Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 527–529
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Editorial

Special issue on Speech Enhancement

1. Introduction

Speech quality is severely degraded in the presence of acoustic noise and the degradation depends largely on the characteristics of the noise and environment. Speech enhancement algorithms which improve the quality of speech and reduce or eliminate the acoustic noise are highly desirable in voice communication systems (e.g., cellular phones). Despite the progress made in reducing stationary noise, there remain many challenges ahead. Depending on the number of microphones used to collect the acoustic signal and noise, different speech enhancement algorithms have been proposed over the years. This Special Issue provides a balanced mix of papers for single-channel (single microphone) enhancement and multiple-channel (multiple microphones) enhancement of speech.

Enhancement of speech in realistic listening environments has been one of the biggest challenges, particularly when only the noisy signal (single microphone) is available. There are a number of aspects in signal enhancement that makes this a challenging task, and this Special Issue addresses some of these aspects. The success of speech enhancement algorithms lies in the choices made in the different processing stages of enhancement. The first stage consists of the analysis stage in which the signal is transformed in some domain via a non-unitary or unitary transformation (e.g., Fourier, Karhunen-Loeve transform). The second stage, which is the heart of most algorithms, consists of the suppression stage in which the transformed signal is multiplied by a gain function designed to attenuate the acoustic noise while preserving the speech signal. Often times, the gain function needs access to information about the noise spectrum density, and continuous estimation/update of the noise spectrum is critical particularly in non-stationary environments wherein the background noise is constantly changing. The last stage is the synthesis stage in which the modified signal is transformed back to the time-domain. The first set of papers focus on single-channel enhancement algorithms and provide various proposals for all three processing stages of enhancement and includes novel methods for constructing gain functions, new noise estimation algorithms suitable for highly-non-stationary environments, and a new analysis/synthesis filter-bank. The final paper provides a comprehensive evaluation and comparison of speech enhancement algorithms, and presents a new noisy speech corpus which can be used to facilitate comparisons between algorithms proposed in various research labs.

Generally, the performance of single-microphone techniques is limited, and in many applications, such as hands-free telephony, voice-controlled systems, teleconferencing, source separation and hearing aids, a substantial gain in performance can be obtained by using multi-microphone systems. Multi-microphone systems enable high-quality, hands-free communication in reverberant and noisy environments due to the spatial filtering capability of suppressing interfering signals coming from undesired directions. Adaptive beamforming for speech signals requires particular consideration of problems that are specific to speech signals and to the acoustic environment. The speech signal is wideband, highly non-stationary, and has a very wide dynamic range. An acoustic enclosure is usually modeled as a filter with very long impulse response due to multiple reflections from the room walls. In a typical office, the length of the filters may reach several thousands of taps. Furthermore, the impulse response is often time-varying due to speaker and objects movements. The second set of papers focus on multi-channel enhancement algorithms and include novel methods for adaptive beamforming, multi-channel post-filtering, simultaneous echo cancellation and noise reduction, and speech source separation from convolutive mixtures.

2. Summary of papers

This Special Issue contains a selection of papers that use a variety of techniques to address the enhancement of speech collected either by a single microphone or multiple microphones. Accordingly, we divide the papers into two sections.

2.1. Single-channel enhancement algorithms

The first paper by Erkelens *et al.* proposes new gain functions for spectral noise suppression. Unlike most gain functions proposed in the literature, the new gain

0167-6393/$ - see front matter
doi:10.1016/j.specom.2007.05.002

functions make no assumptions about the underlying speech distribution (e.g., Gaussian, Laplacian, etc). Instead, the gain functions are derived (trained) from the data at hand by minimizing relevant error criteria using a general optimization procedure. This data-driven approach was shown to perform comparably well with state-of-the art algorithms.

The second paper by You et al. revisits the Kalman filtering approach for speech enhancement and considers a subband filtering scheme that incorporates the auditory masking properties of the human ear. A novel approach is proposed for incorporating the masking threshold in subband Kalman filtering, whereby the estimate of the noise variance used in the Kalman filtering process in each subband is modified according to the masking threshold. The use of Kalman filtering in the subband rather than the full-band domain leads to considerable complexity reduction and performance improvement.

The third paper by Lin et al. addresses the problem of noise estimation in highly non-stationary environments. The authors propose a noise estimation method which utilizes not only spectral information in each subband, but also pitch information in voiced speech segments and statistical properties of durations of unvoiced speech in each frame to estimate the probability of speech presence. The proposed noise estimation method was shown to reduce the adaptation time to as short as 0.3 s for noise bursts without affecting the accuracy of the noise estimation algorithm and without suppressing speech components. Objective measures indicated that the proposed algorithm is 4–6 times faster than other noise-estimation algorithms.

The paper by Lollmann and Vary addresses the choice of filter-bank used in speech enhancement in terms of speech quality, computational complexity and signal delay. A versatile filter-bank for adaptive subband filtering is proposed, which achieves a significantly lower algorithmic signal delay than commonly used analysis-synthesis filter-banks. It performs time-domain filtering with coefficients adapted in the uniform or non-uniform frequency-domain. Results indicated that the proposed filter-bank approach can achieve a very low signal delay and reduced computational complexity with almost no compromise in the perceived speech quality.

The last paper by Hu and Loizou provides a comprehensive subjective evaluation of 13 enhancement techniques encompassing four different classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. The subjective evaluations were performed according to the ITU-T P.835 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion and overall quality. This paper also reports on the development of a noisy speech corpus suitable for evaluation of speech enhancement algorithms. Such a corpus (available at http://www.utdallas.edu/~loizou/speech/noizeus/) can be used to facilitate meaningful comparisons between algorithms proposed in various research labs.

2.2. Multi-channel and multi-modal enhancement algorithms

The first paper by Reuven, Gannot and Cohen presents a performance evaluation of a recently proposed adaptive beamformer, namely the Dual source Transfer-Function Generalized Sidelobe Canceller (DTF-GSC). The DTF-GSC is useful for enhancing speech signals received by an array of microphones in noisy and reverberant environments, when competing speech is present. The authors demonstrate the applicability of the DTF-GSC in some representative reverberant and non-reverberant environments under various noise field conditions, and evaluate the performance based on the power spectral density (PSD) deviation imposed on the desired signal at the beamformer output, the achievable noise reduction, and the interference reduction.

The second paper by Reuven, Gannot and Cohen presents an echo transfer function generalized sidelobe canceller (ETF-GSC), for joint echo cancellation and noise reduction in a reverberant environment. The proposed scheme consists of a primary transfer-function generalized sidelobe canceller (TF-GSC), which is designed for noise suppression, and a secondary modified TF-GSC, which is designed for echo cancellation. The ETF-GSC decouples the noise reduction and echo cancellation tasks, and hence overcomes many of the problems encountered in cascade application of acoustic echo cancellation and TF-GSC.

The third paper by Doclo et al. proposes a multi-channel frequency-domain criterion for a speech-distortion weighted multi-channel Wiener filter (SDW-MWF), which facilitates a trade-off between noise reduction and speech distortion. The authors derive adaptive frequency-domain algorithms for implementing the SDW-MWF with various step-size matrices. They investigate the noise reduction performance, the robustness against signal model errors, and the tracking performance of the proposed algorithms by performing experiments with a small-sized microphone array in a hearing aid.

The fourth paper by Lefkimmiatis and Maragos presents a general method for designing multichannel post-filters under various speech enhancement criteria, including mean-square error (MSE), MSE of the short-time spectral amplitude (STSA), and MSE of log-spectral amplitude (log-STSA). The authors propose robust estimators for the speech and noise power spectral densities by taking into account the noise reduction obtained by a Minimum Variance Distortionless Response (MVDR) beamformer. Experimental results indicate that the proposed technique improves performance compared to other methods in terms of five different objective speech quality measures.

The last paper by Rivet, Girin and Jutten proposes an audio-visual speech source separation method, which combines visual information with source separation techniques to improve the extraction of the speech source of interest from convolutive mixtures. Visual information such as

the temporal dynamics of lip movements are employed for voice activity detection, and the sparseness of speech signals is exploited for a new geometric source separation method. The proposed audiovisual method is shown to be efficient even in the difficult case of convolutive mixtures and highly non-stationary competing sources.

Acknowledgements

We would like to thank all the authors who have contributed to this Special Issue and to the reviewers for their comments and suggestions. We also gratefully acknowledge the support of Mary Lynn van Dijk of the Editorial Production Department.

Philipos C. Loizou
University of Texas-Dallas, USA
E-mail address: loizou@utdallas.edu

Israel Cohen
Technion – IIT, Israel
E-mail address: icohen@ee.technion.ac.il

Sharon Gannot
Bar-Ilan University, Israel
E-mail address: gannot@eng.biu.ac.il

Kuldip Paliwal
Griffith University, Australia
E-mail address: K.Paliwal@griffith.edu.au