Using full-rank spatial covariance models for noise-robust ASR

Dung T. Tran, Emmanuel Vincent, Denis Jouvet, Kamil Adiloglu

USING FULL-RANK SPATIAL COVARIANCE MODELS FOR NOISE-ROBUST ASR

Dung T. Tran^{1,2,3}*, Emmanuel Vincent*^{1,2,3}*, Denis Jouvet*^{1,2,3} *and Kamil Adiloğlu*^{4}

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[4]HörTech gGmbH, Marie-Curie-Str. 2, D-26129 Oldenburg, Germany
dung.tran@inria.fr

ABSTRACT

We present a joint spatial and spectral denoising front-end for Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge based on the Flexible Audio Source Separation Toolbox (FASST). We represent the sources by nonnegative matrix factorization (NMF) and full-rank spatial covariances, which are known to be appropriate for the modeling of small source movements. We then learn acoustic models for automatic speech recognition (ASR) on the enhanced training data. We obtain 40% average error rate reduction due to speech separation compared to multicondition training alone.

Index Terms— speech separation, FASST, noise-robust speech recognition

1. INTRODUCTION

Robust distant-microphone ASR in real-world environments is still a challenging problem, due to reverberation and non-stationary background noise including multiple noise sources. The CHiME Speech Separation and Recognition Challenges[1] were launched to contribute to solving this problem [1] [2].

In the 1st CHiME Challenge, we used the FASST source separation toolbox[2] as a speech enhancement front end [3]. This toolbox models the source spectra by means of multilevel NMF and their spatial properties by means of either rank-1 or full-rank spatial covariance matrices [4]. Based on available knowledge such as the speakers identity, the rough target spatial direction and the temporal location of the target speech utterances within the mixture signal, appropriate constraints can be specified on the model parameters, so as to design a custom speech separation algorithm with little effort.

In the 2nd CHiME Challenge, the difficulty was extended by allowing the target speaker to make small head movements in a zone of ±10 cm. In order to address this issue in FASST, we adopt the same NMF spectral models as in [3] but we model the spatial properties of the sources using full-rank spatial covariance matrices instead. Such matrices, which encode both the spatial direction and the spatial width of the sources, have been shown to be more robust to reverberation and small source movements than the conventional rank-1 spatial covariance model [5]. We then train the ASR acoustic models directly on the enhanced noisy training data, instead of performing maximum a posteriori (MAP) adaptation as in [3].

The rest of this paper is organized as follows. The speech separation algorithm based on FASST is presented in Section 2. The ASR system and its results are discussed in Section 3. We conclude in Section 4.

2. FASST-BASED SPEECH SEPARATION

2.1. Speech spectral model

For each of the 34 speakers, we learn a speaker-dependent NMF model of its short-term power spectrum using 50 utterances randomly picked from the noiseless reverberated training set. The NMF basis spectra are initialized by split vector quantization and re-estimated using FASST.

2.2. Speech spatial model

We also learn an initial speaker-independent full-rank spatial covariance model of the target speech source from the noiseless reverberated training set. Due to the size of this dataset, only 45 utterances are randomly selected from each speaker. The spatial covariance matrices are randomly initialized and re-estimated using FASST.

2.3. Background noise model

The noise is modeled as a sum of 4 sources. Each source is given a full-rank spatial model and a NMF spectral model. This multi-source noise model is trained on the speech-free background samples (5 s before and 5 s after each sentence) of the mixture signals to be separated. The model is randomly initialized and trained using FASST.

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/index.html

[2] http://bass-db.gforge.inria.fr/fasst/

2.4. Mixture separation

After the spatial models and the NMF spectral models have been trained, the utterance to be separated is modeled as a sum of 1 speech source and 4 background noise sources, whose parameters are initialized by those of the corresponding trained models. While the NMF basis spectra of the target and the background are kept fixed, the other parameters (namely, the NMF temporal activation coefficients and the spatial covariance matrices) are re-estimated on that noisy utterance using FASST. Finally, the target speech signal is extracted by multichannel Wiener filtering. This procedure is applied to all noisy utterances in the training, development and test sets.

3. EXPERIMENTS AND RESULTS

We use 500 iterations of FASST for all the above four steps, except in the last step on the development and test data for which 1000 iterations are used. We performed separation with different numbers of NMF components for the speaker and the background noise and found the best number to be 32.

The features used in this experiment are 39-dimensional MFCCs (12 cepstral + log-energy, delta, delta-delta) with cepstral mean subtraction. We use the HTK baseline provided on the CHiME website up to a modification of the ADDDITHER parameter, which governs the amount of noise added to the signal before MFCC calculation, so as to make the MFCCs more robust to zeroes in the speech spectra after source separation. The optimal value of ADDITHER on the development set was found to be 25.

In addition to the baseline noisy and reverberated acoustic models provided on the CHiME website, we train speaker-dependent acoustic models on the enhanced noisy training data using the HTK baseline. Speaker-independent models are learned from all speakers' data and subsequently adapted to each speaker by running a few additional iterations of Baum-Welch and keeping the weights and variances of the Gaussian mixture model (GMM) observation probabilities fixed while re-estimating their means.

We test three possible cases:

- NE+No: without speech enhancement, models trained on noisy data,
- WE+Re: with speech enhancement, models trained on reverberated data,
- WE+En: with speech enhancement, models trained on enhanced noisy data.

The results for test and development sets are given in Tables 1 and 2, respectively. On the test set, we achieve an average error rate reduction (ERR) of 40% compared to multicondition training alone and 10% compared to source separation alone.

The full-rank spatial covariance model resulted in 9% ERR compared to the rank-1 model at the expense of a larger computational cost. Setting ADDDITHER=25 further resulted in 13% ERR compared to ADDITHER=1. As usual [2], multicondition training led to slightly reduced accuracy against reverberated training at higher signal-to-noise ratios.

Table 1. Keyword speech recognition accuracy (in %) for test data

System	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
NE+No	60.17	66.83	75.83	82.67	84.33	87.92
WE+Re	69.42	77.67	84.58	**89.17**	**91.75**	**92.33**
WE+En	**76.42**	**81.00**	**85.33**	89.08	90.67	91.58

Table 2. Keyword speech recognition accuracy (in %) for development data

System	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
NE+No	61.33	66.42	75.00	82.50	86.58	88.83
WE+Re	70.67	76.58	82.50	86.67	89.83	**90.92**
WE+En	**76.00**	**80.25**	**85.00**	**86.75**	**90.08**	90.08

4. CONCLUSION

The results demonstrate the potential of full-rank spatial covariance models combined with NMF as a denoising front end for noise-robust speech recognition. Future work will concentrate on improving the integration of FASST and ASR using uncertainty propagation.

5. REFERENCES

[1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[2] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[3] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *Proc. 1st Int. Conf. on Machine Listening in Multisource Environments (CHiME)*, 2011, pp. 86–87.

[4] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, May 2012.

[5] Ngoc Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Jul 2010.