The CMU-MIT REVERB Challenge 2014 System: Description and Results

Xue Feng[1], Kenichi Kumatani[2], John McDonough[2]

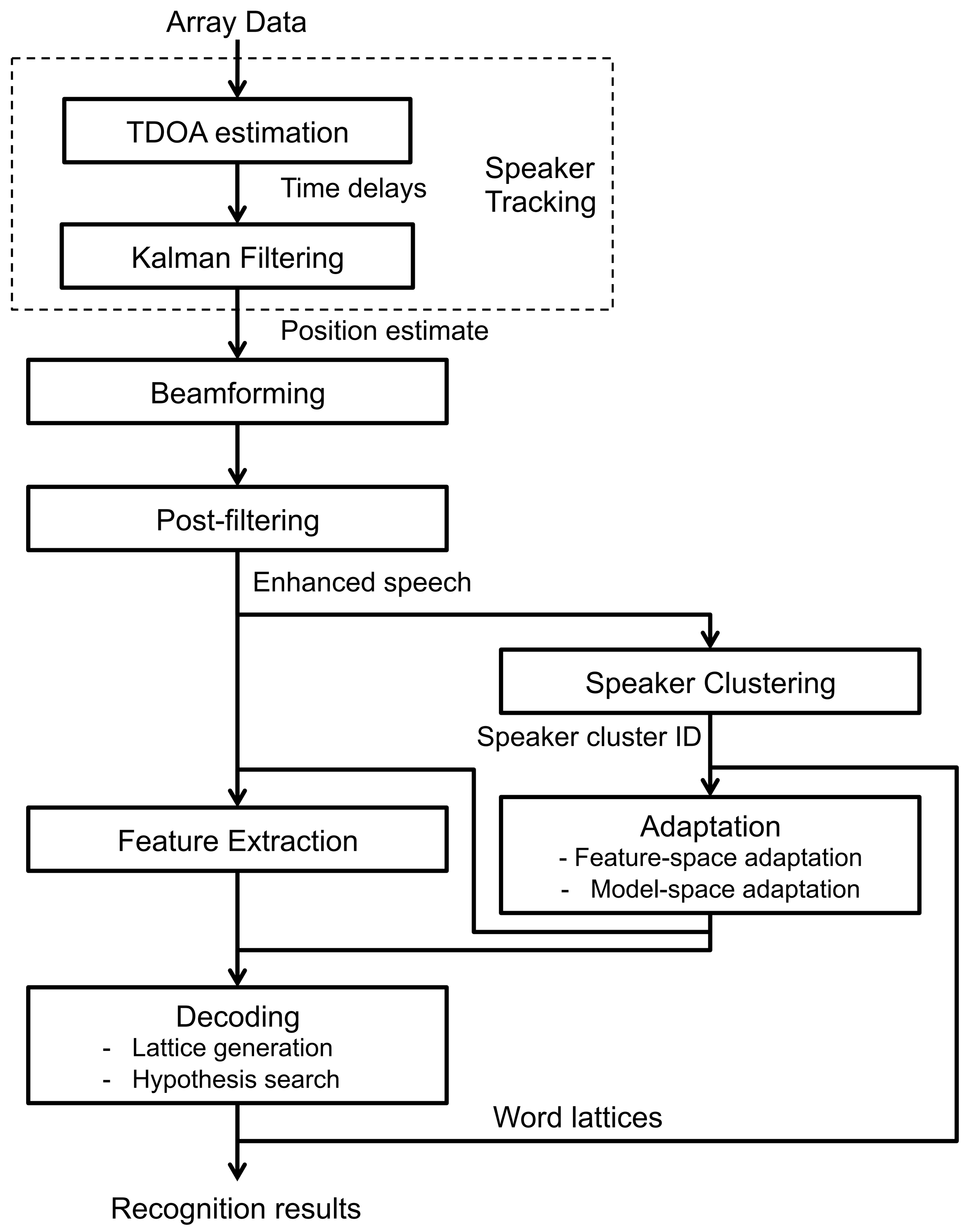
[1]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA.

[2]Languate Technologies Institute, CMU, Pittsburgh, PA, USA.

Introduction

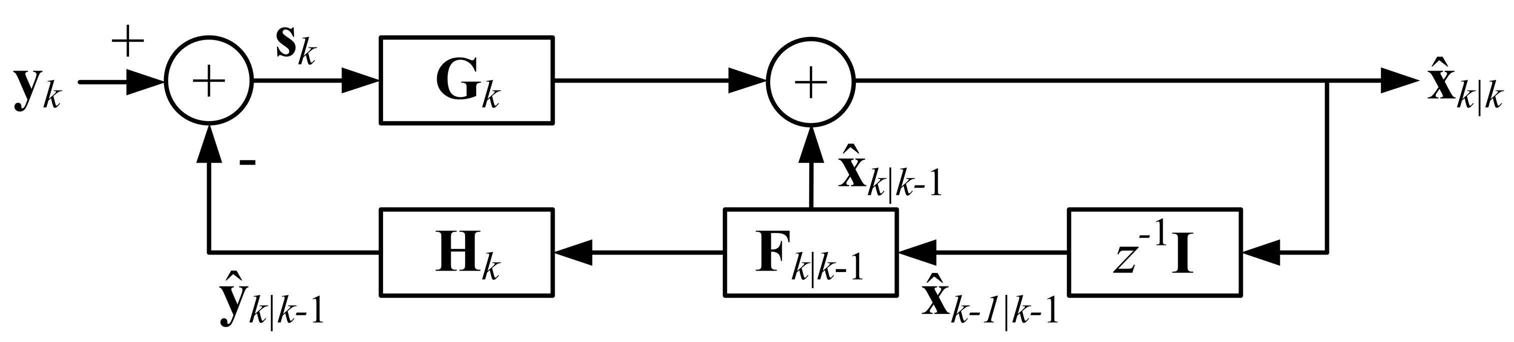
- The CMU-MIT System for RC2014 has four parts:
  1. Speaker tracking to determine speech *direction of arrival*;
  2. Beamforming to enhance speech from microphone array;
  3. Speaker clustering to group utterances for speaker adaptation;
  4. An FST-based speech recognition engine.
- Our system reduced WER from 39.9% with a single array channel to 14.5% with eight channels.

System

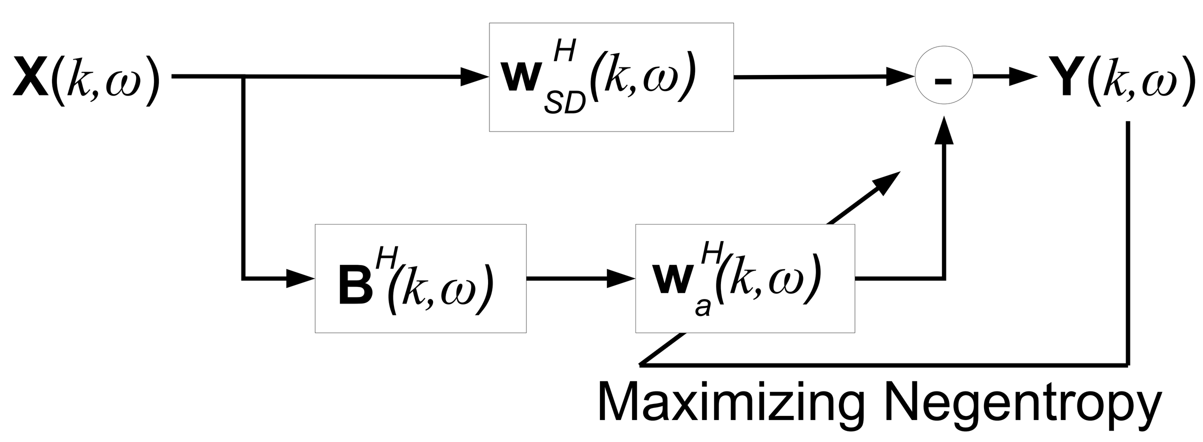


Method

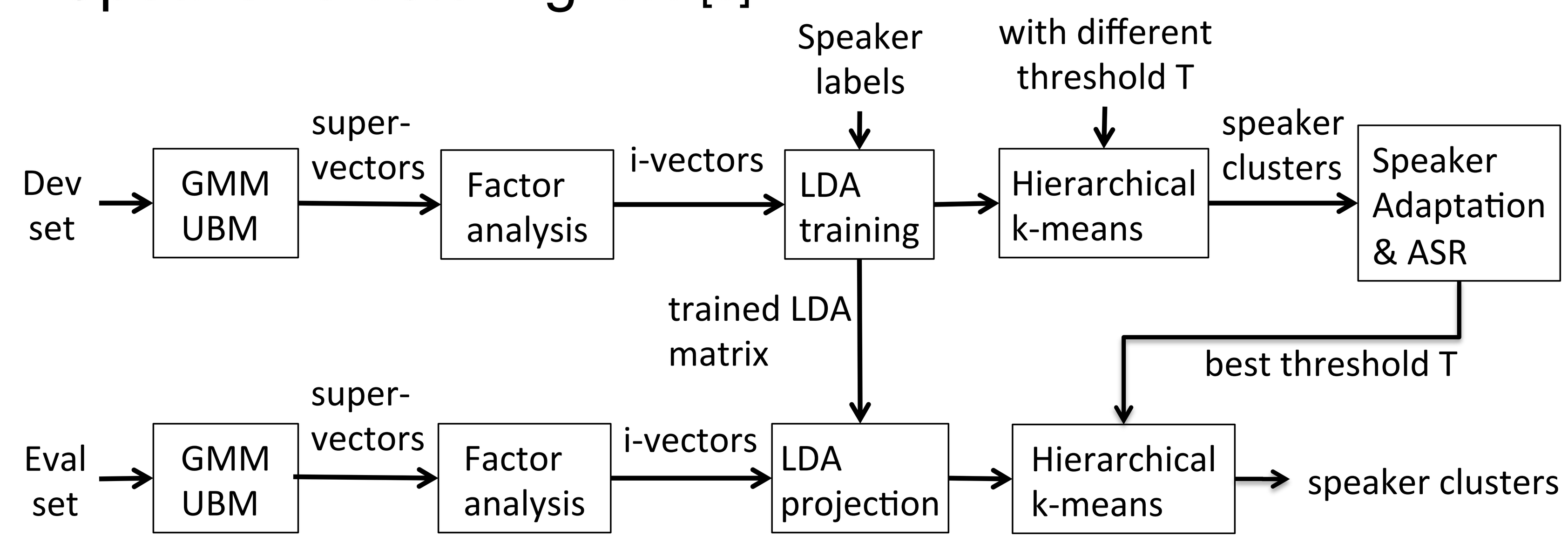
I. Speaker Tracking: see Section 10.2 in [1].



II. Maximum negentropy beamforming: see Kumatani *et al.* [2].



III. Speaker Clustering: see [3].



Cluster division stopped for BIC improvement below a threshold.

Table: Speaker Clustering Results on Dev Set

	Simulated Data		Real Data	
Threshold T	No. of clusters	WER	No.of clusters	WER
30	71	19.4	19	17.12
50	45	18.8	15	16.80
100	11	20.5	10	16.63
180	5	19.4	5	17.35

Experiments

- Speech features were extracted with MVDR cepstral analysis [4].
- Final features obtained by concatenation of 15 successive cepstral frames and performing LDA to reduce feature size to 10.
- The speech recognition engine was based on fast on-the-fly composition of weighted finite-state transducers [5].
- Four passes of recognition were performed with increasing levels of speaker adaptation.
- Unsupervised speaker adaptation was based on word lattices from the prior pass.

	Simulated Data							Real Data		
	Room 1		Room 2		Room 3			Room 1		
System	Near	Far	Near	Far	Near	Far	Ave.	Near	Far	Ave.
Primary	12.89	14.71	14.09	19.38	16.62	31.45	18.68	16.26	16.54	16.46
Contrast A	8.40	10.27	14.1	30.54	17.11	44.65	20.85	38.38	41.41	39.90
Contrast B	8.12	8.93	9.60	12.99	9.73	20.18	11.80	14.76	14.18	14.50
Contrast C	8.17	9.23	10.10	15.00	15.00	29.04	14.42	18.80	11.04	15.95
Contrast D	6.81	6.81	7.59	7.59	7.08	7.08	7.16	7.98	7.36	7.67

Primary: Our official REVERB Challenge 2014 system
Contrast A: Single Array Channel, using true speaker labels
Contrast B: Maximum Negentropy Beamforming, using true speaker labels
Contrast C: Super Directive Beamforming, using true speaker labels
Contrast D: Close Talking Microphone, using true speaker labels

Conclusions and Future Work

- Only experiments on *real* data provide results that reliably predict performance in real environments.
- Maximum negentropy beamforming is more effective than MVDR beamforming for DSR applications.
- Future work:
  - Couple our array processing techniques with a DNN recognizer.
  - Incorporate more speech knowledge into beamforming.
  - Release our array processing tools into the public domain.

References

[1] Wölfel, M. and McDonough, J., *Distant Speech Recognition*, Wiley, New York, 2009.
[2] Kumatani, K., Lu, L., McDonough, J., Ghoshal, A., and Klakow, D., "Maximum negentropy beamforming with superdirectivity," *Proc. of EUSIPCO*, Aalborg, Denmark, 2010.
[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, 2011.
[4] Wölfel, M. and McDonough, J., "Minimum variance distortionless response spectral estimation: Review and refinements," *IEEE Signal Processing Magazine*, 2005.
[5] McDonough, J. and Stoimenov, E., "An algorithm for fast composition of weighted finite-state transducers," *Proc. of ASRU*, Kyoto, Japan, 2007.