Research Journal of Applied Sciences, Engineering and Technology 9(9): 755-759, 2015
DOI:10.19026/rjaset.9.2621
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 05, 2014 Accepted: November 3, 2014 Published: March 25, 2015

Research Article

Noise Robust Speech Parameterization using Relative Spectra and Auditory Filterbank

Youssef Zouhir and Kaïs Ouni
Signals and Mechatronic Systems, SMS, UR13ES49, National Engineering School of Carthage, ENICarthage, University of Carthage, Tunisia

Abstract: In the present study, a new feature extraction method based on relative spectra and gammachirp auditory filterbank is proposed for robust noisy speech recognition. The relative spectra filtering are applied to the log of the output of the gammachirp filterbank which incorporates the properties of the cochlear filter in order to remove uncorrelated additive noise components. The performances of this method have been evaluated on the isolated speech word corrupted by real-world noisy environments using the continuous Gausian-Mixture density Hidden Markov Model. The evaluation of the experimental results shows that the proposed method achieves best recognition rates compared to the conventional techniques like Perceptual Linear Prediction (PLP), Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC).

Keywords: Auditory filterbank, hidden Markov models, noisy speech parameterization

INTRODUCTION

In many practical applications, the performance of Automatic Speech Recognition (ASR) system is limited due to its lack of the robustness in the presence of background noises. ASR relies on speech feature vectors which contain relevant information to distinguish between different speech sounds. To increase the robustness of ASR-systems, the speech feature must be less sensitive in the presence of background noises, while retaining good of distinguished properties (Gajic and Paliwal, 2006). The most commonly used feature extraction algorithms as PLP (Perceptual Linear Prediction) (Hermansky, 1990), LPCC (Linear Prediction Cepstral Coefficients) (Atal, 1974) and MFCC (Mel-Frequency Cepstral Coefficients) (Davis and Mermelstein, 1980), are highly affected in the presence of noisy environments. There are some other algorithms aiming at improving noise robustness by combining the classic algorithms with other technique like the RASTA (Relative Spectra) filtering (Hermansky and Morgan, 1994) or CMN (Cepstral mean normalization) (Liu *et al*., 1993; Shao *et al*., 2007; Droppo and Acero, 2008).

In addition, the auditory system of human has a remarkable ability to recognize the speech signal in noisy environments. This ability has inspired the development of many feature extraction algorithms which take into account certain knowledge on human speech perception (Gajic and Paliwal, 2006). The developed algorithms usually use the gammatone filter as the auditory filter modelling in order to simulate the cochlear filtering (Wang and Brown, 2006; Meddis *et al*., 2010). A new auditory filter known as gammachirp filter is developed by Irino and Patterson (1997, 2006). This filter with an asymmetric amplitude spectrum represents a good approximation to the asymmetry and level dependent characteristics of the cochlea filtering (Meddis *et al*., 2010).

A robust feature extractor for noisy speech recognition is presented in this study. The proposed method is based on relative spectra and gammachirp filterbank. The relative spectra is band-pass time-filtering applied to the log of the output spectral representation of the gammachirp filterbank in order to reduce linear channel distortions which appear as additive components in the logarithmic spectral domain. The used gammachirp filterbank is a filterbank of 34 gammachirp filters covering the frequency range (50 and 8000 Hz) (Zouhir and Ouni, 2013, 2014). The gammachirp filter is used as a model of auditory filter to provide a spectrum reflecting the cochlea spectral behavior (Irino and Patterson, 1997, 2006; Patterson *et al*., 2003).

The HTK (Hidden Markov Model Toolkit) recognizer (Young *et al*., 2009) is employed for isolated-word speech recognition with whole word HMM-GM (HMM with four Gaussian Mixture density) models. Each isolated-word is modeled by a five-state HMM with four mixtures per state.

The isolated speech words extracted from the TIMIT (Garofolo *et al*., 1990) database and corrupted by real-world noisy environments are used for the

Corresponding Author: Youssef Zouhir, Signals and Mechatronic Systems, SMS, UR13ES49, National Engineering School of Carthage, ENICarthage, University of Carthage, Tunisia

performance evaluation of proposed feature extractor. conventional techniques are used: PLP, LPCC and MFCC. Experimental results in the presence of ambient background noises show that the proposed feature extractor outperforms all the classical techniques mentioned above.

Classical feature for speech recognition: The classical feature extractors MFCC, PLP and LPCC are similar in several stages. As shown in Fig. 1, these similar stages are linked by the broken arrows. The procedure to obtain the coefficients of each technique is briefly described here.

The MFCC coefficients: A Discrete Fourier Transform (DFT) is computed for each frame of windowed speech to obtain a short-term power spectrum. Then, the power spectrum of the speech signal is weighted by the magnitude frequency response of a Mel-scale filterbank which uses triangular shaped windows. Logarithmic compression of the Mel-filterbank output is applied. The cepstrum coefficients are then obtained by a Discrete Cosine Transform (DCT) (Davis and Mermelstein, 1980).

The PLP coefficients: Similar to the MFCC procedure, the discrete Fourier transform power spectrum is firstly calculated. Then, the auditory-based warping of the frequency axis is employed to weight the obtained spectrum. The window shape used in PLP analysis is designed to obtain a simulation of the critical-band masking curves. After pre-emphasize the filtered power spectrum by an equal-loudness curve, a cubic root

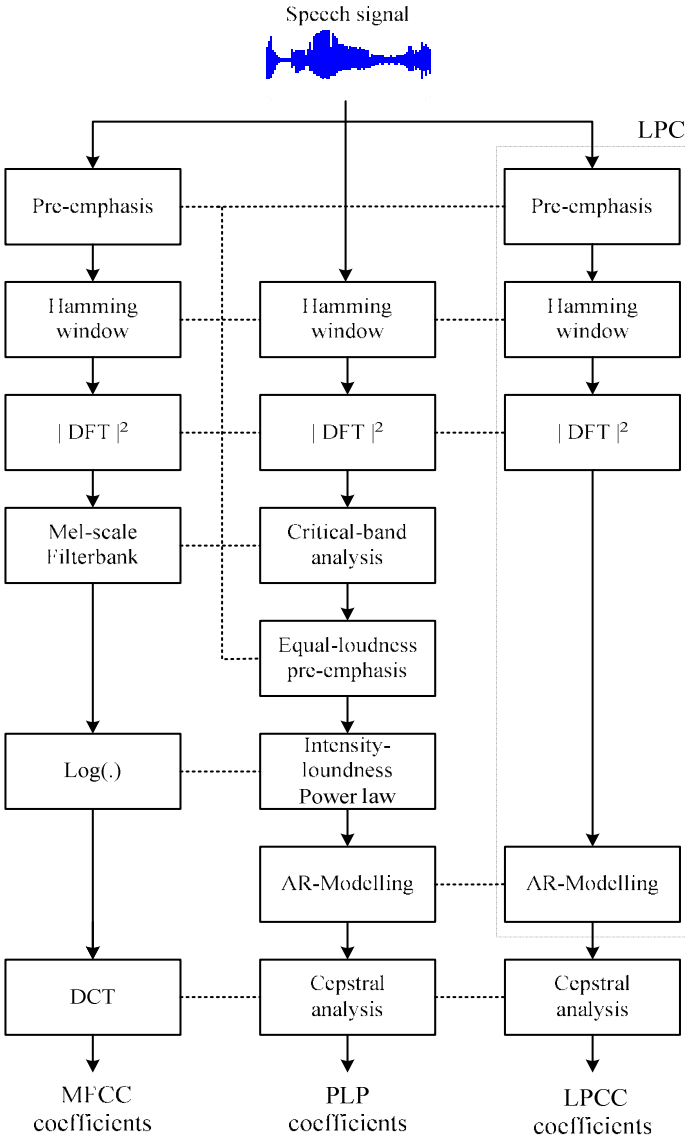

Fig. 1: Flowcharts for MFCC, PLP and LPCC feature extraction techniques

compression of critical-band energies is applied whereas for MFCC logarithmic compression is used. The result spectrum is converted into LP coefficients using Auto-Regression (AR) modelling. The PLP coefficients are computed by applying a cepstral transformation to the LP coefficients (Hermansky, 1990).

The LPCC coefficients: After the extraction of the LPC coefficients from each speech signal frame using autocorrelation method, 12 cepstral coefficients which correspond to LPCC coefficients are computed from the obtained eight coefficients using cepstral transform (Atal, 1974).

METHODOLOGY

Proposed feature extractor: The proposed feature extraction method is based on relative spectra and gammachirp auditory filterbank for robust noisy speech recognition. An illustrative block diagram of the various steps of the proposed feature extractor is shown in Fig. 2.

In the first step, the speech signal is framed (length of analysis frame is 25 msec with a frame shift of 10 msec) and windowed using a Hamming window. Then we apply the square of Discrete Fourier Transform (DFT) for each window segment to obtain the power spectrum. The second step is the Relative spectra-Gammachirp filterbank. In this step, the power spectrum is analyzed using a 34-channel gammachirp filterbank. The latter is characterized by a centre frequencies covering the frequency range of 50-8000 Hz (sampling frequency = 16 kHz) according to the ERB-rate scale. The used filterbank is developed to provide a realistic auditory filterbank for auditory perception models. The complex analytic form of the gammachirp filter is (Irino and Patterson, 1997):

$$g_c(t) = at^{n-1}\exp\left(-2\pi bERB(f_r)t\right)\exp\left(j2\pi f_r t + jc\ln(t) + j\varphi\right) \quad (1)$$

where, a, φ, c and f_r are respectively, the amplitude, the phase, the chirp factor and the asymptotic frequency. b and n are parameters defining the gamma distribution envelope. The time t>0, "ln" is the natural logarithmic operator and ERB (f_r) is the equivalent rectangular bandwidth of the auditory filter at f_r. The *ERB* value at frequency *f* in Hz is defined by (Glasberg and Moore, 1990; Moore, 2012; Wang and Brown, 2006):

$$ERB(f) = 24.7 + 0.108f \quad (2)$$

The equivalent rectangular bandwidth rate (ERBrate (f)) at frequency *f* is given by (Glasberg and Moore, 1990; Moore, 2012; Wang and Brown, 2006):

$$ERBrate(f) = 21.4\log_{10}(0.00437f + 1) \quad (3)$$

The Fourier magnitude spectrum $|H_{Gc}(f)|$ of the gammachirp filter (Irino and Patterson, 2006; Patterson *et al*., 2003) is given by:

$$\left|H_{Gc}(f)\right| = \frac{a\left|\Gamma(n+jc)\right|\cdot\exp(c\cdot\theta)}{\left|2\pi\sqrt{\left(b\,\mathrm{ERB}(f_r)\right)^2 + (f-f_r)^2}\right|^n} \quad (4)$$

where, $\theta = \arctan\left((f-f_r)/(b\,\mathrm{ERB}(f_r))\right)$ and $\Gamma(n+jc)$ is the gamma distribution function.

Afterward, a relative spectral band-pass filtering is applied to the filterbank outputs in the logarithmic domain in order to remove uncorrelated additive noise components. The transfer function of this filter is defined by (Hermansky and Morgan, 1994):

$$\left|H(z)\right| = 0.1z^4\,\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad (5)$$

In the third step, the inverse logarithm of the relative logarithm spectrum is calculated, yielding a

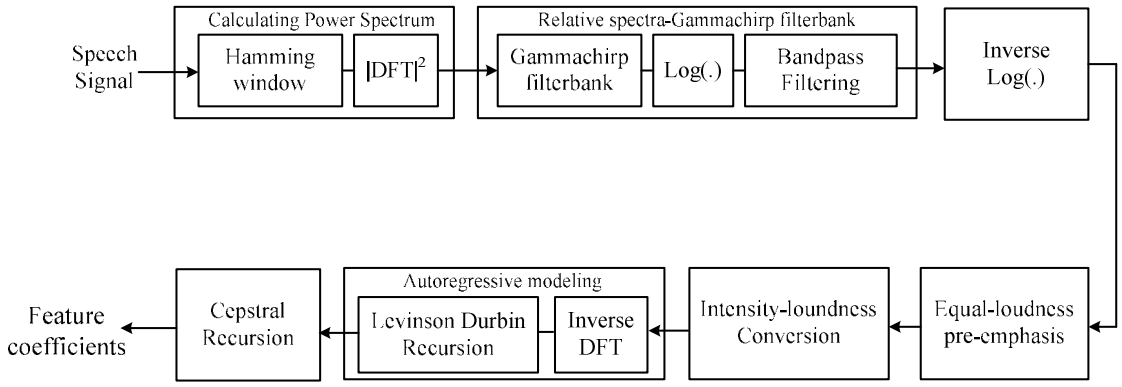

Fig. 2: Diagram of the proposed feature extraction method

relative auditory spectrum (Hermansky and Morgan, 1994). The latter is weighted, in the fourth step, by an equal-loudness pre-emphasis, to compensate for the non equal sensitivity of human hearing system across frequency. Then, the cubic-root compression step which aims at simulating the non-linear relation between sound intensity and its perceived loudness is applied to the pre-emphasis spectrum. The sixth step consists to obtain the autoregressive coefficients of the all-pole model using Inverse-DFT and the Levinson-Durbin Recursion, which is designed to estimate of the auditory-like spectrum of speech (Hermansky, 1990; Zouhir and Ouni, 2014). In the seventh step, the proposed feature are obtained by performs a cepstral transformation.

EXPERIMENTAL RESULTS

This section presents the evaluation results of the experiments that were performed with the various techniques, using an isolated-word speech recognizer, in the presence of various types of the ambient background noises. A total of 13227 isolated-words used in these experiments were manually extracted from the TIMIT database (contains speech signals of 630 speakers from eight English dialect regions and the sampling frequency of these signals is 16 kHz) (Garofolo *et al*., 1990). The training set consisted of 9702 isolated-words and the testing set contains 3525 isolated-words. The all isolated-words used in the testing phase were corrupted by different background noise (Passing-car, Shopping-mall, Rain, Sea waves noise) for various SNR ranging from -3 to 9 dB. These noises were taken from PacDV (PacDV Sound Effects, 2014).

The HTK.3.4.1 toolkit (Young *et al*., 2009) was used to implement an isolated-word based HMM recognizer. Each isolated-word was represented by a simple left-to-right HMM (HMM-GM) of five states with four diagonal Gaussians per state.

The used parameters of gammachirp filter are $n = 4$; $a = 1$; $b = 1.019$; $c = 2$ and $\varphi = 0$.

The recognition performance of the proposed PLPrGc (Perceptual Linear Predictive relative spectra-

Table 1: Comparison of recognition rates of the proposed and the other classical features with passing-car noise at various SNR's for HMM-4-GM

		Feature			
	SNR level (dB)	PLPrGc	PLP	LPCC	MFCC
Passing-car noise	-3	46.24	33.84	25.93	30.72
	0	62.55	48.45	38.04	45.16
	3	76.23	62.95	50.67	61.48
	6	84.43	77.84	68.91	77.56
	9	90.33	87.40	81.19	87.72

Table 2: Comparison of recognition rates of the proposed and the other classical features with shopping-mall noise at various SNR's for HMM-4-GM

		Feature			
	SNR level (dB)	PLPrGc	PLP	LPCC	MFCC
Shopping-mall noise	-3	25.28	17.13	15.86	17.67
	0	45.56	33.11	25.84	32.60
	3	63.66	51.97	40.51	51.46
	6	77.19	71.49	56.45	70.35
	9	84.57	82.95	71.86	82.18

Table 3: Comparison of recognition rates of the proposed and the other classical features with sea waves noise at various SNR's for HMM-4-GM

		Feature			
	SNR level (dB)	PLPrGc	PLP	LPCC	MFCC
Sea waves noise	-3	18.67	10.27	10.72	10.55
	0	32.60	21.56	19.21	19.86
	3	51.52	38.18	31.80	36.65
	6	67.57	56.31	46.95	53.99
	9	79.26	72.57	61.70	70.98

Table 4: Comparison of recognition rates of the proposed and the other classical features with rain noise at various SNR's for HMM-4-GM

		Feature			
	SNR level (dB)	PLPrGc	PLP	LPCC	MFCC
Rain noise	-3	30.21	22.87	14.41	19.26
	0	44.26	31.86	21.36	28.51
	3	59.97	42.67	31.09	41.16
	6	72.88	55.09	42.50	54.01
	9	81.19	66.41	53.96	66.07

Gammachirp) feature has been compared to that of the baseline PLP, LPCC and MFCC feature. The feature vector of each technique consisted of 39 coefficients including 12 static coefficients of feature techniques were added to energy (E), differential coefficients first order (Δ) and second order (A).

Table 1 to 4 represent the recognition rates obtained using the proposed PLPrGc feature and three kinds of PLP, LPCC and MFCC feature for various types of ambient background noise at SNR equal to -3, 0, 3, 6 and 9 dB, respectively. These tables show the effectiveness of proposed feature compared to the other baseline features for the four ambient background noises. It can be observed that PLPrGc feature gives better results of recognition rate for all SNR levels, particularly for low values of SNR values. For example, in the case of passing-car noise at 0 dB SNR, the recognition rate of the PLPrGc is higher than that of the PLP, MFCC and LPCC by 14.1, 17.39 and 24.51, respectively.

CONCLUSION

In this study, we have presented a robust feature extractor based on relative spectra and gammachirp filterbank for noisy speech recognition. Speech recognition results were reported on the isolated speech words TIMIT corrupted by real-world noisy environments and performances were compared with the PLP, LPCC and MFCC. Four different background noises with various SNR ranging from -3 to 9 dB were used. Experimental results show that proposed feature extractor outperformed all the other classical feature extractors for all SNR levels.

REFERENCES

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am., 55(6): 1304-12.

Davis, S.B. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE T. Acoust. Speech, 28(4): 357-366.

Droppo, J. and A. Acero, 2008. Robustness, Environmental, in Springer Handbook of Speech Processing. In: Benesty, J., M.M. Sondhi and Y. Huang (Eds.), Springer, New York, pp: 653-679.

Gajic, B. and K.K. Paliwal, 2006. Robust speech recognition in noisy environments based on sub and spectral centroid histograms. IEEE T. Audio Speech, 14(2): 600-608.

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren, 1990. DARPA, TIMIT acoustic-phonetic continuous speech Corpus. National Institute of Standards and Technology, Technical Report No. NISTIR 4930, Gaithersburg, MD, Speech Data Publish on CD-ROM, NIST Speech Disc 1-1.1.

Glasberg, B.R. and B.C.J. Moore, 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Res., 47(1): 103-138.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87(4): 1738-1752.

Hermansky, H. and N. Morgan, 1994. RASTA processing of speech. IEEE T. Speech Audi. P., 2(4): 578-589.

Irino, T. and R.D. Patterson, 1997. A time-domain, level-dependent auditory filter: The gammachirp. J. Acoust. Soc. Am., 101(1): 412-419.

Irino, T. and R.D. Patterson, 2006. A dynamic compressive gammachirp auditory filterbank. IEEE T. Audio Speech, 14(6): 2222-2232.

Liu, F., R. Stern, X. Huang and A. Acero, 1993. Effcient cepstral normalization for robust speech recognition. Proceeding of ARPA Speech and Natural Language Workshop, pp: 69-74.

Meddis, R., E.A. Lopez-Poveda, R.R. Fay and A.N. Popper, 2010. Computational Models of the Auditory System. Springer Handbook of Auditory Research. Springer, New York, 35: 350.

Moore, B.C.J., 2012. An Introduction to the Psychology of Hearing. 6th Edn., Brill, Leiden.

PacDV Sound Effects, 2014. Retrieved form: http://www.pacdv.com/sounds/.

Patterson, R.D., M. Unoki and T. Irino, 2003. Extending the domain of center frequencies for the compressive gammachirp auditory filter. J. Acoust. Soc. Am., 114(5): 1529-1542.

Shao, Y., S. Srinivasan and D. Wang, 2007. Incorporating auditory feature uncertainties in robust speaker identification. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2007), 4: IV-277-IV-280.

Wang, D.L. and G.J. Brown, 2006. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. Publisher by IEEE Press / Wiley-Interscience.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore *et al*., 2009. The HTK Book Version 3.4.1. Cambridge University Engineering Department, Cambridge, U.K.

Zouhir, Y. and K. Ouni, 2013. Speech signals parameterization based on auditory filter modelling. In: Drugman, T. and T. Dutoit (Eds.), Advances in Nonlinear Speech Processing. LNAI 7911, NOLISP 2013, Springer, Mons, Belgium, Berlin, Heidelberg, pp: 60-66.

Zouhir, Y. and K. Ouni, 2014. A bio-inspired feature extraction for robust speech recognition. SpringerPlus, 3(1): 651.