Using the FASST source separation toolbox for noise robust speech recognition

Alexey Ozerov and Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
{alexey.ozerov, emmanuel.vincent}@inria.fr

Abstract

We describe our submission to the 2011 CHiME Speech Separation and Recognition Challenge. Our speech separation algorithm was built using the Flexible Audio Source Separation Toolbox (FASST) we developed recently. This toolbox is an implementation of a general flexible framework based on a library of structured source models that enable the incorporation of prior knowledge about a source separation problem via user-specifiable constraints. We show how to use FASST to develop an efficient speech separation algorithm for the CHiME dataset. We also describe the acoustic model training and adaptation strategies we used for this submission. Altogether, as compared to the baseline system, we obtain an improvement of keyword recognition accuracies in all conditions. The best improvement of about 40 % is achieved in the worst condition of -6 dB Signal-to-Noise-Ratio (SNR), where 18 % of this improvement is due to the speech separation. The improvement decreases when the SNR increases. These results indicate that audio source separation can be very helpful to improve speech recognition in noisy or multi-source environments.

Index Terms: speech separation, source separation, general flexible framework, noise robust speech recognition

1. Introduction

The CHiME Speech Separation and Recognition Challenge aims to tackle speech separation and recognition in typical everyday listening conditions. The challenge employs noise background that has been collected from a real family living room using binaural microphones. Target speech commands have been mixed into the environment at a fixed position using genuine room impulse responses. The task is to separate the speech and recognize the commands being spoken using systems that have been trained on noise-free commands and room noise recordings. For more details about the challenge and the corresponding datasets, see the challenge web page [1].

We here describe our submission to the challenge based on the Flexible Audio Source Separation Toolbox (FASST), developed in Matlab, that enables the incorporation of prior knowledge about source separation problems to build efficient source separation algorithms. To design a speech separation algorithm for the CHiME challenge we have used the following prior knowledge about the source separation problem:

1. The speaker identity and availability of clean speech signals for each speaker.
2. A rough idea about the target speech source direction.
3. Knowledge that the background noise can involve multiple sources and that there are usually no more than 4 active sources at the same time.
4. Knowledge of the temporal location of the target speech sentences within the noisy background.

Note that points 1 and 4 are used as well by the CHiME baseline recognizer. We here do not give any details about the toolbox and the underlying general flexible audio source separation framework, and rather refer the reader to visit the FASST web page [2] and to read the corresponding article [1].

The rest of this paper is organized as follows. The speech separation algorithm based on FASST is presented in section 2. Section 3 describes the acoustic model training and adaptation strategies we employed. The results are presented in section 4 and the conclusions are drawn in section 5.

2. Speech separation using FASST

Our speech separation algorithm based on FASST consists in the following steps:

1. *Speech spectral power model:* For each of the 34 speakers, a speaker dependent 32-component nonnegative matrix factorization (NMF) model [1] is learned from 60 sentences randomly selected from the training set and converted to mono. The spectral patterns of the NMF model are first initialized by K-means clustering of the short term power spectra, and then reestimated using FASST.
2. *Speech spatial model:* An initial rank-1 convolutive filter is estimated from the clean reverberated speech of the development test set.
3. *Background noise model:* The noise is modeled as a sum of 4 sources. Each source is given a rank-1 convolutive spatial model and an 8-component NMF spectral power model. This multi-source model is initialized randomly and learned using FASST from 20 seconds of speech-free background noise (10 seconds before and 10 seconds after each sentence), which can be extracted thanks to the availability of the temporal location of the target speech sentences.
4. *Mixture model:* A mixture model consisting of 5 source models (1 for speech and 4 for noise) is created for the separation of the target speech. The spatial models and the NMF spectral patterns are initialized by those of the corresponding models described in the three previous steps. The NMF temporal activations are initialized randomly. All the parameters, except the NMF spectral patterns that are kept fixed, are separately reestimated

This work was supported in part by the Quaero Programme, funded by OSEO.

[1] http://www.dcs.shef.ac.uk/spandh/chime/challenge.html

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

Constr. train.	MAP adapt.	Src. sep.	Development set						Test set					
			-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
No	No	No	31.08	36.75	49.08	64.00	73.83	83.08	30.33	35.42	49.50	62.92	75.00	82.42
No	No	Yes	51.08	59.83	69.50	75.92	81.25	84.00	48.50	56.58	66.67	74.33	82.17	86.33
No	Yes	No	44.08	51.17	62.17	71.92	81.08	88.08	43.58	50.08	62.50	73.25	82.08	87.83
No	Yes	Yes	63.42	72.83	79.42	83.50	87.75	89.50	64.92	69.92	77.58	82.75	86.67	87.58
Yes	No	No	47.00	49.83	61.25	73.58	82.83	88.25	44.00	50.08	63.33	73.50	83.25	90.17
Yes	No	Yes	65.75	71.00	78.50	85.17	88.42	90.08	65.50	73.08	80.08	86.25	89.00	**92.83**
Yes	Yes	No	51.67	57.17	70.08	78.17	86.83	89.58	52.92	59.50	69.75	79.92	85.75	91.67
Yes	Yes	Yes	**70.00**	**77.17**	**84.33**	**88.33**	**91.58**	**93.17**	**71.08**	**76.58**	**81.08**	**88.58**	**90.33**	90.67

Table 1: Speech recognition performance expressed in terms of keyword recognition accuracy (in %) for both the development and test sets and for different configurations. Abbreviations: "Constr. train." = constrained acoustic model training, "MAP. adapt." = MAP acoustic model adaptation, and "Src. sep." = using source separation.

from each noisy speech sentence using FASST. Finally the speech source signal is extracted using FASST.

All learning operations are performed in the maximum likelihood (ML) sense.

We did some informal listening of the separated signals. The separated speech is usually quite clean with some artifacts and some interferences, especially from the concurrent speech sources coming from about the same direction. However, there is still some speech in the separated noisy background. Some separation examples are available from the demo web page [3].

3. Training and adaptation

Speech recognition is achieved using speaker-dependent acoustic models estimated from the clean speech of the training set. First, we have tried the models provided within the challenge. Second, we have tried to re-train these models by learning speaker independent models and fixing the acoustic Gaussian mixture model (GMM) variances and weights while adapting them to each speaker, i.e., only the means were reestimated. This constrained training was achieved by adding the `-u m` option to the `HERest` HTK function used for the reestimation of speaker-dependent model parameters.

Since the acoustic models are trained from clean speech and our goal is to recognize either noisy speech or separated speech (corrupted by source separation errors), the speech recognition performance can be improved by adapting the acoustic models to these perturbations. An adaptation set consisting of noisy speech sentences (60 sentences per speaker) was provided within the challenge. However, this set was provided without the corresponding speech-free background. Thus, we were not able to use this set, since it was not possible to separate it by our speech separation algorithm (in order to reproduce similar perturbations) that relies on the knowledge of the neighbouring speech-free background. We have thus created our own adaptation set consisting of 200 sentences per speaker randomly selected from 500 sentences of the training set and randomly mixed with the training set backgrounds (only the mixtures with Signal-to-Noise-Ratio (SNR) between -5 and 10 dB were retained). We then used the maximum a posteriori (MAP) adaptation script provided within the challenge to adapt the acoustic models to this adaptation set as it is, if no speech separation is used, or to its separated version, if speech separation is applied. Knowledge of the test or development SNR was not used. Whatever the SNR, the models were adapted on the whole adaptation set for each speaker.

[3] `http://www.irisa.fr/metiss/ozerov/chime_ssep_demo.html`

4. Results

We have tested all the 8 possible combinations of the proposed options, i.e., with or without constrained training (Sec. 3), with or without MAP adaptation (Sec. 3) and with or without speech separation (Sec. 2). The results for both the development and test sets are given in Table 1. We see that, as compared to the baseline system (first line of Tab. 1), all three tested options lead to some improvement in almost all conditions and whatever their combination. On the development set, the best improvement of about 39 % is achieved in the worst condition of -6 dB SNR, with 16 % due to the constrained training (see lines 1 and 5 of Tab. 1), 5 % due to the MAP adaptation (see lines 5 and 7 of Tab. 1), and the remaining 18 % due to source separation (see lines 7 and 8 of Tab. 1). The improvement decreases but remains positive when the SNR increases.

5. Conclusion

First, our results indicate that audio source separation can be very helpful to improve speech recognition in noisy or multisource environment. Second, building a new algorithm for this particular source separation problem using FASST was really fast since no additional coding is needed. Indeed, it took us only two weeks to prepare this submission, all steps included. This demonstrates once again the usefulness of the FASST source separation toolbox and the underlying general flexible framework [1].

6. Acknowledgements

The authors would like to thank the CHiME'11 organizers for very well-prepared challenge and especially Ning Ma for a constant support, interesting discussions and for his suggestion about the constrained acoustic model training we used.

7. References

[1] 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*, submitted.