INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Modulation Vectors as Robust Feature Representation for ASR in Domain Mismatched Conditions

Samik Sadhu[1], *Hynek Hermansky*[1,2]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, USA

samiksadhu@jhu.edu, hynek@jhu.edu

Abstract

In this work, we demonstrate the robustness of Modulation Vectors, in domain mismatches between the training and test conditions in an Automatic Speech Recognition (ASR) system. Our work focuses on the specific task of dealing with mismatches caused by reverberation. We use simulated data from TIMIT and real reverberant speech from the REVERB challenge data to evaluate the performance of our system. The paper also describes a multistream system to combine information from Mel Frequency Cepstral Coefficient (MFCC) and M-vectors to improve the ASR performance in both matched and mismatched datasets. The proposed multistream system achieves a relative improvement of 25% in recognition accuracy on the mismatched condition, while a M-vector trained hybrid ASR system shows a 7-8% improvement in recognition accuracy, both w.r.t. a MFCC trained hybrid ASR system.

Index Terms: speech recognition, feature extraction, modulation spectrum, multi-stream ASR

1. Introduction

All machine learning systems work under the implicit assumption that the data used for testing the system comes from a distribution that is similar to that of the training data. However, such an assumption is not always satisfied in practical situations. In this work, we address the problem of environmental mismatches in reverberation that is common in speech signals. Apart from common sources of variability, such as gender, age and mood of speaker, etc., different environmental conditions pose additional challenges to extract the core invariable information from speech. It is practically impossible to characterize all the different environmental domains that can be a potential source of target speech data. As a result, Automatic Speech Recognition (ASR) systems can get affected from an infinite variety of noise sources (from microphone distortions, car noise, noise from household electronic equipment etc) and reverberations. To deal with vastly unknown domains of data, and improve the robustness of ASR a variety of approaches have been proposed in the ASR literature. Some of these approaches consist of data-augmentation and multi-condition training [1, 2, 3], model adaptation [3, 4, 5, 6, 7] , robust acoustic modelling [8, 9], speech enhancement [10, 11], multi-stream ASRs [12, 13], providing additional noise information or ”noise aware” ASR [14, 15] , as well as robust acoustic feature representation [16, 17]. In this work, we study the robustness of M-vectors as acoustic feature representation for ASR to deal with mismatches in reverberation for practical ASR systems. We describe the steps for computing M-vectors with a analysis on the meaning of all the parameters involved in Sec. 2, the baseline features are described in Sec. 3, followed by a description of the experimental setup in Sec. 4 and experimental results in Sec. 5.

2. Computing Modulations in Speech

M-vector was introduced as a feature representation [18] that can be used in a multi-stream hybrid or end-to-end ASR system for better recognition performance. These representations follow a novel approach to capture the modulations of speech signal energy w.r.t. time in different frequency sub-bands over a long temporal context. In comparison, MFCC features capture frame level spectral energies in different sub-bands. The time evolution of energy in each sub-band is represented by the Hilbert envelope of the band-passed speech signal. The use of sub-band based modulations as features is inspired by previous experiments performed on the effects of these modulation components on human and machine recognition of speech [19]. The features represent the modulations in the range 0-15 Hz in an autoregressive (AR) approximation of the Hilbert envelope in different sub-bands of the speech signal.

2.1. Approximation to Hilbert envelope by FDLP

The Hilbert Envelope is approximated by the technique of Frequency Domain Linear Prediction (FDLP), introduced by Athineos et al. [20]. The technique is established on a theorem which states that applying Linear Prediction (LP) to the Discrete Cosine Transform (DCT) of an even-symmetrized signal results in autoregressive modelling of the Hilbert envelope of the signal.

2.2. Subband based feature extraction

Since the DCT coefficients effectively represent the frequency domain of the signal, we create triangular “MFCC-like” mel-scaled filter-banks to weight the DCT coefficients over a particular sub-band. These weighted DCT coefficients are used for AR modelling of the Hilbert Envelope for that specific sub-band.

2.3. Recursive cepstral computation

The modulation coefficients are obtained as the magnitude spectrum of the logarithm of the approximated Hilbert Envelope. We use a recursive algorithm to compute these coefficients which is similar to the algorithm to efficiently compute the cepstrum of an all-pole model from the LP coefficients [21].

http://dx.doi.org/10.21437/Interspeech.2019-2723

2.4. Frame-wise feature extraction

ASR systems usually use frame rates of about 100 Hz. To derive M-vectors at 100 Hz frame rate we perform the following operations

1. Hanning window the speech signal over a relatively large window length of T seconds around the frame under consideration
2. Compute Type 2 DCT of the windowed signal
3. Window the DCT by appropriate triangular K number of mel scaled weighting functions
4. Derive the FDLP model with model order p in each of the K sub-bands and perform recursive computation on the filter coefficients to get the modulation coefficients
5. Concatenate the modulation coefficients in an appropriate frequency range $[f_{beg}, f_{end}]$ Hz from all the K sub-bands

2.5. Understanding M-vectors

2.5.1. *What do the parameters mean?*

The window length T represents temporal context for computing modulations. The FDLP model allows us to compute the modulations over a large temporal context to form a single feature vector as an alternative to the standard practice of splicing short time feature vectors over a long context. In order to emphasise on the modulations in the center of this large context, we window the speech signal with a T second long Hanning window that decays to zero at the edges. The window length T, also determines the resolution of the modulation coefficients and is given by $\frac{1}{2T}$ Hz. Hence, for a window length of $T = 0.5$ seconds, the first 15 coefficients correspond to 0-15 Hz of modulation (speculations in reference [18] would incorrectly imply 0-8 Hz).

The FDLP model order p represents the level of approximation of the Hilbert Envelope. A higher model order preserves more details in the envelope, while reducing the model order result in smoother approximations.

The frequency range $[f_{beg}, f_{end}]$ is motivated by previous studies that conclude that a the modulations around 4 Hz are most important for speech recognition [19]. The slow modulations of around 2 Hz also carry information about the Hanning window. The windowing effect results in addition of a constant vector equal to the cepstrum of the Hanning window to the features (see fig. 2). However, we perform cepstral mean subtraction before using the features for ASR which gets rid of this constant windowing effect from the features.

2.5.2. *Motivation for using M-vectors in domain mismatch*

The FDLP technique exploits the power of AR models to obtain varied degrees of approximations of the Hilbert Envelope, projecting p to be an important design parameter to ignore corruptions in the Hilbert Envelope caused by environmental disturbances in the speech signal. Ignoring modulation components that are not important for human speech cognition also minimizes the effect of environmental disturbances on the feature representation. These properties motivate us to examine the robustness of these features.

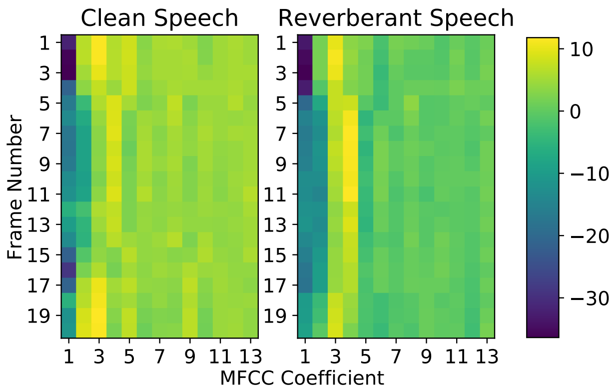

Figure 1: *MFCC coefficients clean Vs reverberation*

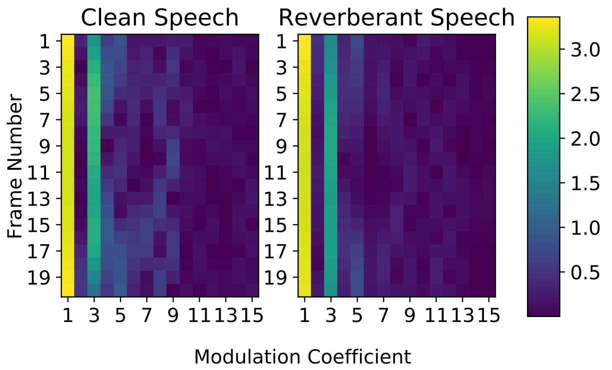

Figure 2: *Modulation Coefficients clean Vs reverberations*

Fig. 1 and 2 show the effect of simulated large room reverberation on MFCC and Modulation Coefficients (from a particular sub-band). Visually, the Modulation Coefficients for reverberant speech appear to preserve similar modulations captured by the clean speech, while the MFCC features for these two conditions look different. The percentage relative change in Frobenius norm for the MFCC features between the two domains shown in fig. 1 is 65.8%, while the same value for Modulation Coefficients in fig. 2 is only 14.08%.

2.5.3. *Suitable parameters for mismatched conditions*

A low model order (in our experiments $15 - 30$ poles) should result in a smoother approximation of the Hilbert Envelope that is expected to ignore domain level details in the Hilbert Envelopes. In section 5.1.1, we support this hypothesis with HMM-GMM ASR models. However, the ASR performance degrades on the matched condition for a lower model order.
In previous studies on modulations in speech [22], it has been shown that the DC or zero frequency component is harmful in reverberant conditions when computing the DC over the entire utterance and a "gain normalization" of the FDLP model is recommended. However, for windows of $T = 0.5$ second, we observed that the inclusion of the DC component improves the recognition accuracy by a small margin (see section 5.1.2).

Table 1: *Effect of Model order on matched and mismatched test sets for TIMIT*

System	MV order 10 (WER %)		MV order 15 (WER %)		MV order 30 (WER %)		MFCC	
	Mismatched	Clean	Mismatched	Clean	Mismatched	Clean	Mismatched	Clean
LDA-MLLT	61.6	36.7	63.2	31.1	64.7	30.3	70.1	24.5
SAT	58.2	34.6	58.8	28.8	60.8	27.4	66.4	22

3. Baseline MFCC Features

To compare the robustness of M-vectors we use the traditional MFCC features computed as follows

1. Window the time domains signal with hamming window of 20 ms length
2. Compute the power spectrum of the signal
3. Compute the log of spectral energy in 20 overlapping triangular mel-scaled filters from 0-$\frac{F_s}{2}$ Hz, F_s being the sampling frequency of the signal
4. Compute DCT of the 20 Mel-bank energies and preserve the first 13 coefficients

4. Experimental Setup

We demonstrate the robustness of M-vectors on HMM-GMM systems, Hybrid HMM-DNN systems as well as a multistream system, all trained using the Kaldi [23] ASR toolkit.

4.1. HMM-GMM

We follow the standard Kaldi recipe to obtain LDA-MLLT and SAT HMM-GMM models with context dependent (triphone) states. We use the alignments from the LDA-MLLT model to train the hybrid systems. With MFCC features, we use a splicing of ± 4 feature frames, however, for M-vectors we perform no splicing operation of the features. We use standard cepstral mean subtraction as implemented in Kaldi for both the features.

4.2. Hybrid HMM-DNN System

We use two different hybrid systems for our experiments, the standard Kaldi `nnet2` TANH network and the recently introduced Time Domain Neural Network (TDNN) model with LF-MMI training [24]. The alignments for the hybrid models are derived form the LDA-MLLT systems. The TANH network is trained with 5 hidden layers, each with 512 hidden units. We use the same TDNN architecture as described in [24]. However, the splicing at the first TDNN layer is changed to zero for M-vectors, and ± 1 for 200 dimensional multistream features (see sec 4.3). The TDNNs are trained without speed-perturbed data augmentation, i-vectors and high resolution features which are included by default in the standard Kaldi recipe. We also use a 10% Cross entropy regularization over the MMI training.

4.3. Multistream System

We use a very simple multistream system [25, 12, 13] without any stream selection algorithm. The pre-softmax state posteriors from the TANH hybrid model in each stream on the training data are reduced to 100 dimensions by Principal Component Analysis (PCA) and concatenated to form 200 dimensional features for a second stage of hybrid ASR training. We use both `nnet2` TANH network and TDNN to train our second stage hybrid model. We also do cepstral mean normalization for this second stage of ASR training and no feature splicing for the TANH system. However, we observed a ± 1 feature splicing for the TDNN model worked the best.

4.4. M-vector Configuration

We use a window length $T = 0.5$ seconds and experiment with the model order on the matched and mismatched tasks for the TIMIT database. For the REVERB data, however, the model order is fixed at $p = 30$. We also use 15 modulation coefficients for a window length of 0.5 seconds and correspondingly change it for higher and lower window lengths to represent the same range of modulation frequency. The number of sub-bands are fixed at $K = 20$ for both MFCC and M-vector features.

5. Results

5.1. Experiments on TIMIT

We train LDA-MLLT and SAT HMM-GMM models with all the 3696 utterances in the clean TIMIT training set. A simulated reverbant test set is generated by convolving the 192 utterances in the TIMIT core test set with a large room reverberation impulse response (Reverberation time $\approx$ 700 ms) from the REVERB challenge [26]. We call this the "mismatched" set and keep the original 192 matched test utterances as the "clean" set. Since hybrid systems do not show any significant improvement for small datasets like TIMIT, we show the results on two standard HMM-GMM systems from the Kaldi toolkit.

5.1.1. Effect of model order

Table 1 shows how variation in model order results in a compromise between the performance on mismatched and clean test sets. The results show that with a decrease in model order, the recognition accuracy for the mismatched set improves, while the performance on the clean case becomes progressively worse. For mismatched conditions, M-vectors can improve the WER from 70.1% for MFCC to 61.6% for the LDA-MLLT system, however at the loss of the matched condition WER increasing from 24.5 % to 36.7 %. A similar trend can be observed for the SAT system also. This motivates us to look at a multi-stream architecture to fuse the information from both the MFCC and M-vector features as described in section 4.3. We will look at such a multi-stream system for our experiments with the REVERB challenge data in section 5.2.

5.1.2. Effects of gain coefficient

It has been reported in the literature that gains capture the linear distortions in speech and it is beneficial to remove it from feature representations, particularly for reverberant conditions [27, 28]. For other similar studies, the gain usually has been

Table 2: *Results on the REVERB challenge data*

Model	Type	et_simu_1ch (WER %)				et_real_1ch (WER %)			
		Avg Far		*Avg Near*		*Far*		*Near*	
		MFCC	**MV**	**MFCC**	**MV**	**MFCC**	**MV**	**MFCC**	**MV**
HMM-GMM	LDA-MLLT	26.97	25.7	16.43	18.98	52.9	47.47	54.14	47.52
	SAT	22.31	22.67	13.01	16.98	44.46	40.38	41.78	42.99
Hybrid	TANH	16.81	17.29	11.31	11.91	42.98	39.91	40.72	37.27
	TDNN	9.9	9.2	6.2	6.5	58.7	53.9	56.9	53.5
Hybrid Multi-stream	TANH	15.58		10.97		36.12		33.95	
	TDNN	11.31		7.88		31.97		30.31	

computed over the entire utterance [22, 29]. In table 3 we show the recognition accuracy for the LDA-MLLT system for different window lengths T with or without the DC coefficient. It can be seen that the effect of the gain on performance decreases as we increase the window length from 0.25 seconds to 1.5 seconds, thereby corroborating to earlier studies. However, since for most of our experiments we choose an optimum window length of 0.5 seconds, we choose to preserve the gain coefficient.

Table 3: *Effect of gain on mismatched set recognition accuracy*

T (in secs)	WER (%)	
	With Gain	**Without Gain**
0.25	63.2	83.4
0.5	63.2	70.7
1	63.4	66.7
1.5	66.5	64.2

5.2. Experiments on REVERB

We test our features with the single channel data from the REVERB challenge [26], which creates a specially interesting and practical mismatched condition, using simulated reverberant speech for training an ASR and real reverberant speech for testing it. The simulated data is generated by convolving the speech signal with six different types of reverberation conditions consisting of three room sizes and two types of distances to the microphone (near $\approx$ 50 cm and far $\approx$ 200 cm), in addition to adding small amount of noise. In Kaldi naming conventions, we use the single channel simulated data set `tr_simu_1ch` for training the ASR models and test the system on single channel real and simulated reverberant speech from `et_real_1ch` and `et_simu_1ch` respectively. The real testing data has been collected in two real reverberant scenarios in one single room with two different distances to the microphone. The results in table 2 are organized into performance for the near and far microphone scenarios. For the simulated set `et_simu_1ch` we show the average performance for near and far microphone for all three room sizes.

5.2.1. Observations

The low model order of the M-vector features results in sightly worse performance on the matched test set `et_simu_1ch` compared to MFCC features. However, there is a relative gain of 7-8% WER for the mismatched test set `et_real_1ch` for the near room and far room scenarios using the TANH Hybrid model. The multi-stream hybrid system, effectively sees the posteriors from both the MFCC and M-vector streams and this results in better recognition accuracy for both matched and mismatched conditions. The results for the mismatched condition being particularly noticeable with a 25% relative improvement in WER from 40.72 % to 30.31 % of the multi-stream TDNN system over a TANH Hybrid system with MFCC features.

5.2.2. Behaviour of TDNN

The TDNN models trained on individual streams appear to have severe over-fitting on the simulated data domain for both MFCC and M-vector features. This results in the large disparity in performance between the two domains as can be seen in table 2. It shows that neural networks discriminatively trained with raw features and without a considerable variety of data do not generalize well across domains. However, the TDNN for the second stage in the multistream system with "better behaved" posterior features, performs much more reasonably across domains, and in fact, achieves the best performance recorded in our experiments.

6. Conclusions

In this work, we show the usefulness of M-vectors for domain mismatches in ASR, particularly mismatches in reverberation. We show that we can reduce the FDLP model order to very low values of $p = 10$ for a 0.5 second window to get as much as a 12 % relative improvement in WER on simulated reverberation mismatches with the TIMIT database. For a moderately low model order of $p = 30$, we can also get a 7-8 % WER improvement on real reverberant data. We also show that a simple multi-stream system can result in a significant improvement for both matched and mismatched conditions, with a 25 % relative improvement in WER on the single channel REVERB dataset.

7. Acknowledgements

The work was supported by the National Science Foundation under EAGER Grant No. 1743616 and by a gift from Google Inc.

8. References

[1] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5220–5224.

[2] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] R. Hsiao, J. Ma, W. Hartmann, M. Karafiát, F. Grézl, L. Burget, I. Szöke, J. H. Černockỳ, S. Watanabe, Z. Chen *et al.*, "Robust speech recognition in unknown reverberant and noisy conditions," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 533–538.

[4] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba *et al.*, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, no. 3, pp. 851–873, 2013.

[5] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7893–7897.

[6] R. Hsiao, T. Ng, S. Tsakalidis, L. Nguyen, and R. Schwartz, "Unsupervised adaptation for deep neural network using linear least square method," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[8] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU aspire system: Robust LVCSR with TDNNS, ivector adaptation and RNN-LMs," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 539–546.

[9] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5532–5536.

[10] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[11] 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.

[12] S. H. Mallidi and H. Hermansky, "Novel neural network based fusion for multistream asr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5680–5684.

[13] S. H. Mallidi, T. Ogawa, K. Veselỳ, P. S. Nidadavolu, and H. Hermansky, "Autoencoder based multi-stream combination for noise robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] D. Dimitriadis, S. Thomas, and S. Ganapathy, "An investigation on the use of i-vectors for robust asr." in *INTERSPEECH*, 2016, pp. 3828–3832.

[15] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[16] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4117–4120.

[17] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4286–4289.

[18] S. Sadhu, R. Li, and H. Hermansky, "M-vectors: Sub-band based energy modulation features for multi-stream automatic speech recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[19] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 140–147.

[20] M. Athineos and D. P. Ellis, "Frequency-domain linear prediction for temporal features," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 261–266.

[21] K. Tokuda, T. Kobayashi, and S. Imai, "Recursive calculation of mel-cepstrum from lp coefficients," *Trans. IEICE*, vol. 71, pp. 128–131, 1994.

[22] S. H. Mallidi, S. Ganapathy, and H. Hermansky, "Modulation spectrum analysis for recognition of reverberant speech," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[25] S. H. R. Mallidi and H. Hermansky, "A framework for practical multistream asr." in *INTERSPEECH*, 2016, pp. 3474–3478.

[26] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[27] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[28] H. Hermansky, N. Morgan, and H.-G. Hirsch, "Recognition of speech in additive and convolutional noise based on rasta spectral processing," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 83–86.

[29] S. Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," Ph.D. dissertation, Johns Hopkins University, 2012.