

Real-time Prototype for Integration of Blind Source Extraction and Robust Automatic Speech Recognition

Francesco Nesta, Marco Matassoni, HariKrishna Maganti

Fondazione Bruno Kessler-Irst
 via Sommarive 18, 38123 Trento, Italy
 {nesta,matassoni,maganti}@fbk.eu

Abstract

This demo presents a real-time prototype for automatic blind source extraction and speech recognition in presence of multiple interfering noise sources. Binaural recorded mixtures are processed by a combined Blind/Semi-Blind Source Separation algorithm in order to obtain an estimation of the target signal. The recovered target signal is segmented and used as input to a real-time automatic speech recognition (ASR) system. Further, to improve the recognition performance, noise robust features based on Gammatone filters are used. The demo utilizes the data provided for the CHiME Pascal speech separation and recognition challenge and also real-time mixtures recorded on-site. Users will be able to listen to the recovered target signal and compare it with the original mixture and ASR output.

Index Terms: blind source separation, speech enhancement, robust speech recognition

1. Introduction

Ubiquitous computing aims at facilitating the user to communicate and interact naturally with applications; as such, speech is an appealing modality and the human-machine interaction using automatic speech processing technologies is a diversified research area, which has been investigated actively [1]. Speech acquisition, processing and recognition in a non-ideal acoustic environments are complex tasks due to presence of noise, reverberation and interfering speakers.

Recent achievements in the field of Blind Source Separation (BSS) have shown that binaural mixtures can be successfully processed by BSS methods in order to remove diffuse background noise from a given source of interest [2]. To date, ASR and BSS research areas have been investigated independently. With the increasing abilities of approaches, it is timely to integrate the combination of BSS and ASR in real environments, to validate the potential advantages that BSS can bring for the recognition task, particularly in distant-talking scenarios.

In particular, CHiME is a recent speech corpus designed for investigating robust speech processing and recognition to compare the achievements obtained in both speech enhancement and recognition communities [3]. The recorded data includes background recordings from a head simulator positioned in a domestic setting as well as binaural impulse responses collected in the same environment. By means of the measured responses, utterances from the Grid corpus have been added to this setting and mixed with the background noise to produce controlled and natural audio data [4]. Note that the noise in CHiME can be generated by any competing sources (e.g. other speakers, tv, radio, home appliances, etc.) and then it is in general highly non-stationary.

In this demo, a real-time prototype to process the CHiME challenge data is presented. A Blind Source Extraction (BSE) system, based on the combination of Blind and Semi-Blind source separation [5][6] estimates the noise and target spectra which is later used to control the coefficients of a Wiener filter. This BSE architecture allows a high suppression of the noise while maintaining the quality of the target source signal at an acceptable level. The processed output is then used as input to a real-time ASR system which uses robust features to further improve the recognition performance.

2. System Architecture

The hardware set up of the demo is described in figure 1. A loudspeaker, immersed in noise due to multiple interfering sources, plays a sequence of utterances which has to be recognized by the ASR system. The signals acquired by two microphones are sampled at 16kHz through an external USB audio card which is connected to a laptop. The recorded mixtures are then processed and the recovered target signal is reproduced in the headphones.

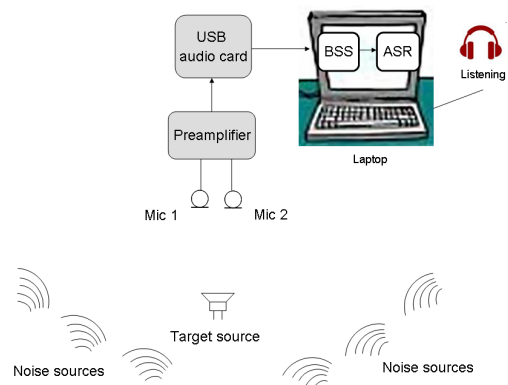


Figure 1: Demo setup: the target source is reproduced by a loud speaker and recorded with 2 microphones, the system processes the two channel input providing the separated source and the corresponding ASR result.

2.1. Blind Source Extraction

The global architecture of the BSE is depicted in figure 2 and described as follows. The sampled time-domain signals are transformed in a discrete time-frequency representation applying a Short-Time Fourier Transform (STFT) with overlapped Hanning windows (e.g. 75% of overlapping) in order to ob-

tain frames with a certain degree of continuity in time. Since the accuracy of ICA depends on the amount of observed data, STFT time-series are obtained by grouping the estimated STFT frames in different segments. A simplified implementation of the Recursively-Regularized ICA in frequency-domain [5] is applied to all the batch segments in order to have an estimate of the mixing parameters $\mathbf{H}(k)$. The estimated mixing matrices $\mathbf{H}(k)$ are transformed in monodimensional observations of the acoustic propagation as in [7][8] and the directional property of the acoustic propagation is estimated through the GSCT function. A Semi Blind Source Separation (based on the same paradigm discussed in [6]) algorithm, which uses as a prior the estimated target mixing parameters, is adopted in order to have an estimation of the noise and target spectra. The estimated spectra is used to compute the coefficient of a Wiener filter in order to get the spatial images of the target source at microphones. The enhanced signals are then beamformed according to the source direction and fed to the ASR system.

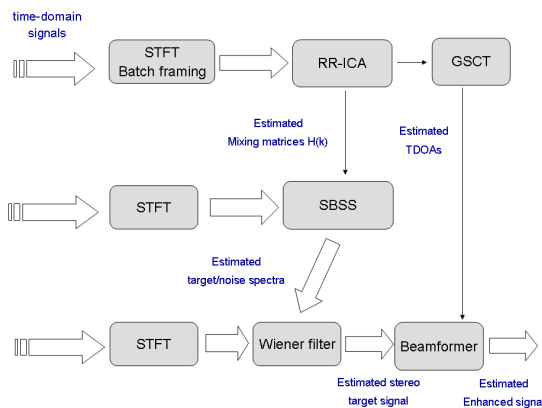


Figure 2: BSE system architecture

2.2. Robust ASR

The ASR system uses whole-word HMMs with topology described in [9], trained with the reverberated Grid training data. Speaker-dependent models are derived from an advanced features extraction module, based on gammatone analysis and spectral modulation features [10]. Such auditory features are inspired by some properties of the human auditory system and have been recently applied to ASR system to enhance their robustness. Model adaptation is also applied to account for the mismatch between the training and test models. More details on the baseline recognizer and on the proposed strategies can be found in [11]. The small-vocabulary task is characterized by a dictionary of 51 words and performance is measured as word accuracy related to (only) two keywords for utterance. In the Figure 3 the results obtained with the baseline recognizer [3] and with the proposed strategies are compared.

3. Conclusions

This paper presents a real-time demo in which an integrated system comprising blind source enhancement and speech recognition functioning in presence of multiple interfering noise sources is demonstrated. Stereo recordings are processed by a combined Blind/Semi-Blind Source Separation algorithm that provides an estimation of the desired speech reducing the unwanted noise sources. A real-time recognizer, based on suitable robust acoustic features, decodes the resulting segmented sig-

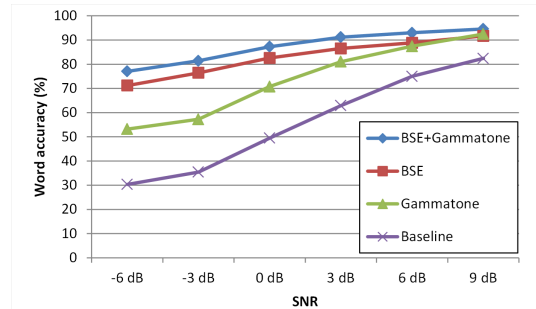


Figure 3: Word accuracy of baseline and proposed methods.

nal.

The setup is similar to the CHiME Pascal speech separation and recognition challenge, considering that this real-time prototype requires a voice activity detection (VAD) stage; based on energy and direction of arrival information the recognizer processes only speech chunks in which the desired speaker is active. As a result, expected performance of the real-time system depends also on the capability of this component, not taken into account in the CHiME challenge.

Another interesting feature of the proposed prototype is obviously the real-time implementation, since it is required to present the output synchronously with the input (apart from a fixed processing delay).

4. References

- [1] G. Abowd and et al., "Living laboratories: The future computing environments group at the georgia institute of technology," in *Proc. Conf. Human Factors in Comp. Sys. (CHI)*, 2000.
- [2] Y. Takahashi and et al., "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Tran. on Audio, Speech and Lang. Proc.*, vol. 17, no. 4, pp. 650–664, May 2009.
- [3] H. Christensen, J. Barker, N. Ma, and P. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Proceedings of Interspeech*, Makuhari, Japan, 2010.
- [4] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.
- [5] F. Nesta, P. Svaizer, and M. Omologo, "Convolutional bss of short mixtures by ica recursively regularized across frequencies," *IEEE Tran. on Audio, Speech and Lang. Proc.*, vol. 19, no. 3, pp. 624–639, 2011.
- [6] F. Nesta, T. Wada, and B.-H. Juang, "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE Tran. on Audio, Speech and Lang. Proc.*, vol. 19, no. 3, pp. 583–599, 2011.
- [7] F. Nesta and M. Omologo, "Approximated kernel density estimation for multiple TDOA detection," to appear in the proceedings of ICASSP 2011.
- [8] —, "Generalized state coherence transform for multidimensional localization of multiple sources," in *Proc. of WASPAA*, October 2009.
- [9] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.
- [10] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 570–573.
- [11] F. Nesta and M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *Proc. of CHiME Workshop*, Florence, Italy, September 2011.