Using the FASST Source Separation Toolbox for Noise Robust Speech Recognition

Alexey Ozerov and Emmanuel Vincent

INRIA, Rennes Bretagne Atlantique, Rennes, France

e-mail: {alexey.ozerov,emmanuel.vincent}@inria.fr

Abstract

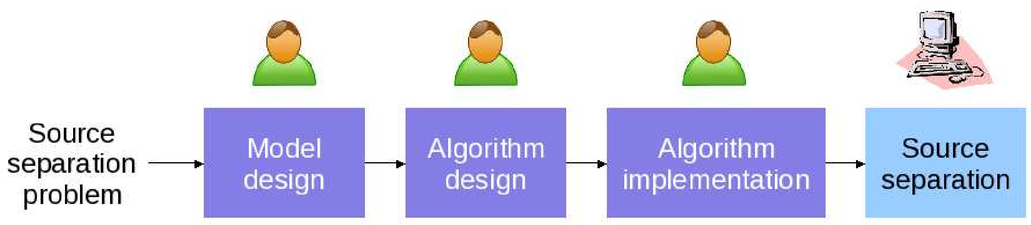
- We present our submission to the 2011 CHiME Speech Separation and Recognition Challenge.
- Our speech separation algorithm was built using the Flexible Audio Source Separation Toolbox (FASST) [1] we developed recently. This toolbox is an implementation of a general flexible framework [2] based on a library of structured source models that enable the incorporation of prior knowledge about a source separation problem via userspecifiable constraints.

Prior knowledge used for speech separation

- The speaker identity and availability of clean speech signals for each speaker.
- A rough idea about the target speech source direction.
- Knowledge that the background noise can involve multiple sources.
- Availability of "clean" background noise signals before/after each target speech sentence.

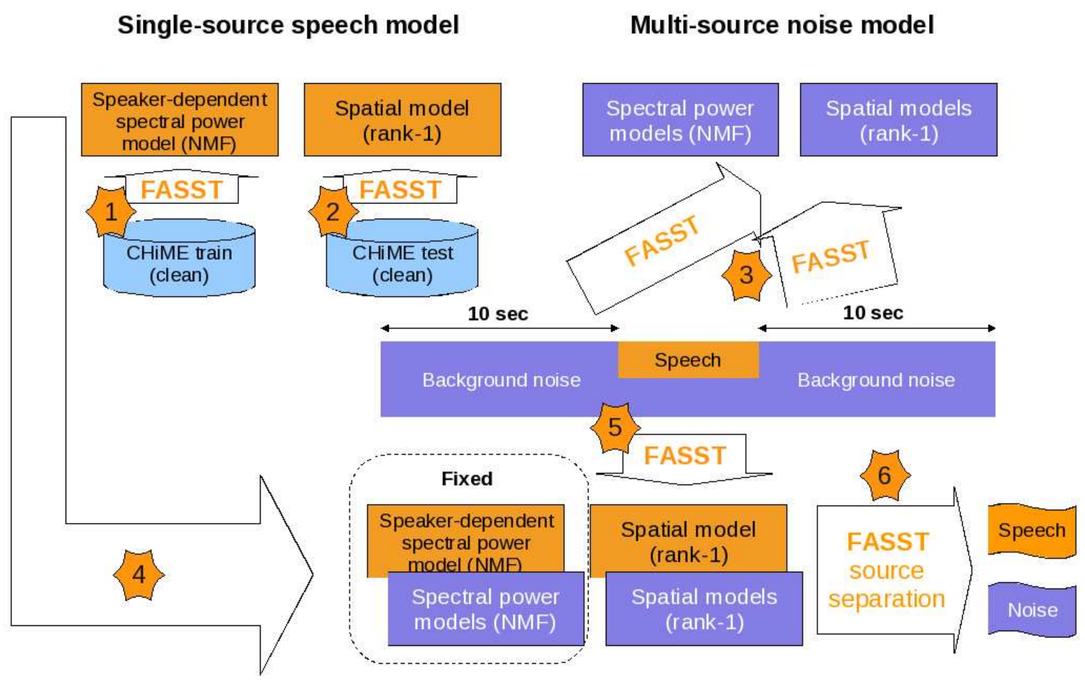
Flexible audio source separation is FASST

Current approach for audio source separation SLOW !!!



Flexible audio source separation framework FASST !!!

Speech separation using FASST



Results

- Constrained training: constrained acoustic model training (only GMM means are adapted to each speaker).
- MAP adaptation: MAP acoustic model adaptation (to noisy or separated speech).
- Source separation: using speech separation.

Constrained training	MAP adaptation	Source sseparation	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.

Conclusions

- Results indicate that audio source separation can be very helpful to improve speech recognition in noisy or multisource environment.
- Building a new algorithm for this particular source separation problem using FASST was really fast since no additional coding was needed (only 2 weeks were spent to prepare this submission).

References

[1] A. Ozerov, E. Vincent, and F. Bimbot, "Flexible Audio Source Separation Toolbox (FASST)." Available: http://bass-db.gforge.inria.fr/fasst/

[2] ——, "A general flexible framework for the handling of prior information in audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, submitted. Available: http://hal.inria.fr/inria-00536917/