SPEECH/NON-SPEECH CLASSIFICATION USING MULTIPLE FEATURES FOR ROBUST ENDPOINT DETECTION

Won-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee and Jong-Seok Lee

Information Technology Lab.
LG Corporate Institute of Technology
{whshin,lbs,yklee,ljs}@lgcit.com

ABSTRACT

In this paper, we describe a new speech/non-speech classification method that improves the endpoint detection performance for speech recognition in noisy environments. The proposed method uses multiple features to increase the robustness in noisy environments, and the classification and regression tree(CART) technique is applied to effectively combine these multiple features for classification of each frame.

We evaluate the performance of the proposed method by conducting speech/non-speech classification experiments on noisy speech. We also investigate the importance of various features on speech/non-speech classification in noisy environments

In particular, the proposed method is applies to the endpoint detection algorithm for isolated speech recognition of voice-dialing cellular phone. We simulate the speech recognition experiments in various noise environments, and the effects of proposed method on speech recognition performance are evaluated.

1. INTRODUCTION

Currently the research on continuous speech recognition is more predominantly carried out than other areas, such as isolated word recognition, in the speech recognition community. However, there are still numerous markets which require the application of the isolated word recognition systems, e.g., in computer telephony integration(CTI) and other voice controlled products. One of these examples is voice dialing in cellular phones. In this application, one major concern is to overcome the degradation of performance due to the background noise.

A major cause of errors in isolated word speech recognition systems is the inaccurate detection of the beginning and ending boundaries of test and reference patterns[1]. Especially, the performance of endpoint detection severely degrades in noisy environments. Many endpoint detection algorithms have been proposed which are based on features of short-time energy and zero-crossing rate[2]. However, these features do not work well in noisy environments.

To improve the speech recognition performance in noisy environments, there have been many studies on new features[3],[4] for robust endpoint detection, such as linear prediction error energy, pitch[5], and band energy, etc. However, each feature does not always guarantee the reliable endpoint detection in the various kinds of environments. Therefore, it is necessary to use these parameters together to improve the robustness in various noisy environments. This paper focuses on using the multiple features effectively in endpoint detection.

Since it is very difficult to design the rules that combine multiple features efficiently, statistical approach is considered. We used the CART[6] algorithm for speech/non-speech classification of each frame using multiple features.

We performed speech/non-speech classification experiments on various noisy speech. We also investigated the usefulness of various features for speech/non-speech classification in noisy environments. We applied the proposed method on speech recognition algorithm of voice dialing cellular phone and its performance is evaluated.

This paper is organized as follows. In Section 2, a review on the the general structure of endpoint detection algorithm is presented. In Section 3, we explain the proposed method of speech/non-speech classification using multiple features. The experimental results are shown in section 4 and the conclusion is given in Section 5.

2. STRUCTURE OF ENDPOINT DETECTION ALGORITHM

Fig. 1 shows a block diagram of a conventional endpoint detection process. Input speech signals are segmented into frames using window function and feature parameters such as

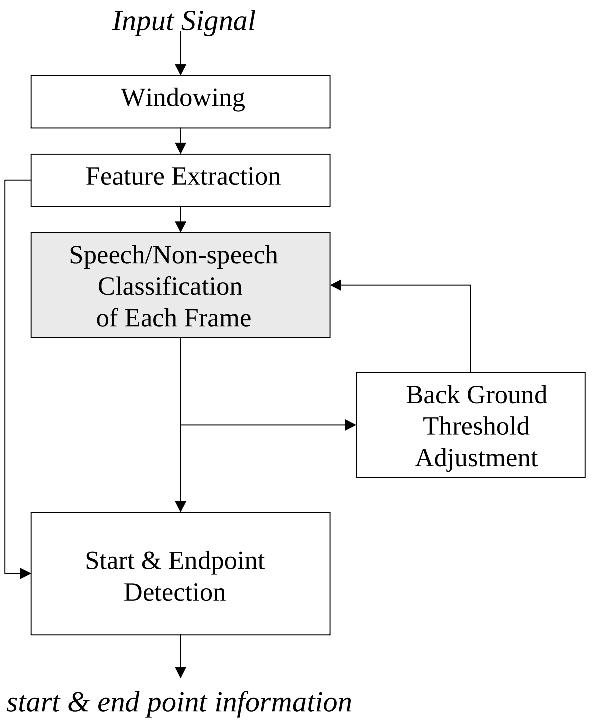
Input Signal
Windowing
Feature Extraction
Speech/Non-speech Classification of Each Frame
Back Ground Threshold Adjustment
Start & Endpoint Detection
start & end point information

Figure 1. The block diagram of general endpoint detection algorithm.

energy, and zero-crossing rates are extracted. During the non-speech period of several mili-seconds, a few background thresholds are calculated, which are adjusted continuously using the following non-speech frames. Each frame is classified into speech or non-speech using the background thresholds. Endpoint detection logic extracts the start and end point information using the speech/non-speech class information of each frame, the number of successive frames of each class, and the parameter values.

In this paper we propose a new method to improve the performance of 'Speech/Non-Speech Classification of Each Frame' part of fig. 1.

3. SPEECH/NON-SPEECH CLASSIFICATION USING MULTIPLE FEATURES

3.1 CART

Tree-based modeling, which is a non-parametric statistical modeling, successively divides the region of feature vector *X* to minimize the prediction error. The CART technique[6], a special case of tree-based modeling, produces classification or regression trees with the type of predicted variable *Y*. The process of training CART proceeds by partitioning the root and the following subtrees. The partitioning process is applied recursively to each descendant, continuing until all leaves of the resulting tree have pre-determined number of entries or impurity.

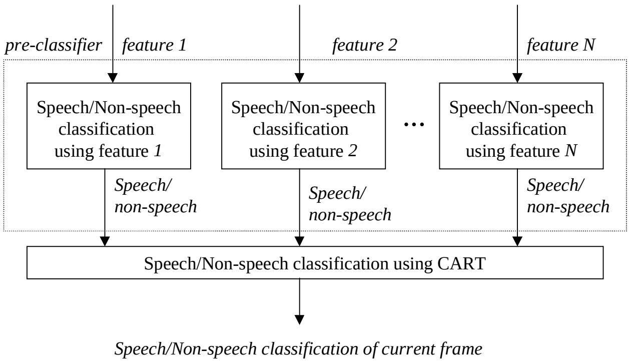
pre-classifier
feature 1
feature 2
feature N
Speech/Non-speech classification using feature 1
Speech/Non-speech classification using feature 2
...
Speech/Non-speech classification using feature N
Speech/ non-speech
Speech/ non-speech
Speech/ non-speech
Speech/Non-speech classification using CART
Speech/Non-speech classification of current frame

Figure 2. The block diagram of the proposed speech/non-speech classification method (This figure replaces the grey-colored block of Figure 1)

The classification process is as follows. Staring at the root node, the binary function is used to test the given input *X*. If the result is '0', the left branch is followed: if the result is '1', the right branch is followed. The process is repeated until a leaf is reached, at which point the associated label *Y* is the output.

3.2 Speech/Non-speech Classification Using CART

To consistently extract islands of reliability, even in very noisy conditions, the proposed method uses the following multiple features for speech/non-speech classification. The features are full-band energy, band energy of audible frequency range(300-3700Hz) and higher frequency range(2-4kHz), peakyness, LPC-residual energy, and noise-filtered energy. The full-band and audible frequency energy are conventionally frequently. We use higher frequency range band energy to detect the consonants more accurately. However we don't use lower frequency band energy, because some noises such as car noise are centered on lower bands. The peakyness is useful for detection of the voiced parts. LPC residual energy is robust to low-band noise such as car noise. Finally the noise-filtered energy is used to remove the influence of the background noise. Input signal is filtered by Wiener filter which is modeled using the non-speech part in the beginning of the input signal. Zero crossing ratio is not used, because it is not immunized against noise.

Fig. 2 shows the structure of the proposed speech/non-speech classification process. There are multiple pre-classifiers using each feature independently. Each pre-classifier generates speech/non-speech class information for every frame. Since some features make good decision and some do not for each frame, it is necessary to design a rule to make a final decision using the multiple outputs of pre-classifiers. CART is a convenient tool for designing decision logic, since it does not need a priori knowledge of the input features. Using the speech/non-speech class information from multiple pre-classifiers, CART makes a final decision whether the current frame is a speech or non-speech.

4. EXPERIMENTAL RESULTS

In this experiment, we evaluated the performance of the proposed speech/non-speech classification method. We also performed speaker dependent isolated word recognition experiments in noisy environments to demonstrate the effects of the proposed method on speech recognition performance.

For speech recognition experiments, eight people pronounced twenty words, once for reference and three times for test. We also recorded four kinds of noise signals: car, subway station, train, and street. To construct the test database, we artificially generated noisy speech of 5, 0, -5dB by mixing the noise signals with speech signals. The total number of tokens of the test database is 5760. Every token has sufficient non-speech period at the start and end of utterance.

For training the CART, another database is prepared which has the same organization as the test database, and the start and end points are manually labeled for training.

The CART is trained using the input parameters which consist of class information obtained from pre-classifiers and the target class information. We trained the CART in two different modes: one uses class information of current frame only and the other uses those of current frame, previous two frames and the next two frames.

We evaluated speech/non-speech classification performance of the proposed method. For the comparison, we also evaluated the classification performance using each feature separately. Speech/non-speech classification rate was calculated as follows:

$$\text{speech/nonspeech classification rate} = \frac{\text{number of frames coincide with manual classification}}{\text{number of total frames}}$$

Table 1 shows that the proposed methods using multiple features performed better than using a single feature by about 4 to 10 %. We can also find that it is more effective to use the information of the previous and the next frames together. The improvements in performance demonstrate that the CART performs well in combining multiple features for decision of speech or non-speech

Table 1. Speech/non-speech classification rates

	Speech/non-speech Classification Rate(%)		
	5dB	0dB	-5dB
I	84.81	80.30	73.93
II	86.77	82.50	76.36
III	81.69	77.27	72.70
VI	88.85	84.07	76.80
V	84.35	79.52	74.13
VI	87.20	81.81	75.55
VII	89.45	84.97	78.21
VIII	91.11	87.73	81.16

Used features and method, I: full-band energy, II: band energy of audible frequency range(300-3700Hz), III: band energy of higher frequency range(2~4kHz), VI: peakyness, V: LPC residual energy, VI: noise-filtered energy, VII: the proposed method (current frame only), VIII: the proposed method (including previous and next frames)

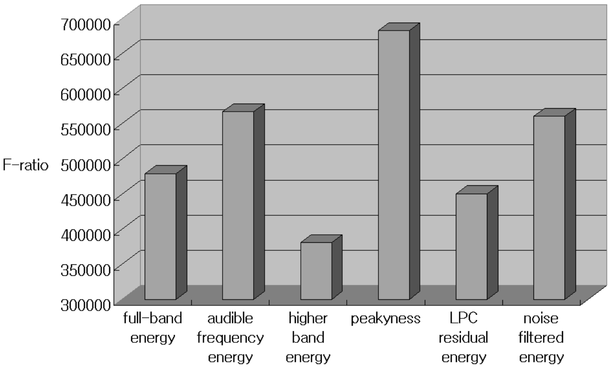

Figure 3. The effectiveness of various features for speech/non-speech classification in noisy environments.

Table 1 also shows the effectiveness of each feature on the speech/non-speech classification in various noise environments. The results show that the peakyness, the energy of audible frequency range, and the noise-filtered energy performed better than other features. The full band energy, the conventional feature for endpoint detection for clean speech, does not work well for noisy speech.

Table 2. Recognition result using the proposed method

Recognition Rate(%) / Used Feature	SNR		
	5dB	0dB	-5dB
Energy	98.80	93.28	69.21
The proposed method I	99.38	95.10	70.46
The proposed method II	99.58	96.40	74.84

The proposed method I: use current frame only, The proposed method II: including previous and next frames

And the performance of the noise-filtered energy does not reach our expectation, since the noise spectral characteristics change in time. We computed F-ratios[7] of the features to evaluate their effectiveness, witch compares the inter- and intra-cluster variances. Fig. 3 shows similar results as those of Table 1.

We also evaluate the effects of the proposed method on speech recognition. First, we performed the speaker dependent isolated word recognition experiments with conventional endpoint detection algorithm using full band energy. The matching process uses DTW algorithm. Then, we substituted the speech/non-speech classification part with the proposed one and compared the performances. Table 2 shows that the proposed method improves the speech recognition performance. Especially in low SNR environments, the recognition rate increases by about 5%.

5. CONCLUSION

In this paper, we proposed robust speech/non-speech classification method in noisy environments. The proposed method uses CART technique to combine the multiple features effectively.

We evaluated the performance of the proposed method by speech/non-speech classification experiments. The results showed that the proposed method using multiple features performed better than using a single feature by 4 to 10 %. The CART works well in combining multiple features for decision of speech or non-speech. We also investigated the effectiveness of various features on speech/non-speech classification in noisy environments.

We applied the proposed method to the isolated speech recognition algorithm of voice dialing cellular phone. The simulation results showed that the proposed method increases the recognition rate by about 5% in low SNR environments.

6. REFERENCES

[1] J-C. Junqua, “Robustness and cooperative multimodel man-machine communication application,” in *Proc. Second Venaco Workshop and ESCA ETRW*, Sep. 1991.

[2] M. H. Savoji, “A robust algorithm for accurate endpointing of speech,” *Speech Commin.*, vol.8, pp. 45-60, 1989.

[3] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, “Endpoint detecion of isolated utterances based on a modified teager energy measure,” In *Proc. IEEE ICASSP*, vol. 2, pp. 732-735, 1993.

[4] H. Kobetake, K. Tawa, and A. Ishida, “Speech/nonspeech discrimination for speech recognition system under real life noise environments,” In *Proc. IEEE ICASSP*, pp. 365-368, 1989.

[5] M. Hamada, Y. Takizawa, and T. Norimatsu, “A noise robust speech recognition” in *Proc. ICSLP*, pp. 893-896, 1990.

[6] P. A. Chou, “Optimal Partioning for Classification and Regression Trees,” *IEEE Trans. PAMI*, vol. 13. No. 4, pp. 340-354, 1991.

[7] J. Wolf, “Efficient acoustic parameters for speaker recognition,” *J. Acoust. Soc. Am.*, 51, pp. 2044-2056, 1972.