Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 84–99

www.elsevier.com/locate/specom

Compositional model for speech denoising based on source/filter speech representation and smoothness/sparseness noise constraints

P. Cabañas-Molero*, D. Martínez-Muñoz, P. Vera-Candeas, F.J. Cañadas-Quesada, N. Ruiz-Reyes

Department of Telecommunication Engineering, University of Jaén, Polytechnic School, Linares, Jaén, Spain

Received 10 January 2015; received in revised form 14 October 2015; accepted 14 October 2015
Available online 2 December 2015

Abstract

We present a speech denoising algorithm based on a regularized non-negative matrix factorization (NMF), in which several constraints are defined to describe the background noise in a generic way. The observed spectrogram is decomposed into four signal contributions: the voiced speech source and three generic types of noise. The speech signal is represented by a source/filter model which captures only voiced speech, and where the filter bases are trained on a database of individual phonemes, resulting in a small dictionary of phoneme envelopes. The three remaining terms represent the background noise as a sum of three different types of noise (smooth noise, impulsive noise and pitched noise), where each type of noise is characterized individually by imposing specific spectro-temporal constraints, based on sparseness and smoothness restrictions. The method was evaluated on the 3rd CHiME Speech Separation and Recognition Challenge development dataset and compared with conventional semi-supervised NMF with sparse activations. Our experiments show that, with a similar number of bases, source/filter modeling of speech in conjunction with the proposed noise constraints produces better separation results than sparse training of speech bases, even though the system is only designed for voiced speech and the results may still not be practical for many applications.

Keywords: Audio source separation; Speech separation; Speech enhancement; Non-negative matrix factorization; Compositional models.

1. Introduction

Speech separation from background noise and other acoustical interferences (a problem often referred to as *speech enhancement*) is one of the most popular lines of research in signal processing. Applications include hands-free communications systems, automatic speech recognition, hearing aids and, in general, every situation where a contaminated speech signal must be restored to its original form. The problem is specially difficult for one-channel mixtures, where spatial information is unavailable as a cue for separating sound sources. Traditionally, speech enhancement has been accomplished by using filter-based algorithms, in which the clean speech spectrum is retrieved based on the estimation of the power spectral density (PSD) of the undesired sound (Boll, 1979; Cohen and Berdugo, 2001; Ephraim and Malah, 1984; Wiener, 1964). More recently, algorithms based on computational auditory scene analysis (CASA) (Bregman, 1994) have been proposed to separate speech without requiring prior knowledge about the interfering sources (Hu and Wang, 2010).

A solution that has gained considerable attention in the last years is the use of model-driven methods, in which speech and noise components are modeled through parametric descriptions that characterize the behavior of each component (Le Roux et al., 2007; Virtanen et al., 2015). The separation process consists then in estimating the parameters of these models, usually by resolving a minimization problem (an example can be found in Le Roux et al. (2007)). Among all the model-driven methods, probably the most popular are those based on *compositional* models, specially due to their easy formulation and fast computation (Virtanen et al., 2015). In compositional models, the spectrogram of each source signal is modeled by a combination of spectral bases, which represent spectral patterns (which may be unknown) from which that source can be constructed. The observed mixed signal can then be expressed as a constructive

* Corresponding author. Tel.: +34 953 648581; fax: +34 953 648508.
E-mail addresses: pcabanas@ujaen.es (P. Cabañas-Molero), damian@ujaen.es (D. Martínez-Muñoz), pvera@ujaen.es (P. Vera-Candeas), fcanadas@ujaen.es (F.J. Cañadas-Quesada), nicolas@ujaen.es (N. Ruiz-Reyes).

http://dx.doi.org/10.1016/j.specom.2015.10.008
0167-6393/

combination of the different basis spectra corresponding to the underlying sources, and the separation is accomplished by decomposing the input spectrogram into these bases and their corresponding gains in each time instant. The success of this model relies on the fact that many common sounds can be approximated as a time-varying combination of repetitive fixed patterns. For this reason, compositional models have been widely applied to music signals, which are typically constructed from repetitive structures (notes, chords) that combine along time with different degrees of intensity. Another reason for the success of these models is the existence of mathematical tools that enable to estimate their parameters with fast converging iterative algorithms, most of them derived from the field of non-negative matrix factorization (NMF) (Lee and Seung, 2001). During the last years, powerful NMF-based algorithms for music analysis or separation have been developed, based either on formulating appropriate signal models (Carabias-Orti et al., 2011; Virtanen and Klapuri, 2006) or imposing constraints to the decomposition method (Cañadas Quesada et al., 2014; Virtanen, 2007).

Recently, some efforts have been made to extend the applicability of NMF-based methods for the analysis and separation of speech signals. Since speech is not as intrinsically repetitive as music, mainly due to the high number of possible pronunciations and intonations, the majority of the methods in the literature need to use large dictionaries of speech and noise patterns, which may be composed by thousands of bases without any particular high level meaning. These dictionaries are usually learned from training material imposing sparsity on the activations, such that at test time, the mixture is factorized keeping the bases fixed and optimizing the activations, also enforcing sparsity or any other appropriate constraint. For instance, in Wilson et al. (2008) a regularized NMF is proposed for speech denoising, where the activations are imposed to preserve the same statistics found during training. In Schmidt and Olsson (2006), a sparse NMF decomposition is used to separate concurrent speakers from a given mixture, based on speaker-dependent dictionaries which are also learned enforcing sparsity. In Weninger et al. (2014), a discriminative training approach with separate bases for analysis and reconstruction is proposed, where the reconstruction bases (trained with material including mixed sources) are optimized to recover the sources with Wiener filtering. Recently, a strategy that has become popular for acquiring basis functions is to use exemplar-based approaches, in which the bases are randomly selected as a subset of the training data, without performing any training. This approach is reported to produce good results for speech separation and recognition (Geiger et al., 2013; Gemmeke et al., 2011), specially when the exemplars cover several time frames. Other methods try to exploit the structure of speech to construct speech bases with a certain high level meaning. For example, a method is proposed in Raj et al. (2011) that employs separate bases for each phoneme, learned from a corpus of individual phonemes. Although the bases trained in this way provide a good separation, the system requires prior knowledge about the location of the phonemes in the recording. In Hurmalainen et al. (2013), speech is modeled using trained spectro-temporal template atoms, such that an atom is trained for each state label of a recognizer. The method described in Virtanen and Cemgil (2009) relies on a considerably different approach. Instead of learning basis vectors for each source, the method trains parameters of prior distributions defined for these basic vectors, following a Bayesian perspective. During separation, the basis vectors can be updated to better approximate the input signal, as long as they fit the learned distributions. Most of the algorithms for speech and noise separation are *supervised*, meaning that both speech and noise bases have to be trained. Recently, there have been an interest to develop robust *semi-supervised* algorithms, where the noise model can be learned online. One example is the work by Mohammadiha et al. (2013), where the priors for noise bases are updated from the data to separate speech and noise with a Bayesian NMF.

The first semi-supervised method in the literature designed to decompose vocal sounds into bases with an explicit higher level meaning is described in Durrieu et al. (2011). In Durrieu et al. (2011), a source/filter signal model is proposed to represent the source of interest, such that, at each frame, the source is assumed to have an excitation part, approximated by combining a dictionary of excitation bases, and a filter part, approximated by combining a dictionary of filter bases. This representation is assumed discriminative enough to allow the separation of the target source from the remaining content, without requiring any training or further constraints. Although the model proposed in Durrieu et al. (2011) is generic and potentially applicable to a wide range of music applications, it is interesting to explore which modifications would be useful for speech and noise separation. Specifically, three important aspects can be observed. First, in speech utterances the speaker produces a higher number of pitches than in music, due to the natural intonation present in common speech. Second, since the number of phonemes is limited, it is possible to define a specific set of spectral filters for each phoneme. Instead of using generic smooth functions as in Durrieu et al. (2011), these filters can be learned from actual phonemes in a previous training stage. Since the filter and source contributions are decoupled in the model, it is possible to characterize each phoneme with a small number of filters. And third, background noise can be characterized imposing certain mathematical restrictions to its bases and gains. For instance, it is known that most real noises exhibit a relatively smooth spectrogram in comparison with the target speech. In this case, if the noise matrix is constrained to be smooth, it will capture the background sound more effectively, thus avoiding the inclusion of speech components. Following this strategy, different types of noises (or even other interferences, such as music) can be jointly captured provided a mathematical restriction describing their behavior, as long as these restrictions are distinguishable from speech. The idea of incorporating constraints to the parameters in addition to source/filter modeling is not new in the context of NMF, and has been applied before for semi-supervised speech/noise separation. In Simsekli et al. (2014), a similar probabilistic non-negative source/filter model is proposed for separating speech from noise, in which the constraints are focused on characterizing the dynamics of speech. The generic framework by Ozerov et al. (2012) also enables implicitly to define sources under a source/filter representation, and to incorporate constraints to the parameters of the model.