Speech Communication 102 (2018) 100–111

Contents lists available at ScienceDirect

Speech Communication

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

Sequence discriminative training for deep learning based acoustic keyword spotting☆

Zhehuai Chen, Yanmin Qian*, Kai Yu*

Computer Science and Engineering Department, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Keywords:
ASR
KWS
Sequence discriminative training
Generative sequence model
Discriminative sequence model

ABSTRACT

Speech recognition is a *sequence prediction* problem. Besides employing various deep learning approaches for frame-level classification, sequence-level discriminative training has been proved to be indispensable to achieve the state-of-the-art performance in large vocabulary continuous speech recognition (LVCSR). However, keyword spotting (KWS), as one of the most common speech recognition tasks, almost only benefits from frame-level deep learning due to the difficulty of getting competing sequence hypotheses. The few studies on sequence discriminative training for KWS are limited for fixed vocabulary or LVCSR based methods and have not been compared to the state-of-the-art deep learning based KWS approaches. In this paper, a sequence discriminative training framework is proposed for both fixed vocabulary and unrestricted acoustic KWS. Sequence discriminative training for both sequence-level generative and discriminative models are systematically investigated. By introducing word-independent phone lattices or non-keyword *blank* symbols to construct competing hypotheses, feasible and efficient sequence discriminative training approaches are proposed for acoustic KWS. Experiments showed that the proposed approaches obtained consistent and significant improvement in both fixed vocabulary and unrestricted KWS tasks, compared to previous frame-level deep learning based acoustic KWS methods.

1. Introduction

Keyword spotting (KWS) is one of the most widely used speech-related techniques, which requires a highly accurate and efficient recognizer specializing in the detection of some words or phrases of interest in continuous speech. KWS has broad applications, such as speech data mining (Zhou et al., 2005), low resource audio indexing (Shen et al., 2009), spoken document retrieval (Garofolo et al., 2000) and wakeup-word recognition (Chen et al., 2014a). The last two applications are considered in this paper.

KWS techniques can be categorized into two groups: (i) Unsupervised *query-by-example* (QbyE) (Zhang and Glass, 2009; Barakat et al., 2012; Chen et al., 2015a), which utilizes keyword audio samples to generate a set of keyword templates and matches them against testing audio samples to spot keywords. (ii) Supervised text-based method, which can be further divided into *large vocabulary continuous speech recognition* (LVCSR) based methods (Garofolo et al., 2000; Ng and Zue, 2000) and *acoustic KWS* (Mandal et al., 2014).[1] For LVCSR based methods, in training stage, a word or sub-word recognition system is constructed. Acoustic and language models are used to transcribe speech into a database of text or lattice during testing stage. Keyword searching is conducted on the database to get the final result. Acoustic KWS models the target keywords or sub-word sequences using an acoustic model without a language model. Some methods further include a series of non-keyword elements in the model (Sukkar et al., 1996). QbyE is mainly used in low resource audio indexing, which is not the focus of this paper. In spoken document retrieval, LVCSR based methods often show better performance than acoustic keyword spotting based method. However, LVCSR based methods have some inevitable shortcomings: requirement of large vocabulary coverage in training dataset, large computational resource requirement in both training and testing stage,[2] and out-of-vocabulary (OOV) problem, etc. These shortcomings limit its deployment in many practical applications such as wakeup-word recognition. Furthermore, LVCSR based KWS methods

☆ This work was supported by the National Key Research and Development Program of China (Grant No.2017YFB1002102), the China NSFC project (No. 61603252), the China NSFC project (No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

* Corresponding authors.
E-mail addresses: chenzhehuai@sjtu.edu.cn (Z. Chen), yanminqian@sjtu.edu.cn (Y. Qian), kai.yu@sjtu.edu.cn (K. Yu).

[1] A branch of newly proposed end-to-end methods (Kintzley et al., 2011; Audhkhasi et al., 2017) can also be viewed as a variant of it.
[2] except for low resource speech recognition, e.g. Babel project (Gales et al., 2014).

https://doi.org/10.1016/j.specom.2018.08.001
Received 28 September 2017; Received in revised form 16 May 2018; Accepted 1 August 2018
Available online 08 August 2018

ignore the special characteristics of KWS discussed in Section 2.1, and the performance improvements mainly rely on the advances of acoustic and language model in LVCSR. Therefore, this paper is focused on acoustic KWS.

In acoustic keyword spotting, models are typically trained to classify individual frames. Recent advances include two folds. First, applying a stronger frame-level classifier, deep neural network, yields significant improvements (Chen et al., 2014a; Sainath and Parada, 2015). Second, as speech recognition is inherently a sequence prediction problem, traditional GMM-HMM based systems achieve significantly better performance when trained using sequence discriminative criteria like discriminatively trained sub-word verification function (Sukkar and Lee, 1996), minimum classification error (MCE) (Sandness and Hetherington, 2000) and performance-related discriminative training (Keshet et al., 2009). Recently, within the deep learning framework, word-based connectionist temporal classification (CTC) model has also been used for KWS (Fernández et al., 2007). In all above sequence discriminative training methods, the complete search space modeling, i.e. hypothesis modeling, is the key of the success. However, in KWS, the in-domain search space specified by keyword sequences is much smaller. Thus the out-of-domain search space should be modeled by specific non-keyword elements as competitors. The difficulties in getting competing sequence hypotheses limit the usage of sequence discriminative training in KWS. Especially in unrestricted KWS, the possible competing words are usually not enumerable and the competing hypotheses generation is computationally expensive if using the same procedure as in LVCSR (Povey, 2005).

This paper proposes a sequence discriminative training framework for deep learning based unrestricted acoustic KWS. According to whether the model is defined for sequence conditional likelihood or sequence posterior probability, there are two types of sequence models: *generative sequence models* (GSM) such as HMM, and *discriminative sequence models* (DSM) such as CTC. For GSM, sequence discriminative training requires applying Bayes' theorem at sequence level to derive sequence conditional likelihood to posterior probability, while for DSM, sequence posterior probability can be used.

For both frameworks, competing hypotheses handling is the key difficulty. The paper proposes two methods to solve the problem: implicitly modeling a sub-word level language model and explicitly modeling non-keyword symbols. In HMM, inspired by the success of applying a pruned phone level language model to replace the word lattices in LVCSR discriminative training (Povey et al., 2016; Chen et al., 2006), the keyword sequences are modeled by a sub-word level acoustic model, and a corresponding language model is used to model the complete search space. To strengthen the discrimination ability of keywords, their gradients are weighted more significantly than those on non-keywords. Moreover, various neural network architectures and discriminative training criteria are compared. In CTC, non-keyword model units are introduced explicitly. Namely, the search space of sub-word level CTC based KWS is composed of keywords, phone boundaries (blank) and word boundaries (wb). Additional non-keyword spans (filler) are introduced in word level CTC based KWS. Lastly, an efficient post-processing algorithm is proposed to include phone confusions in the hypothesis searching.

The major contributions are summarized as follows: (i) The first work to systematically investigate sequence discriminative training for both generative and discriminative sequence models. (ii) Propose novel methods to construct competing hypotheses for sequence discriminative training for acoustic KWS and significantly improve the performance. (iii) Propose efficient post-processing methods to include phone confusion in hypotheses search.

The rest of the paper is organized as follows. In Section 2, the acoustic modeling in KWS is briefly reviewed. In Section 3, the traditional discriminative training methods are summarized. In Section 4 and Section 5, the proposed sequence discriminative training methods for deep learning based KWS are introduced respectively in CTC framework and HMM framework. Experiments are conducted on unrestricted KWS (spoken document retrieval task), and fixed vocabulary KWS (wakeup-word recognition task) in Section 6, followed by the conclusion in Section 7.

2. Acoustic modeling for keyword spotting

2.1. Comparison between LVCSR and KWS

LVCSR and acoustic KWS are two related but different speech recognition tasks. LVCSR focuses on accurately transcribing of the whole utterance, whereas KWS focuses on detecting some specific words or phrases of interest. Although some common techniques can be shared by the two tasks, they have different requirements on acoustic modeling. To show that it is not trivial to apply the sequence discriminative training techniques (originally developed for LVCSR) to KWS, it is necessary to discuss the special requirements of acoustic modeling for KWS.

- *Search space*. Due to extremely small vocabulary size, the in-domain search space of KWS is much smaller. Meanwhile, there are much more non-keywords in KWS than the out-of-vocabulary (OOV) words in LVCSR. Hence specific non-keyword models should be added into the search space of KWS system (Sukkar et al., 1996; Sukkar and Lee, 1996) to represent out-of-domain search space.
- *Model granularity*. Since the vocabulary in LVCSR is large, acoustic model granularities smaller than word are usually used [3], e.g., clustered tri-phones, which enhances both data efficiency and robustness (Young and Woodland, 1994). However, there is no such consideration for KWS, thus the model granularity can be keyword, sub-word, phone, tri-phone, etc.
- *Decoding*. In LVCSR, decoding refers to the search process to find the most likely sequence of labels given acoustic and language models. In contrast, acoustic KWS usually does not require a language model but needs post-processing after the frame-level acoustic model inference. The post-processing method can be categorized into three groups: (i) Posterior smoothing (Chen et al., 2014a). (ii) Model based inference (Ge and Yan, 2017). (iii) filler based decoding.[4] The first two groups aim to filter out the noise posterior output by heuristic or data-driven methods, respectively. The third group attempts to model the previously described out-of-domain search space, which will be explained in Section 4.2 in detail.

2.2. Acoustic modeling for KWS

The acoustic keyword spotting based method are typically trained to classify individual frames. In a deep learning based HMM hybrid system (NN-HMM) whose model granularity is the tri-phone state, a neural network is trained to calculate posterior probabilities of HMM states. Specifically, for an observation $\mathbf{o}_{ut}$ corresponding to time t in utterance u, $y_{ut}(s) = P(s|\mathbf{o}_{ut})$ is the output of the neural network for the HMM state s. The formulation is similar to traditional GMM-HMM based systems (Young and Woodland, 1994), except for the pseudo log-likelihood $\log p(\mathbf{o}_{ut}|s)$ of HMM states s,

$$p(\mathbf{o}_{ut}|s) \propto \frac{y_{ut}(s)}{P(s)} \tag{1}$$

where $P(s)$ is the prior probability of state s. In deep learning based

[3] Recent progress in end-to-end system makes word or sub-word level modeling become competitive (Graves et al., 2013; Chan, 2016; Chen et al., 2018b) and efficient (Chen et al., 2016). But the techniques have not been widely adopted.

[4] In some recent works (Chen et al., 2014b; 2017a), a small language model can be applied in the filler modeling and shows moderate improvement.