Speech Communication 83 (2016) 1–9

Contents lists available at ScienceDirect

Speech Communication

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

Feature mapping using far-field microphones for distant speech recognition

Ivan Himawan[a,*], Petr Motlicek[a], David Imseng[a], Sridha Sridharan[b]

[a] Idiap Research Institute, Martigny, Switzerland
[b] Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 29 June 2015
Revised 9 June 2016
Accepted 18 July 2016
Available online 19 July 2016

Keywords:
Deep neural network
Bottleneck features
Distant speech recognition
Meetings
AMI corpus

ABSTRACT

Acoustic modeling based on deep architectures has recently gained remarkable success, with substantial improvement of speech recognition accuracy in several automatic speech recognition (ASR) tasks. For distant speech recognition, the multi-channel deep neural network based approaches rely on the powerful modeling capability of deep neural network (DNN) to learn suitable representation of distant speech directly from its multi-channel source. In this model-based combination of multiple microphones, features from each channel are concatenated and used together as an input to DNN. This allows integrating the multi-channel audio for acoustic modeling without any pre-processing steps. Despite powerful modeling capabilities of DNN, an environmental mismatch due to noise and reverberation may result in severe performance degradation when features are simply fed to a DNN without a feature enhancement step. In this paper, we introduce the nonlinear bottleneck feature mapping approach using DNN, to transform the noisy and reverberant features to its clean version. The bottleneck features derived from the DNN are used as a teacher signal because they contain relevant information to phoneme classification, and the mapping is performed with the objective of suppressing noise and reverberation. The individual and combined impacts of beamforming and speaker adaptation techniques along with the feature mapping are examined for distant large vocabulary speech recognition, using a single and multiple far-field microphones. As an alternative to beamforming, experiments with concatenating multiple channel features are conducted. The experimental results on the AMI meeting corpus show that the feature mapping, used in combination with beamforming and speaker adaptation yields a distant speech recognition performance below 50% word error rate (WER), using DNN for acoustic modeling.

1. Introduction

Automatic speech recognition from distant microphones is a challenging task, because the speech signals to be recognized are degraded by the presence of interfering signals and reverberation due to large speaker-to-microphone distance (Yoshioka et al., 2012). The conventional multi-channel enhancement techniques, such as beamforming, are widely employed to suppress noise and reverberation from the desired speech when multiple microphones (e.g., microphone arrays) are used to capture audio signals (Anguera et al., 2007; Veen and Buckley, 1988).

In the context of ASR, the conventional speech enhancement methods are typically used as a pre-processing step to reduce mismatch between a model trained using clean speech and the noisy features. Since these methods are designed to improve signal-to-noise ratio (SNR), or signal-to-interference-plus noise ratio, the performance of the speech recognizer will be sub-optimal. In case of multi-channel ASR, there have been studies on designing a beamformer with the aim of optimizing ASR performance. A technique such as likelihood maximizing beamforming (LIMABEAM) (Seltzer et al., 2004; Seltzer and Stern, 2006) specifically optimizes array parameters using gradient descent to maximize the likelihood of the recognized hypothesis under an ASR speech model, given the filtered acoustic data. Recent research on LIMABEAM suggests no significant improvement using the standard LIMABEAM on large vocabulary distant speech recognition on the AMI meeting corpus and it is recommended to use a better optimization strategy for any LIMABEAM implementation (Fox and Hain, 2014).

Further, it is also possible to perform recognition from microphone arrays without employing any pre-processing steps. For example, each individual channel can be separately recognized, and the recognition hypotheses are combined using a confusion

* Corresponding author.
E-mail addresses: ihimawan@idiap.ch, ivan.himawan@idiap.ch (I. Himawan), pmotlic@idiap.ch (P. Motlicek), dimseng@idiap.ch (D. Imseng), s.sridharan@qut.edu.au (S. Sridharan).

http://dx.doi.org/10.1016/j.specom.2016.07.003
0167-6393/

network combination to select a word sequence with the highest probability (Metze et al., 2014; Wölfel and McDonough, 2005). Channel selection approaches such as finding the channel producing the maximum acoustic likelihood (Shimizu et al., 2000), or selecting the channel with the maximum confidence from its decoded sequence (Wolf and Nadeu, 2014), may be particularly useful when microphones are loosely specified in users' environments. Since recognition needs to be performed before any hypothesis is selected or combined, these decoder-based approaches for recognizing multiple microphones are computationally demanding (i.e., multi-pass-systems) .

Recently, acoustic models based on DNN have been shown to significantly improve the ASR performance on a variety of tasks when compared to the conventional Gaussian mixture model hidden Markov model (GMM/HMM) systems. Several international challenges have recently been organized to attract researchers' interest in providing the ASR solution in reverberant environments, such as the ASpIRE (Harper, 2015) and CHiME challenge series (Barker et al., 2015; 2013). In those challenges, participants were encouraged to build state-of-the-art speech recognition systems that are robust to various environmental factors and recording scenarios, while minimizing the impact of mismatch between training and testing conditions. For example, the recent 3rd CHiME challenge specifically addressed the far-field recordings from a mobile tablet device, captured using six microphones positioned around the tablet frame in real-world environments. It was reported that one of the most effective techniques, where significant gains have been achieved, is to transform the DNN features using feature-space maximum likelihood linear regression (fMLLR) (Hori et al., 2015; Sivasankaran et al., 2015), and some of the best scoring systems have used baseline DNN configurations for acoustic modeling (Barker et al., 2015). Apart from DNN's superior modeling capacity in acoustic modeling, a DNN which is trained with context-dependent phonetic targets can be used to produce neural-network-based features or bottleneck (BN) features. These features have been shown to be effective in improving the performance of ASR systems especially when exploited in combination with traditional short-term spectral features, such as MFCCs or PLPs (Seltzer et al., 2013; Yu and Seltzer, 2011). The BN features are usually extracted from one of the internal layers of DNN (with a small number of hidden units in comparison to the size of the other layers) and represent a nonlinear transformation (while usually reducing dimensionality) of the input features (Grézl et al., 2007; Yu and Seltzer, 2011). The stacked BN features which are extracted from the cascaded DNN structures have been investigated on several ASR tasks, such as speech recognition of Cantonese spontaneous telephone conversations (Karafiát et al., 2013) and speech recognition with minimum resource (Zhang et al., 2014). In Liu et al. (2014), the BN features were also used for far-field speech recognition.

This paper introduces a nonlinear BN feature mapping approach by using the BN feature of a close-talking microphone (referred to as the individual headset microphone (IHM)) as a target for distant speech input. The DNN is used to map the noisy and reverberant features to the BN-based features extracted from the close-talking input. Once the mapping is completed, the transformed BN features are extracted for training a new acoustic model (Himawan et al., 2015). The model-based combination of multiple microphones using the transformed BN features is proposed to integrate the multi-channel inputs for acoustic modeling. For the feature mapping approach, the fMLLR for speaker adaptation is applied to the features prior to DNN training and to the transformed BN features in the stacked hybrid fashion (Yoshioka et al., 2014). The fMLLR has been shown to be effective in both hybrid and tandem DNN-based systems for removing speaker variabilities and variations in the recording process, due to speaker-to-microphone distances and the use of different microphone channels (Bell et al., 2013; Swietojanski and Renals, 2014; Yoshioka et al., 2014). Although many recent speaker adaptation techniques for DNN have been proposed such as learning hidden unit contributions (LHUC) (Swietojanski and Renals, 2014), providing speaker identity vectors (i-vectors) along with regular ASR features as input to neural nets (Saon et al., 2013; Senior and Lopez-Moreno, 2014), and incorporating i-vectors to project the speech features into a speaker-normalized space (Miao et al., 2014a,b), it is straightforward to use fMLLR in DNN/HMM hybrid acoustic models. The GMM/HMM models which are usually trained to generate the alignment with context-dependent phone states for DNN training can further be used to estimate speaker transforms. This paper investigates the feature mapping approach for far-field microphones by examining the individual and preferably combined impacts of beamforming and fMLLR for robust ASR. The comparison to multi-condition training is also presented.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes the DNN-based mapping approach. The experimental setup is described in Section 4. The ASR results, employing the BN feature mapping approach using far-field microphones, are presented in Section 5. Section 6 discusses the results. Finally, the study is concluded in Section 7.

2. Related work

2.1. Speech enhancement using DNN

In a noisy and reverberant room, the reverberated speech $x(t)$ is represented in time domain as the convolution of the clean speech signal $s(t)$ and the room impulse response $h(t)$, corrupted by additive noise $n(t)$, as

$$x(t) = s(t) * h(t) + n(t). \tag{1}$$

The effect of early reflection and late reverberation on the reverberant signal is considered as a separate process in many studies. The late reverberation part of the room impulse response is often modeled as an exponentially damped Gaussian noise process and treated as additive noise. Hence, the observed reverberant signal $x(t)$ can be written by using the notation in Yoshioka et al. (2012) as

$$x(t) = s(t) * h_e(t) + r(t) + n(t), \tag{2}$$

where $h_e(t)$ is the early reflection part of the impulse response and $r(t)$ is the late reverberation component of $x(t)$.

The conventional methods to recognize reverberated speech captured from distant microphones is to first reconstruct a clean version of the speech. This may be performed with a blind dereverberation method, such as estimating the inverse filter solely on the observed signals capable to cancel out the reverberation effects (Miyoshi and Kaneda, 1988; Nakatani et al., 2005). Since the late reverberation is often treated as additive noise, speech enhancement methods, such as spectral subtraction (Boll, 1979) and minimum mean-square error (MMSE)-based techniques (Ephraim and Malah, 1984; 1985), may be used to mitigate the impact of reverberation. If two or more microphones are used to capture speech, multi-channel speech enhancement techniques such as multi-channel Wiener filter (Meyer and Simmer, 1997), beamforming followed by post-filtering (McCowan and Bourlard, 2003), or blind speech separation (Makino et al., 2007) can be used for improving the quality of speech. One drawback of these conventional speech enhancement methods is that they often fail to track the non-stationary noise signals in real-world scenarios.

One of the emerging speech enhancement approaches is based on deep architectures. In Xu et al. (2015), the DNN-based regression model was trained using noisy data and their corresponding