NORMALIZED AMPLITUDE MODULATION FEATURES FOR LARGE VOCABULARY NOISE-ROBUST SPEECH RECOGNITION

Vikramjit Mitra, Horacio Franco, Martin Graciarena, Arindam Mandal

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.
{vmitra, hef, martin, arindam}@speech.sri.com

ABSTRACT

Background noise and channel degradations seriously constrain the performance of state-of-the-art speech recognition systems. Studies comparing human speech recognition performance with automatic speech recognition systems indicate that the human auditory system is highly robust against background noise and channel variabilities compared to automated systems. A traditional way to add robustness to a speech recognition system is to construct a robust feature set for the speech recognition model. In this work, we present an amplitude modulation feature derived from Teager's nonlinear energy operator that is power normalized and cosine transformed to produce normalized modulation cepstral coefficient (NMCC) features. The proposed NMCC features are compared with respect to state-of-the-art noise-robust features in Aurora-2 and a renoised *Wall Street Journal* (WSJ) corpus. The WSJ word-recognition experiments were performed on both a clean and artificially renoised WSJ corpus using SRI's DECIPHER large vocabulary speech recognition system. The experiments were performed under three train-test conditions: (a) *matched,* (b) *mismatched,* and (c) *multi-conditioned*. The Aurora-2 digit recognition task was performed using the standard HTK recognizer distributed with Aurora-2. Our results indicate that the proposed NMCC features demonstrated noise robustness in almost all the training-test conditions of renoised WSJ data and also improved digit recognition accuracies for Aurora-2 compared to the MFCCs and state-of-the-art noise-robust features

Index Terms— *Noise-Robust Speech Recognition, Large Vocabulary Speech Recognition, Modulation Features.*

1. INTRODUCTION

Recent advances in LVCSR research have demonstrated high levels of recognition performance under clean or high signal-to-noise ratios (SNRs). Unfortunately, these LVCSR systems suffer from environmental degradations stemming from background noises and/or channel degradations.

Several approaches exist for realizing noise-robust automatic speech recognition (ASR) systems, such as 1) feature-space based, 2) model-space based, and 3) missing feature theory based approaches. While the model-space and marginalization-based missing-feature approaches tend to adapt the acoustic model to reduce the mismatch between training and testing utterances, the feature-space approaches achieve the same by generating relatively cleaner features for the acoustic model. Such approaches can be grouped into two subcategories. In the first subcategory, the noisy speech signal is enhanced by reducing noise corruption (e.g., spectral subtraction [1], computational auditory scene analysis [2], and so on). In the second subcategory, noise-robust features are employed in ASR systems (e.g., ETSI [European Telecomm. Standards Institute] advanced [3] front end, power normalized cepstral coefficients [PNCCs] [4], fepstrum features [5], perceptually motivated minimum variance distortion-less response [PMVDR] features [6], etc.

The most widely used mel-frequency log-energy based acoustic features (MFCCs) perform quite appreciably in clean matched conditions and have been used in several state-of-the-art ASR systems. Unfortunately, MFCCs are susceptible to frequency localized random perturbations, to which human perception is largely insensitive [7] and their performance degrades drastically in the presence of noise and channel degradations. This has led to a new quest for obtaining perceptually motivated noise robust acoustic features.

Studies [8, 9] have shown that amplitude modulation (AM) of the speech signal plays an important role in speech perception and recognition. Hence, recent studies [5, 10] have treated the speech signal as a sum of amplitude-modulated, narrow-band signals. The Discrete Energy Separation Algorithm (DESA) proposed in [11] uses the nonlinear Teager-Kaiser Energy Operator (TKEO) to demodulate the AM/FM components of a narrow-band signal. TKEO has been used in [12] to create mel-cepstral features that demonstrated robustness against car noise and improved ASR performance. As shown in [10], for matched training and testing in clean conditions, performance of the TKEO-based features was similar to that of ASR as standard mel-cepstral features. A recent study [7] showed that TKEO-based cepstral features offered a 60% relative word error rate (WER) improvement over MFCCs for mismatched conditions in Aurora-3 speech recognition task. The nonlinear DESA tracks the instantaneous AM energies quite reliably [11], which in turn provide better formant information [12] compared to conventional power spectrum-based approaches.

In this work, we present a perceptually motivated feature: Normalized Modulation Cepstral Coefficient (NMCC), that treats speech as a combination of AM/frequency modulation [FM] signals. We used DESA to transform speech into AM/FM components. The significance of DESA is twofold: (a) it doesn't impose a linear model to analyze speech and (2) it tracks the frequency and amplitude variations at the sample level without imposing any stationary assumption as done by linear prediction or Fourier transform. For DESA to give good AM/FM estimates the input signal has to be sufficiently bandlimited [10]; for which we used a perceptually inspired gammatone filter-bank (with configuration as in [13]). The AM components obtained from DESA were normalized (similar to [4]) as it helps to reduce the mismatch between training and testing spectral representation [4] under noisy conditions. We used root compression on the normalized AM power spectrum as conventional log compression is known to be more susceptible to noise corruption [19]. The final NMCC feature set is obtained by taking Discrete Cosine Transform (DCT) of the root compressed AM power spectrum.

The proposed features were compared with traditional MFCC features and some state-of-the-art noise-robust features in two different word recognition tasks: (a) noisy digit recognition with Aurora-2 dataset and (b) renoised WSJ word recognition. For task (a) an Aurora-2 *mismatched* train-test setup was used, where the whole word models were trained with clean speech and were tested with noise and channel degraded speech. For task (b) the WSJ corpus was corrupted synthetically with the noise and channel degradations similar to those of the Linguistic Data Consortium (LDC) DARPA RATS (Robust Automatic Transcription of Speech) rebroadcast examples using the renoiser tool of the International Computer Science Institute (ICSI) and the University of Columbia [14]. For this task the word recognition experiments were performed in (a) *matched* (training and testing under similar conditions), (b) *mismatched,* and (c) *multi-conditioned* training setups. The proposed normalized modulation cepstral coefficient (NMCC) features were found to outperform the MFCCs and most state-of-the-art features used in our experiments.

2. MODULATION FEATURES

Teager [15] introduced a nonlinear energy operator, Ψ, that tracks the instantaneous energy of a signal, where he assumed that a signal's energy is not only a function of its amplitude but also its frequency. Considering a discrete sinusoid $x[n]$, where A = const. amplitude, Ω = digital frequency, f = frequency of oscillation in hertz, f_s = sampling frequency in hertz, and θ = initial phase angle.

$$x[n] = A\cos[\Omega n + \theta];\ \Omega = 2\pi\,(f/f_s) \qquad (1)$$

If $\Omega \leq \pi/4$ and sufficiently small, then Ψ takes the form

$$\Psi\{x[n]\} = \{x^2[n] - x[n-1]x[n+1]\} \approx A^2\Omega^2 \qquad (2)$$

where, the maximum energy estimation error in Ψ will be 23% if $\Omega \leq \pi/4$, or $f/f_s \leq 1/8$ [11] used Ψ to formulate the discrete energy separation algorithm (DESA) and showed that it can instantaneously separate the AM/FM components of a narrow-band signal using

$$\Omega_i[n] \approx \cos^{-1}\left\{1 - \frac{\Psi(x[n]) + \Psi(x[n+1])}{4\Psi(x[n])}\right\} \qquad (3)$$

$$|a_i[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1 - [\cos(\Omega_i[n])]^2}} \qquad (4)$$

Note that in (2) $x^2[n] - x[n-1]x[n+1]$ can be less than zero if $x^2[n] < x[n-1]x[n+1]$, while the right hand side is strictly non-negative, $A^2\Omega^2 \geq 0$, so we have modified (2) to

$$\Psi\{x[n]\} = |\{x^2[n] - x[n-1]x[n+1]\}| \approx A^2\Omega^2 \qquad (5)$$

which now tracks the magnitude of energy changes. Also, the AM/FM signals computed from (3) and (4) may contain discontinuities [16] (that substantially increases their dynamic range), for which traditionally median filters have been used. To prevent such discontinuities we have modified the AM estimation equation (4), as we are only interested in the AM signal.

Let $s^w[n]$ be a windowed speech signal, after applying a time window $w[n]$. Let $a_{k,j}[n]$ be the AM time signal obtained from the DESA algorithm on the k^{th} gamma-tone filter-bank time signal at the j^{th} time window, $s_{k,j}[n]$ of the windowed speech sample $s_j^w[n]$. Let $A_{k,j} = max_n(s_{k,j}[n])$, then we can assume that

$$0 \leq a_{k,j}[n] \leq \theta A_{k,j} \ \text{ where } \theta \geq 1 \qquad (6)$$

θ determines the deviation allowed in the AM time signal $a_{k,j}[n]$ from the peak time signal amplitude, $A_{k,j}$. Then (4) can be written as

$$|a_{k,j}[n]| = \begin{cases} \sqrt{\dfrac{\Psi\{s_{k,j}[n]\}}{1-\left[1-\frac{\Psi(s_{k,j}[n])+\Psi(s_{k,j}[n+1])}{4\Psi(s_{k,j}[n])}\right]^2}} & if\ |a_{k,j}[n]| \leq \theta A_{k,j} \\ \frac{1}{N}\sum_{i=1}^{N}|s_{k,j}[n]| & o.w \end{cases} \qquad (7)$$

Equation (7) is imposing a hard upper bound on $|a_k^w[n]|$ by replacing the $|a_{k,j}[n]|$ outliers beyond the bound by the mean absolute magnitude of $s_{k,j}[n]$. Note that there still exist discontinuities at the time indexes of the outliers; to smooth out those discontinuities $|a_{k,j}[n]|$ is low-pass filtered with a cut-off frequency π/M and then downsampled by M. In our experiments we have used $\theta = 1.5$and $M = 4$. Note that having $\theta \gg 1$ would result in retaining the peaky discontinuities while having $\theta \leq 1$ will result in missing the proper peaks of the AM signal. Fig. 1 shows the overlaying plot of the windowed narrow-band time signal and its corresponding AM magnitude.

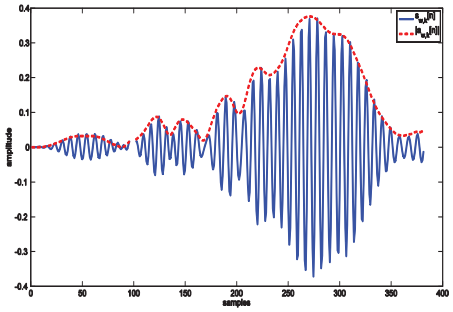

Fig. 1. A windowed narrow-band speech signal (in blue) and its corresponding AM signal (red) from the modified DESA algorithm.

The steps involved in obtaining the NMCC features are shown in Fig. 2. At the onset, speech signal is pre-emphasized (using a pre-emphasis filter of coefficient 0.97) and then analyzed using 25.6ms hamming window with 10ms frame rate. The windowed speech signal $s^w[n]$ is passed through a gamma-tone filter-bank having 40 channels between 200Hz and 7500Hz. The distribution of the center frequencies and the configuration of the filter-banks are used as specified in [13]. The AM time signals $a_{k,j}[n]$ are then obtained for each of the 40 channels using the modified DESA algorithm. The total AM power of the windowed time signal for k^{th} channel and j^{th} frame is given as

$$P_{k,j}^{AM} = a_{k,j}^T a_{k,j} \qquad (8)$$

Let $P_{0.95}$ denote the 95$^{\text{th}}$ percentile power across all j and k, the AM power for the k^{th} channel and j^{th} frame is normalized using

$$\hat{P}_{k,j}^{AM} = P_{k,j}^{AM}/P_{0.95} \qquad (9)$$

The normalized AM powers were then bias subtracted using a similar approach as specified in [4]. Fig. 3 shows the spectrogram of a noise corrupted signal (at SNR 15.6 dB), it's normalized AM power spectrum and its corresponding bias subtracted version.

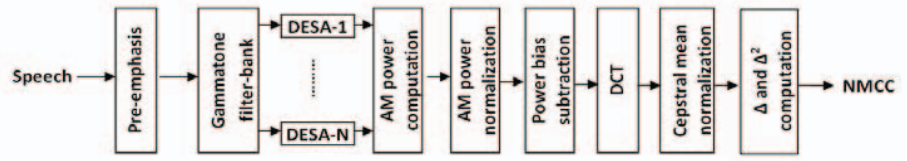

Fig. 2. Flow-diagram of NMCC feature extraction from speech.

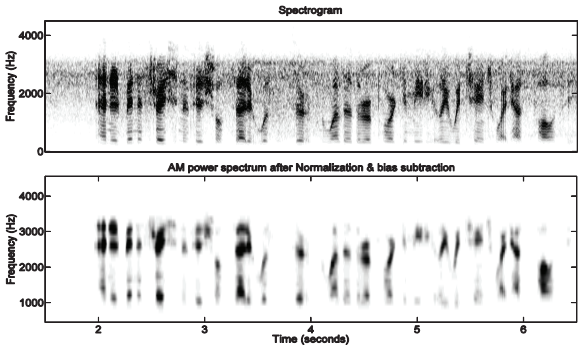

Fig. 3. Spectrogram of a renoised WSJ utterance corrupted with 15.6dB noise and Channel A characteristics (refer to Table 1) and it's normalized & bias subtracted amplitude modulation spectrum

$1/15^{th}$ root power compression was performed on the bias subtracted AM power spectrum and DCT was performed on the resultant. The first 13 coefficients were retained (including C_0) and cepstral mean normalization was performed at the utterance level. These 13 coefficients along with their Δs and Δ^2s resulted in a 39D NMCC feature set.

3. DATA USED FOR ASR EXPERIMENTS

The DARPA WSJ1 CSR dataset was used in the experiments presented in this paper. For training a set of 35990 speech utterances (77.8hrs) from the WSJ1 collection having 284 speakers was used. For testing the WSJ-eval94 dataset composed of 424 waveforms (0.8hrs) from 20 speakers was used. The data was artificially corrupted with noise and channel distortions using Dan Ellis's (Univ. of Columbia & ICSI) renoiser tool [14]. The renoiser tool estimates the noise/filter characteristics along with the SNR and frequency-shifts for the DARPA RATS Rebroadcast Example (RATS-RE) signals (LDC2011E20) and combines the same with the WSJ1 dataset. The renoiser filters each utterance with the filter characteristics that it has learnt from the RATS-RE and adds similarly estimated noise which is 'laundered' via LPC analysis-synthesis over a 1.0 sec window, at the estimated SNR (that it has learnt from the RATS-RE). Eight different versions of the WSJ1 train-test data were created with the renoiser tool corresponding to the eight channels specified in RATS-RE. The specifications for the RATS-RE data are shown in Table 1, where the SNRs and frequency shifts were estimated using the renoiser tool.

We also used the digit-corpus of Aurora-2 [17] to perform digit recognition experiments in clean, noisy and channel-degraded conditions. Aurora-2 is created from the TIdigits corpus and consists of connected digits spoken by American English speaker. The speech data in Aurora-2 are sampled at 8 kHz and have three test sections: A, B and C. Test-set A and B each have four subparts representing four different noise types; hence A and B altogether eight different noise types. Section C involves two noise types borrowed each from sections A and B and have different channel (MIRS) than the training set

Table 1. Channel specifications used in renoising *WSJ1*

Channel	Microphone	SNR (dB)	Frequency-shift (Hz)
A	Motorola HT1250	15.6	0
B	Midland GXT1050	6.0	0
C	Midland GXT1050	6.2	0
D	Galaxy DX2547	3.5	180.9
E	Icom IC-F70D	0.9	0
F	Trisquare TSX300	3.0	0
G	Vostek LX-3000	18.7	0
H	Magnum 1012 HT	3.0	120.7

4. DESCRIPTION OF THE ASR SYSTEMS USED

For the Aurora-2 experiments we used the HTK-based speech recognizer distributed with Aurora-2, which uses eleven whole word Hidden Markov Models (HMMs) with 16 states per word and three mixture components per state and two pause models for 'sil' and 'sp' with six mixture components per state. The ASR experiment was based on training on clean condition and testing on multi-SNR noisy data.

For the renoised WSJ data, we used SRI's DECIPHER LVCSR system, which uses a common acoustic front-end that computes 13 MFCCs (including energy) and their Δs and Δ^2s. The acoustic models were trained as cross-word, triphone HMMs with decision tree-based state clustering that resulted in 2048 fully tied states, and each state was modeled by a 32-component Gaussian mixture model, i.e. a total of 64K Gaussians for the entire acoustic model. The model uses three states (left-to-right) per phone. For the experiments presented in this work, all models were trained with maximum likelihood estimation. The system uses a bigram language model (LM) with 20K unigrams and 1 million 2-grams. There is a second pass decoding using model space MLLR speaker adaptation with an average of 7 regression classes and a third pass with a 3-gram LM to re-score the lattices from the second pass. The 3-gram LM had 20k unigrams, 12 million 2-grams and 20 million 3-grams. A detailed description of the ASR system is provided in [18].

5. EXPERIMENTS AND RESULTS

For Aurora-2 noisy digit recognition experiments we have used three different feature sets: (a) standard MFCCs (distributed with Aurora-2 database), (2) ETSI-2 [3] and (3) the proposed NMCCs. Please note, in all the experiments presented in this paper we have used the original feature generation source code shared with us by their authors/distributors. Table 2 shows the averaged word recognition accuracies for the three test sets of Aurora-2. In Table 2 we can see that the proposed NMCC features performed better than ETSI-2 almost always above 5dB. The bolded numbers in the table represent the best recognition accuracy for that test category.

For the renoised WSJ data, the proposed NMCC features were compared against MFCC (SRI's MFCC implementation), PMVDR [6], PNCC [4], ETSI-2 [3] and modulation fepstrum features [5]. The WERs were obtained under three training-testing conditions: (1) Matched: Acoustic models were trained and tested under the same condition (i.e., if trained with clean data, then tested with clean data only); (2) Mismatched: Acoustic models were trained with clean data only and then evaluated under different conditions; (3) Multi-conditioned: Acoustic models were trained with the same training data size as from clean data and the eight channels (randomly selected, but channels were selected uniformly for each speaker) and then evaluated under the different testing conditions.

The acoustic features were mean and variance normalized (at the speaker level) prior to training and testing the models. WERs were obtained under three conditions: (a) using a bigram LM, (b) maxim -um likelihood linear regression (MLLR) scoring, and (c) lattice-

Table 2. Word recognition accuracies for Aurora-2 mismatched condition

	Set-a			Set-b			Set-c		
	MFCC	ETSI2	NMCC	MFCC	ETSI2	NMCC	MFCC	ETSI2	NMCC
clean	99.00	99.09	**99.12**	99.00	99.09	**99.12**	99.00	**99.06**	99.03
20 dB	90.38	98.15	**98.28**	86.67	98.09	**98.66**	94.83	97.53	**97.96**
15 dB	75.29	96.93	**97.07**	69.44	96.48	**97.76**	88.66	95.55	**96.50**
10 dB	53.11	93.16	**93.39**	47.96	92.53	**94.56**	75.23	90.47	**92.00**
5 dB	32.15	**84.36**	80.89	28.35	82.05	**83.20**	50.85	78.60	**79.07**
0 dB	14.89	**60.88**	46.50	12.21	**58.68**	51.30	23.83	**51.91**	46.87
0-20dB	*53.16*	***86.70***	*83.23*	*48.92*	***85.56***	*85.09*	*66.68*	***82.81***	*82.50*

rescoring. Since lattice rescoring always gave the best result for all the features tested in our work, we report only the lattice rescoring result in this paper. Tables 3 through 5 show the WERs obtained from all the features used in our experiments in matched, mismatched, and multi-conditioned cases.

Table 2 shows that the proposed NMCC features performed better than the ETSI-2 features for SNRs≥5dB. On average, NMCCs improved word recognition accuracy by 0.4% and 22.3% relative to ESTI-2 and MFCCs. Table 5 shows that under the multi-conditioned case, NMCCs offered the best mean WER and gave the best WER for most channels. For matched and mismatched conditions, NMCCs were second best, offering best WERs for at least three out of eight channels. Note that for the mismatched condition, NMCCs gave the least WER for half of the renoised data.

Table 3. WER for matched training-testing conditions

	MFCC	PMVDR	PNCC	ETSI2	Fepstrum	NMCC
Clean	6.7	**6.4**	6.9	7.8	6.7	6.8
A	15.5	17.2	14.9	16.9	15.5	**14.6**
B	30.3	31.4	30.8	30.5	29.9	**29.5**
C	30.5	31.0	30.8	30.0	**28.7**	28.8
D	38.4	33.1	36.0	**29.8**	30.3	35.3
E	60.3	53.3	56.2	**52.2**	53.7	52.5
F	48.5	46.3	44.8	49.4	**38.9**	43.0
G	9.7	9.8	9.6	10.1	9.3	**9.1**
H	35.4	34.0	32.9	29.5	**29.2**	32.3
mean	***30.6***	***29.2***	***29.2***	***28.5***	***26.9***	***28.0***

Table 4. WER for mismatched training-testing conditions

	MFCC	PMVDR	PNCC	ETSI2	Fepstrum	NMCC
Clean	6.7	**6.4**	6.9	7.8	6.7	6.8
A	23.0	28.4	**18.0**	26.1	32.5	18.6
B	51.7	59.2	**46.4**	52.7	59.0	47.7
C	51.7	58.7	45.2	53.7	57.5	**45.1**
D	70.0	76.4	72.5	72.5	**69.9**	75.7
E	86.6	86.6	83.4	84.4	86.9	**79.2**
F	75.3	76.7	67.9	80.2	71.9	**66.9**
G	12.7	13.0	11.2	12.9	13.3	**10.9**
H	70.4	74.1	**60.8**	65.5	66.1	64.8
mean	***49.8***	***53.3***	***45.8***	***50.6***	***51.5***	***46.2***

Table 5. WER for multi-conditioned training-testing conditions

	MFCC	PMVDR	PNCC	ETSI2	Fepstrum	NMCC
Clean	16.7	11.4	11.1	12.4	13.1	**10.4**
A	26.5	24.1	20.2	25.2	26.1	**20.0**
B	44.8	35.5	34.3	35.3	41.9	**33.8**
C	43.4	36.2	33.6	36.0	42.4	**32.8**
D	65.0	**42.2**	46.7	43.1	50.7	47.3
E	87.8	68.3	66.2	**63.8**	75.6	**63.8**
F	70.0	52.4	48.4	60.8	52.9	**46.3**
G	16.5	14.2	**13.8**	15.6	15.0	14.1
H	58.1	40.1	39.5	40.0	44.5	**38.7**
mean	***47.6***	***36.0***	***34.9***	***36.9***	***40.2***	***34.1***

6. CONCLUSION

We have proposed an amplitude modulation based noise-robust feature for ASR. The proposed feature was found to outperform ETSI-2 features in Aurora-2 experiments and also provided best overall WER for multi-conditioned test cases of renoised WSJ data. For matched and mismatched conditions, it performed competitively and provided the least WER for almost half of the channels. The proposed features overall demonstrated sufficient noise robustness compared to state-of-the-art noise-robust features used in our experiments. Channels D and H overall showed higher WERs, which may be due to their frequency shift characteristics. Currently, NMCC does not have a way to deal with these frequency shifts, and future study should account for this in order to improve its performance in those cases.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The authors sincerely thank Dan Ellis (Columbia University and the International Computer Science Institute [ICSI]) for sharing his renoiser code and Vivek Tyagi (IBM-IRL) for sharing his fepstrum feature generation code.

8. REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. Speech Audio Process.* vol.7, no.2, pp. 126–137, 1999.

[2] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition", *IEEE Trans Audio, Speech, Lang. Process.,* vol. 15, no. 7, pp. 2130–2140, 2007.

[3] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI ES 202 050 Ver. 1.1.5, 2007.

[4] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", in *Proc. ICASSP,* pp. 4574–4577, 2010.

[5] V. Tyagi, *Fepstrum features: Design and application to conversational speech recognition,* IBM Research Report, 11009, 2011.

[6] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition", *Speech Comm.,* vol.50, iss. 2, pp. 142–152, 2008.

[7] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition", in *Proc of Interspeech*, pp. 3013–3016, 2005.

[8] 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.

[9] 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.

[10] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition", *IEEE Trans. Speech & Audio Proc.,* vol. 9, no. 3, pp. 196–200, 2001.

[11] 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.

[12] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise", *IEEE Sig. Proc. Letters,* vol. 6, no. 10, pp. 259–261, 1999.

[13] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research,* vol. 47, pp.103–138, 1990.

[14] http://labrosa.ee.columbia.edu/projects/renoiser/create_wsj.html

[15] H. Teager, "Some observations on oral air flow during phonation", *IEEE Trans. ASSP,* pp. 599–601, 1980.

[16] J.H.L. Hansen, L. Gavidia-Ceballos, and J.F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment", *IEEE Trans. Biomedical Engineering,* vol. 45, no. 3, pp. 300–313, 1998.

[17] D. Pearce and H.G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *Proc. ICSLP,* Beijing, China, 2000.

[18] M. Akbacak, H. Franco, M. Frandsen, S. Hasan, H. Jameel, A. Kathol, S. Khadivi, X. Lei, A. Mandal, S. Mansour, K. Precoda, C. Richey, D. Vergyri, W. Wang, M. Yang, and J. Zheng, "Recent advances in SRI's IraqComm(tm) Iraqi Arabic-English speech-to-speech translation system", in *Proc. IEEE ICASSP* (Taipei), pp. 4809–4813, April 2009.

[19] S. Ravindran, D. V. Anderson and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," in SAPA, Pittsburgh, PA, September 2006.