Coupled Dictionary-based Speech Enhancement with Adaptively Learned Atoms for the CHiME-3 Challenge

Deepak Baby†  Tuomas Virtanen†  Hugo Vanhamme†

†Department ESAT, KU Leuven, Belgium
‡Department of Signal Processing, Tampere University of Technology, Finland
{Deepak.Baby, Hugo.Vanhamme}@esat.kuleuven.be, Tuomas.Virtanen@tut.fi

The CHiME-3 challenge [1] targets the performance of an automatic speech recognition (ASR) setting in a real world, commercially motivated scenario where the recordings are obtained using a tablet fitted with a six-channel microphone array. This paper investigates the previously proposed exemplar-based speech-enhancement technique using coupled dictionaries [2] as a front-end for various ASR back-ends on the CHiME-3 dataset. In addition, this work extends this approach by adding an adaptive dictionary where the input and the coupled output atoms are learned online from the test data. These adaptive dictionaries are particularly useful to model the unseen noise present in the test data that are not available in the training set. In our framework, the six-channel data is converted to a single-channel signal using a beamformer, enhanced using the exemplar-based technique and fed to different ASR decoders that use GMM, DNN and CNN-DNN-based ASR back-ends, respectively. Since it is found that the CNN-DNN-based setting significantly outperforms the other two back-ends, the rest of the evaluations are reported using this back-end.

Experiments using the enhanced speech using the coupled dictionary-based speech enhancement without adaptive dictionaries gave WERs of 24.4%, 24.9% and 27.2% for the Mel-DFT, DFT-DFT and the MS-DFT settings, respectively. These results reveal that the exemplar-based techniques can be effectively used to further improve the ASR performance. It is also observed that the modulation spectrogram features are more sensitive to variations in the environments when compared to the Mel and DFT features.

The evaluations using the proposed adaptive dictionary learning yielded WERs of 23.1%, 22.8% and 25.1% for the Mel-DFT, DFT-DFT and the MS-DFT settings, respectively. It can thus be seen that the proposed adaptive learning of coupled atoms can yield a better speech and noise separation and further introduce WER improvements. Noticeably, the inclusion of the adaptive atoms is found to benefit the settings with the higher dimensional features more when compared to the lower dimensional Mel features. We assume that there is a higher risk of the adaptive atoms to model speech as well when used with lower dimensional features.

To conclude, this work presented an evaluation of a highly realistic ASR task by combining an established beamforming technique with a state-of-the-art single channel speech enhancement setting and a CNN-DNN-based ASR setting. The best WER thus obtained in this work is with the DFT-DFT setting to-gether with the adaptive dictionaries which yielded an absolute WER improvement of 4.2% (15.6% relative) over the baseline setting.

1. Acknowledgements

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (IN-SPIRE).
2. References


