Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 69–80

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Application of non-negative spectrogram decomposition with sparsity constraints to single-channel speech enhancement

Kyogu Lee *

Music and Audio Research Group, Graduate School of Convergence Science and Technology, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea

Received 24 April 2013; received in revised form 17 November 2013; accepted 18 November 2013
Available online 27 November 2013

Abstract

We propose an algorithm for single-channel speech enhancement that requires no pre-trained models – neither speech nor noise models – using non-negative spectrogram decomposition with sparsity constraints. To this end, before staring the EM algorithm for spectrogram decomposition, we divide the spectral basis vectors into two disjoint groups – speech and noise groups – and impose sparsity constraints only on those in the speech group as we update the parameters. After the EM algorithm converges, the proposed algorithm successfully separates speech from noise, and no post-processing is required for speech reconstruction. Experiments with various types of real-world noises show that the proposed algorithm achieves performance significantly better than other classical algorithms or comparable to the spectrogram decomposition method using pre-trained noise models.

Keywords: Single-channel speech enhancement; Non-negative spectrogram decomposition; Sparsity constraint; Unsupervised source separation

1. Introduction

The quality and intelligibility of speech are severely degraded in the presence of acoustic noise, and the amount of degradation depends on the type of noise and environment. Speech enhancement, which aims to improve the quality and intelligibility of speech degraded by background noise, is required in many speech applications, including voice communication and automatic speech recognition, to name a few.

Single-channel speech enhancement is a particularly challenging problem, where only a single microphone is used to collect the acoustic signal and noise, and numerous algorithms have been proposed over decades to tackle this problem. Classical algorithms for single-channel speech enhancement algorithms include statistical model-based (Ephraim and Malahd, 1985), Wiener filter (Scalart and Filho, 1996), spectral subtraction (Kamath and Loizou, 2002), and subspace algorithms (Hu and Loizou, 2003). All these algorithms require noise-only excerpts to estimate the noise-related parameters which are later used to enhance the speech with noise.

More recently, a great deal of research effort has been put into non-negative spectrogram decomposition for many audio and music applications, including speech enhancement, due to its ability to learn a dictionary or basis functions to model speech or audio signals. Lauberg et al. proposed a structured NMF algorithm for blind source separation without using training data for individual sources, and applied it to music source separation (Lauberg et al., 2008). Although they did not require pre-trained models for individual sources, the main assumption of "uniqueness" is limited because it is not likely hold in noisy speech, particularly for non-stationary noises.

Joder et al. presented an online semi-supervised algorithm for real-time speech enhancement using Non-negative Matrix Factorization (NMF) (Joder et al., 2012).

* Tel.: +82 318889139; fax: +82 318889148.
E-mail address: kglee@snu.ac.kr

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.11.008

While they proved via experiments that their system performed as well as a supervised NMF algorithm, the major drawback is that it needs to learn speech bases from training data.

Duan et al. also used a non-negative spectrogram decomposition technique to enhance monaural speech signals degraded by non-stationary noises (Duan et al., 2012a,b). Unlike an NMF algorithm proposed by Joder et al. above where they pre-learned speech bases, Duan et al. required noise-only excerpts to learn a dictionary of noise, or spectral basis functions of noise. These pre-trained noise models are then used to separate speech from noise. Using various types of non-stationary noise signals, Duan et al. showed that the spectrogram decomposition method significantly outperformed the classical algorithms.

In this paper, we apply a non-negative spectrogram decomposition technique with sparsity constraints to single-channel speech enhancement. In doing so, we first divide the basis spectra into two disjoint groups of speech and noise, and impose sparsity constraints, by means of entropic prior, only on the basis spectra in the speech group. This is based on the observation that the spectral distribution of a speech signal is in general more *sparse* with conspicuous peaks while that of a noise signal is more broadband, yielding less sparsity.

This can be seen in Fig. 1. At the top is shown a magnitude spectrum of a female speech signal, and that of a keyboard noise signal is displayed at the bottom, respectively. The difference in their spectral distributions is obvious, and the entropy of each distribution is $\mathcal{H} = 0.59$ and $\mathcal{H} = 0.95$, respectively, when a uniform distribution has the entropy $\mathcal{H} = 1.0$.

Therefore, when updating the parameters in the spectrogram decomposition algorithm, we divide the spectral basis functions into two groups – one group for speech and the other for noise – and selectively impose a sparsity constraint only on the spectral bases in the speech group. This will ensure, after convergence, that the spectral bases in the speech group will have low entropy, which better explains the pitched or harmonic signal, and the spectral bases in the noise group will have high entropy, which is more adequate to describe the broadband signal.

The proposed approach has a few advantages. First of all, unlike the above-mentioned algorithms no pre-trained noise models are required because speech-noise separation is done simultaneously in the spectrogram decomposition framework. This eliminates time and labor to obtain the training excerpts and to build the models.

The second advantage of our approach is that no post-processing or classification is needed after the decomposition. Because we pre-label the harmonic and the percussive basis vectors before decomposition begins, and the spectrogram decomposition algorithm works in an unsupervised way, all is needed after the algorithm converges is just regroup the basis vectors according to the corresponding label.

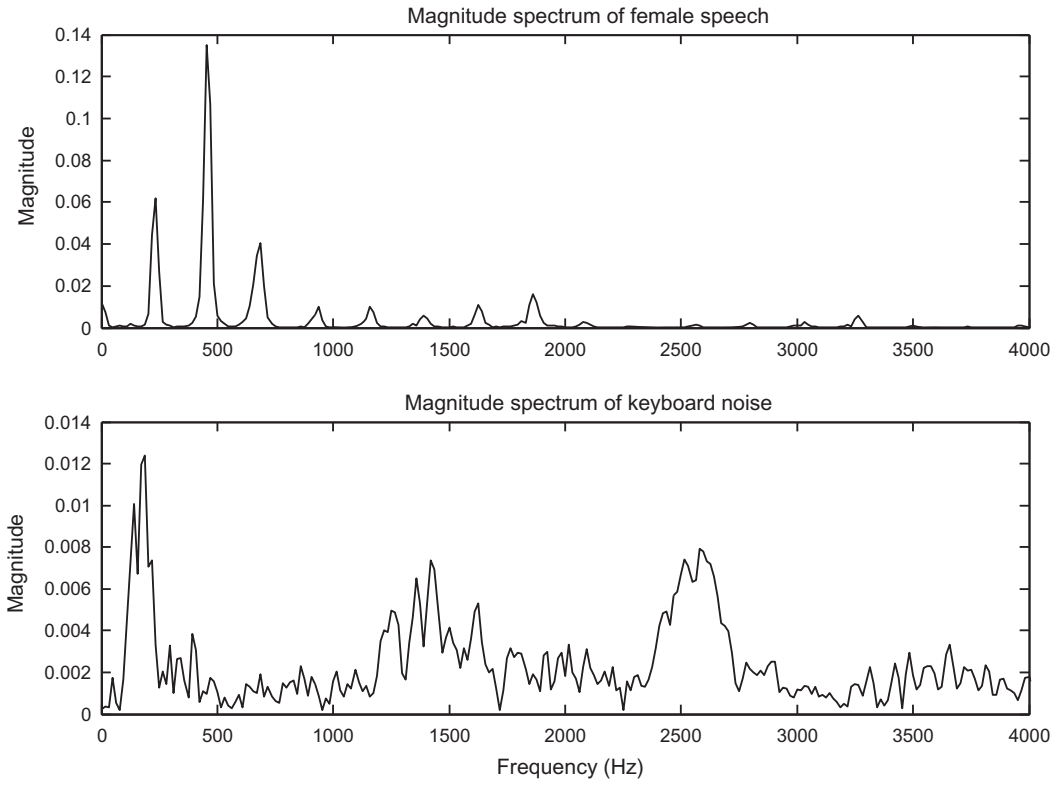

Fig. 1. Magnitude spectrum of (a) a female speech signal ($\mathcal{H} = 0.59$) and (b) a keyboard noise signal ($\mathcal{H} = 0.95$).