Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language 00 (2017) 1–12
www.elsevier.com/locate/csl

A generic neural acoustic beamforming architecture for robust multi-channel speech processing☆

Jahn Heymann*, Lukas Drude, Reinhold Haeb-Umbach

Paderborn University, Warburger Straße 100, Paderborn, Germany

Received 8 April 2016; received in revised form 27 September 2016; accepted 13 November 2016

Abstract

Acoustic beamforming can greatly improve the performance of Automatic Speech Recognition(ASR) and speech enhancement systems when multiple channels are available. We recently proposed a way to support the model-based Generalized Eigenvalue beamforming operation with a powerful neural network for spectral mask estimation. The enhancement system has a number of desirable properties. In particular, neither assumptions need to be made about the nature of the acoustic transfer function (e.g., being anechonic), nor does the array configuration need to be known. While the system has been originally developed to enhance speech in noisy environments, we show in this article that it is also effective in suppressing reverberation, thus leading to a generic trainable multi-channel speech enhancement system for robust speech processing. To support this claim, we consider two distinct datasets: The CHiME 3challenge, which features challenging real-world noise distortions, and the REVERBchallenge, which focuses on distortions caused by reverberation. We evaluate the system both with respect to a speech enhancement and a recognition task. For the first task we propose a new way to cope with the distortions introduced by the Generalized Eigenvalue beamformer by renormalizing the target energy for each frequency bin, and measure its effectiveness in terms of the PESQ score. For the latter we feed the enhanced signal to a strong DNN back-end and achieve state-of-the-art ASR results on both datasets. We further experiment with different network architectures for spectral mask estimation: One small feed-forward network with only one hidden layer, one Convolutional Neural Network and one bi-directional Long Short-Term Memory network, showing that even a small network is capable of delivering significant performance improvements.

Keywords: Robust speech recognition; Acoustic beamforming; Multi-channel speech enhancement; Deep neural network

1. Introduction

Acoustic beamforming has been considered as a front-end processing technique for Automatic Speech Recognition (ASR) for many years. As early as 1990 Compernolle et al. showed that significant word error rate (WER) improvements are achievable by acoustic beamforming (Compernolle et al., 1990). Research on acoustic beamforming has made great progress since then, including the use of novel objective functions, such as the multi-channel Wiener filter, and the consideration of arbitrary Acoustic Transfer Functions (ATFs) from the speech source to the

☆ This paper has been recommended for acceptance by Prof. Roger Moore

* Corresponding author.
E-mail address: jahnheymann@gmail.com (J. Heymann).

http://dx.doi.org/10.1016/j.csl.2016.11.007

microphones, thus giving up the assumption of an anechoic delay-only propagation path, see, e.g., Gannot and Habets (2013) for a tutorial.

While these modern beamforming concepts have been employed for speech communication tasks, their use as a front-end in ASR was rather limited. Further, with the recent success of ASR back-ends relying on Deep Neural Networks (DNNs), the front-end acoustic beamforming needs reconsideration.

An obvious approach to handle multi-channel signals is to first employ a conventional beamforming approach to condense the multiple signals into one signal which is then fed into a Deep Neural Network (DNN) back-end. Delcroix et al. have shown that a strong DNN back-end can be significantly improved with a sophisticated beamformer based on the Minimum Variance Distortionless Response (MVDR) criterion (Delcroix et al., 2014).

While this work showed the effectiveness of acoustic beamforming in a DNN-based ASR system, only few multi-channel approaches exist which directly employ DNNs. Swietojanski et al. employed the logarithmic Mel filterbank features of multiple acoustic channels as a parallel input to a CNN. They explored different weight sharing approaches and found that channel-wise convolution followed by a cross-channel max-pooling performed better than multi-channel convolution (Swietojanski et al., 2014). This approach, however, has the intrinsic drawback that the information on the relative phases between the channels is lost, since current feature extraction methods are agnostic to the phase. On the other hand it is well-known that in geometrically compact microphone array configurations the main difference between the signals of the individual channels reside in their phases, not in their magnitudes.

An alternative approach to make use of multiple input channels for ASR is to leverage temporal difference information between channels by directly working on the raw waveform, i.e., feeding the time domain signals into the DNN. Hoshen et al. reported noticeable performance gains over single-channel input (Hoshen et al., 2015). Following works are even able to achieve better results than a MVDR beamformer (Sainath et al., 2015; 2016).

Others proposed to jointly train a MVDR beamformer and the acoustic model (Xiao et al., 2016). Thereby, they use a DNN to estimate the beamforming weights for the MVDR beamformer given the Time Differences of Arrival (TDOA), perform the beamforming operation, extract the features and finally use these features to train an acoustic model. During this training, they are able to backpropagate the cross-entropy error down to the network estimating the beamforming weights.

In this paper we adhere to the conventional approach of first condensing multiple input channels to a single enhanced output signal to be fed to the ASR back-end. However, we still make use of the recent progress in DNNs by employing a neural network component in the estimation of the beamformer coefficients. We consider the acoustic beamformer to be a multiple-input single-output (MISO) linear time-invariant filter. A key concern is how to estimate the filter coefficients to extract the target signal while suppressing interferences, exploiting the different spatial and spectral properties of the target and the distortions. For the Delay-and-Sum Beamformer (DSB), the filter coefficients can be derived from an estimate of the Direction-of-Arrival (DoA), if the geometry of the microphone array is known. Note that the assumption underlying the DSB is that of an anechoic acoustic environment. If reverberation is to be accounted for, the (relative) ATFs between source and sensors are estimated, which usually requires an estimation of the statistics of the target speech signal (Gannot et al., 2001). Further, advanced beamforming concepts also require an estimate of the Cross-Power Spectral Density (PSD) matrix of the noise signal.

These statistics can be obtained by estimating spectral masks for speech and noise which are typically obtained by model-based methods, i.e. Sawada, Araki, Makino, (2011), Ito et al. (2014), Vu and Haeb-Umbach (2010), Ito et al. (2013), Yoshioka et al. (2015), Araki and Nakatani (2011), Araki et al. (2016). Instead of using a model-based approach, we recently proposed to use a DNN to estimate those masks. A distinctive advantage of the proposed neural network based mask estimation is that we explicitly account for time and frequency dependencies during mask estimation whereas most model-based approaches treat individual frequencies independently. This improves the accuracy of the estimated signal statistics and hence the overall results (Heymann et al., 2016). Additionally, making no assumptions about the distribution of the data for masking but rather inferring it from the training data, we expect this approach to be more robust against different noise types and reverberation. Further, by carrying out mask estimation for each channel separately and relying on microphone array independent signal statistics renders the trained neural network parameters independent of the microphone array configuration. Thus our approach can be applied to arbitrary array configurations. We can even cope with array configurations at test time, which are different from those at training time but still employ a powerful DNN in the multi-channel processing pipeline.

DNNs for mask estimation have been used in single channels speech enhancement for a while (e.g. Narayanan and Wang, 2013) and even extended to include the phase (Williamson et al., 2016). Although similar, the overall

Please cite this article as: J. Heymann et al., A generic neural acoustic beamforming architecture for robust multi-channel speech processing, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2016.11.007