Different Filler Models for Keyword Recognizer

Mansoo Park, Hoirin Kim
School of Engineering, ICU
E-mail : mansoo@icu.ac.kr

Abstract

Our goal of this work is to find the adequate acoustic modeling for the keyword spotting recognizer. We use three types of training data composed of only word DB, both sentence DB and word DB, or only sentence DB for modeling the non-keyword. To enhance the discrimination between the keyword model and non-keyword model, we trained the keyword models with context-dependent tri-phone unit and the filler models with context-independent mono-phone unit. Also, we compared of the performance according to the variation of the number of mixtures in model states.

1. Introduction

In a keyword spotting (KWS) system, the keywords in a pre-defined vocabulary must be detected if they are spoken in the input sentence. So, the speech to be recognized can be classified into words that belong or do not belong to the pre-defined vocabulary. Several approaches have been proposed in order to model the words outside the vocabulary. For example, there are filler or garbage models. In this work, we use filler model for the non-keyword model.

As for performance evaluation, we use the following measures[1].

- In the case of input speech including the keyword
  (1) CA : Correct Acceptance rate for keyword
  (2) FAK : False Acceptance rate for Keyword
  (3) FR : False Rejection for rate keyword

- In the case of input speech not including the keyword
  (1) CR : Correct Rejection rate for out-of-vocabulary
  (2) FAO : False Acceptance rate for out-of-vocabulary

2. KWS Recognizer Structure

2.1 KWS system

Figure 1 shows the block diagram of KWS system based on continuous HMM[2]. Feature vectors are extracted after detecting the speech segment using the endpoint detector when the speech is inputted. And then, it recognizes the keyword using the Viterbi algorithm[2].

2.2 Network structure

Figure 2 shows the recognition network involving the keyword models and the filler models. The filler model set used in the recognition system consists of 47 mono-phone units involving “silence”, and the keyword model set consists of 23 tri-phone units.

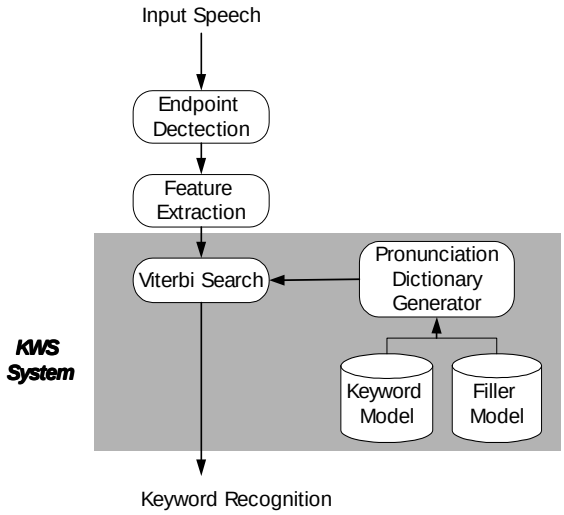

Figure 1. Block diagram of KWS system

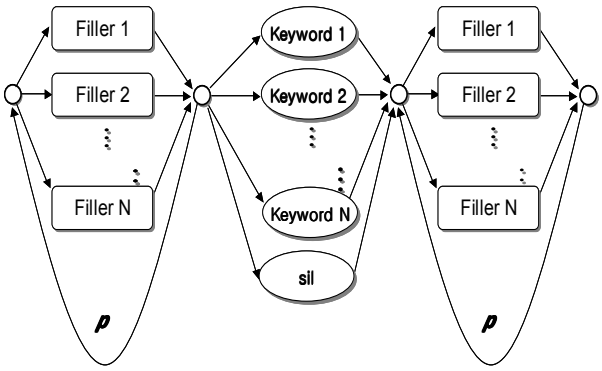

Figure 2. KWS network

Here, we use only 3 keywords that are "컴퓨터", "보이스포탈", and "음성포탈". At the part of keyword network, the model "sil" is inserted to process the input speech not including any keyword. At the part of filler network, we use the word penalty, p, for reducing deletion error.

3. Experiments and Results

3.1 Baseline system

For the performance evaluation of our task in the KWS system, the following databases are used.

- PBW452 DB : consisted of 452 isolated Korean words which are spoken two times by 72 speakers. The speech signal is sampled at 16 kHz and quantized with 16 bits. We have down-sampled it to 8 kHz to cope with telecommunication channel.
- Speech Recognition DB of KT Multimedia Center : consisted of 800 Korean sentences that are divided into 8 set, spoken by 100 speakers, and recorded through 3 stand MICs located with 3 different distances(20cm, 50cm, 100cm), Laptop computer, and Headset. The speech signal is sampled at 16 kHz and quantized with 16 bits. We have down-sampled it to 8 kHz.

For modeling the filler models, we use about 90 % of PBW452 DB and 200 types of sentence DB that are extracted in Speech Recognition DB of KT Multimedia Center. The sentence DB is concerned with railroad service and stock service. For modeling the keyword models, we use 3 kinds of keyword DB that are extracted in Speech Recognition DB of KT Multimedia Center. The test database consists of Speech Recognition DB of KT Multimedia Center not involved in the training database.

At first, the input speech is pre-emphasized using the first-order FIR filter with a coefficient of 0.97. The samples are blocked into overlapping frames of 16ms and each frame is shifted at the rate of 8ms. Each frame is windowed with a Hamming window. Every frame is characterized by total 39^{th} order feature vectors which are 12 mel frequency cepstral coefficients (MFCC), their first-order temporal regression coefficients (ΔMFCC), their second-order temporal regression coefficients (ΔΔMFCC), and the log-energy and its first- and second-order temporal regression coefficients. In order to remove the recording condition dependencies, for each feature vector we have performed the cepstral mean subtraction. And, each phone is modeled to a three state left-to-right continuous density HMM.

3.2 Performance comparison according to the different training DB

In this section we compare of the performance according to the different training DB for modeling the filler. We trained the filler models using only PBW452 DB, both PBW452 DB and sentence DB, and only sentence DB. To enhance the discrimination between the keyword model and non-keyword model, we trained the keyword models with context-dependent tri-phone units and the filler models with context-independent mono-phone units. In the above acoustic modeling, mono-phone models trained by PBW452 DB were used as initial models.

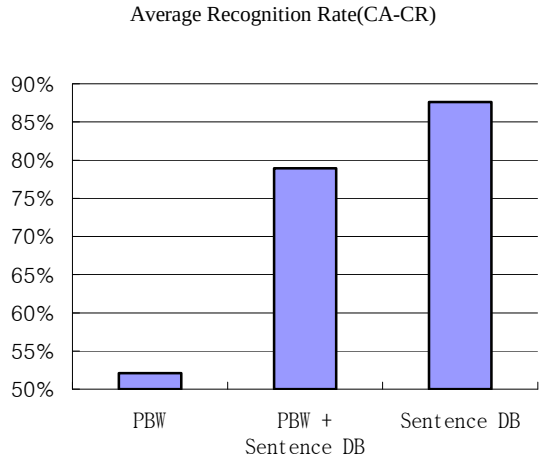

Figure 3. Recognition performance according to the type of training DB

Figure 3 shows the arithmetic average recognition rate of CA and CR rates for different training DB in modeling the fillers when both the keyword and filler models have 5 Gaussian mixtures. As shown in the figure, we can conclude that the filler should be trained only with the sentence DB since the DB reflects more well the characteristics of garbage sentence in test speech.

3.3 Performance evaluation according to the number of Gaussian mixture.

Here, we trained the filler models using only the sentence DB from the previous results. As previously described, the keyword models are based on context-dependent tri-phone units and the filler models are based on context-independent mono-phone units. Now, we investigate on the performance variation according to increase of the number of mixture in each state of acoustic models.

Table 1 to 4 show the performances according to variation in the number of Gaussian mixture in keyword and filler models. As the number of mixture increases, the performances are substantially improved except Table 1. And, FAK and FAO rates are relatively high.

Table 1. Recognition performance when the filler model has only one mixture (%)

Measure / Keyword	CA	FAK	FR	CR	FAO
Mixture1	92.03	7.66	0.31	71.64	28.36
Mixture3	93.73	6.19	0.08	64.72	35.28
Mixture5	94.01	5.91	0.08	64.47	35.53
Mixture7	94.09	5.83	0.08	63.26	36.74

Table 2. Recognition performance when the filler model has 3 mixtures (%)

Measure / Keyword	CA	FAK	FR	CR	FAO
Mixture1	91.48	7.14	1.38	77.41	22.59
Mixture3	93.98	5.68	0.34	78.59	21.41
Mixture5	94.52	5.17	0.31	79.21	20.79
Mixture7	94.52	5.17	0.31	79.03	20.97

Table 3. Recognition performance when the filler model has 5 mixtures (%)

Measure / Keyword	CA	FAK	FR	CR	FAO
Mixture1	91.85	6.74	1.41	79.41	20.59
Mixture3	94.15	5.51	0.34	79.18	20.82
Mixture5	94.67	4.99	0.34	80.55	19.45
Mixture7	94.70	4.99	0.31	80.11	19.89

Table 4. Recognition performance when the filler model has 7 mixtures (%)

Measure / Keyword	CA	FAK	FR	CR	FAO
Mixture1	91.59	6.72	1.69	79.50	20.50
Mixture3	94.18	5.51	0.31	79.91	20.09
Mixture5	94.52	5.16	0.32	80.42	19.58
Mixture7	94.65	5.01	0.34	80.27	19.73

Figure 4 to 5 show the variation of the average recognition rate for different number of mixtures in keyword and filler models. As the results show, the best performance was obtained at mixture 5 for both keyword model and filler model.

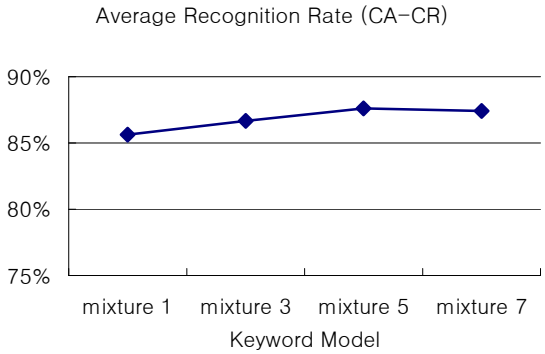

Figure 4. Average recognition rate according to the number of keyword mixture (Filler : 5 mixtures)

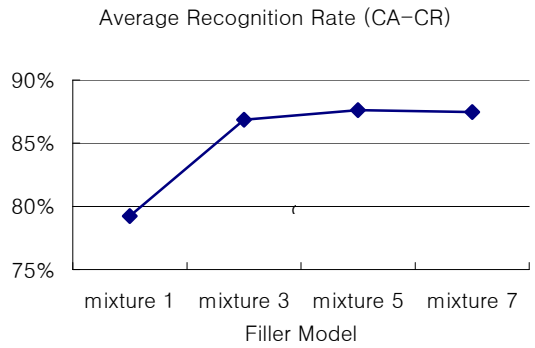

Figure 5. Average recognition rate according to the number of filler mixture (Keyword : 5 mixtures)

4. Conclusion

The experimental results show that the performance of KWS recognizer is quiet different according to the type of training DB and the number of Gaussian mixture. As shown in the above experiments, for improving the performance of KWS recognition, we should choose sentence DB for training the filler models since the DB involves a various and sufficient vocabulary, and the characteristics of training DB are consistent with the test DB. In the future work, we will try to change KWS network for improving the performance, and study on a post-processing method to reduce the FAK and FAO rates.

References

1. Ki-Tae Kim, Kwang-Sik Moon, Hoi-Rin Kim, Young-Jik Lee, Jae-Ho Chung, “Performance Comparison of Out-Of-Vocabulary Word Rejection Algorithms in Variable Vocabulary Word Recognition,” Jour. Of ASK, Vol. 20, No. 2, pp. 27-34, 2001.
2. Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition,” Prentice-Hall International, Inc. 1993.