International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **1**, January **2011**

 ISSN 1349-4198
pp. 1–11

A NOVEL FRONTEND WITH FREQUENCY MASKING FOR AUTOMATIC SPEECH RECOGNITION

Ja-Zang Yeh, Bo-Feng Wu and Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
70 Lien-Hai Road, Kaohsiung, Taiwan
{ycc97m,cpchen}@mail.cse.nsysu.edu.tw
m983040035@student.nsysu.edu.tw

(Communicated by Chia-Ping Chen)

Abstract. *In this paper we investigate modifying the commonly-used feature extraction process for automatic speech recognition systems. A novel frequency masking curve, which is based on modeling the basilar membrane as a cascade system of damped simple harmonic oscillators, is used to replace the critical-band masking curve to compute the masking threshold. We mathematically analyze the coupled motion of the oscillator system (basilar membrane) when they are driven by short-time stationary (speech) signals. Based on the analysis, we derive the relation between the amplitudes of neighboring oscillators, and accordingly insert a masking module in the front-end signal processing stage to modify the speech spectrum. Evaluated on the Aurora 2.0 noisy-digit speech database, the proposed methodology shows significant improvements. In summary, we propose the novel model, make the detailed mathematical analysis, realize the masking-curve implementation, and achieve improved recognition accuracy in noisy environments.*

Keywords: frequency masking, noise robustness, speech recognition

1. **Introduction.** Most of the time our world presents us with a multiple of sounds simultaneously. We automatically accomplish the task of distinguishing each of the sounds and finding out the ones of greatest importance. Unless we want to hear to get the information of all sounds in detail, we probably do not consider all the sounds we do not hear in the course of a day. Statistical automatic speech recognition (ASR) systems show severe performance degradation when they are used in mismatched acoustic environments [1]. The recognition model learned from the training data simply does not generalize well to unseen data.

There are many approaches developed to improve this speech robust problem. Robustness techniques in general fall into many aspects either from feature domain or from filter domain or from their corresponding probability distributions. The feature domain, VoIP (Voice over Internet Protocol) is technology that helps people to communicate via voice using the IP protocol instead of traditional telephone lines. The computer extracts the feature of speech as same as the human ear to analysis the external voice singnal [2]. For the corresponding probability distributions of speech model process, it is very important step to take the parameter estimation [3] or perceptual speech [4] with effective algorithm. Besides, there are some research of ASR robust by observing how the emotion of human being react [5]. According to the volume, rate, pitch, pausing and silence attributes, we can make the best recognition result explicitly. In human speech, emotion plays a very important role for us to enhance the recognition accuracy.

As human speech recognition (HSR) significantly outperforms ASR in noisy environments [6], quite a few ideas stemming from the experimental findings on psycho-acoustics and auditory systems have been proposed to improve the noise-robustness of ASR systems [7, 8, 9, 10]. We were well-known the basilar membrane in our inner ear vibrates in response to sound. Low frequencies displace the basilar membrane much more: the distance from stapes is about 30mm at 25 Hz compared to 20mm at 800 Hz. Additionally, as frequency increases, the location of maximum displacement along the basilar membrane moves from the farthest section of the inner ear toward the middle ear. In this paper, we study methods based on the simultaneous masking (also known as the frequency masking) effect [11], referring to phenomenon that a tone masks (makes inaudible) another tone adjacent in frequency. Note that certain properties of the simultaneous masking effects, such as asymmetry and locality, have been further discovered in psychoacoustic experiments [12, 13, 14, 15].

The existence of the masking effect is arguably paradoxical. From an information-theoretic point of view, the frequency masking effect means that the neural signals collected at different positions in the basilar membrane are not independent, and renders the intrinsic information capacity of the neural channels sub-optimal. The frontend of an ASR system, on the other hand, does not have to have this "redundancy", and therefore it may achieve the optimal information rate given the spectral information of speech in principle.

We state our argument for the frequency masking mechanism from the perspective of recognition as follows. While the masking effects reduce the information for the discrimination between speech classes (e.g. the phones, words, or longer linguistic units), it also alleviates the distortion due to noises if they are present. That is, the masking effect can be viewed as a *preventative* mechanism to reduce the mismatch between the clean and noisy environments. Essentially, the phsyical effect of masking is to emphasize the high-energy part and de-emphasize the low-energy part of the speech spectrum. Since the high-energy (strong) part is less vulnerable to random noises than the low-energy (weak) part, such a emphasis/de-emphasis mechanism does make very good sense.

The organization of this paper is as follows. In Section 2, we describe the details of our implementation of the frequency masking effects in the ASR frontend. In Section 3, we define and analyze our string-of-oscillators model for the basilar membrane, and derive the coupled motion of the oscillators under the influence of external speech signal, which acts as the external driving force function to the oscillator system. The experimental results on the Aurora 2.0 database are presented and commented in Section 4. In Section 5, we draw our conclusions and outline our future works.

2. **The Implementation of the Frequency Masking Effect.**

How the frequency masking effects are implemented is very crucial in the resulting effectiveness in noise-robustness. The basic ideas of the scheme used in this paper are outlined as follows. The discrete spectral bins of a speech frame are treated as an array of simultaneous tones. These tones assume the roles of both maskers and probes, masking each other and being masked by each other. From the intensity levels (spectral energy) of these bins, we can compute an audible threshold curve. As a result of the masking effect, those bins with intensity levels under the threshold curve are not audible. For inaudible bins, we replace the raw intensity levels by the threshold values. Note that it is clear that the degree of smoothness of the modified spectrum in our method is better than the alternative of resetting the under-threshold values to 0.

The main steps in our implementation for the frequency masking effect are described below, which are similar to [15].

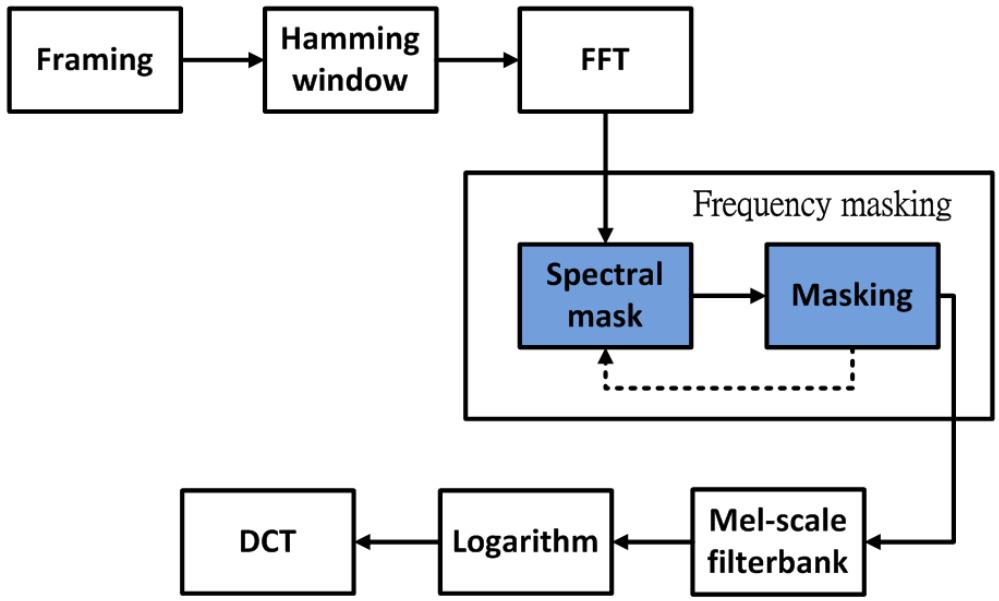

FIGURE 1. Block diagram of an MFCC front-end inserted with the proposed novel frequency masking module. Note that the frequency masking operation can be iterated.

1. **warping power spectrum**: One class of critical band [16] scales is called bark-frequency scale which included 24 critical bands. The power spectrum of N-point discrete Fourier transform (DFT) is warped according to the bark-frequency Ω. The formula as follow:

$$\Omega(\omega) = 6 \ln \left\{ \frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right\}. \tag{1}$$

where ω is the angular frequency in rad/s. This special Bark-hertz transformation which is for improvement over the linear scale of the conventional LP analysis, yet more accurate approximation of the nonlinear frequency scale of hearing of human. That is,

$$p(\Omega_n) \leftarrow p[n], \quad \Omega_n = \Omega \left(\frac{n f_s}{N} \right), \ n = 0, \ldots, \frac{N}{2}. \tag{2}$$

where $p[n]$ is the DFT power spectrum and f_s is the sampling frequency.

2. **computing audible threshold**: To compute the audible threshold curve, a frequency masking curve is required. Frequently, the well-known critical-band masking curve [17]

$$\psi(\Omega) = \begin{cases} 0, & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)}, & -1.3 \le \Omega \le -0.5, \\ 1, & -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)}, & 0.5 \le \Omega \le 2.5, \\ 0, & \Omega > 2.5 \end{cases} \tag{3}$$

is used. Note that one of the main point of this work is to replace $\psi(\Omega)$ by novel functions based on our analysis of the assumed physical models. Given $\psi(\Omega)$ and $p[n]$, the discrete masking threshold is approximated by a normalized convolution in

the warped-frequency domain. That is,

$$M(\Omega_n) = \frac{1}{\sum_l \psi_{n,l}} \sum_{l=0}^{N/2} p(\Omega_l)\psi_{n,l}, \tag{4}$$

where

$$\psi_{n,l} = \psi(\Omega_n - \Omega_l). \tag{5}$$

3. **modifying spectrum**: Those values under the audible threshold curve is replaced by the threshold values, while those values above the curve remain unchanged. That is,

$$\hat{p}[n] \leftarrow \hat{p}(\Omega_n) = \max\left\{p(\Omega_n), M(\Omega_n)\right\}. \tag{6}$$

The modified spectrum $\hat{p}[n]$ is used in subsequent processing in the ASR frontend.

The block diagram is shown in Figure 1. Note that the frequency masking stage can be applied iteratively. That is, we let

$$p[n] \leftarrow \hat{p}[n], \tag{7}$$

and repeat the three steps iteratively. (7) does lead to performance improvements, as will be shown in Section 4.

3. **Analysis of the Proposed Model.** It is seen in Section 2 that a masking curve is required for our proposed implementation of the frequency masking effect. The critical-band masking curve (3) is based on the experimental discovery of psychoacoustics [18, 19]. Contrarily, in this paper, a novel masking curve is derived based on our analysis of a physical model for the basilar membrane [20]. We model the basilar membrane as a cascade system of simple harmonic oscillators and each oscillator is driven by a force function which is related to the speech spectrum. This is illustrated in Figure 2.

This model assumption is consistent with the well-known *tonotopic* property of the basilar membrane, that the position of the maximum displacement in the basilar membrane depends on the spectral content of the inducing speech signal. An empirical formula for this dependency has been observed by Bèkèsy [21, 22]. It is also consistent with the physiology of the inner ear. Essentially, the basilar membrane is stiff and narrow at the base (near *stapes* and the oval window), loose and wide at the apex. To accommodate such physical variation, we allow the masses (denoted by m_i) and the spring constants (denoted by k_i) of the oscillators (denoted by o_i) to be dependent on the longitudinal position (i.e., along the cochlea). In addition, damping coefficients (denoted by γ_i) are introduced since the motion of the basilar membrane is influenced by the liquids, *endolymph* and *perilymph*, in the two tubes (*scala media* and *scala tympani*) it separates.

In order to clearly describe our ideas and approaches based on the proposed model, we distinguish between two parts of the overall motions of the oscillators.

1. According to the tonotopic property of the basilar membrane, an oscillator in the model is basically not responsive to all spectral components. Instead, it mainly responds to the component with frequency closest to its resonant frequency. We will call this main part of response the *primary* response.
2. However, since the segments of the basilar membrane modeled by oscillators are physically connected, the motion of an oscillator is bound to be influenced by the surrounding oscillators. This part of response is called the *secondary* response.

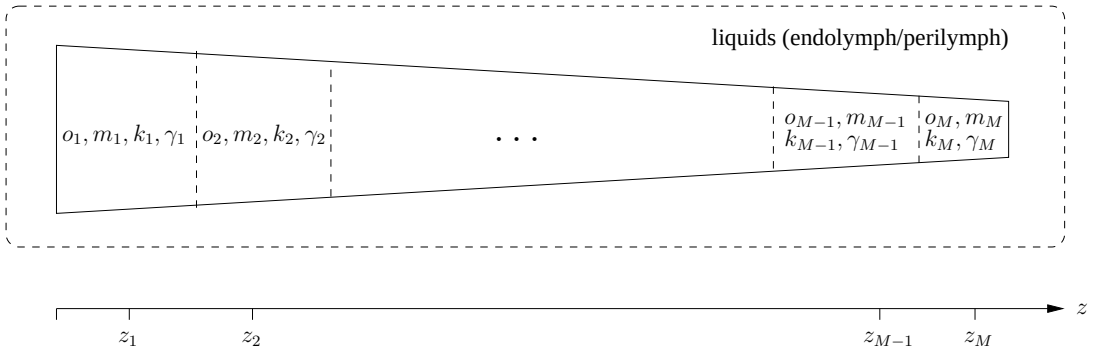

FIGURE 2. Modeling the basilar membrane as a cascade system of simple harmonic oscillators. k_i, m_i, γ_i are the spring constant, the mass, and the damping coefficient of the i^{th} oscillator o_i, respectively. z is the distance from the stapes of the middle ear.

The *amplitude* is approximated by the sum of the primary and the secondary responses. Without loss of generality, we can write the amplitude of oscillator o_i as

$$A_i = A_{i\leftarrow i} + \sum_{i' \neq i} A_{i\leftarrow i'}, \quad i = 1, \ldots, M, \tag{8}$$

where the primary response of o_i is denoted by $A_{i\leftarrow i}$, the secondary response of o_i due to oscillator $o_{i'}$ is denoted by $A_{i\leftarrow i'}$, and the amplitude of o_i is denoted by A_i. For the secondary response $A_{i\leftarrow i'}$, we call o_i the target oscillator and $o_{i'}$ the source oscillator.

(8) is our fundamental equation for the coupled motions of the oscillators. We next describe how to compute both $A_{i\leftarrow i}$ and $A_{i\leftarrow i'}$.

3.1. **The Primary Response and the Spectrum.** The oscillating motion of the basilar membrane from its equilibium position is driven by the air pressure waves caused by the speech or other sound signals. Using the short-time stationary assumption, we denote the driving force for the modeled oscillators as time-varying short-time stationary signal $f(t;\tau)$, where τ is the frame index, and t is the continuous time. Since $f(t;\tau)$ is of finite duration, it can be treated as one period of a periodic function. According to the Fourier analysis, the periodic extension of $f(t;\tau)$, denoted by $\tilde{f}(t;\tau)$, can be decomposed by the sinusoidal functions

$$\tilde{f}(t;\tau) = \sum_n c_n(\tau) \cos\left(2\pi n F t + \theta_n(\tau)\right), \tag{9}$$

where W is the period and $F \triangleq \frac{1}{W}$ is the corresponding fundamental frequency.

The decomposition of $\tilde{f}(t;\tau)$ given by (9) is indeed related to the discrete power spectrum $p[n]$ obtained by the DFT. The discrete frequency of the n^{th} freqeuncy bin corresponds to continuous frequency ω_n by

$$\omega_n = \left(\frac{n}{N}\right) 2\pi f_s, \quad n = 0, \ldots, \frac{N}{2}, \tag{10}$$

where f_s is the sampling frequency. Since there are N samples in W, we also have

$$2\pi n \frac{f_s}{N} = 2\pi n \frac{1}{W} = 2\pi n F, \quad n = 0, \ldots, \frac{N}{2}. \tag{11}$$

That is, the DFT power spectrum $p[n]$ is related to the magnitude of the n^{th} sinusoid in (9) since both are related to the same linear frequency $\omega_n = 2\pi n F$. Therefore, we assume

that the magnitude c_n is proportional to the square root of $p[n]$

$$c_n \propto \sqrt{p[n]} \quad \Rightarrow \quad c_n = b\sqrt{p[n]}, \quad n = 0, \ldots, \frac{N}{2}, \tag{12}$$

where b is a constant. Moreover, we approximate the resonant frequency of oscillator o_i by its natural resonant frequency $\omega_{0,i}$ without damping. Since $\omega_{0,i}$ is inversely proportional to its mass m_i,

$$\omega_{0,i}^2 = \frac{k_i}{m_i}, \tag{13}$$

we have

$$m_i \propto \frac{1}{\omega_i^{*2}} \quad \Rightarrow \quad m_i = \frac{d}{\omega_i^{*2}}, \quad i = 1, \ldots, M, \tag{14}$$

where d is a constant.

In our earlier work (please refer to [23] for more details), we show that the primary response is given by

$$A_{i \leftarrow i} = \frac{c_{r(i)}}{m_i} \frac{1}{|\omega_{0,i}^2 - \omega_i^{*2} + j2\gamma_i\omega_i^*|}, \quad i = 1, \ldots, M, \tag{15}$$

where m_i is the mass of oscillator o_i, $\omega_{0,i}$ is the natural resonant frequency of oscillator o_i, ω_i^* is the resonant frequency of oscillator o_i, γ_i is the damping coefficient of oscillator o_i, and $r(i)$ is the component in the driving force function with frequency closest to ω_i^*, i.e.

$$r(i) = \arg\min_n \ |\omega_n - \omega_i^*|, \quad i = 1, \ldots, M, \tag{16}$$

and $c_{r(i)}$ is the corresponding magnitude. It has been derived that the resonant frequency for oscillator o_i is

$$\omega_i^* = \sqrt{\omega_{0,i}^2 - 2\gamma_i^2}, \quad i = 1, \ldots, M, \tag{17}$$

so we can rewrite the primary response (15) as

$$A_{i \leftarrow i} = \frac{c_{r(i)}}{m_i} \frac{1}{\sqrt{4\gamma_i^2(\omega_i^{*2} + \gamma_i^2)}}, \quad i = 1, \ldots, M. \tag{18}$$

With (12) and (14), (18) can be re-written as

$$A_{i \leftarrow i} = \frac{b}{d} \frac{\sqrt{p[r(i)]} \cdot \omega_i^{*2}}{\sqrt{4\gamma_i^2(\omega_i^{*2} + \gamma_i^2)}} = u \cdot g_i \cdot \sqrt{p[r(i)]}, \quad i = 1, \ldots, M, \tag{19}$$

where the parameters are define as

$$u \triangleq \frac{b}{d}, \tag{20}$$

and

$$g_i \triangleq \frac{\omega_i^{*2}}{\sqrt{4\gamma_i^2(\omega_i^{*2} + \gamma_i^2)}}, \quad i = 1, \ldots, M. \tag{21}$$

3.2. **The Coupling Between Oscillators.** As a first-order approximation, we assume that the secondary response of the target oscillator is proportional to the amplitude of the source oscillator. To be specific, we assume that

$$A_{i \leftarrow i'} = \alpha_{ii'} A_{i'}, \quad i, i' = 1, \ldots, M, \quad i' \neq i, \tag{22}$$

where $\alpha_{ii'}$, $i \neq i'$ are called the *coupling* coefficients which quantify the coupling effect. With (22), the total response A_i of oscillator o_i can be written as

$$A_i = A_{i \leftarrow i} + \sum_{j \neq i} A_{i \leftarrow j} = A_{i \leftarrow i} + \sum_{j \neq i} \alpha_{ij} A_j, \quad i = 1, \ldots, M. \tag{23}$$

Given (23), we need to come up with the proportional constants α_{ij}. In principle, one can work out the equations and express α_{ij} in terms of the parameters of the model, such as the magnitudes c_i of the signal, the masses m_i, the spring constants k_i, and the damping conditions of the oscillators. Since it is reasonable to assume that the secondary effects tend to be strong between neighbors and fades away with distance, we model the coupling constants as follows in this study [24].

- Rectangular. This scheme sets the coupling coefficients to 1 within a range in agreement with the critical band. That is,

$$\alpha_{ij} = \begin{cases} 0, & \Omega_i - \Omega_j < -1.3 \\ 1, & -1.3 \leq \Omega_i - \Omega_j \leq 2.5 \\ 0, & \Omega_i - \Omega_j > 2.5 \end{cases} \tag{24}$$

- Triangular. This scheme sets the coupling coefficients based on a triangular function within a range in agreement with the critical band. That is,

$$\alpha_{ij} = \begin{cases} 0, & \Omega_i - \Omega_j < -1.3 \\ \frac{\Omega_j - \Omega_i}{1.3}, & -1.3 \leq \Omega_i - \Omega_j < 0 \\ \frac{\Omega_i - \Omega_j}{2.5}, & 0 < \Omega_i - \Omega_j < 2.5 \\ 0, & \Omega_i - \Omega_j > 2.5 \end{cases} \tag{25}$$

- Standard Normal. This scheme uses the standard normal distribution. That is,

$$\alpha_{ij} = \frac{1}{\sqrt{2\pi}} e^{\frac{(i-j)^2}{2}}. \tag{26}$$

- Gaussian. This scheme is similar to (26), but it uses a Gaussian distribution with zero mean and a variance depending on the critical bandwidth. That is,

$$\alpha_{ij} = \frac{1}{\sqrt{2\pi}} e^{\frac{(i-j)^2}{2\sigma_i^2}}, \tag{27}$$

  where σ_i is given by

$$\sigma_i = \frac{n_i}{10}, \tag{28}$$

  and n_i is the number of oscillators within the critical band surrounding oscillator o_i in the Bark scale, i.e.,

$$n_i = \sum_{\{j \mid -1.3 \leq \Omega_i - \Omega_j \leq 2.5\}} 1. \tag{29}$$

Combining (19) for the primary response and α_{ij} from (24) to (27) for the secondary response, we can express the amplitude of the oscillators according to (23) using the parameters we have defined. Collecting the amplitudes into a vector, we have a consistency equation

$$\mathcal{A} = \mathcal{A}_p + \mathbf{M}\mathcal{A}, \tag{30}$$

where $\mathcal{A}$ is the amplitude vector, $\mathcal{A}_p$ is the primary response vector, and $\mathbf{M}$ is a matrix whose off-diagonal entries are the coupling coefficients α_{ij}. The solution is

$$\mathcal{A} = (I - \mathbf{M})^{-1}\mathcal{A}_p. \tag{31}$$

In our implementation, the amplitude of oscillator o_i is further normalized by the sum of coefficients, i.e.,

$$\tilde{A}_i = \frac{A_i}{1 + \sum_{j \neq i} \alpha_{ij}}, \quad i = 1, \ldots, M. \tag{32}$$

Note that the parameter u is cancelled out in the normalization step (32), so the relationship between the normalized magnitude $\tilde{A}_i$ and the input power spectrum can be written as

$$\bar{p}[r(i)] = \left(\frac{\tilde{A}_i}{g_i} \right)^2 . \tag{33}$$

Finally, the masked spectrum is computed by

$$\check{p}[n] = \max \{p[n], \bar{p}[n]\} . \tag{34}$$

4. Experiments.

4.1. Experimental Setup. The proposed frequency masking curve is evaluated on the Aurora 2.0 noisy-digit speech database [25]. The front-end feature vector consists of the log energy and 12 mel-frequency cepstral coefficients (MFCCs) for the static features, as well as the corresponding dynamic velocity and acceleration features. In the back-end recognizer, each whole-word digit model consists of 16 states, and the state-dependent observation density is a Gaussian mixture model with 3 mixture components. The 3-state silence model has 6 Gaussians per state. The 1-state short-pause model is tied to the middle state of silence model.

In our experiments we set ω_i to be the resonant frequency ω_i^* of the ith oscillator,

$$r(i) = i, \quad \omega_i \approx \omega_i^*, \tag{35}$$

and the damping coefficient is assumed to be

$$\gamma_i = 0.1\ \omega_i. \tag{36}$$

TABLE 1. *Percentage word accuracy rates of the Aurora 2.0 clean-training tasks with the* $0 - 20$ *dB SNR test data, using the proposed method. Here the average (Avg.) is weighted by the data set sizes* $2 : 2 : 1$. *The relative improvements (rel. imp.) is based on the word error rates.*

	Set A	Set B	Set C	Avg.	rel. imp.
baseline	61.3	55.8	66.1	60.1	=
CMS	66.2	70.8	64.9	67.8	19.3
CMS+CBMC	68.1	72.2	67.1	69.6	23.8
CMS+HOMC	68.9	72.6	68.2	70.2	25.9
CMS+COM_r	66.4	71.2	64.9	68.0	19.9
CMS+COM_t	68.2	72.4	67.4	69.7	24.2
CMS+COM_s	66.9	71.6	65.6	68.5	21.2
CMS+COM_g	67.3	71.7	66.0	68.8	21.9

4.2. Results and Discussion. Table 1 lists the recognition accuracies of the mismatched (clean-training) Aurora tasks averaged over the noisy test data with signal-to-noise ratios (SNR) ranging from 0 to 20 dB. The **baseline** experiments use the raw features extracted by the Aurora 2.0 standard frontend. The commonly used post-processing scheme of the cepstral mean subtraction (**CMS**) [26] achieves 19.3% relative improvement over the baseline. The frequency masking implementation via the critical-band masking curve (3) as described in Section 2 is denoted by **CBMC**. The implementation via our previous proposed masking curve, the harmonic oscillator masking curve as presented in [23], is denoted by **HOMC**. The current implementation of our proposed model as analyzed in Section 3.2 is called causal oscillator masking (COM). The results where the coupling coefficients α_{ij} are given in (24) - (27) are denoted respectively by

- **COM_r** for the rectangular coefficients;
- **COM_t** for the triangular coefficients;
- **COM_s** for the standard normal coefficients;
- **COM_g** for the Gaussian coefficients;

Combined with **CMS**, **CBMC** achieves a relative improvement of 23.8% over the baseline, while **COM_t** achieves 24.2%, which outperforms **CBMC**. Note that **HOMC** achieves 25.9% and also outperforms **CBMC**.

TABLE 2. *Percentage word accuracy rates of the Aurora 2.0 clean-training tasks with the* $0-20$ *dB SNR test data with iterative frequency masking using the different CAMC.*

	no. iter.	Set A	Set B	Set C	Avg.	rel. imp.
CMS	=	66.2	70.8	64.9	67.8	19.3
CMS+CBMC	5	70.5	73.0	69.9	71.2	28.3
CMS+HOMC	4	71.2	73.6	71.3	72.2	30.3
CMS+COMC_r	4	70.7	73.2	72.4	72.0	30.0
CMS+COMC_t	5	70.0	72.4	71.2	71.2	27.9
CMS+COMC_s	10	67.6	72.2	66.3	69.2	22.8
CMS+COMC_g	10	70.4	73.4	69.7	71.5	28.5

Table 2 lists the recognition accuracies for iterative frequency masking operation, as specified in (7), using **CBMC** and **COM**. One can see that the performance is further improved. For **CBMC**, the optimal improvement of 28.3% is reached with 5 iterations. For **COM**, the optimal improvement of 30.0% is reached with 4 iterations in the case of **COM_r**. The other cases also show improvements. For **HOMC**, the optimal improvement of 30.3% is reached with 4 iterations. The results show that the mismatch between the clean and the noisy features can be reduced with iterative masking.

TABLE 3. *Percentage word accuracy rates of the Aurora 2.0 clean-training tasks with the* $0-20$ *dB SNR test data, using masking implementations combined with AFE. Here the average (Avg.) is weighted by the data set sizes* $2:2:1$*. The relative improvements (rel. imp.) is based on the word error rates.*

	Set A	Set B	Set C	Avg.	rel. imp.
baseline	61.3	55.8	66.1	60.1	=
AFE	87.5	87.0	85.6	86.9	67.2
AFE+CBMC	88.0	87.2	86.5	87.4	68.5
AFE+HOMC	87.3	86.4	85.3	86.5	66.3
AFE+COM_r	87.8	87.1	86.2	87.2	68.0
AFE+COM_t	88.0	87.3	86.6	87.4	68.5
AFE+COM_s	87.6	87.1	85.9	87.1	67.6
AFE+COM_g	87.8	87.1	86.1	87.2	68.0

Table 3 lists the recognition accuracies for our proposed method combine with the advanced front end (**AFE**) [27, 28], using **CBMC**, **HOMC** and **COM**. One can see that the performance of **COM** and **CBMC** are better than the **baseline** and **AFE**. Furthmore, when combined with AFE, the **COM** implementation of masking effect outperforms **HOMC**.

5. **Conclusion and Future Work.** In this paper we investigate a novel front-end with frequency masking effects to improve the noise-robustness of ASR systems. We model the basilar membrane as a cascade system of simple harmonic oscillators. Based on an analysis for the motion of the oscillators under the influence of speech signal, we design a frequency masking method which is used in modifying the speech spectrum. Evaluation on the Aurora 2.0 database shows that the proposed frequency masking implementation can outperform the well-known critical-band masking curve. Moreover, iterative frequency masking operation can lead to further improvement.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] X. Wang, J. Lin, and Y. Sun, "Applying feature extraction of speech recognition on voip auditing," *International Journal of Innovative Computing, Information and Control*, vol. 5, no. 7, pp. 1851–1856, July 2009.

[3] J. Dong, X. Wei, Q. Zhang, and L. Zhao, "Speech enhancement algorithm based on higher-order cumulants parameter estimation," *International Journal of Innovative Computing, Information and Control*, vol. 5, no. 9, pp. 2725–2733, September 2009.

[4] Y. Jiao, L. Ji, and X. Niu, "Perceptual speech hashing and performance evaluation," *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 3, pp. 1447–1458, March 2010.

[5] T. Pao, Y. Chen, and J. Yeh, "Emotion recognition and evaluation from mandarin speech signals," *International Journal of Innovative Computing, Information and Control*, vol. 4, no. 7, pp. 1695–1709, July 2008.

[6] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[8] D. S. Kim, S. Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.

[9] L. Turicchia and R. Sarpeshkar, "A bio-inspired companding strategy for spectral enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 243–253, 2005.

[10] K. Paliwal, B. Shannon, J. Lyons, and K. Wójcicki, "Speech-Signal-Based Frequency Warping," *IEEE Signal Processing Letters*, vol. 16, no. 4, p. 319, 2009.

[11] E. Zwicker, "Masking and psychological excitation as consequences of ear's frequency analysis," *Frequency Analysis and Periodicity Detectionin Hearing*, 1970.

[12] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, no. 1, pp. 47–65, 1940.

[13] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.

[14] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation." *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.

[15] K. K. Paliwal and B. T. Lilly, "Auditory masking based acoustic front-end for robust speech recognition," in *Proceedings of IEEE TENCON*, vol. 1, 1997, pp. 165–168.

[16] D. D. Greenwood, "Auditory masking and the critical band," *the Acoustical Society of America*, vol. 33, pp. 484–502, October 1961.

[17] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[18] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, December 1979.

[19] M. A. Krasner, *Digital encoding of speech and audio signals based on the perceptual requirements of the auditory system.* PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1979.

[20] C. J. Plack and A. J. Oxenham, "Basilar-membrane nonlinearity and the growth of forward masking," *the Acoustical Society of America*, vol. 103, pp. 1598–1608, October 1998.

[21] G. Von Békésy, *Experiments in hearing*. McGraw-Hill New York, 1960.
[22] L. E. Kinsler, *Fundamentals of acoustic*. John Wiley and Sons, 1982.
[23] J. Z. Yeh and C. P. Chen, "Auditory front-ends for noise-robust automatic speech recognition," *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing*, 2010.
[24] T. E. Lee, G. Refael, O. Kogan, J. L. Rogers, and M. C. Cross, "Universality in the one-dimensional chain of phase-coupled oscillators," *The American Physical Society*, vol. 80, pp. 1–13, October 2009.
[25] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSA ITRW ASR2000*, September 2000.
[26] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
[27] ETSI Standard, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," *ETSI ES 202 050*, 2007.
[28] J. Ramrez, J. C. Segura, C. Bentez, A. D. L. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, October 2004.