NMF based front-end processing in multichannel distant speech recognition

Nikhil Mohanan[1], Premanand Nayak[1], Rajbabu Velmurugan[1], Preeti Rao[1], Sonal Joshi[2], Ashish Panda[2], Meet Soni[2], Rupayan Chakraborty[2], Sunilkumar Kopparapu[2]

[1]Indian Institute of Technology Bombay, India
[2]Tata Consultancy Services, India

- ❑ Our system focuses on implementing a better front-end for the Automatic Speech Recognition (ASR) system
- ❑ Single-channel enhancement using non-negative matrix factorization (NMF) followed by multi-channel minimum variance distortionless response (MVDR) beamformer
- ❑ Alternate model to enhance the MVDR output signal by a novel NMF based enhancement.

Challenge Setup And Baseline

- ➢ Distant speech recognition with natural conversational speech [1] :
  - Microsoft Kinects arrays, 4 microphones each, placed at different locations.
  - Session has 6 such arrays, 2 each at locations: living, kitchen and dining.
  - Session has 4 speakers, in the same room at a particular instant wearing a close-talking binaural mic.
- ➢ Our results are for the single-array track (Ranking A) and focuses on acoustic robustness.
- ➢ We use baseline acoustic model (AM) and language model (LM)
- ➢ Baseline enhancement system
  - Single channel noise filtering using Weiner Filtering
  - Source localization by GCC-PHAT followed by Viterbi algorithm.
  - Delay Sum Beamformer (DSB)

Proposed System

❑ MVDR + NMF:

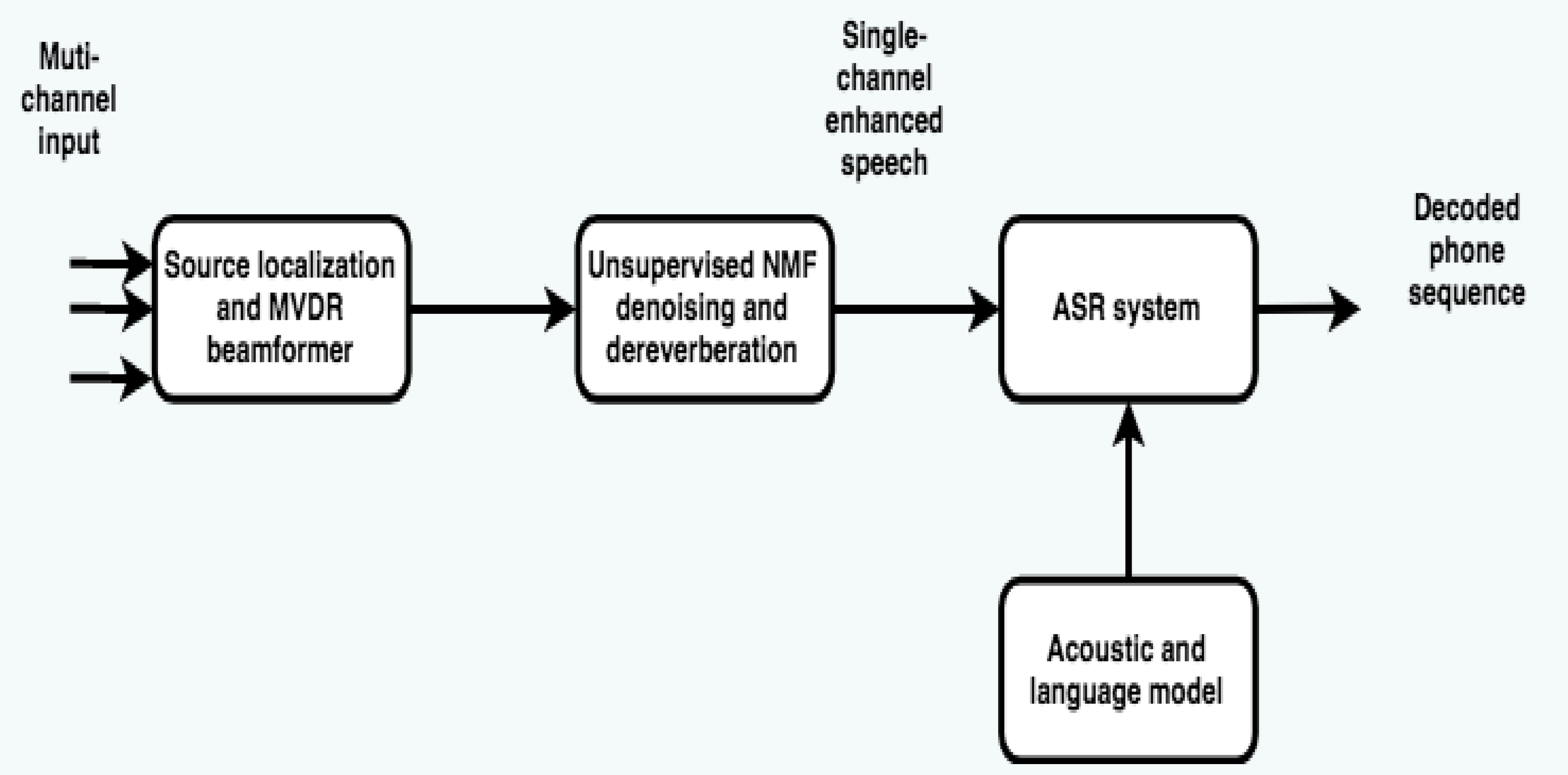

Figure 1: Block diagram of MVDR+NMF system

- ➢ GCC-PHAT compute TDOA's.
- ➢ Minimum Variance Distortionless Response Beamforming (MVDR)
  - For removal of directional noise
  - Covariance matrix computed using noisy frames located using VAD
- ➢ Non-negative Matrix Factorization (NMF) [3] used to enhance MVDR output.
- ➢ Drawback:
  - No improvement in terms of ASR.
    - Possible reason: noisy TDOA's fed as steering vector
- ➢ Modified system : enhance each channel using NMF filtering followed by MVDR beamforming

❑ NMF + MVDR system:

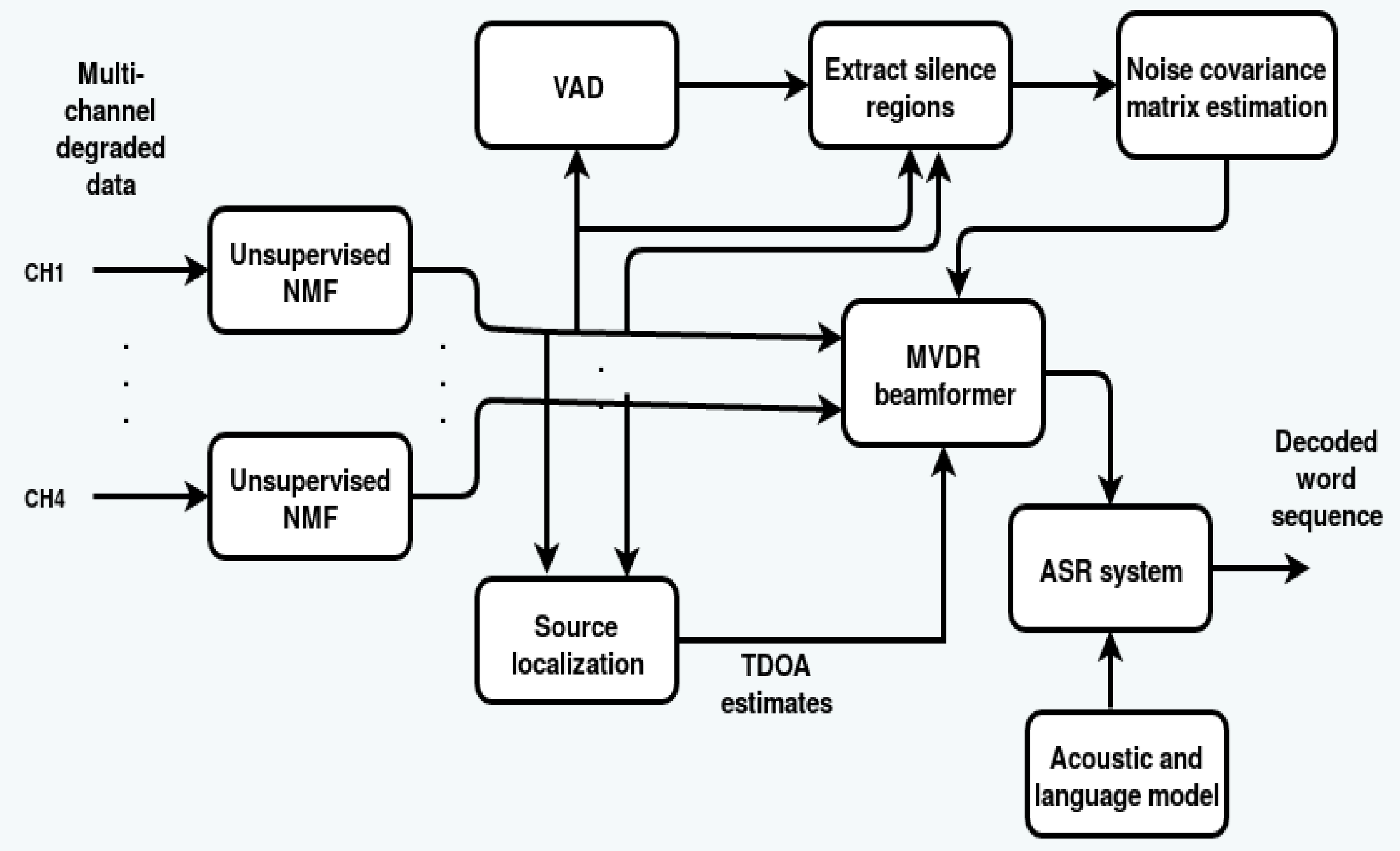

Figure 2: Block diagram of NMF+MVDR system

- ➢ Input array signals were using NMF and fed to MVDR.
- ➢ Supervised approach: clean speech and noise bases learnt from the degraded data

Single Channel NMF

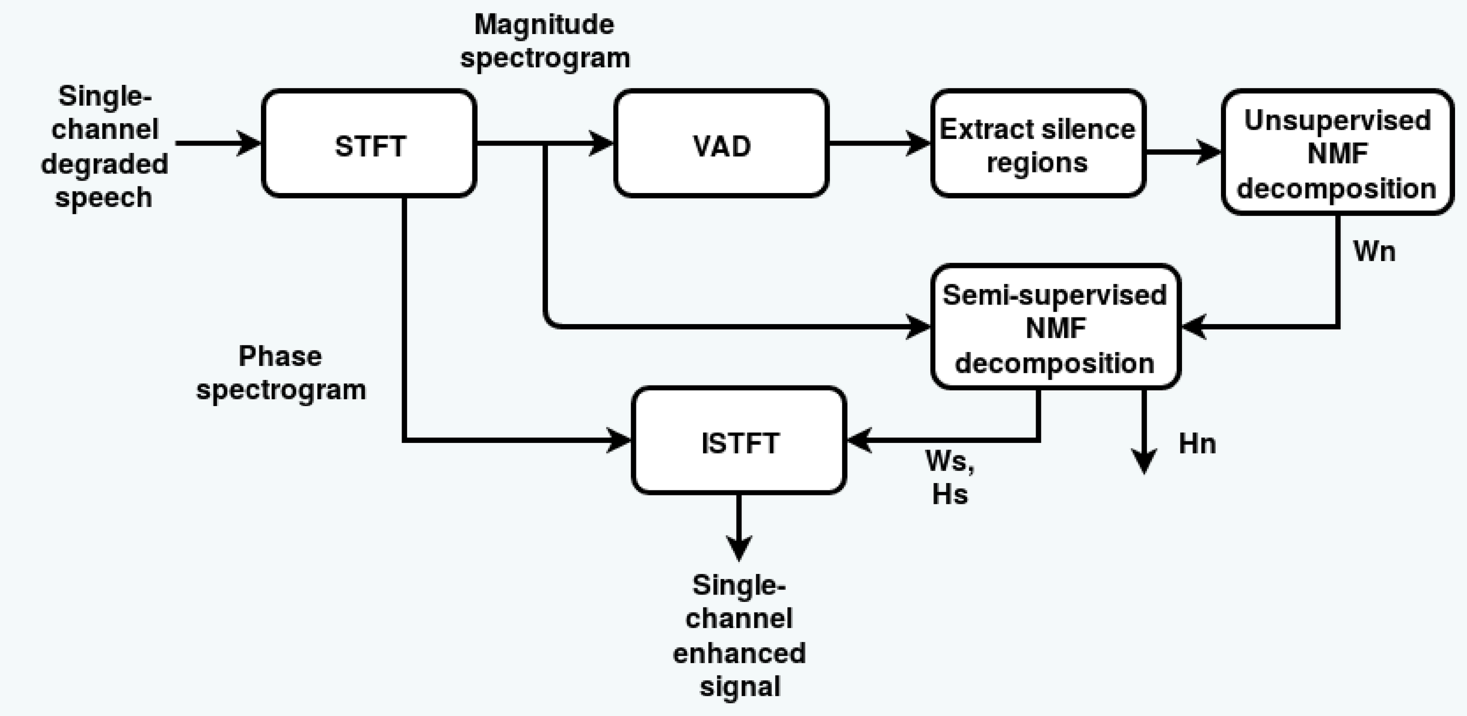

Figure 3: Block diagram for the unsupervised NMF block

- ➢ Noise bases learning
  - Clean speech bases learned using unsupervised approach
  - MVDR output used for feature extraction and decoded by ASR system.
- ➢ Degraded (reverb and noisy)speech spectrogram : $Y = Y_r + Z = [W_r \,|W_n][X_r^T \mid X_n^T]^T$
- ➢ Reverb spectrogram $Y_r = W_r\, X_r$, Noise spectrogram $Z = W_n\, X_n$
- ➢ Reverb bases and activations related to corresponding clean bases and activations

Results and Analysis

- ➢ Training using the baseline AM, a mixture of both close-talking microphones and array channels data.
- ➢ Total of 100k (61349 close talking and 38651 array) utterances of this mixture
- ➢ Magnitude spectrogram obtained using a 64ms Hamming window with a 32ms hop.
- ➢ TDOA estimates obtained from NMF filtered channel Beamformit used compute steering vector for MVDR
- ➢ Enhanced utterance used for ASR.

Track	System	WER
Single-Microphone Array	Degraded (single-channels)	92.18
	Beamformit (Baseline)	91.33
	Beamformit+NMF	93.94
	Beamformit+RNMF	95.51
	MVDR	96.68
	NMF+MVDR	*95.56*
	MVDR+NMF	96.80

Table 1: Overall WER (%) for the GMM-HMM based systems tested on the development test set using baseline AM and LM.

- ➢ Enhancements done(GMM-HMM acoustic model):
  - Beamformit: Baseline enhancement by DSB beamforming
  - Beamformit+NMF: Beamformit followed by NMF de-noising for noise suppression
  - Beamformit+RNMF
  - MVDR:MVDR beamforming with TDOA's computed via GCC-PHAT
  - NMF+MVDR:NMF de-noising followed by MVDR beamforming
  - MVDR+NMF:MVDR beamforming followed by NMF.

Track	Session	Kitchen	Living	Dining	Overall
Single Microphone Array	S02	97.58	96.47	94.56	*95.56*
	S09	94.77	95.30	94.43	

Table 2: Results on development dataset of the NMF+MVDR for GMM-HMM based systems

Track	Session	Kitchen	Living	Dining	Overall
Single Microphone Array	S02	95.46	90.17	91.98	*92..85*
	S09	93.41	93.98	93.13	

Table 3: Results on development dataset of the NMF+MVDR for TDNN based systems

- ➢ WER is poor for all the locations and the .
- ➢ Poor performance is train-test mismatch.
- ➢ Attempt was made to remove residual noise and reverberation in MVDR output by NMF and RNMF post filtering.
- ➢ Proposed methods however did not shown improvement in WER

Acknowledgements

Part of the work supported by Bharti Centre for Communication in IIT Bombay, Council of Scientific and Industrial Research (CSIR), India and Tata Consultancy Services (TCS), India

References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in Interspeech 2018 Hyderabad, India, Sep. 2018.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2011–2022, 2007.

[3] N. Mohanan, R. Velmurugan, and P. Rao, "A non-convolutive NMF model for speech dereverberation," in Interspeech, 2018.