Model-driven Speech Enhancement for Multisource Reverberant Environment: Signal Separation Evaluation Campaign (SiSEC) 2011

Pejman Mowlaee†, Rahim Saeidi
pejman.mowlaee@rub.de, rahim.saeidi@let.ru.nl *

†Institute of Communication Acoustics (IKA),Ruhr-Universität Bochum (RUB),
*Centre for Language and Speech Technology, Radboud University Nijmegen

Abstract

We present a low complexity speech enhancement technique for real-life multisource environment. The idea is to incorporate the speaker model information for enhancing the target signal corrupted in non-stationary noise in a reverberant scenario. Most of the current speech enhancement methods work based on stationary assumption of background noise and therefore are limited as in practice the interfering noise signal is time-varying and unpredictable. The idea of incorporating speaker model into a speech enhancement framework, helps to improve the limited performance of noise-tracking based speech enhancement methods under unpredictable or non-stationary noise scenarios and provides enhanced speech estimate of the single-channel speech signal by effectively rejecting other interfering sources. According to experimental results on SiSEC, we observed that the proposed approach is successful in rejecting the interference signal in the noisy input and providing an enhanced output signal.

Proposed Method

Posing the problem into a model-driven one, the problem to be solved is: given the noisy observation, find the speech estimate which best explains the observed signal in an optimality criterion. We solve the problem under the constraint that the speech estimates are to be selected from speaker codebook used to model the characteristic of the target source and learned by training codebook on the available training set. For taking into account the effect of the distortion channel, we train the speaker models on the channel distorted speech signals, which *a priori* available for each speaker in the dataset [4].

The proposed approach consists of three steps: (1) noise estimation; (2) identifying unreliable components, and (3) enhancement stage.

* The work of Pejman Mowlaee was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699. The work of Rahim Saeidi was funded by the European Community's Seventh Framework Programme (FP7 2007-2013) under grant agreement no. 238803.

Noise estimation We assume that in the first five frames of the noisy signal no speech component exist. We use these frames to obtain an initialized estimate for noise amplitude denoted by $|\hat{N}[k]|$. For noise estimation we use improved minima controlled recursive averaging (IMCRA) proposed in [1].

Identifying unreliable components At spectral regions with low signal-to-noise ratio, the noise spectrum masks the speech spectral peaks that leads to wrong codebook inference. The codevectors selected for these unreliable regions are not representing the speech spectrum but the noise spectrum. To estimate the missing spectral components of target speech, we employ the negative energy criterion that is helpful in the sense that it leads to a better decision making in the codebook inference stage by rejecting those candidate codewords which violate the bounding inequality. A binary mask is produced according to the negative SNR energy.

Enhancement Stage The binary mask prouced in previous stage is used to filter mixed signal. The filtered mixed signal is then given to the speaker model to find the matching codeword. The found speech estimate together with noise estimate are used to build another binary mask to provide enhanced speech. This enhanced speech estimate is then used together with the noise estimate to produce the final enhanced speech based on Wiener filtering.

Speech separation results

As processing strategy, we process each mixture (isolated sentence) alone. The proposed method was applied to the development data and the test data (24+24 utterances) and the enhanced wave files can be found at URL [6]. The system info was as follow: RAM: 4.00 GB, CPU: Intel(R) Core(TM) i5 with 3.2 GHz. The averaged running time for the algorithm is 1.84 × RT.

We evaluate the separation performance of the proposed method in a reverberant adverse noisy scenario as described in CHIME dataset [4] and [5]. We report signal-to-noise ratio (SNR) and segmental signal and segmental SNR:

- SNR: On average, the proposed method achived 3.20 dB SNR compared to -0.98 of noisy input, 4.95 of ideal binary mask, -0.66 by exampler-based [3] and 1.23 by Koldovsky et al. [2].
- SSNR: The proposed method achived 1.60 dB SSNR compared to -1.51 of noisy input and 3.15 provided by ideal binary mask.

References

1. Cohen, I. *Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging*, IEEE Transactions on Speech and Audio Processing, pp.466–475, vol. 11, no. 5. (2003)

2. Z. Koldovsky, J. Malek, J. Nouza,and M. Balik, *CHiME Data Separation Based on Target Signal Cancellation and Noise Masking,* Proc. 12th Annual Conference of the International Speech Communication Association, pp.47–50, (2011)
3. H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen, and K. J. Palomaki *Uncertainty measures for improving exemplar-based source separation,* Proc. 12th Annual Conference of the International Speech Communication Association, pp.53–57, (2011)
4. H. Christensen, J. Barker, N. Ma, and P. Green, *The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments,* Proc. Interspeechg (2010)
5. The third community-based Signal Separation Evaluation Campaign (SiSEC 2011), `http://sisec.wiki.irisa.fr/tiki-index.php`
6. `cs.joensuu.fi/pages/saeidi/Sisec2011_wavFiles.tar.gz`