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

nikhilm@ee.iitb.ac.in

Abstract

The submitted system for CHiME-5 challenge focuses on implementing a better front-end for an automatic speech recognition (ASR) system trained on the data provided by CHiME-5. In this work, we focus on using non-negative matrix factorization (NMF) based technique to denoise and dereverberation. In Approach 1, the degraded single-channel speech utterances were enhanced using multi-channel Weighted prediction error (WPE) or NMF followed by a minimum variance distortionless response (MVDR) beamformer to obtain an enhanced signal. In Approach 2, we used multi-channel MVDR followed by a NMF based single-channel enhancement. Using the baseline acoustic model (AM), these enhanced speech utterances did not provide improved WER compared to the baseline Beamformit based system. So, we retrained the AM using WPE enhanced data for training (Approach 3). These approaches were able to improve the ASR results as compared to baseline. We are submitting results for the single-array track and only focus on acoustic robustness (i.e., ranking A).

1. Degradation model

The CHiME-5 recordings were done for conversational speech happening in a dinner party scenario [1]. Four participants were present at each of these dinner parties. The speakers were asked to have normal conversations. The speech was recorded using six Kinect microphone arrays placed in different locations in the room. The duration of each dinner party was at least 1.5 hours. The recordings were degraded by the presence of non-stationary noise, reverberation, overlapping speakers and speaker movements.

1.1. Model for reverberation and noise

In the proposed framework, it is assumed that at any time only one speaker is active. Further, it is assumed that the clean speech is degraded due to reverberation and noise. Other degradations like the presence of interfering speakers and speaker movements are not handled. Reverberation in the time domain is modeled as the convolution of the original source with the room impulse response (RIR). Noise is assumed to be additive to reverberant speech. Time domain speech recorded by each microphone $y^i(n)$ is written as,

$$y^i(n) = y^i_R(n) + z^i(n) = s(n) * h^i(n) + z^i(n) \qquad (1)$$

where, $s(n)$ is the clean utterance, and $y^i_R(n)$, $h^i(n)$ and $z^i(n)$ are the reverberated speech, RIR and noise at the i-th microphone, respectively.

The proposed NMF enhancements are based on the magnitude spectrogram model for degraded speech in [2]. The NMF enhancement can be performed for any one channel of the microphone array recording or to the output of a multi-channel enhancement method. The input degraded spectrogram $\mathbf{Y} \in \mathbb{R}^{K\times T}$ is modeled using NMF. Such a model is obtained by utilizing NMF models for clean speech and noise spectrograms along with a separability assumption on RIR spectrogram $H(k,n) = H_1(k)H_2(n)$. The NMF models for clean speech $\mathbf{S}$ and noise spectrograms $\mathbf{Z}$ are shown in (2).

$$\begin{aligned}\mathbf{S} &= \mathbf{W}_s\mathbf{X}_s \\ \mathbf{Z} &= \mathbf{W}_n\mathbf{X}_n\end{aligned} \qquad (2)$$

where, $\mathbf{W}_s$, $\mathbf{W}_n$ represents the bases for clean speech and noise spectrograms, respectively. $\mathbf{X}_s$, $\mathbf{X}_n$ represents the activations for clean speech and noise spectrograms, respectively. Using a separability assumption on RIR spectrogram and the NMF models for noise and clean speech spectrograms, the degraded speech spectrogram can be represented as,

$$\begin{aligned}\mathbf{Y} &= \mathbf{H} *_n \mathbf{S} + \mathbf{Z} \\ &= [\mathbf{W}_R|\mathbf{W}_n][\mathbf{X}_R^T|\mathbf{X}_n^T]^T\end{aligned} \qquad (3)$$

where, $\mathbf{W}_R$ and $\mathbf{X}_R$ represent the bases and activation matrix for reverb spectrogram, and $*_n$ represents convolution along time index. The reverb bases and activations are related to the corresponding clean bases and activations.

$$\begin{aligned}W_R(k,r) &= W_s(k,r)H_1(k) \\ X_R(r,n) &= X_s(r,n) *_n H_2(n)\end{aligned} \qquad (4)$$

The model for reverberation used in WPE is discussed in section 2.2.

2. Proposed system for CHiME-5

This section discusses various multi-channel and single-channel enhancement methods in the proposed framework for enhancing the CHiME-5 data. Multi-channel enhancement includes beamforming and multi-channel WPE. Single-channel enhancement uses NMF. The last subsection discusses various combinations of multi-channel and single-channel methods used in this work.

2.1. Beamforming

Beamforming is commonly used to enhance a degraded multi-channel recording. The algorithm acts as a spatial filter and enhances source in a particular direction. The performance of algorithm depends on an accurate estimate of source position. This work compares the performance of two types of beamforming - delay-sum beamforming (DSB) and MVDR. DSB is implemented using Beamformit [3]. The MVDR beamformer uses information about noise covariance matrix and source location to perform beamforming. Figure 1 shows the approach

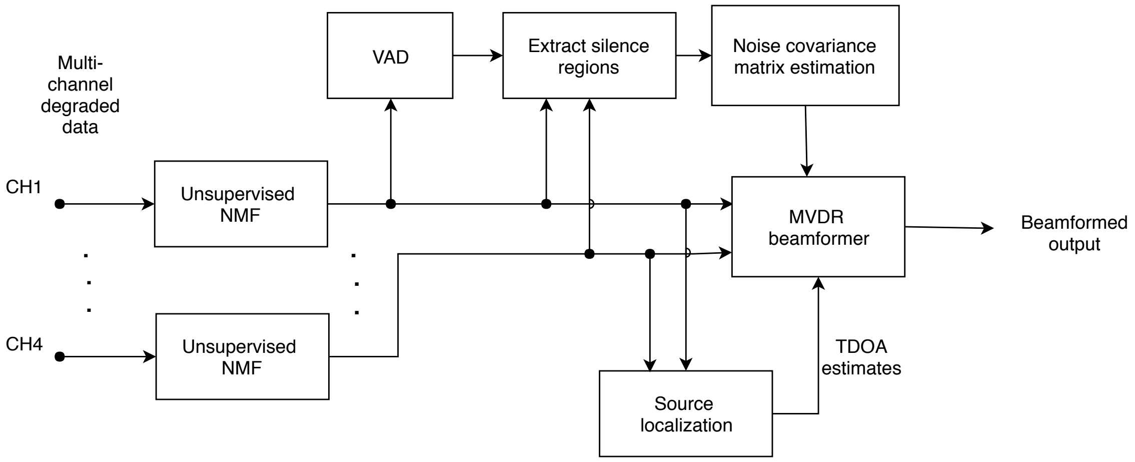

Figure 1: *Block diagram of the MVDR based front-end system implemented for CHiME-5 challenge.*

taken to estimate these parameters. An energy-based voice activity detector (VAD) is used to estimate the silence regions. Noise covariance is estimated from these silence regions. The source localization provided by Beamformit is used in the implementation of MVDR.

2.2. WPE

The CHiME-5 data consists of recording from a near realistic setting of a typical home. The data is affected by reverberation prevailing inside the rooms. Reverberant data degrades the performance of source localization and hence affect the ASR results. The algorithm [4] uses a statistical approach, to remove the late part of reverberation using the multi-microphone signal, without any prior information of the RIR. The speech signal is assumed to be generated using a Gaussian modeled process and the estimate is achieved using a delayed linear prediction with maximum likelihood estimation (MLE). The time-varying characteristic of the speech is compensated in the estimate to an extent by normalizing each speech frame. The algorithm estimates an inverse system to cancel the effects of late reverberation. The estimator is robust such that the convergence is achieved within a few seconds of utterance.

$$x_m(n) = \sum_{k=0}^{L_h-1} h(k,m)s(n-k) \tag{5}$$

$$d_m(n) = \sum_{k=0}^{D-1} h(n,m)s(n-k) \tag{6}$$

$$r_m(n) = \sum_{k=D}^{L_h-1} h(k,m)s(n-k) \tag{7}$$

$$x_m(n) = d_m(n) + r_m(n) \tag{8}$$

$$\hat{d}_m(n) = x_m(n) - (\hat{C})^T x_m(n-D) \tag{9}$$

The degradation and enhancement obtained using this approach is briefly discussed next. The observed signal at the m-th channel $x_m(n)$ can be modeled as (5) where m, L_h corresponds to microphone index and RIR length, respectively. $h(n,m)$ and $s(n)$ represent the time domain RIR for m-th channel and clean speech, respectively. $d_m(n)$ in (6) corresponds to the received clean speech plus the early reverberation part. $r_m(n)$ in (7) is the undesirable late reverberation and $x_m(n)$ in (8) express the observed signal as the sum of both. The early and late part of the reverberation is separated by using a D sample index, which splits the impulse response into two parts. (9) shows the desired signal can be estimated from the previously observed samples, where $(\hat{\mathbf{C}})^{\mathbf{T}}$ is the estimated regression coefficients using MLE. This method is referred to as WPE.

2.3. Unsupervised NMF

The speech enhancement method proposed in [2] cannot be directly used for CHiME-5 data due to the following reasons. Firstly, train data is needed to learn clean speech and noise bases. This is unavailable for CHiME-5 data. Secondly, the algorithm needs the knowledge of temporal variation of RIR spectrogram $H_2(n)$, which is also unavailable. Hence it is appropriately modified to handle CHiME-5 data.

Figure 2 shows the block diagram for the unsupervised NMF algorithm used in the work. The basis vectors representing both clean speech and noise are learned from the degraded data. Hence, there is a need to separate out the clean speech bases from the noise bases. For this purpose, noise bases are learned first from noise-only regions (silence regions) of degraded speech. The silence regions are estimated using an energy-based VAD. An unsupervised NMF decomposition is performed in these silence regions to estimate the noise bases ($\mathbf{W}_n$). With knowledge of $\mathbf{W}_n$, a semi-supervised NMF decomposition is then performed on the entire degraded magnitude spectrogram, with $\mathbf{W}_n$ fixed, to estimate the clean speech bases $\mathbf{W}_s$ and corresponding clean speech activations $\mathbf{X}_s$. The semi-supervised NMF decomposition is discussed in Section 2.4.

For CHiME-5 data, the clean speech and noise bases are updated every 2 minute. This is done to handle the non-stationary behavior of background noise. The enhanced speech is reconstructed using W_s, X_s and the phase information from the degraded speech.

2.4. Semi-supervised NMF

The proposed algorithm follows a two-stage approach as used in [2]. In the first stage, the reverb bases $\mathbf{W}_R$, reverb activations $\mathbf{X}_R$, clean speech bases $\mathbf{W}_s$ and RIR frequency envelope $H_1(k)$ are learned from the degraded data. In the second stage, clean speech activations $\mathbf{X}_s$ and temporal variation

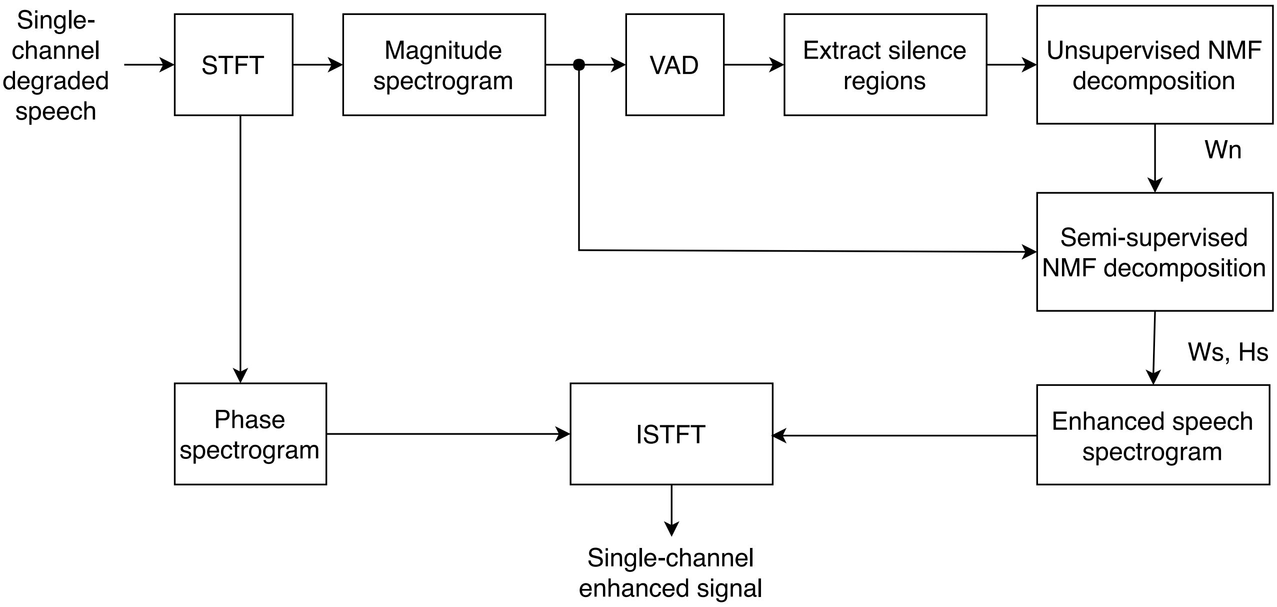
Single-channel degraded speech
STFT
Magnitude spectrogram
VAD
Extract silence regions
Unsupervised NMF decomposition
Wn
Semi-supervised NMF decomposition
Ws, Hs
Phase spectrogram
ISTFT
Enhanced speech spectrogram
Single-channel enhanced signal

Figure 2: *Block diagram of the* ***unsupervised NMF*** *based enhancement used in the MVDR framework shown in Figure 1.*

of RIR spectrogram $H_2(n)$ are learned from $\mathbf{X}_R$ based on a model for $H_2(n)$. The first stage uses Kullback–Leibler (KL) divergence as the cost function as shown in (10). The second term in the cost function is added to make sure that the solution obtained for $W_s(k,r)$ follows the model in (4). Multiplicative updates are used to estimate the parameters $W_R(k,r)$, $W_s(k,r)$, $X_R(r,n)$, $H_1(k)$ from the cost function. The update rules obtained are listed in equations shown in (11).

$$\begin{aligned} C =& \frac{1}{KT}\sum_{n,k} KL(\hat{Y}_D(k,n)|Y_D(k,n))+ \\ & \frac{\alpha}{KR}\sum_{k,r}(W_R(k,r) - W_s(k,r)H_1(k))^2 \end{aligned} \tag{10}$$

In the second stage, clean activations are estimated from reverb activations. In order to avoid the trivial solution, as observed in [2], the temporal variation $H_2(n)$ is modeled using an exponential envelope, i.e., $H_2(n) = e^{-\alpha n}$. Figure 3 plots the estimated temporal variation for different measured RIRs. From the figure, it is observed that the variation of $H_2(n)$ can be approximated using an exponential with the decay rate controlled by α. The modified cost function for the second stage is given in (12). Multiplicative updates can be obtained to find the solution. The proposed unsupervised NMF based dereverberation and denoising method is referred to as R-NMF+NMF and the proposed unsupervised NMF based dereverberation method is referred to as R-NMF.

$$\begin{aligned} W_R(k,r) =& W_R(k,r) \\ & \times \left[\frac{\frac{2\alpha T}{R} W_s(k,r)H_1(k) + \sum_n \frac{\hat{y}(k,n)}{y(k,n)} X_R(r,n)}{\frac{2\alpha T}{R} W_s(k,r)H_1(k) + \sum_n X_R(r,n)} \right] \\ W_s(k,r) =& \frac{W_R(k,r)}{H_1(k)} \\ H_1(k) =& H_1(k)\frac{\sum_r W_R(k,r)W_s(k,r)}{\sum_r W_s^2(k,r)H_1(k)} \\ X_R(r,n) =& X_R(r,n) \\ & \times \left[\frac{\frac{2\beta K}{R} X_s(r,n) *_n H_2(n) + \sum_n \frac{\hat{y}(k,n)}{y(k,n)} W_R(k,r)}{\frac{2\beta K}{R} X_R(r,n) + \sum_k W_R(k,r)} \right] \\ X_n(r,n) =& X_n(r,n)\frac{\sum_k \frac{\hat{y}(k,n)}{y(k,n)} W_n(k,r)}{\sum_k W_R(k,r)} \end{aligned} \tag{11}$$

$$C_1 = \sum_{r,n}(X_R(r,n) - X_s(r,n) *_n e^{-\alpha n})^2 \tag{12}$$

2.5. Approaches taken

This section discusses the various frameworks experimented. The baseline method using Beamformit followed by an enhancement using R-NMF+NMF. This approach is taken to remove the residual reverberation and noise present in the output of Beamformit. This method is referred to as Bemformit+R-NMF+NMF. The implementation of MVDR discussed in Sec. 2.1 is referred to as MVDR. Muti-channel MVDR followed by enhancement using R-NMF+NMF is referred to as MVDR+R-NMF+NMF. Alternatively, each channel of multi-channel data can be enhanced using R-NMF+NMF followed

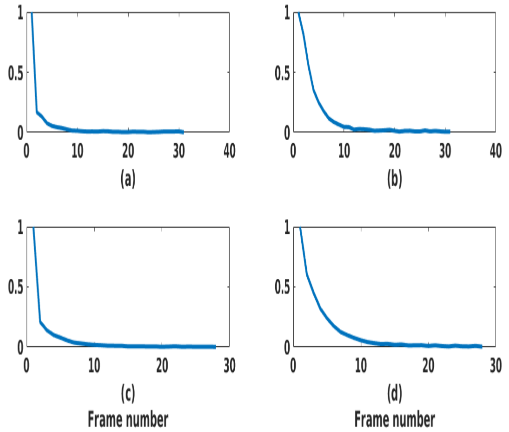

Figure 3: *The estimated temporal variation $H_2(n)$ for various measured RIRs. It is observed that the variations have an exponential-like decay with time.*

by MVDR. This is referred to as R-NMF+NMF+MVDR. The multi-channel WPE enhancement followed by Beamformit is referred to as WPE+Beamformit. Multi-channel WPE followed by R-NMF followed by Beamformit is referred to as WPE+R-NMF+Beamformit.

3. Contributions

Most algorithms used in this work were already available in the literature [3, 2, 4]. The implementation of the NMF based algorithms (R-NMF and R-NMF+NMF) were modified accordingly to handle the CHIME-5 data. Different combinations of multi-channel and single-channel methods are tried. The WER for different combinations were compared.

4. Experimental setup and evaluation

The CHiME-5 array recordings are severely degraded by the presence of multiple speakers, speaker movements, non-stationary noise, and reverberation. Presence of these degradations makes building an enhancement system challenging. In this work, we tried to account for these aspects.

The parameters used in various algorithms is discussed next. The magnitude spectrogram was obtained using a 64 ms Hamming window with a 32 ms hop. The TDOA estimates provided by Beamformit was used in MVDR. The noise covariance matrix was updated for every 2 minutes. The window is selected so that we have sufficiently large silence regions for obtaining the noise covariance matrix, but small so that noise can be assumed to be stationary. The single channel NMF enhanced signal enhancement using NMF is also applied for every 2 minutes. For NMF, 200 bases vectors are learned for clean speech and background. The noise bases vectors are learned from the silence regions of each 2 minutes recordings. Silence regions were obtained using an energy-based VAD. The clean speech bases are learned by NMF decomposition of the original degraded spectrogram, with the noise bases fixed to the value obtained earlier. The enhanced speech magnitude spectrogram is obtained from the clean speech bases and activations. The enhanced speech is obtained by using the enhanced magnitude spectrogram and phase of the original degraded speech. WPE enhancement was performed on 32 ms window with 8 ms hop size. L_h and D were chosen to be 10 and 3. For WPE+R-NMF+Beamformit, the R-NMF is applied for every 10 s data. The shorter window is chosen to account for the change in RIR due to speaker movement.

Enhancement methods in Approach 1 apply a single-channel enhancement on each channel of the microphone array recordings to reduce the effects of reverberation and noise. Beamforming is performed on the enhanced data. The methods include WPE+Beamformit, WPE+R-NMF+Beamformit, and R-NMF+NMF+MVDR. Approach 2 methods use single-channel enhancement as a post-processing step for the beamformed output. These methods include Beamformit+NMF, Beamformit+R-NMF+NMF, MVDR+R-NMF+NMF, WPE+Beamformit, and WPE+R-NMF+Beamformit. Approach 1 and 2 focus on enhancing the the front-end. These algorithms use baseline acoustic and language models. The baseline AM is discussed next.

The baseline acoustic model uses a mixture of both close-talking microphones and array channels data for training. A total of 100k (61349 close-talking utterances and 38651 array utterances) of this mixture were used for training the model, with the utterance timestamps obtained from the transcripts. Following the Ranking A part of the challenge, we focus on using the baseline Gaussian Mixture Model (GMM) - Hidden Markov Model (HMM) and time-delayed neural network (TDNN) acoustic model. Testing is done using the enhanced development data. The enhancement was done for the 'reference' array for a given session of the development data and for each such enhancement, hypothesis transcripts were obtained by feature extraction and decoding of each utterance of this enhanced recording file.

5. Results and discussions

Table 1 discusses the results obtained for various enhancement methods. We were able to reproduce the baseline results. The Beamformit enhanced output has residual noise and reverberation. Approach 1: we performed NMF based enhancement of the individual channels and then used a MVDR beamforming as shown in Figure 1. We expected the method (R-NMF+NMF+MVDR) to give better results based on a sample 10 minutes recording from the given data, as less background noise will have better TDOA estimates. However, contrary to this we obtained poor results when considering the entire data. The performance of MVDR and NMF based approaches depends heavily on the noise estimates. The provided data has very few silence regions and noise estimates are poor. This may have resulted in degraded ASR performance. Enhancement using multi-channel WPE (WPE) was able to produce better the ASR results. The improvement was due to reduced effects of reverberation. This is because WPE does not have a model to handle noise. Due to this observation, we hypothesized that WPE followed by Beamformit (WPE+Beamformit) should handle both reverberation and moise and hence give improved ASR results. However, WPE+Beamformit degraded the performance.

Approach 2: We suppressed the noise using a NMF based denoising method (referred to as Beamformit+NMF) and a NMF based dereverberation and denoising method [2] (referred to as Beamformit+R-NMF). The ASR results do not show any improvement in WER even though perceptually the speech is enhanced.

Table 2 shows the location-based WER for different sessions for dev data. The WER is degraded for all the locations.

Table 1: *Overall WER (%) for the GMM-HMM systems tested on the development test set using baseline acoustic and language model.*

Track	System	WER
Single	Degraded (single-channels)	92.18
	Beamformit (Baseline)	91.33
	Beamformit+NMF	93.94
	Beamformit+R-NMF+NMF	95.51
	MVDR	94.68
	R-NMF+NMF+MVDR	95.56
	MVDR+R-NMF+NMF	94.80
	WPE	92.01
	WPE+Beamformit	94.49
	WPE+R-NMF+Beamformit	97.22

We include the results for R-NMF+NMF+MVDR in Table 2. All these results were obtained without any retraining of the acoustic model, i.e. mismatched conditions.

Table 2: *Results on the development dataset for the GMM-HMM system for R-NMF+NMF+MVDR enhancement technique. WER (%) per session and location together with the overall WER.*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	97.58	96.47	94.56	95.56
		S09	94.77	95.30	94.43	
	Eval	S01	97.06	91.45	108.24	97.36
		S21	96.82	94.30	99.49	

Table 3: *Overall WER (%) for the TDNN systems tested on the development test set using baseline acoustic and language model.*

Track	System	WER
Single	Beamformit (Baseline)	81.1
	R-NMF+NMF+MVDR	92.85
	WPE	83.49
	WPE+Beamformit	86.49
	WPE+R-NMF+Bemformit	84.67

Table 4: *Overall WER (%) for the TDNN systems tested on the development test set using acoustic model obtained from WPE enhanced training data.*

Track	System	WER
Single	WPE	80.16
	WPE+R-NMF+Bemformit	81.23

We have also obtained the TDNN results for the various enhancement methods. Table 3 shows the ASR results obtained for these approaches. We were able to reproduce the baseline. However, the enhancement approaches failed to produce improved results. One of the reasons for the poor performance of the ASR system discussed in Table 1, 3 can be attributed to train-test mismatch (apart from presence of highly non-stationary noise). The training data is the CHiME-5 data and only the Dev data is enhanced. To avoid this mismatch, we modified the AM. The AM was obtained from training data which was enhanced using the same preprocessing technique as done for testing (Approach 3). Table 5 shows results with the matched condition for GMM-HMM system. The use of Beamformit enhanced train data were not able to significantly improve the WER as compared using training done with degraded data. One of the reasons can be that the enhancement methods are not able to completely remove the non-stationary noise which is severely degrading the ASR performance. Such noise is predominantly present in the kitchen and is less in the living room. Based on this observation, we trained a new acoustic model using WPE enhanced data. The utterances from last one hour of the 6-th array (which is placed in the living room) were used for training. The WPE enhanced Dev data was able to show significant improvement in WER as compared to baseline. The results obtained are shown in Tabe 4. The WPE+R-NMF+Beamformit method was also tried out. This method was expected to further improve the ASR as the algorithm had models to handle reverberation and noise. However, the results were poor as compared to WPE.

Table 5: *Results on GMM-HMM system trained and tested using the same enhancement technique (matched condition).* 100 *k utterances of the array processed data was used for training.*

Track	System	WER
Single	Degraded (single-channels)	93.00
	Beamformit	92.14

6. Summary

Three different approaches have been tried to improve the ASR performance on CHiME-5 data. In Approach 1, each channel of the microphone array is enhanced using NMF or WPE based approaches, followed by a beamformer. The motivation was to reduce the effects of noise and reverberation before beamforming is performed. The enhanced channel should give better TDOAs and hence better beamformer performance. Approach 2 uses NMF based enhancement is performed as a post-filtering step to beamforming. Single channel enhancement is expected to remove the residual noise and reverberation present after beamforming. However, these methods give unconvincing results. Train-test mismatch and lack of silence regions to correctly estimate noise were the reasons for the poor performance. In Approach 3 we tried to avoid these shortcomings. Testing is done in a matched condition. The AM is build on enhanced training data, which uses the same enhancement technique as in testing. The Beamformit enhanced AM did not improve the ASR results. However, building an AM using WPE enhanced data of living room alone were able to improve the overall WER for the entire WPE enhanced data. The improvements were better than baseline. Moreover, WPE+R-NMF+Beamformit also performed better using this AM.

7. Acknowledgments

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

8. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

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

[3] 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.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[5] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4029–4032.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.