Information Sciences 423 (2018) 157–171

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Discriminative keyword spotting using triphones information and N-best search

Shima Tabibian[a,*], Ahmad Akbari[b], Babak Nasersharif[b,c]

[a] Aerospace Research Institute, Ministry of Science, Research and Technology, Tehran 14665-834, Iran
[b] Audio & Speech Processing Lab, Computer Engineering Department, Iran University of Science & Technology, Tehran, Iran
[c] Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 6 June 2016
Revised 2 August 2017
Accepted 18 September 2017
Available online 20 September 2017

Keywords:
Discriminative keyword spotting
Hidden Markov model
Phone recognizer
Triphone
One-best search
N-best search

ABSTRACT

Keyword Spotting (KWS) systems can be divided into two main groups: Hidden Markov Model (HMM)-based and Discriminative KWS (DKWS) systems. In this paper, we propose an approach to improve a DKWS system using advantages of HMM-based systems. The proposed DKWS system contains feature extraction and classification (that includes a classifier and a search algorithm) parts. The focus of this paper is on the feature extraction part and the search algorithm. At first, we propose a method for using the advantages of a triphone-based HMM system and improving the monophone-based feature extraction, (proposed in our previous works), to triphone-based one. Then, we propose an N-best search algorithm instead of one-best algorithm. The results on TIMIT database indicate that the true detection rate of the triphone-based Evolutionary DKWS (EDKWS) system with N-best search (Tph-EDKWS-N-Best), in false alarm rate per keyword per hour greater than two, is 4.6% higher than that of the monophone-based EDKWS system with one-best search (Mph-EDKWS-1-Best). This improvement costs about 0.4 unit degradation in Real Time Factor (a common metric of measuring the speed of an automatic speech recognition system). Additionally, Figure of Merit (average true detection rate for different false alarm per keyword per hour from 1 to 10) of the Tph-EDKWS-N-Best system is noticeably higher than that of HMM-based KWS systems. However, the computational complexity of the Tph-EDKWS-N-Best system is considerably higher than that of the HMM-based KWS systems.

1. Introduction

This section presents the required introductory information for keyword spotting (KWS). The section is devided into four sub-sections. The first subsection defines the KWS and presents its applications. The second sub-section contains a literature review of KWS. The main motivations of the paper are discussed in the third sub-section. Finally, the last sub-section presents the paper structure.

* Corresponding author at: Postal address: Aerospace Research Institute, Ministry of Science, Research and Technology, Aerospace Research Center lane, Mahestan Street, Iran Zamin Street, Tehran Postal Code: 14657-74111, Iran.
E-mail addresses: tabibian@ari.ac.ir, shimatabibian@gmail.com (S. Tabibian), akbari@iust.ac.ir (A. Akbari), bnasersharif@kntu.ac.ir (B. Nasersharif).
URL: http://aspl.iust.ac.ir/ (S. Tabibian)

https://doi.org/10.1016/j.ins.2017.09.052
0020-0255/

1.1. Definition and applications

Nowadays, according to the increase of digital audio data volume, the use of speech in interaction between humans and machines becomes a critical necessity. Conversation between human and machine requires speech recognition and speech production systems. Speech recognition systems are used in a wide range of applications such as receiving and understanding a set of simple commands and even extracting all information from the speech signal. In some specific applications, the goal is to detect only particular keywords or phrases uttered by a speaker. In such cases, if the speaker utters other words or phrases rather than the special keywords or phrases, speech recognition problem is converted to KWS problem. KWS refers to discovering a set of target keywords in continuous speech utterances. Five major applications of KWS are keyword monitoring, audio document indexing, command control devices, dialogue systems and routing multimedia files and streams according to their content. We discuss some of these applications with more detail in the following.

Keyword monitoring applications [61]: A real time audio stream is continuously monitored to discover any occurrences of a keyword in it. Special keyword monitoring applications are telephone tapping, listening device monitoring and broadcast monitoring. Telephone tapping and listening device monitoring are used in the security organizations to detect criminal or spiteful activities. Broadcast commercial companies actively perform broadcast monitoring to locate the segments that may be interesting to a client.

Audio document indexing [18,61]: It is the task of quick searching an audio document for attractive keywords. It is similar to text search engine such as Google. However, it operates on audio documents instead of text archives.

Command-controlled devices [37,61]: They monitor the audio stream and react when a specific command is detected. Some examples of such command-controlled devices are: Speech-enabled mobile phones, voiced-controlled and command-controlled factory machinery, computer games, automatic teller banks, helping peoples with disabilities, online forms, appliances and software control via speech.

Dialogue systems [19,53,61]: They are exploited usually in the commercial environments instead of human operated call centers.

1.2. Literature review

There are two approaches for designing KWS systems. In the first approach, a Large Vocabulary Continuous Speech Recognition (LVCSR) system converts the input speech utterance to the text and then a text search algorithm determines the target keywords occurred in the utterance [9,43–45,55,65,66]. Therefore, we call the first approach, “LVCSR-based KWS”. In the second approach, KWS is considered as a function or a classifier without passing from the speech recognition step. We call it “direct KWS” (Fig. 1).

LVCSR-based KWS consists of two phases. First, a large vocabulary speech recognizer converts large audio archives to phone or word lattices. Then, in the second phase, lattice-based search looks for the set of target keywords. The first phase of LVCSR-based KWS is offline while the second one is online. This approach has three major drawbacks. Firstly, a large amount of labeled data is required to train LVCSR-based KWS. Secondly, the computational complexity implied by large vocabulary decoding is high. The third drawback is decrease of performance in Out Of Vocabulary (OOV) words. Different methods have been proposed to solve OOV problem [4,8,39,49,57].

In the second approach (direct KWS), KWS is considered completely independent from the speech recognition. We can consider keyword spotter as a function “KWS as a function” or as a binary classifier “KWS as a binary classifier”. In the first group [2,31,58,61], keyword spotter is a function with two input and two output arguments. The input arguments are input speech utterances and the set of target keywords. The output arguments consist of the confidence measure of keywords occurrences in the input speech utterances and the keywords positions. In some applications, detecting the keyword occurrence is more important than determining its position. In such cases, the output of the keyword spotter is only the confidence measure of keyword occurrence in the input speech. We can use one of several methods [6,17,35,46,64] to compute the confidence measure of keyword occurrence in the input speech. If the calculated confidence measure is greater than a predefined threshold, the keyword spotter confirms the occurrence of target keyword in the corresponding position of the input speech utterance.

In the second group, keyword spotter is a binary classifier [5,32,59,60] that separates the class of sentences with target keywords from the class of sentences without target keywords. Each classification method consists of two important parts: feature extraction and classification. In the feature extraction part, some discriminative features are extracted from the input speech utterances. These features model the confidence measure of occurrence of the target keyword in the input utterance and its position. In the classification part, a pre-trained hyper-plane (classifier) is exploited to separate the two mentioned classes with a minimum error rate, according to an evaluation measure.

Regardless of the approach used for solving KWS problem, it is necessary to determine an appropriate method for training the KWS systems. Hidden Markov Models (HMMs) are common methods and tools for training the KWS systems. We divide HMM-based KWS training methods into two main groups: LVCSR-based KWS (as discussed before) and phone-based KWS that considers KWS as a direct task independent of LVCSR systems. In the phone-based KWS [22,33,50–52,56,62,64], the keyword model is built from monophones, diphones or triphones sub-models. Phone-based KWS has several drawbacks. The main drawback is low phone recognition rate due to insertion, deletion and substitution errors. In some cases, in order to detect non-keyword parts of the input speech, a model called Garbage (Filler) model is trained and used along with