MEDIUM-DURATION MODULATION CEPSTRAL FEATURE FOR ROBUST SPEECH RECOGNITION

Vikramjit Mitra, Horacio Franco, Martin Graciarena, Dimitra Vergyri

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.

ABSTRACT

Studies have shown that the performance of state-of-the-art automatic speech recognition (ASR) systems significantly deteriorate with increased noise levels and channel degradations, when compared to human speech recognition capability. Traditionally, noise-robust acoustic features are deployed to improve speech recognition performance under varying background conditions to compensate for the performance degradations. In this paper, we present the Modulation of Medium Duration Speech Amplitude (MMeDuSA) feature, which is a composite feature capturing subband speech modulations and a summary modulation. We analyze MMeDuSA's speech recognition performance using SRI International's DECIPHER® large vocabulary continuous speech recognition (LVCSR) system, on noise and channel degraded Levantine Arabic speech distributed through the Defense Advance Research Projects Agency (DARPA) Robust Automatic Speech Transcription (RATS) program. We also analyzed MMeDuSA's performance against the Aurora-4 noise-and-channel degraded English corpus. Our results from all these experiments suggest that the proposed MMeDuSA feature improved recognition performance under both noisy and channel degraded conditions in almost all the recognition tasks.

Index Terms— *noise-robust speech recognition, large vocabulary continuous speech recognition, modulation features.*

1. INTRODUCTION

Current Large-Vocabulary Continuous Speech Recognition (LVCSR) systems demonstrate high levels of recognition accuracy under clean condition or at high signal-to-noise ratios (SNRs). However, these systems are very sensitive to environmental degradations such as background noise, channel mismatch, and/or distortions. Hence, robust speech analysis has become an important research area, not only for enhancing the noise/channel robustness of automatic speech recognition (ASR) systems, but also for other speech applications, such as voice-activity detection, speaker identification, etc.

Traditionally, ASR systems use Mel-frequency cepstral coefficients (MFCCs) as the acoustic observation. These perform quite well in clean matched conditions and have been used in several state-of-the-art ASR systems. Unfortunately, MFCCs are susceptible to noise [1], and their performance degrades dramatically with increases in noise levels and channel degradations. To account for MFCC's vulnerability to noise and channel degradations, researchers have actively sought to obtain a robust acoustic feature set. Research on noise robust acoustic features typically aims to generate noise compensated features for the acoustic-model training and such features can be generated in two ways: (1) using speech-enhancement-based approaches, where the noisy speech signal is enhanced by reducing noise corruption (e.g., spectral subtraction [2], computational auditory scene analysis [3], etc.) followed by cepstral feature extraction; or (2) by using noise robust speech-processing approaches, where noise-robust transforms and/or human perception based speech analysis methodologies are deployed for acoustic-feature generation (e.g., ETSI [European Telecomm. Standards Institute] advanced frontend [4], power normalized cepstral coefficients [PNCC] [5], modulation based features [6, 7], and several others).

Studies [8, 9] have shown that amplitude modulation of the speech signal plays an important role in speech perception and recognition; hence, recent studies [6, 7, 10, 11] have modeled the speech signal as a combination of amplitude-modulated narrow-band signals. Literature [7, 10, 12, 23] have demonstrated that modulation based features are robust to noise. In this paper we present the Modulation of Medium Duration Speech Amplitude (MMeDuSA) feature, which is similar to previously proposed modulation based features such as the normalized modulation cepstral coefficients (NMCC) [7] or the mean Hilbert energy coefficients (MHEC) [23] features but has an extra layer of summary modulation information and uses a novel approach to estimate the amplitude modulations (AM) of bandlimited subband speech signals. The summary modulation information provides a noise robust estimate about the overall speech modulation and is geared to capture voicing information along with information about vowel stress and prominence. MMeDuSA is a noise and channel-robust acoustic feature that uses a medium-duration analysis window to obtain instantaneous estimates of subband speech AM signals.

One of the conventionally used techniques to estimate AM signals from a subband speech signal is the Discrete Energy Separation Algorithm (DESA) [11], which uses the nonlinear Teager's Energy Operator (TEO) to demodulate the AM/FM components of a narrow-band signal. Prior studies [12] have used TEO [10, 11, 12] to create mel-cepstral features that demonstrated robustness by improving ASR performance in noisy conditions. Note that AM signals computed from DESA may contain discontinuities [10, 13] that introduce artifacts in TEO based acoustic features. In this paper we directly use the subband TEO to have a crude estimate of the AM signals, where the resulting estimates are free from any unusual discontinuities. In MMeDuSA processing the AM estimates are used to compute the AM power over a medium duration window of length 51.2 ms; then it performs bias subtraction (using a similar approach outlined in [5]) followed by nonlinear root compression to generate an AM power spectrum. Discrete Cosine Transform (DCT) is performed on the root compressed AM power spectrum to yield a cepstra-like feature.

To analyze the performance of the proposed MMeDuSA feature we compare its speech recognition accuracy with respect to traditional MFCC features and some state-of-the-art noise-robust features, where we explored two different noisy speech recognition

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tasks: (1) a noisy English speech recognition task with the Aurora4 dataset and (2) a noisy and channel distorted Levantine Arabic dataset. The latter is available from the Linguistic Data Consortium (LDC) [20] through the DARPA Robust Automatic Transcription of Speech (RATS) program. Note that RATS data is unique in the sense that the noise and channel degradations were not artificially introduced by performing simple mathematical operations on the speech signal, but by transmitting clean source signals through different radio channels [20], where variations amongst different channels introduced an array of distortion modes. The data also contained distortions, such as frequency shifting, speech modulated noise, non-linear artifacts, no transmission bursts etc., which made robust signal processing approaches even more challenging compared to traditional noisy corpora available in the literature.

2. THE MMEDUSA FEATURE PIPELINE

In [15] Teager proposed an energy operator, popularly known as the Teager's energy operator or TEO. The TEO operates on a bandlimited signal and is a function of the signals amplitude and its frequency. In [16] Kaiser analyzed the nonlinear TEO, Ψ, and presented some of its salient properties.

Considering a discrete sinusoid $x[n]$, where A = constant amplitude, Ω = digital frequency, f = frequency of oscillation in Hertz, f_s = sampling frequency in Hertz, and θ = initial phase angle-

$$x[n] = Acos[\Omega n + \theta];\ \Omega = 2\pi\,(f/f_s). \quad (1)$$

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

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

where the maximum energy estimation error in Ψ will be 23% if $\Omega \leq \pi/4$, or $f/f_s \leq 1/8$. [17] 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

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

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

Note that in (2) $x^2[n] - x[n-1]x[n+1]$ can be less than zero when $x^2[n] < x[n-1]x[n+1]$, while $A^2\Omega^2$ is strictly non-negative. Thus, we have modified (2) to

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

which now tracks the magnitude of energy changes. Also, the AM/FM signals computed from (3) and (4) may contain discontinuities [18] which can substantially increase their dynamic range. In order to remove such artifacts from the DESA algorithm, we propose an estimate of the instantaneous AM signal by assuming that the instantaneous FM signal will be approximately equal to the center frequency of the gammatone filterbank when the subband signals are sufficiently bandlimited

$$\Omega_i \approx f_c \quad (6)$$

Given (6), the estimation of the instantaneous AM signal from (5) becomes very simple

$$A_i \approx \sqrt{\frac{|x^2[n]-x[n-1]x[n+1]|}{{\Omega_i}^2}} \quad (7)$$

The steps involved in obtaining the MMeDuSA feature are shown in Fig. 1. In the MMeDuSA pipeline the speech signal is first pre-emphasized and then analyzed using a Hamming window of 51.2-ms with a 10-ms frame rate. The windowed speech signal $s[n]$ is passed through a gammatone filter-bank having 30 critical bands, with center frequencies spaced equally in the equivalent rectangular bandwidth (ERB) scale between 250 Hz and 3800 Hz. Note that for all experiments presented in this paper, we assume that the input speech signal has useful information up to 4000 Hz. The filters' bandwidths are characterized by the ERB scale, where the ERB for channel c (where $c = 1 \ldots 34$) given by-

$$ERB_c = \frac{f_c}{Q_{ear}} + BW_{min} \quad (8)$$

where f_c represents the center frequency for filter c and Q_{ear} and BW_{min} are constants set to 9.26449 and 24.7 according to Glasberg & Moore specifications [21]. The time signal from the c^{th} gammatone filter with impulse response $h_c(n)$ is given as

$$s_c(n) = s(n) * h_c(n). \quad (9)$$

For each of these 30 subband signals, their AM signals are computed using (7).

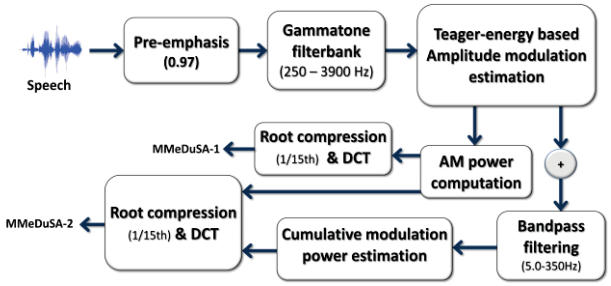

Fig. 1. MMeDuSA1 and MMeDuSA2 feature extraction pipeline

The power of the estimated AM signals was computed (refer to Fig. 1) and non-linear compression (we have used $1/15^{th}$ root compression as it was found to be more noise robust compared to logarithmic compression and other root compression coefficients) was performed on it. The power of the AM signal $a_{k,j}[n]$ for k^{th} channel and j^{th} frame is given as

$$P_{k,j}^{AM} = a_{k,j}^T a_{k,j}. \quad (10)$$

For a given analysis window, 30 power coefficients were obtained for each of the 30 channels, which were then transformed using DCT, and their first 13 coefficients were retained. In our experiments we have used these 13 coefficients by themselves along with their derivatives; we identify this feature as MMeDuSA1. Note that the feature operates in 'medium duration' as it uses an analysis window of size 52 ms compared to the traditionally used 10-ms~25ms windows.

In parallel, each of the 30 estimated AM signals (as shown in Fig. 1) were band-pass filtered using DCT, retaining information only within 5 Hz to 350 Hz. These are the medium duration modulations (represented as: $a\_mod_{k,j}[n]$), which were summed across the frequency scale to obtain a medium duration modulation summary

$$\overline{a\_mod_j} = \sum_{k=1}^{N} a\_mod_{k,j}[n]. \quad (11)$$

The power signal of the medium duration modulation summary was obtained, followed by $1/15^{th}$ root compression. The resultant was transformed using DCT and the first n coefficients were retained. These n coefficients were combined with the cepstral coefficients and their derivatives obtained from the other branch (MMeDuSA1) (refer to Fig. 1) of the feature processing and the resulting feature set is named as the MMeDuSA2. Both MMeDuSA1 and MMeDuSA2 features were used in our ASR experiments presented below. Please note that we have used the value of n as 4 for Aaurora-4 experiments and 3 for Levantine

Arabic experiments as those were the optimal values found based on the results from the development set.

3. DATA USED FOR ASR EXPERIMENTS

For the English LVCSR experiments, the Aurora4 database was used, which contains six additive noise versions with channel matched and mismatched conditions. It is created from the standard 5K Wall street Journal (WSJ0) database and has 7180 training utterances of approximately 15 hours duration, and 330 test utterances each with an average duration of 7 seconds. The acoustic data (both training and test sets) come with two different sampling rates (8 kHz and 16 kHz) and the experiments reported here uses only the 8 kHz data. Two different training conditions were specified: (1) clean training, which is the full SI-84 WSJ train-set without any added noise; and (2) multi-condition training, with about half of the training data recorded using one microphone, the other half recorded using a different microphone (hence incorporating two different channel conditions), with different types of added noise at different SNRs. The noise types are similar to the noisy conditions in test. The Aurora4 test data include 14 test-sets from two different channel conditions and six different added noises (in addition to the clean condition). The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were (1) car; (2) babble; (3) restaurant; (4) street; (5) airport and (6) train (set07) along with clean condition. The evaluation set comprised 5K words in two different channel conditions. The original audio data for test conditions 1-7 was recorded with a Sennheiser microphone while test conditions 8-14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones (more details in [19]). The different noise types were digitally added to the clean audio data to simulate noisy conditions.

For the Levantine Arabic LVCSR experiments, the training and test data were taken from DARPA RATS Rebroadcast Example Levantine Arabic signals (LDC2011E111 and LDC2011E93) for the Arabic KWS distributed by LDC [20]. The data was collected by retransmitting Levantine Arabic telephone conversation speech through eight different communication channels marked A though H. These channels have a range of distortions associated with them that were similar to that observed in air traffic controller radio communication channels and have characteristics such as sideband mistuning, tonal interference, intermittent no-transmission bursts, multi path interference etc. [20]. The total data for training the systems from all eight channels, plus the original clean speech, was 173 hours. The test set (dev-1) was created in a similar manner by retransmitting the data through the eight channels and contained between 2.0~2.5 hours of test data. The DARPA RATS dataset is unique in the sense that noise and channel degradations were not artificially introduced by performing mathematical operations on the clean speech signal, but the signals were in fact rebroadcasted through a channel and noise degraded ambience and then rerecorded. The data contained several artifacts such as nonlinearity, frequency shifts, modulated noise, intermittent bursts, extremely low SNRs etc., and traditional noise robust approaches developed in the context of additive noise may not work so well.

4. ASR SYSTEM DESCRIPTION

For the Aurora4 LVCSR experiments, we used SRI International's DECIPHER® LVCSR system, which uses a common acoustic front-end that computes 13 MFCCs (including energy) and their Δs, Δ^2s and Δ^3s. The 52 dimensional MFCC features were transformed to 39 dimensions using heteroscedastic linear discriminant analysis (HLDA) transform. From our experiments we observed that using HLDA on 52 dimensional MFCC features (with up to Δ^3 information) gives lower error rates than using 39 dimensional MFCC features (with up to Δ^2 information). Speaker-level mean and variance normalization is performed on the acoustic features prior to acoustic model training. 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 Aurora-4 system used in our experiments uses the 5K non-verbalized closed vocabulary set language model (LM), where a bigram LM is used on the initial pass of decoding and second-pass decoding with model space maximum likelihood linear regression (MLLR) speaker adaptation followed by trigram LM rescoring of the lattices is used. A detailed description of the ASR system is provided in [22].

DECIPHER® was also used for the Levantine Arabic acoustic model training. We used a version of the system that achieves competitive performance on conversational telephone speech (CTS) tasks in multiple languages. The ASR system's acoustic model is trained as a speaker-independent model by pooling sentences from all speakers and clustering them into speaker clusters in an unsupervised manner using standard maximum likelihood loss measures for splitting Gaussian distributions. Then using the acoustic features, mean and variance normalization factors are estimated for each speaker cluster. Next, the system collects sufficient statistics for three-state hidden Markov models (HMM) of triphones and then clusters these states into distinct groups using a decision tree that applies a clustering criterion based on the linguistic properties of phonetic contexts. Finally, using the Baum-Welch algorithm, triphones are estimated as three-state HMMs with Gaussian mixture models as the output probability distributions of the HMMs. During decoding we used a bigram language model and phone-loop MLLR to produce an initial set of ASR hypotheses that served as references to perform model-space MLLR speaker adaptation of the speaker-independent acoustic model on the unsupervised speaker cluster hypothesized earlier. A full matrix transformation with an offset vector for the Gaussian means and a diagonal variance transformation vector are estimated using a regression-class tree, where each node with a minimum adaptation count of 2200 frames resulted in a set of MLLR transformations that were shared by all triphone HMM states in that cluster. An average of eight regression classes of MLLR transformation was computed for each unsupervised speaker cluster. Subsequently, the MLLR-adapted acoustic models were used for a second pass of ASR decoding to produce bigram lattices. Lastly, the lattices were expanded using a tri-gram language model to produce more accurate hypotheses.

5. EXPERIMENTS AND RESULTS

We used Aurora4 LVCSR experiments to analyze different components of the MMeDuSA pipeline, where the 8 kHz clean training set was used to train the acoustic model and part of the 8 kHz noisy training data (from the Aurora4 multi-condition setup) was used as the development set. We explored replacing the TEO based AM estimation with a Hilbert envelope, but that resulted in ~1.5% increase in word error rate (WER). Finally, we also observed a significant degradation in performance when power compression was replaced with standard log-compression (around

8% relative degradation in WER). From our experiments using a wide range of power coefficients we observed that a power compression of $1/15^{th}$ root was the optimal choice. Based on the above set of observations we finalized the configuration of the MMeDuSA features that were used in our final ASR experiments.

In Aurora4 experiments we used only mismatched conditions (i.e., train with clean data [clean training] and test with noisy and different channel data) at 8 kHz sampling rate. Six different feature sets were used: (1) MFCCs, (2) PNCC [5], (3) NMCC [7], (4) ETSI-AFE [4], and the two different versions of the proposed feature: MMeDuSA1 and MMeDuSA2. In all the experiments presented below, we used the original feature generation source code shared with us by their authors or available from their websites (for ETSI-AFE). We also explored using a Frequency-modulation component (refer to equation 3) from the DESA algorithm as a possible competitor for the AM-based MMeDuSA feature, but their results were much worse compared to any of the features used in our experiments. Tables 1-2 show the WERs from the Aurora4 experiments.

Table 1. WER on Aurora-4 for clean training, matched channel condition

		MFCC	PNCC	NMCC	ETSI	MMeDuSA1	MMeDuSA2
1	Clean	11.7	13.2	13.4	13.4	12.6	**11.3**
2	Car	16.6	17.7	17.3	18.5	17.5	**15.6**
3	Babble	37.9	32.5	32.6	**31.8**	33.1	32.1
4	Restaurant	41.5	33.7	35.3	35.5	34.9	**33.5**
5	Street	45.1	34.9	34.7	33.9	36.4	**33.2**
6	Airport	33.2	31.4	30.1	29.9	30.6	**28.9**
7	Train	45.7	**32.6**	34.8	33.0	34.6	33.9
	Avg. (2-7)	36.7	30.5	30.8	30.4	31.2	**29.5**

Table 2. WER on Aurora-4 for clean training, mismatched channel condition

		MFCC	PNCC	NMCC	ETSI	MMeDuSA1	MMeDuSA2
1	Clean	15.0	16.6	17.4	17.9	16.8	**14.7**
2	Car	20.6	23.2	21.7	23.4	21.3	**19.3**
3	Babble	44.7	38.0	37.0	36.5	37.7	**35.4**
4	Restaurant	48.3	42.5	41.4	41.2	40.7	**38.9**
5	Street	52.9	41.0	40.3	41.2	42.1	**39.9**
6	Airport	39.8	38.0	35.7	34.9	35.1	**34.1**
7	Train	51.4	39.1	39.2	39.0	40.5	**38.4**
	Avg. (2-7)	42.9	36.9	35.9	36.0	36.2	**34.3**

From table 2 we see that the proposed MMeDuSA2 feature performed better for the channel-mismatched conditions than any other feature, closely followed by MMeDuSA1, NMCC and ETSI. It is quite interesting to note that the difference between the MMeDuSA2 and MMeDuSA1 is only a few (4 in the case of Aurora4 experiments) coefficients that capture the summary modulation and that played an important role in reducing the relative overall WER by 5.4% for matched channel conditions and 5.2% for mismatched channel conditions. The summary modulation information in MMeDuSA2 is geared to capture voicing information while capturing information such as vowel stress and prominence. It also provides a more noise robust estimate about the overall speech modulation. The above attributes of the summary modulation part of MMeDuSA2 may be the main factor behind MMeDuSA2's superior performance compared to MMeDuSA1. We observed a similar trend in our recent exploration of MMeDuSA features on speaker recognition evaluation [24].

For matched channel conditions (table 1) at 8 kHz MMeDuSA2 provided the best results in five out of seven conditions and also provided the lowest overall WER in noisy condition, closely followed by ETSI, PNCC and NMCC. In summary MMeDuSA2 helped to lower the relative overall WER by 19.6% compared to the baseline MFCC features and 3% compared to the 2^{nd} best performing ETSI features in channel matched condition. For channel mismatched condition MMeDuSA2 helped to lower the relative overall WER by 20% compared to the baseline MFCC features and 4.5% compared to the 2^{nd} best performing NMCC features. These results suggest the MMeDuSA2 is both channel and noise robust compared to some of the state-of-the-art features used in this work.

Table 3 presents the WERs from the Levantine Arabic ASR on the RATS data. Note that due to the difficulty in properly transcribing dialectal Arabic, the WERs on clean conversational telephone speech are quite higher (around 40%~50% [22]) than in English, hence adding noise, channel degradations and other distortions easily worsens the WERs. WERs in table 3 confirm that the MMeDuSA feature performs well for channel and noise degraded Levantine Arabic speech as well, where MMeDuSA2 provided the lowest WER for four out of eight channels.

Table 3. WER from RATS Arabic speech recognition task

	MFCC	RASTA-PLP	PNCC	NMCC	MMeDuSA1	MMeDuSA2
Clean	58.5	**56.3**	68.4	58.5	62.2	60.1
Chan. A	84.3	82.8	85.2	82.6	83.0	**82.0**
Chan. B	84.2	83.7	88.0	83.5	84.9	**82.7**
Chan. C	84.4	**82.9**	88.1	83.6	83.9	83.2
Chan. D	72.9	**71.1**	81.2	73.0	74.4	72.1
Chan. E	86.8	85.9	90.5	86.4	86.8	**85.8**
Chan. F	75.2	74.0	81.4	74.2	75.0	**73.9**
Chan. G	65.1	**63.4**	74.2	65.7	68.2	66.3
Chan. H	81.7	**79.9**	87.2	80.6	81.9	81.0
Avg.(A-H)	79.3	**78.0**	84.5	78.7	79.8	78.4

6. CONCLUSION

We presented an amplitude-modulation-based noise-robust feature for ASR and demonstrated that it offered noise robustness for both English and Levantine Arabic LVCSR systems. For English, we performed mismatched acoustic-model training; whereas for Levantine Arabic, we used a multi-condition model. In both cases the proposed MMeDuSA feature demonstrated overall lower WERs under noisy conditions compared to the other features.

The experiments presented in this paper dealt with ASR tasks for speech degraded with real-world noise and channel artifacts using recognition tasks from two different languages. Given the difficulty of the task the proposed feature provided consistent improvement with respect to the baseline features and other state-of-the-art robust features. In the future we intend to explore the summary modulation part in details and try multi-resolution approaches in obtaining them. We also intend to explore these features under noisy and reverberant conditions.

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.
Approved for Public Release, Distribution Unlimited

8. REFERENCES

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

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. Speech Audio Process.*, 7(2), pp. 126–137, 1999.

[3] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition", IEE*E Trans Audio, Speech, Lang. Process.*, 15(7), pp. 2130–2140, 2007.

[4] *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.

[5] 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.

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

[7] V. Mitra, H. Franco, M. Graciarena and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition", *in Proc. of ICASSP*, pp. 4117-4120, Japan, 2012.

[8] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception", *J. Acoust. Soc. of Am.*, 95(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 Am.*, 110(3), pp. 1628–1640, 2001.

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

[11] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition", *IEEE Trans. Speech & Audio Proc.*, 9(3), pp. 196–200, 2001.

[12] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise", *IEEE Sig. Proc. Letters*, 6(10), pp. 259–261, 1999.

[13] 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*, 45(3), pp. 300–313, 1998.

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

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

[16] J.F. Kaiser, "Some useful properties of the Teager's energy operator", *in Proc of IEEE*, Iss. III, pp. 149-152, 1993.

[17] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Signal Processing*, 41, pp. 3024–3051, 1993.

[18] 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*, 45(3), pp. 300–313, 1998.

[19] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task", *ETSI STQ-Aurora DSR Working Group*, June 4, 2001.

[20] K.Walker and S. Strassel, "The RATS radio traffic collection system," *in Proc. of ISCA*, Odyssey, 2012.

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

[22] D. Vergyri, K. Kirchhoff, R. Gadde, A. Stolcke, and J. Zheng, "Development of a conversational telephone speech recognizer for Levantine Arabic," *in Proc. Interspeech*, 2005.

[23] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin and J. H. L. Hansen, "Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA", *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, Atlanta, GA, USA, 2011.

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