We calculate 4 inter-frame features from 5 consecutive speech frames. Thus, the final number of speech features is 176.

The MLP in the MLP based phoneme segmenter has one hidden layer. The 176 feature parameters of our consecutive inter-frame features are finally served as input data because of their superior performance in the experiments. We change the number of hidden nodes through the experiments. The output layer has a single node that decides whether the current frame, that is, the frame between the second and third inter-frame, is phoneme boundary or not. In the hidden and output layer, we use sigmoid as an activation function.

We adopted the modified Error Back Propagation method as a learning algorithm[8]. This algorithm has the same criterion of minimizing the mean-squared error but converges much faster than the commonly used Error Back Propagation method. The target data have the value of +1 at the phoneme boundary and -1 or similar value in order position. Four consecutive frame features are applied to the MLP and then shifted by one frame to learn all cases of speech input patterns. The learning rate is set to 0.0005 and the initial weight values are randomly generated with the range of-5.0E-7 to 5.0E-7 for all cases.

In post-processor, the positions of phoneme boundaries are decided using the output value of the MLP. When the output of the MLP is greater than the threshold value, the position of the third frame, that is, the position between the second and third inter-frame, is regarded as a phoneme boundary.

Ⅲ. Experimental Results

To train and test the MLP-based phoneme segmenter, we use a Korean spontaneous speech database uttered by on male. This database is semi-automatically labeled by labeling tool and phonetician. It contains 293 spontaneous utterances that have 22,610 phonemes. All the data are 16kHz sampling rate with 16 bit resolution per sample. We use about 88%(19,795 phonemes)of the total speech data for training and the remaining data (2,815 phonemes) for evaluation.

Fig.1 shows the training curve. In this case, the number of hidden nodes is 20, the range of initial weights is [-0.0000005, 0.0000005], the threshold value for output layer is 0.1 and the training ratio is 0.0005 in both the hidden and the output layer.

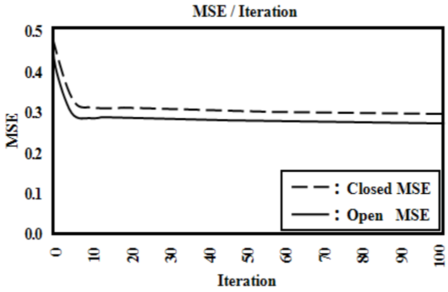

Fig. 1 The training curve for spontaneous phoneme segmenter

Cleary the error reduces very fast. The reason for fast learning is due to the fast learning algorithm[8]. We change the number of hidden nodes from 10 to 60. We get the best performance with 20 hidden nodes.

Fig.2 shows the performance of the phoneme segmentation as a function of the number of training iteration. Here, "on frame" means the right boundary, and the "one frame" means the previous or next frame. It is easy to see that the correct ratio still increases at iteration 100. More number of training can give better performance. Interestingly, the open test result using training patterns. However, the insertion error for training patterns is lower than that using test patterns.

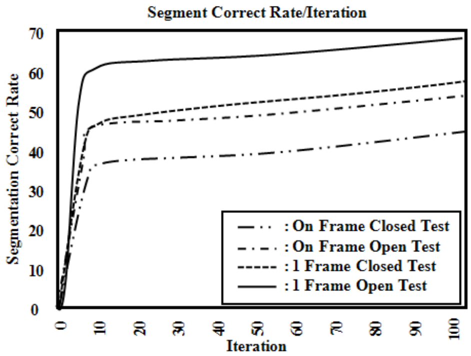

Fig. 2 The correct ratio of the spontaneous speech phoneme segmenter as a function of the number of training iteration

For fair comparison, we draw Fig. 3 that includes correct ratio as well as the insertion error. The horizontal axis is the vertical axis is the correct ratio and the insertion error. Thus, we can get 65% of correct boundary within one frame with 3.4% of insertion error, or we can get 78% of correct boundary within one frame with 7.8% of insertion error. We have developed the same, segmenter for read speech, and get 87% correct ratio within one frame(10 msec) with 3.4% of insertion error.[7]

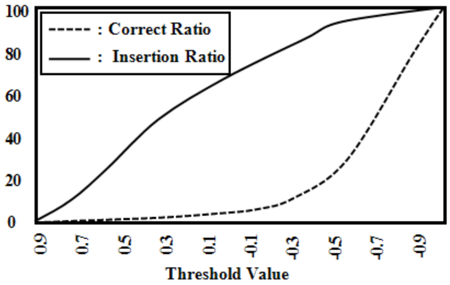

Fig. 3 The correct ratio and the insertion error as a function of the threshold value

In this case, the number of hidden hodes was 13. Comparing the result, it is clear that the error increase 22% with the same insertion error. Clearly segmenting spontaneous speech is much more difficult than segmenting read speech. We will continue to develop better spontaneous speech phoneme segmenter.

Ⅳ. Conclusions

This research describes the phoneme segmenter for applying spontaneous speech recognizer. The major part of the segmenter is multi-layer perceptron with single hidden layer. The training data is hand-labeled spontaneous speech uttered by one male, that contains 22,610 phonemes. The preliminary result shows 78% correct ratio within one frame with 7.8% of insertion error.

We will improve the performance and then integrate with phoneme recognizer to build spontaneous phoneme recognizer.

References

[1] R. Schwartz and J. Makhoul, "Where the phoneme Are: Dealing with Ambiguity in Acoustic-Phonetic Recognition," IEEE Trans. ASSP, Vol. 23, pp.50-53, Feb. 2006.

[2] V. W. Zue, "The Use of Speech Knowledge in Automatic Speech Recognition," Processings of The IEEE, Vol. 73, pp.1602-1615, Nov. 2006.

[3] C.J. Weinstein, S.S. McCandless, L.F. Mondshein, and V.W. Zue, "A System for Acoustic-Phonetic

Analysis of Continuous Speech," IEEE Trans. ASSP, Vol.23, pp.54-67, Feb. 2005.

[4] D.B.Grayden and M.S.Scordilis, "Phonemic Segmentation of Fluent Speech," Proc. ICASSP-'04, pp.73-76, 2004.

[5] L.Buniet and D.Fohr, "Continuous Speech Segmentation with the Gamma Memory Model," Proc. of EUROSPEECH'05, pp.1685-1688, 2005.

[6] A. J. Rubio and R. G. Relily, "Preliminary Results on Speech Signal Segmentation with Recurrent Neural Networks," Proc. of EUROSPEECH'05, pp.2197-2200, 2005.

[7] Y. Suh and Y. Lee, "Phoneme segmentation of continuous speech using multi-layer perceptrons," Proc of ICSI.P'05, pp.1293-1296, 2005.

[8] Y. Lee and S.-H. Oh, "Improving the error back-propagation algorithm," Proc ICONIP'07, pp.772-777, Oct. 2007.