A Front-End Technique for Automatic Noisy Speech Recognition

Hay Mar Soe Naing
University of Computer Studies, Thaton
Thaton, Myanmar
haymarsoenaing@ucsy.edu.mm

Risanuri Hidayat
Gadjah Mada University
Yogyakarta, Indonesia
risanuri@ugm.ac.id

Rudy Hartanto
Gadjah Mada University
Yogyakarta, Indonesia
rudy@ugm.ac.id

Yoshikazu Miyanaga
Hokkaido University
Sapporo, Japan
miya@ist.hokudai.ac.jp

Abstract—**The sounds in a real environment not often take place in isolation because sounds are building complex and usually happen concurrently. Auditory masking relates to the perceptual interaction between sound components. This paper proposes modeling the effect of simultaneous masking into the Mel frequency cepstral coefficient (MFCC) and effectively improve the performance of the resulting system. Moreover, the Gammatone frequency integration is presented to warp the energy spectrum which can provide gradually decaying the weights and compensate for the loss of spectral correlation. Experiments are carried out on the Aurora-2 database, and frame-level cross entropy-based deep neural network (DNN-HMM) training is used to build an acoustic model. While given models trained on multi-condition speech data, the accuracy of our proposed feature extraction method achieves up to 98.14% in case of 10dB, 94.40% in 5dB, 81.67% in 0dB and 51.5% in -5dB, respectively.**

Index Terms—**Feature Extraction, Gammatone Filterbank, Psychoacoustics, Simultaneous Masking, Speech Recognition**

I. INTRODUCTION

Speech recognition is the critical technology in the human-computer interfaces (HCI) system and translate the human auditory function [1] [2]. In parallel to computer technology development, automatic speech recognition (ASR) is invading our lives, and its applications are more and more pervasive. It is built into medical purpose, banking system, tourism and information inquiry, speech to speech translation and other service systems [3]. The recognition performance of the ASR system is unavoidably affected by the interruption of channel and unwanted background noise. Hence, to improve the performance of the system, we necessitate to remove the corrupted noise and enhance the quality of the speech signal [4]. Noise reduction techniques can be implemented in a different perspective to the ASR system, such as speech enhancement upon the signal level, extracting the robust feature vectors and adjusting the back-end acoustic models. In the real environment, the situation of ambient noise cannot consider in the prior stage and hard to predict. The noise reduction techniques should not depend upon the assumptions about noisy conditions or training parameters and work well under specific noise scenarios. The primary intention of a noise-robust feature extractor is to cause few or without assumptions about the noise information. This is one of the challenging tasks in favor of recent and ongoing research areas. Appropriate and relevant speech features can differentiate the different speech classes under the disturbance of environmental noise and variability of speaker characteristics [5].

Some of the earlier works have been conducted in different perceptive to increase the noise immunity of Mel frequency cepstral coefficients (MFCC) under noisy situations. MFCC simulates the human auditory system and captures the main characteristics of phonemes in speech. In [6], the spectral subtraction (SpecSub) is used in conventional MFCC by estimating and subtracting the noise spectrum of non-speech region from noisy speech. SpecSub algorithm is considerably good under noise situations; however, it could not function well in clean speech signal. In [7], a psychoacoustic model of frequency masking has been suggested and introduced the transformation of power spectral density in multiple fundamental harmonics frequencies into MFCC. A front-end [8], spectral subtraction algorithm was presented to pre-filter the noise in the speech signal and analyzed various frequency warping scales with a non-perceptual scale. One study proposed the power normalized cepstral coefficients (PNCC) replacing with q-log power function and applied the mean normalization after logarithm to remove the effect of convolution noise [9].

This paper proposes a modified method of MFCC to extract the relevant and appropriate acoustic feature vectors from noisy speech by applying the simultaneous masking effect of human hearing mechanism. The minimum masking threshold is calculated based on the psychoacoustic model, which commonly use in audio watermarking technology can represent the most sensitive limit for distortion of the signal. With the use of the masking effect model into conventional MFCC, the noise effect can be lessened without substantial loss and perceiving deterioration of speech signal.

The organization of this paper is as follows. This paper propose the simultaneous masking effect based cepstral feature and Gammatone filterbank analysis in section 2. In Section 3, the feature recognition engine based on Gaussian Mixture Model-based Hidden Markov Model (GMM-HMM) and hybrid Deep Neural Network (DNN)-HMM are briefly explained. The experimental results and discussion part will present in section 4. The last section concludes the whole paper.

978-1-7281-9896-5/20/$31.00 © 2020 IEEE

II. Methodology

The automatic speech recognition system may lead to low recognition performance due to the variation among speakers, different channels, or surrounding corrupted noise. Thus, the robustness has been a crucial problem in signal and speech processing area [10]. Psychoacoustic is a study of sound perception in the human auditory system includes the concept of auditory masking, how human response in different frequencies, the relation between loudness and sound pressure level. Typically, the concept of examining and modeling the human hearing mechanism is a logical approach to enhance the accuracy of the speech recognition system. One weak audible sound becomes uncleared in the existence of another louder sound. This is called the effect of auditory masking, and which is fundamental in the psychoacoustic modeling process. This masking effect is the relation to the selectivity of auditory processing and how human ears response to different complex sounds in real-life environment. Simultaneous masking occurs between two close frequencies sound components. The low-power signal (maskee) is unhearable by the concurrent existence of another louder sound component (masker). Both this masker and maskee are possibly a tone or narrow-band noise.

Figure 1 shows the nature of simultaneous or frequency masking, where the stronger signal S_0 is the masker. Due to the presence of masker, the absolute hearing threshold is elevated as the new hearing threshold. This is called the masking threshold and is a type of noticeable limit for distortion in the hearing mechanism. Any sound components below this curve cannot be heard and masked by the masker. The faint sounds S_1 and S_2 are wholly unhearable because their sound pressure levels are at a lower place of masking threshold. The sound S_3 is partly masked by S_0 masker and perceivable only the above portion of the masking threshold. The masker produces a sufficient strength excitation patterns on the basilar membrane in the human cochlea. This excitation keeps the catching of a weaker sound excitation within the same critical band [11]. Simultaneous masking is usually happening in a real environment. The idea behind the audio watermarking technology utilize the minimum masking threshold (MMT) to hide the watermark information. Depend on this consideration, this paper introduces a modified MFCC by combining with the masking effect to extract the robust acoustic features from noisy speeches. With the use of masking models into feature extraction, can figure out which frequency components commit more to the masking threshold and how much noise effect can mix into the signal without being anticipated. Moreover, it can also figure the amplitude of speech signal.

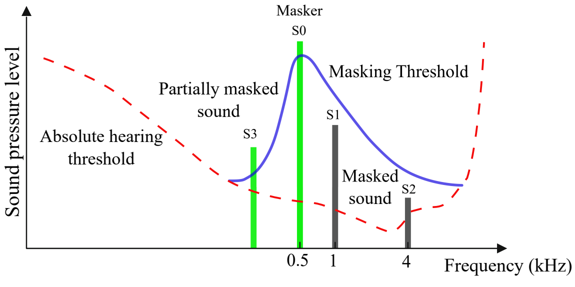


Fig. 1: Nature of Simultaneous or Frequency Masking.

A. Modeling Minimum Masking Threshold

There are six documented processing steps to implement the modeling of minimum masking threshold in psychoacoustic model [11] [12] [13].

Step 1: *Perform FFT analysis*: Spectral estimation is computed for each frame by applying the Fast Fourier Transform (FFT) to produce the spectral coefficients. The power spectral density estimate of $\tilde{x}(n)$ is defined as follows.

$$PSD(k) = 10\log_{10}\left|\frac{1}{N}\left[\sum_{n=0}^{N-1}\tilde{x}(n)\exp\left(\frac{-j2\pi nk}{N}\right)\right]\right|^2 \quad (1)$$

where $0 < k \leq N/2$. Then, power spectral density estimate $PSD(k)$ is normalized to 65dB sound pressure level due to the maximum sensitive part of human conversation speech.

$$P(k) = 65 - \max\{PSD(k)\} + PSD(k) \quad (2)$$

Step 2: *Finding tonal and non-tonal components*: The tone-like, noise-like frequency components are selected from the maximum of normalized power spectral density estimate within two neighbors, which called the local maxima. If the local maxima value is minimum 7dB exceeding than its neighboring components within a specific Bark range D_k, such a maxima is denoted as a tonal masker. Otherwise, treated as a non-tonal masker.

$$S_{\text{TM}} = P(k) \mid [P(k) - P(k \pm D_{\text{k}})] \geq 7dB \quad (3)$$

$$S_{\text{NM}} = P(k) \notin S_{TM} \quad (4)$$

where S_{TM} is a set of tonal maskers and S_{NM} is a set of non-tonal maskers. As the masking effect is additive in logarithmic domain, the sound pressure level of each masker is computed as follow:

$$P_{TM,NM}(k) = 10\log_{10}\sum_{j}\left[10^{\frac{P(j)}{10}}\right]\forall P(j) \quad (5)$$

where $P(j)$ is the set of tonal and non-tonal components, $P(j) \in S_{TM,NM}$.

Step 3: *Determination of valid maskers*: The magnitude of each masker must exceed the absolute threshold of hearing. Any group of maskers must be taking place within 0.5 Bark distance. Only the masker with the highest sound pressure level value can preserve and the rest can eliminate.

$$P_{TM,NM}(k) \quad \geq \quad ATH(k) \quad (6)$$

$$P_{TM,NM}(k) = arg \max_{k_0 \in [-0.5,0.5]} P_{TM,NM}(k + k_0) \quad (7)$$

Step 4: *Figuring individual masking thresholds*: The tonal and non-tonal masking threshold expresses a masking contribution at frequency index i to another masker located at frequency index j. The individual masking thresholds, $T_{(TM,NM)}(i,j)$ are given by:

$$T_{TM,NM}(i,j) = P_X[z(j)] + \Delta_X[z(j)] + SF(i,j) \qquad (8)$$

where $P_X[z(j)]$ refers to the sound pressure level of the tonal or non-tonal masker in frequency index j, $z(j)$ is the Bark frequency of j. The term Δ_X is masking index of tonal and non-toanl masker, and $SF(i,j)$ denotes the spreading function of masking contribution from masker at j to maskee at i.

Step 5: *Figuring of global masking threshold*: The powers corresponding to the upper and lower slopes of individual sub-band masking curves and a given absolute hearing threshold are summed to form a composite global masking contour.

$$T_g(i) = 10^{\frac{ATH(i)}{10}} + \sum_{j=1}^{N_{TM}} 10^{\frac{T_{TM}(i,j)}{10}} + \sum_{j=1}^{N_{NM}} 10^{\frac{T_{NM}(i,j)}{10}} \qquad (9)$$

$$T_g(i) = 10 \log_{10} \quad T_g(i) \qquad (10)$$

Step 6: *Figuring of minimum masking threshold*: The minimum masking threshold is gained from the global masking threshold. The spectral subsamples of global masking threshold are mapped onto n^{th} uniform sub-bands $(1 \leq n \leq 32)$.

$$T_M(m) = \min_{f_{id}(i) \in n} T_g(i) \qquad m = [8(n-1)+1] : 8n \qquad (11)$$

The MMT, which is figured from a psychoacoustic model represents the most sensitive limit or just a noticeable distortion of the signal. Any sounds or frequency components lie under the threshold that cannot be heard and masked by a masker [13]. This conception is analyzed and integrated into the conventional MFCC feature extraction to tolerate the noise impact. The spectrum value of each frame is compared with the minimum masking threshold. If the spectrum value is lower than the masking threshold, this spectrum is set to the value of the threshold limit. In such a way, the modified spectrum magnitude is worked out on each frame and passed through the process of filterbank analysis.

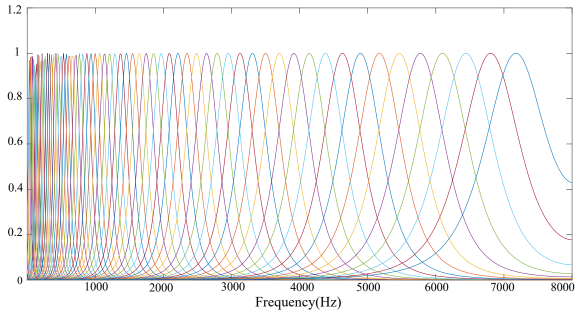


Fig. 2: Frequency Integration of Gammatone Filterbank on 16kHz.

B. Gammatone Filterbank Analysis

The speech signal causes the vibration on the basilar membrane of the inner ear's cochlea in the human auditory system. The localized frequency information of the speech signal is reacted to each part of the basilar membrane. Similarly, the digital filterbank resembles the processing of basilar membrane in auditory modeling. Each bandpass filter is simulated the frequency characteristics of the basilar membrane [14]. The human hearing is the most sensitive to frequencies between 2000-5000Hz and less sensitive in high frequency region. This will affect the performance of the ASR system. The triangular shape Mel filterbank is used to warp the spectrum envelop in conventional MFCC to overcome this problem. Triangular shape filterbank is symmetrically tapered at the ends, and it cannot render the weight outsides the sub-bands. As a consequence, the correlation between sub-bands and nearby spectral information from adjacent sub-bands may lead to losing. Gammatone filterbank has gaussian shape and allows for gradually decreasing the weights at both ends and tolerate for compensating the possible loss of spectral information correlation [15]. The frequency response of a Gammatone filterbank is illustrated in Figure 2. This gaussian shape of filterbank is substituted in place of the triangular Mel filterbank. The Gammatone filterbank is physiologically motivated to imitate the structure of the human auditory system. The frequency response of the Gammatone function in the time domain is specified as follows [16]:

$$g(t) = at^{n-1} cos(2\pi f_c + \phi) e^{-2\pi bt} \qquad (12)$$

where a is the amplitude; n is the order of filter which determines the slope of each filter; f_c is the center frequency; π is the phase shift and b is the bandwidth of filter which specifies the duration of impulse response.

III. Perceptual based Gammatone Frequency Cepstral Coefficients (PGFCC)

The proposed feature extraction involves the modeling of minimum masking threshold based on the presence of masker and maskee on every frame. Each frame has 512 points, and the frame-shift is 384 points. Firstly, the Hamming windowing process is done in order to keep the continuities at the beginning and end of each frame. Then, the minimum masking threshold is figured out based on the simultaneous or frequency masking of psychoacoustic model. The tone-like and noise-like components are determined from normalized power spectral density based on minimum local maxima value (7dB) and also takes out the irrelevant elements. After that, the individual and global masking thresholds are calculated. After figuring the global masking threshold, the spectral subsamples are mapped onto 32 uniform subbands to generate the minimum masking threshold. Then, the normalized FFT outputs are compared with the minimum masking threshold on each frame. If the spectral value is lower than the threshold, assigns with the minimum limit value. These modified spectral features are passed through the gaussian shape 64 channels Gammatone filterbank to wrap the spectrum envelop. After getting the

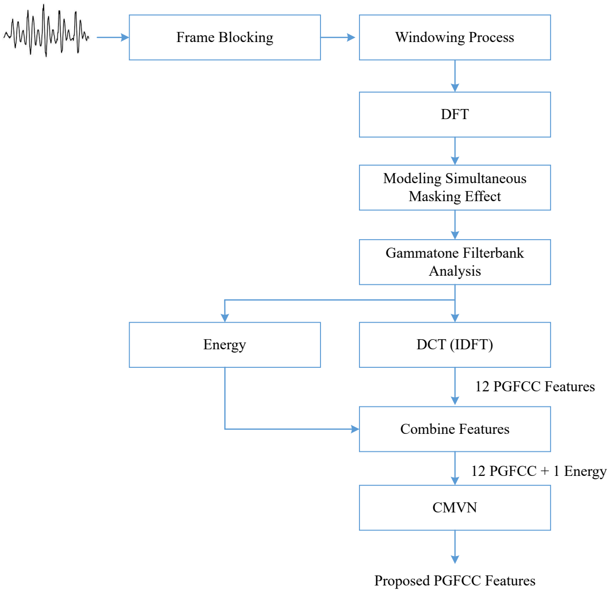


Fig. 3: Process Flow of Proposed Perceptual Simultaneous Masking Effect based Cepstral Feature Technique.

results of filterbank process, the discrete cosine transform (DCT) is done on the coefficients to achieve the maximum decorrelation among the feature vectors. The amplitudes of speech signal are varied in time. These amplitude variations represent the short-term energy which gives information about the time-dependent characteristics. The short-time spectral energy is calculated on the log filterbank energy of each frame. Finally, the first (2-13) cepstral coefficients and short-term spectral energy (PGFCC+Energy) are defined as the proposed front-end feature vectors. Figure 3 illustrates the detail process flow of proposed feature extraction technique.

IV. Feature Recognition Engine

After extracting the cepstral features from the speech signal, the feature recognition engine judges the possible word sequences within the feature space from training utterances. In this section will describe the statistical approaches called Gaussian Mixture Model (GMM) based hidden Markov model (HMM) and cross entropy-based hybrid Deep Neural Network (DNN) - HMM model as the recognition engines. Kaldi toolkit is used to implement the acoustic model training for this study.

A. GMM-HMM Model

Gaussian Mixture Model-based Hidden Markov Model (GMM-HMM) is one of the most popular statistical models to interpret the sequential structure of the speech. Each HMM state applies a mixture of Gaussian to model a spectral representation of speech. Each training process is followed by the alignment between the acoustic feature vector and sound units [17].

Triphone training (PGFCC+Δ+$\Delta\Delta$): This training concentrates on the contextual information between left and right phonemes. In general, the delta and delta-delta features represent the first order and second-order derivatives.

Triphone training using LDA+MLLT: The linear discriminant analysis (LDA) uses the spliced PGFCC features and reduces feature dimension into 40 for all data to produce the HMM states. MLLT process applies to enforces the linear transformation to get a significant change for individual speaker.

Speaker Adaptive Training using fMLLR: The primary goal of speaker adaptation is to modify the acoustic model parameters to match the features of actual audible. Features are linearly transformed to normalize the variability of speakers with the use of feature space maximum likelihood linear regression (fMLLR).

B. DNN-HMM Model

This phase trains a DNN to provide posterior probability estimates for each HMM state from given observation sequence. The networks are trained to optimize a given training objective function using the standard error back-propagation procedure. Typically, frame-level cross-entropy is employed as the objective, and optimization is through with the mini-batch stochastic gradient descent (SGD) [18]. The parameter learning with a cross-entropy criterion is processed as three steps. Firstly, the GMM-HMMs model is trained with the use of the maximum likelihood estimation, which is applied in DNN-HMM training that contains the state's prior and transition probabilities. Then, the forced alignment is generated by matching the acoustic features vectors with the corresponding labels from the Viterbi algorithm with GMM-HMM model. State labels are the learning target of the DNN output layer and the weight values of DNN training is done by minimizing the cost function of cross-entropy given as below,

$$C = -\sum_{Q}^{i=1} q_i \; log \; p_i \qquad (13)$$

where C denotes as the cross entropy function, Q is the set of states, p_i is the softmax layer output and q_i is the targeted state.

V. Results and Discussion

AURORA database is designed by the European Telecommunications Standards Institute (ETSI) to evaluate and standardize the performance of distributed speech recognition (DSR) systems in noisy environment. AURORA-2 [19] is based on the original version of TIDigits English connected digits database launched from Linguistic Data Consortium (LDC) [20]. The total 8,440 utterances are taken from the training part of TIDigits, participating with 55 male and 55 female adult speakers. These utterances are equally split into twenty subsets and contain 422 utterances in each subset. These twenty subsets represent four different noise situations; subway (recording in a moving suburban train), exhibition (atmosphere in a classical exhibition hall with mixture of voices), babble (atmosphere in mixture of several chatter voices) and car noise (recording inside a running car) and different SNRs levels (clean, 20dB, 15dB, 10dB, 5dB). 1,001 utterances of 52 male and 52 female speakers are taken from the testing part of TIDigits database. The summary of detail data usage was described in Table I.

TABLE I: Detail Description of Aurora2 Speech Corpus

Category	**Description**	
Vocabulary	continuous digits sequences (0-9) plus ’oh’	
Sampling	44.1kHz, 16bits, mono channel	
Participants	Male 111 spks	21-70 ages
	Female 114 spks	17-59 ages
Training	8,440 utts.	Multi-condition*
Testing	1,001 utts.	Subway,Babble,Exhibition,Car

*Multi-condition training under subway, exhibition hall, car and babble noise at clean condition and SNR of 20dB, 15dB, 10dB and 5dB.

In this paper, the recognition of connected digit has been carried out to evaluate the noise robustness of proposed features extraction technique. The feature extraction part is implemented using the MATLAB software. These extracted acoustic features are passed through the Kaldi toolkit to build the speaker adaptive triphone model using GMM-HMM and hybrid DNN-HMM techniques. Firstly, the speaker adapted training (SAT), i.e., train on fMLLR adapted features is built. In this training, we use the number of leaves is 300 and the total gaussian is 3000. Then, the frame-level cross entropy-based DNN-HMM training is built with three hidden layers before softmax. Each hidden layer has 378 neurons and the network has 272 output units. The initial learning rate is 0.008, and the mini-batch size is 256 as default.

TABLE II: Recognition accuracy (%) of MFCC feature extraction on different noise situations and different SNR levels

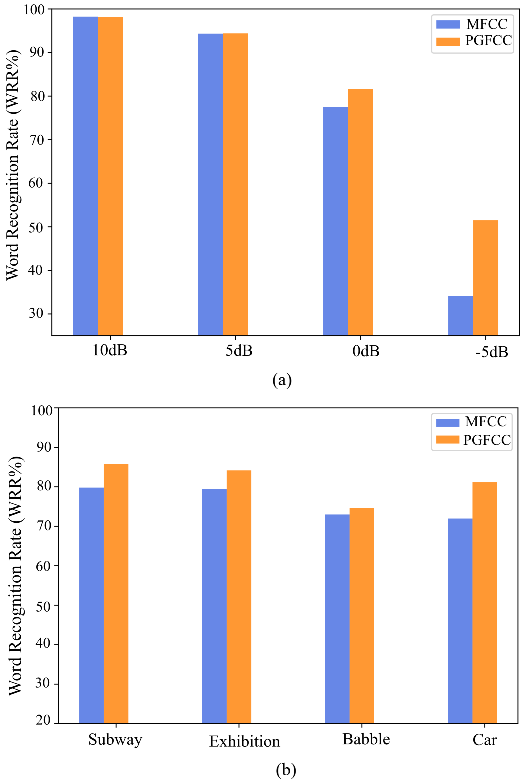
Model	SNR	Subway	Exhibition	Babble	Car	Avg.
GMM	10dB	96.68	93.37	96.61	95.56	95.5
	5dB	91.56	84.45	86.94	87.06	87.5
	0dB	71.72	62.82	54.87	46.82	59.05
	-5dB	23.95	25.64	16.9	9.90	19.09
DNN	10dB	98.71	97.72	97.97	98.54	98.24
	5dB	95.92	93.86	93.02	94.54	94.34
	0dB	82.56	80.99	71.19	75.40	77.54
	-5dB	42.00	45.26	29.78	19.36	34.10

Table II expresses the word recognition rate (%) of MFCC technique over four different situations of noises, namely, subway, exhibition, babble, and car noises. With the use of GMM-HMM model, the average recognition accuracy of overall noise situations takes a value of 95.55% under SNR of 10dB, 87.50% under 5dB, 59.05% under 0dB and 19.09% for SNR of -5dB, respectively. Additionally, the cross entropy-based DNN-HMM model is evaluated to boost the performance of the system. The hybrid DNN-HMM systems have the advantages of DNN’s strong learning power and HMMs sequential modeling ability to outperform the existing GMM-HMM systems. According to this experiment, the average recognition accuracy achieved up to 98.24% in SNR of 10dB, 94.34% in 5dB, 77.54% in 0dB and 34.10% in -5dB, respectively. However, the conventional MFCC functions well in a quiet environment while the result is degrading under background noise. The implementation of simultaneous masking effect using the psychoacoustic model is integrated into MFCC to defeat the drawback of conventional MFCC. By introducing the simultaneous masking effect into feature extraction technique, the noise effect of speech signal can be lessened while the noise is higher in speech signal, and it can also minimize the irrelevant feature components.

TABLE III: Recognition Accuracy (%) of Proposed PGFCC Feature Extraction on different noise situations and SNRs

Model	SNR	Subway	Exhibition	Babble	Car	Avg.
GMM	10dB	97.36	95.16	95.19	95.82	95.88
	5dB	92.08	89.63	85.13	87.92	88.69
	0dB	76.11	71.46	55.50	57.08	65.04
	-5dB	36.97	32.27	21.55	16.91	26.93
DNN	10dB	98.89	98.52	96.86	98.30	98.14
	5dB	95.73	95.40	91.05	95.41	94.40
	0dB	86.64	85.22	71.80	83.00	81.67
	-5dB	61.74	57.51	38.81	47.93	51.50

According to this experiment, our proposed feature extractor against the noise and outperformed the MFCC, especially in low SNRs 0dB and -5dB. Table III expresses the recognition accuracy of proposed PGFCC technique. With the use of GMM-HMM model, the recognition performance was achieved 95.88% in SNR of 10dB, 88.69% in 5dB, 65.04% in 0dB and 26.93% in -5dB respectively. When the DNN-HMM



(a)

(b)

Fig. 4: Overall noise validation of MFCC and proposed PGFCC (a) Mean accuracy at different SNR level (b) Mean accuracy separated for different noise types using cross-entropy based DNN-HMM acoustic model.

model has been building, the average accuracy gained up to 98.14% in case of 10dB, 94.40% in 5dB, 81.67% in 0dB and 51.50% in -5dB. As illustrated in Figure 4(a), when comparing with conventional MFCC, the average relative improvements were 6.44% in SNR of 5dB, 25.57% in 0dB, and 91.2% in -5dB respectively with the use of hybrid DNN-HMM acoustic model. However, the recognition performance decreased in terms of 10dB. The relative decrements were 0.1% using DNN-HMM. Besides, the overall accuracy (%) was followed up to find out how much the results will grant in different noise situations under all SNRs level. As seen in Figure 4(b), the average accuracy of proposed PGFCC outperformed in all types of noise situations. Using the DNN-HMM model, the relative improvement was 7.5%, 5.91%, 2.24%, and 12.78% under subway, exhibition, babble and car noises, respectively.

Although the relative improvement is not too much significant in higher SNR level comparing with conventional MFCC, we observed that the proposed algorithm achieved the more substantial improvement in lower SNR levels especially at 0dB and -5dB. Moreover, the proposed algorithm cannot be recognized well in types of babble noise in terms of 10dB and 5dB. As the nature of babble noise is the atmosphere in a mixture of various chatter voices, such kind of sound may lead to diverge the human perception in the auditory system. Our proposed method finds the sensitive limit threshold for perceiving in the hearing mechanism based on auditory masking and influence the amplitude of the signal. As a consequence, in the situation of higher SNR level (10dB and 5dB) under babble noise, our proposed method may lead to a substantial loss of spectral information and possibly distort the original clean signal.

VI. Conclusion

In this paper, we propose a modified front-end algorithm based on MFCC which is successfully implemented with simultaneous masking and Gammatone frequency integration. This method imitates the nature of human auditory system to handle noise-adverse situations. Aurora-2 database is used to carry out the experiments. The GMM-HMM and DNN-HMM recognizers are used to prove the robustness of proposed front-end algorithm. Although the word recognition rates of our proposed method have achieved comparable results with MFCC for high SNRs, they are significantly outperformed in lower SNRs, especially at 0dB and -5dB. The highest accuracy is gained by DNN-HMM, which provides 98.14% in 10dB, 94.40% in 5dB, 81.67% in 0dB and 51.50% in -5dB. Yet while the proposed modification algorithm still managed to improve the satisfying accuracy in high SNR of babble noise, the recognition rate was higher in more noisy conditions over MFCC. The future effort will assess a detailed analysis on proposed algorithm with new improvement in clean condition and babble noise. Also, the performance will analyze on large vocabulary continuous speech recognition system.

Acknowledgment

This work was financially supported by ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net), JICA. This study was also supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B) (18H0321) and the Ministry of Internal Affairs and Communications for SCOPE Program (185001003).

References

[1] S. A. Majeed, H. Husain, S. A. Samad, and T. F. Idbeaa, “Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study,” J. Theor. Appl. Inf. Technol., vol. 79, no. 1, pp. 38–56, 2015.

[2] B. T. Sai, I. C. Yadav, S. Shahnawazuddin, and G. Pradhan, “Enhancing pitch robustness of speech recognition system through spectral smoothing,” 12th Int. Conf. Signal Process. Commun. SPCOM, pp. 242–246, 2018.

[3] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A Review on Speech Recognition Technique,” Int. J. Comput. Appl., vol. 10, no. 3, pp. 16–24, 2010.

[4] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, “Robust speech recognition with feature extraction using combined method of RSF and DRA,” in IEEE International Symposium on Communications and Information Technologies: ISCIT, 2004, vol. 2, pp. 1001–1004.

[5] S. J. Arora and R. P. Singh, “Automatic speech recognition: A Review,” Int. J. Comput. Appl., vol. 60, no. 9, pp. 34–44, 2012.

[6] A. L. Georgescu, H. Cucu, C. Burileanu, “SpeeD’s DNN approach to Romanian speech recognition,” in International Conference on Speech Technology and Human-Computer Dialogue (SpeD),” 2017, pp. 1–8.

[7] K. K. Tomchuk, “Spectral Masking in MFCC Calculation for Noisy Speech,” in Wave Electronics and its Application in Information and Telecommunication Systems, WECONF, 2018, pp. 1–4.

[8] N. Upadhyay and H. G. Rosales, “Robust Recognition of English Speech in Noisy Environments Using Frequency Warped Signal Processing,” Natl. Acad. Sci. Lett., vol. 41, no. 1, pp. 15–22, Feb. 2018.

[9] H. F. Pardede, “On noise robust feature for speech recognition based on power function family,” in International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS, 2015, pp. 386–391.

[10] D. Darabian, H. Marvi, and M. S. Noughabi, “Improving the performance of MFCC for Persian robust speech recognition,” Journal of Artificial Intelligence. Data Mining, vol. 3, no. 2, pp. 149–156, 2015.

[11] Y. Lin and W. H. Lin, “Audio watermark: A comprehensive foundation using MATLAB,” 2015.

[12] H. K. Maganti and M. Matassoni, “A perceptual masking approach for noise robust speech recognition,” Eurasip Journal of Audio, Speech, Music Processing, vol. 2012, no. 1, pp. 1–9, 2012.

[13] H. M. S. Naing, R. Hidayat, B. Winduratna, and Y. Miyanaga, “Psychoacoustical masking effect-based feature extraction for robust speech recognition,” International Journal of Innovative Computing, Information and Control, vol. 15, no. 5, pp. 1641–1654, 2019.

[14] M. Russo, M. Stella, M. Sikora, and V. Pekić, “Robust cochlear-model-based speech recognition,” Computers, vol. 8, no. 1, 2019.

[15] G. K. Liu, “Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech,” arXiv preprint arXiv:1806.09010, 2018.

[16] J. Qi, D. Wang, Y. Jiang, and R. Liu, “Auditory features based on Gammatone filters for robust speech recognition,” in Proceedings - IEEE International Symposium on Circuits and Systems, 2013, pp. 305–308.

[17] P. Upadhyaya, S. K. Mittal, Y. V. Varshney, O. Farooq and M. R. Abidi, “Speaker adaptive model for hindi speech using Kaldi speech recognition toolkit,” International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), 2017, pp.222-226.

[18] V. V. Vegesna, K. Gurugubelli, H. K. Vydana, B. Pulugandla, M. Shrivastava, and A. K. Vuppala, “DNN-HMM acoustic modeling for large vocabulary Telugu speech recognition,” in International Conference on Mining Intelligence and Knowledge Exploration, 2017, pp. 189–197.

[19] D. Pearce and H. G. Hirsch, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in 6th International Conference on Spoken Language Processing, ICSLP, 2000.

[20] R. Leonard, “A database for speaker-independent digit recognition,” in ICASSP ’84. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1984, vol. 9, no. 1, pp. 328–331.