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

Deepak Baby[⋆] *Tuomas Virtanen*[†] *Hugo Van hamme*[⋆]

[⋆]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 acoustic modelling.

A schematic outline of the investigated setting is shown in Figure 1. In this setting, the 6-channel noisy speech is converted to a single channel data using a SRP-PHAT pseudo-spectrum-based beamformer [3]. This is then enhanced using the coupled dictionary-based speech enhancement described in [2]. Here, the noisy speech is decomposed as a weighted sum of atoms in an input dictionary containing exemplars sampled from a domain of choice, and the resulting weights are applied to a coupled output dictionary containing exemplars sampled in the short-time Fourier transform (STFT) domain to directly obtain the speech and noise estimates for speech enhancement. The use of exemplars is motivated from the episodic memory nature present in the human auditory processing and has been successfully used to separate speech from noise [4]. Finally, the enhanced speech data is fed to the various ASR back-ends for decoding.

This paper investigates the use of three exemplar spaces for creating the input dictionaries and obtaining the decomposition; Mel-scaled magnitude spectra (denoted as *Mel-DFT* setting), magnitude spectra (denoted as *DFT-DFT* setting) and the perceptually motivated modulation spectrogram features (denoted as *MS-DFT* setting). The output of the beamformer is used for obtaining the baseline ASR results. Average word error rates (WER) in % are used to compare the various settings. The Mel, DFT and MS feature dimensions are the same as used in [2].

The dictionaries used are comprised of 10000 speech and 5000 noise exemplars. The number of adaptive atoms learned is set to be $\lceil 0.03 \cdot F \rceil$, where F is the number of frames in the input noisy speech. The NMF-based decomposition is obtained such that the Kullback-Leibler divergence between the input noisy speech spectrogram and its approximation as a non-negative linear combination of the dictionary atoms is minimized.

For the baseline system, the evaluation on the real recordings subset of the database yielded WERs of 37.7%, 34.5% and 27.0% using the 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 together 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 (INSPIRE).

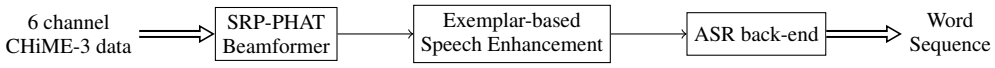

Figure 1: *Outline of the investigated setting for ASR evaluation on the CHiME-3 data.*

2. References

[1] J. Barker, R. Marxer, E. Vincent, and S. Watanbe, "The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *IEEE ASRU*, 2015.

[2] D. Baby, T. Virtanen, J. F. Gemmeke, and H. Van hamme, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 1788 –1799, Nov. 2015.

[3] L. B. and B. Yang, "Adaptive Segmentation and Separation of Determined Convolutive Mixtures under Dynamic Conditions," in *Latent Variable Analysis and Signal Separation - 9th International Conference, LVA/ICA Proceedings*, 2010, pp. 41–48.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.