Signal Processing 93 (2013) 321–325

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast communication

Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization

F. Mustière [a,b,*], M. Bouchard [b], H. Najaf-Zadeh [a], R. Pichevar [a], L. Thibault [a], H. Saruwatari [c]

[a] Communications Research Centre, 3701 Carling Avenue, Ottawa, Ontario, Canada K2H 8S2
[b] School of Electrical Engineering and Computer Science, 800 King Edward Avenue, Ottawa, Ontario, Canada K1N 6N5
[c] Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-Shi, Nara 630-0192, Japan

ARTICLE INFO

Article history:
Received 29 October 2011
Received in revised form
27 June 2012
Accepted 30 June 2012
Available online 7 July 2012

Keywords:
Speech enhancement
Multichannel
Frequency-domain
MMSE
Spatial cues
Spectral distance

ABSTRACT

It is often very important for multichannel speech enhancement systems, such as hearing aids, to preserve spatial impressions. Usually, this is achieved by first designing a particular speech enhancement algorithm and later or separately constraining the obtained solution to respect spatial cues. Instead, we propose in this paper to conduct the entire system's design via the minimization of statistical spectral distances seen as functions of a real-valued, common gain to be applied to all channels in the frequency-domain. For various spectral distances, we show that the gain derived is expressible in terms of optimal multichannel spectral amplitude estimators (such as the multichannel Minimum Mean Squared Error Spectral Amplitude Estimator, among others). In addition, we report experimental results in complex environments (i.e., including reverberation, interfering talkers, and low signal-to-noise ratio), showing the potential of the proposed methods against recent state-of-the-art multichannel enhancement setups which preserve spatial cues as well.

1. Introduction

The demand for speech enhancement systems is currently very high with applications including cellular phones, teleconferencing, and hearing aids. A popular strategy for noise reduction is to employ an approach based on frequency-domain analysis, modification of spectral amplitudes, and resynthesis in the time-domain. The main advantage of this type of frequency-domain processing is that each step can be carried out very fast; moreover, signal and noise models remain easy to formulate in the transformed domain. Many current-generation frequency-domain enhancers are designed based on several microphones, as it is recognized that multichannel processing has the potential to outperform monaural processing given the increased amount of information available [1]. In the literature, the most popular methods are based on a combination of beamforming, non-optimally merged monaural algorithms (i.e., single-channel methods that are then combined based on heuristic criteria), or Wiener-type multichannel algorithms (e.g. [2–4]). In terms of noise estimation, recent binaural schemes have also been shown to be superior (see [5] for diffuse noise environments).

In many situations where output signal acoustic images must be recreated at multiple sounding devices, it can be crucial to preserve spatial impressions (this is for example the case for hearing aids). Various solutions exist to try and satisfy this constraint, either *a posteriori* or by augmentation of the initial problem. One simple strategy in the context of frequency-domain noise reduction algorithms is to aim for a

* Corresponding author at: Communications Research Centre, 3701 Carling Avenue, Ottawa, Ontario, Canada K2H 8S2.
Tel.: +1 613 562 5800x6190.
E-mail address: mustiere@site.uottawa.ca (F. Mustière).

0165-1684/$ - see front matter Crown Copyright © 2012 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.024

single, real-valued frequency-dependent gain that is applied to all incoming measurements [2]. Using such a zero-phase gain avoids introducing any dispersion or group delay distortion in the signals, while in the binaural context, using a common gain on both measurements ensures that the binaural cues (interaural level and time differences, onsets) are preserved. Currently, the solutions used to preserve spatial cues using such a single real-valued gain are based on non-optimal or heuristic criteria or approximated solutions that combine single-channel estimates (e.g. [2,6–8]). Alternatively, instead of trying to determine a common gain, in a specific class of optimal multichannel frequency-domain algorithms (namely the Multichannel Wiener Filter), some solutions exist ([9–11], and other works cited in [11]) which instead try and constrain complex Wiener gains to minimize additional components in the global cost function which are aimed at reducing interaural time or level differences; however, we see some drawbacks in [9–11]: first the solutions are tailored to a binaural hearing aid configuration; next they require a potentially complex iterative optimization scheme for each incoming frame of noisy samples [11]. Finally, as reported in [11] in an extensive analysis, these solutions may not be robust to multiple-noise-source scenarios.

In this paper, we employ the real-valued common gain strategy, and propose to let the spatial cues preservation constraint dictate the entire multichannel enhancement design. To do this, the problem is simply viewed as the optimization of statistical spectral distances between the noisy spectral components, and the clean components to which the single, real-valued gain has been applied in some arbitrary way. The solutions obtained are therefore presented as unapproximated results using true multichannel criteria shown to require multichannel estimates (as opposed to approximate results only requiring a combination of single-channel estimates based on single-channel criteria). In fact, it is seen that the optimal real-valued common gain expressions obtained make use of a wide variety of multichannel frequency-domain estimators, depending on the spectral distance used and on the assumptions regarding the statistical nature of the speech and noise signals. Accordingly, for completeness we recapitulate some of the most general multichannel estimators available yet (including several non-published solutions) at www.eecs.uottawa.ca/~bouchard/papers/SP2012_Multich_Derivations.pdf, which also includes complete derivations. Some of the proposed algorithms are tested in an experimental setup, with real recordings in complex environments, and some conclusions are formulated.

2. Problem formulation, distance functions and their solutions

2.1. Notation

As in [12], define the kth spectral component $Y(k)$ of a time-domain signal $y(t)$ over the interval $[0;T]$ as

$$Y(k)=\frac{1}{T}\int_0^T y(t)\exp\left(-j\frac{2\pi}{T}kt\right)dt \tag{1}$$

This form corresponds to samples of the continuous time Fourier transform (with an additional scaling factor $1/T$). Focusing on one frequency bin and dropping the subscript k, define the following M-channel model

$$Z_m=H_mS+N_m \tag{2}$$

where $\{N_m\}_{m=1}^M$ and $\{Z_m\}_{m=1}^M$ are the noise and noisy components at each channel. H_m is the transfer function between the fully coherent parts of the target speech (represented by S) and channel m (e.g. in the binaural domain, an HRTF in the case of non-reverberant environments with head shadow effects). We also use the notation $S_m=H_mS$ to denote the target speech as received at microphone m.

Alternatively $\{H_m\}_{m=1}^M$ may represent the frequency ratios between all components and an arbitrarily chosen channel j, so that $H_j=1$ and the signal to estimate is the speech received at channel j. In vector form, we use $\mathbf{z}$, $\mathbf{h}$, and $\mathbf{n}$, with mth element respectively equal to Z_m, H_m, and N_m, to write the single equation

$$\mathbf{z}=\mathbf{h}S+\mathbf{n} \tag{3}$$

2.2. Problem formulation as spectral distance functions optimization

By using a real-valued common gain G to be applied to all channels, for each directional source the gain ratio between channels and the phase difference between channels remains unchanged. In the case of a binaural system, this means that the binaural acoustic cues for localization (interaural level differences and interaural time differences) remain unchanged. A real-value gain G also ensures a constant (and zero) group delay across frequencies, so that no dispersion is introduced in the process.

In contrast with several previous solutions working towards preserving spatial cues with a common gain method, the whole enhancement design is here centered on well-defined multichannel objectives, allowing system designers to have better awareness of the properties of the common gain sought. Based on the above notation, the multichannel criteria are of the form of an expected distance D between a function of the target speech spectral component S and a function of the measurements on which a real-valued gain G has been applied, conditioned on the knowledge of $\mathbf{z}$. The main variable in this distance is G, and the optimal value of G that minimizes the distance D must be found. In the context of speech and signal processing, we can take inspiration from [14,15] that specialize in sensible spectral distance measures, and adapt some of these distances to our context to form the following examples:

$$D_1(G)=\sum_{m=1}^{M}\mathcal{E}\{(|S_m|-G|Z_m|)^2|\mathbf{z}\} \tag{4}$$

$$D_2(G)=\sum_{m=1}^{M}\mathcal{E}\{(\log|S_m|-\log G|Z_m|)^2|\mathbf{z}\} \tag{5}$$

$$D_3(G)=\sum_{m=1}^{M}\mathcal{E}\{|S_m-GZ_m|^2|\mathbf{z}\} \tag{6}$$