EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Auditory processing-based features for improving speech recognition in adverse acoustic conditions

Hari Krishna Maganti* and Marco Matassoni

Abstract

The paper describes an auditory processing-based feature extraction strategy for robust speech recognition in environments, where conventional automatic speech recognition (ASR) approaches are not successful. It incorporates a combination of gammatone filtering, modulation spectrum and non-linearity for feature extraction in the recognition chain to improve robustness, more specifically the ASR in adverse acoustic conditions. The experimental results with standard Aurora-4 large vocabulary evaluation task revealed that the proposed features provide reliable and considerable improvement in terms of robustness in different noise conditions and are comparable to those of standard feature extraction techniques.

Introduction

Present technological advances in speech processing systems aim at providing robust and reliable interfaces for practical deployment. Achieving robust performance of these systems in adverse and noisy environments is one of the major challenges in applications such as dictation, voice-controlled devices, human-computer dialog systems and navigation systems. Speech acquisition, processing and recognition in non-ideal acoustic environments are complex tasks due to presence of unknown additive noise and reverberation. Additive noise from interfering noise sources and convolutive noise arising from acoustic environment and transmission channel characteristics mostly contribute to the degradation of speech intelligibility as well as the performance of speech recognition systems. This article addresses the problem of achieving robustness in large vocabulary automatic speech recognition (ASR) systems by incorporating principles inspired by cochlea processing in the human auditory system.

The human auditory processing system is a robust front-end for speech recognition in adverse conditions. In the recently conducted PASCAL CHiME challenge [1], which aimed at source separation and robust speech recognition in noisy conditions similar to that of daily life, the performance of a human was much better than that of the ASR standard baseline for different signal-to-noise ratios (SNRs). As seen from Figure 1, the performance of a human is more robust and consistent than the ASR baseline. Further, the performance of both ASR baseline and human improved in line with the increase in SNR. This plot shows how susceptible the present systems are compared with a human listener with latest noise experimental setup.

The degradation of recognition accuracy for ASR systems in noisy environments is mostly due to the discrepancy between training and testing conditions. The training data are recorded in clean conditions, and the accuracy gets degraded when it is tested against data acquired in noisy conditions. Various speech signal enhancement, feature normalization and model parameterization techniques are applied at various phases of processing to reduce the mismatch between training and testing conditions [2,3]. Spectral subtraction-, Wiener filtering-, statistical model- and subspace-based speech enhancement techniques aim at improving the quality of speech signal captured through a single microphone or microphone array [4,5]. Feature normalization attempts to represent parameters that are less sensitive to noise by modifying the extracted features. Common techniques include cepstral mean normalization (CMN) which forces the mean of each element of the cepstral feature vector to be zero for all utterances. Other variants include

*Correspondence: maganti@ieee.org
Center for Information and Communication Technology, Fondazione Bruno Kessler, via Sommarive 18, Trento 38123, Italy

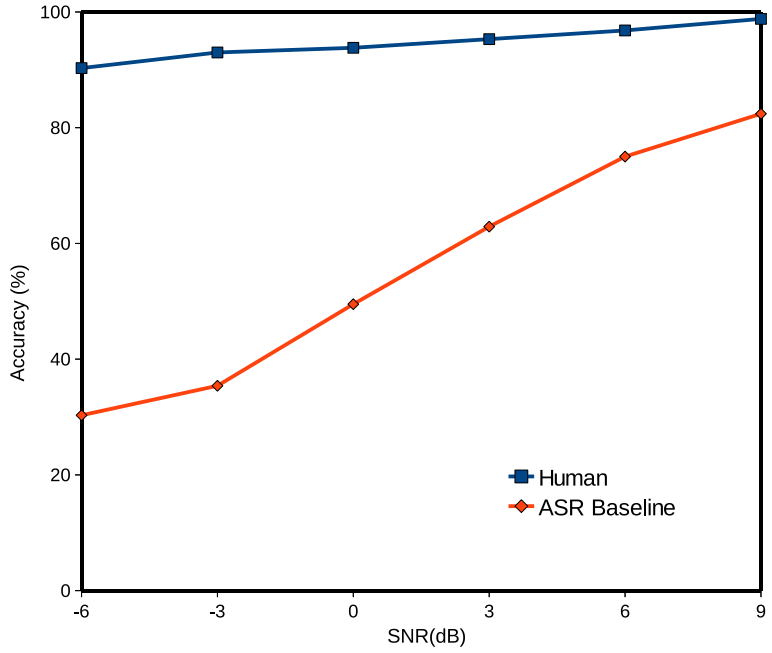

Figure 1 Comparison of performance accuracy for human and ASR baseline for different SNRs.

mean-variance normalization (MVN), cepstral mean subtraction and variance normalization (CMSVN) and relative spectral (RASTA) filtering [2,6]. Model adaptation approaches modify the acoustic model parameters' match with the observed speech features [4,7].

The auditory system-based techniques have been used in speech recognition to improve the robustness [8-15]. Examples of non-uniform frequency resolution in popular speech analysis techniques include Mel frequency-based features and perceptual linear prediction which attempt to emulate human auditory perception. The gammatone filter bank with non-uniform bandwidths and non-uniform spacing of center frequencies provided better robustness in adverse noise conditions for speech recognition tasks [12-15].

Another important characteristic, the modulation spectrum of speech, represents low temporal modulation components and is important for speech intelligibility [16,17]. Similar to the perceptual ability of human auditory system, the relative prominence of slow temporal modulations is different at various frequencies. The gammatone filter bank-derived modulation spectral features have shown to improve the robustness for far-field speaker identification [18]. Our previously described auditory-based modulation spectral feature is a combination of gammatone filtering and modulation spectral features for robust speech recognition [19].

The present paper describes an alternate approach, in which the gammatone filtering, non-linearity and modulation spectrum for feature extraction are combined. The enhanced speech signal improved the accuracy of the system by reducing the sensitivity. The features derived from the combination are used to provide robustness, particularly in the context of mismatch between training and testing in noisy environments. The studied features are shown to be reliable and robust to various noises for a large vocabulary task. For comparison purposes, the recognition results obtained by using conventional features are tested, and the usage of the proposed features is proved to be efficient.

The paper is organized as follows: Section Related work gives an overview of the auditory-inspired features including gammatone filter bank processing, modulation spectrum and non-linearity processing. Section Auditory processing-based features describes the methodology for feature extraction. Section Database description and experiments presents database description and experiments. Section Recognition results discusses the results, and finally, Section Conclusions concludes the paper.

Related work

Most state-of-the-art ASR systems perform far below the human auditory system in the presence of noise. Auditory modeling, which simulates some properties of the human auditory system, has been applied to speech recognition systems to enhance robustness. The information coded in auditory spike trains and the information transfer processing principles found in the auditory pathway are used in [20]. The neural synchrony is used for creating noise-robust representations of speech. The model parameters

are fine-tuned to conform to the population discharge patterns in the auditory nerve which are then used to derive estimates of the spectrum on a frame-by-frame basis. This was extremely effective in noise and improved performance of the ASR dramatically. Various auditory processing-based approaches were proposed to improve robustness, and in particular, the works described in [13,20] were focused to address the additive noise problem. Further, in [21], a model of auditory perception (PEMO) developed by Dau et al. [22] is used as a front end for ASR, which performed better than the standard Mel-frequency cepstral coefficient (MFCC) for an isolated word recognition task. The auditory processing-related principles attempted to model human hearing to some extent have been applied for speech recognition [6,17]. The modulation spectrum is an important psychoacoustic property which represents a slow temporal modulation which is significant for determining speech intelligibility. For improving robustness, the normalized modulation spectra have been proposed in [23]. Similar work in the context of large vocabulary speech recognition such as noisy Wall Street Journal (New York, NY, USA) and GALE task as reported in [24,25].

Feature extraction at different stages of the auditory model output to determine which component has the highest impact on the accuracy of recognition has been studied [26]. The study also evaluated the contribution of each stage in auditory processing for improving robustness on the resource management database by using SPHINX-III speech recognition system (Carnegie Mellon University, Pittsburgh, PA, USA). Particularly, the effects of rectification, non-linearities, short-term adaptation and low-pass filtering were shown to contribute the most to robustness at low SNRs.

In another study [8], the techniques motivated by human auditory processing are shown to improve the accuracy of automatic speech recognition systems. It was shown that non-linearities in the representation, especially non-linear threshold effect, played important role in improving robustness. Other important aspect was the impact of time-frequency resolution based on the observations that the best estimates of attributes of noise are obtained by using relatively long observation windows and frequency smoothing provides significant improvements to robust recognition.

In the context of speaker identification, auditory-based features have been extensively studied [27]. The contrasts of MFCC and gammatone frequency cepstral coefficients (GFCC) have been compared, and the noise robust improvements by GFCC has been explained in [28].

In our earlier studies [19], several auditory processing-motivated features have shown considerably improved robustness for both additive noise and reverberation. However, all these above studies are confined to small and medium vocabulary tasks. In that direction, it is an attempt to apply these techniques for large and complex vocabulary task, namely Aurora-4, which is based on Wall Street Journal database. Artificially added noises ranged from SNRs of 5 to 15 dB with a variety of noises which include babble, car, street and restaurant. The effects at different stages of processing are analyzed to study the contribution of each stage for improving robustness. A preliminary version of our work was presented earlier [29].

Auditory processing-based features

In this section, a general overview of gammatone filter bank-, non-linearity- and modulation spectrum-based auditory features is presented.

Gammatone filter bank

Gammatone filters are linear approximation of physiologically motivated processing performed by the cochlea [30]. It is commonly used in modeling the human auditory system and consists of a series of bandpass filters. In the time domain, the filter is defined by the following impulse response:

$$g(t) = at^{n-1} cos\left(2\pi f_c t + \phi\right) e^{-2\pi bt}, \tag{1}$$

where n is the order of the filter, b is the bandwidth of the filter, a is the amplitude, f_c is the filter center frequency, and ϕ is the phase.

The filter center frequencies and bandwidths are derived from the filter's equivalent rectangular bandwidth (ERB) as detailed in [30]. In [31], Glasberg and Moore relate center frequency and the ERB of an auditory filter as

$$\mathrm{ERB}\left(f_c\right) = 24.7\left(\frac{4.37 f_c}{1000} + 1\right) \tag{2}$$

The filter output of the mth gammatone filter, X_m can be expressed by

$$X_m(k) = x(k) * h_m(k), \tag{3}$$

where $h_m(k)$ is the impulse response of the filter.

The frequency response of the 32-channel gammatone filter bank is as shown in Figure 2.

Non-linearity

The sigmoid non-linearity that represents physiologically observed rate-level non-linearity is the same as that described in [26] and given by

$$y_i(t) = \frac{w_2}{1 + e^{(w_1 x_i[t] + w_0)}}, \tag{4}$$

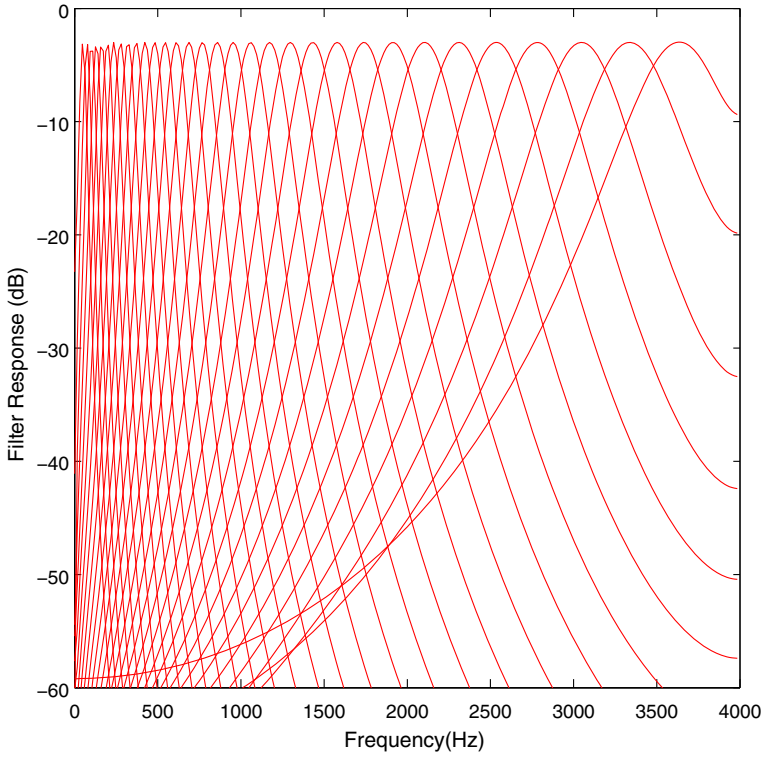

Figure 2 Frequency response for the 32-channel gammatone filter bank.

where $x_i[t]$ is the ith channel log gamma spectral value, and $y_i[t]$ is the corresponding sigmoid compressed value. The optimal parameters are derived from evaluation of resource management development set in additive noise at 10 dB [26].

Modulation spectrum

The long-term modulations examine the slow temporal evolution of the speech energy with time windows in the range from 160 to 800 ms, contrary to the conventional short-term modulations studied with time windows of 10 to 30 ms which capture rapid changes of the speech signals. The modulation spectrum $Y_m(f,g)$ is obtained by applying Fourier transform on the running spectra, obtained by taking absolute values $|Y(t,f)|$ at each frequency, where $Y(t,f)$ is the time-frequency representation after short-time Fourier analysis, expressed as

$$Y_m(f,g) = FT\left[|Y(t,f)|\right]|_{t=1,\ldots T}, \tag{5}$$

where T is the total number of frames, and g is the modulation frequency. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [16,17]. In [17], it has been shown that for noisy environments, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility, particularly the band below 1 Hz which contains mostly information about the environment. Therefore, the recognition performance can be improved by suppressing this band in the feature extraction. Figures 3 and 4 show the spectrogram, gammatonegram and gammatonegram with non-linearity plots for two types of noise-corrupted utterance. It can be observed that the gammatonegram with non-linearity plots for babble and restaurant noises provide cleaner spectral information which is important for speech recognition.

Database description and experiments

The Aurora 4 evaluation task provides a standard database for comparing the effectiveness of robust techniques in LVCSR tasks in the presence of mis-matched channels and additive noises. It is a part of the ETSI standardization process and derived from the standard 5k WSJ0 Wall Street Journal database. It has 7,180 training utterances of approximately 15 h and 330 test utterances with an average duration of 7 s.

The acoustic data (both training and test) are also available in two different sampling frequencies (8 and 16 kHz), compressed or uncompressed. Two different

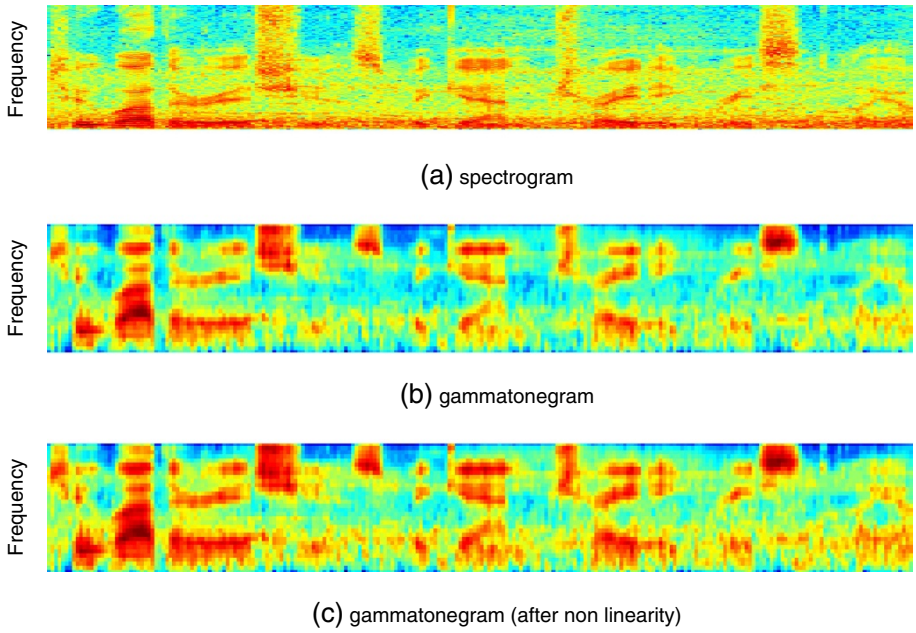

Figure 3 Spectrogram (a), gammatonegram (b) and gammatonegram with non-linearity (c) plots for babble noise-corrupted utterance.

training conditions were specified. Under clean training (clean train), the training set is the full SI-84 WSJ train set processed with no noise added. Under multicondition training (multi-train), about half of the training data were recorded using one microphone; the other half were recorded under a different channel (also used in some of the test sets) with different types of noise and different SNRs added. The noise types are similar to the noisy conditions in test.

The Aurora 4 test data include 14 test sets from two different channel conditions and six different added noises (in addition to the clean environment). The SNR was

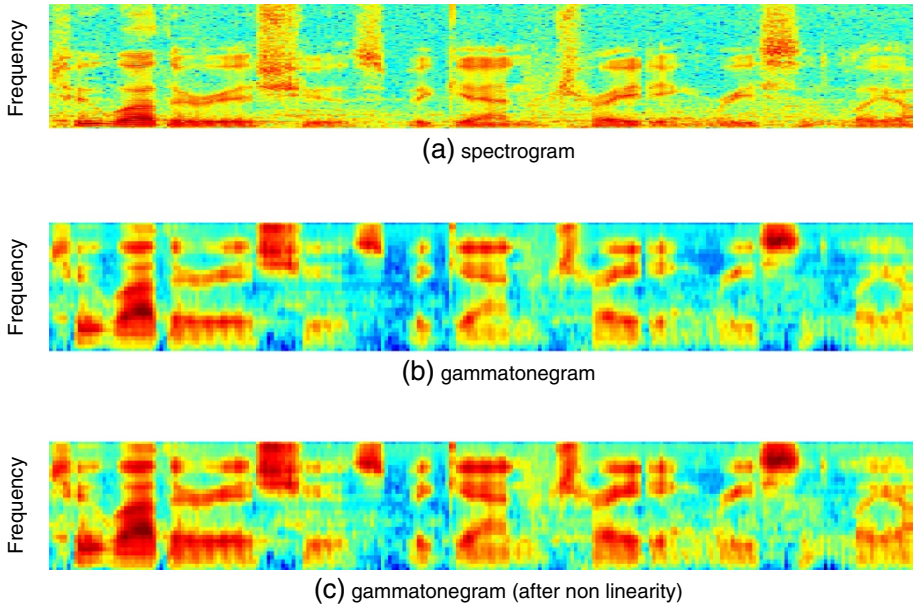

Figure 4 Spectrogram (a), gammatonegram (b) and gammatonegram with non-linearity (c) plots for restaurant noise-corrupted utterance.

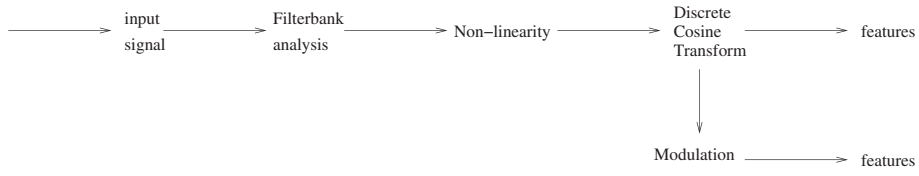

Figure 5 Processing stages of the non-linear spectro-temporal feature extraction.

randomly selected between 0 and 15 dB on an utterance-by-utterance basis. Six noisy environments and one clean environment no noise (set01), car (set02), babble (set03), restaurant (set04), street (set05), airport (set06) and train (set07) are considered in the evaluation set which comprises 5,000 words under two different channel conditions. The original audio data for test conditions 1 to 7 was recorded with a Sennheiser microphone (Lower Saxony, Germany), while test conditions 8 to 14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones. These included such common types as a Crown PCC-160 (Elkhart, IN, USA), Sony ECM-50PS (New York, NY, USA) and a Nakamichi CM100 (Tokyo, Japan). Noise was digitally added to this audio data to simulate operational environments.

The block schematic for the feature extraction technique is shown in Figure 5. The speech signal first undergoes pre-emphasis (with a coefficient of 0.97), which flattens the frequency characteristics of the speech signal. The signal is then processed by a gammatone filter bank which uses 32 frequency channels equally spaced on the equivalent ERB scale as shown in Figure 2. The computationally efficient gammatone filter bank implementation as described in [32] is used. The gammatone filter bank transform is computed over L ms, and the segment is shifted by n ms. The log magnitude resulting coefficients are then decorrelated by applying a discrete cosine transform (DCT). The computations are made over all the incoming signal, resulting in a sequence of energy magnitudes for each band sampled at $1/n$ Hz. Then, frame-by-frame analysis is performed, and a N-dimensional parameter is obtained for each frame. The modulation energy of each coefficient, which is defined as the Fourier transform of its temporal evolution, is computed. In each band, the modulations of the signal are analyzed by computing FFT over the P ms Hamming window, and the segment is shifted by p ms. The energies for the frequencies between the 2 and 16 Hz, which represent the important components for the speech signal are computed. For the experiments and gammatonegrams shown in Figures 3 and 4, the values of L, n and N are 25 ms, 10 ms and 32, respectively, and modulation parameters of P and p with 160 and 10ms, respectively, are used.

Recognition results

The HTK setup followed is three-state cross-word triphone models tied to approximately 3,000 tied states, each represented by four-component Gaussian mixtures with diagonal covariance, together with the 5,000 closed vocabulary bigram language model (LM) [33]. Triphone states were tied using the linguistic-driven top-down decision-tree clustering technique, resulting in a total of 3,135 tied states in clean train and 3,068 tied states in multi-train. The CMU dictionary was used to map lexical items into phoneme strings, and the 5,000-word closed vocabulary bigram LM was used. The LM weights, pruning thresholds and insertion penalties were based on [33].

In order to analyze the effect of the non-linearity (Equation 4) on phone recognition rate, small subsets with a random number of utterances from AURORA-4 multi-condition training data are used. Experiments with training on clean condition are considered, because the purpose is precisely to test robustness in the presence of noise while retaining similar performance in clean conditions. All experiments have been performed with 16-kHz

Table 1 Accuracy rate (%) baselines for different feature extraction techniques

Channel	MFCC	PLP	GFCC
1	89.3	90.5	88.43
2	77.6	77.1	80.7
3	61.7	63.8	64.5
4	53.8	55.5	58.4
5	57.9	57.5	64.4
6	62.5	62.2	63.6
7	53.2	53.5	61.6
8	72.9	73.0	71.8
9	58.4	61.8	63.2
10	45.2	47.7	48.6
11	41.1	43.2	45.9
12	36.2	38.9	46.1
13	47.0	48.0	49.8
14	38.3	40.2	46.9
Average	56.7	58.0	61.0

Table 2 Accuracy rates (%) for the different extraction techniques

Channel	GFCC-MS	GFCC-MS-NL	GFCC-NL
1	87.3	89.9	89.6
2	80.7	83.4	81.1
3	62.6	72.8	70.9
4	56.0	69.2	65.6
5	64.5	71.6	69.4
6	62.6	73.2	70.1
7	62.0	66.9	68
8	69.7	74.4	71
9	61.7	63.8	62.4
10	48.2	52.3	51.9
11	44.5	51.2	50.1
12	43.0	50.8	48.4
13	47.2	54.3	51.6
14	46.8	50.4	50.2
Average	59.8	64.3	66.0

data of the Aurora-4 database. Table 1 shows the results in percent accuracy for the different features. The average performance for all the noise conditions for the different features is shown at the last row of the table. MFCC, perceptual linear prediction (PLP) and GFCC are the standard 39-dimensional Mel-frequency, perceptual linear prediction and gammafrequency cepstral coefficient features along with their delta and acceleration derivatives. From Table 1, it is clear that the traditional MFCC features have the lowest accuracy indicating inefficiency of these features for noisy environments. Also, it can be seen that GFCC has the best performance compared to PLP which, in turn, was better when compared to MFCC which is consistent with earlier studies [13,14].

Table 2 shows the results for gammafrequency with modulation spectrum (GFCC-MS), gammafrequency with modulation spectrum and non-linearity (GFCC-MS-NL) and gammafrequency with non-linearity (GFCC-NL) feature extraction techniques. For this task, we can see that the GFCC-MS do not provide any improvement which is contrary to our earlier study [29]. In our earlier study, the combination of GFCC and modulation spectrum was better than GFCC alone for isolated word recognition in reverberant environment of around 0.3 to 0.5 s.

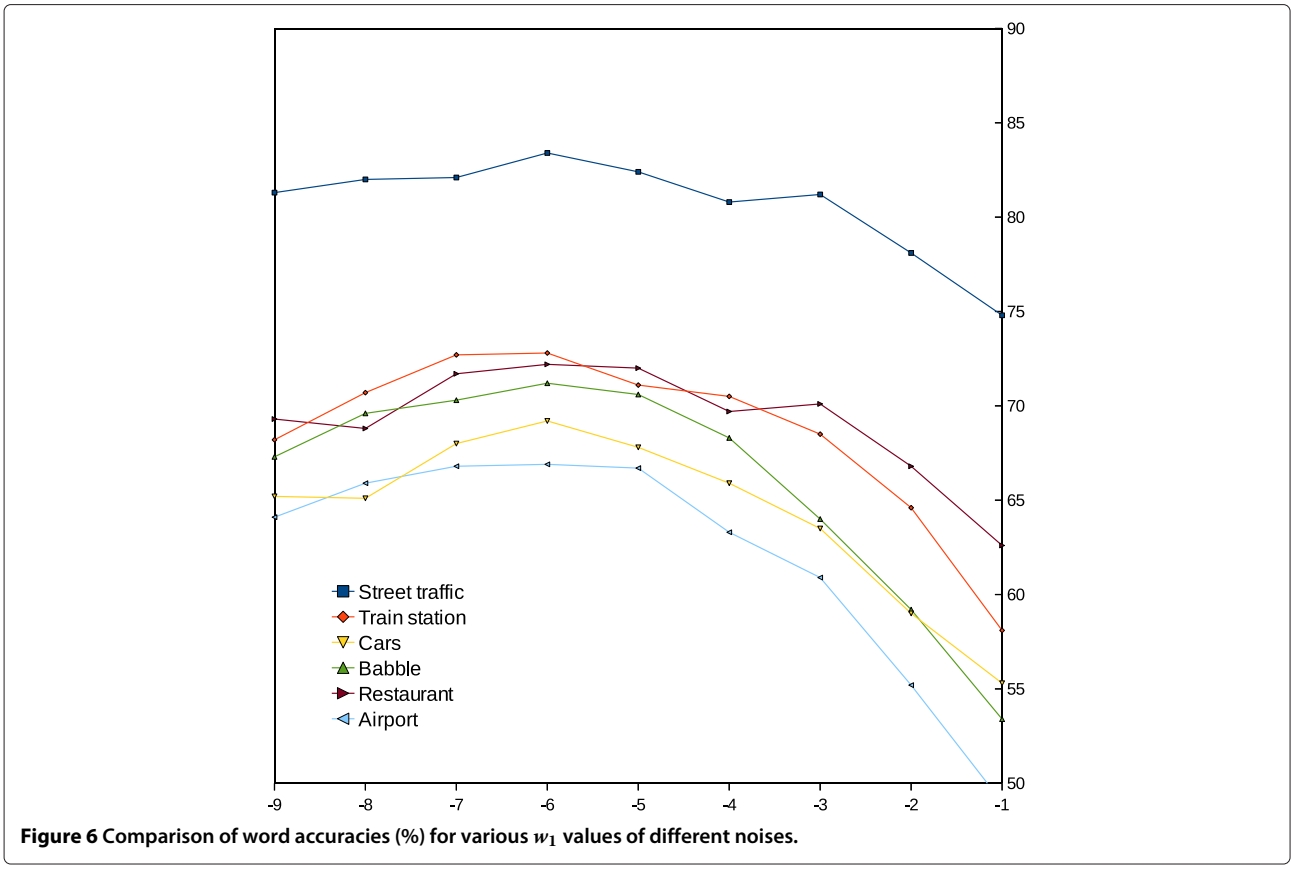

Figure 6 Comparison of word accuracies (%) for various w_1 values of different noises.

We hypothesize that we do not observe the similar effect in this case due to different task (large vocabulary with triphones) and different environment (only additive). However, from the table, we can see that the optimized non-linearity improved the performance of GFCC and GFCC-MS considerably. Further, we can also be observe that the contribution towards improved performance from the non-linearity is consistent for all types of noises. This clearly demonstrates that including a non-linearity is significantly beneficial for improving robustness in noisy environment.

The features are computed with w_2 = 1.0 and various w_0 and w_1 combinations. As seen from Figure 6, the selection of the weights is crucial for improving the recognition performance. It can also be observed that for w_1 ranging from -0.7 to -1.8, the performance is better than those of GFCC-MS and GFCC. The best performance for this task is obtained with $w_0 = 1$ and $w_1 = -0.9$ which are used for the experiments reported in Table 2.

Conclusions

The features proposed in the present study are derived from auditory characteristics, which include gammatone filtering, non-linear processing and modulation spectral processing, emulating the cochlear and the middle ear to improve robustness. In earlier studies, several auditory processing-motivated features have improved robustness for small and medium vocabulary tasks. The paper has studied the application of these techniques to large and complex vocabulary task, namely, the Aurora-4 database. The results have shown that the proposed features considerably improved the robustness in all types of noise conditions. However, the present study is essentially confined to handle noise effects on speech and has not considered reverberant conditions. The selected weights for the non-linearity were heuristic, and automatic selection of optimal weights from the evaluation data is desirable. For the future, we would like to investigate these issues and evaluate the performance of the proposed features for reverberant environments and large vocabulary tasks.

Competing interests
The authors declare that they have no competing interests.

Received: 17 January 2012 Accepted: 16 April 2014
Published: 6 May 2014

References

1. J Barker, E Vincent, N Ma, C Christensen, P Green, The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. **27**(3), 621–633 (2013)
2. J Droppo, A Acero, robustness, Environmental, in *Springer Handbook of Speech Processing*, ed. by byJ Benesty, MM Sondhi, and Y Huang (Springer New York, 2008), pp. 653–679
3. MJF Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
4. M Omologo, P Svaizer, M Matassoni, Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Commun. **25**, 75–95 (1998)
5. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
6. H Hermansky, N Morgan, RASTA processing of speech. IEEE Trans. Speech Audio Proc. **2**(4), 578–589 (1994)
7. MJF Gales, SJ Young, A fast and flexible implementation of parallel model combination. ICASSP, **1**, 133–136 (1995)
8. C Kim, Signal processing for robust speech recognition motivated by auditory processing. Ph.D. Thesis, CMU, 2010
9. GJ Brown, KJ Palomaki, A reverberation-robust automatic speech recognition system based on temporal masking. J. Acoustical Soc. Am. **123**(5), 2978 (2008)
10. O Ghitza, Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans. Speech Audio Proc. SAP-2(1), 115–132 (1994)
11. D-S Kim, S-Y Lee, RM Kil, Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Trans Speech Audio Proc. **7**, 55–69 (1999)
12. D Dimitriadis, P Maragos, A Potamianos, On the effects of filterbank design and energy computation on robust speech recognition. IEEE Trans. Audio Speech Lang. Proc. **19**, 1504–1516 (2011)
13. R Flynn, E Jones, A comparative study of auditory-based front-ends for robust speech recognition using the Aurora 2 database. Paper presented at the IET Irish signals and systems conference, Dublin, Ireland, 28–30, June 2006 pp. 28–30
14. R Schluter, I Bezrukov, H Wagner, H Ney, Gammatone features and feature combination for large vocabulary speech recognition. Paper presented in the IEEE international conference on acoustics, speech, and signal processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007 pp. 649–652
15. Y Shao, Z Jin, DL Wang, S Srinivasan, An auditory-based feature for robust speech recognition. Paper presented at the IEEE international conference on acoustics, speech, and signal processing (ICASSP), Taipei, Taiwan, 19–24 April 2009 pp. 4625–4628
16. R Drullman, J Festen, R Plomp, Effect of reducing slow temporal modulations on speech reception. J. Acoustical Soc. Am. **95**, 2670–2680 (1994)
17. N Kanedera, T Arai, H Hermansky, M Pavel, On the importance of various modulation frequencies for speech recognition. Paper presented at the Eurospeech, Rhodes Greece, 22–25 Sept 1997 pp. 1079–1082
18. TH Falk, WY Chan, Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Lang. Process. **18**(1), 90–100 (2010)
19. HK Maganti, M Matassoni, An auditory based modulation spectral feature for reverberant speech recognition. Paper presented at the 13th annual conference of the International Speech Communication Association (Interspeech), Makuhari, Japan, 26–30 Sept 2010 pp. 570–573
20. L Deng, H Sheikhzadeh, Use of temporal codes computed from a cochlear model of speech recognition, chapter 15, in *Listening to Speech: An Auditory Perspective*, ed. by Greenberg S, W Ainsworth (Lawrence Erlbaum Mahwah, 2006), pp. 237–256
21. M Kleinschmidt, J Tchorz, B Kollmeier, Combining speech enhancement and auditory feature extraction for robust speech recognition. Speech Commun. **34**, 75–91 (2001)
22. T Dau, D Pueschel, A Kohlrausch, A quantitative model of the effective signal processing in the auditory system. J. Acoustical Soc. Am. **99**, 3615–3622 (1996)
23. X Xiong, C Eng Siong, L Haizhou, Normalization of the speech modulation spectra for robust speech recognition. IEEE Trans. Audio Speech Lang. Proc. **16**(8), 1662–1674 (2008)
24. V Mitra, H Franco, M Graciarena, A Mandal, Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP), Kyoto, Japan, 25–30 March 2012, pp. 4117–4120
25. F Valente, M Magimai-Doss, C Plahl, SV Ravuri, Hierarchical processing of the modulation spectrum for GALE Mandarin LVCSR system. Paper presented at the meeting of the International Speech Communication Association (Interspeech), Brighton, UK, 6–10 Sept 2009, pp. 2963–2966

26. Y-HB Chiu, B Raj, RM Stern, Learning-based auditory encoding for robust speech recognition. Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP), Dallas, TX, USA, 14–19 March 2010, pp. 4278–4281
27. X Zhao, Y Shao, DL Wang, CASA-based robust speaker identification. IEEE Trans. Audio Speech Lang. Proc. **20–25**, 1608–1616 (2012)
28. X Zhao, DL Wang, Analyzing noise robustness of MFCC and GFCC features in speaker identification. Paper presented at the IEEE international conference on acoustics, speech and signal processing (ICASSP), Vancouver, Canada, 26–31 May 2013, pp. 7204–7208
29. M Matassoni, HK Maganti, M Omologo, Non-linear spectro-temporal modulations for reverberant speech recognition. Paper presented at the joint workshop on hands-free speech communication and microphone arrays (HSCMA), Edinburgh, Scotland, 30 May–1 June 2011, pp. 115–120
30. M Slaney, *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,* in Apple technical report, Perception Group, 1993
31. B Glasberg, B Moore, Derivation of auditory filter shapes from notched-noise data. Hearing Res. **47**, 103–108 (1990)
32. Ellis DPW, Gammatone-like spectrograms, http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/. Accessed 6 June 2011.
33. N Parihar, J Picone, D Pearce, HG Hirsch, Performance analysis of the Aurora large vocabulary baseline system. Paper presented at the 12th European signal processing conference (EUSIPCO)n Vienna, Austria, 6–10 Sept 2004, pp. 553–556

doi:10.1186/1687-4722-2014-21
Cite this article as: Maganti and Matassoni: **Auditory processing-based features for improving speech recognition in adverse acoustic conditions.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:21.