Auditory motivated front-end for noisy speech using spectro-temporal modulation filtering

Sriram Ganapathy[a] **and Mohamed Omar**
IBM T.J. Watson Research Center, Yorktown Heights, New York 10562
ganapath@us.ibm.com, mkomar@us.ibm.com

Abstract: The robustness of the human auditory system to noise is partly due to the peak preserving capability of the periphery and the cortical filtering of spectro-temporal modulations. In this letter, a robust speech feature extraction scheme is developed that emulates this processing by deriving a spectrographic representation that emphasizes the high energy regions. This is followed by a modulation filtering step to preserve only the important spectro-temporal modulations. The features derived from this representation provide significant improvements for speech recognition in noise and language identification in radio channel speech. Further, the experimental analysis shows congruence with human psychophysical studies.

PACS numbers: 43.72.Ne, 43.72.Ar [DOS]
Date Received: July 3, 2014 **Date Accepted:** September 12, 2014

1. Introduction

Even with several advancements in the practical application of speech technology, the performance of the state-of-the-art systems remain fragile in high levels of noise and other environmental distortions. On the other hand, various studies on the human auditory system have shown good resilience of the system to high levels of noise and degradations (Greenberg *et al.*, 2004). This information shielding property of the auditory system may be largely attributed to the signal peak preserving functions performed by the cochlea and the spectro-temporal modulation filtering performed in the cortical stages.

In the auditory periphery, there are mechanisms that serve to enhance the spectro-temporal peaks, both in quiet and in noise. The work done in Palmer and Shamma (2004) suggests that such mechanisms rely on automatic gain control (AGC), as well as the mechanical and the neural suppression of those portions of the signal which are distinct from the peaks The second aspect in our analysis relates to the importance of spectro-temporal modulation processing. The importance of spectral modulations (Keurs *et al.*, 1992) and temporal modulations (Drullman *et al.*, 1994) for speech perception is well studied. Furthermore, the psychophysical experiments with spectro-temporal modulations illustrate that modulation filtering is an effective tool in enhancing the speech signal for human speech recognition in the presence of high levels of noise (Elliott and Theunissen, 2009).

Given these two properties of human hearing, we investigate the emulation of these techniques for feature extraction in automatic speech systems. The auditory filter based decomposition like mel/bark filter banks (for example, Davis and Mermelstein, 1980) have been widely used for at least three decades in many speech applications with normalization techniques like mean-variance normalization (Chen and Bilmes, 2007) or short-term Gaussianization (Pelecanos and Sridharan, 2001). Additionally, the modulation filtering approaches have also been proposed for speech feature extraction with RASTA filtering (Hermansky and Morgan, 1994) and multi-stream combinations (Chi *et al.*, 2005; Nemala *et al.*, 2013).

[a]Author to whom correspondence should be addressed.

In this paper, we propose a feature extraction scheme which is based on the understanding of the important properties of the auditory system. The initial step is the derivation of a spectrographic representation which emphasizes the high energy peaks in the spectro-temporal domain. This is achieved by using two dimensional (2-D) autoregressive (AR) modeling of the speech signal (Ganapathy *et al.*, 2014). The next step is the modulation filtering of the 2-D AR spectrogram using spectro-temporal filters.

The automatic speech recognition (ASR) experiments are performed on the noisy speech from the Aurora-4 database using a deep neural network (DNN) acoustic model. We study the effect of temporal as well as spectral smearing using the modulation filters for noise robustness. The results from these experiments, which are similar to the conclusions from the human psychophysical studies reported in Elliott and Theunissen (2009), indicate that the important modulations in the temporal domain are band-pass in nature while they are low-pass in the spectral domain. Furthermore, language identification (LID) experiments performed on highly degraded radio channel speech (Walker and Strassel, 2012) confirm the generality of the proposed features for a wide range of noise conditions.

The rest of the paper is organized as follows. Section 2 describes the two stages of the proposed feature extraction approach—the derivation of the 2-D AR spectrogram followed by the application of modulation filtering. The speech recognition and language identification experiments are reported in Sec. 3 and Sec. 4, respectively. In Sec. 5, we summarize the important contributions from this work.

2. Feature extraction

The block schematic of the proposed feature extraction scheme is shown in Fig. 1. The input speech signal is processed in 1000 ms analysis windows and a long-term discrete cosine transform (DCT) is applied. The DCT coefficients are then band-pass filtered with Gaussian shaped mel-band windows and used for frequency domain linear prediction (FDLP) (Athineos and Ellis, 2007). The FDLP technique attempts to predict $X[k]$ with a linear combination of $X[k-1]$, $X[k-2]$,…, $X[k-p]$, where $X[k]$ denotes the DCT value at frequency index k and p denotes the order of FDLP. This prediction process estimates an AR model of the sub-band temporal envelope.

The sub-band FDLP envelopes are then integrated in short-term windows (25 ms with a shift of 10 ms). The integrated envelopes are stacked in a column-wise manner as shown in Fig. 1 and the energy values across the frequency sub-bands for each frame provides an estimate of the power spectrum of the signal (Ganapathy *et al.*, 2014). These estimates generate autocorrelation values which can be used in the conventional time domain linear prediction (TDLP) (Makhoul, 1975) framework to model the power spectrum. At the end of this two stage process, we obtain the 2-D AR spectrogram which emulates the peak preserving property of the human auditory system and suppresses the low energy regions of the signal which are vulnerable to noise.

The final step is the modulation filtering of the spectrogram to extract the key dynamics in the temporal modulations [rate frequencies (Hz)] and spectral modulations [scale frequencies (cycles per kHz)]. This is achieved by windowing the 2-D DCT

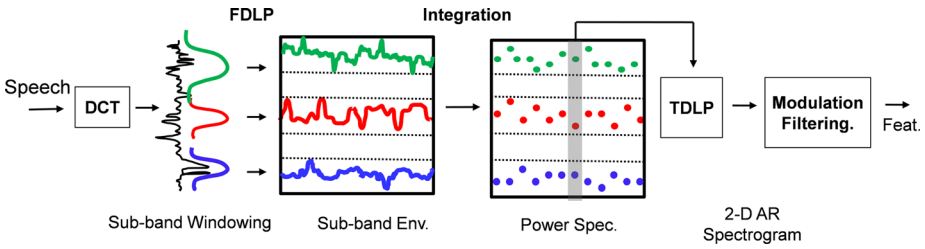

Fig. 1. (Color online) Block schematic of the proposed feature extraction scheme using modulation filtering of 2-D AR spectrograms.

transform of the spectrogram (similar to image filtering using window functions). The AR model spectrogram from the previous step with the temporal context of the entire recording and the full spectral context (0–4 kHz) is transformed using 2-D DCT. The 2-D DCT space contains the amplitude value for each rate of change (modulation) in the spectral and temporal dimension. We design window functions in this 2-D DCT space which have a passband value of unity in the spectro-temporal patch of interest and a smooth Gaussian shaped decay at the transition band. For example, a temporal band-pass (0.25–15 Hz), spectral low pass (0–1.0 cycles per kHz) filter is designed by mapping this range of modulations to the corresponding range in the 2-D DCT space. A unity value is assigned to the pass-band range with a smooth transition to a value of zero outside this range. Since each audio recording has a different length, the window functions are derived separately for each audio file. The application of these windows on the 2-D DCT space implies a modulation filtering of the spectrogram. The windowed 2-D DCT is transformed with inverse 2-D DCT function to obtain the modulation filtered spectrogram.

The illustration of the robustness achieved by the proposed approach is shown in Fig. 2. Here, we plot the spectrographic representation of the speech signal in three conditions—clean speech, noisy speech [additive babble noise at 10 dB signal-to-noise ratio (SNR)], and radio channel speech [from channel C in the RATS database (Walker and Strassel, 2012)]. The plots compare the representation from the conventional mel frequency analysis with the representation obtained from the modulation filtering of the 2-D AR spectrograms. As seen here, the proposed approach yields a representation focusing on important regions of the clean signal. For the degraded conditions, the representation provides a good match with the clean signal suppressing the effects of noise. As shown in the experiments, this is useful in improving the robustness of speech applications in mismatched conditions.

3. Noisy speech recognition experiments

We perform automatic speech recognition (ASR) experiments in the Aurora4 database using a deep neural network (DNN) system. We use the clean training setup which contains 7308 clean recordings (14 h) for training the acoustic models using the Kaldi toolkit (Povey *et al.*, 2011). The system uses a tri-gram language model with 5000 vocabulary size. The test data consist of 330 recordings each from six noisy conditions which include train, airport, babble, car, restaurant, and street noise at 5–15 dB SNR.

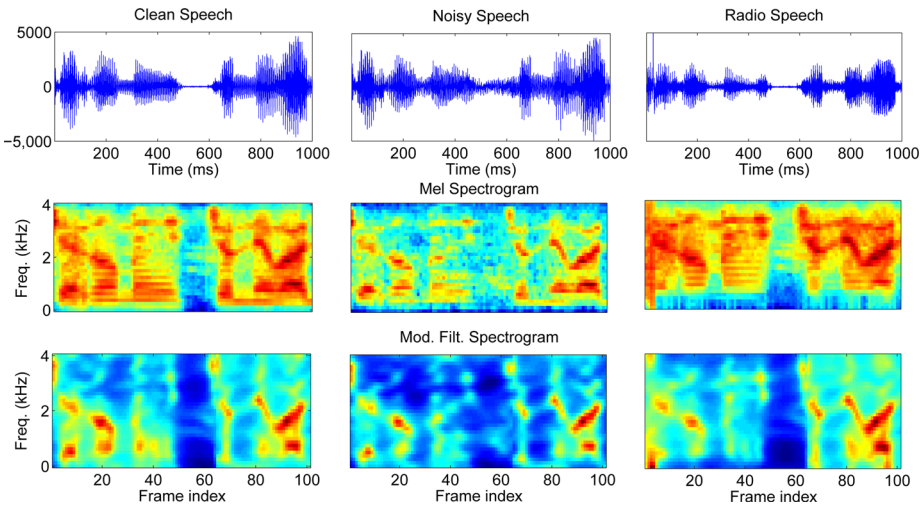

Fig. 2. (Color online) Comparison of the spectrographic representation provided by mel frequency analysis and the proposed modulation filtering approach for a clean speech signal, noisy speech signal (additive babble noise at 10 dB SNR) and radio channel speech (non-linear noise from channel C).

For the proposed features, we use a 200 ms context of the sub-band energies decorrelated by a DCT. The features from each sub-band are spliced together with their frequency derivatives to form the input for the DNN. We use a DNN with four hidden layers of 1024 activations and uses context dependent phoneme targets. The performance of the ASR system is measured in terms of word error rate (WER).

In order to determine the important modulations in the spectral and temporal domain, we use the average ASR performance on the six additive noisy conditions. The performance as a function of the rate frequency is shown in the top panel of Fig. 3. The first observation is that the performance improves by a band-pass filtering compared to low-pass filtering. The results with band-pass filtering indicate that an upper cut-off frequency of 15 Hz gives the best speech recognition performance on noisy speech.

The ASR performance as a function of the scale frequency is shown in the bottom panel of Fig. 3. Unlike the variation with respect to the rate frequency, the ASR performance is significantly better with a low-pass filtering in the spectral modulation domain. The best performance is achieved with a scale filtering in the 0–1 cycles per kHz range. It is also important to note that the ASR results shown in Fig. 3 follow a similar trend to the human speech recognition results on noisy speech reported in Elliott and Theunissen (2009) where it was shown that the modulation transfer function (MTF) for speech comprehension lies in the band-pass temporal modulations with an upper cut-off frequency of 12 Hz and low pass spectral modulations below 1 cycle per kHz. This interesting similarity is observed even with a stark difference between the ASR back-end using a DNN and the auditory cortex.

In Table 1, we compare the performance of the proposed approach with various feature extraction methods, namely, mel filter bank energies (MFBE) (Davis and Mermelstein, 1980), power normalized cepstral coefficients (PNCC) based filter bank energies (PNFBE) (Kim and Stern, 2012) and Advanced ETSI front-end (ETSI, 2002). In order to understand the impact of the two steps involved in the proposed approach, namely, the derivation of 2-D spectrogram and the modulation filtering, we experiment with features generated with each one of these individually, namely, the 2-D AR spectrogram alone without the modulation filtering (2-D AR) as well as the features derived from the modulation filtering of mel spectrogram (MFBE + Mod.Filt.).

Among the baseline features, the PNFBE method provides the best performance on clean conditions and the ETSI features provide the best performance on additive noise conditions. The methods of 2-D AR modeling provided by 2-D AR features

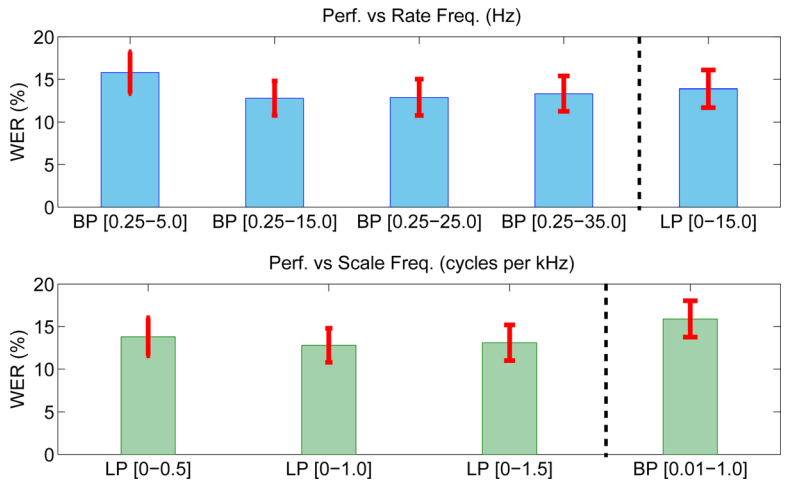

Fig. 3. (Color online) ASR performance in terms of word error rate [WER (%)] with standard deviation (error bar) as a function of the rate frequency (Hz) and scale frequency (cycles per kHz). Here, LP denotes low-pass filtering, BP denotes band-pass filtering, and the two frequencies in the *x* axis indicate the lower and upper cut-off frequency.

as well as the modulation filtering with mel filter bank energies (MFBE + Mod.Filt.) improve the performance on the noisy conditions without degrading the performance on clean conditions. The best performance is achieved by using the proposed scheme of using these two steps in sequence, namely, the derivation of 2-D AR spectrogram from the speech signal followed by the modulation filtering with band-pass representation in the temporal domain and low pass filtering in the spectral domain (average relative improvements on 17% on the additive noise conditions with the same microphone and 10% on the additive noise conditions with different microphone over the ETSI features). For the noisy conditions, the relative improvement of the proposed approach over the MFBE + Mod.Filt. features is statistically significant (p-value < 0.01), which shows that the combination of the 2-D AR modeling and modulation filtering improves robustness.

4. Language identification of radio speech

The development and test data for the LID experiments use the LDC releases of RATS LID evaluation (Walker and Strassel, 2012). This consists of clean speech recordings passed through noisy radio communication channels with each channel inducing a degradation mode to the audio signal based on specific device non-linearities, carrier modulation types and network parameter settings. In the RATS initiative, a set of eight channels (channels A-H) is used with specific parameter settings and carrier modulations. The five target languages are Levantine-Arabic, Farsi, Dari, Pashto, and Urdu. In order to investigate the effects of an unseen communication channel (not seen in training), we divide the eight channels to two groups—channels B,E,G,H used in the training and the channels A,C,D,F used in testing.

The training data consist of 24 123 recordings with 270 h of data from each of the four noisy communication channels (B,E,G,H) and the test set consists of 7164 recordings with about 15 h of data from each of the eight channels (A–H). The training and test recordings have speech segments with 120, 30, and 10 s of speech. The features are processed with feature warping (Pelecanos and Sridharan, 2001) and are used to train a Gaussian mixture model-Universal background model (GMM-UBM) with

Table 1. Word error rate (%) in Aurora-4 database with clean training for various feature extraction schemes.

Cond.	MFBE	ETSI	PNFBE	2-D AR	MFBE + Mod. Filt.	Prop.
			Clean Same Mic			
Clean	3.1	3.1	**2.8**	3.1	2.9	3.3
			Clean Diff. Mic			
Clean	14.9	14.8	**11.3**	**11.3**	11.7	**11.3**
			Additive Noise Same Mic			
Airport	23.6	13.6	17.6	15.4	14.4	**13.3**
Babble	20.7	14.1	15.9	15.2	14.9	**13.5**
Car	8.0	8.7	5.9	5.6	**5.1**	5.2
Restaurant	26.3	19.4	21.9	19.1	19.0	**17.2**
Street	19.8	18.3	16.9	14.8	14.1	**13.0**
Train	20.8	16.9	16.0	14.9	**13.9**	14.2
Avg.	19.9	15.2	15.7	14.2	13.6	**12.7**
			Additive Noise Diff. Mic			
Airport	41.5	29.9	35.6	31.2	30.9	**30.0**
Babble	38.4	31.3	34.3	31.1	32.4	**30.4**
Car	25.8	23.9	20.7	17.8	**17.7**	18.4
Restaurant	41.3	34.0	37.4	32.4	32.7	**30.9**
Street	38.1	33.5	33.1	29.2	29.3	**28.1**
Train	37.3	32.1	31.7	29.2	29.3	**28.9**
Avg.	37.1	30.8	32.1	28.5	28.7	27.8

Table 2. LID performance [equal error rate (EER %)] for various features on the RATS database using an LID system trained on channels B,E,G,H and tested on seen channels B,E,G,H as well as unseen channels A,C,D,F with 120, 30, and 10 s speech duration.

Cond.	MFCC	MVA	PNCC	Prop.
		120 s		
Avg. Seen	3.1	**2.3**	2.4	**2.3**
Chn. A	21.0	12.5	15.0	**7.0**
Chn. C	14.5	16.6	13.9	**12.8**
Chn. D	18.5	16.6	13.1	**12.0**
Chn. F	12.4	19.9	7.7	**5.0**
Avg. Unseen	16.6	16.4	12.4	**9.2**
		30 s		
Avg. Seen	3.7	3.7	**3.4**	3.9
Chn. A	21.0	13.3	17.5	**10.8**
Chn. C	13.8	15.4	10.9	**10.3**
Chn. D	22.0	19.1	16.1	**13.6**
Chn. F	11.5	16.7	10.1	**6.7**
Avg. Unseen	17.1	16.1	13.7	**10.4**
		10 s		
Avg. Seen	9.1	**8.8**	8.9	8.9
Chn. A	24.5	20.0	23.6	**14.8**
Chn. C	20.0	22.1	19.4	**16.9**
Chn. D	24.3	22.9	**19.5**	**19.5**
Chn. F	17.3	23.2	14.5	**13.1**
Avg. Unseen	21.3	22.1	19.3	**16.1**

1024 mixture components. Then, an i-vector projection model of 300 dimensions is trained (Dehak *et al.*, 2011). The back-end classifier is a multi-layer perceptron (MLP) having a single hidden layer of 2000 units. The MLP is trained with the input i-vectors and the language labels as the targets. The performance of the LID system is measured in terms of equal error rate (EER).

We experiment with various feature extraction schemes like MFCC features, MVA features (Chen and Bilmes, 2007), PNCC features (Kim and Stern, 2012), and the proposed features which involve 2-D AR modeling followed by modulation filtering and cepstral transformation. All the features are processed with delta and acceleration coefficients before training the GMM.

The performance of the various features for the seen conditions {channels B,E,G,H} and unseen conditions {channels A,C,D,F} for different speech segment durations is reported in Table 2. The proposed approach of using modulation filtered 2-D AR spectrograms provides significant improvements for unseen radio channel conditions (average relative improvements of 17%–25% in terms of EER) compared to the baseline PNCC system. These results are in conjunction with the ASR results and indicate the consistency of the proposed approach for variety of speech applications involving various types of artifacts like additive noise, convolutive noise as well as non-linear radio channel distortions.

5. Summary

The main contributions from the paper are the following:

(1) Identifying the key modulations in the spectral and temporal domain for robust speech applications—bandpass filtering in the temporal domain and low-pass filtering in the spectral domain.

(2) Peak picking in the spectro-temporal domain using 2-D AR modeling yields a robust spectrogram of the speech signal.
(3) Combining the above steps by modulation filtering of 2-D AR spectrogram provides significant improvements to unseen conditions without assuming any model of the noise or channel.

Acknowledgments

This work was supported by the DARPA Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the authors and do not reflect the official policy of the Department of Defense or the U.S. Government. The authors would like to thank the contributions of Sri Harish Mallidi and Vijayaditya Peddinti for the software fragments used in the experiments.

References and links

Athineos, M., and Ellis, D. P. W. (**2007**). "Autoregressive modelling of temporal envelopes," IEEE Trans. Signal Proc. **55**, 5237–5245.
Chen, C., and Bilmes, J. A. (**2007**). "MVA processing of speech features," IEEE Trans. Audio Speech Lang. Process. **15**(1), 257–270.
Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906.
Davis, S., and Mermelstein, P. (**1980**). "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Proc. **28**, 357–366.
Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (**2011**). "Front-end factor analysis for speaker verification," IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798.
Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**(2), 1053–1064.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**(3), e1000302.
ETSI (**2002**). "ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," http://www.etsi.org/deliver/etsi_es/202000_202099/202050/01.01.05_60/es_202050v010105p.pdf.
Ganapathy, S., Mallidi, S. H., and Hermansky, H. (**2014**). "Robust feature extraction using modulation filtering of autoregressive models," IEEE Trans. Audio Speech Lang. Process. **22**(8), 1285–1295.
Greenberg, S., Ainsworth, W. A., Popper, A. N., and Fay, R. R. (**2004**). *Speech Processing in the Auditory System* (Springer, New York), Vol. 18, Chap. 1, pp. 17–20.
Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Proc. **2**(4), 578–589.
Keurs, T. M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am. **91**(5), 2872–2880.
Kim, C., and Stern, R. M. (**2012**). "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proceedings of Int. Conf. on Acoust. Speech and Signal Proc.* (IEEE), pp. 4101–4104.
Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580.
Nemala, S. K., Patil, K., and Elhilali, M. (**2013**). "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," IEEE Trans. Audio Speech Lang. Proc. **21**(2), 416–426.
Palmer, A., and Shamma, S. (**2004**). *Physiological Representations of Speech: Speech Processing in the Auditory System* (Springer, New York), Chap. 4, pp. 163–230.
Pelecanos, J., and Sridharan, S. (**2001**). "Feature warping for robust speaker verification," in *Proc. IEEE Odyssey Speaker Lang. Recognition Workshop* (IEEE), pp. 213–218.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsk, J., Stemmer, G., and Vesel, K. (**2011**). "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recog. and Understanding* (IEEE), 1–4.
Walker, K., and Strassel, S. (**2012**). "The RATS radio traffic collection system," in *Proc. IEEE Odyssey Speaker Lang. Recog. Workshop* (IEEE).