International Journal of Innovative Computing, Information and Control
Volume **15**, Number **5**, October **2019**

ICIC International ©2019 ISSN 1349-4198
pp. **1641–1654**

PSYCHOACOUSTICAL MASKING EFFECT-BASED FEATURE EXTRACTION FOR ROBUST SPEECH RECOGNITION

Hay Mar Soe Naing[1,2], Risanuri Hidayat[1], Bondhan Winduratna[1] and Yoshikazu Miyanaga[2]

[1]Department of Electrical Engineering and Information Technology
Gadjah Mada University
Jln. Grafika 2, Yogyakarta 55281, Indonesia
haymarsoenaing@ucsy.edu.mm; {risanuri; windurat}@ugm.ac.id

[2]Graduate School of Information Science and Technology
Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
miya@ist.hokudai.ac.jp

Received January 2019; revised May 2019

Abstract. *A new approach for speech feature extraction in automatic speech recognition (ASR) is proposed in this paper. It is based on the human auditory system. Generally, the mel frequency cepstral coefficients (MFCC) are the most widely used speech features in ASR systems, but one of their main drawbacks is background noise, which can affect and hamper the results. This paper proposes noise robust speech features which improve upon the MFCC. A psychoacoustic model-based feature extraction that simulates the perception of sound in the human auditory system is investigated and integrated into the MFCC. The complexity of the signal can be reduced by using a masking effect during feature extraction, minimizing the feature components without any significant loss in perceiving quality of sound. Moreover, it can reduce the noise effect of speech signal. In this paper, a hidden Markov model is employed to recognize English isolated digits. These experiments verify that the proposed modified method effectively improves the recognition under adverse situations. With respect to the use of perceptual masking effect-based cepstral features, the accuracy reached up to 97.16% in signal to noise ratio at 10dB, 95.02% at 5dB, 90.34% at 0dB, 77.08% at −5dB and 62.76% at −10dB, respectively.*
Keywords: Hidden Markov model, Masking effect, Mel frequency cepstral coefficient, Psychoacoustic model, Speech recognition

1. **Introduction.** Speech is the most efficient communication medium for human beings. With the recent appreciable improvement of computer technology, people are comfortable interacting with computers via speech to receive efficient real-time responses without the need for extraneous devices. This fact has been an ongoing challenge in maturing an automatic speech recognition system [1, 2]. ASR has the potential to support multiple applications in various markets, such as medical transcription, assisting the disabled, home automation, automobile audio, telephony, telecommunications and speech-to-speech translation [3]. However, speech recognition systems remain far from the desired target.

Conventional ASR systems cannot consistently recognize spoken words correctly under any acoustic noise environment [4, 5, 6]. In practice, many marketed ASR systems can recognize speech well in a silent environment, but recognition performance is degraded significantly for noisy speech. There are several aspects to ASR tasks including speech

DOI: 10.24507/ijicic.15.05.1641

quality, vocabulary size, accents and speaker characterstics, but the most significant factors in the recognition process are channel and noise circumstances [7]. During any kinds of noise with a low signal to noise ratio (SNR), ASR systems have difficulty recognizing correct word sequences.

Speech feature extraction is regarded as the most crucial phase in a speech recognition task. The primary goal of feature extraction in noise robust ASR is to extract the most relevant and effective information from noisy speech [8]. The more important and better features to represent the speech can derive accurate recognition results. In the literature, some studies have been conducted to produce noise robust speech features for ASR under noisy situations. One study [9] evaluated three types of noise robust features: power normalized cepstral coefficients (PNCC), mel frequency cepstral coefficients (MFCC), and Sphinx-embedded denoising features on French broadcast news corpus. Their results showed that PNCC features are a promising set of noise robust features and gave the best results. However, they are strongly dependent on the initialization of the normalization factor, especially for short duration speech segments.

The implementation of a denoising wavelet in the MFCC feature extraction process on 0-9 Indonesian digits has been presented [10]. Eleven wavelet methods and a ten-level decomposition were investigated during the denoising process to improve the recognition rate on input signals with 0-10dB SNRs. All the tested wavelet transformation-based denoising processes were able to achieve higher accuracy compared to a system without denoising on SNRs of 0-10dB. In one paper, Qi et al. [11] presented the implementation of a time-domain and a frequency-domain gammatone filter-based cepstral feature (GFCC) to improve the robustness of speech recognition in noisy situations. They expressed that the time-domain GFCC was able to attain performance superior to MFCC and perceptual linear prediction features while providing a significantly faster processing speed. Nonetheless, it is inadequate to model the asymmetric non-linear frequency selective characteristics of both basilar membranes and auditory nerve responses.

The implementation of a psychoacoustic model in simultaneous and temporal masking that comes from frequency and time-domain was investigated by Dai et al. [12]. They compared the different implementation styles of psychoacoustic models and explored the effect of a 2D psychoacoustic P-filter on removing noise and increasing the SNR. However, the most efficient feature in recognizing speech under very noisy environment (at a low SNR level, such as -5dB and -10dB) has always been a popular debating subject.

This study proposes a perceptual masking effect-based feature extraction technique in the development of an ASR system. The primary intention of this work is to improve the robustness of a recognizer in very noisy situations. The human hearing system can work properly in adverse situations, such as background noise, channel distortion and speaker variability. The idea of analyzing and modeling the perception of human ears is a logical approach to improve the recognition accuracy of an ASR system [13]. The extraction of spectral information can improve greatly by modeling the non-linear perceptions of human hearing sensation [14]. Thus, the minimum masking threshold (MMT) computed from a psychoacoustic model which represents the most sensitive limit of the signal is explored and combined into the MFCC technique. In applying this model to a conventional MFCC, the irrelevant feature components can depreciate while the noise effect of the speech signal lowers. The effectualness of the proposed robust features presents with a suite of experiments on TIDIGITs isolated English digits corpus, which are corrupted with different noises to simulate the real-world circumstances.

This paper is organized as follows. A brief explanation of how an auditory masking effect occurs in our natural environment is described in Section 2. A detailed algorithm description of the psychoacoustic masking model which we used in the acoustic feature

extraction process is given in Section 3. This is followed by the hidden Markov model that we used for feature recognition in Section 4. The experimental database and results are given in Section 5, and we conclude our proposed work in Section 6.

2. **Auditory Masking.** The sounds in our natural world rarely occur in isolation because sounds are complex and usually occur simultaneously. Typically, a masking effect concerns the perceptual interaction between sound elements. Auditory masking means that a weak but audible sound becomes imperceptible to the human ear due to the existence of other louder sound elements [15, 16]. The study of masking is useful concerning the selectivity of the auditory system and how human beings process complex sounds in our environment. As illustrated in Figure 1, there are two kinds of auditory masking that influence human hearing sensation: simultaneous and non-simultaneous. Auditory masking occurring in the frequency domain is called simultaneous, frequency or spectral masking, while masking occurring in the time domain is called non-simultaneous or temporal masking [17].

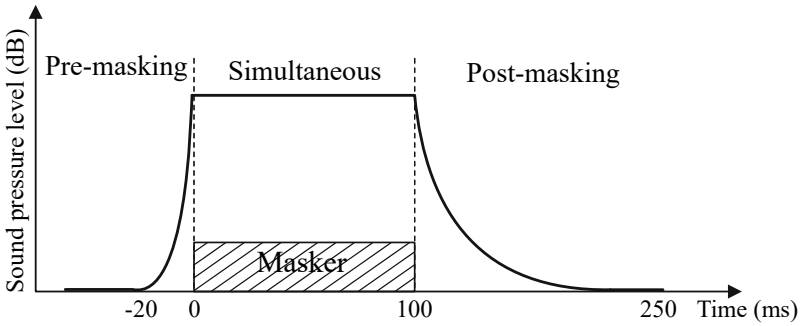

FIGURE 1. Two types of auditory masking: simultaneous and non-simultaneous

2.1. **Non-simultaneous masking.** Non-simultaneous masking occurs in a situation where the signal and masker present relatively delayed in time. One sound component is made inaudible by another sound which originates immediately preceding or following the source sound [13]. The temporal masking can be divided into two types, namely pre-masking and post-masking. Backward masking or pre-masking is unexpected, taking place immediately preceding the presence of masker. Masking which obscures a sound immediately following the masker is called forward masking or post-masking [18]. Figure 2 illustrates the typical phenomenon of non-simultaneous or temporal masking.

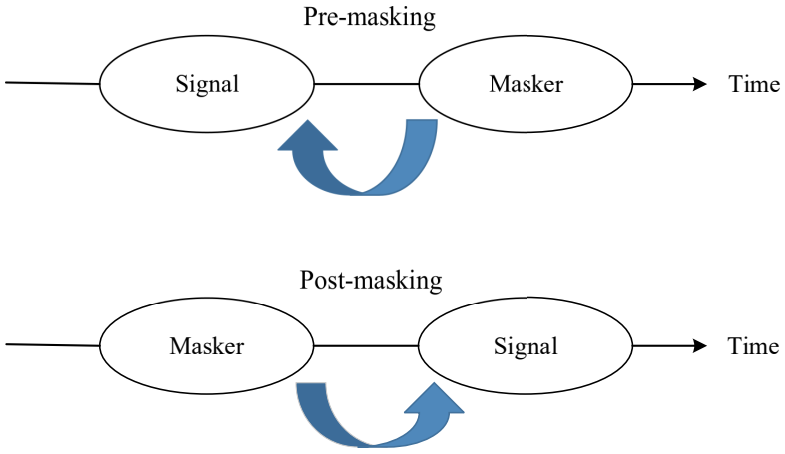

FIGURE 2. Paradigm of non-simultaneous or temporal masking

2.2. **Simultaneous masking.** Simultaneous masking occurs when a sound is inaudible due to the presence of other unwanted sounds occurring in the same duration. This type of masking is generally used in audio compression and speech enhancement [19]. The paradigm of simultaneous masking is described in Figure 3, where the sound component S_0 is a masker. Due to the existence of S_0, the threshold in quiet rises to a new hearing threshold, which is a limit for just noticeable distortion. Any sound components below this masking threshold (S_1 and S_2) are rendered inaudible by S_0, since their sound pressure level lies below the threshold. S_3 is partially masked by S_0 with only the portion above the masking threshold being perceivable [19].

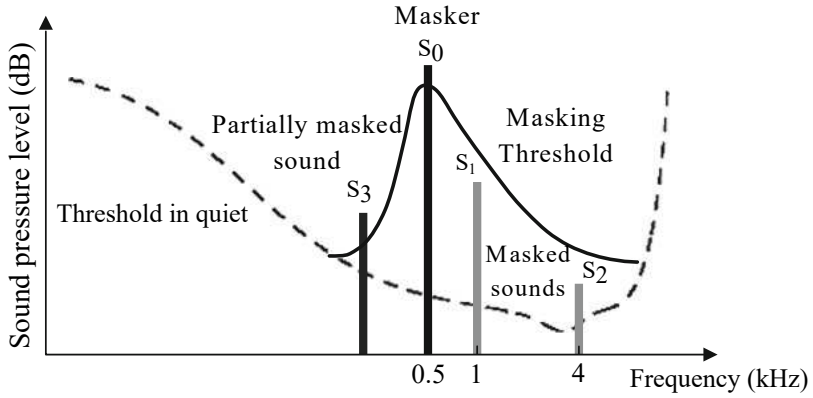

FIGURE 3. Paradigm of simultaneous masking

3. **Psychoacoustic Masking Model.** Speech parameterization is the main component in the speech recognition process, extracting the acoustic feature vectors from the speech signal. All speech is composed of spectrum information and different individual utterance characteristics. As the extraction of the most important and useful features can help to achieve the high recognition performance, this paper proposes a new approach to feature extraction that generates such features for isolated digit recognition tasks. By integrating the masking effect of a psychoacoustic model used to compress audio into conventional MFCC, the more robust and important features can be extracted from noisy speech utterances.

A psychoacoustic model is a form of the quantitative model which imitates the human sensation of sound in the auditory system. It accomplishes this by manipulating the auditory masking characteristic of human ears [17]. A mixture of various sounds surrounded by a real-world environment and the perception of one sound becomes obscured by the presence of another. This phenomenon is called auditory masking and is fundamental to psychoacoustical modelling [16]. Modelling and analyzing the effect of simultaneous masking are major tasks in psychoacoustic modelling and mostly occur in a real-world environment. The MMT calculated from a psychoacoustic model used in an audio watermarking technique can represent the distortion limit while its psychoacoustic principle allows the signal and computation of noise effects to be examined [19].

3.1. **Modeling the effect of simultaneous masking.** There are a series of six documented steps to model the effect of simultaneous masking [19]. They are:

- Conversion from a time to a frequency domain;
- Identification of tonal and non-tonal components from the speech signal;
- Decimation of invalid tonal and non-tonal maskers;
- Figuring of individual masking thresholds;

- Determination of the global masking threshold;
- Determination of the minimum masking threshold for each sub-band.

A detailed explanation of each step is provided in the following sub-sections.

3.1.1. *Conversion from a time to a frequency domain.* The fast Fourier transform is executed to obtain the high-resolution spectral estimate from each frame $x(n)$ for an accurate analysis of the frequency components. A Hamming window $w(n)$ is applied to incoming frames to minimizing the spectral leakage effect. The power spectral density (PSD) estimate values are computed as follows:

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

where the value of k is between 0 and N/2 and j is the square root of -1. Then, $PSD(k)$ is normalized to a sound pressure level of 65dB because the maximal is limited to 65dB for average human conversation speech.

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

3.1.2. *Identification of tonal and non-tonal components.* Tonal maskers are picked out from the local maxima values of the normalized power spectral density estimate. The local maxima is defined as the maximum *PSD* between two neighbors.

$$P(k) \geq P(k+1) \quad \text{and} \quad P(k) \geq P(k-1) \tag{3}$$

If the value of the local maxima is greater than 7dB of its neighboring components within a certain Bark range D_k, these maxima can be regarded as tonal components.

$$S_{TM} = \{P(k) \mid [P(k) - P(k \pm D_k)] \geq 7\text{dB}\} \tag{4}$$

where D_k is a Bark range. The value of D_k can be varied using different frequency indices:

$$D_k \in \begin{cases} \{\pm 2\}, & 2 < k < 63 \quad \leftrightarrow \quad \dfrac{2F_s}{N} \sim \dfrac{63F_s}{N}\,\text{kHz} \\ \{\pm 2, \pm 3\}, & 63 \leq k < 127 \quad \leftrightarrow \quad \dfrac{63F_s}{N} \sim \dfrac{127F_s}{N}\,\text{kHz} \\ \{\pm 2, \ldots, \pm 6\}, & 127 \leq k \leq 250 \quad \leftrightarrow \quad \dfrac{127F_s}{N} \sim \dfrac{250F_s}{N}\,\text{kHz} \end{cases} \tag{5}$$

As the effect of masking is a logarithmic scale, the sound pressure level of each tonal component is computed as follows:

$$P_{TM}(k) = 10\log_{10}\left[10^{\frac{P(k-1)}{10}} + 10^{\frac{P(k)}{10}} + 10^{\frac{P(k+1)}{10}}\right] \tag{6}$$

Otherwise, the remaining components within each critical band can be denoted as non-tonal components.

$$P_{NM}(k) = 10\log_{10}\sum_j \left[10^{\frac{P(j)}{10}}\right] \quad \forall\, P(j) \notin S_{TM} \tag{7}$$

3.1.3. *Decimation of invalid tonal and non-tonal maskers.* The sets of invalid tonal and non-tonal maskers are analyzed and discarded based on the following two criteria.

- Any tonal and non-tonal maskers below the threshold in quiet can be discarded.
- Any group of maskers occurs within a distance of 0.5 Bark. It means only the masker with the highest SPL preserve and the rest can eliminate.

3.1.4. *Figuring of individual masking thresholds.* After discarding invalid maskers, calculate the individual masking threshold for each tonal and non-tonal masker as follows:

$$L_{TM}[z(j), z(i)] = P_{TM}[z(j)] + \Delta_{TM}[z(j)] + SF[z(j), z(i)] \tag{8}$$

$$L_{NM}[z(j), z(i)] = P_{NM}[z(j)] + \Delta_{NM}[z(j)] + SF[z(j), z(i)] \tag{9}$$

where $L_{TM}[z(j), z(i)]$ and $L_{NM}[z(j), z(i)]$ are the individual masking thresholds for tonal and non-tonal maskers, respectively. The masker at frequency index j leads to the masking effect on maskees at frequency index i. $z(j)$ and $z(i)$ are the respective masker and maskee frequencies on the Bark scale. $P_{TM}[z(j)]$ and $P_{NM}[z(j)]$ are the respective sound pressure levels of the tonal and non-tonal maskers on Bark scale $z(j)$. The term $SF[z(j), z(i)]$ is the spreading function that can be defined as follows:

$$10 \log_{10} SF = \begin{cases} 17dz - 0.4P_X[z(j)] + 11, & -3 \le dz < -1 \\ (0.4P_X[z(j)] + 6)\, dz, & -1 \le dz < 0 \\ -17dz, & 0 \le dz < 1 \\ -17dz + 0.15\, P_X[z(j)]\, (dz - 1), & 1 \le dz < 8 \end{cases} \tag{10}$$

where dz is the distance from maskee to masker $dz = z(i) - z(j)$. Δ_{TM} and Δ_{NM} are the offsets between the excitation pattern and actual masking threshold. The masking indices of tonal and non-tonal maskers can be specified as follows:

$$\Delta_{TM}[z(j)] = -6.025 - 0.275\ z(j) \tag{11}$$

$$\Delta_{NM}[z(j)] = -2.025 - 0.175\ z(j) \tag{12}$$

3.1.5. *Determination of the global masking threshold.* The global masking threshold is calculated by combining the individual masking threshold and absolute threshold of hearing, as shown in Figure 4. Practically, this threshold is an estimation on the concurrent masking effect for all spectral components. Since the mixture of masking is additive, the global masking threshold at frequency index i is calculated as follows:

$$L_G(i) = 10 \log_{10} \left[10^{\frac{ATH(i)}{10}} + \sum_{j=1}^{N_{TM}} 10^{\frac{L_{TM}[z(j),z(i)]}{10}} + \sum_{j=1}^{N_{NM}} 10^{\frac{L_{NM}[z(j),z(i)]}{10}} \right] \tag{13}$$

where $ATH(i)$ is the sound pressure level of the absolute threshold of hearing at frequency index i, N_{TM} and N_{NM} are the number of tonal and non-tonal maskers, and $L_{TM}[z(j), z(i)]$

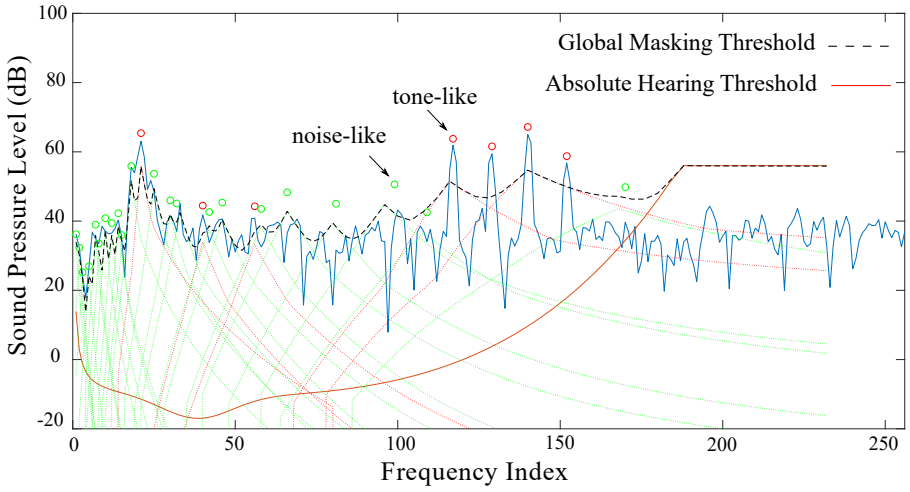

FIGURE 4. Global masking threshold

and $L_{NM}[z(j), z(i)]$ are the correspondent individual masking thresholds of tonal and non-tonal elements, respectively.

3.1.6. *Determination of minimum masking threshold in each sub-band.* The minimum masking threshold can be derived from the global masking threshold in Section 3.1.5. The global masking threshold is computed on a subset of samples over the frequency spectrum. Then, the MMT is calculated by mapping these spectral subsamples onto the n^{th} sub-bands ($1 \leq n \leq 32$), as illustrated in Figure 5.

$$L_{Min}(n) = \min_{f_{id}(i) \in subband\, n} L_G(i) \tag{14}$$

where $f_{id}(i)$ is the frequency index corresponding to the i^{th} subsample ($1 \leq i \leq 106$).

$$L_{MMT}(m) = L_{Min}(n) \quad m = [8(n-1)+1] : 8n \tag{15}$$

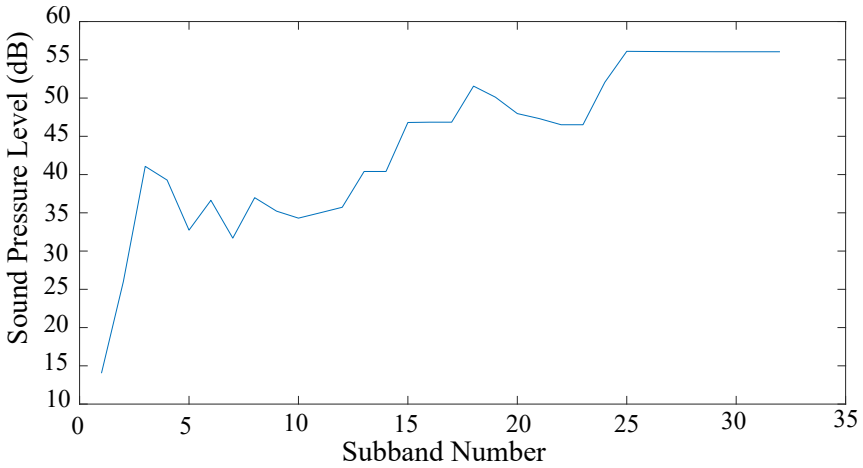

FIGURE 5. Minimum masking threshold for each sub-band

3.2. **Structure of MFCC and psychoacoustic model-based feature extraction.** The detailed process flow of conventional MFCC and the proposed feature extraction approaches are carried out referring to the block diagram illustrated in Figure 6. An acoustic front-end converts the speech signal into acoustic feature vectors, which provide useful and important spectrum information for the recognition process. The MFCC represent the short-time power spectrum of a speech signal, commonly used in ASR and speaker recognition systems. The frequency domain feature using the mel scale is considered more accurate than time domain features [4, 20].

The most common approaches in speech signal processing are performed using short-time analysis. Typically, the speech signal is divided into multiple time frames of 10∼30ms. Each frame underwent the windowing process with the aid of a Hamming window to hold the continuity at the edges. Then, the time domain data was transformed into the frequency domain by applying the discrete Fourier transform (DFT) to obtain the spectral information for each window. This magnitude spectrum contained considerable information that did not require for feature matching process. In fact, the feature matching algorithms were unable to distinguish the difference between frequency components spaced too closely together. Hence, the group of spectral bins was summed up to obtain the existent energy levels for each distinct frequency region.

A filterbank analysis using the mel scale aims to imitate the non-linear human ear perception of sound in nature, being more discriminative at lower frequencies and less discriminative at higher frequencies [21]. Triangular bandpass filters with a mel frequency

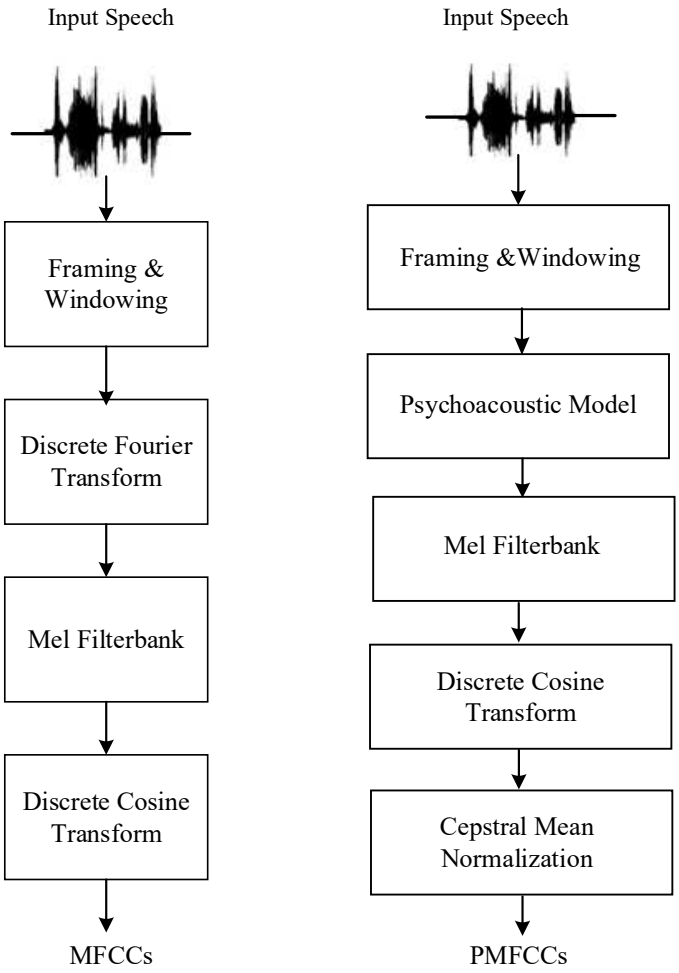

FIGURE 6. Process flow of conventional MFCC and the proposed psychoacoustical masking effect-based cepstral feature extraction (PMFCC)

warping scale are used to extract the spectral envelope, which is constituted using the dominant frequency components in the speech signal. After taking the mel filterbank analysis, the discrete cosine transform (DCT) is applied to transforming the log mel spectrum into time domain [22, 23]. Twenty filterbank energies are applied to the mel scale, and the first 13 coefficients are extracted as the cepstral features. MFCC can capture the main characteristics of phones in speech, and can approximate the human hearing system much more closely than other systems. However, MFCC are not very robust in adverse situations where the presence of background noise can affect and hamper the quality of results. Hence, recognition systems are usually normalized to lessen the influence of noise effects [24].

The psychoacoustic model is the study of sound perception, i.e., how humans respond to different frequencies, the masking effect, and the relationship between sound pressure levels and loudness. To hide watermark information, some audio watermarking techniques utilize the MMT from the psychoacoustic model. Based on this concept, this study proposed a psychoacoustical masking effect-based feature extraction (PMFCC) to extract the important features from the noise-corrupted speech signal. This model can analyze not only which frequency components lead more to the masking threshold, but also how much noise can be mixed in the utterance. Moreover, it can also shape the amplitude of the audio signal.

The basic algorithm follows these procedures: frame blocking, windowing, applying a psychoacoustic model, filterbank analysis, DCT and normalizing as the mean cepstral features. The final PMFCC is the amplitude of the resulting spectrum. After obtaining the magnitude spectrum from the DFT, the MMT is determined in each sub-band by applying the psychoacoustic model. The masking threshold is the limit for just noticeable distortion, which means any sounds or frequency components below this threshold are masked by the presence of masker. For this reason, the magnitude spectrum of each frame is compared with the value of the threshold. If the value of the spectrum is less than the MMTs, this spectrum value raised to the value of the minimum threshold. In this way, the modified spectrum can be calculated for each frame. Finally, these modified spectrum values are passed through a filterbank analysis. By applying this psychoacoustical masking effect into a conventional MFCC technique, the irrelevant feature components can be removed without any significant loss in perceived sound quality while also diminishing the noise effect from the speech signal.

4. **Speech Recognition Engine.** The speech recognition engine estimates the classes within the feature space from a training dataset after extracting the cepstral features from the speech signal. In this section, we will describe a statistical approach called the hidden Markov model (HMM). The approach is highly prized for its applications to ASR due to its mathematical structure, theoretical basis and computationally feasible use. The HMM is a model of the probability for a sequence of observable events and has a collection of states connected by transitions. The states are not visible directly, but each state s_i has a probability density p_i. $p(o|s_i)$ represents the probability density for an acoustic observation o in state s_i. Thus, the sequence of tokens rendered by an HMM can supply information about the sequence of hidden states [25]. An HMM can be characterized by the following elements [4, 26, 27]:

$Q = q_1\ q_2\ \ldots\ q_N$ = A set of states

$A = a_{01}\ a_{02}\ \ldots\ a_{n1}\ \ldots\ a_{nn}$ = Transition probability matrix; (a_{ij}) represents the probability of a transaction from state i to state j.

$O = o_1\ o_2\ \ldots\ o_N$ = A set of observations

$B = b_i(o_t)$ = A set of emission probabilities, where each is the probability of an observation o_t being generated from state i.

q_0, q_{end} = Non-emission states which are not associated with observations

In ASR, the hidden states can be phones or words. Each observation is a vector of real-valued features which indicate the spectrum information and energy of the speech signal at time t. The decoding process maps these sequences of acoustic information to phones or words. There are three fundamental problems that must be solved using HMMs [28, 29].

- *Computing the Likelihood:* Given an HMM $\lambda = (A, B)$ and a sequence of observation O, compute the likelihood $P(O|\lambda)$.
- *Finding the Hidden Sequence:* Given an HMM $\lambda = (A, B)$ and a sequence of observation O, find the hidden sequence of states $Q = q_1, q_2, q_3, \ldots, q_T$ that is most likely to generate O.
- *Estimating the Parameters:* Given a sequence of observation O and the set of possible states in an HMM, estimate the transition and emission probabilities A and B that are most likely to give O.

The likelihood $P(O|A, B)$ in an HMM from given observation sequence $O = (o_1, o_2, \ldots, o_T)$ and known transition and emission probabilities matrices A, B can be computed using the forward algorithm. That algorithm computes the observation probability $P(O|A, B)$ by summing over the probabilities of all possible hidden states Q that could generate the observation sequence. The Viterbi algorithm is most commonly used for decoding algorithms in HMM. It uses dynamic programming, like the forward algorithm, that can be applied to estimating the best-hidden state sequence. The Viterbi gives the state path through the HMM that assigns the maximum likelihood value to given observation sequence O. HMMs learn on data samples that include a sufficient number of speakers with Baum-Welch algorithm, which is similar to expectation-maximization (EM) algorithm when determining optimal HMM parameters for a speaker independent model. Expectation-maximization is an iterative algorithm that starts by randomly initializing the values for transition probabilities A and emission probabilities B of the HMM, then iteratively improving them. Each update iteration has an expectation phase and maximization phase [29, 30].

5. **Experimental Setup and Results.** This section will discuss a suite of experiments for our proposed work. In this study, eleven isolated words from the English vocabulary ("zero" to "nine" and "oh") were used for experiments. The spoken digit utterances were extracted from the TIDIGITS speech corpus of several thousand continuous digit utterances (available from the Linguistic Data Consortium) [31]. This experiment focused mainly on female speakers, and each digit was repeated twice by one hundred and fourteen female speakers. These isolated digits were recorded in a clean environment and encompassed 2,508 utterances (11 digits $\times$ 114 speakers $\times$ 2 times) for all digits. The authors mixed background noise into the clean isolated digit speech corpus to simulate real-world environments. This corpus covered five different types of noise, categorized as subway (atmosphere in moving subway train), babble (mixture of several voices), restaurant (atmosphere in a typical restaurant), car (inside a moving car) and street (atmosphere on a busy street) that were based on the constructions for AURORA noisy speech evaluation [32]. All the utterances were up-sampled to 44.1kHz. Finally, this dataset was constructed with a total of 12,540 utterances for all five different noise atmospheres at various SNR levels of 10dB, 5dB, 0dB, -5dB and -10dB that will be used in further experiments.

First of all, the conventional MFCC approach was applied to the noise-corrupted speech signal and the results were analyzed. The masking effect of the psychoacoustical model was then integrated into a conventional MFCC approach for robustness in recognition. This experiment was carried out on each noise situation and compared for recognition performance. The utterances of first fifty-seven speakers (1,254 utterances) worked as the training and the utterances of the remaining fifty-seven speakers utilized for evaluation. When using a conventional MFCC approach, the recognition accuracy (%) was significantly high, approximately 90.00% in an SNR of average (0-10)dB under subway noise, 90.56% under babble noise, 91.36% under car noise, 95.93% under street noise and 92.00% under restaurant noise. However, when more noise was introduced to the clean signal, the recognition performance degraded rapidly to 81.02%$\sim$63.07% for all five types of noise at the SNR of -5dB condition and 61.72%$\sim$46.33% in SNR of -10dB, as illustrated in Figure 7. Although the conventional MFCC can reduce the frequency information of speech utterance into a small number of coefficients, they function well in a clean environment. When speech is recorded under background noise, MFCC results will degrade. The detailed experimental results are shown in Table 1.

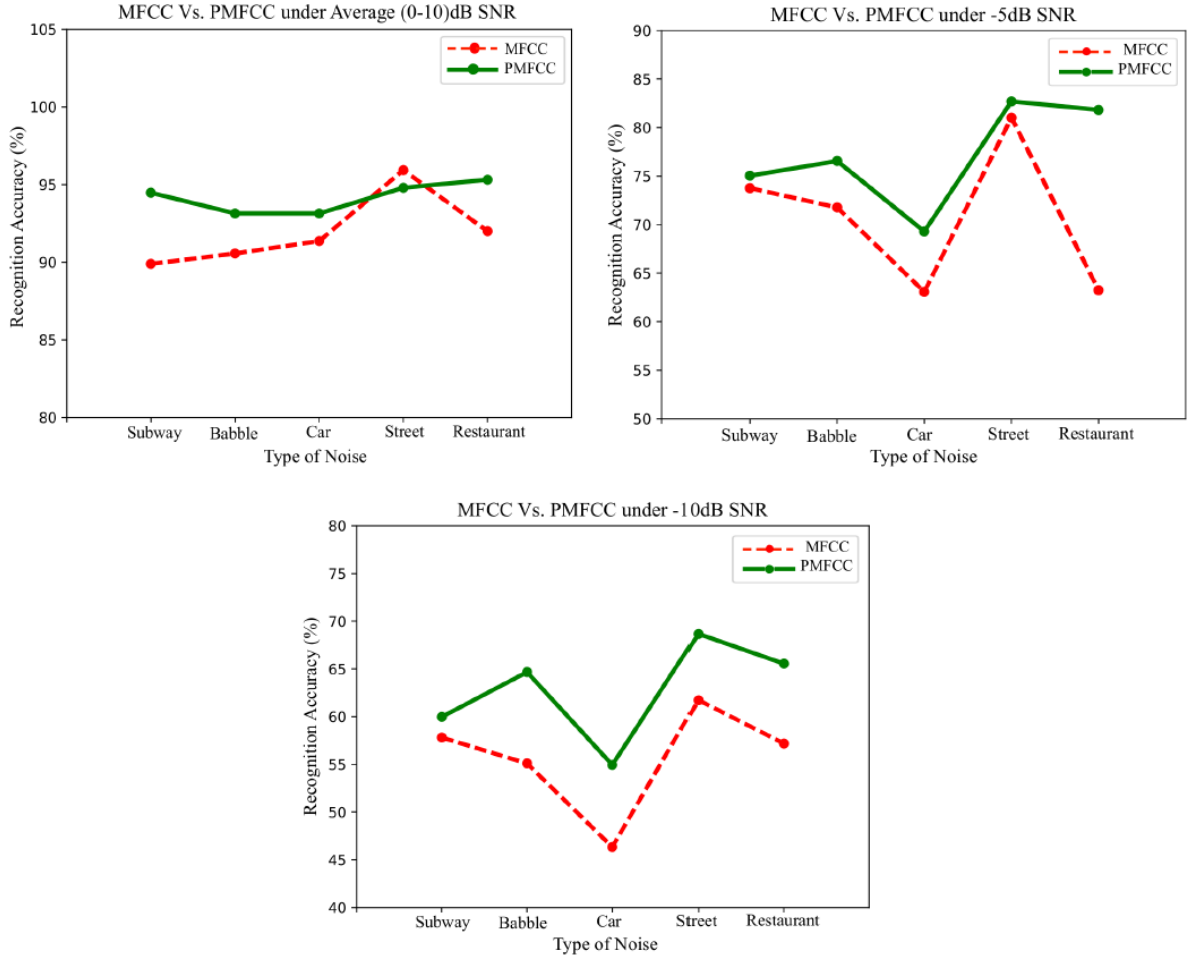

FIGURE 7. The recognition accuracy of conventional MFCC and the proposed PMFCC feature extraction under different noise situations and SNRs

TABLE 1. Recognition accuracy (%) of conventional MFCC features

Types	10dB	5dB	0dB	−5dB	−10dB
Subway	96.57	95.69	77.43	73.76	57.81
Babble	96.01	91.94	83.73	71.77	55.1
Car	96.96	93.22	83.89	63.07	46.33
Street	98.41	96.17	93.22	81.02	61.72
Restaurant	97.61	97.84	80.54	63.24	57.18

To overcome these phenomena, the auditory masking effect of the psychoacoustic model was investigated and integrated into conventional MFCC. This proposed method simulated the human perception of sound in the auditory system to obtain the robust features for noisy speech utterances. Compared with conventional MFCC, our PMFCC achieved better results in SNR of average (0-10)dB under all types of noise except street noise. The relative improvements were 4.9% in SNR of 0-10dB under subway noise, 2.8% under babble noise, 1.9% under car noise and 3.6% under restaurant noise, as shown in Figure 7. Furthermore, in terms of −5dB and −10dB SNR levels, the proposed PMFCC algorithm gave relatively good results under all types of noise. The relative improvements were 1.73%, 6.67%, 9.86%, 2.07% and 29.3% for subway, babble, car, street and restaurant noises, respectively, in SNR of −5dB. For the SNR −10dB condition, the relative

TABLE 2. Recognition accuracy (%) of psychoacoustical based (PMFCC) features

Types	10dB	5dB	0dB	−5dB	−10dB
Subway	96.89	95.45	91.07	75.04	59.97
Babble	96.65	94.34	88.44	76.56	64.67
Car	96.57	94.42	88.44	69.29	54.94
Street	97.69	94.66	92.03	82.70	68.66
Restaurant	98.00	96.25	91.70	81.82	65.55

improvements were 3.74%, 17.36%, 18.58%, 11.24% and 14.43% for all types of noise, respectively. Table 2 summarizes the detailed recognition performance with the use of PMFCC on different noise situations.

Additionally, the overall accuracy (%) of recognition was carried out to determine how the results will be yielded in all types of noise conditions and at various SNR levels. The improvement in accuracy when the auditory masking effect of the psychoacoustic model is integrated into a conventional MFCC approach is assured, as shown in the experimental results presented in Figure 8. According to these results, the proposed PMFCC algorithm achieved improvement of 0.051% and 0.052% at SNR 10dB and 5dB, respectively. However, the recognition of speech sample at SNR of 0dB, −5dB and −10dB performed better than the MFCC with a more significant improvement. Compared with the accuracy of the MFCC, the relative improvements were 7.85% in SNR of 0dB, 9.22% in −5dB and 12.81% in −10dB. It can be observed that the speech utterances under SNR −5dB and −10dB were very noisy, yet while the proposed algorithm still managed to improve the satisfying accuracy, the recognition rate was higher in more noisy condition compared to the MFCC.

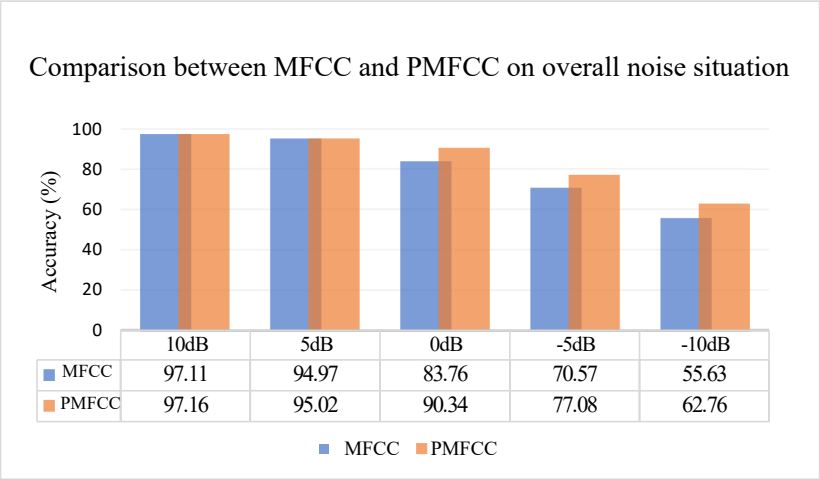

FIGURE 8. The results of overall noise validation with conventional MFCC and the proposed PMFCC at various SNR of 10dB, 5dB, 0dB, −5dB and −10dB

6. **Conclusions.** In this study, the auditory masking effect of the psychoacoustic model was integrated into conventional MFCC features and applied to developing a robust ASR system. Instead of removing the inaudible frequency components, the speech signals amplified to the limit for noticeable distortion or MMT. Using the auditory masking effect in feature extraction can reduce the complexity of the speech signal and minimize the irrelevant feature components without any significant loss in the perception of sound.

Consequently, it can also reduce the noise effect of the signal. Several works have been performed using a front-end algorithm on an isolated English digit dataset. The PMFCC has been shown to be more robust than a conventional MFCC approach in different noise environments and at various SNR levels. The obtained results assure that our proposed algorithm effectively improves recognition accuracy when speech is recorded under background noise. This present work was implemented based on a simple design using the simultaneous masking effect in feature extraction. The complete design of the psychoacoustic model will be developed into a speech recognition system in the future.

Acknowledgment. We would like to express our gratitude to the ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net), JICA for financial supporting this research. We also thank our colleagues from Gadjah Mada University, Indonesia, who assisted greatly with critical insights and valuable remarks for this research. Furthermore, we would like to thank Tetsuya Nakagoshi from Laboratory of Information Communication Networks, Graduate School of Information Science and Technology, Hokkaido University, Japan. This study is supported in part by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Scientific Research (B) (18H0321).

REFERENCES

[1] D. O'Shaughnessy, Automatic speech recognition: History, methods and challenges, *Pattern Recognition*, vol.41, no.10, pp.2965-2979, 2008.

[2] S. Yoshizawa, Y. Miyanaga, N. Wada and N. Yoshida, A lowpower LSI design of Japanese word recognition system, *Proc. of IEICE International Technical Conference on Circuits/Systems, Computers and Communications*, vol.1, no.1, pp.98-101, 2002.

[3] S. Karpagavalli and E. Chandra, A review on automatic speech recognition architecture and approaches, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol.9, no.4, pp.393-404, 2016.

[4] J. H. Martin and D. Jurafsky, Speech and language processing: An introduction to natural language processing, *Computational Linguistics and Speech Recognition*, Pearson/Prentice Hall, 2009.

[5] R. E. Gruhn, W. Minker and S. Nakamura, *Statistical Pronunciation Modeling for Non-native Speech Processing*, Springer Science & Business Media, 2011.

[6] S. Sahoo and A. Routray, MFCC feature with optimized frequency range: An essential step for emotion recognition, *International Conference on Systems in Medicine and Biology (ICSMB)*, pp.162-165, 2016.

[7] N. Wada, Y. Miyanaga, N. Yoshida and S. Yoshizawa, A consideration about an extraction of features for isolated word speech recognition in noisy environments, *ISPACS2002, DSP2002-33*, pp.19-22, 2002.

[8] P. V. Janse, S. B. Magre, P. K. Kurzekar and R. Deshmukh, A comparative study between MFCC and DWT feature extraction technique, *International Journal of Engineering Research and Technology*, vol.3, no.1, pp.3124-3127, 2014.

[9] T. Fux and D. Jouvet, Evaluation of PNCC and extended spectral subtraction methods for robust speech recognition, *The 23rd European Signal Processing Conference (EUSIPCO)*, pp.1416-1420, 2015.

[10] R. Hidayat et al., Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system, *The 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp.280-284, 2018.

[11] J. Qi, D. Wang, Y. Jiang and R. Liu, Auditory features based on gammatone filters for robust speech recognition, *International Symposium on Circuits and Systems (ISCAS)*, pp.305-308, 2013.

[12] P. Dai, F. Rudzicz, Y. Soon, A. Mihailidis and H. Ding, 2D psychoacoustic modeling of equivalent masking for automatic speech recognition, *Signal Processing*, vol.115, pp.9-19, 2015.

[13] P. Dai and Y. Soon, A temporal frequency warped (TFW) 2D psychoacoustic filter for robust speech recognition system, *Speech Communication*, vol.54, no.3, pp.402-413, 2012.

[14] B. Winduratna, FM analysis/synthesis-based audio coding, *Audio Engineering Society Convention*, vol.104, 1998.

[15] D. Naveen and A. Jhansi rani, Implementation of psychoacoustic model in audio compression using Munich and Gammachirp wavelets, *International Journal of Engineering Science and Technology*, vol.2, no.5, pp.1066-1072, 2010.
[16] H. K. Maganti and M. Matassoni, A perceptual masking approach for noise robust speech recognition, *Journal on Audio, Speech, and Music Processing*, 2012.
[17] S. Desai, P. D. Khandekar and K. J. Raut, 2-D psychoacoustic modeling for automatic speech recognition in noisy environment, *Conference on Advances in Signal Processing (CASP)*, pp.129-132, 2016.
[18] T. S. Gunawan, *Audio Compression and Speech Enhancement Using Temporal Masking Models*, Ph.D. Thesis, University of New South Wales, Sydney, Australia, 2007.
[19] Y. Lin and W. H. Abdulla, Principles of psychoacoustics, in *Audio Watermark*, Springer, 2015.
[20] W. Zhang and G. Li, The research of feature extraction based on MFCC for speaker recognition, *The 3rd International Conference on Computer Science and Network Technology (ICCSNT)*, pp.1074-1077, 2013.
[21] E. S. Wahyuni, Arabic speech recognition using MFCC feature extraction and ANN classification, *Proc. of the 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp.22-25, 2017.
[22] S. B. Magre, R. R. Deshmukh and P. P. Shrishrimal, A comparative study on feature extraction techniques in speech recognition, *International Conference on Recent Advances in Statistics and their Applications*, https://www.researchgate.net/publication/278549945.pdf, 2013.
[23] M. A. Imtiaz and G. Raja, Isolated word automatic speech recognition (ASR) system using MFCC, DTW & KNN, *Proc. of Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast)*, pp.106-110, 2016.
[24] K. Gupta and D. Gupta, An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system, *The 6th International Conference on Cloud System and Big Data Engineering*, pp.493-497, 2016.
[25] A. G. Adami, Automatic speech recognition: From the beginning to the Portuguese language, *The International Conference on Computational Processing of Portuguese (PROPOR)*, Porto Alegre, Rio Grande do Sul, 2010.
[26] F. Rosdi and R. N. Ainon, Isolated malay speech recognition using hidden Markov models, *International Conference on Computer and Communication Engineering (ICCCE)*, pp.721-725, 2008.
[27] R. Ranjan and R. K. Dubey, Isolated word recognition using HMM for Maithili dialect, *International Conference on Signal Processing and Communication (ICSC)*, pp.323-327, 2016.
[28] H. Z. Muhammad, M. Nasrun, C. Setianingsih and M. A. Murti, Speech recognition for English to Indonesian translator using hidden Markov model, *International Conference on Signals and Systems (ICSigSys)*, pp.255-260, 2018.
[29] M. Aymen, A. Abdelaziz, S. Halim and H. Maaref, Hidden Markov models for automatic speech recognition, *International Conference on Communications, Computing and Control Applications (CCCA)*, pp.1-6, 2011.
[30] R. E. Gruhn, W. Minker and S. Nakamura, Automatic speech recognition, in *Statistical Pronunciation Modeling for Non-Native Speech Processing*, Springer, 2011.
[31] R. Leonard, A database for speaker-independent digit recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84)*, vol.9, pp.328-331, 1984.
[32] D. Ellis, *Sound Examples for Projects*, http://www.ee.columbia.edu/∼dpwe/sounds/, 2002.