KEYWORD SPOTTING IN BROADCAST NEWS

Yung-Hwan Oh, Jeong-Sik Park and Kyung-Mi Park

Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology, Daejeon, Korea
Phone:+82-42-869-3516, Fax:+82-42-869-3510
Email:{yhoh, dionpark, kmpark}@speech.kaist.ac.kr

ABSTRACT

In this presentation, we introduce several research topics related with the keyword spotting system on broadcast news. The system searches the keyword speech in online broadcast news and extracts the articles including the keyword(s). To obtain a stable recognition performance, the system uses several speech processing techniques such as utterance verification, out-of-vocabulary rejection, audio classification, and noise reduction. We will show an overall configuration for our system and report technical advances as well as some experimental results. We also describe future works for the system improvements and the application for further complicate spoken documents like interviews or movies.

1. INTRODUCTION

The keyword spotting, which searches the keyword(s) from continuous speech, has been advanced as an essential field of speech technologies. Since the keyword spotting system gives more reliable performance than general speech recognition system, which recognizes overall speech data, it has been widely used in various applications [1].

One of principal applications is the multimedia data retrieval. In the internet and mobile era, fairly amount of users consume extensive multimedia data and even produce their own contents. As such, data retrieval technology, which collects just the information they want correctly and fast, became an essential research topic. By the way, it is more difficult to retrieve and manage spoken documents than written documents because speech data must be translated into text data prior to the retrieval. Since there are few multimedia data containing text information, keyword spotting plays a major role in the spoken data retrieval.

This paper concentrates on the keyword spotting in the broadcast news. Since the news data consists of speech correctly pronounced by newscasters or reporters, it gives better recognition results than other kinds of speech data such as dialogs. In this reason, broadcast news is mainly applied for continuous speech recognition [2].

Moreover, a great demand for topic selection requires the keyword spotting technology. Increasingly numbers of people expect to watch just the articles in which they have a great interest. Searching the interesting articles by topics or keywords can be realized from keyword spotting system [3].

This paper presents an overall configuration for keyword spotting system on broadcast news and reports technical advances. Section 2 introduces our system, and section 3 reports several techniques for system improvements. Finally, section 4 and 5 present experimental results and conclusions, respectively.

2. KEYWORD SPOTTING SYSTEM ON BROADCAST NEWS

Figure 1 represents the configuration of keyword spotting system on broadcast news. The simple structure of keyword spotting system consists of feature extraction and search module. In short, feature parameters are extracted from the broadcast news and applied to keywords and garbage HMM (Hidden Markov Model), which are pre-trained speech models. Once the output probability of each HMM for an input utterance is estimated, the decision is made whether to reject or accept the utterance.

There are several kinds of search methods, which are based on the LVCSR (Large Vocabulary Continuous Speech Recognition), the phoneme

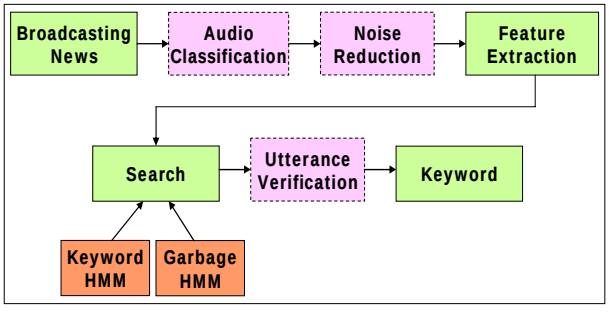

Fig. 1 Configuration of Keyword Spotting System on Broadcast News

recognition, and the whole-word model. Prior to a search step, the LVCSR and phoneme recognition approaches produce text scripts for overall input speech, using word and phoneme-level transcriptions, respectively. As such, the LVCSR recognizer requires tens of hours of word-level transcriptions while preserving good performance. Meanwhile, the phoneme recognizer requires very less hardware resources than the LVCSR system but gives low performance. Accordingly, we use the whole-word model based search process, which takes advantage of above two systems.

The search process based on phoneme recognizer only requires several word HMMs for each keyword and garbage, i.e., non-keywords. The keyword HMM helps to detect the keyword while the garbage HMM is used to reject non-keyword regions. Consequently, the system performance mainly depends on training two kinds of HMMs accurately. We use the real data extracted from the broadcast news for HMM training, and make connections between garbage, keyword models and a silence model.

Although the structure is simple and fast, two kinds of detection errors may be more frequently occurred than other approaches, thus degrading the system performance. One is the false alarm, which means incorrectly accepted data, and the other error comes from incorrectly rejected data, which is called the false reject.

The errors result from several characteristics of broadcast news data. First, news data include various types of audio streams. In contrast to the speech data of news reporters, background music or commercials may cause more detection errors. Second, background noises included during the outdoor reporting degrade the speech quality, thus resulting in recognition errors. Finally, various speakers such as male/female newscasters or interviewees are frequently changed and may cause recognition error.

To preserve the keyword spotting system from the unexpected errors, we employ several speech processing techniques, presented as dotted line in Figure 1. Audio classification and noise reduction are expected to solve the first and second problem, respectively. In addition, utterance verification can be essentially employed for reduction of false alarm and false reject.

3. TECHNIQUES FOR SYSTEM IMPROVEMENT

3.1 Utterance verification

Utterance verification decides whether the recognition result is accepted or rejected depending on a decision criterion called the confidence measure [4]. Since the results detected as the keyword(s) contain a lot of false alarm and false reject, utterance verification is essentially applied to post-processing of keyword spotting system.

The most commonly used confidence measure is based on N-best and likelihood, i.e. output probability, which is defied as

$$CM = \frac{L_1}{\sum_{i=1}^{N} L_i} \quad (1)$$

$$CM = L_1 - \sum_{i=1}^{N} \frac{L_i}{N} \quad (2)$$

where L_i is the likelihood of i-th result in N-best list [5]. N-best and likelihood means the N hypotheses, or recognition results, scored for an utterance. CM calculated in equation (1) or (2) is compared with empirically determined threshold. If the value is greater than the threshold, the utterance is accepted as a keyword; otherwise, it is rejected.

Though N-best and likelihood based confidence measure gives good performance with low computation, it has some limitations when applied to whole-word model based keyword spotting. First, the N-best result is not reliable since the system depends on just a few word models. Second, the constant threshold deciding acceptance or rejection is not appropriate for rapidly varied utterances of news data.

For an advanced utterance verifier, we measure a distance between the detected region and the real data based on the DTW (Dynamic Time Warping) algorithm. The DTW, which finds the optimal path from start to end for an input utterance, gives superior performance on the isolated word recognition system [6]. The simple and effective pattern matching algorithm requires less data (reference patterns) and lower computation than HMM-based technique. Accordingly, it is further applicable to the verification of the detected utterance.

Figure 2 explains the DTW based utterance verification. Two input patterns, i.e., a detected region as keyword (test pattern) and real keyword data (reference pattern), are warped dynamically and the distance between them is estimated. To get a correct distance, we move the detected region from former 5 frames to rear 5 frames and obtain several test

patterns. For example, if the detected region is frame 'A' to 'B', we use several adjacent regions ((A-5,B-5),…,(A+5,B+5)) as test patterns. Next, each pattern is applied to the warping algorithm and the distance from each reference pattern is calculated. Reference patterns consist of keywords and non-keywords

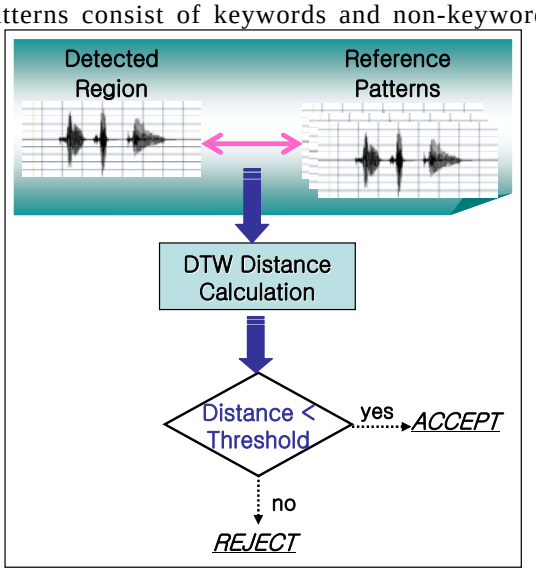

Fig. 2 DTW-based Utterance Verification

(garbage) obtained from real news data. Finally, the minimum distance is compared with the predetermined threshold, and then acceptance or rejection is decided.

3.2 Speech and audio classification

The goal of speech and audio classification is to partition and label an input audio stream into speech, music, commercials, or other acoustic types [7]. This preliminary process is a challenging research topic in ASR and spoken document retrieval, and necessary for keyword spotting in broadcast news which contains various audio types. Most of audio classification techniques focus on two aspects: the particular feature and the statistical model. Feature-based classification is derived from different distribution characteristics between speech and non-speech segments in both the time and frequency domains. This method depends mainly on the discriminative features and are implemented either in a complex threshold-dependent scheme or with some pattern classification method (Euclidean distance, nearest neighbor, nearest feature line, etc.). Model-based methods make a specific model (Gaussian mixture model (GMM), multi-layer perceptron (MLP), etc.) for speech, speech with music backgrounds, and only music. By the way, features used in each method are quite different. As such, most researchers treat the two kinds of methods separately and do not consider them in an integrated manner.

Real-time keyword spotting in broadcast news need to reduce the time consumption as much as possible. Accordingly, this paper investigates the feature-based approach rather than model-based since the feature extraction requires less computation capacity than the model construction. The features are represented in the time domain (zero-crossing rate (ZCR), energy, etc.), or in the frequency domain (sub-band power, low short-time energy ratio, etc.). We consider a technique combining some advanced features including VSF and VZCR.

SF (Spectrum Flux) is the ordinary Euclidean norm of the delta spectrum magnitude, which is calculated as

$$SF = \| S_i - S_{i-1} \|_2 = \frac{1}{N}\left(\sum_{k=0}^{N-1}(S_i(k) - S_{i-1}(k))^2\right)^{\frac{1}{2}} \quad (3)$$

where S_i is the spectrum magnitude vector of frame i and N is the size of the window [8]. Music or environmental sounds are periodic or monotonic and have more constant rates of changes than speech. Consequently, the variance of spectrum flux (VSF) of speech is larger than that of music or environmental sounds.

VZCR (Variance of Zero Crossing Rate) is based on the ZCR as follows.

$$Z_s(m) = \frac{1}{N}\sum_{n=m-N+1}^{m}\frac{|\mathrm{sgn}\{s(n)\} - \mathrm{sgn}\{s(n-1)\}|}{2} w(m-n) \quad (4)$$

where N is the length of the frame, m is the endpoint of the frame, and $w(n)$ is the window function [9]. Since music and environmental sounds are more periodic than speech, their ZCR will be more constant with fewer fluctuations. This denotes that the variance of ZCR of speech is larger than that of music or environmental sounds.

For the improvement of classification performance, we combine the time domain features (energy, VZCR) and frequency domain features (sub-band power, VSF). Since each feature is independent and uncorrelated, distribution-based combining technique is not appropriate. The simplest method is making all features for each audio block as a vector and classifying vectors into a few audio groups. In this paper, we use the SVM-based approach for vector clustering.

SVM (Support Vector Machine) is a classifier that estimates decision surfaces directly rather than modeling a probability distribution across training data. This classifier has demonstrated good

performance on several pattern recognition tasks [10].

The feature vectors composed by outputs of various feature based classifications can be input into the support vector machine to decide whether each audio block presents speech or non-speech region. Considering the problem of typical two-class classification, the training set $\{(x_1, y_1),...,(x_s, y_s)\}$ where $x_i \in R^n$ is a feature vector and $y_i \in \{+1,-1\}$ is a class label ($i = 1,2,...,s$), are separated by a hyper-plane of equation $w \cdot x + b = 0$. The optimal hyper-plane separates two classes, thus maximizing the margin, i.e., the distance from the nearest samples to the classification boundary. We apply the SVM function derived as in [10] to feature vector classification, which makes a decision over whether the segment is speech or not.

3.3 Noise reduction

There exist various kinds of noises including background noise (babble, car, factory, etc.) and channel noise, and they decrease recognition accuracy significantly due to mismatches in training and operating environments. Since the spoken data of news reporters or interviewees is easily contaminated by various background noises, noise reduction is essentially required for broadcast news.

Over the last several decades, various techniques for noisy speech recognition have been reported, classifying in three categories: noise resistant features, speech enhancement, and speech model compensation for noise. Recently, significant works for noise reduction in the distributed speech recognition have been introduced and ETSI (European Telecommunication Standards Institute) published noise robust feature extraction scheme [11]. We apply this approach to our system.

The front-end proposed by ETSI is based on Wiener filter and it is performed in two stages. Figure 4 shows the main components of the noise reduction block of the front-end. The input signal is first de-noised in the first stage and the output of the first stage then enters the second stage. In the second stage, an additional, dynamic noise reduction is performed, which is dependent on the signal-to-noise ratio (SNR) of the processed signal.

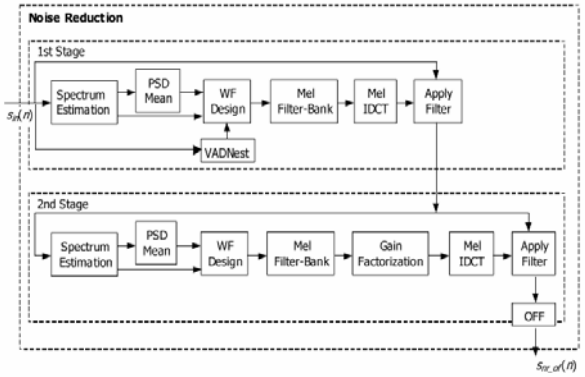

Fig. 3 Noise Reduction of ETSI Front-end

4. EXPERIMENTS

This section presents the recognition performance to verify the efficiency of our system. The system was evaluated based on keyword spotting accuracy, which means correctness in detection of keyword regions.

4.1 Experimental setup

We performed the keyword spotting experiments on Korean broadcast news data. For training a keyword HMM, about 100 utterances (each 50 utterances per a male and a female) were extracted from real news data. And about 40 utterances were used for reference patterns in DTW-based confidence measure. The reference patterns were extracted from the news and personally recorded in clean environments. We used 20 kinds of words as keywords and used about 15 hours' news data for evaluation. For feature parameters, we obtained 12 MFCCs (Mel-Frequency Cepstral Coefficients) and log energy with their first and second derivatives from all training and testing data.

4.2 Experimental result

The first experiment evaluates the efficiency of DTW-based confidence measure. We made a comparative study on six different groups according to data set used for reference patterns and distance criteria, as shown in Table 1. The reference patterns are divided into two groups, that is, the news data only ('News') and the news and personally recorded data ('News+Recorded'). Distance criteria indicates a value used as the distance between a test pattern and references. 'Minimum' means the smallest value among the distances between two patterns while 'average' means the average of all distances. Figure 4 shows the correct detection result according to false acceptance rate for each group in Table 1. The false acceptance rate means the ratio of the number of incorrectly detected utterances to the number of all accepted utterances while the correct acceptance rate

is calculated as the number of correctly detected utterances divided by the number of all utterances to be accepted, that is, the number of all keywords existing in the data. As shown in the result, the best group is 'DTW-1', where uses only the news data as reference patterns and the minimum distance value as distance criterion. In the 68.5% of correct acceptance rate, 'DTW-1' represents 31.1% of false acceptance rate, while 'no-DTW' shows 62.3%. 'no-DTW' means only applying N-best and likelihood-based confidence measure, excluding DTW-based method. According to the result, DTW-based confidence measure can reject about 50% of incorrectly detected utterances, while preserving the recognition result.

Tab. 1 Data groups used for evaluation of DTW-based confidence measure

Group	Reference Pattern	Distance Criteria
DTW-1	News	minimum
DTW-2	News+Recorded	minimum
DTW-3	News	average
DTW-4	News+Recorded	average
DTW-5	News	average except minimum & maximum
no-DTW	without DTW-based CM	

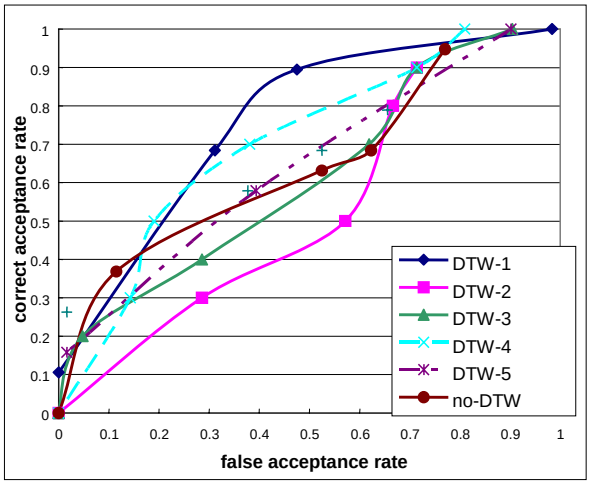

Fig. 4 Result of DTW-based confidence measure (ROC curve)

The second experiment evaluates the performance of our keyword spotting system. We investigated the detection result in about 15 hours' test data, and analyzed the number of false alarm and false reject, while changing the threshold of confidence measure.

Table 2 represents the experimental results of our system. Baseline system ('Baseline') just has keyword spotting module while our system ('Proposed') includes speech processing techniques explained in section 3. Compared to baseline system, our system shows 14.1% reduction of equal error rate (EER), which means the error rate where false rejection rate equals to false acceptance rate. Moreover, detection rate and detection accuracy was improved by 11.9% and 25%, respectively. Detection rate means the ratio of the number of correctly detected utterances to the number of overall keyword utterances, and detection accuracy is calculated as the number of correct detection divided by the number of detected utterances.

Another experimental result is presented as Figure 5. The DET curve, the representative performance measure of keyword spotting system, shows the variation of false rejection rate according to false acceptance rate. The better system performance is, the closer the curve comes to the origin.

Tab. 2 Experimental result of our system

	Baseline	Proposed	Relative Improvement
EER	35.3%	21.2%	21.8%
Detection Rate	76.2%	85.3%	11.9%
Detection Accuracy	62.8%	78.5%	25%

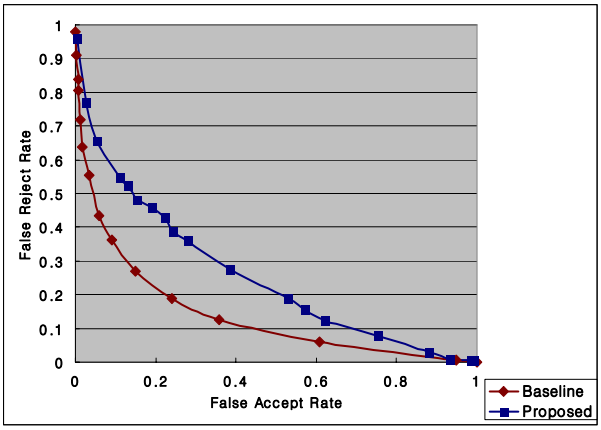

Fig. 5 Keyword spotting performance (DET curve)

5. CONCLUSIONS

This article addressed our keyword spotting system on broadcast news and several research topics for system improvements. While the system searches the articles containing the keyword(s) based on whole-word model, detection errors such as false alarm and false reject may degrade system performance. To reduce the errors, several techniques are applied to the system. Audio classification selects speech regions

except music or commercials, and we employed the feature-based method for the classification. To reduce various types of noises, we applied the ETSI-based noise reduction technique to our system. Utterance verification operates further critically, since it decides whether to accept or reject the detected utterances. For an improved verifier, we used DTW-based confidence measure, which verifies the detected utterance by the distance between the utterance and real keyword data. Keyword spotting experiments based on Korean broadcast news demonstrated superior performance of our system to the baseline. Especially, DTW-based confidence measure can reject about 50% of incorrectly detected utterances, while preserving the recognition result. We have a goal of improving our system and achieving the keyword spotting in more complicate spoken documents like interviews or movies, by applying further improved speech techniques.

REFERENCES

[1] James D., "The Application of Classical Information Retrieval Techniques to Spoken Documents", PhD thesis, Downing College, UK, 1995.

[2] Jean-Luc G., Lori L., Gilles A., "The LIMSI Broadcast News transcription system," Speech Communication, vol.37, pp.89-108, May 2002.

[3] M. G. Brown, J. T. Foote, "Automatic content-based retrieval of broadcast news," Proc. of ACM international conference on Multimedia, pp.35-43, 1995

[4] H. Jiang, "Confidence measures for speech recognition : a survey," Speech Communication, vol.45, no.4, pp.455-470, Apr.2005

[5] Gang G., Chao H., "A comparative study on various confidence measures in large vocabulary speech recognition," Proc. ISCSLP, pp.9-12, 2004

[6] Cory Myers, Lawrence R. Rabiner, Aaron E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition", IEEE Transactions on Acoustics, speech, and Signal Processing, Vol. ASSP-28, No. 6, 1980.

[7] Liao L., Gregory M.A., "Algorithms for speech classification," Proc. of International Symposium on Signal Processing and Its Applications, pp.623-627, 1999

[8] A.H.Gray Jr., J.D.Markel, "Distance measures for speech processing," IEEE Trans.Acoust.,Speech Signal Process., vol.ASSP-24, no.5, pp.380-391, Oct.1976.

[9] L.R.Rabiner, R.W.Shafer, Digital signal processing of speech signals Englewood Cliffs, NJ: Prentice-Hall, 1978.

[10] Ganapathiraju A., Hamaker J.E., Picone J., "Applications of support vector machines to speech recognition," Signal Processing, IEEE Transactions on, vol.52, pp.2348-2355, Aug. 2004.

[11] ETSI standard document, "Extended advanced front-end feature extraction algorithm" in ETSI ES 202 212 v1.1.1, 2003.8