Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 1–12

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection☆

Inyoung Hwang[a], Hyung-Min Park[b], Joon-Hyuk Chang[a,*]

[a] *School of Electronic and Computer Engineering, Hanyang University, Seoul 133-791, Republic of Korea*
[b] *School of Electronic Engineering, Sogang University, Seoul 121-742, Republic of Korea*

Received 2 March 2015; received in revised form 5 September 2015; accepted 5 November 2015
Available online 30 November 2015

Abstract

In this paper, we investigate the ensemble of deep neural networks (DNNs) by using an acoustic environment classification (AEC) technique for the statistical model-based voice activity detection (VAD). From an investigation of the statistical model-based VAD, it is known that the traditional decision rule is based on the geometric mean of the likelihood ratio or the support vector machine (SVM), which is a shallow model with zero or one hidden layer. Since the shallow models cannot take an advantage of the diversity of the space distribution of features, in the training step, we basically build the multiple DNNs according the different noise types by employing the parameters of the statistical model-based VAD algorithm. In addition, the separate DNN is designed for the AEC algorithm in order to choose the best DNN for each noise. In the on-line noise-aware VAD step, the AEC is first performed on a frame-by-frame basis using the separate DNN so the *a posteriori* probabilities to identify noise are obtained. Once the probabilities are achieved for each noise, the environmental knowledge is contributed to allow us to combine the speech presence probabilities which are derived from the ensemble of the DNNs trained for the individual noise. Our approach for VAD was evaluated in terms of objective measures and showed significant improvement compared to the conventional algorithm.

Keywords: Voice activity detection; Statistical model; Acoustic environment classification; Deep neural network; Ensemble

1. Introduction

Voice activity detection (VAD) which classifies the period of speech and non-speech from an input speech signal is an essential part of speech signal processing in tasks such as speech recognition, speech enhancement, and efficient variable-rate speech coding. Among the various VAD methods, we focus on a statistical model-based approach, which was originated from the work of Ephraim and Malah Ephraim and Malah (1984) for speech enhancement due to its high detection accuracy as well as low computational complexity. Sohn *et al.* Sohn et al. (1999) devised the VAD based on a Gaussian statistical model by employing the decision rule based on the geometric mean of the likelihood ratio (LR), which can be considered as a heuristic way. The novelty of the statistical model-based VAD was extensively reviewed,

☆ This paper has been recommended for acceptance by Hugo Van Hamme.
* Corresponding author. Tel.: +82 2 2220 0355.
E-mail addresses: iyhwang88@hanyang.ac.kr (I. Hwang), hpark@sogang.ac.kr (H.-M. Park), jchang@hanyang.ac.kr (J.-H. Chang).

http://dx.doi.org/10.1016/j.csl.2015.11.003

so that further improved methods based on the LR test according to two hypotheses for speech presence and absence have been presented in many studies. Kang et al. (2008) proposed a decision rule based on a discriminative weight training method which fuses the LR through a linear weighed combination in which weights are optimized by the minimum classification error (MCE) training method based on the gradient descent algorithm. Yu and Hansen (2010) further improved the method of Kang et al. (2008) by applying a multiple observation technique to the decision rule which reflects the LR of not only the current frame, but also previous few frames. On the other hand, Jo et al. (2009) found that the LR corresponding to speech absence and presence cannot be separated by a linear decision function such as the geometric mean and a linear weighted combination due to its considerable class overlap in the feature space. They thus applied a support vector machine (SVM) to the statistical model-based VAD as a robust decision function since the SVM makes it possible to build an optimal hyper-plane among many possible hyper-planes separating the two classes of speech absence and presence and especially has the advantages on addressing nonlinear properties of the input feature vector by applying the kernel function. However, the SVM cannot be considered as an effective method especially for the statistical model-based VAD since it cannot fully take the advantage of multiple features due to its shallow properties. Shallow architecture lacks the ability to take into account the diversity of the nonlinear distribution of the feature vector since it has zero or at most one hidden layer.

Recently, the deep belief network (DBN) has been proposed, by Hinton and Salakhutdinov (2006), as a powerful hierarchical generative model not only for feature representation but also for classification by taking multiple-layered deep architecture. It is noted that the DBN is known to avoid the poor local-minima and over-fitting by the greedy layer-wise unsupervised learning process called pre-training. The superiority of the DBN compared to conventional shallow architecture-based machine learning techniques including the SVM has been reviewed, and thus the DBN has been successfully applied to various pattern recognition applications such as speech recognition (Mohamed et al., 2009, 2012; Hinton et al., 2012) and hand-written character recognition (Hinton and Salakhutdinov, 2006; Lee et al., 2007) as a state-of-the-art technique. The key idea behind the method of Zhang and Wu (2013) is to extract new features by transferring the acoustic features through deep nonlinear hidden layers since the deep model can combine multiple features in a nonlinear way to discover the regularity among the features. Thereafter, they proposed the denoising deep neural network (DNN)-based VAD approach (Zhang and Wu, 2013) which differs from Zhang and Wu (2013) in that it tries to minimize the reconstruction cross-entropy loss between the input noisy feature and its corresponding clean feature at the target while (Zhang and Wu, 2013) uses the noisy feature at the target in the pre-training step. In addition, Ryant et al. (2013) proposed the DNN-based VAD for web video such as YouTube by using 13 normalized MFCCs as feature vector. However, the DNN-based VAD is far from fully investigated yet in the area of the statistical model-based VAD under various noise environments, which is a main topic of interest in this study.

Before presenting our work, it is worthwhile to mention the performance of the VAD by incorporating the acoustic environment classification (AEC) technique since it is useful to build and use a different DNN scheme for various acoustic environments. In the literature, Sangwan et al. (2007) proposed a technique to use the SVM for environmentally aware VAD and to find the best operating point for the competitive Neyman–Pearson VAD. Gaussian mixture model (GMM), in the method of Choi and Chang (2012), was applied to perform the AEC for speech enhancement to adaptively select the optimal parameters for a given noise type. Recently, Xia and Bao (2014) applied the GMM-based noise classification to speech enhancement, which is different from the work of Choi and Chang (2012) in that the weighted denoising auto-encoder (WDA) model is chosen among a number of models which were trained for each kind of noise in the training data-set. Unfortunately, these methods are restricted in tracking the subtle changes in acoustics, which cause nonlinear change in the feature space since the acoustic features are extracted through the fixed filters such as a Mel-filter bank or linear prediction filter. In addition, both the SVM and GMM belong to the shallow method, and they thus cannot represent the diversity of the feature yet in this AEC task. We note that the DNN can be successfully used in representing raw speech data to encapsulate the underlying information associated with various acoustic scenes.

In this paper, we develop the statistical model-based VAD by employing the DNN with a multiple layer deep architecture as a novel decision rule in classifying the signal into speech or noise. The first step is to establish the baseline of the DNN by which the improved speech presence probability (SPP) is obtained based on the novel features of the statistical model-based VAD, namely the LR, the *a priori* signal-to-noise ratio (SNR), and the *a posteriori* SNR. As the key point that contributes to the success of DNN-based VAD, distinct DNNs according to different noise types are established via the separate training. Then different SPPs which correspond to a number of total noises are derived. As for the environmental awareness, an independent DNN module is constructed by a separate training process to