Speech Communication 86 (2017) 97–106

Contents lists available at ScienceDirect

Speech Communication

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

Multichannel acoustic echo cancellation exploiting effective fundamental frequency estimation

Laura Romoli[1],*, Stefania Cecchi[1], Francesco Piazza[1]

Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 19 October 2016
Accepted 21 November 2016
Available online 28 November 2016

Keywords:
Fundamental frequency estimation
Multichannel acoustic echo cancellation
Multichannel decorrelation
Multichannel double-talk detection
Multichannel voice activity detection

ABSTRACT

Multichannel teleconferencing systems exploit multichannel acoustic echo cancellers to weaken the echo replicas due to the acoustic coupling among loudspeakers and microphones. Many issues have to be dealt with to ensure an effective echo cancellation in such systems, including the reduction of the interchannel coherence among channels, the detection of an active remote human speaker, and the identification of the well-known double-talk scenario. In this paper, a comprehensive solution for dealing with the aforementioned aspects is discussed. The core of this solution is the estimation of the fundamental frequency of the audio signal. It is exploited for weakening the linear relation among channels and for tracking the presence of an active human speaker both in the remote room and in the local room of a teleconferencing system. The computational complexity of the presented approach is also reported to support its feasibility in a real scenario. Then, its real-time implementation is presented and validated on the NU-Tech framework, showing its fundamental frequency tracking capability and echo reduction performance both in simulated and real scenarios.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Multiple microphones and loudspeakers are exploited in audio/video teleconferencing systems to improve the listening experience of its participants placing them inside the acoustic scenario (Benesty et al., 1998; Cecchi et al., 2016; Romoli et al., 2015, 2014b, 2014a). However, the presence of multiple channels implies the need for multichannel acoustic echo cancellers (MAECs) to reduce the echo replicas on the microphone signals arising from the coupling between each loudspeaker and each microphone. In the context of acoustic echo cancellation, several issues have to be dealt with. More specifically, an effective identification of multiple echo paths requires the reduction of the correlation among channels since convergence problems could arise from the linear relationship existing among them. This aspect is well-known in the literature as “non-uniqueness” problem (Benesty et al., 1998). Then, remote human speaker activity tracking and the identification of the well-known double-talk scenario are other important aspects to deal with, based on the discrimination of vocal frames, silence segments, and noisy periods (Romoli et al., 2015). Indeed, voice activity detection (VAD) procedures and double-talk detection (DTD) algorithms assume an important role in practical implementation of such systems since filter coefficients update is typically suppressed when there is no active human speaker in the remote room and/or when there are active human speakers both in the remote room and in the local room (Benesty et al., 2001; Tashev, 2009). A comprehensive work combining all these aspects in a unique solution is not present in the state of the art up to authors’ knowledge, while multiple approaches focusing on each aspect have been discussed in the literature.

Regarding channel decorrelation, some approaches, previously focused on the stereophonic scenario, have been generalized to the multichannel case, to reduce the interchannel coherence and thus, to improve the performance of multichannel systems (Cecchi et al., 2012). Among these, the phase modulation approach for surround sound systems (Herre et al., 2007) and the half-wave rectifier distortion (Benesty et al., 1998) were proposed. Recently, a novel solution for multichannel decorrelation that is suitable for both speech and music signals has been proposed. It is based on fundamental frequency estimation and removal in order to exploit psychoacoustic criteria and to avoid audio quality worsening (Romoli et al., 2014b, 2014a; Cecchi et al., 2011; Romoli et al., 2012, 2010).

Regarding VAD procedures, several approaches have been proposed in the literature as presented in Ma and Nishihara (2013) and Babu and Vanathi (2009) based on signal energy and

* Corresponding author.
E-mail addresses: l.romoli@univpm.it (L. Romoli), stefania.cecchi@gmail.com (S. Cecchi).

[1] Fax Number: +390712204453

http://dx.doi.org/10.1016/j.specom.2016.11.009
0167-6393/© 2016 Elsevier B.V. All rights reserved.

zero-crossing rate (Atal and Rabiner, 1976) and on more robust acoustic features (Davis et al., 2006; Shuyin et al., 2009; Marzinzik and Kollmeier, 2002) or their combination with pitch analysis in the presence of higher levels of noise (Atal and Rabiner, 1976). Recently, a new method for detecting voiced segments of the far-end signal has been proposed exploiting the fundamental frequency tracking through second-order adaptive notch filters in a multichannel scenario (Romoli et al., 2015).

Regarding DTD algorithms, multichannel procedures are typically based on cross-correlation (Benesty and Gansler, 2002; Iqbal et al., 2009). More specifically, the approach described in Benesty and Gansler (2002) is based on the cross-correlation between the far-end signal and the microphone signal while the technique described in Iqbal et al. (2009) is based on the cross-correlation between the residual echo signal and the microphone signal. In both cases, the detection is made according to the value assumed by a decision variable with respect to a predefined threshold. Recently, a new DTD method has been proposed based on human speaker activity tracking through second-order adaptive notch filters (Cecchi et al., 2016).

In this paper, starting from the obtained promising results (Cecchi et al., 2016; Romoli et al., 2015, 2014b), a low-complexity comprehensive solution for multichannel acoustic echo cancellation based on fundamental frequency estimation is investigated. Although the exploitation of pitch analysis for active human speaker detection is a known topic in the literature, the proposed approaches are intended to be combined in an exhaustive solution for MAEC systems, since decorrelation, VAD, and DTD procedures are performed based on the same variables. Several results are reported to show the performance of pitch tracking and echo reduction. Moreover, the real-time implementation of the solution is discussed with the aim of applying the approach in real scenarios. The computation of the number of multiplications required for multichannel decorrelation, VAD procedure, and DTD algorithm is also reported in order to underline the advantage of using the same parameters to face different problems. Indeed, in this way, a self contained solution can be derived.

The paper is organized as follows. The role of fundamental frequency estimation in MAEC system is presented in Section 2 discussing the exploitation of the fundamental frequency estimation for decorrelation (Section 2.2), VAD algorithm (Section 2.3), and DTD procedure (Section 2.4). Then, the computational complexity of the aforementioned procedures (Section 4) is reported to prove its feasibility in real scenarios. The real-time implementation of the approach and its validation are reported in Section 5 in terms of effective tracking of fundamental frequency for remote and local human speakers and of satisfactory echo reduction. Finally, concluding remarks are presented in Section 6.

2. The role of fundamental frequency estimation

The proposed solution for multichannel acoustic echo cancellation that includes multichannel decorrelation, voice activity detection, and double-talk detection is based on fundamental frequency estimation. The adopted scenario is reported in Fig. 1, with L loudspeakers and P microphones. The remote signal is the residual error $r_l(n)$ (being $l = 1, \dots, L$ and n the time index) and it is processed by the decorrelation block providing the decorrelated signals $x_l(n)$ and the estimated frequency $f_l^*(n)$ (Romoli et al., 2014b, 2014a). Then, the estimated echo signal $y_p(n)$ ($p = 1, \dots, P$) is calculated by the system identification procedure and the error signal $e_p(n)$ is computed by subtracting it from the microphone signal $d_p(n)$. The system identification procedure is controlled by the VAD and DTD blocks, that can enable or disable the filter coefficients update. The former makes its decision according to the remote fundamental frequency $f_l^*(n)$, whereas the latter detects the double-talk situation considering both the remote fundamental frequency $f_l^*(n)$ and the local fundamental frequency $\varphi_p^*(n)$. In the following sections, a brief review of the algorithm adopted for estimating and removing the fundamental frequency is reported as previously discussed in Cecchi et al. (2016) and Romoli et al. (2015, 2014b). Then, its application to multichannel decorrelation, VAD procedure, and DTD algorithm is presented.

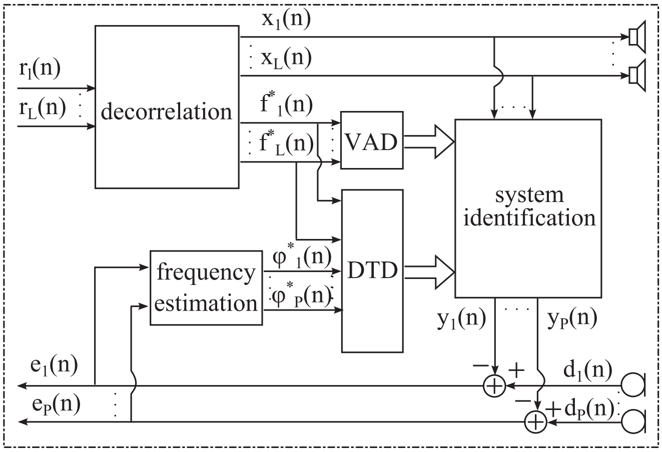

Fig. 1. Block diagram of multichannel audio teleconferencing system.

2.1. Fundamental frequency estimation and removal

A block diagram summarizing the main steps for fundamental frequency estimation and removal is reported in the upper branch of Fig. 2. The algorithm is reported in this section for M input signals but this procedure will be exploited in the following sections for estimating and refining the fundamental frequency of both the L loudspeaker signals ($M = L$) and the P microphones signals ($M = P$).

The M input signals $s_m(n)$ ($m = 1, \dots, M$) are decimated by a factor D to improve spectral resolution at low frequencies and then filtered by pre-processing second-order shelving filters H_{pre} with gain G_s and cut-off frequency f_{cuts} (Zölzer, 2002). Assuming that the fundamental frequency is typically bounded between $f_{\text{inf}} = 60$ Hz and $f_{\text{sup}} = 600$ Hz (Cecchi et al., 2016), G_s and f_{cuts} have to be chosen to apply a spectral decay that ensures a correct tracking and reduces possible mighty harmonic components. Then, the M signals are filtered by the following adaptive notch filters:

$$H_m(z, \tilde{n}) = \frac{1 + 2k_m(\tilde{n})z^{-1} + z^{-2}}{1 + k_m(\tilde{n})[1 + \alpha_m(\tilde{n})]z^{-1} + \alpha_m(\tilde{n})z^{-2}}, \tag{1}$$

where $\tilde{n}$ is the index of the decimated time instant, the contraction factor $\alpha_m(\tilde{n})$ controls the m-th filter bandwidth, and $k_m(\tilde{n})$ is an adaptive coefficient related to the fundamental frequency to be tracked (Cecchi et al., 2012). More specifically, the coefficient $k_m(\tilde{n})$ is bounded to the range $(-1, +1)$ according to the following sigmoid function:

$$k_m(\tilde{n}) = \frac{2}{1 + e^{-g_m(\tilde{n})}} - 1, \tag{2}$$

where $g_m(\tilde{n}) \in R$ has to be adapted in order to minimize the time average of the notch filter output as fully described in Cecchi et al. (2012) and Romoli et al. (2012). In this way, the signals $x_m(\tilde{n})$ without the fundamental frequency are obtained. The signals $x_m(n)$ are then derived applying post-processing second-order shelving filters H_{post} with the same gain G_s and cut-off frequency f_{cuts} and upsampling by a factor D. Differently, the mth fundamental frequency is a function of the adaptive parameter $k_m(\tilde{n})$ as given

Download English Version:

https://daneshyari.com/en/article/4977839

Download Persian Version:

https://daneshyari.com/article/4977839

Daneshyari.com