International Journal of Innovative Computing, Information and Control
Volume **11**, Number **3**, June **2015**

 ISSN 1349-4198
pp. 955–963

ROBUST FEATURES FOR NOISY SPEECH RECOGNITION USING JITTER AND SHIMMER

Hajer Rahali, Zied Hajaiej and Noureddine Ellouze

Laboratory of Signal, Image and Information Technologies
National Engineering School of Tunis
BP 37, Le Belvédère, 1002 Tunis, Tunisie
{hajer.rahali; zied.hajaiej; N.ellouze}@enit.rnu.tn

Received March 2014; revised July 2014

Abstract. *In this paper we introduce a robust feature extractor, dubbed as Modified Function Linear Prediction (MODFLP), based on gammachirp filterbank and psychoacoustic model. The goal of this work is to improve the robustness of speech recognition systems in additive noise and real-time reverberant environments. In speech recognition systems, Perception Linear Prediction (PLP) and Mel Frequency Cepstral Coefficient (MFCC) are the two main techniques used. In our work the effectiveness of proposed changes to PLP were tested and compared against the PLP and MFCC features. The above-mentioned techniques were tested with impulsive signals under various noisy conditions within TIMIT databases.*
Keywords: Auditory filter, Impulsive noise, PLP, RASTA filter, ARMA filter

1. **Introduction.** Speech parameterization is an important step in modern automatic speech recognition systems (ASR). The speech parameterization block is used to extract from the speech waveform the relevant information for discriminating between different speech sounds. The information is presented as a sequence of parameter vectors. In this paper, two acoustic features are found: PLP and MFCC. Generally, these methods are based on three similar processing blocks: firstly, basic short-time Fourier analysis which is the same for both methods; secondly, auditory based filterbank, and thirdly, cepstral coefficients computation. In addition, temporal filtering is also useful to enhance speech features, such as RASTA filtering [16]. The principle of RASTA method comes from the human auditory perception which indicates the relative insensitivity of human hearing to slowly and quickly varying auditory stimuli [16]. Thus, the RASTA band-pass filter is designed with an IIR filter with a sharp spectral zero at the zero frequency in the modulation frequency domain. PLP are used extensively in ASR. There are many similarities between these methods. The difference however lies in the shape of the filterbank. In this paper we present the proposed modifications of PLP method, and it will be shown that the performance of PLP and MFCC, are also compared to MODFLP which integrate a new model. In the current paper, prosodic information is first added to a spectral system in order to improve their performance. Such prosodic characteristics include parameters related to the fundamental frequency such as the jitter and shimmer. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on HMM\GMM. We propose a study of the performance of parameterization techniques using jitter and shimmer features in the presence of different impulsive noises. The sounds are added to the word with different signal-to-noise SNRs (20 dB, 15 dB, 10 dB, 5 dB, 0 dB and –5 dB). Note that the robustness is shown in terms of correct recognition rate (CRR) accuracy. The evaluation is done on the TIMIT Database. This paper is organized

as follows. In the next section we describe the proposed modifications of PLP. An experimental study performed to compare the performance of the different parameterization methods in various acoustic environments is described in Section 3. Finally, the major conclusions are summarized in Section 4.

2. New Proposed Technique.

2.1. MFCC and PLP features. Figure 1 shows a block diagram for extracting PLP and MFCC features. Physiological studies have shown that human auditory system does not follow a linear scale. Using Mel frequency scaling well approximates the frequency response of human auditory systems and it can be used to capture the phonetically important characteristics of speech. One approach for simulating the subjective spectrum is to use a filterbank, spaced uniformly on the Mel scale.

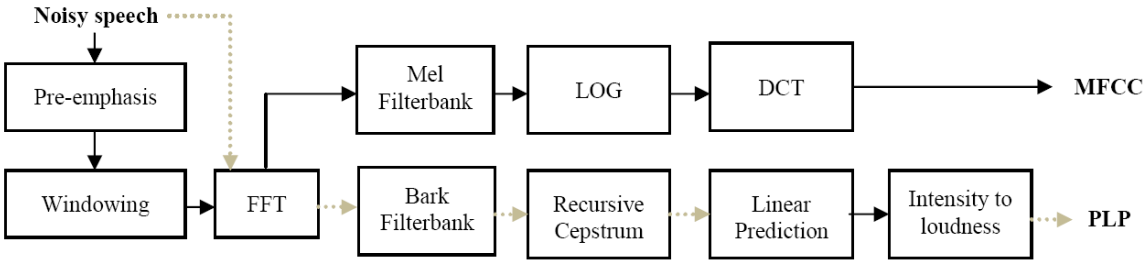

FIGURE 1. Block diagram for extracting MFCC and PLP

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Typically the Hamming window is used. The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. After that the scale of frequency is converted from linear to Mel scale. Then logarithm is taken from the results. In final step, the log Mel spectrum is converted back to time domain. The result is called the Mel frequency cepstrum coefficients (MFCC).

In PLP analysis, a Fourier transform is first applied to compute the short-term power spectrum, and the perceptual properties are applied while the signal is represented in this filterbank form. The spectrum is transformed to a Bark scale, and this spectrum is pre-emphasized by a function that approximates the sensitivity of human hearing at different frequencies. The output is compressed to approximate the non-linear relationship between the intensity of a sound and its perceived loudness.

2.2. Proposed modifications of PLP. In this paper, we present some modifications of the standard PLP feature extraction method. The proposed modifications are presented in the following section. A schematic diagram of the proposed technique is shown in Figure 2. In this proposed algorithm MODFLP an application of pre-emphasis is applied to the speech signal before the short term spectral analysis. In the second step, the digitized noisy speech is segmented into overlapping frames, each of length 20 ms with 10 ms overlap, in speech processing a Hamming window are mostly used. Next, the FFT is taken of these segments. Afterwards, the segmented signal is filtered using the non linear model of the external and middle ear (Figure 3). The next processing step applies a filterbanks. Many different types of filterbanks exist but for MODFLP features the gammachirp filterbank is used. In this study, our objective is to introduce new speech features that are more robust in noisy environments. We propose a robust speech feature which is based on the method with gammachirp filterbank. The output signal of the

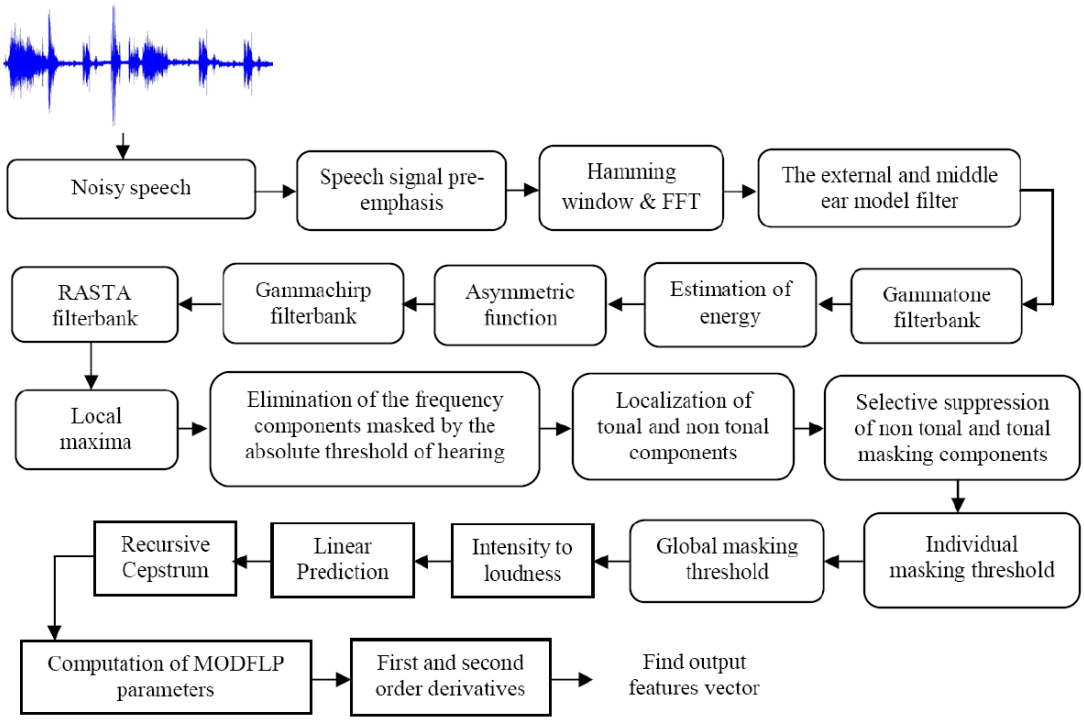

FIGURE 2. The structure of MODFLP features extraction

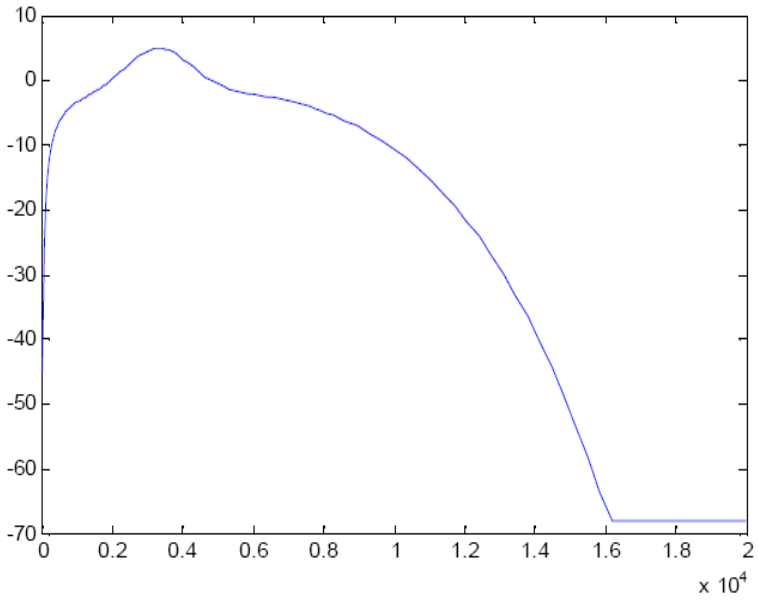

FIGURE 3. Model of the external and middle ear

outer and middle ear model filter is applied to a gammatone filterbank. On each sub-band we calculate the sound pressure level Ps (dB) in order to have the corresponding sub-band chirp term C. Those values of chirp term C corresponding to each sub-bands of the gammatone filterbank lead to the corresponding gammachirp filterbank (Figure 4). The proposed auditory use filter that is smoother and broader than the Bark filterbank. The main differences between the proposed filterbank and the typical one used for PLP estimation are the type of filters used and their corresponding bandwidth. In this paper, we experiment with one parameter to create a family of gammachirp filterbanks: the

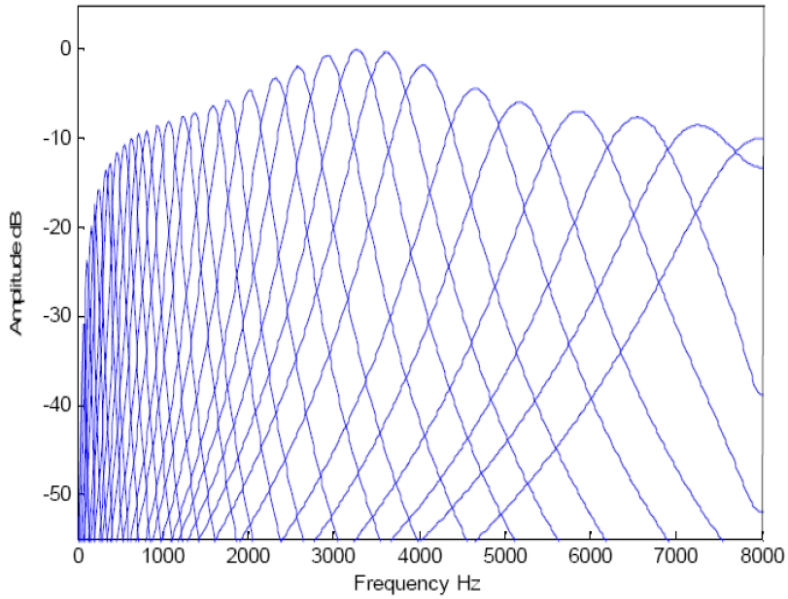

FIGURE 4. Output of gammachirp filterbank

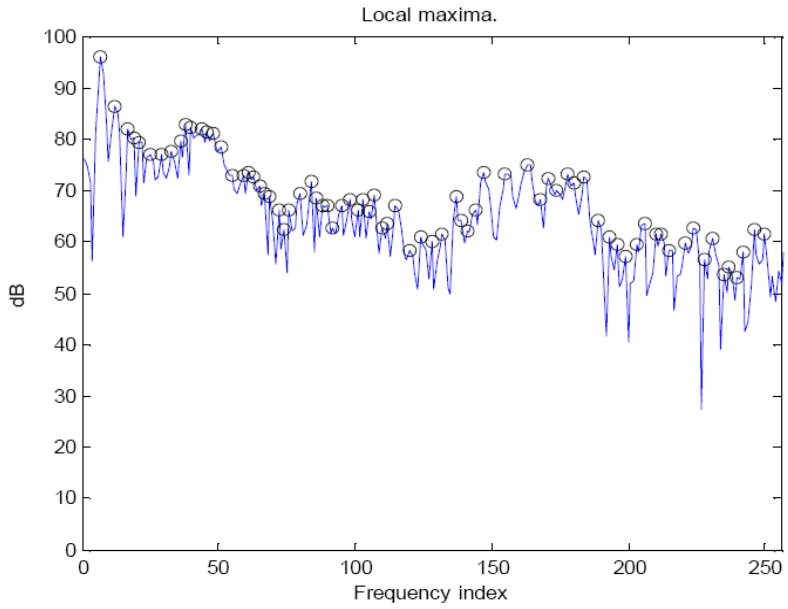

FIGURE 5. Local maxima

number of filters. The free parameter in gammachirp filterbank as noted above is the number of filters. By increasing the number of filters they become narrow but with a small number of filters the loss of information is introduced. The RASTA filter removes variations in the signal that are outside the rate of change of speech by filtering the log-spectrum at each frequency band. Both very slow and very fast changes in sound are ignored by the human ear, so RASTA processing attempts to filter these components out. The filter also helps to eliminate noise due to channel variation in the data. That is why we use the temporal filtering technique to process the cepstral coefficients, and then we get the features coefficients which we need. In the next stage, we calculate tonal and non tonal components. This step begins with the determination of the local maxima, followed by the extraction of the tonal components (speech) and non tonal components (impulsive

noise), in a bandwidth of a critical band. If frequency exceeds neighboring components within a bark distance by at least 6 dB then it will be treated as "speech", otherwise it will be considered as "noise". The local maximum of word "greasy" is shown in Figure 5.

The selective suppression of tonal and non tonal components of masking is a procedure used to reduce the number of maskers taken into account for the calculation of the global masking threshold. The tonal and non tonal components remaining are those which are above the hearing absolute threshold. Individual masking threshold takes into account the masking threshold for each remaining component.

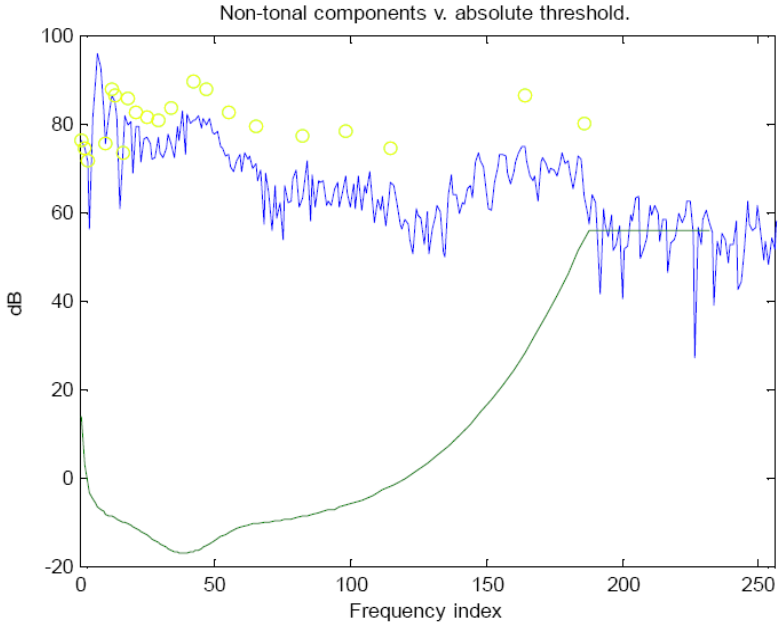

FIGURE 6. Non tonal components with absolute threshold

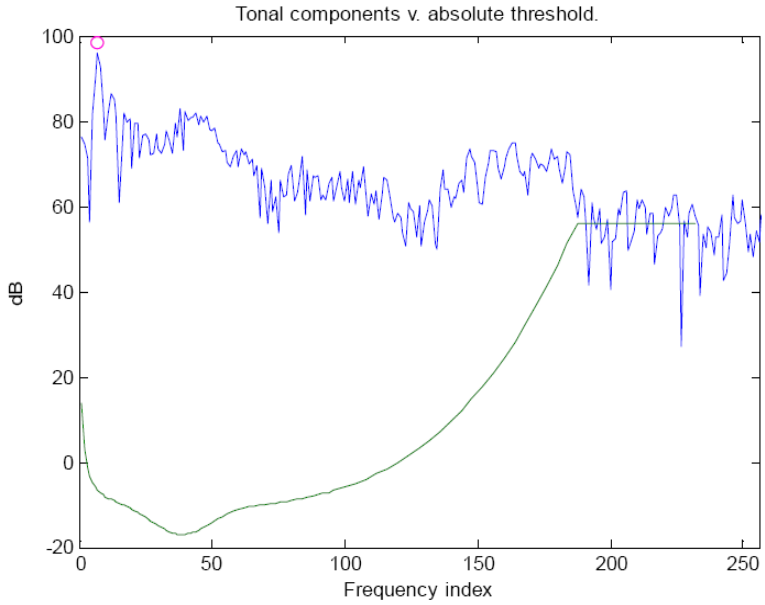

FIGURE 7. Tonal components v. absolute threshold

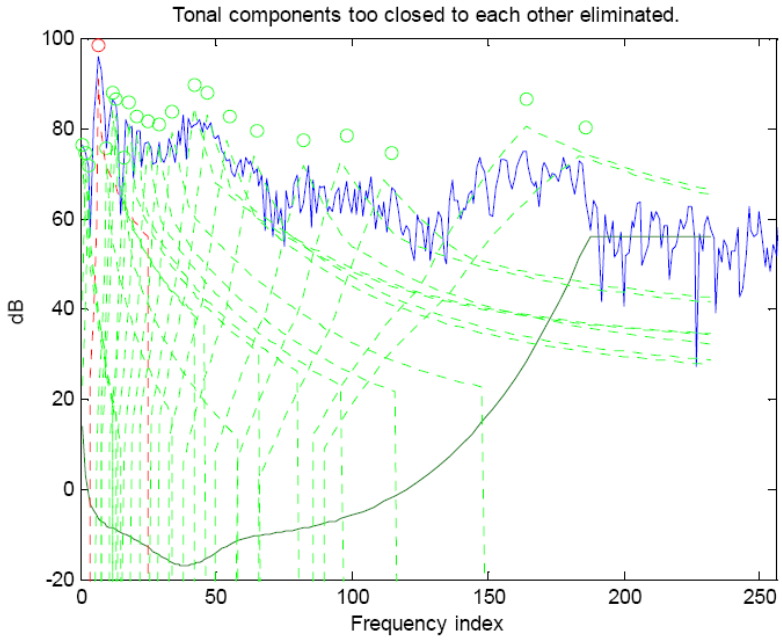

FIGURE 8. Masking threshold

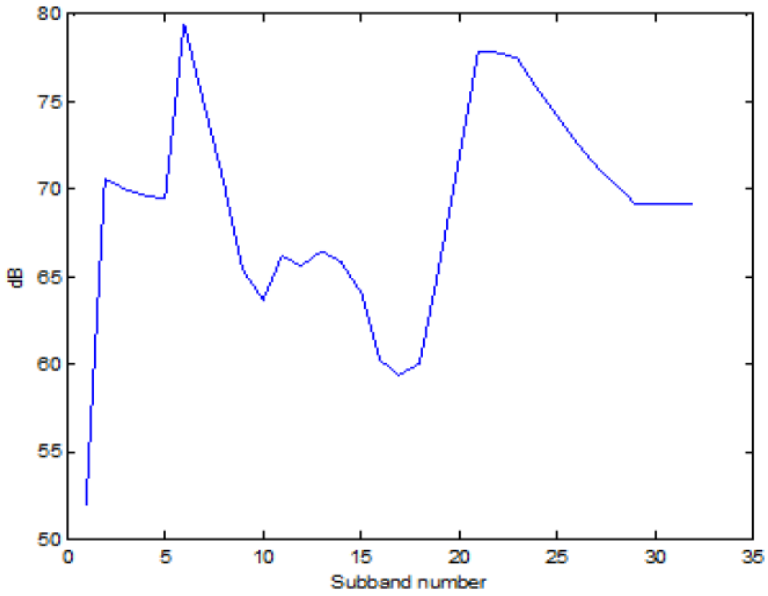

FIGURE 9. Global masking threshold

Then, the PLP are applied to these channels to extract features characteristics. PLP as previously stated has the advantage that they can represent sound signals in an efficient way because of the frequency warping property. In this way, the advantages of this technique are combined in the proposed method. For the final acoustic modeling we extended the modified PLP-cepstral representation with derived delta and delta-delta features. The following features were extracted: 39 PLP and 39 MODFLP. The energy of the frame, first (Δ) and second temporal derivatives ($\Delta\Delta$) extracted from each enumerated parameter. At the end, we have 9 distinct features vectors that can be categorized into three categories according to its length. Firstly, feature vector has length 13: (12 MODFLP and Energy). Secondly, feature vector has length 26: (12 MODFLP, Energy,

and 13 Δ). Thirdly, feature vector has length 39: (12 MODFLP, Energy, 13 Δ, and 13 $\Delta\Delta$). We note that the addition of delta-cepstral features to the static 13 dimensional MODFLP features strongly improves speech recognition accuracy, and a further (smaller) improvement is provided by the addition of double delta-cepstral features. The feature vector constructed on the basis of MODFLP is applied in the statistical classifier. This classifier is based on Gaussian mixture model (GMM) and Hidden Markov Model (HMM).

In the next section, we investigate the robustness and compare the performance of the proposed features to that PLP and MFCC with the different prosodic parameters by artificially introducing different levels of impulsive noise to the speech signal and then computing their correct recognition rate.

3. **Experimental and Results.**

3.1. **TIMIT database.** The Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems [17]. TIMIT contains a total of 6300 sentences, consisting of 10 sentences spoken by each of 630 speakers (438 male and 192 female) from 8 major dialect regions of the United States. Each speaker recorded ten speech utterances with duration of about 3 seconds. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions, as well as speech waveform data for each spoken sentence. For the experiments, the data were split into a training set, a validation set and a test set. The training set contains 3696 sentences (462 speakers), the validation set contains 400 sentences (50 speakers) and the test set contains 192 sentences (24 speakers).

3.2. **Experimental setup.** To evaluate the suggested techniques, we carried out a comparative study with different baseline parameterization techniques of PLP and MFCC implemented in HTK. The features extracted from clean and noisy database have been converted to HTK format using "VoiceBox" toolbox [4] for MATLAB. In our experiment, there were 10 HMM models trained using the selected feature MODFLP, PLP and MFCC. Each model had 5 by 5 states left to right. The features corresponding to each state occupation in an HMM are modeled by a mixture of 12 Gaussians. In all the experiments, 39 vectors are used as the baseline feature vector. Jitter and shimmer are added to the baseline feature set both individually and in combination.

3.3. **Results and discussion.** The performance of the suggested parameterization methods is tested on the TIMIT databases using HTK. We use the percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the feature extractors considered in this paper. Comparison of phoneme recognition rates is shown in Tables 1, 2, 3 and 4.

In the first test, we vary the number value N of filter, and we observe that with the validation with a test database, the MODFLP coefficients showed the best results in terms of recognition rates calculated with a value of about 90% for N equal to 44. Comparing

TABLE 1. Word accuracy (%) using different parameterization techniques

N	**12**	**18**	**24**	**34**	**39**	**44**
MFCC	52.56	57.34	60.59	65.64	67.12	71.67
PLP	50.21	55.13	59.54	63.98	67.65	70.75
MODFLP	64.08	86.23	87.09	88.65	90.72	90.11

TABLE 2. Word accuracy (%) of PLP using Jitter and Shimmer

Features	Explosions							Door slams						
SNR (dB)	∞	20	15	10	5	0	–5	∞	20	15	10	5	0	–5
PLP (Baseline)	85.45	82.25	78.29	78.44	66.38	50.65	33.34	84.34	80.56	78.98	77.76	70.12	55.32	42.76
PLP+Jitter	88.05	84.85	80.89	80.04	68.98	53.25	35.94	86.94	83.16	81.58	80.36	72.72	57.92	45.36
PLP+Shimmer	88.45	85.25	81.29	81.44	69.38	53.65	36.34	87.34	83.56	81.98	80.76	73.12	58.32	45.76
PLP+Jitter+Shimmer	89.55	87.47	82.39	82.54	70.48	54.75	37.44	88.44	84.66	82.76	81.86	74.22	59.42	46.86

TABLE 3. Word accuracy (%) of MODFLP using Jitter and Shimmer

Features	Explosions							Door slams						
SNR (dB)	∞	20	15	10	5	0	–5	∞	20	15	10	5	0	–5
MODFLP (Baseline)	92.43	90.17	88.20	85.74	77.21	71.54	50.07	90.35	89.96	88.98	87.70	78.99	70.22	68.43
MODFLP+Jitter	95.05	92.77	90.80	88.34	79.81	74.74	52.67	92.95	92.56	91.58	90.30	81.59	72.82	71.03
MODFLP+Shimmer	95.43	93.17	91.20	88.74	80.21	74.54	53.07	93.35	92.96	91.98	91.70	84.59	75.82	74.03
MODFLP+Jitter+Shimmer	96.53	95.28	92.30	89.84	81.31	75.64	54.17	94.45	94.06	93.08	92.80	85.69	76.92	75.13

TABLE 4. Recognition rate (%) of MODFLP using delta and delta-delta parameters

Features	Explosions							Door slams						
SNR (dB)	∞	20	15	10	5	0	–5	∞	20	15	10	5	0	–5
MODFLP (13)	81	80.34	79.98	75.74	71.65	65.34	43.87	82.54	80.67	80.54	79.70	65.87	54.02	34.74
MODFLP+Δ (26)	83.23	83.78	82.47	80.76	79.65	70.45	59.54	90.95	89.56	88.98	85.30	80.45	72.45	66.87
MODFLP+Δ+ΔΔ (39)	90.64	91.17	91.20	90.09	89.56	85.67	77.87	94.21	92.87	91.98	90.10	81.90	78.10	72.76

the recognition rates of system based coefficients PLP with those achieved with MODFLP coefficients, we note that the performance of the latter is better for all numbers of filter.

Tables 2 and 3 present the performance of two voice features in presence of various levels of additive noise. We note that the MODFLP features exhibit the best CRR. Also, it is observable that the performance of the two features decreases when the SNR decreases too, that is, when the speech signal becoming more noisy. Jitter and shimmer are added to the baseline feature set both individually and in combination. The absolute accuracy increase is 2.6% and 3.0% after appending jitter and shimmer individually, while there is 4.1% increase when used together. As we can see in these tables, the identification rate increases with speech quality, for higher SNR we have higher identification rate, the MODFLP based parameters are slightly more efficiencies than standard PLP for noisy speech (95.28% vs 87.47% for 20 dB of SNR with jitter and shimmer) but the results change the noise of another. From the above Table 4, it can be seen that the recognition rates are above 90%; this recognition rates are due to the consideration of using 39 MODFLP features.

4. **Conclusion.** The proposed features called MODFLP have been shown to be more robust than PLP and MFCC in noise environments for different SNRs values. Adding jitter and shimmer to baseline spectral and energy features in an HMM\GMM based classification model resulted in increased word accuracy across all experimental conditions. The results gotten after application of this features show that this method gives acceptable and better results by comparison at those gotten by other methods of parameterization.

REFERENCES

[1] J. O. Smith III and J. S. Abel, Bark and ERB bilinear transforms, *IEEE Trans. Speech and Audio Processing*, vol.7, no.6, 1999.

[2] H. G. Musmann, Genesis of the MP3 audio coding standard, *IEEE Trans. Consumer Electronics*, vol.52, pp.1043-1049, 2006.

[3] H. G. Hirsch and D. Pearce, The AURORA experiment framework for the performance evaluations of speech recognition systems under noisy condition, *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, France, 2000.
[4] M. Brookes, *VOICEBOX: Speech Processing Toolbox for MATLAB*, Software, 2011.
[5] E. Ambikairajah, J. Epps and L. Lin, Wideband speech and audio coding using gammatone filter banks, *Proc. of ICASSP*, Salt Lake City, USA, vol.2, pp.773-776, 2001.
[6] M. N. Viera, F. R. McInnes and M. A. Jack, Robust F0 and Jitter estimation in the pathological voices, *Proc. of ICSLP*, Philadelphia, pp.745-748, 1996.
[7] L. Salhi, Design and implementation of the cochlear filter model based on a wavelet transform as part of speech signals analysis, *Research Journal of Applied Sciences*, vol.2, no.4, pp.512-521, 2007.
[8] F. Weber, L. Manganaro, B. Peskin and E. Shriberg, Using prosodic and lexical information for speaker identification, *Proc. of ICASSP*, Orlando, FL, USA, 2002.
[9] J. W. Pitton, K. Wang and B. H. Juang, Time-frequency analysis and auditory modeling for automatic recognition of speech, *Proc. of IEEE*, vol.84, pp.1199-1214, 1996.
[10] E. Loweimi and S. M. Ahadi, A new group delay-based feature for robust speech recognition, *Proc. of IEEE Int. Conf. on Multimedia & Expo*, Barcelona, pp.1-5, 2011.
[11] T. Irino, E. Okamoto, R. Nisimura, H. Kawahara and R. D. Patterson, A gammachirp auditory filterbank for reliable estimation of vocal tract length from both voiced and whispered speech, *The 4th Annual Conference of the British Society of Audiology*, Keele, UK, 2013.
[12] 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, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp.4574-4577, 2010.
[13] D. P. W. Ellis and B. S. Lee, Noise robust pitch tracking by subband autocorrelation classification, *The 13th Annual Conference of the International Speech Communication Association*, 2012.
[14] D. Povey and L. Burget, The subspace gaussian mixture model – A structured model for speech recognition, *Computer Speech & Language*, vol.25, no.2, pp.404-439, 2011.
[15] C.-P. Chen, J. Bilmes and K. Kirchhoff, Low-resource noise-robust feature post-processing on aurora 2.0, *Proc. of ICSLP*, pp.2445-2448, 2002.
[16] H. Hermansky and N. Morgan, RASTA processing of speech, *IEEE Trans. Speech and Audio Processing*, vol.2, no.4, 1994.
[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena and V. Zue, TIMIT acoustic-phonetic continuous speech corpus, *Linguistic Data Consortium*, 1993.