Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 402–413

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A temporal frequency warped (TFW) 2D psychoacoustic filter for robust speech recognition system

Peng Dai *, Ing Yann Soon

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 19 May 2011; received in revised form 7 October 2011; accepted 11 October 2011
Available online 20 October 2011

Abstract

In this paper, a novel hybrid feature extraction algorithm is proposed, which implements forward masking, lateral inhibition, and temporal integration with a simple 2D psychoacoustic filter. The proposed algorithm consists of two key parts, the 2D psychoacoustic filter and cepstral mean variance normalization (CMVN). Mathematical derivation is provided to show the correctness of the 2D psychoacoustic filter based on the characteristic functions of masking effects. The effectiveness of the proposed algorithm is tested on the AURORA2 database. Extensive comparison is made against lateral inhibition (LI), forward masking (FM), CMVN, RASTA filter, the ETSI standard advanced front-end feature extraction algorithm (AFE), and the temporal warped 2D psychoacoustic filter. Experimental results show significant improvements from the proposed algorithm, a relative improvement of nearly 46.78% over the baseline mel-frequency cepstral coefficients (MFCC) system in noisy conditions.

Keywords: Automatic speech recognition; 2D mask; Simultaneous masking; Temporal masking; Temporal frequency warping; Temporal integration

1. Introduction

Noise robustness has always been a hot topic in speech processing. Automatic speech recognition (ASR) system can work very well with clean speech, where the recognition rates can reach 99%. However, with noise added in, the performance of ASR systems falls dramatically. Therefore, the development of noise robust speech recognition algorithm is very important. Human auditory system can work properly in very adverse situations such as environmental noise, channel distortion and speaker variability (Hermansky, 1998). Therefore, the idea of analyzing and modeling human auditory system is a logical approach to improve the performance of ASR systems.

For human, hearing is not simply a mechanical phenomenon of wave indexing. It is more like a sensory and perceptual process. Actually, before reaching the human brain, speech has already been processed by the human auditory system, which means the signal that reaches our brain is not the physical signal produced by the sound source. So the basic idea of this paper is to recognize speech based on the 'new' version of speech. The science of studying how humans perceive sounds including relationships between sound pressure level and loudness, human response to different frequencies, and a number of masking effects is called psychoacoustics (Gold and Morgan, 2000). Previous research shows that successful application of human auditory property can greatly improve the performance of ASR system. The mel-frequency cepstral coefficients (MFCC) method can serve as a good example. In MFCC, an auditory-based warping of the frequency axis, called critical-band, is implemented which is so successful that MFCC has become a standard speech feature for ASR system (Davis and Mermelstein, 1980).

In order to get the above mentioned 'new' version of speech which reaches human brain, the psychoacoustic phenomenon that needs more attention is the masking effect. Masking effect is the phenomenon which describes

* Corresponding author.
E-mail addresses: daip0001@e.ntu.edu.sg, daipengmay@gmail.com (P. Dai), eiysoon@ntu.edu.sg (I.Y. Soon).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.10.004

how a clearly audible sound is influenced by another sound. Masking effects can be classified as simultaneous or temporal depending on the dimension it works in. The one describing how two simultaneously occurring sounds affect each other is simultaneous masking. Similarly, the one describing the masking effect generated from another sound different in time is temporal masking. In temporal masking, a signal can be masked by both the one earlier and later in time, known as forward masking and backward masking respectively (Shamma, 1985). Forward masking can generate much stronger masking effect than backward masking, so research is more focused on forward masking (Jesteadt et al., 1982; Oxenham and Plack, 2000).

In our previous work (Dai and Soon, 2009), an original 2D psychoacoustic filter is proposed, which makes use of the temporal frequency property of speech and yields reasonable results. Then temporal warping is adopted to improve the performance of 2D psychoacoustic filter (Dai and Soon, 2010). In this paper, the 2D psychoacoustic filter is further extended, and new insights are provided to show the theoretical basis of the proposed algorithm. It is developed based on psychoacoustics, including simultaneous masking, temporal masking, and temporal integration. Mathematical derivation is provided to show the validity of 2D psychoacoustic filter. Another important and unique aspect of the proposed algorithm is the extension of masking effect to multi-dimensional area. A novel mathematical function is provided to calculate the so called 'diagonal parameter'. Verification tests are carried out on the AURORA2 database. Extensive comparison is made against MFCCs, forward masking (FM), lateral inhibition (LI), cepstral mean and variance normalization (CMVN), the ETSI standard advanced front-end feature extraction algorithm (AFE), RASTA filter, and the temporal warped 2D filter (Dai and Soon, 2009, 2010; ETSI, 2007; Hermansky and Morgan, 1994). Significant improvements can be seen from the experimental results.

2. 2D psychoacoustic modeling

2.1. Masking effects

2.1.1. Simultaneous masking

Psychoacoustics is the scientific study of sound perception. Simultaneous masking studies how the amount of masking changes with frequency. The amount of masking refers to the difference in power level between actual speech and what is perceived by the human. Usually it is defined as a subtractive energy level. Simultaneous masking has been widely used in many areas such as audio compression and speech enhancement (Cheng and O'Shaughnessy, 1991). However, its implementation in speech recognition is relatively new. Lateral inhibition (LI) is one effective approach for implementing simultaneous masking (Shamma, 1985; Luo et al., 2008).

In neurobiology, lateral inhibition is used to measure the capacity of an excited neuron to reduce the activity of its neighbors. It is a common phenomenon in sensory reception of biological systems. In speech processing, it stands for how two speech signals with different frequencies occurring at the same time affects each other. LI helps to sharpen spectral changes. Fig. 1 shows the sketch of the psychoacoustic data given in Houtgast's paper (Cheng and O'Shaughnessy, 1991; Houtgast, 1972). The curve describes how the amount of LI masking changes with frequency.

Assume that a speech signal, $x(t)$, is corrupted by noise, $n(t)$, resulting in a noisy speech, $s(t)$. The relationship is given by,

$$s(t) = x(t) + n(t) \tag{1}$$

where t is the time index.

The speech signal is cut into frames and transformed into frequency domain using DFT. Then Eq. (1) becomes

$$S(f,t) = X(f,t) + N(f,t) \tag{2}$$

where f is the frequency index; $S(f,t)$, $X(f,t)$, $N(f,t)$ refer to the time frequency domain signal of noisy speech, clean speech, and additive noise.

By assuming the additivity on the powers of the components in the frequency domain (Gold and Morgan, 2000; Ishizuka and Nakatani, 2010), the power spectrum of the noisy speech is given by

$$|S_{f,t}|^2 = |X_{f,t}|^2 + |N_{f,t}|^2 \tag{3}$$

Let $M_{LI}(f)$ represent the lateral inhibition masker. The lateral inhibition masker is modeled to satisfy the following constraint (Cheng and O'Shaughnessy, 1991; Dai and Soon, 2010),

$$\int_{-\infty}^{\infty} M_{LI}(f)df = 0 \tag{4}$$

The LI masker is very effective in removing stationary noise. Assuming $|N(f,t)|^2$ in Eq. (3) is stationary, after applying the given lateral inhibition masker in Eq. (3) to the noisy speech power spectrum, the processed speech, $|\hat{S}(f,t)|^2$, can be calculated by

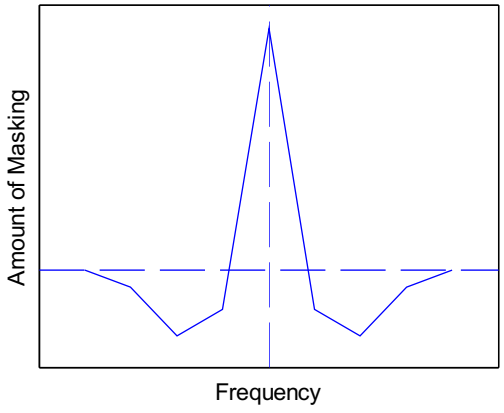


Fig. 1. Sketch of psychoacoustic data (Cheng and O'Shaughnessy, 1991).

$$
\begin{aligned}
|\hat{S}(f,t)|^2 &= \int_{-\infty}^{\infty} |S(f,t)|^2 M_{LI}(f)df \\
&= \int_{-\infty}^{\infty} |X(f,t)|^2 M_{LI}(f)df \\
&\quad + \int_{-\infty}^{\infty} |N(f,t)|^2 M_{LI}(f)df \\
&= |\hat{X}(f,t)|^2 + |N(f,t)|^2 \int_{-\infty}^{\infty} M_{LI}(f)df \\
&= |\hat{X}(f,t)|^2
\end{aligned} \tag{5}
$$

where $|\hat{X}(f,t)|^2$ stands for the clean speech that is processed by the LI filter.

Since it is widely known that speech signal is highly non-stationary, the above mentioned LI filter has little impact on speech. However, from Eq. (5) it can be found out that the stationary noise is completely removed.

2.1.2. Temporal masking

Temporal masking occurs when a masker sound makes inaudible other sounds which are present immediately preceding or following the masker. It consists of forward masking and backward masking. Forward masking (FM) describes the phenomenon that over short durations the effective dynamic range of the human auditory system depends on the characteristics of the previous sound. The characteristic curve of forward masking is frequency dependent. Besides, like many other masking effects it is different among different people. Backward masking shares similar characteristic but the masker comes after the signal rather than before it. The characteristic curve of temporal masking is shown in Fig. 2, which shows how the amount of masking changes with time. The detailed equations will be given in Section 2.3.1.

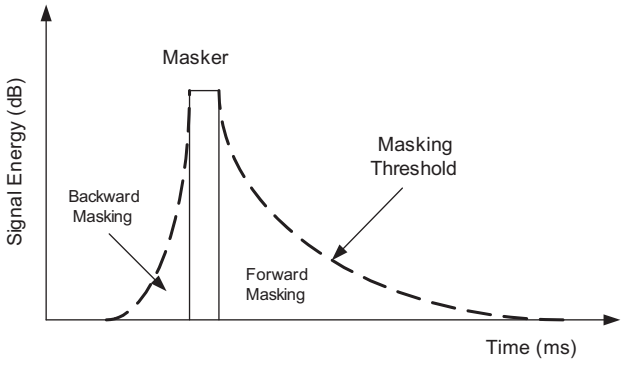


Fig. 2. Temporal masking.

For a given speech utterance, speech is more likely to possess stronger power than noise. As is shown in Fig. 3, despite the presence of noise the high power regions still belong to speech in the speech active period. Temporal masking is actually a psychoacoustic phenomenon that describes how a stronger sound masks another sound earlier or later in time. When it comes to speech, speech is usually stronger than noise. Therefore, noise in return becomes the one that is masked. In other words, temporal masking helps to remove noise, in the human hearing process.

2.2. Theoretical analysis

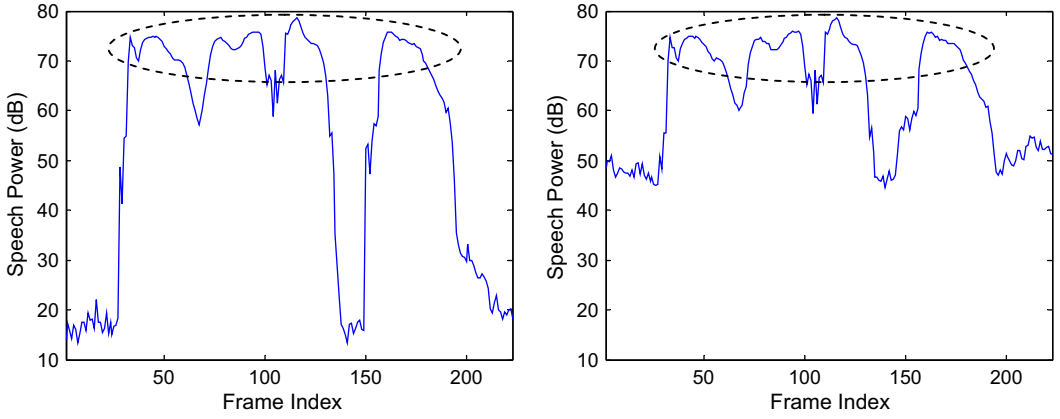
2.2.1. Temporal masking and simultaneous masking

As described in the previous section, there are mainly two kinds of masking effect, temporal masking and simultaneous masking. During speech feature extraction, the speech signal is cut into frames and transformed into time frequency domain, which is denoted as $S_{f,t}$.

According to Jesteadt et al. (1982), forward masking (FM) can be modeled by the following equation.

$$M_{fm} = F_{fm}(S_{f,t}, \Delta t), \quad \Delta t > 0 \tag{6}$$

where M_{fm} is the amount of masking, $F_{fm}(\cdot)$ denotes the FM characteristic function, and Δt is the signal delay. Eq. (6) describes how a speech signal, $S_{t,f}$, affects other acoustic signal that is Δt later in time.



(a) Clean Speech (b) SNR 10 dB (Subway Noise)

Fig. 3. Speech power temporal distribution for '3Z82' from the AURORA2 database.

Besides, there is also the so called backward masking (BM) effect.

$$M_{bm} = F_{bm}(S_{f,t}, \Delta t), \quad \Delta t < 0 \tag{7}$$

Strictly speaking, Eqs. (6) and (7) describes how a speech signal $S_{t,f}$ affects another signal. However, in order to imitate how human auditory system acquires the speech signal, the more important part is to study how a specific speech signal is affected by other signals 'nearby'.

For a sound element, $S_{f,t}$, in time frequency domain, theoretically, it is affected by all the sound near it in time, as is shown in Fig. 4.

Therefore, the overall influence on the target element is a joint effect of all the elements 'before' and 'after' it in time.

$$M^t_{total} = \sum_{\Delta t=1}^{+\infty} F_{fm}(S_{f,t}, \Delta t) + \sum_{\Delta t=-\infty}^{-1} F_{bm}(S_{f,t}, \Delta t) \tag{8}$$

Similarly, for frequency masking, the overall influence on the target element is also a joint effect of all the elements 'before' and 'after' it in frequency.

$$M^f_{total} = \sum_{\Delta f=1}^{+\infty} F_{li}(S_{f,t}, \Delta f) + \sum_{\Delta f=-\infty}^{-1} F_{li}(S_{f,t}, \Delta f) \tag{9}$$

2.2.2. Overall joint effect

In time frequency domain, all speech elements are influenced by the surrounding elements. It means for a speech signal $S_{t,f}$ it is affected by all the other speech signals within a certain range, $\{Y_{f,t} | T_1 \leqslant t \leqslant T_2, F_1 \leqslant f \leqslant F_2\}$. T_1, T_2, F_1, and F_2 are the effective range of FM, BM and LI. Fig. 5 gives a better view of the above mentioned effect. All the speech elements within the rectangle contributes to the final masking effect on $S_{t,f}$.

As is shown in Fig. 5, the slashed area can be modeled using temporal masking, Eq. (8), and the crossed area can be modeled by frequency masking (LI), Eq. (9). Finally, remaining area in the rectangle can be modeled

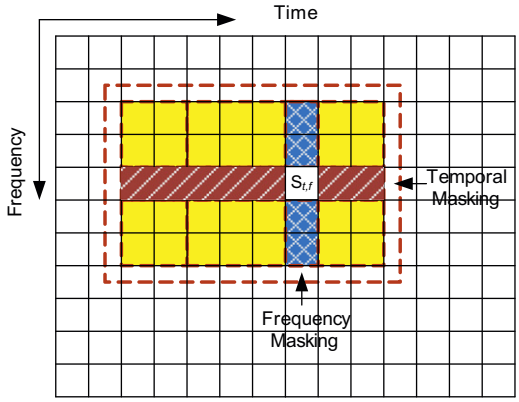


Fig. 5. Joint masking effect.

using the frequency dependent forward masking function from Jesteadt's paper (Jesteadt et al., 1982), which will be discussed in detail in later parts. Therefore, the overall joint masking effect can be described as follows:

$$\begin{aligned} M_{total} = & \sum_{\Delta t=1}^{\infty} F_{fm}(S_{f,t}, \Delta t) + \sum_{\Delta t=-\infty}^{-1} F_{bm}(S_{f,t}, \Delta t) \\ & + \sum_{\Delta f=1}^{\infty} F_{li}(S_{f,t}, \Delta f) + \sum_{\Delta f=-\infty}^{-1} F_{li}(S_{f,t}, \Delta f) \\ & + \sum_{\substack{\Delta t \\ \Delta t \neq 0}} \sum_{\substack{\Delta f \\ \Delta f \neq 0}} F_{diag}(S_{f,t}, \Delta t, \Delta f) \end{aligned} \tag{10}$$

In the proposed method, the function which decides the relationship between the masking effect and the initiating sound is simplified to linear function. Then taking into account the effective range of masking effect, we get

$$M_{total} = \sum_{\Delta t=-T_1}^{T_2} \sum_{\Delta f=-F_1}^{F_2} a_{\Delta f, \Delta t} S_{f-\Delta f, t-\Delta t} - a_{0,0} S_{f,t} \tag{11}$$

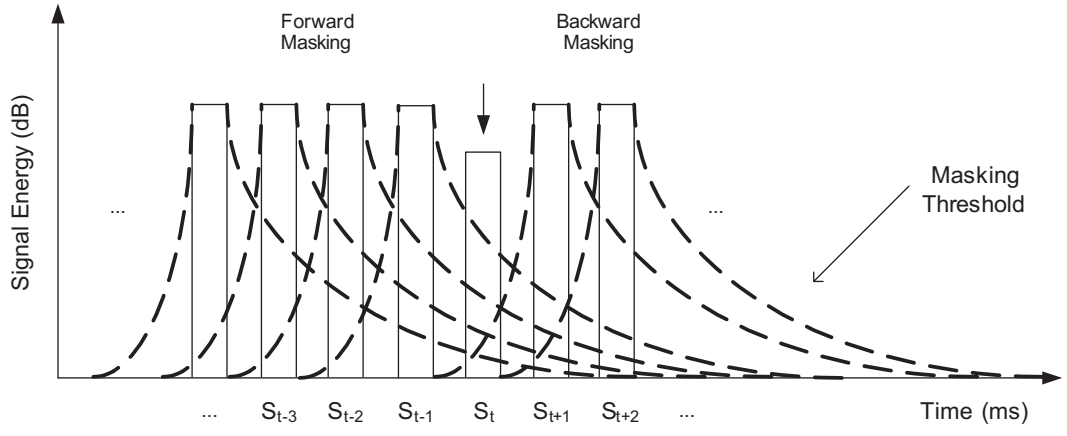


Fig. 4. Joint masking effect.

$$M_{total} = S_{f,t} \otimes \begin{bmatrix} \mathbf{0}_{F_1-F_2,T_1-T_2} & & & & \mathbf{0}_{F_2-F_1,T_2+T_1+1} & & & \\ & a_{F_2,T_2} & & & a_{F_2,0} & & & a_{F_2,-T_1} \\ & & \ddots & & \vdots & & \ddots & \\ & & & a_{1,1} & a_{1,0} & a_{1,-1} & & \\ \mathbf{0}_{F_1+F_2+1,T_1+T_2-1} & a_{0,T_2} & \cdots & a_{0,1} & 0 & a_{0,-1} & \cdots & a_{0,-1} \\ & & & a_{-1,1} & a_{-1,0} & a_{-1,-1} & & \\ & & \ddots & & \vdots & & \ddots & \\ & a_{-F_1,T_2} & & & a_{-F_1,0} & & & a_{-F_1,-T_1} \end{bmatrix} \tag{12}$$

Noting that $T_1 > T_2$ and $F_1 > F_2$ according to the effective range of masking effects.

where $\otimes$ stands for convolution; $\mathbf{0}$ denotes a zero matrix.

Eq. (12) gives the total masking effect introduced by the surrounding speech elements. Because masking effects tend to weaken speech, Eq. (12) describes how surrounding speech elements weaken the center element. Therefore, the target speech element, $S_{t,f}$, changes to

$$\begin{aligned} S'_{f,t} &= S_{f,t} - M_{total} = \sum_{\Delta t=-T_1}^{T_2} \sum_{\Delta f=F_1}^{F_2} (-a_{\Delta t,\Delta f}) S_{f-\Delta f,t-\Delta t} \\ &= S_{f,t} \otimes Mask_{f,t} \end{aligned} \tag{13}$$

where

$$Mask_{\Delta f,\Delta t} = \begin{bmatrix} \mathbf{0}_{F_1-F_2,T_1-T_2} & & & & \mathbf{0}_{F_2-F_1,T_2+T_1+1} & & & \\ & -a_{F_2,T_2} & & & -a_{F_2,0} & & & -a_{F_2,-T_1} \\ & & \ddots & & \vdots & & \ddots & \\ & & & -a_{1,1} & -a_{1,0} & -a_{1,-1} & & \\ \mathbf{0}_{F_1+F_2+1,T_1+T_2-1} & -a_{0,T_2} & \cdots & -a_{0,1} & 1 & -a_{0,-1} & \cdots & -a_{0,-T_1} \\ & & & -a_{-1,1} & -a_{-1,0} & -a_{-1,-1} & & \\ & & \ddots & & \vdots & & \ddots & \\ & -a_{-F_1,T_2} & & & -a_{-F_1,0} & & & -a_{-F_1,-T_1} \end{bmatrix} \tag{14}$$

The designing of parameters in Eq. (14) will be discussed in the following sections.

2.3. 2D psychoacoustic filter design

2.3.1. Temporal frequency modeling

Because backward masking is very weak compared with FM, it is not taken into consideration in the design of 2D psychoacoustic filter. According to Jesteadt et al. (1982), forward masking follows the following equation

$$M_{fm} = a[b - \log(\Delta t)](S_{f,t} - c) \tag{15}$$

where a, b, and c are parameters listed in Table 1, and Δt is signal delay.

Because c is relatively small compared with $S_{t,f}$, Eq. (15) is further simplified to

$$M_{fm} = a[b - \log(\Delta t)]S_{f,t} \tag{16}$$

Fig. 6 shows the characteristic curve of forward masking, which describes how the amount of masking changes with time.

The sampling frequency of the AURORA2 database is 8 kHz. The frame length is chosen to be 200 (according to the Aurora2 database demo scripts). After transformation into time–frequency domain, the frequency bins that are covered are 40∼4000 Hz. Theoretically, forward masking parameters are different for different frequency. However, strictly follow the above mentioned property will make the proposed filter too complicated for practical

Table 1
Forward masking parameters (Jesteadt et al., 1982).

Frequency	Parameter values		
	a	b	c
125	0.140	5.583	5.36
250	0.334	2.697	1.02
1000	0.372	2.278	7.20
4000	0.351	2.252	4.19

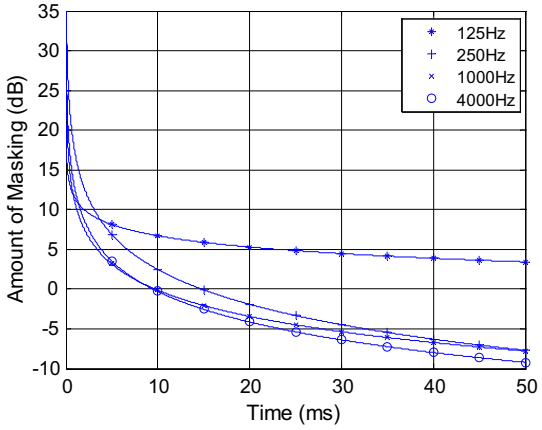


Fig. 6. Characteristic curve of FM.

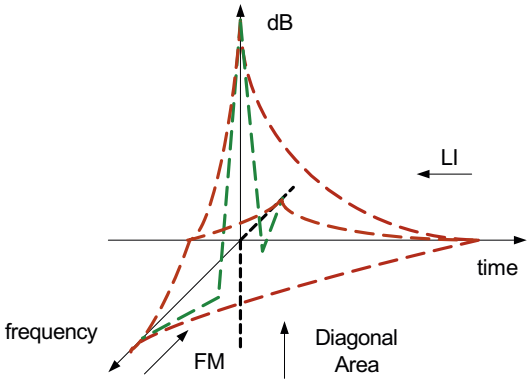


Fig. 7. Masking effect.

implementation. Therefore, for simplicity the parameters are chosen to be fixed for different frequencies. In this paper, the 4000 Hz parameter set is adopted for filter design.

Similarly, for lateral inhibition the following function can be used to calculate the amount of masking

$$M_{li} = \alpha_{li} S_{f,t} \tag{17}$$

where α_{li} is the LI parameter given in Table 2.

Theoretically LI parameters should change with frequency. For simplicity only the simplified version of LI is adopted. The LI parameters from Cheng's paper are directly incorporated in this paper. Following analysis steps in Section 3.1, the masking parameter can be achieved. It has to be noted that the LI parameters from Cheng's paper is symmetric. Nevertheless, in speech processing lateral inhibition is well known to be asymmetrical (Houtgast, 1972). Therefore, frequency warping is incorporated to generate a set of better parameters. The warping algorithm in our previous paper (Dai and Soon, 2010) is adopted to calculate the warped LI parameter. The detailed parameters are given in Table 2.

2.3.2. Full parameter design

The diagonal parameters are one of the most important characteristics of the proposed algorithm. Most of the current algorithms focus on the relationship between the two parameters. For example, as is shown in Fig. 7 FM focuses on the relationship of masking effect over different time difference, and LI represents the relationship between masker level and frequency. However, it has to be noted that even two signal with different frequency still possess masking effect.

A two step process is developed to calculate the amount of masking between two speech signals with different frequency specified by $\Delta t, f_1$ and f_2. Firstly, the forward masking result for Δt is calculated using FM characteristic function. Then the amount of masking from FM is used to generate the LI result. Eq. (18) gives a better view of the above mentioned process. Similarly, the diagonal result can also be calculated by the LI-first process shown in Eq. (19) and Fig. 8 (Route 2).

Route 1:

$$M_{f,t}^{diag} = F_{li}[F_{fm}(S_{f,t}, \Delta t), \Delta f] \tag{18}$$

Route 2:

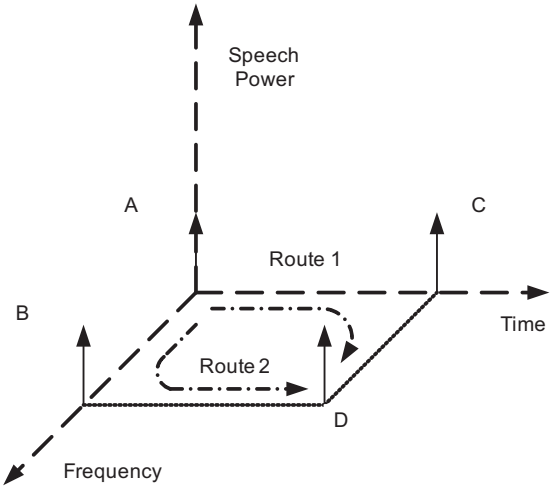


Fig. 8. Design of 2D psychoacoustic filter.

Table 2
Design of LI parameters.

Frequency index	−3	−2	−1	0	1	2	3	4	5
LI filter	−0.07	−0.27	−0.16	1	−0.16	−0.27	−0.07		
Masking parameter	0.07	0.27	0.16	0	0.16	0.27	0.07		
Warped parameter			0.0137	0	0.0914	0.1757	0.2386	0.2129	0.0986
Warped LI filter			−0.0137	1	−0.0914	−0.1757	−0.2386	−0.2129	−0.0986

$$M_{f,t}^{diag} = F_{fm}[F_{li}(S_{f,t}, \Delta f), \Delta t] \tag{19}$$

Eqs. (18) and (19) can be viewed as going along two different routes in the time frequency domain. As is shown in Fig. 8, route 1 corresponds to Eq. (18) while route 2 shows Eq. (19). The diagonal parameter can be calculated by

$$M_{t,f}^{diag} = F_{fm}[F_{li}(S_{f,t})] = F_{li}[F_{fm}(S_{f,t})] = \alpha_{li}\alpha_{fm}S_{f,t} \tag{20}$$

where α_{fm} is the FM parameter.

Then taking into account of the physical meaning of the proposed 2D filter

$$M_{t,f}^{diag} = -\alpha_{li}\alpha_{fm}S_{f,t} \tag{21}$$

It has to be noted that under the above mention assumption Route 1 and Route 2 should achieve the same results.

$$F_{fm}[F_{li}(S_{f,t})] = F_{li}[F_{fm}(S_{f,t})] \tag{22}$$

where F_{fm} stands for the FM characteristic function and F_{li} is the LI characteristic function.

FM and LI share similar form in the characteristic function is the sufficient condition of Eq. (22)'s validity. It is well known that FM and LI share similar property (Houtgast, 1972; Jesteadt et al., 1982; Park and Lee, 2003). Based on the above mentioned fact, the two-step calculation approach for diagonal parameters is very reasonable since it follows the basic properties of psychoacoustics. The detailed design process mainly consists of two steps. Firstly, the basic LI and FM parameters are given by Eqs. (16) and (17). Then the diagonal parameters can be calculated based on Eq. (21), which are just the initial parameters given in Appendix A (Table A1).

2.3.3. Temporal integration

Masking effect mainly describes how the amount of masking is affected by time, Δt, and frequency, f. However, further study shows that the duration of speech signal also greatly affects the total masking, which is the so called temporal integration (TI). According to Oxenham's paper (Oxenham, 2001), when signal duration increases there is great decrease in the mean thresholds (or the amount of masking). For example, in the experimental data presented by Oxenham shown in Fig. 9 (Oxenham, 2001, Fig 1, pp. 735), at an offset of 9 ms mean thresholds decreased by nearly 14 dB as the signal duration increased from 2 to 7 ms. In other words, in Oxenham's experiment an increase of 5 ms (7 ms − 2 ms) result in 14 dB decrease in the amount of masking. Noting that at duration of 2 ms the amount of masking is about 56 dB, it can be found out that the amount of masking drops about 25% due to a slight increase (5 ms) in the signal duration.

As is known to all, since speech has active/non-active periods its power is more concentrated in certain areas, both larger in energy and longer in duration. Therefore, temporal integration tends to impose more influence on speech. Then Eq. (12) should be modified to

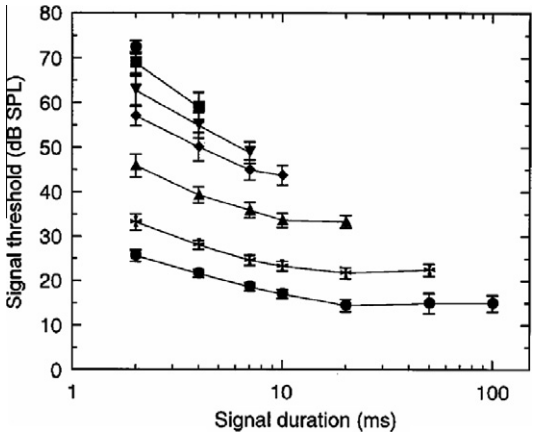


Fig. 9. TI experimental results (Oxenham, 2001).

$$\tilde{M}_{total} = \begin{cases} \tilde{M}_{total}, & nonspeech \\ \tilde{M}_{total} - M_{TI}, & speech \end{cases} \tag{23}$$

where M_{TI} stands for the decrease of masking caused by temporal integration. Then

$$\tilde{Mask}_{\Delta f,\Delta t} = \begin{cases} Mask_{\Delta f,\Delta t}, & nonspeech \\ Mask_{\Delta f,\Delta t} + Mask_{TI}, & speech \end{cases} \tag{24}$$

where $Mask_{TI}$ denotes the filter for TI.The relationship between noisy speech power, clean speech power, and noise power given in Eq. (3). Therefore, after processed by the new 2D filter $\tilde{M}ask_{f,t}$

$$\begin{aligned} |\tilde{S}_{f,t}|^2 &= |S_{f,t}|^2 \otimes \tilde{M}ask_{\Delta f,\Delta t} \\ &= (|X_{f,t}|^2 + |N_{f,t}|^2) \otimes \tilde{M}ask_{\Delta f,\Delta t} \\ &= |X_{f,t}|^2 \otimes (Mask_{\Delta f,\Delta t} + Mask_{TI}) + |N_{f,t}|^2 \\ &\quad \otimes Mask_{\Delta f,\Delta t} \end{aligned} \tag{25}$$

After processed by the proposed filter, the clean speech becomes $|X_{f,t}|^2 \otimes (Mask_{t,f} + Mask_{TI})$ denoted as $|\tilde{X}_{f,t}|^2$. Then

$$\begin{aligned} \frac{|\tilde{X}_{f,t}|^2}{|\tilde{S}_{f,t}|^2} &= \frac{|X_{f,t}|^2 \otimes (Mask_{\Delta f,\Delta t} + Mask_{TI})}{|X_{f,t}|^2 \otimes (Mask_{\Delta f,\Delta t} + Mask_{TI}) + |N_{f,t}|^2 \otimes Mask_{\Delta f,\Delta t}} \\ &> \frac{|X_{f,t}|^2}{|X_{f,t}|^2 + |N_{f,t}|^2} \end{aligned} \tag{26}$$

From Eq. (26), it can be easily shown that temporal integration can help to increase the SNR of the noisy speech. In practical implementation, a perfect design of a filter following the above mentioned rule is too complicated. The key difficulty is the choice of M_{TI}. It is determined by too many parameters, while experimental data are very limited. Therefore, approximation has to be made to achieve an easier implementation.

Now that M_{TI} is dependent on whether the frame is speech active, in our present implementation it is fixed to be

$$M_{TI} = \alpha S_{t,f} \tag{27}$$

Then Eq. (14) becomes

$$Mask_{\Delta f,\Delta t} = \begin{bmatrix} \mathbf{0}_{F_1-F_2,T_1-T_2} & & & & \mathbf{0}_{F_2-F_1,T_2-T_1+1} & & & \\ & -a_{F_2,T_2} & & & -a_{F_2,0} & & & -a_{F_2,-T_1} \\ & & \ddots & & \vdots & & \ddots & \\ & & & -a_{1,1} & -a_{1,0} & -a_{1,-1} & & \\ \mathbf{0}_{F_1+F_2+1,T_1+T_2-1} & -a_{0,x_2} & \cdots & -a_{0,1} & 1+\alpha & -a_{0,-1} & \cdots & -a_{0,-T_1} \\ & & & -a_{-1,1} & -a_{-1,0} & -a_{-1,-1} & & \\ & & \ddots & & \vdots & & \ddots & \\ & -a_{-F_1,T_2} & & & -a_{-F_1,0} & & & -a_{-F_1,-T_1} \end{bmatrix} \tag{28}$$

By doing this, not only temporal integration is successfully incorporated, but another problem is also satisfyingly solved. That is the filter generated by Eq. (21) given in Appendix A (Table A1) does not satisfy the well known energy consistency property. The overall sum of the total masking effect coefficients is nearly −5, while a neutral filter should counts to one. Mathematically, this process can also be viewed as a weighted combination of the original signal and the filtered result. According to experiments, 5 is an optimal value for $Mask_{0,0}$. Since CMVN is implemented in the proposed algorithm, the scaling problem can be left to CMVN. Therefore, in implementation the 2D psychoacoustic filter is not normalized. Appendix A (Table A2) gives the center enhanced 2D psychoacoustic filter (without normalization). The proposed 2D psychoacoustic filter enhances the high frequencies and helps to sharpen the spectral peaks so as to improve the performance of the recognition system.

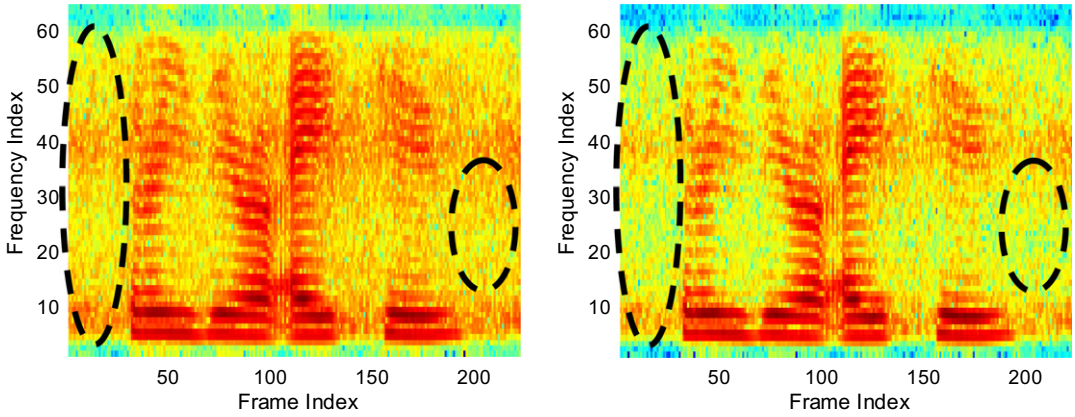
2.4. Summary

The proposed algorithm intends to implement forward masking (FM), lateral inhibition (LI) and temporal integration (TI) with simply a 2D psychoacoustic filter. The mathematical derivations and detailed design procedure are discussed in previous sections. In order to give a better view of the proposed algorithm, an example is given to show what spectral changes it has on speech articulations. Fig. 10 shows the effect of the proposed filter. The sample speech is selected from the AURORA2 database. It can be easily found out that the proposed 2D psychoacoustic filter can successfully remove noise.

3. Design of experiments

3.1. Database

In order to show the performance of the proposed algorithm evaluation tests are carried out on the AURORA2 database. The AURORA2 database is modified from the original TIDigits by down sampling to 8 kHz (Hirsch and Pearce, 2000). There are four different sets of speech data, Set A, Set B, Set C, and TRAIN set. In the TRAIN set, two different series of speech data are provided, which are used for the so called clean and noisy training. In the clean training condition, only the clean speech is adopted for HMM training. In the multi training condition, noisy speech together with clean speech is used for HMM training.



(a) Noisy speech (b) Processed

Fig. 10. Digit string '3Z82' from theAURORA2 database.

Sets A and B are the noisy testing data. Eight different kinds of noise (babble, restaurant, train, car, street, airport, exhibition, and subway) recorded from real environment are added at different SNRs from −5 dB to 20 dB with 5 dB step size. For Set C, it is designed to test the performance of ASR system in a telecommunication terminal. The speech data in this set are all processed by the MIRS (modified intermediate reference system) filter, which is used to simulate the characteristics of a telecommunication terminal. Two noises (subway and street) are then added.

3.2. System description

The proposed algorithm is developed based on the mel-frequency cepstral coefficients (MFCCs). It is a type of representation of the audio frame. The MFCC program from VoiceBox Toolkit (Brookes, 1997) is adopted to be the baseline system. The baseline results are based on the standard 13 mel-frequency cepstral coefficients (MFCC) together with the corresponding velocity and acceleration parameters. Hence there are a total of 39 parameters, denoted by MFCC(39). The proposed front-end feature extractor is modified from the MFCC model by integrating a 2D psychoacoustic modeling shown in Fig. 11.

The same recognizer is used for both the proposed front-end feature extraction algorithm and the baseline system for a meaningful comparison. Each digit is modeled by a simple left-to-right 18 states (including two non-emitting states) HMM model, with 3 Gaussian mixtures per state. Two pause models are defined. One is "sil", which has three HMM states and models the pauses before and after each utterance. The other one is "sp", which is a single state model (tied with the middle state of "sil") and models the pauses among words. Extensive comparison will be made against a series of relevant peer work. That is FM (Park and Lee, 2003), LI (Cheng and O'Shaughnessy, 1991), the relative spectra (RASTA) filter (Hermansky and Morgan, 1994), the original 2D filter (Dai and Soon, 2009), AFE (ETSI, 2007), and the temporal warped 2D filter (Dai and Soon, 2010).

4. Results and discussion

4.1. Experimental results

Experimental results are given in Tables 3–5. MFCC(39) stands for MFCC with velocity and acceleration components. All the comparison methods are developed based on MFCC(39). Experimental results are averaged over SNR 0–20 dB denoted as Avg. 0–20. 'Rel. Imp.' stands for relative improvements in terms of word error rate (WER).

4.2. Results analysis

4.2.1. Clean training condition results

The first comparison is made against the baseline MFCCs. The detailed recognition results are given in

Table 4
Recognition rate (%) of comparison targets for clean training condition.

SNR/dB	Clean	Avg. 0–20	Rel. Imp	−5	Rel. Imp.
MFCC(39)	99.36	71.29	**46.78**	13.04	**13.51**
FM	99.03	77.34	**32.57**	21.30	**4.43**
LI	99.42	73.97	**41.30**	17.07	**9.31**
CMVN	99.32	77.78	**31.23**	13.90	**12.65**
RASTA	99.08	77.94	**30.73**	19.93	**6.07**
Original 2D	99.32	77.64	**31.66**	13.85	**12.70**
AFE	99.20	82.23	**14.01**	24.77	**0.03**
Temporal warped 2D	99.33	80.36	**22.20**	14.42	**12.12**

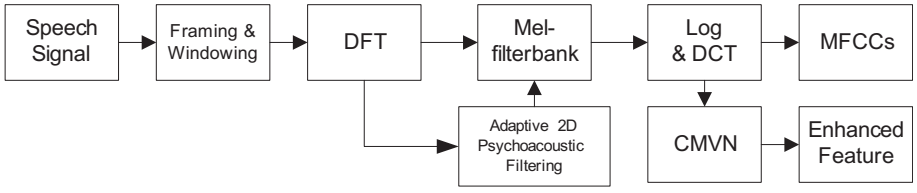


Fig. 11. System diagram.

Table 3
Detailed recognition rates (%) for the proposed algorithm.

Condition	SNR	Set A				Set B				Set C		Avg
		Subway	Babble	Car	Exhibition	Station	Restaurant	Street	Airport	Restaurant	Street	
Clean	Clean	99.45	99.21	99.34	99.63	99.45	99.21	99.34	99.63	99.36	99.27	99.39
	Avg. 0–20	85.57	85.03	85.76	82.20	85.66	85.76	86.39	85.31	82.20	83.35	84.72
	−5 dB	27.23	23.64	22.93	25.33	28.74	26.57	25.41	23.51	21.25	23.31	24.79
Multi	Clean	98.93	98.82	98.72	99.14	98.93	98.82	98.72	99.14	98.83	98.76	98.88
	Avg. 0–20	92.83	92.00	92.45	90.77	92.23	91.93	92.88	92.34	91.06	91.08	91.96
	−5 dB	49.71	42.96	44.23	43.72	44.73	45.59	46.85	44.25	41.94	40.21	44.42

Table 5
Recognition rate (%) of comparison targets for multi training condition.

SNR/dB	Clean	Avg. 0–20	Rel. Imp	−5	Rel. Imp.
MFCC(39)	99.11	87.85	**33.83**	26.83	**24.04**
FM	98.74	87.48	**35.78**	25.46	**25.44**
LI	99.13	88.26	**31.52**	26.59	**24.29**
CMVN	98.94	91.74	**2.66**	42.64	**3.10**
RASTA	98.60	91.27	**7.90**	44.35	**0.13**
Original 2D	99.02	91.65	**3.71**	41.69	**4.68**
AFE	98.96	91.83	**1.59**	42.33	**3.62**
Temporal warped 2D	99.05	91.45	**5.96**	38.57	**9.52**

Table 7
Relative Improvements for clean and multi training condition.

Condition	Clean		Multi	
SNR/dB	Avg. 0–20	−5	Avg. 0–20	−5
MFCC(39)	46.78	13.51	33.83	24.04
FM	32.57	4.43	35.78	25.44
LI	41.30	9.31	31.52	24.29
CMVN	31.23	12.65	2.66	3.10
RASTA	30.73	6.07	7.90	0.13
Original 2D	31.66	12.7	3.71	4.68
Temporal warped 2D	22.20	12.12	5.96	9.52
AFE	14.01	0.03	1.59	3.62

Table 4 and Table 6. The advantage of the proposed algorithm is very obvious. The experimental results show that the proposed algorithm yields better recognition rate for all SNR levels. Even for clean set, at 99% recognition rate level the proposed algorithm still performs better, though it is meaningless to discuss the difference between 99.3% and 99.4%. For Avg. 0–20, a relative improvement of 46.78% is achieved. Furthermore, at SNR –5 dB the relative improvements amazingly reaches 13.51%.

The second set of comparison is made against FM, LI and CMVN since they are the indispensible parts of the proposed algorithm. The proposed algorithm is designed to give FM and LI effects at the same time using a simple 2D psychoacoustic filter. From Table 4 it can be easily found out that the proposed algorithm yields significant improvements at all SNR levels. At Avg. 0–20, the relative improvements are 32.57% over FM, 41.3% over LI, and 31.23% over CMVN. For SNR −5 dB, the relative improvements are also very impressive, 4.43%, 9.31%, and 12.65% over FM, LI, and CMVN, respectively.

The final comparison is made against relevant work, the RASTA filter, original 2D filter, AFE, and temporal warped 2D filter. For Avg. 0–20, the relative improvements reach 30.73% over RASTA, 31.66% for the original 2D filter, 14.01% over AFE and 22.2% over the temporal warped 2D filter. For SNR −5 dB, the relative improvements are 6.07% over the RASTA filter, 12.7% over the Original 2D filter, 0.03% over AFE, and 12.12% over the temporal warped 2D filter. A better view of the advantage of the proposed algorithm over other comparison targets can be seen from Table 7. For all SNR levels significant improvements can be observed. From the above discussion, the advantage of the proposed algorithm is obvious.

Table 6
Word error rate (WER) for clean and multi training condition.

Condition	Clean		Multi	
SNR/dB	Avg. 0−20	−5	Avg. 0−20	−5
MFCC(39)	28.71	86.96	12.15	73.17
FM	22.66	78.70	12.52	74.54
LI	26.03	82.93	11.74	73.41
CMVN	22.22	86.10	8.26	57.36
RASTA	22.06	80.07	8.73	55.25
Original 2D	22.36	86.15	8.35	58.31
Temporal warped 2D	19.64	85.58	8.55	61.43
AFE	17.77	75.23	8.17	57.67
Proposed algorithm	15.28	75.21	8.04	55.58

4.2.2. Multi training condition results

There are two training conditions in the AURORA2 database, clean and multi training condition. For the multi training condition, since noisy speech is used to train HMMs the recognition results are all very good, even achieves about 80% recognition rate at SNR 5 dB. Therefore very large improvements at this level are not very possible to achieve. However, the proposed algorithm still manages to get very promising results. Table 7 shows the relative improvements of the proposed algorithm over the comparison targets. Tables 5 and 6 show the detailed numbers.

In terms of Avg. 0–20, the relative improvements are 33.83% over MFCC, 35.78% over FM, 31.52% over LI, 2.66% over CMVN, 7.90% over RASTA, 3.71% over the Original 2D filter, 1.59% over AFE, and 5.96% over the temporal warped 2D filter. At SNR −5 dB, the relative improvements are 24.04% over MFCC, 25.44% over FM, 24.29% over LI, 3.10% over CMVN, 0.13% over RASTA, 4.68% over the original 2D, 3.62% over AFE, and 9.52% over the temporal warped 2D filter.

5. Conclusions

In this paper, a hybrid feature extraction algorithm is proposed, which is successfully implemented in a MFCC based speech recognition system. It is developed based on MFCC. The fundamental operation in the implementation of the proposed algorithm is 2D convolution. Therefore, the proposed algorithm shares similar computational complexity like 2D convolution. The key feature of the proposed algorithm is that it successfully implements FM, LI and TI with a simple 2D psychoacoustic filter. It manages to reflect the asymmetrical nature of human auditory system. Moreover, the proposed method does not need any additional training process, making the computational burden very small. Besides, due to the simplicity of the proposed algorithm it can be easily combined with other algorithm. Extensive comparison is made based on the AURORA2 database. Significant improvements are achieved according to experimental results.

Table A1
Initial parameter (FM and LI).

Freq./time	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
−1	**−0.0137**	−0.0065	−0.005	−0.0041	−0.0034	−0.0029	−0.0025	−0.0022	−0.0019	−0.0017	−0.0014	−0.0012	−0.001	−0.0008	−0.0007	−0.0005	−0.0004
0	**1**	**−0.4736**	**−0.3622**	**−0.2971**	**−0.2508**	**−0.215**	**−0.1857**	**−0.1609**	**−0.1395**	**−0.1205**	**−0.1036**	**−0.0883**	**−0.0743**	**−0.0614**	**−0.0495**	**−0.0384**	**−0.0281**
1	**−0.0914**	−0.0433	−0.0331	−0.0272	−0.0229	−0.0196	−0.017	−0.0147	−0.0127	−0.011	−0.0095	−0.0081	−0.0068	−0.0056	−0.0045	−0.0035	−0.0026
2	**−0.1757**	−0.0832	−0.0636	−0.0522	−0.0441	−0.0378	−0.0326	−0.0283	−0.0245	−0.0212	−0.0182	−0.0155	−0.0131	−0.0108	−0.0087	−0.0068	−0.0049
3	**−0.2386**	−0.113	−0.0864	−0.0709	−0.0598	−0.0513	−0.0443	−0.0384	−0.0333	−0.0288	−0.0247	−0.0211	−0.0177	−0.0147	−0.0118	−0.0092	−0.0067
4	**−0.2129**	−0.1008	−0.0771	−0.0632	−0.0534	−0.0458	−0.0395	−0.0343	−0.0297	−0.0257	−0.0221	−0.0188	−0.0158	−0.0131	−0.0105	−0.0082	−0.006
5	**−0.0986**	−0.0467	−0.0357	−0.0293	−0.0247	−0.0212	−0.0183	−0.0159	−0.0138	−0.0119	−0.0102	−0.0087	−0.0073	−0.0061	−0.0049	−0.0038	−0.0028

Table A2
The final parameter (FM, LI, and TI).

Freq./time	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
−1	**−0.0137**	−0.0065	−0.005	−0.0041	−0.0034	−0.0029	−0.0025	−0.0022	−0.0019	−0.0017	−0.0014	−0.0012	−0.001	−0.0008	−0.0007	−0.0005	−0.0004
0	**5**	**−0.4736**	**−0.3622**	**−0.2971**	**−0.2508**	**−0.215**	**−0.1857**	**−0.1609**	**−0.1395**	**−0.1205**	**−0.1036**	**−0.0883**	**−0.0743**	**−0.0614**	**−0.0495**	**−0.0384**	**−0.0281**
1	**−0.0914**	−0.0433	−0.0331	−0.0272	−0.0229	−0.0196	−0.017	−0.0147	−0.0127	−0.011	−0.0095	−0.0081	−0.0068	−0.0056	−0.0045	−0.0035	−0.0026
2	**−0.1757**	−0.0832	−0.0636	−0.0522	−0.0441	−0.0378	−0.0326	−0.0283	−0.0245	−0.0212	−0.0182	−0.0155	−0.0131	−0.0108	−0.0087	−0.0068	−0.0049
3	**−0.2386**	−0.113	−0.0864	−0.0709	−0.0598	−0.0513	−0.0443	−0.0384	−0.0333	−0.0288	−0.0247	−0.0211	−0.0177	−0.0147	−0.0118	−0.0092	−0.0067
4	**−0.2129**	−0.1008	−0.0771	−0.0632	−0.0534	−0.0458	−0.0395	−0.0343	−0.0297	−0.0257	−0.0221	−0.0188	−0.0158	−0.0131	−0.0105	−0.0082	−0.006
5	**−0.0986**	−0.0467	−0.0357	−0.0293	−0.0247	−0.0212	−0.0183	−0.0159	−0.0138	−0.0119	−0.0102	−0.0087	−0.0073	−0.0061	−0.0049	−0.0038	−0.0028

Appendix A

The proposed psychoacoustic filter has the following form,

$$Mask_{\Delta f \cdot \Delta T} = \begin{bmatrix} \mathbf{0}_{F_1-F_2,T_1-T_2} & \mathbf{0}_{F_2-F_1,T_2-T_1+\mathbf{1}} \\ \mathbf{0}_{F_1+F_2+\mathbf{1},T_1+T_2-\mathbf{1}} & \hat{M} \end{bmatrix}$$

$\hat{M}$ are given in Tables A1 and A2.

References

Brookes, M., 1997. Voicebox. Available: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

Dai, P., Soon, I.Y., 2009. 2D psychoacoustic filtering for robust speech recognition. In: Proc. ICICS, December.

Dai, P., Soon, I.Y., 2010. A temporal warped 2D psychoacoustic modeling for robust speech recognition system. Speech Comm. 53 (2), 229–241.

Cheng, Y.M., O'Shaughnessy, D., 1991. Speech enhancement based conceptually on auditory evidence. IEEE Trans. Signal Process. 39, 1943–1954.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Speech Signal Process. 28 (4), 357–366.

ETSI, 2007. European Telecommunications Standards Institute (ETSI). ETSI ES 202 050 V1.1.5.

Gold, B., Morgan, N., 2000. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. John Wiley, New York.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Hermansky, H., 1998. Should recognizers have ear? Speech Comm. 25 (1), 3–27.

Hirsch, H., Pearce, D., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR 181, 188.

Houtgast, T., 1972. Psychophysical evidence for lateral inhibition in hearing. J. Acoust. Soc. Amer. 51 (6B), 1885–1894.

Ishizuka, K., Nakatani, T., 2010. Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Comm. 52, 41–60.

Jesteadt, W., Bacon, S., Lehman, J., 1982. Forward masking as a function of frequency, masker level, and signal delay. J. Acoust. Soc. Amer. 71, 950–962.

Luo, X., Soon, I.Y., Yeo, C.K., 2008. An auditory model for robust speech recognition. Proc. ICALIP 1105, 1109.

Oxenham, A.J., Plack, C.J., 2000. Effects of masker frequency and duration in forward masking: further evidence for the influence of peripheral nonlinearity. Hearing Res. 150, 258–266.

Oxenham, A.J., 2001. Forward masking: adaptation or integration? J. Acoust. Soc. Amer. 109 (2), 732–741.

Park, K.Y., Lee, S.Y., 2003. An engineering model of the masking for the noise-robust speech recognition. Neurocomputing 52 (4), 615–620.

Shamma, S.A., 1985. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. J. Acoust. Soc. Amer. 78 (5), 1622–1632.