Confidence Measure for Utterance Verification in Keyword Spotting System

Jeong-Sik Park

Department of Intelligent Robot Engineering, Mokwon University, Daejeon, South Korea
parkjs@mokwon.ac.kr

Abstract

In this article, we propose an utterance verification technique for keyword spotting. The keyword spotting system analyzes a given spoken content and searches every speech segment in which one of pre-defined keywords is uttered. To maintain a stable recognition performance in the system, we propose an utterance verification technique that verifies whether a found utterance, or a candidate keyword segment, can be categorized as a keyword. The proposed approach employs a confidence measure based on the recognition results (N-best log-likelihood). In keyword spotting experiments using spoken broadcast news, our approach achieved superior performance compared to the conventional approach.

Keywords: *Keyword spotting, Utterance verification, Confidence measure.*

1. Introduction

Keyword spotting provides the best solution for multimedia retrieval [1]. The keyword spotting in spoken data analyzes a given content and searches every speech segment in which one of pre-defined keywords is uttered. In general, the keyword spotting system provides more reliable performance than that of the continuous speech recognition system, while reducing computational time and intensity. Due to this efficiency, keyword spotting plays an important role in the retrieval of spoken multimedia contents [2].

The keyword spotting system consists of several modules affecting its performance. Among them, this paper concentrates on utterance verification that verifies whether each candidate keyword segment is finally determined as a keyword utterance.

The remainder of this paper is organized as follows. Section 2 introduces a framework of the standard keyword spotting system. Section 3 proposes an utterance verification approach for keyword spotting. Section 4 explains the experimental setup and results. Finally, Section 5 presents our conclusions.

2. Keyword Spotting System

The standard keyword spotting system consists of two stages: model training and keyword spotting [3]. The model training stage aims at constructing two kinds of speech models, respectively called keyword models and garbage models. A keyword model indicates acoustic characteristics of the corresponding keyword, which are estimated from a set of keyword utterances. On the other hand, a garbage model, also known as a filler model, is used to absorb non-keyword segments. Thus, the garbage model is trained using a set of non-keyword segments.

A framework of the standard keyword spotting stage is described in Figure 1. First, acoustic feature parameters are extracted from each of consecutive speech segments of a given spoken content. The parameters are then applied to each of keyword models and garbage models in the search step. If a certain speech segment indicates acoustic characteristics of one of keyword models rather than those of garbage models, the segment is regarded as a keyword candidate. In the post-processing step, the candidate segment is verified whether it can be finally determined as a keyword utterance. This step accepts the candidate segment, categorizing as a keyword, or rejects it, categorizing as a non-keyword.

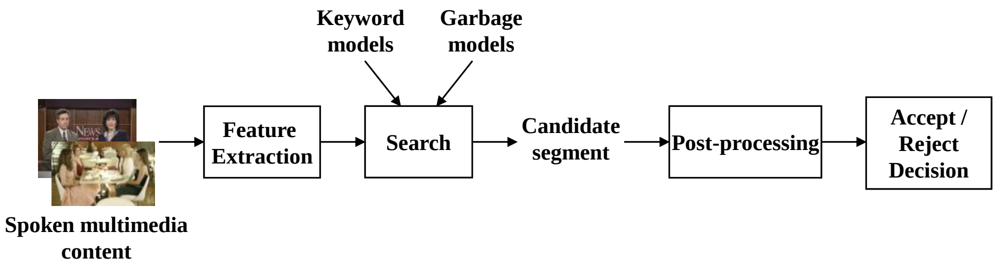

Figure 1. Framework of Keyword Spotting for Spoken Multimedia Content

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 [4][5]. The LVCSR and phoneme recognition approaches produce text scripts for overall speech segments prior to the search step, using word and phoneme-level transcriptions, respectively. The LVCSR-based approach provides reliable performance but requires tens of hours of word-level transcriptions. Meanwhile, the phoneme recognizer requires very less hardware resources than the LVCSR but gives poor performance. The whole-word model-based search method takes advantage of above two systems. In this method, word-level keyword and garbage models are trained and only the keyword-level transcription is generated.

In general, the performance of the keyword spotting system is determined by how frequently detection errors occur. There are two kinds of detection errors: false alarm and false rejection. The false alarm means a case that the system categorizes a non-keyword segment as a keyword and accepts it, whereas the false rejection occurs from a case that the system regards a keyword segment as a garbage and rejects it.

3. Confidence Measure for Utterance Verification in Keyword Spotting

To preserve the keyword spotting system from the two kinds of detection errors, this paper concentrates on utterance verification in the post-processing step that determines whether each segment recognized as a keyword candidate is finally accepted or rejected.

3.1. Confidence Measure using Keyword Recognition Results

Utterance verification has been applied to speech recognition tasks such as [6] for the purpose of improving the system reliability. This technique decides whether the recognition result is accepted or rejected depending on a decision criterion called the confidence measure (CM) [7]. We strongly believe that the measures can be very applicable to post-processing of keyword spotting tasks, as both the speech and keyword spotting tasks cope with equivalent problems in the post-processing.

A very large portion of CM-related works aim to search for a feature that is informative to distinguish correctly recognized results from other possible recognition errors. Some common features are related with: *N*-best recognition results, acoustic stability, duration, language model, etc. Among them, the *N*-best results-based CM is the most commonly used measure and provides reliable verification performance without intensive computation [7, 8].

The *N*-best results mean a list of *N* hypotheses and recognition result of each hypothesis, scored for a candidate keyword segment. In a Hidden Markov Model (HMM)-based recognition system, the *N*-best results indicate *N* hypotheses ranked according to output probability called log-likelihood. Let us denote $R_r(x)$ as the model index (ranging from 1 to *N*) at the *r*-th rank in the *N*-best list obtained from all *N* models (including both keyword and garbage models) with a given candidate keyword segment *x*. Two representative conventional measures based on *N*-best results are respectively described as

$$CM_1(x) = \frac{\log P(x \mid \lambda_{R_1(x)})}{\sum_{r=1}^{N} \log P(x \mid \lambda_{R_r(x)})}, \quad (1)$$

$$CM_2(x) = \log P(x \mid \lambda_{R_1(x)}) - \frac{1}{N} \sum_{r=1}^{N} \log P(x \mid \lambda_{R_r(x)}), \quad (2)$$

where $\lambda_{R_r(x)}$ and $\log P(x \mid \lambda_{R_r(x)})$ indicate the model corresponding to the index and the log-likelihood result at the *r*-th rank in the *N*-best list, respectively. These measures compute a relative distance between the log-likelihood at the first rank and overall log-likelihood results, on the assumption that the distance becomes larger for more confident segment. This assumption seems to be reasonable in general speech recognition tasks that process speech inputs belonging to pre-defined recognition units. But these conventional measures may induce an incorrect verification in keyword spotting tasks in which a limited number of garbage models cannot absorb all of non-keyword segments and each of those are constructed by mixed utterances, thus generating unreliable log-likelihood.

Considering the limitation of garbage models, we propose two new confidence measures. One of them is similar to (2) but we exclude the log-likelihood results for garbage models, as follows.

$$CM_3(x) = \log P(x \mid \lambda_{R_1(x)}) - \frac{1}{N_k} \sum_{r=1}^{N_K} \log P(x \mid \lambda^{*k}_{R_r(x)}), \quad (3)$$

where N_K indicates the number of keyword models and $\lambda^{*k}_{R_r(x)}$ means one of keyword models that corresponds to the model index.

The second measure concentrates on the distance between the log-likelihood at the first rank and that at the last rank in the recognition results for the keyword models as follows.

$$CM_4(x) = \log P(x \mid \lambda_{R_1(x)}) - \log P(x \mid \lambda^{*k}_{R_{N_K}(x)}). \quad (4)$$

This measure ignores log-likelihood results at the other ranks as well as results for garbage models. The standard keyword spotting system conducts a recognition process for a vast variety of unknown utterances with a limited number of acoustic models.

Thus, the log-likelihood results at the other ranks may negatively affect the verification of the candidate segment in comparison of those at the first and the last ranks.

4. Experimental Results and Analysis

To evaluate the effectiveness of our proposed approach, we conducted keyword spotting experiments. The experiments were performed on broadcast news data collected from a Korean news channel. Since broadcast news data consists of the sequence of read speech correctly pronounced by newscasters or reporters, the data has been widely adopted for the verification of continuous speech recognition systems. In particular, the broadcast news retrieval is a representative application of keyword recognition.

4.1. Experimental Setup

We extracted speech signals from about ten hours' broadcast news for evaluation. Then, we selected three representative keywords related with issues that are currently making headlines and attempted to search every speech segment, in which one of the keywords are uttered. To construct keyword models, we searched articles related with each keyword from Internet news providers and collected speech segments corresponding to the keyword utterances.

Each keyword HMM was trained using about twenty utterances spoken by male and female speakers. In addition, we constructed three garbage models using about one hundred utterances among the non-keyword segments. In short, three keyword HMMs and three garbage HMMs were constructed in the training stage. Acoustic feature parameters are configured as 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs) and log energy, and their first and second derivatives.

4.2. Experimental Results

As addressed in Section 2, the performance of the keyword spotting system is generally determined by two kinds of detection errors, that is, false alarm and false rejection. Most conventional studies investigate the keyword detection accuracy when the two kinds of errors indicate the equal error rate (EER) [9].

We investigated the accuracy of keyword detection for the ten hour's evaluation data, by applying each of confidence measures for utterance verification. Hence, the performances of the proposed measures, CM_3 (in (3)) and CM_4 (in (4)), were compared with those of the conventional measures, CM_1 (in (1)) and CM_2 (in (2)). We obtained the EER for each confidence measure, as shown in Table 1. The proposed measures successfully reduced the EER when compared to the conventional measures. This result demonstrates that our measures well verify the candidate keyword segments by ignoring the results for garbage models, but concentrating on the results for keyword models. It should be noted that CM_4 achieved the best performance, which means that the results at the first and the last ranks provide more useful criterion of keyword verification than results at the other ranks.

Table 1. The Equal Error Rate (EER) of our broadcast news retrieval system for each confidence measure.

Confidence Measure	Equal Error Rate
CM_1 (Conventional)	38.6%
CM_2 (Conventional)	35.8%
CM_3 (Proposed)	31.3%
CM_4 (Proposed)	28.6%

5. Conclusions

This paper proposed an efficient utterance verification technique for keyword spotting. Our approach applies a confidence measure based on recognition results for verifying candidate keyword segments. For this work, we advanced the conventional confidence measure adopted to speech recognition tasks. To verify the efficiency of our approach, we conducted keyword spotting experiments on broadcast news data. The advanced confidence measure successfully improved the accuracy of keyword detection.

For future works, we will verify our approach using other types of multimedia contents like interviews or movies and further improve the performance of the post-processing module using another utterance verification technique.

Acknowledgements

This study was financially supported by academic research fund of Mokwon University in 2012 and the NAP (National Agenda Project) of the Korea Research Council of Fundamental Science & Technology.

References

[1] H. J. Kim and J. Chang, "Discovering News Keyword Associations Using Association Rule Mining", J. of IWIT (The Institute of Webcasting, Internet and Telecommunication). 11, 63-71 **(2011)**.

[2] C. Chelba, T. J. Hazen and M. Saraclar, "Retrieval and Browsing of Spoken Content", IEEE Signal Process. Mag. 25, 39-49 **(2008)**.

[3] Z. Chenyan, L. Shuqin and S. Chengli, "Study of the Design and Implementation of Speech Keyword Recognition System based on Streaming Media", Proceedings of the 8th International Conference on Signal Processing, **(2006)**.

[4] I. Szoke, P. Schwarz, L. Burget, M. Karafiat and J. Cernocky, "Phoneme based Acoustics Keyword Spotting in Informal Continuous Speech", TSD 2005. LNCS. 3658, 302-309 **(2005)**.

[5] P. Cardillo, M. Clements and M. Miller, "Phonetic Searching vs LVCSR: How to Find What You Really Want in Audio Archives", Int. J. of Speech Technology. 5, 9-22 **(2002)**.

[6] J. Y. Ahn, S. B. Kim, S. H. Kim and K. I. Hur, "A Study on Voice Recognition using Model Adaptation HMM for Mobile Environment", J. of IWIT (The Institute of Webcasting, Internet and Telecommunication). 11, 175-180 **(2011)**.

[7] H. Jiang, "Confidence Measures for Speech Recognition: A Survey", Speech Communication. 45, 455-470 **(2005)**.

[8] G. Guo, C. Huang, H. Jiang and R. H. Wang, "A Comparative Study on Various Confidence Measures in Large Vocabulary Speech Recognition", Proceedings of the 4th International Symposium on Chinese Spoken Language Processing, **(2004)**.

[9] Z. Pengyuan, S. Jian, Z. Qingwei and Y. Yonghong, "Keyword Spotting Based on Syllable Confusion Network", Proceedings of the 3rd International Conference on Natural Computation, **(2007)**.

Authors

Jeong-Sik Park

Dr. Jeong-Sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. From 2010 to 2011, he was a Post-Doc. researcher in the Computer Science Department, KAIST. He is now a professor in the Department of Intelligent Robot Engineering, Mokwon University. His research interests include speech emotion recognition, speech recognition, speech enhancement, and voice interface for human-computer interaction.