Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 68 (2015) 97–106

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Distant speech separation using predicted time–frequency masks from spatial features

Pasi Pertilä *, Joonas Nikunen

Tampere University of Technology, Department of Signal Processing, P.O. Box 553, 33101-FI, Finland

Received 27 August 2014; received in revised form 8 December 2014; accepted 28 January 2015
Available online 7 February 2015

Abstract

Speech separation algorithms are faced with a difficult task of producing high degree of separation without containing unwanted artifacts. The time–frequency (T–F) masking technique applies a real-valued (or binary) mask on top of the signal's spectrum to filter out unwanted components. The practical difficulty lies in the mask estimation. Often, using efficient masks engineered for separation performance leads to presence of unwanted musical noise artifacts in the separated signal. This lowers the perceptual quality and intelligibility of the output.

Microphone arrays have been long studied for processing of distant speech. This work uses a feed-forward neural network for mapping microphone array's spatial features into a T–F mask. Wiener filter is used as a desired mask for training the neural network using speech examples in simulated setting. The T–F masks predicted by the neural network are combined to obtain an enhanced separation mask that exploits the information regarding interference between all sources. The final mask is applied to the delay-and-sum beamformer (DSB) output.

The algorithm's objective separation capability in conjunction with the separated speech intelligibility is tested with recorded speech from distant talkers in two rooms from two distances. The results show improvement in instrumental measure for intelligibility and frequency-weighted SNR over complex-valued non-negative matrix factorization (CNMF) source separation approach, spatial sound source separation, and conventional beamforming methods such as the DSB and minimum variance distortionless response (MVDR).

Keywords: Speech separation; Microphone arrays; Neural networks; Beamforming; Time–frequency masking

1. Introduction

Source separation refers to a process of estimating individual sound sources from an observed mixture of sources. In the distant talker scenario, investigated in this work, multiple speech sources are away from the array and the captured mixture contains reverberated target signals embedded in noise, which makes the separation task non-trivial. The applications of speech separation include automatic speech recognition (ASR), hearing aids, teleconferencing and post-processing of recordings.

Speech can be approximated as a sparse signal, i.e. that simultaneous speech consist of non-overlapping T–F points of each speaker. The assumption is also known as the w-disjoint orthogonality (Ylmaz and Rickard, 2004). Given a stimulus, auditory masking (Fastl and Zwicker, 1990) can cause inaudibility of lower amplitude frequencies inside a critical band. Auditory masking occurs also after the end of a stimulus, and to some degree even before the onset. Therefore, even non-overlapping interference can reduce the perception of the target signal.

* Corresponding author. Tel.: +358 40 8490 786; fax: +358 3 364 1352.
E-mail addresses: pasi.pertila@tut.fi (P. Pertilä), joonas.nikunen@tut.fi (J. Nikunen).

http://dx.doi.org/10.1016/j.specom.2015.01.006
0167-6393/

Four main approaches for source separation can be identified (Hummersone et al., 2014): blind source separation (BSS) methods, computational audio scene analysis (CASA) approaches, spatial filtering methods such as beamforming, and the non-negative matrix factorization (NMF) approaches.

The independent component analysis (ICA) (Hyvärinen and Oja, 2000) is a popular BSS scheme that can be used to separate independent and non-Gaussian source signals. ICA can be applied in the frequency domain to deal with reverberation, but since each frequency is processed separately the source permutations needs to be solved in addition (Smaragdis, 1998; Sawada et al., 2004).

T–F separation methods apply a mask on top of the observed spectrogram to separate the desired (speech) signal from interference. The ideal binary mask (IBM) assigns each T–F point a value of 1 or 0 depending whether or not the point belongs to target, and it can be considered as the goal of CASA (Wang, 2005). The real-valued ideal ratio mask (IRM), presented in Srinivasan et al. (2006), contains values in the range [0, 1]. The well known Wiener filter (WF) can be considered as a T–F mask and it is an important special case of the IRM since it minimizes the estimation error of reconstruction. Recently, Hummersone et al. (2014) argued that IRM may be more closely related to auditory processes than IBM and reviewed studies favoring IRM over IBM in certain ASR tasks and speech intelligibility (SI) measurements. In a listening experiment performed by Madhu et al. (2013) the WF based mask achieved higher SI in noisy conditions over the IBM. For hearing impaired listeners with cochlear implants the WF based mask and IBM resulted in equal SI (Koning et al., 2015). The normal hearing listeners preferred the use of WF based mask over IBM in a pairwise comparison test.

Largely the difficulty in T–F based masking is the actual mask estimation. Machine learning techniques have been researched in speech enhancement and separation intensively for mono signals. Kim et al. (2009) propose a Bayesian classifier to predict the IBM using the Gaussian mixture model (GMM), trained in three types of noise. The algorithm provided SI improvement for normal hearing listeners in noise conditions of -5 dB, and 0 dB. Healy et al. (2013) utilize DNNs to predict IBM, and demonstrate that the method can significantly improve SI for hearing impaired listeners. Weninger et al. (2014) used a long short-term memory (LSTM) recurrent neural network (RNN) to predict a T–F mask for speech enhancement. Narayanan and Wang (2013) use spectral features such as Mel-frequency cepstral coefficients and their delta components as the features of a DNN for obtaining an IRM. Furthermore, using NMF on the DNN output has been proposed for enhancing the perceptual quality of separated speech by Williamson et al. (2014). A combination of DNNs and support vector machines (SVMs) for speech enhancement using binary classification of T–F bands is proposed by Wang and Wang (2013). Deep recurrent autoencoder neural network is used to denoise input features in noise robust automatic speech recognition (ASR) by Maas et al. (2012). Swietojanski et al. (2013) reported that using a DNN in distant speech recognition by concatenating the amplitude based feature values from multiple channels resulted in increased speech recognition results over using the DNN with features from a single channel or a beamformer output. The approach however did not consider spatial features between multiple channels.

The NMF and non-negative matrix deconvolution (NMD) have been widely applied for monaural speech enhancement (Raj et al., 2010; Mohammadiha et al., 2013) and directly for noise-robust feature extraction for ASR (Gemmeke et al., 2011; Schuller et al., 2010). The single channel speech enhancement with NMF requires training the signal model based on a prior knowledge about the noise context or speaker identity. For source separation in a blind and unforeseen scenario the multichannel and complex-valued extensions of NMF have been proposed by Ozerov and Fevotte (2010) and by Sawada et al. (2013). The CNMF based separation proposed by Sawada et al. (2013) uses the spatial cues in conjunction with spectral redundancy in estimation of the parameters of the sources, i.e. mixing and magnitude spectrogram. The reconstruction of the sources is based on multichannel Wiener filtering, which can be regarded as a MVDR beamformer with a single channel post-filter applied to its output (Simmer et al., 2001). A comparison of separation performance of CNMF and conventional frequency-domain ICA is reported by Nikunen and Virtanen (2014a). The NMF based methods in general operate off-line both in single and multichannel case.

Introducing a second microphone enables spatial cues. The human head and torso cause source position and frequency specific modifications to the received signal. Such recordings from microphones placed in the ears are referred as binaural, and the typical cues are interaural time delay (ITD) and interaural level difference (ILD). The degenerate un-mixing estimation technique (DUET) clusters each T–F point based on its cue values (Ylmaz and Rickard, 2004), and obtains a binary mask for each cluster. Supervised learning via kernel-density estimation is used to obtain binary T–F mask values by Roman et al. (2003). The IRM prediction from binaural features was proposed by Srinivasan et al. (2006), who used it for speech enhancement based ASR. The approach was compared to missing data ASR using the IBM prediction method by Roman et al. (2003). The small vocabulary task was marginally more successfully recognized with the missing data approach, while the large vocabulary recognition task was significantly more successful using the IRM. Ayllon et al. (2013) utilize the generalized discriminant analysis (GDA) to classify binaural features to predict IBM for hearing aids. The authors analyze different types of non-linear feature transforms, using varying amounts of adjacent T–F points, and propose an evolutionary quantization scheme to achieve low transmission rate to exchange feature values between the two hearing aid devices. Woodruff and Wang (2013) propose the use of spa-