Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 606–621

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Combining localization cues and source model constraints for binaural source separation

Ron J. Weiss *, Michael I. Mandel, Daniel P.W. Ellis

LabROSA, Dept. of Electrical Engineering, Columbia University, New York, NY 10027, USA

Available online 12 January 2011

Abstract

We describe a system for separating multiple sources from a two-channel recording based on interaural cues and prior knowledge of the statistics of the underlying source signals. The proposed algorithm effectively combines information derived from low level perceptual cues, similar to those used by the human auditory system, with higher level information related to speaker identity. We combine a probabilistic model of the observed interaural level and phase differences with a prior model of the source statistics and derive an EM algorithm for finding the maximum likelihood parameters of the joint model. The system is able to separate more sound sources than there are observed channels in the presence of reverberation. In simulated mixtures of speech from two and three speakers the proposed algorithm gives a signal-to-noise ratio improvement of 1.7 dB over a baseline algorithm which uses only interaural cues. Further improvement is obtained by incorporating eigenvoice speaker adaptation to enable the source model to better match the sources present in the signal. This improves performance over the baseline by 2.7 dB when the speakers used for training and testing are matched. However, the improvement is minimal when the test data is very different from that used in training.

© 2011 Elsevier B.V. All rights reserved.

Keywords: Source separation; Binaural; Source models; Eigenvoices; EM

1. Introduction

Human listeners are often able to attend to a single sound source in the presence of background noise and other competing sources. This is partially a result of the human auditory system's ability to isolate sound sources that arrive from different spatial locations, an effect of the fact that humans have two ears (Cherry, 1953). Localization is derived from low-level acoustic cues based on the time and level differences of the sounds arriving at a listener's ears (Blauert, 1997). The use of these perceptual localization cues has had much success in the development of binaural source separation algorithms (Yilmaz and Rickard, 2004; Mandel et al., 2007). Unlike competing source separation approaches such as independent component analysis, localization-based algorithms are often able to separate mixtures containing more than two sources despite utilizing only binaural observations.

In contrast to binaural source separation based on the same principles used by the human auditory system, the most successful approaches to separating sources given a single channel observation have been model-based systems which rely on pre-trained models of source statistics (Cooke et al., 2010). Such monaural source separation algorithms generally require relatively large, speaker-dependent (SD) models to obtain high quality separation. These supervised methods therefore have the disadvantage of requiring that the identities of all sources be known in advance and that sufficient data be available to train models for each them. In contrast, most binaural separation algorithms based on localization cues operate without any prior knowledge of the signal content. The only assumption they make is that the sources be spatially distinct from one another. However, it is to be expected that incorporating some prior knowledge about the source characteristics would be able to further improve separation performance.

* Corresponding author.
E-mail addresses: ronw@ee.columbia.edu (R.J. Weiss), mim@ee.columbia.edu (M.I. Mandel), dpwe@ee.columbia.edu (D.P.W. Ellis).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.01.003

In this paper, we describe a system for source separation that combines inference of localization parameters with model-based separation methods and show that the additional constraints derived from the source model help to improve separation performance. In contrast to typical model-based monaural separation algorithms, which require complex SD source models to obtain high quality separation, the proposed algorithm is able to achieve high quality separation using significantly simpler source models and without requiring that the models be specific to a particular speaker.

The remainder of this paper is organized as follows: Section 2 reviews previous work related to the algorithms we describe in this work. Section 3 describes the proposed signal model for binaural mixtures and Section 4 describes how this model is used for source separation. Experimental results comparing the proposed systems to other state of the art algorithms for binaural source separation are reported in Section 5.

2. Previous work

In this paper, we propose an extension of the model-based expectation maximization source separation and localization (MESSL) algorithm (Mandel et al., 2010), which combines a cross-correlation approach to source localization with spectral masking for source separation. MESSL is based on a model of the interaural phase and level differences derived from the observed binaural spectrograms. This is similar to the degenerate unmixing estimation technique (DUET) algorithm for separating underdetermined mixtures (Yilmaz and Rickard, 2004) and other similar approaches to source localization (Nix and Hohmann, 2006) which are based on clustering localization cues across time and frequency. These systems work in an unsupervised manner by searching for peaks in the two dimensional histogram of interaural level difference (ILD) and interaural time, or phase, difference (ITD or IPD) to localize sources. In the case of DUET, source separation is based on the assumption that each point in the spectrogram is dominated by a single source. Different regions of the mixture spectrogram are associated with different spatial locations to form time–frequency masks for each source.

Harding et al. (2006) and Roman et al. (2004) take a similar but supervised approach, where training data is used to learn a classifier to differentiate between sources at different spatial locations based on features derived from the interaural cues. Unlike the unsupervised approach of Yilmaz and Rickard (2004) and Nix and Hohmann (2006), this has the disadvantage of requiring labeled training data. MESSL is most similar to the unsupervised separation algorithms, and is able to jointly localize and separate spatially distinct sources using a parametric model of the interaural parameters estimated directly from a particular mixture.

A problem with all of these methods is the fact that, as we will describe in the next section, the localization cues are often ambiguous in some frequency bands. Such regions can be ignored if the application is limited to localization, but the uncertainty leads to reduced separation quality when using spectral masking. Under reverberant conditions the localization cues are additionally obscured by the presence of echoes which come from all directions. Binaural source separation algorithms that address reverberation have been proposed by emphasizing onsets and suppressing echoes in a process inspired by the auditory periphery (Palomäki et al., 2004), or by preprocessing the mixture using a dereverberation algorithm (Roman and Wang, 2006).

In this paper, we describe two extensions to the unsupervised MESSL algorithm which incorporate a prior model of the underlying anechoic source signal which does not suffer from the same underlying ambiguities as the interaural observations and therefore is able to better resolve the individual sources in these regions. Like the supervised separation methods described above, this approach has the disadvantage of requiring training data to learn the source prior (SP) model, but as we will show in Section 5, such a prior can significantly improve performance even if it is not perfectly matched to the test data. Furthermore, because the source prior model is trained using anechoic speech, it tends to de-emphasize reverberant noise and therefore improves performance over the MESSL baseline, despite the fact that it does not explicitly compensate for reverberation in a manner similar to Palomäki et al. (2004) or Roman and Wang (2006).

The idea of combining localization with source models for separation has been studied previously in Wilson (2007) and Rennie et al. (2003). Given prior knowledge of the source locations, Wilson (2007) describes a complementary method for binaural separation based on a model of the magnitude spectrum of the source signals. This approach combines a model of the IPD based on known source locations with factorial model-based separation as in Roweis (2003) where each frame of the mixed signal is explained by the combination of models for each of the underlying source signals. The system described in Wilson (2007) models all sources using the same source-independent (SI) Gaussian mixture model (GMM) trained on clean speech from multiple talkers. Such a model generally results in very poor separation due to the lack of temporal constraints and lack of source-specific information available to disambiguate the sources (Weiss and Ellis, 2010). In this case, however, the localization model is able to compensate for these shortcomings. Per-source binary masks are derived from the joint IPD and source model and shown to improve performance over separation systems based on localization cues alone.

Rennie et al. (2003) take a similar approach to combining source models with known spatial locations for separation using microphone arrays. Instead of treating the localization and source models independently, they derive a model of the complex speech spectrum based on a prior on the speech magnitude spectrum that takes into account

Download English Version:

https://daneshyari.com/en/article/567529

Download Persian Version:

https://daneshyari.com/article/567529

Daneshyari.com