ROBUST FEATURES AND SYSTEM FUSION FOR REVERBERATION-ROBUST SPEECH RECOGNITION

Vikramjit Mitra[1], Wen Wang[1], Yun Lei[1], Andreas Kathol[1], Ganesh Sivaraman[2], Carol Y. Espy-Wilson[2]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
[2]University of Maryland, College Park, MD
[1]{vmitra, wwang, yunlei, kathol}@speech.sri.com, [2]{ganesa90, espy}@umd.edu

ABSTRACT

Reverberation in speech degrades the performance of speech recognition systems, leading to higher word error rates. Human listeners can often ignore reverberation, indicating that the auditory system somehow compensates for reverberation degradations. In this work, we present robust acoustic features motivated by the knowledge gained from human speech perception and production, and we demonstrate that these features provide reasonable robustness to reverberation effects compared to traditional mel-filterbank-based features. Using a single-feature system trained with the data distributed through the REVERB 2014 challenge on automatic speech recognition, we show a modest 12% and 0.2% relative reduction in word error rate (WER) compared to the mel-scale-feature-based baseline system for simulated and real reverberation conditions. The reduction is more pronounced when three systems are combined, resulting in a relative 20% reduction in WER for the simulated reverberation condition and 11.7% for the real reverberation condition compared to the mel-scale-feature-based baseline system. The WER was found to reduce even further with addition of more systems trained with robust acoustic features. HLDA transform of features and MLLR adaptation of speaker clusters were also explored in this study and both of them were found to improve the recognition performance under reverberant conditions.

Index Terms— *feature combination, robust speech recognition, reverberation robustness, robust features.*

1. INTRODUCTION

Automatic speech recognition (ASR) systems are sensitive to speech signal degradations such as background noise and/or reverberation, which can result in significant performance drops. Studies [1] have demonstrated that ASR performance is seriously affected by reverberation, which is a very common artifact introduced by the speaker's environment. The room or enclosed environment where the speech signal is recorded primarily defines the character of the reverberation. Reverberation is typically caused by multiple reflections of the source sound from the ambient enclosure, and such distortions seriously degrade speech signal quality. Recently, de-reverberation or mitigation of artifacts from reverberation has been an important research topic, with microphone array processing [2]; acoustic echo cancellation [3]; reverberation-robust signal processing [4]; and speech enhancement [5] serving as major research areas.

Typically, when ASR systems trained on clean data (i.e., data without noise or reverberation artifacts) are deployed in reverberant conditions (where the acoustic signal is significantly distorted by the reverberant background), ASR performance sharply falls. This effect can be mitigated by training the ASR systems with reverberant data, which usually reduces the performance mismatch between the training and evaluation data [6].

Robust acoustic features have been explored in [7, 8] for improving reverberation robustness of ASR systems. Robust signal-processing techniques have been found to mitigate the effect of reverberation and to improve speech recognition performance. In this work, we present a set of robust acoustic features that demonstrate appreciable robustness to reverberation compared to standard mel-frequency cepstral coefficients (MFCCs). Human listeners can overlook reverberation effects, indicating that human auditory processing involves filtering that helps to mask (to some extent) the reflections from the reverberant room or environment. In our approach, we propose to use acoustic features motivated by human auditory perception and to demonstrate that these features improve the reverberation robustness of conventional ASR systems. We also explore articulatory features, where the articulatory information is derived from acoustic observations by using a deep neural network (DNN), and our results indicate that these features also provide some degree of reverberation robustness through system fusion. We also demonstrate that heteroscedastic linear discriminant analysis (HLDA) transform of large dimensional acoustic features (features that have more contextual information than what is traditionally used) can help to reduce WERs significantly compared to the same system trained with lower dimensional feature (cepstral features with velocity (Δ) and acceleration (Δ^2) coefficients) containing less contextual information. Model space maximum likelihood linear regression (MLLR) speaker-cluster adaptation of the acoustic models (where the speaker clusters were learnt from unsupervised clustering of the acoustic data (for each test/train condition separately) using gaussian mixture models (GMMs)) also contributed significantly to the reduction in the WERs.

In this work, we used the data distributed through the REVERB (REverberant Voice Enhancement and Recognition Benchmark) 2014 challenge [1] to train and evaluate our systems. We trained separate systems with different features and observed that all of these systems performed better than the MFCC-baseline system in almost all reverberant conditions. Fusing these systems by using ROVER [9] further improved performance, resulting in a significant reduction in the overall word error rate (WER) compared to the baseline system.

The paper is structured as follows: First, in Section 2, we briefly describe the REVERB 2014 challenge dataset used in our experiments. In Section 3 we present the different feature-extraction strategies used in our work. In Section 4, we present results from our individual systems and from the system combination experiments. Finally, in Section 5, we present our conclusions.

2. DATASET AND TASK

Our experiments used the REVERB 2014 challenge speech dataset, which contains single-speaker utterances recorded with one-channel, two-channel, or eight-channel circular microphone arrays. The dataset includes a training set, a development set, and an evaluation set. The training set consists of the clean WSJCAM0 [10] dataset, which was convolved with room impulse responses (with reverberation times from 0.1 sec to 0.8 sec) and then corrupted with background noise. The evaluation and development data contain both real recordings (real data) and simulated data (sim data). The real data is borrowed from the MC-WSJ-AV corpus [11], which consists of utterances recorded in a noisy and reverberant room. For the sim data, reverberation effects were artificially introduced. More details about the dataset are provided in [1]. We used the channel-1 training data to build our acoustic models, which contained altogether 7861 utterances (5699 unique utterances). The simulated dev set had 742 utterances in each of far and near microphone conditions almost equally spread in three room types (1, 2, and 3) and the real dev set had 179 utterances almost equally spread into near and far microphone conditions. The simulated evaluation set contained 1088 utterances in each of the far and near microphone conditions, each of which were split into three room conditions (1, 2 and 3). The real evaluation set contained 372 utterances split equally between near and far microphone conditions.

The REVERB 2014 challenge consists of two parts: (1) speech enhancement and (2) ASR. This paper presents the performance of our system for the ASR task only. In our work, we used SRI's DECIPHER large-vocabulary, continuous speech recognition (LVCSR) system to train and test the acoustic models. We explored different feature-based acoustic models and used only one-channel training data to build these models. No speaker information was used during acoustic model training and testing, and all processing was independent of the room impulse responses and the relative position of the speakers with respect to the recording device. We report our results in terms of WER using conditions identical to those of the baseline system distributed with the REVERB 2014 challenge data.

3. ACOUSTIC FEATURES

We explored an array of robust features for our experiments, motivated by human auditory perception and speech production. The features explored are briefly outlined in this section.

3.1 Damped Oscillator Cepstral Coefficients (DOCC)

DOCC [12] aims to model the dynamics of the hair cells within the human ear. The hair cells detect the motion of incoming sound waves and excite the neurons of the auditory nerves. In DOCC processing, the incoming speech signal is analyzed by a bank of gammatone filters (in this work, we used a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale), which splits the signal into bandlimited subband signals. In turn, these subband signals are used as the forcing functions to an array of damped oscillators whose response is used as the acoustic feature. More details about damped oscillator processing and the DOCC pipeline can be obtained in [12]. We analyzed the damped oscillator response by using a Hamming analysis window of ~26 ms with a frame rate of 10 ms. The power signal from the damped oscillator response was computed, then root compressed using 15^{th} root and discrete cosine transformed (DCT). We retained the first 13 DCT coefficients and appended those with their Δ, Δ^2, and Δ^3 coefficients, yielding a 52-dimensional DOCC feature vector.

3.2 Normalized Modulation Cepstral Coefficients (NMCC)

NMCC [13] is motivated by the fact that amplitude modulation (AM) of subband speech signals plays an important role in human speech perception and recognition [14, 15]. These features were obtained from tracking the amplitude modulations of subband speech signals in a time domain using a Hamming window of ~26 ms with a frame rate of 10 ms. In this processing, the speech signal was analyzed using a time-domain gammatone filterbank with 34 channels equally spaced on the ERB scale. The subband signals from the gammatone filterbanks were then processed using the Discrete Energy Separation algorithm (DESA) [16], which produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed using 15^{th} root and their DCT coefficients were generated, from which only the first 13 coefficients were selected. These 13 coefficients along with their Δ, Δ^2, and Δ^3 coefficients yielded a 52-dimensional NMCC feature vector that was used in our experiments.

3.3 Modulation of Medium Duration Speech Amplitudes (MMeDuSA)

Like the NMCC features, the MMeDuSA [24] features aim to track the subband AM signals of speech, but they use a medium duration analysis window and also track the overall summary modulation. The summary modulation plays an important role in tracking speech activity as well as in locating events such as vowel prominence/stress, etc. [17]. The MMeDuSA-generation pipeline used a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. It employed the nonlinear Teager energy operator [18] to crudely estimate the AM signal from the bandlimited subband signals. The MMeDuSA pipeline used a medium duration hamming analysis window of ~51 ms with a 10 ms frame rate and computed the AM power over the analysis window. The powers were root compressed and then their DCT coefficients were obtained, out of which the first 13 coefficients were retained. These 13 cepstral coefficients and their Δ, Δ^2, and Δ^3 coefficients resulted in a 52-dimensional feature set. Additionally, the AM signals from the subband channels were bandpass filtered to retain the modulation information within the range 5 to 200 Hz, which was then summed across the frequency scale to produce a summary modulation signal. The power signal of the modulation summary was obtained, followed by 15^{th} root compression. The result was transformed using DCT and the first three coefficients were retained and combined with the previous 52-dimensional features to produce the 55-dimensional MMeDuSA features.

3.4 Gammatone Cepstral Coefficients (GCCs)

The gammatone filters are a linear approximation of the auditory filterbank performed in the human ear. In GCC processing speech is analyzed using a bank of 40 gammatone filters equally spaced on the ERB scale. The power of the band limited time signals within an analysis window of ~26 ms was computed at a frame rate of 10 ms. Subband powers were then root compressed using 15^{th} root and DCT was performed on the resultant. The first 13 DCT

coefficients were retained and their Δ, Δ^2, and Δ^3 were appended to generate a 52-dimensional feature vector.

Note that we also explored unsupervised speaker clustering based vocal tract length normalization (VTLN), where the training, development and test data were individually clustered using unsupervised Gaussian mixture models (GMMs) and the speaker clusters were used to compute the vocal tract lengths. The speaker clusters in our experiments were learnt separately for each train/test conditions, hence no cross condition information was used in obtaining the speaker clusters.

3.5 Articulatory Features

Articulatory features aim to model the co-articulatory variabilities in spontaneous speech, and previous studies [19-20] have demonstrated that they provide a sufficient degree of robustness for ASR tasks. We trained a thin deep neural network (DNN) with 150, 200, 100, 80, 60, and 40 neurons, where the input observations were time-contextualized NMCC features (generated from the acoustic waveform), and the outputs were time-domain vocal-tract constriction variables also abbreviated as TVs. Due to a lack of any natural speech dataset containing parallel data of acoustic waveforms and TVs, we used Haskins Laboratories' Task Dynamic model (known as TADA [21]) to generate a synthetic English isolated-word speech corpus along with the TVs. TADA was used to generate a synthetic word corpus of 111,929 words, where the words were borrowed from the CMU dictionary. TADA generated the corresponding TVs (eight vocal-tract constriction variables corresponding to Lip Aperture (LA); Lip Protrusion (LP); Tongue Tip Constriction Degree (TTCD); Tongue Tip Constriction Location (TTCL); Tongue Body Constriction Degree (TBCD); Tongue Body Constriction Location (TBCL); Velic Opening (VEL); and Glottal Opening (GLO)). 80% of the synthetic data was used for training the DNN; 10% was used as the cross-validation set; and the remaining 10% was used to test the DNN. More details regarding the architecture are provided in [22].

Note that we trained the DNN using an artificially generated synthetic word corpus (clean condition), then these networks were deployed on the REVERB 2014 challenge data. No information regarding noise or reverberation was used while training the DNN. The modulation information from the eight estimated TVs (estimated from the DNN) was combined with the standard MFCC features (as detailed in [19, 22, 25]), and after principal component analysis (PCA)-based dimensionality reduction the resultant MFCC+ModTV feature was used as the feature in our experiments. After PCA, the resultant feature had 30 dimensions, and we call these features MFCC+ModTV_pca30.

4. ACOUSTIC MODEL

For the acoustic model training, we used SRI International's DECIPHER® LVCSR system, which employs a common acoustic frontend that computes 13 MFCCs (including energy) and their Δs, Δ^2s, and Δ^3s. Global mean and variance normalization was performed on the acoustic features prior to acoustic model training. The acoustic models were trained as crossword triphone HMMs with decision-tree-based state clustering that resulted in 2048 fully tied states, and each state was modeled by a 64-component Gaussian mixture model. The model used three states (left-to-right) per phone. For the experiments presented here, all models were trained with maximum likelihood estimation. The system used the 5K non-verbalized punctuation, closed vocabulary set language model (LM), where a bigram LM is used in the initial pass of decoding to generate the first-pass hypotheses. In the second pass, model-space MLLR adaptation of the cross-word acoustic models was conducted based on the first-pass hypotheses. The adapted acoustic models were used for generating bigram HTK lattices, which were then rescored by the trigram LM. Both bigram and trigram are the 5K closed vocabulary non-verbalized-punctuation LMs trained on the WSJ CSR LM training data [26]. A detailed description of the ASR system is provided in [23]. Note that we performed HLDA transform on all features (MFCC [52D], NMCC [52D], DOCC [52D], GCC [52D], and MMeDuSA [55D]) to reduce their dimension to 39D before training the acoustic model. For the 30D MFCC+ModTV_pca30 features, the HLDA transform was not used. The classes for HLDA were obtained from the forced-alignment of training data and the HLDA matrix was learnt using the 1-channel training data only. We trained our models using only the one-channel training data.

5. RESULTS

We present our results for two baseline systems: (1) the MFCC-HTK system distributed through the REVERB 2014 challenge website and (2) the DECIPHER-MFCC system that we trained. We used the development data to analyze the impact of the HLDA transform on WER. In the first case, we used 39D MFCCs (13 cepstra with their first two Δs) where HLDA transform was not applied; in the second case, we used 52D MFCCs that were HLDA transformed to 39D. Our experiments revealed an average 7% or more relative reduction in WER on the real development data when HLDA was applied. Table 1 presents the results from all the baseline MFCC systems that were trained.

Table 1. WERs from different baseline systems using real development data for decoding

Acoustic model	Feature	adaptation	WER (%)		
			far	near	avg.
HTK [REVERB2014]	MFCC	none	51.5	53.7	52.6
	MFCC	cMLLR	46.0	47.3	46.7
DECIPHER [SRI]	MFCC	none	50.1	52.3	51.2
	MFCC	MLLR	43.7	45.3	44.5
	MFCC+HLDA	none	46.2	49.1	47.7
	MFCC+HLDA	MLLR	41.2	39.9	40.5

Table 1 shows that our baseline system (DECIPHER) trained with 39D MFCCs behaved similarly as the HTK baseline system distributed with the REVERB2014 challenge data, where our results are slightly better (rel. WER reduction of 2.6%) than the HTK baseline for the non-adapted conditions. We used MLLR adaptation in our experiments and with 39D MFCCs the relative WER was 4% lower than the cMLLR results from the HTK baseline. The most interesting result is the role of the HLDA transform, in both the adapted and non-adapted conditions, HLDA transform of 52D MFCCs brought about a significant reduction in WERs compared to their non HLDA counterparts, which were roughly 7% and 9% relative WER reductions for non-adapted and MLLR adapted cases. This result indicates that using a larger context (i.e., additional Δ^3 coefficients) is beneficial for reverberation robustness, as the extra contextual information followed by HLDA transform reduces the time–scale distortions introduced by the reverberation effects. Based on these

observations, we always used Δ^3 coefficients in our features followed by HLDA transform, except for the 30D MFCC+ModTV_pca30 feature, where HLDA was not performed.

Table 1 shows that our baseline system performed similarly as the REVERB2014-HTK system, our MLLR adaption brought about similar gains as the cMLLR adaptation for HTK models. HLDA transform on features and MLLR adaption brought in significant improvement to our baseline and hence forth we will be treating DECIPHER with HLDA and MLLR adaptation as our final baseline system.

Table 2 presents result from the baseline systems and the individual systems that was used in our REVERB 2014 submission. We used full batch processing in all our experiments, where no prior information about the speakers, room conditions, or background noise was employed. In full batch processing (as defined by the REVERB 2014 guidelines) all utterances from a single test condition (room type and near/far position in the room) can be used to improve the performance, which allows multiple passes on the data of a single test condition until the final results are achieved.

We used two-way ROVER to combine the outputs from the DOCC and MFCC-TV_pca30 systems (shown as DOCC+MFCC-TV_pca30-ROVER in table 1). We also performed a three-way ROVER combination of the MFCC, DOCC, and MFCC-TV_pa30 systems (represented as MFCC+DOCC+MFCC-TV_pa30-ROVER in table 2). As is evident from table 3, the three-way combination gave the best overall WER for all conditions. Note that in both of these ROVER combinations, the individual subsystems were weighed equally.

Table 2 shows that, by using DECIPHER with HLDA-transformed MFCCs and model space MLLR-based speaker cluster adaptation (where we did unsupervised speaker clustering using GMM models), we could reduce the relative WER by 26% and 13% for the simulated and real test conditions, respectively, compared to the MFCC-HTK cMLLR system [1] distributed through the REVERB 2014 challenge. Henceforth, we will refer to the DECIPHER-MFCC system as the final baseline with which we will compare the performance both of our individual feature-based systems and of the ROVER-fused systems.

Table 2 shows that the DOCC features provide a relative reduction of 13% and 0.2% WER for the simulated and real test conditions compared to the final baseline system. Note that, although the MFCC-TV_pca30 system did not show any relative reduction in WER compared to the final baseline system, it provided significant improvement in accuracy when combined using ROVER with the DOCC features, where the ROVER combination of these two systems show significant reduction in relative WER of 16% and 6% compared to the final baseline for the simulated and real test data. The two-way fusion of the DOCC and MFCC-TV_pca30 features gave relatively lower WERs compared to the standalone DOCC-feature system, indicating that these two systems have complementary. Finally, the best result was obtained from the three-way fusion of the final baseline MFCC, DOCC, and MFCC-TV_pca30 feature-based systems, where we obtained a relative WER reduction of 20% and 12% for the simulated and real test data, respectively, relative to the final baseline.

After submitting our results to the REVERB 2014 challenge, we explored the remaining three features (NMCC, MMeDuSA and GCC) and found that they also contributed to WER reduction compared to the final baseline. The post-submission results are shown in tables 3 and 4, where we show the performance of the individual NMCC, MMeDuSA and GCC systems and their ROVER combination with other systems. We also explored VTLN on each feature, where the VTLs were learnt through unsupervised speaker clustering of the data; surprisingly we did not observe any reduction in WER due to VTLN.

Table 3 shows the individual feature based system performance on the real dev. data, where the models were trained with SRI's DECIPHER system. Note that in our actual REVERB submission NMCC, MMeDuSA and GCC features were not used, only MFCC, DOCC and MFCC-TV_pca30 features were used. From table 3 we can see that for real development data NMCC overall performed the best amongst all the features. Table 4 shows that for evaluation data both GCC and DOCC performed much better than the other features.

We performed m-way ROVER combination amongst all the systems trained during pre- and post-REVERB challenge submission (where m=2, 3, 4, 5 and 6) and we show the results in the last 5 rows of tables 3 and 4. We did an exhaustive search for the best ROVER combination for each m-way ROVER and selected the m systems that produced the best m-way ROVER on the development set. Then we conducted ROVER on the same optimized combination on the evaluation set. We did not tune ROVER weights for the development set nor the evaluation set; all systems in m-way ROVER were assigned equal weights. For each individual system, we also did not tune any weights for the development set and the evaluation set. The weights for acoustic model, language model, and word insertion penalty all stayed constant for all test sets.

The ROVER results from the best m-way ROVER for the real dev. set is shown in table 3. Note that for the actual REVERB submission we did not perform any selection of best systems for ROVER fusion, we had three systems (based on MFCC, MFCC-TV_pca30 and DOCC) and submitted the results from the three-way fusion of these three systems.

From table 3 we can see that for the real dev. set, the lowest WER was obtained from the four-way ROVER fusion of DOCC, GCC, MFCC and NMCC systems, where the relative overall WER reduction was 12% compared to the final MFCC-baseline system. The best single feature system (NMCC) provided an average 6% relative WER reduction over the final baseline system for the real dev. data.

Table 4 shows that for the evaluation data we obtained the lowest WER for simulated data from the 6-way ROVER combination system, where the relative WER reduction was 21% compared to the final baseline. Interestingly for the real data the lowest overall WER was obtained from the 5-way ROVER-combination system, where we obtained a relative WER reduction of 15% compared to the final baseline system. Besides the ROVER-based combination of systems trained on different front-end features, we also plan to investigate the efficacy of the feature-level combination on the various features in the future work.

6. CONCLUSION

In this work, we presented several different features that successfully demonstrated reverberation robustness for the REVERB 2014 challenge dataset. We demonstrated that robust features motivated by human speech perception and production can improve reverberation robustness without using any prior information about the room/environment or its impulse response. We also demonstrated that using higher order delta (Δ^3) followed by HLDA-based dimensionality reduction can significantly reduce

Table 2. WERs on the evaluation set from the different systems (one-channel training and full-batch processing) submitted to the REVERB 2014 challenge.

FEATURES	WER (%)									
	sim data							real data		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
MFCC(39)-HTK	14.21	17.45	21.07	37.19	22.73	40.28	25.49	46.97	47.37	47.17
MFCC	12.83	12.10	13.99	25.49	16.81	32.61	18.97	41.90	39.87	40.89
DOCC	8.64	9.88	12.85	23.43	14.08	30.32	16.53	40.85	40.75	40.80
MFCC-TV_pca30	9.79	11.22	14.20	29.02	18.36	40.44	20.51	43.53	44.16	43.85
DOCC+MFCC-TV_pca30-ROVER	7.42	8.98	11.83	22.87	14.06	30.98	16.02	38.61	38.45	38.53
MFCC+DOCC+MFCC-TV_pa30-ROVER	**7.83**	**8.71**	**11.14**	**21.34**	**13.27**	**28.51**	**15.14**	**36.44**	**35.75**	**36.10**

Table 3. WERs on the real data of the **development set** from the different systems (one-channel training and full-batch processing) from work after the REVERB 2014 challenge, and the combination configuration producing the best m-way ROVER for each m = 2, 3, 4, 5 and 6.

FEATURES	WER (%)		
	real data		
	Room 1		Ave.
	Near	Far	
MFCC(39)-HTK [REVERB 2014 baseline]	46.00	47.29	46.65
MFCC [SRI's baseline]	39.93	41.15	40.54
MFCC-TV_pca30	46.04	46.75	46.40
DOCC	41.05	40.94	41.00
NMCC	36.81	39.10	37.96
MMeDuSA	44.04	50.17	47.11
GCC	38.24	40.40	39.32
2-way-ROVER (opt: GCC+MFCC)	34.56	38.28	36.42
3-way-ROVER (opt: GCC+MFCC+NMCC)	33.44	37.94	35.69
4-way-ROVER (opt: DOCC+GCC+MFCC+NMCC)	33.69	37.32	**35.50**
5-way-ROVER (opt: DOCC+GCC+MFCC+MFCCTV+NMCC)	34.62	**36.91**	35.77
6-way-ROVER (all subsystems)	35.50	37.87	36.69

Table 4. WERs on the **evaluation set** from the different systems (one-channel training and full-batch processing) from post-submission to the REVERB 2014 challenge, and the combination configuration producing the best m-way ROVER for each m = 2, 3, 4, 5 and 6.

FEATURES	WER (%)									
	sim data							real data		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
MFCC(39)-HTK [REVERB 2014 baseline]	14.21	17.45	21.07	37.19	22.73	40.28	25.49	46.97	47.37	47.17
MFCC [SRI's baseline]	12.83	12.10	13.99	25.49	16.81	32.61	18.97	41.90	39.87	40.89
MFCC-TV_pca30	9.79	11.22	14.20	29.02	18.36	40.44	20.51	43.53	44.16	43.85
DOCC	8.64	9.88	12.85	23.43	14.08	30.32	16.53	40.85	40.75	40.80
NMCC	10.03	11.32	14.29	29.78	17.62	39.57	20.44	42.03	40.95	41.49
MMeDuSA	9.62	11.06	13.11	26.40	16.31	34.53	18.51	46.92	45.17	46.05
GCC	9.78	11.78	13.12	27.09	16.77	36.34	19.15	39.12	41.22	40.17
2-way-ROVER (opt: GCC+MFCC)	8.91	9.61	11.43	22.17	13.89	30.17	16.03	35.84	36.36	36.10
3-way-ROVER (opt: GCC+MFCC+NMCC)	8.56	9.62	11.64	23.40	14.04	32.09	16.56	36.09	36.87	36.48
4-way-ROVER (opt: DOCC+GCC+MFCC+NMCC)	7.52	8.69	10.74	21.11	12.92	29.45	15.07	34.78	35.18	34.98
5-way-ROVER (opt: DOCC+GCC+MFCC+MFCCTV+NMCC)	7.25	8.34	10.66	21.51	12.92	29.53	15.04	34.40	35.01	**34.71**
6-way-ROVER (all subsystems)	7.22	8.40	10.55	21.24	12.92	29.62	**14.99**	35.20	35.45	35.33

the WER of the baseline systems in reverberant conditions. Finally, we have shown that the fusion of multiple systems with a diverse set of features motivated by speech perception and production can significantly improve ASR performance, where we obtained 21% and 15% relative reduction in overall WER for simulated and real evaluation data.

7. ACKNOWLEDGMENT

This research was partially supported by NSF Grant # IIS-1162046.

8. REFERENCES

[1] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[2] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer Verlag, 2001.

[3] R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation and Noise Reduction: A Two Microphone Approach," *Journal of Annales des Télécommunications*, Vol. 49, Iss. 7–8, pp. 429–438, 1994.

[4] K. Ohta and M. Yanagida, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-Wise Time Sequences of Frequency Components," *Proc. of IWAENC*, pp. 1–4, 2006.

[5] M. Wu and D.L. Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement," *IEEE Trans. Aud. Speech & Lang. Process.*, Vol. 14, No. 3, pp. 774–784, 2006.

[6] L. Couvreur and C. Couvreur, "Robust Automatic Speech Recognition in Reverberant Environments by Model Selection," *Proc. of HSC*, pp. 147–150, 2001.

[7] A. Sehr and W. Kellermann, "A New Concept for Feature-Domain Dereverberation for Robust Distant-Talking ASR," *Proc. of ICASSP*, pp. 369–372, 2007.

[8] M. Delcroix and S. Watanabe, "Static and Dynamic Variance Compensation for Recognition of Reverberant Speech with Dereverberation Preprocessing," *IEEE Trans. on Aud. Speech & Lang. Process.*, Vol. 17, No. 2, pp. 324–334, 2009.

[9] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction. (ROVER)," *Proc. of ASRU*, pp. 347–354, 1997.

[10] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," *Proc. ICASSP*, pp. 81–84, 1995.

[11] M. Lincoln, I. McCowan, J. Vepa and H.K. Maganti, "The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments," proc. of *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[12] V. Mitra, H. Franco and M. Graciarena, "Damped Oscillator Cepstral Coefficients for Robust Speech Recognition," *Proc. of Interspeech*, pp. 886–890, 2013.

[13] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition," *Proc. of ICASSP*, pp. 4117–4120, 2012.

[14] R. Drullman, J. M. Festen and R. Plomp, "Effect of Reducing Slow Temporal Modulations on Speech Reception," *J. Acoust. Soc. of Am.*, Vol. 95, No. 5, pp. 2670–2680, 1994.

[15] O. Ghitza, "On the Upper Cutoff Frequency of Auditory Critical-Band Envelope Detectors in the Context of Speech Perception," *J. Acoust. Soc. of America*, vol. 110, no. 3, pp. 1628–1640, 2001.

[16] P. Maragos, J. Kaiser and T. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Processing*, Vol. 41, pp. 3024–3051, 1993.

[17] V. Mitra, M. McLaren, H. Franco, M. Graciarena and N. Scheffer, "Modulation Features for Noise Robust Speaker Identification," *Proc. of Interspeech*, pp. 3703–3707, 2013.

[18] H. Teager, "Some Observations on Oral Air Flow During Phonation," in *IEEE Trans. ASSP*, pp. 599–601, 1980.

[19] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan and M. Liberman, "Articulatory Features for Large Vocabulary Speech Recognition," *Proc. of ICASSP*, pp. 7145–7149, 2013.

[20] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition," *IEEE Trans. on ASLP*, Vol. 19, Iss. 7, pp. 1913–1924, 2010.

[21] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, Portable Task Dynamics Model in Matlab," *J. of Acoust. Soc. Am.*, 115(5), p. 2430, 2004.

[22] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson and E. Saltzman, "Articulatory Features from Deep Neural Networks and Their Role in Speech Recognition," *Proc. of ICASSP*, 2014.

[23] D. Vergyri, K. Kirchhoff, R. Gadde, A. Stolcke and J. Zheng, "Development of a Conversational Telephone Speech Recognizer for Levantine Arabic," *Proc. Interspeech*, 2005.

[24] V. Mitra, H. Franco, M. Graciarena, D. Vergyri, "Medium duration modulation cepstral feature for robust speech recognition," *Proc. of ICASSP*, Florence, 2014.

[25] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," *Proc. of ICASSP*, Florence, 2014.

[26] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proc. of HLT*, pp 357-362, 1991.