Technical Issues for Automatic Article Retrieval from Spoken Broadcast News

[1]Jeong-Sik Park, [2] Kyung-Mi Park, [*3] Yung-Hwan Oh

[1] *Department of Intelligent Robot Engineering, Mokwon University, Daejeon, Republic of Korea, parkjs@mokwon.ac.kr*
[2] *Samsung Electronics, Suwon, Republic of Korea, kmpark@speech.kaist.ac.kr*
[3] *Computer Science Department, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, yhoh@speech.kaist.ac.kr*

Abstract

This article introduces several research topics related to the keyword spotting-based article retrieval system in broadcast news data. The system searches for the keyword speech from online broadcast news and retrieves all articles in which the corresponding keyword is spoken. For stable performance, various techniques are adopted including: utterance verification, out-of-vocabulary rejection, audio classification, and noise reduction. This article surveys several useful methods for the techniques and reports some experimental results. And then, several future works are addressed.

Keywords*: Automatic article retrieval, Keyword spotting, Spoken broadcast news, Speech recognition, online broadcast news*

1. Introduction

During the past years, voice search has been recognized as one of the most brilliant technologies adopted on smart devices [1, 2]. Capability of searching online information via voice became a representative application of voice interface, providing great convenience for users of smart devices. This advanced voice technology stimulated users' desire for spoken document retrieval that searches useful information or contents from a vast number of online multimedia data. However, it is still a challenge to retrieve and manage spoken documents rather than written documents, because speech data should be translated into text prior to retrieval.

Keyword recognition, also known as keyword spotting, provides the best solution for spoken document retrieval [3]. This technique analyzes a given spoken document and searches every speech segment in which one of pre-defined keywords is uttered. In general, the keyword recognition system provides more reliable performance than that of the continuous speech recognition system, while reducing computational time and intensity. Due to this efficiency, keyword recognition plays an important role in spoken document retrieval [4]. A representative application of keyword recognition is broadcast news retrieval [5, 6].

This paper concentrates on keyword spotting in broadcast news. Since the broadcast news data consist of speech utterances correctly pronounced by newscasters or reporters, they give better recognition results compared to other types of speech data such as conversations or natural voices, thus providing a higher possibility of a more reliable voice search application. In addition, a great demand for topic selection requires the keyword spotting technology. In general, people tend to select and read only the articles that they're interested in among a huge number of news. The keyword spotting system supports people when searching for such articles via a topic or a keyword [6].

* Corresponding Author
Received: Sep. 12, 2013, Revised: Oct. 11, 2013, Accepted: Dec. 18, 2013

This paper introduces a configuration for the keyword spotting system on broadcast news and reports several technical issues. The remainder of this paper is organized as follows. Section 2 introduces a framework of the standard keyword recognition system. Section 3 reports several technical issues for performance improvement. Section 4 explains the experimental setup and results. Finally, Section 5 presents our conclusions.

2. Keyword Spotting-based Article Retrieval from Spoken Broadcast News

Figure 1 represents the configuration of the keyword spotting system on broadcast news. The simple structure of the keyword spotting system consists of feature extractions and search processes. In short, acoustic feature parameters are continuously extracted from audio streams of the broadcast news. Then the features are compared with keyword models and garbage models to determine whether the corresponding stream is close to the keyword or not. The two types of models are obtained during the acoustic model training process. The keyword models and garbage models construct an amount of keyword speech data and non-keyword speech data, respectively. The garbage models are used to reduce non-keyword streams in the keyword-spotting system. Once a candidate keyword stream is found, the stream is verified based on the post-processing procedure and finally decided as a keyword or non-keyword. Figure 2 illustrates an example of the keyword spotting-based article retrieval.

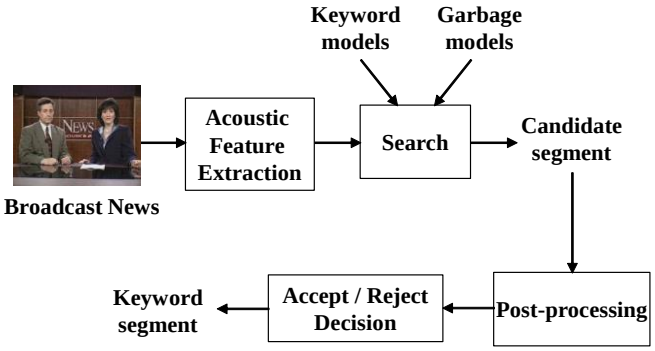

Figure 1. Configuration of Keyword Spotting System on Broadcast News

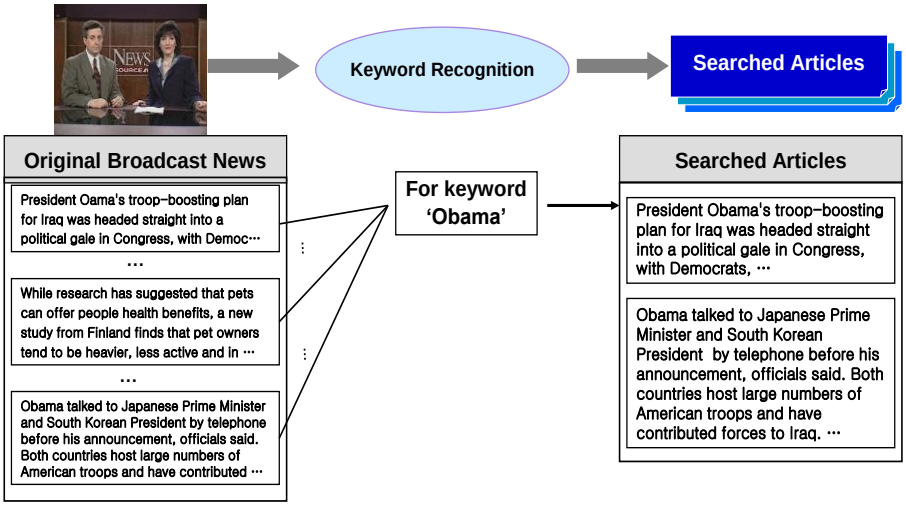

Figure 2. An Example of Keyword Spotting-based Article Retrieval

There are several kinds of search methods, which are based on the large vocabulary continuous speech recognition (LVCSR), the phoneme recognition, and the whole-word model. The LVCSR and phoneme recognition approaches produce text scripts for overall input speech prior to the search step, using word and phoneme-level transcriptions, respectively. The LVCSR-based approach provides reliable performance but requires hours of word-level transcriptions. Meanwhile, the phoneme recognizer requires less hardware resources than the LVCSR but gives poor performance. The whole-word model based search method takes advantage of the above two systems.

Although the keyword spotting system has a simple structure and operates fast compared to the continuous speech recognition system, two kinds of detection errors may occur more frequently, thus degrading the system performance. One is the false alarm that comes from incorrectly accepted data, and the other is the false rejection that comes from incorrectly rejected data. These two 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 when outdoor reporting degrade the speech quality, thus resulting in recognition errors. Finally, speakers such as male/female newscasters or interviewees are frequently changed, causing recognition errors. To preserve the keyword spotting system from unexpected errors, this article introduces several advanced techniques.

3. Techniques for Article Retrieval from Spoken Broadcast News

For the improvement of performance of the keyword spotting-based article retrieval system, this article concentrates on several techniques: speech and audio classification, noise reduction, and utterance verification. Both speech and audio classification and noise reduction operate the front-end of the system, whereas utterance verification performs for the post-processing. Figure 3 explains how these essential techniques work for the keyword spotting system.

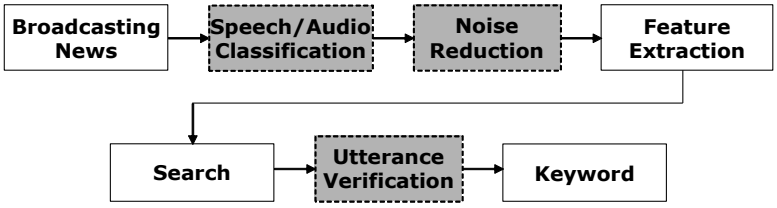

Figure 3. Essential Techniques for Keyword Spotting-based Article Retrieval

3.1. 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, 8]. This preliminary process is a challenging research topic in ASR and spoken document retrieval, and is necessary for keyword spotting in broadcast news which contains various audio types. Most of the audio classification techniques focus on two approaches: 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. Hence, 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. Accordingly, the feature-based approach is more suitable for the real-time system than the model-based approach, since the feature extraction requires less computational capacity than the model construction. Several features have been considered for the time domain (zero-crossing rate (ZCR), energy, etc.) and the frequency domain (sub-band power, low short-time energy ratio, etc.). Among them, several advanced features such VSF and VZCR are useful for speech and audio classification.

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

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

where S_i is the spectrum magnitude vector of frame i and N is the size of the window [9]. 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.

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

Here N is the length of the frame, m is the endpoint of the frame, and $w(n)$ is the window function [10]. 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.

3.2. Noise Reduction

Acoustic noises are one of the main factors that interfere with the speech processing systems. There exists various types of noises including background noise (babble, car, factory, etc.) and channel noise, and they deteriorate recognition accuracy due to mismatches in training and operating environments. Since broadcast news data also contain a vast amount of background noises, noise reduction is essentially required for the article retrieval system.

Over the last several decades, various techniques for noisy speech recognition have been investigated, which can be classified into 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]. The front-end proposed by ETSI is based on the Wiener filter and is performed through two stages. Figure 4 shows the main components of the noise reduction block of the front-end. The noise signals are eliminated from input speech in the first stage and the output of the first stage then enters the second stage. In the second stage, an additional and dynamic noise reduction is performed which is dependent on the signal-to-noise ratio (SNR) of the processed signal.

Conventional noise cancellation methods aim to eliminate stationary noises; consequently, their performance with speech-like noises is very poor. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [12, 13].

Among them, there is beamforming, a family of algorithms utilizing the geometric information, and Blind Signal Separation(BSS) [14] which is another family of algorithms that does not require any information of the source characteristics and microphone locations. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions.

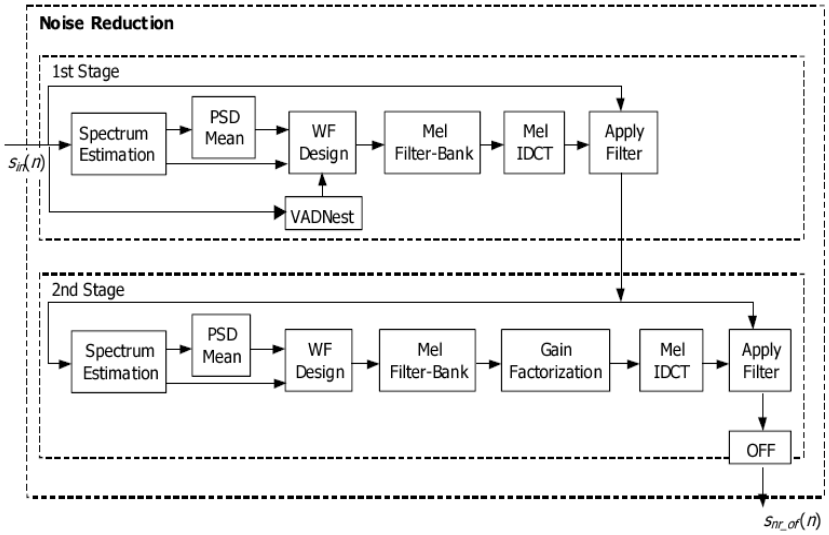

Figure 4. Noise Reduction of ETSI Front-end

3.3. Utterance Verification

Utterance verification has been used to improve system reliability in speech recognition tasks. Such techniques accept or reject recognition results using a decision criterion called a Confidence Measure (CM) [15]. We strongly believe that CMs can be productively applied to the post-processing of keyword recognition tasks, as both these and speech recognition tasks cope with equivalent post-processing problems.

Much CM-related research involves searching for features to distinguish correctly recognized results from recognition errors. Features that might be used include N-best recognition results, acoustic stability, duration, and the language model. Among these, N-best results-based CM is the most commonly used measure as it provides reliable verification performance without intensive computation [15, 16].

The N-best results involve a set of N hypotheses for a candidate keyword segment and scored recognition results for each hypothesis. For an HMM-based recognition system, the N-best results indicate N hypotheses ranked according to their log-likelihood output probabilities. The most representative confidence measure based on N-best results is defined as

$$CM = \frac{L_1}{\sum_{i=1}^{N} L_i} \tag{3}$$

$$CM = L_1 - \sum_{i=1}^{N} \frac{L_i}{N} \tag{4}$$

where L_i is the likelihood of i-th result in the N-best list [16]. N-best and likelihood means the N hypotheses, or recognition results, scored for an utterance. CM calculated in equation (3) or (4) is compared with the empirically determined threshold. If the value is greater than the threshold, the utterance is accepted as a keyword; otherwise, it is rejected.

4. Performance Evaluation

We verified the efficiency of the keyword spotting system investigated in this article. 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. To construct keyword models and garbage models, we trained Hidden-Markov Models (HMMs) from news data. For training each keyword HMM, about 100 utterances (each 50 utterances per a male and a female) were extracted from real news data. 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 evaluated the performance of the keyword spotting system employing speech processing techniques addressed in this article. We investigated the detection result in about 15 hours' test data, and analyzed the number of false alarms and false rejects, while changing the threshold of confidence measure.

Table 1 represents the experimental results of our system. Baseline system ('Baseline') just has the keyword spotting module while our system ('Proposed') includes speech processing techniques explained in Section 3. Compared to the baseline system, our system shows 14.1% reduction of equal error rate (EER), which means the error rate where false rejection rate equals the 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.

Table 1. Keyword Recognition Performance

	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%

The next experimental result shows the performance of our keyword spotting system using the Detection Error Trade-off (DET) curve. The DET curve measures the False Rejection Rate (FRR) against the False Acceptance Rate (FAR) for binary classification systems [3]. Figure. 5 represents the result. As shown in this figure, the proposed approach provided lower error rates than the baseline system. This result explains that the speech processing techniques investigated in this article positively affect the keyword spotting performance.

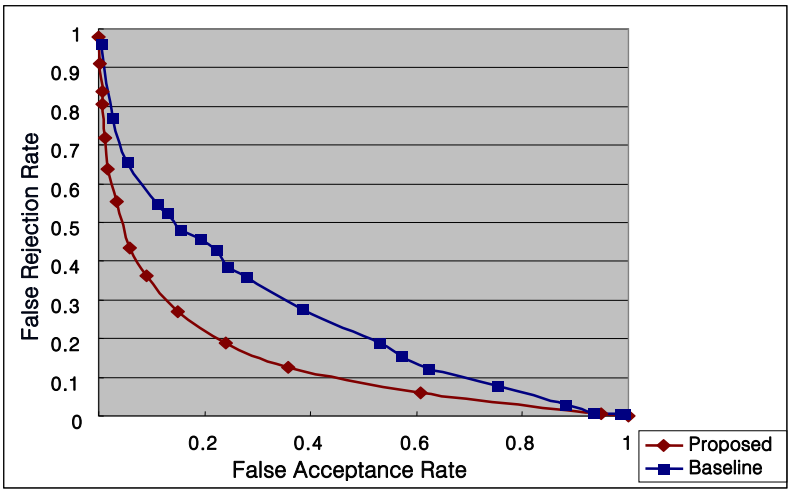

Figure 5. Detection Error Tradeoff (DET) curve for each approach

5. Conclusion

This article addressed several technical issues for the keyword spotting-based article retrieval system. While the system searches for articles corresponding to pre-defined keywords, detection errors such as false alarm and false reject may degrade system performance. To reduce the errors, several speech processing techniques were investigated: speech and audio classification, noise reduction, and utterance verification. To verify the efficiency of this approach, we conducted keyword-spotting experiments on Korean broadcast news data. The experimental result demonstrated superior performance of the proposed approach to the baseline.

6. Acknowledgments

This research was supported by the Converging Research Center Program through the Ministry of Science, ICT and Future Planning, Korea (2013K000358), and NAP (National Agenda Project) of the Korea Research Council of Fundamental Science and Technology.

7. References

[1] K. M. Park, J. S. Park, J. H. Bae, and Y. H. Oh, "Online speaker diarization for multimedia data retrieval on mobile devices," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 26, No. 8, pp. 1–22, Dec. 2012.

[2] G. J. Jang, J. S. Park, J. H. Kim, and Y. H. Seo, "Line spectral frequency-based noise suppression for speech-centric interface of smart devices," *Advances in Electrical and Computer Engineering*, Vol. 11, No. 4, pp. 3–8, Nov. 2011.

[3] J. S. Park, G. J. Jang, and J. H. Kim, "Multistage utterance verification for keyword recognition-based online spoken content retrieval," *IEEE Trans. Consum. Electron.*, Vol. 58, No. 3, pp. 1000–1005, Sept. 2012.

[4] D. James, "The application of classical information retrieval techniques to spoken documents," Ph.D. dissertation, Downing College, UK, 1995.

[5] G. Jean-Luc, L. Lori, and A. Gilles, "The LIMSI Broadcast News transcription system," *Speech Communication*, Vol. 37, pp. 89–108, 2002.

[6] M. G. Brown and J. T. Foote, "Automatic content-based retrieval of broadcast news," in Proc. ACM International Conference on Multimedia, San Francisco, CA, USA, Nov. 1995, pp. 35–43.

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

[8] H. G. Kim, G. J. Jang, J. S. Park, J. H. Kim, and Y. H. Oh, "Particle filtering based pitch sequence correction for monaural speech segregation," *International Journal of Imaging Systems and Technology*, Vol. 23, No. 1, pp.64–70, 2013.

[9] A. H. Gray Jr., and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 24, No. 5, pp.380–391, 1976.

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

[11] Extended advanced front-end feature extraction algorithm, ETSI standard document in ETSI ES 202 212 v1.1.1, 2003.

[12] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002, pp. 881–884.

[13] J. S. Park, G. J. Jang, J. H. Kim, and S. H. Kim, "Acoustic interference cancellation for a voice-driven interface in smart TVs," *IEEE Trans. Consum. Electron.*, Vol. 59, No. 1, pp. 244–249, 2013.

[14] I. Lee and G. Jang, "Independent vector analysis using densities represented by chain-like overlapped cliques in graphical models for separation of convolutedly mixed signals," *Electronic Letters*, Vol. 45, No. 13, pp. 710–711, 2009.

[15] H. Jiang, "Confidence measures for speech recognition: a survey," *Speech Communication*, Vol. 45, No. 4, pp. 455–470, 2005.

[16] G. Guo, C. Huang, H. Jiang, and R. H. Wang, "A comparative study on various confidence measures in large vocabulary speech recognition," in Proc. International Symposium on Chinese Spoken Language Processing, Hong Kong, Dec. 2004, pp. 9–12.