Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 229–241
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A temporal warped 2D psychoacoustic modeling for robust speech recognition system

Peng Dai *, Ing Yann Soon

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

Received 12 November 2009; received in revised form 14 July 2010; accepted 16 September 2010

Abstract

Human auditory system performs better than speech recognition system under noisy condition, which leads us to the idea of incorporating the human auditory system into automatic speech recognition engines. In this paper, a hybrid feature extraction method, which utilizes forward masking, backward masking, and lateral inhibition, is incorporated into mel-frequency cepstral coefficients (MFCC). The integration is implemented using a warped 2D psychoacoustic filter. The AURORA2 database is utilized for testing, and the Hidden Markov Model (HMM) is used for recognition. Comparison is made against lateral inhibition (LI), forward masking (FM), cepstral mean and variance normalization (CMVN), the original 2D psychoacoustic filter and the RASTA filter. Experimental results show that the word recognition rate is significantly improved, especially under noisy conditions.

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

1. Introduction

Environment noise, channel distortion and speaker variability can all degrade automatic speech recognition system's performance. However, human auditory system can easily deal with such situation (Allen, 1994; Hermansky, 1998). Hence analysis and modeling of human auditory system can help to improve automatic speech recognition system (ASR).

Hearing is not only a simple mechanical phenomenon of wave indexing, but it is also a sensory and perceptual process. Psychoacoustics is just the science of analyzing the human perception of sounds, which consist of relationships between sound pressure level and loudness, human response to different frequencies, and a number of masking effects (Golden and Morgan, 2000). Through the analysis and integration of these nonlinear perceptions, the performance of speech recognition system can be greatly improved. The famous mel-frequency cepstral coefficients (MFCC) method is a good example. An auditory-based warping of the frequency axis, called critical-band, is adopted for the modeling of the frequency sensitivity of human hearing (Davis and Mermelstein, 1980).

Masking effect is a kind of phenomenon in which a clearly audible sound can be masked by another sound, known as the masker. Masking effects can be classified as simultaneous or temporal depending on the time of occurrence of the signals. If two sounds occur simultaneously or at the same time, the masking effect between the two is referred to as simultaneous masking. In temporal masking, signals can also be masked by the preceding sound that occurs earlier in time, which is known as forward masking. If the masker occurs after the signal, the masking effect is called backward masking (Shamma, 1985; Kvale and Schreiner, 2004). Because forward masking is much more effective than backward masking, work has been mainly focused on the modeling and application of forward masking (Jesteadt et al., 1982; Strope and Alwan, 1997; Oxenham and Plack, 2000; Oxenham, 2001).

The novelty of this paper lies in that with a simple integration of a 2D psychoacoustic filter the speech recognition

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

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.09.004

results can be greatly improved. The AURORA2 database is used for verification tests. It is a widely used, standard English database, which contains isolated digits as well as digit serials. Comparison is made against ordinary MFCCs, forward masking (FM), lateral inhibition (LI), cepstral mean and variance normalization (CMVN), RASTA filter, and the original 2D filter. Experimental results show the excellent performance of the proposed method.

2. Masking effects

2.1. Simultaneous masking

A lot of work has been conducted for the implementation of simultaneous masking in audio compression and speech enhancement (Ambikairajah et al., 1997; Cheng and O'Shaughnessy, 1991; Luo et al., 2008; Virag, 1999). However, speech recognition is a relatively new area for its application. One effective approach of simultaneous masking is lateral inhibition (LI) (Shamma, 1985; Luo et al., 2008). Lateral inhibition is a common phenomenon involved in sensory reception of biological systems, e.g., the uniform luminosity of an object is emphasized at its edges in human vision, an excitation on one's skin can be diminished by the presence of other spatially neighboring excitations, etc. (Cheng and O'Shaughnessy, 1991). In other words, the general function of lateral inhibition is to sharpen input changes. Lateral inhibition exists in both time domain and spectral domain.

Firstly, the effects of lateral inhibition in the spectral domain are discussed. Fig. 1(a) shows the characteristic curve of lateral inhibition. The 1D Mexican hat is utilized for the parameterization of lateral inhibition, as is shown in Fig. 1(b) (Dai and Soon, 2009; Luo et al., 2008).

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

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

where k is the time index. $n(k)$ is further assumed to be uncorrelated with speech signal, $s(k)$.

In power spectra, Eq. (1) can be written as:

$$P_x(\omega) = P_s(\omega) + P_n(\omega) \tag{2}$$

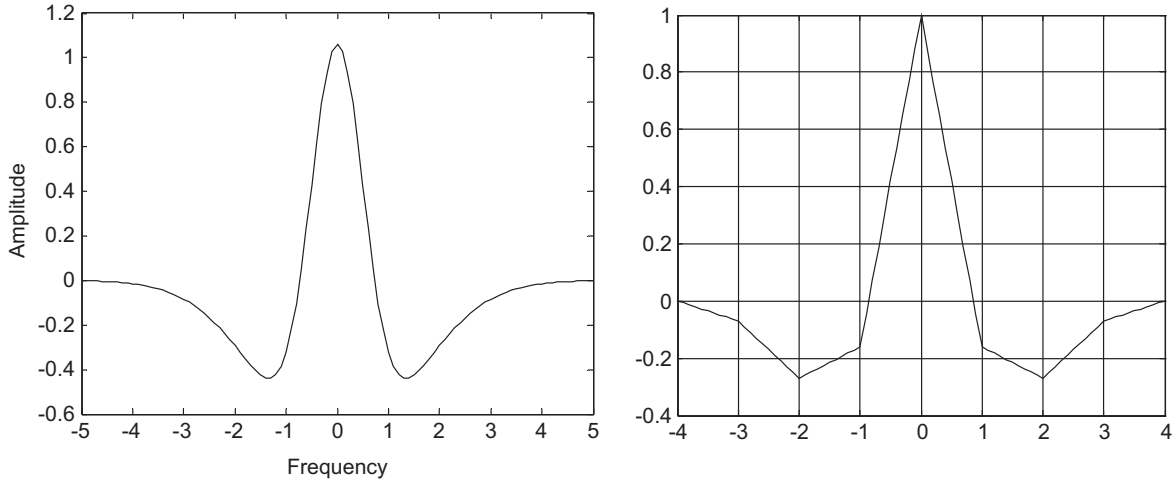
where ω is the radian frequency index, $P_x(\omega)$, $P_s(\omega)$ and $P_n(\omega)$ are the power spectral density of $x(k)$, $s(k)$ and $n(k)$, respectively.

Define $P_m = \min\{P_n(\omega)\}$. Therefore, $P_n(\omega) = P_m + P'_n(\omega)$, and $P'_n(\omega) \leqslant P_n(\omega)$. Let $M_{\mathrm{LI}}(\omega, \Omega)$ represent the lateral inhibition masker, where Ω is the centre frequency of the masker. The lateral inhibition masker is modeled to satisfy the following constraint (Cheng and O'Shaughnessy, 1991):

$$\int_{-\infty}^{\infty} M_{\mathrm{LI}}(\omega, \Omega) d\omega = 0 \quad \forall \Omega \tag{3}$$

By applying the given lateral inhibition masker in Eq. (3) to the noisy speech power spectrum, the processed power spectrum, $\hat{P}_x(\Omega)$, is

$$\begin{aligned}
\hat{P}_x(\Omega) &= \int_{-\infty}^{\infty} P_x(\omega) M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&= \int_{-\infty}^{\infty} P_s(\omega) M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&\quad + \int_{-\infty}^{\infty} P_n(\omega) M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&= \hat{P}_s(\Omega) + \int_{-\infty}^{\infty} [P_m + P'_n(\omega)] M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&= \hat{P}_s(\Omega) + \int_{-\infty}^{\infty} P_m M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&\quad + \int_{-\infty}^{\infty} P'_n(\omega) M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega \\
&= \hat{P}_s(\Omega) + \int_{-\infty}^{\infty} P'_n(\omega) M_{\mathrm{LI}}(\Omega - \omega, \Omega) d\omega
\end{aligned} \tag{4}$$



(a) Characteristic Curve of LI (b) 1D Mexican Hat

Fig. 1. Characteristic curve of lateral inhibition.

As is shown in Eq. (4), after processed by the lateral inhibition masker, part of the noise is removed, which effectively increases the SNR. It has to be noted that if the noise is white noise $P'_n(\omega) = 0$, then

$$\begin{aligned}\hat{P}_x(\Omega) &= \int_{-\infty}^{\infty} P_x(\omega) M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega \\ &= \int_{-\infty}^{\infty} P_s(\omega) M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega \\ &\quad + \int_{-\infty}^{\infty} P_n(\omega) M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega \\ &= \hat{P}_s(\Omega) + \int_{-\infty}^{\infty} [P_m + P'_n(\omega)] M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega \\ &= \hat{P}_s(\Omega) + \int_{-\infty}^{\infty} P_m M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega \\ &\quad + \int_{-\infty}^{\infty} P'_n(\omega) M_{\mathrm{LI}}(\Omega-\omega,\Omega)d\omega = \hat{P}_s(\Omega) \end{aligned} \tag{5}$$

Theoretically, the power spectrum of processed speech with white noise equals to that of the input clean speech. From Eqs. (4) and (5), the algorithm not only sharpens the spectrum of the input signal but also removes noise.

2.2. Temporal masking

A weak sound emitted soon after the end of a louder sound can be masked by the louder sound. In fact, even a weak sound occurring just before a louder sound can also be masked by the louder sound. These two effects are called forward and backward temporal masking, respectively. Forward masking (FM) reveals that over short durations, the usable dynamic range of the human auditory system depends on the spectral characteristics of the previous stimuli. Forward masking can be viewed as a consequence of auditory adaptation (Strope and Alwan, 1997).

The parameters for forward masking are frequency dependent. Backward masking has the same characteristic but the masker comes after the signal rather than before it. Fig. 2 shows the characteristic curve of temporal masking. Forward masking can be viewed as a consequence of auditory adaptation (Park and Lee, 2003; Strope and Alwan, 1997).

3. 2D psychoacoustic modeling

3.1. Fundamentals of 2D filtering

Although forward masking is frequency dependent, the variation is relatively small. By noting these facts, a simplified 2D psychoacoustic algorithm can be implemented (Dai and Soon, 2009). The speech signal is cut into frames and transformed into frequency domain, as is shown in Fig. 3. Hence there is a time interval, Δt, between each frame. Within each frame, the elements are in frequency domain. Hence lateral inhibition can be applied within each frame. Between frames, because of the existence of Δt, temporal masking effects should be taken into consideration. In our system, the frame length is chosen to be 128. Besides, there is 50% overlapping between each frame.

The proposed 2D psychoacoustic filter is designed to give the effect of both lateral inhibition and Temporal Masking at the same time. As stated in previous section, masking effect measures how a certain speech signal is affected by its neighboring sound. Therefore, after taking into account masking effect, the speech signal becomes

$$\hat{P}_x(f,t) = \sum_{\Delta f}\sum_{\Delta t} F[P_x(f+\Delta f, t+\Delta t)] + P_x(f,t) \tag{6}$$

where $F(\cdot)$ stands for the function for masking effects (Park and Lee, 2003; Strope and Alwan, 1997; Jesteadt et al., 1982).

In our proposed method, the relationship between the masking effect and the initiating sound is assumed to be linear. Then taking into account the effective range of masking effect, Eq. (6) becomes

$$\begin{aligned}\hat{P}_x(f,t) &= \sum_{\Delta f=-F_1}^{F_2}\sum_{\Delta t=-T_1}^{T_2} a_{\Delta f,\Delta t} P_x(f+\Delta f, t+\Delta t) \\ &= P_x(f,t) \otimes M_{\Delta t,\Delta f}\end{aligned} \tag{7}$$

where $\otimes$ stands for convolution, $a_{0,0} = 1$; F_1, F_2, T_1, T_2 are the boundaries for masking effect

$$M_{\Delta t,\Delta f} = \begin{bmatrix} a_{T_2,F_2} & & & a_{0,F_2} & & & a_{-T_1,F_2} \\ & \ddots & & \vdots & & & \\ & & a_{-1,-1} & a_{0,-1} & a_{-1,1} & & \\ a_{T_2,0} & \cdots & a_{-1,0} & 1 & a_{-1,0} & \cdots & a_{-T_1,0} \\ & & a_{1,-1} & a_{0,1} & a_{-1,-1} & & \\ & & \vdots & & \ddots & & \\ a_{T_2,-F_1} & & & a_{0,-F_1} & & & a_{-T_1,-F_1} \end{bmatrix} \tag{8}$$

The detailed design process for the parameters in Eq. (8) will be discussed in the following sections.

3.2. The original 2D filter

The 1D Mexican-hat, as is shown in Fig. 1(b), is suitable for the modeling of lateral inhibition (Luo et al., 2008)

$$[-0.07 \quad -0.27 \quad -0.16 \quad 1 \quad -0.16 \quad -0.27 \quad -0.07]$$

The sampling frequency of AURORA2 database is 8000 Hz. Therefore, the time difference, Δt, between each frame is 8 ms. From the characteristic of forward masking, it is reasonable to set the parameters for FM to cover 7 frames, 56 ms. Besides, from Strope's paper and Park's paper (Strope and Alwan, 1997; Park and Lee, 2003) the shape of the characteristic curve of FM is similar to the 1D Mexican-hat. Therefore, by assuming that lateral inhibition and temporal masking share the same set of

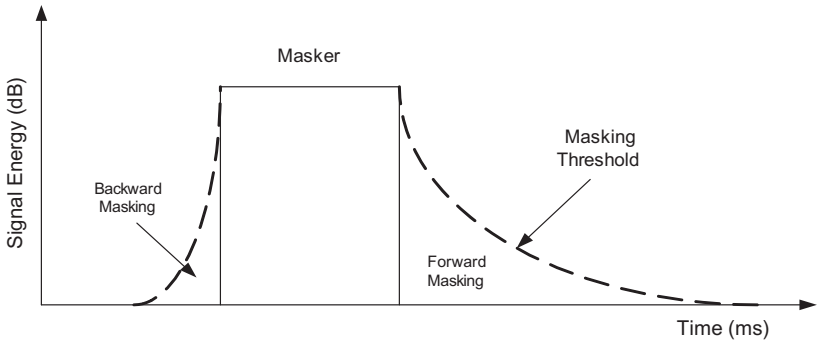


Fig. 2. Temporal masking.

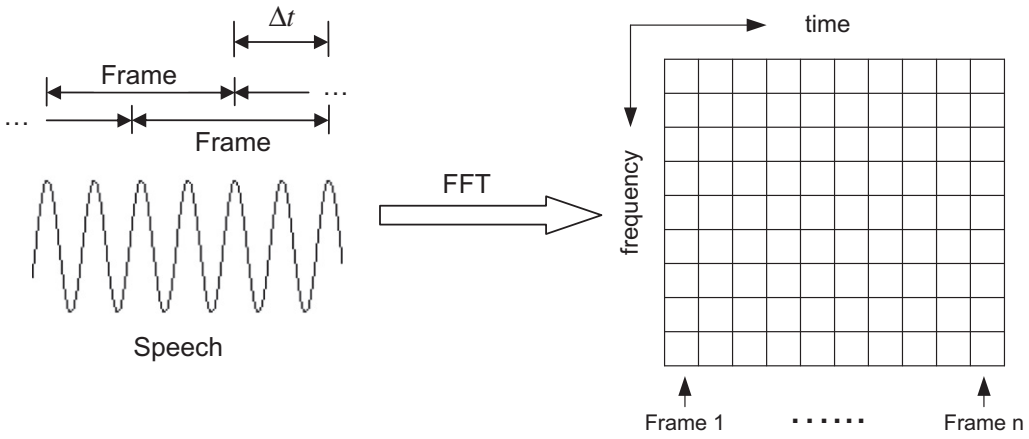


Fig. 3. Processing of speech signals.

parameters, the initial parameter set of 2D filter is shown in Table 1.

The parameters in diagonal directions are calculated from Table 1 using linear interpolation. Fig. 4 shows the characteristic curve used to calculate the parameters.

The diagonal coefficient value is a function of d and it can be obtained from Fig. 4 using linear interpolation as described by Eq. (10). The distance of the diagonal coefficients, $mask(d)$, from the centre coefficients is denoted as d. d can be calculated using Eq. (9)

$$d = \sqrt{t^2 + f^2} \tag{9}$$

Because the masker is symmetric, the calculation can be made using just one side of the curve shown in Fig. 4. Take the right side of the curve as an example. The function of the characteristic curve is

$$mask(d) = \begin{cases} 1 - 1.16d, & 0 \leqslant d < 1 \\ -0.05 - 0.11d, & 1 \leqslant d < 2 \\ -0.67 + 0.2d, & 2 \leqslant d < 3 \\ -0.07 + \frac{0.07}{3\sqrt{2}-3}(d-3), & 3 \leqslant d \leqslant 3\sqrt{2} \end{cases} \tag{10}$$

It has to be noted that the last section is $3 \leqslant d \leqslant 3\sqrt{2}$ making the total range of d in Eq. (10) $-3\sqrt{2} \leqslant d \leqslant 3\sqrt{2}$, which is slightly different from the Mexican hat in Fig. 1(b). The objective is to have four zeros at the corners of the 2D mask, $mask(\pm 3, \pm 3)$, as shown in Table 2.

Empirically, 40 is an optimal value for the centre element (Dai and Soon, 2009), as is shown in Table 2. The original 2D filter is shown in Fig. 5(b).

3.3. Temporal warping

The original 2D filter is developed based on the assumption that the lateral inhibition and temporal masking share the same set of parameters. However, further study shows that the validity of the assumption depends on lots of things such as sampling rate, frame rate and so on. Hence temporal warping is required to improve the 2D psychoacoustic filter.

There are different mathematical models for describing temporal masking effect. All of them (Lois, 1962; Strope and Alwan, 1997; Park and Lee, 2003; Nghia et al., 2008) come to a common conclusion that forward masking is more effective than backward masking, thus the parameters of temporal masking cannot be symmetric. It has to be warped to get a new set of temporal masking parameters, which should follow the characteristic curve in Fig. 2. As stated in Section 1, the 1D Mexican Hat is chosen to be the primary parameter for temporal masking. Each side

Table 1
Initial parameter set of 2D filter.

Freq \ time ($f \setminus t$)	−3	−2	−1	0	1	2	3
−3				−0.07			
−2				−0.27			
−1				−0.16			
0	−0.07	−0.27	−0.16	1	−0.16	−0.27	−0.07
1				−0.16			
2				−0.27			
3				−0.07			

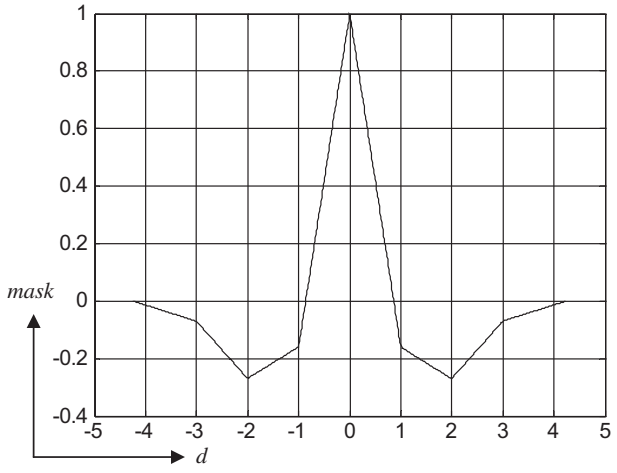


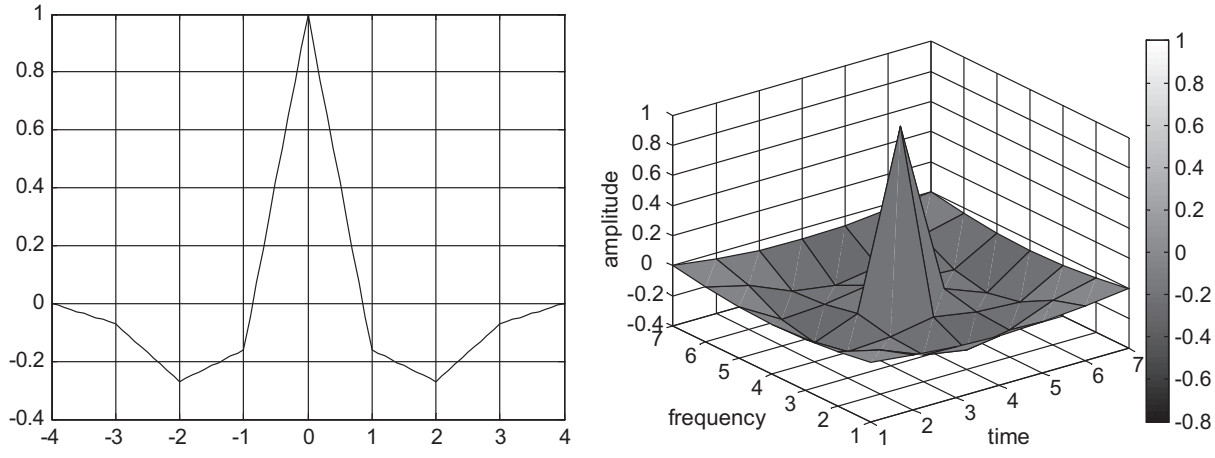
Fig. 4. Characteristic curve.

of the mask is linearly warped. Fig. 6 shows the warped mask. The detailed steps for calculation are as follows.

Starting from the 1D Mexican hat in Fig. 1(b), the warping incorporates two steps. Firstly, each side of the Mexican hat is modified proportionally, making the right side become 7/4 times the original length and the left side become 1/4 the length of the original. Fig. 6 shows the warped parameter and the original parameter. Each side of the Mexican hat contains four sections (right: [0, 1) [1, 2) [2, 3) [3, 4]; left: (−1, 0) (−2, −1] (−3, −2] [−4, −3]). The length of each interval Δd is 1. After the first step, the sections become [0, 1.75) [1.75, 3.5) [3.5, 5.25) [5.25, 7] of the right side and (−0.25, 0) (−0.5, −0.25] (−0.75, −0.5] [−1, −0.75] of the left side. The length of intervals on the left side is 0.25, and the interval length of the right side is 1.75. Like Eq. (10), the two sections at the ends

Table 2
The original 2D filter.

Freq \ time ($f \setminus t$)	−3	−2	−1	0	1	2	3
−3	0	−0.0359	−0.0609	−0.07	−0.0609	−0.0359	0
−2	−0.0359	−0.1043	−0.2228	−0.27	−0.2228	−0.1043	−0.0359
−1	−0.0609	−0.2228	−0.2056	−0.16	−0.2056	−0.2228	−0.0609
0	−0.07	−0.27	−0.16	40	−0.16	−0.27	−0.07
1	−0.0609	−0.2228	−0.2056	−0.16	−0.2056	−0.2228	−0.0609
2	−0.0359	−0.1043	−0.2228	−0.27	−0.2228	−0.1043	−0.0359
3	0	−0.0359	−0.0609	−0.07	−0.0609	−0.0359	0



(a) 1D Mexican Hat (b) The Original 2D Filter

Fig. 5. The original 2D psychoacoustic filter.

([5.25, 7] and [−1, −0.75]) needs further modification. In Eq. (10) the last interval is extended to $3\sqrt{2}-3$ times the original length. Similar operation is made to the warped Mexican hat.

The second step is to extend interval [5.25, 7] and interval [−1, −0.75] to $\left[5.25, 5.25+1.75\times\frac{3\sqrt{2}-3}{1}\right]$ and $\left[-0.75-0.25\times\frac{3\sqrt{2}-3}{1}, -0.75\right]$, respectively. Hence, the linear interpolation equation for the warped Mexican hat is

$$mask(d)=\begin{cases}-0.239-0.2253d, & -1.06\leqslant d\leqslant -0.75\\ -0.67-0.8d, & -0.75< d\leqslant 0.5\\ -0.05+0.44d, & -0.5< d\leqslant -0.25\\ 1+4.64d, & -0.25< d<0\\ 1-0.6629d, & 0\leqslant d<1.75\\ -0.05-0.0629d, & 1.75\leqslant d<3.5\\ -0.67+0.1143d, & 3.5\leqslant d<5.25\\ -0.239+0.0322d, & 5.25\leqslant d\leqslant 7.42\end{cases} \tag{11}$$

The warped parameters are calculated using Eq. (11), setting $d=-1, 0, 1, 2, 3, 4, 5$. The warped parameters are as follows.

$$[-0.0137 \quad 1 \quad 0.3371 \quad -0.1757 \quad -0.2386 \quad -0.2129 \quad -0.0986]$$

The initial parameter set of the warped 2D filter is shown in Table 3.

In order to develop an equation for the diagonal parameters, two things have to be made clear. Firstly, the proposed filter is implemented by convolution in the time frequency domain. Another thing that has to be mentioned is that the physical meaning of the sideward parameters. The centre element stands for the target speech element. According to temporal masking and simultaneous masking, in time frequency domain all the neighboring sound weakens the target sound. Therefore, all the other 48 elements of the mask are used to measure how much the signal at that location weakens the target signal. For that reason except the centre all the parameters should be negative, which accounts for the weakening effect. Besides, as is shown in Figs. 1 and 2, masking effect will become weaker as two sounds become farther from each other. Therefore, the equation for diagonal parameters should possess the following form,

$$mask(t,f)=\begin{cases}1, & f=0 \text{ and } t=0\\ -\frac{[x^2(t)+y^2(f)]^{1/2}}{(t^2+f^2)^{1/2}}, & \text{others}\end{cases} \tag{12}$$

$$x(t)=\{-0.0800, 1, 0.3371, -0.1757, -0.2386, -0.2129, -0.0986\},\quad t=-1,0,\ldots,4,5$$

$$y(f)=\{-0.07, -0.27, -0.16, 1, -0.16, -0.27, -0.07\},\quad f=-3,-2,\ldots,2,3$$

where f stands for the frequency index of the mask, and t refers to the time index of the mask; $x(f)$ and $y(t)$ are the basic parameters; the minus sign corresponds to the fact that all the surrounding speech signal in the time frequency domain tend to weaken the centre element, $mask(0, 0)$; $1/(t^2+f^2)^{1/2}$ corresponds to the fact that masking effect becomes weaker as two sounds become farther from each other. With the initial parameters from Table 3, parameters of the mask are calculated using Eq. (12). The warped mask is shown in Table 4.

About $[x^2(t)+y^2(f)]^{1/2}$, it is chosen because it can yield good property for the final parameter. As is shown in Table 4, similar shape can be found by cutting from any direction crossing the centre. For example, crossing the centre with angle 45°, the parameters are as follows

$$[-0.1136 \quad 1 \quad -0.2639 \quad -0.1139 \quad -0.0586]$$

The parameters satisfy the properties of masking effect pretty well. The two parameters near the centre are the most negative on their own sides, −0.1136, −0.2639. As the element lies farther from the centre, the parameter becomes closer to 0, which corresponds to the property described in Figs. 1 and 2.

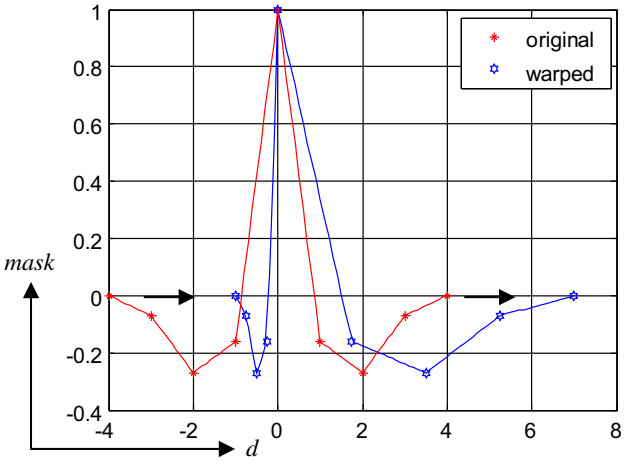


Fig. 6. 2D psychoacoustic filter with temporal warping.

3.4. Centre enhancement

The overall sum of the mask element, without the centre element is −10.3766. A neutral filter should have a centre element of 11.3766, so that the overall weighting is one. Direct implementation of this mask yields poor recognition results, as the masking effects are too dominant. Hence, we would like to reduce the effect by boosting the centre element which is also the original signal. This process can also be viewed as a weighted combination of the original signal and the filtered result. According to experiments, 40 is an optimal value for $mask(0, 0)$. Table 5 shows the centre enhanced 2D psychoacoustic filter.

The proposed 2D psychoacoustic filter enhances the high frequencies and helps to sharpen the spectral peaks

Table 3
Initial parameter set of warped 2D filter.

Freq \ time ($f \setminus t$)	−1	0	1	2	3	4	5
−3		−0.07					
−2		−0.27					
−1		−0.16					
0	−0.0137	1	0.3371	−0.1757	−0.2386	−0.2129	−0.0986
1		−0.16					
2		−0.27					
3		−0.07					

Table 4
Temporal warped 2D filter.

Freq \ time ($f \setminus t$)	−1	0	1	2	3	4	5
−3	−0.0226	−0.3341	−0.1089	−0.0525	−0.0586	−0.0448	−0.0207
−2	−0.1209	−0.5179	−0.1932	−0.1139	−0.0999	−0.0769	−0.0534
−1	−0.1136	−1.0127	−0.2639	−0.1063	−0.0908	−0.0646	−0.0369
0	−1.0001	1	−1.0553	−0.5077	−0.3427	−0.2556	−0.201
1	−0.1136	−1.0127	−0.2639	−0.1063	−0.0908	−0.0646	−0.0369
2	−0.1209	−0.5179	−0.1932	−0.1139	−0.0999	−0.0769	−0.0534
3	−0.0226	−0.3341	−0.1089	−0.0525	−0.0586	−0.0448	−0.0207

so as to improve the performance of the recognition system. The frequency response of the proposed 2D psychoacoustic filter is shown in Fig. 7.

4. Design of experiments

4.1. Database

The AURORA2 database is adopted to evaluate the performance of the proposed method. The AURORA2 data are based on a version of the original TIDigits (as available from LDC) downsampled to 8 kHz (Hirsch and Pearce, 2000; Leonard, 1984).

The training set in this database has no noise added and it consists of 8440 utterances recorded from 55 male and 55 female adults. 4004 utterances from 52 male and 52 female speakers are split equally into 4 subsets with 1001 utterances each, with all speakers being present in each subset. In the multi-condition training set, four types of noises have been added at various SNR levels.

In test set A, subway, babble, car and exhibition hall noises are added to the four clean data subsets at SNRs from −5 dB to 20 dB with 5 dB step size. So there are $4 \times 7 \times 1001 = 28{,}028$ utterances in this test set. Similarly, in test set B, another four different types of noises (restaurant, street, airport and train station) are added to the same subsets of clean data with the same SNR levels. For test set C, two noises (subway and street) processed by MIRS

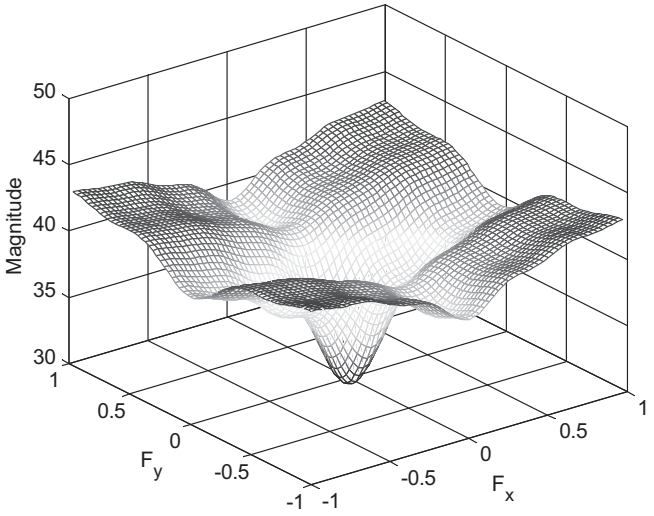


Fig. 7. Frequency response of the 2D psychoacoustic filter.

Table 5
Final 2D parameters.

Freq \ time ($f \setminus t$)	−1	0	1	2	3	4	5
−3	−0.0226	−0.3341	−0.1089	−0.0525	−0.0586	−0.0448	−0.0207
−2	−0.1209	−0.5179	−0.1932	−0.1139	−0.0999	−0.0769	−0.0534
−1	−0.1136	−1.0127	−0.2639	−0.1063	−0.0908	−0.0646	−0.0369
0	−1.0001	40	−1.0553	−0.5077	−0.3427	−0.2556	−0.201
1	−0.1136	−1.0127	−0.2639	−0.1063	−0.0908	−0.0646	−0.0369
2	−0.1209	−0.5179	−0.1932	−0.1139	−0.0999	−0.0769	−0.0534
3	−0.0226	−0.3341	−0.1089	−0.0525	−0.0586	−0.0448	−0.0207

(Modified Intermediate Reference System) filter are added, which simulates the frequency characteristics of a telecommunication terminal. The data size of test set C is thus half that of test sets A and B, since there are only two types of noise added.

4.2. Baseline system and comparison target

Four set of comparisons will be made. The first part is made against MFCC. The second is against LI and FM. The third and the fourth are made against some peer work which appeared recently, the RASTA filter (Hermansky and Morgan, 1994) and the original 2D filter (Dai and Soon, 2009), which can help to show the effectiveness of the proposed method. Results are averaged over the noisy test sets with SNRs from 0 to 20 dB.

4.2.1. MFCC

Mel frequency cepstral coefficients (MFCCs) are derived from a type of cepstral representation of the audio clip. Fig. 8 shows the block diagram of MFCC. 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 MFCC code in VoiceBox Toolkit is adopted to calculate the 13 cepstral coefficients including log energy term (logE). Furthermore, delta coefficients and acceleration coefficients are computed using

$$d_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2} \tag{13}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t+\theta}$ to $c_{t-\theta}$. In the experiments, Θ is set to be 2.

4.2.2. Cepstral mean and variance normalization (CMVN)

A typical problem with speech recognition systems is that the extracted features may be greatly different from one another. Cepstral mean and variance normalization (CMVN) is usually adopted to minimize the effect of these differences on recognizer performance. CMVN is typically used for removing convolution distortion. The mean of MFCCs across frames, $\bar{c}_k$, is calculated by

$$\bar{c}_k = \frac{1}{T}\sum_{t=1}^{T} c(t,k), \quad k = 1, 2, \ldots, K \tag{14}$$

where T is the number of frames and K is the number of coefficients in a vector. Then $\bar{c}_k$ is subtracted from the feature and then divided by the standard deviation, σ_k

$$\hat{c}(t,k) = \frac{c(t,k) - \bar{c}_k}{\sigma_k}, \quad k = 1, 2, \ldots, K \tag{15}$$

4.2.3. Relative spectra filter (RASTA)

The relative spectra (RASTA) was suggested by Hermansky in 1994. It is based on the fact that human perception tends to react to the relative value of an input (Hermansky and Morgan, 1994). Fig. 9 shows the diagram of RASTA method.

The transfer function of the RASTA filter is

$$H(z) = 0.1z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{16}$$

The difference between the proposed method and the RASTA filter is that the 2D psychoacoustic filter is applied in the frequency domain while RASTA filter is in the cepstral domain and log frequency domain.

4.3. Recognition engine

The proposed front-end feature extractor is modified from the MFCC model provided by Voicebox Toolkit (Brookes, 1997) by integrating a 2D psychoacoustic modeling. As is shown in Fig. 10, MFCC model with 2D psychoacoustic effects, shown in Fig. 10(b), has only two additional steps as compared to the traditional MFCC model, shown in Fig. 10(a).

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 3 HMM states and models the pauses before and after each

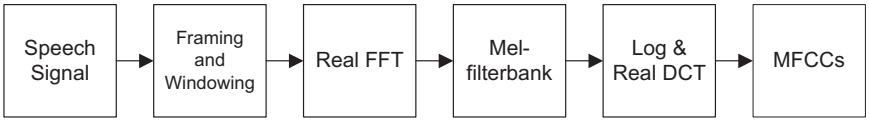


Fig. 8. Diagram of MFCC.

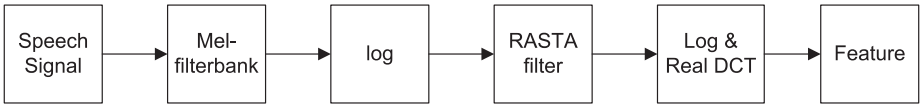


Fig. 9. Diagram of RASTA.

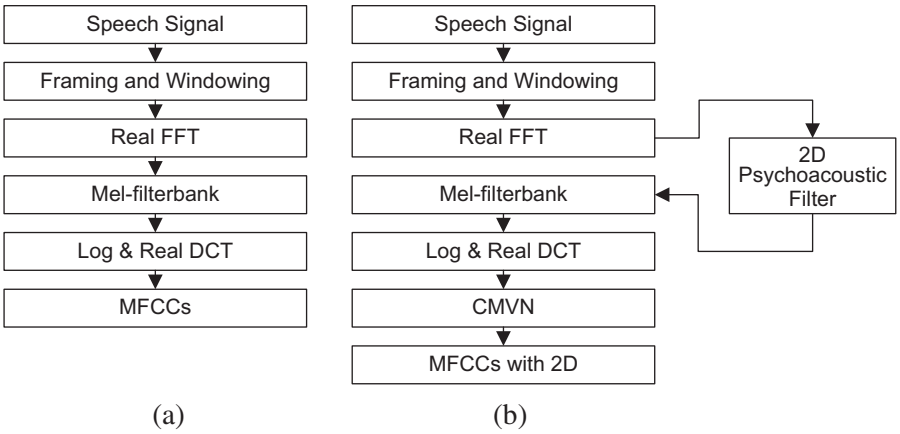


Fig. 10. Comparison of MFCC and 2D psychoacoustic models.

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.

5. Results and discussion

Experiments are conducted using clean training condition, which means that the HMMs are trained using only the clean speech data. Comparisons are made against the baseline system, forward masking, lateral inhibition, FM and LI together, cepstral mean and variance normalization, and the original 2D psychoacoustic filter. Results are averaged over the noisy test sets with SNRs from 0 to 20 dB. All the comparison methods are developed based on MFCC(39). In the following discussion, the proposed 2D psychoacoustic filter is denoted as warped 2D. LI stands for MFCC(39) with LI auditory filter, FM denotes MFCC(39) with forward masking filter, FM + LI stands for FM together with LI, CMVN denotes MFCC(39) processed by CMVN, and original 2D is the 2D psychoacoustic filter without temporal warping.

Table 6 shows the experimental results, which is also shown in Fig. 11. In the table, Avg stands for average. Avg 0–20 refers to the average over SNR 0–20 dB. Avg All means the average over all the SNRs, which contains clean and SNR 20 dB to −5 dB. For each method, the table shows the detailed experimental results for all three test sets at various SNRs.

5.1. Baseline system

As stated in above sections, the proposed 2D psychoacoustic filter is developed based on mel frequency cepstral coefficients (MFCCs). Hence, the first comparison is made against MFCC(39). The detailed recognition results of MFCC and 2D filter under clean training condition are given in Table 6.

The experimental results show that the proposed algorithm yields better recognition rate for all SNR levels. And as SNR decreases the improvement becomes more and more significant. At SNR 20 dB the recognition rate relatively improves 0.10%. As SNR falls down from 20 dB to 0 dB, the relative improvement goes from 2.22% to the impressive 50.91%. For SNR −5 dB, the relative improvement reaches 10.54%. The detailed improvement rates are shown in Fig. 12. The largest relative improvement occurs at SNR 0 dB.

In terms of the absolute improvement, the advantage of the proposed 2D psychoacoustic filter is still quite significant. The largest improvement comes at SNR 5 dB, which is about 19.68%. At SNR 20 dB 15 dB 10 dB 9 dB and −5 dB, the increases are 0.10%, 2.08%, 9.06%, 14.45% and 1.38%, respectively. Although the experimental results at clean state show some disadvantage of the proposed method with a slight decrease of 0.035%, the drop is quite small compared with the improvement. For the overall average of speech recognition rate, the relative

Table 6
Recognition rate (%) of warped 2D filter vs. MFCC(39), FM, LI, CMVN, RASTA filter, and the original 2D filter.

SNR (dB)	Clean	20	15	10	5	0	−5	Avg 0–20	Avg All
MFCC(39)	99.3617	97.3683	93.5142	81.1600	56.0225	28.3933	13.0433	71.2917	66.9805
FM	99.0283	97.0158	93.9117	85.8933	68.2442	41.6533	21.2983	77.3437	72.4350
LI	99.4217	97.1925	94.2317	83.2933	60.9217	34.2117	17.0692	73.9702	69.4774
FM + LI	99.0617	96.6308	93.5850	86.2625	69.9675	40.9358	16.4708	77.4763	71.8449
CMVN	99.3200	96.9692	94.3225	87.5900	71.2000	38.8408	13.9025	77.7845	71.7350
RASTA	99.0783	96.4483	93.5050	86.4133	70.5700	42.7808	19.9267	77.9435	72.3746
Original 2D	99.3150	97.0000	94.3192	87.4075	70.8217	38.6675	13.8492	77.6432	71.6258
Warped 2D	99.3267	97.4650	95.5933	90.2150	75.7008	42.8475	14.4183	80.3643	73.6524

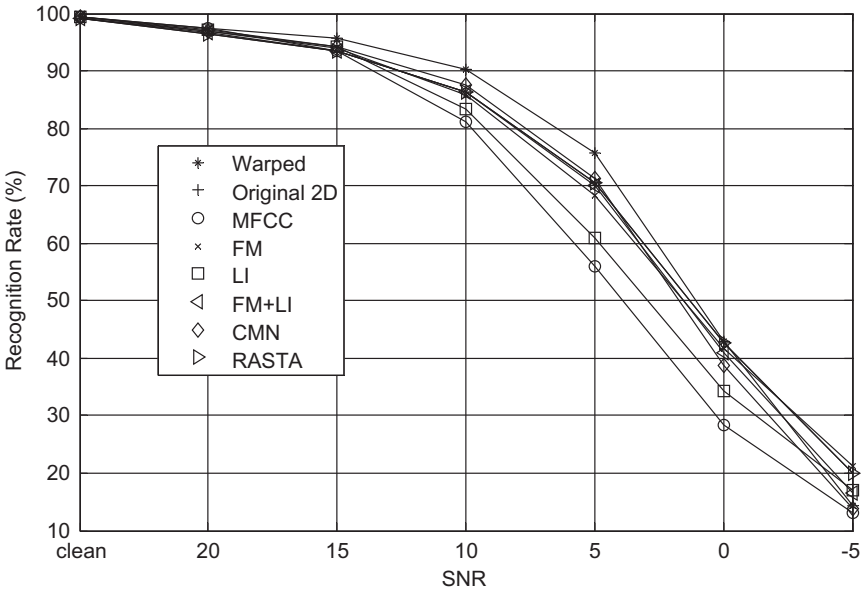


Fig. 11. Recognition rate (%) of warped 2D filter vs. MFCC(39), FM, LI, FM + LI CMVN, RASTA filter, and the original 2D filter.

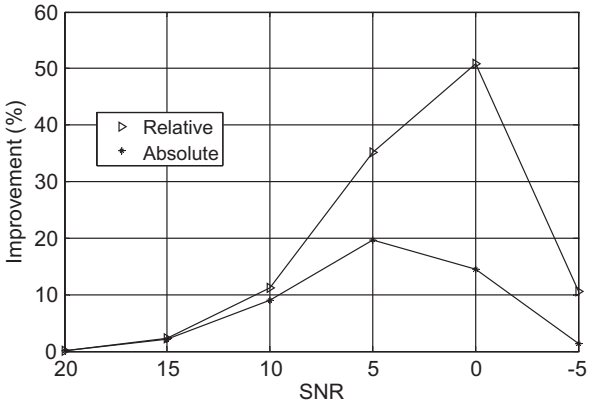


Fig. 12. Improvement of speech recognition rate.

improvement is quite impressive, 10%. Also the average of recognition rates of SNRs from 0 dB to 20 dB is given in Table 6. The relative improvement reaches 12.73%.

From the experimental results, it can be easily seen that the proposed warped 2D psychoacoustic filter works very well for input signals with mid-range SNRs, such as 5 dB and 0 dB, which is the most commonly encountered SNR range in actual environments. For example, in Table 6, the proposed algorithm yields 11.2% improvement at 10 dB, 35.1% improvement at 5 dB and 50.9% improvement at 0 dB.

5.2. FM, LI, FM + LI and CMVN

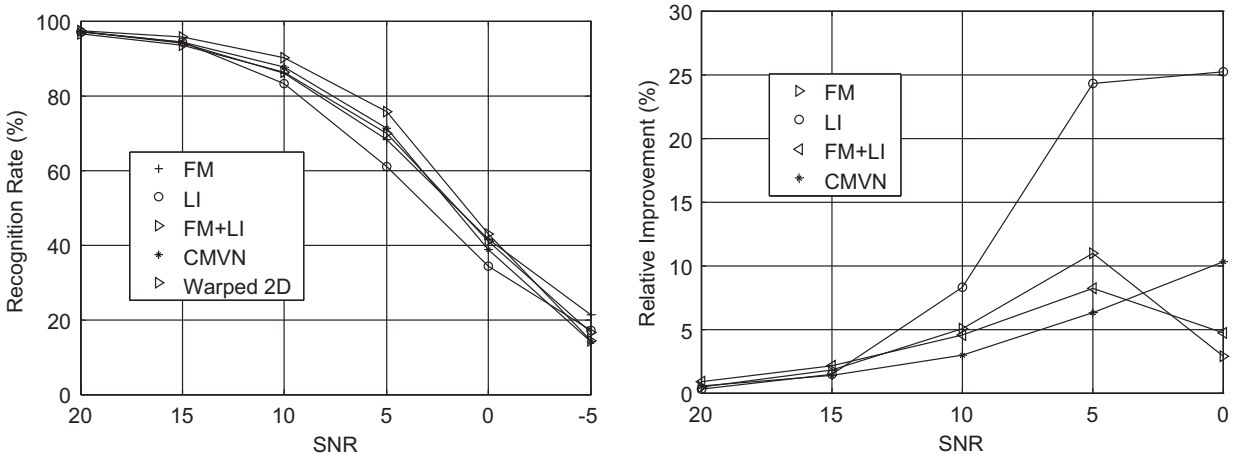
The second set of comparisons is made with forward masking (FM), lateral inhibition (LI), FM + LI and cepstral mean and variance normalization (CMVN). The detailed experiment results are listed in Table 6. Fig. 13 shows the average recognition rate over three test sets at various SNRs.

Fig. 13(a) shows the recognition rate for FM, LI, FM + LI, CMVN and the warped 2D. It can be easily seen that the proposed method hold great advantage over the others. For example, at SNR from 20 dB to 0 dB, the proposed method achieves a relative improvement of 0.51%, 1.35%, 3.00%, 6.32% and 10.32% over CMVN. Also for FM, there is great improvement, 0.46%, 1.79%, 5.03%, 10.93% and 2.87% at SNRs 20–0 dB, respectively. The most impressive increase comes from the comparison with LI, which yields 25.24% for SNR 0 dB, 24.26% for SNR 5 dB, 8.31% for SNR 10 dB, 1.44% for SNR 15 dB and 0.28% for SNR 20 dB. For FM + LI, the relative improvements are 0.8633%, 2.146%, 4.5819%, 8.1942%, and 4.67%. Undoubtedly, the proposed warped 2D psychoacoustic filter is pretty good at SNRs 0–20 dB. The relative improvement in terms of the SNR 0 dB to SNR 20 dB average is 3.91% over FM, 8.64% over LI, 3.73% over FM + LI, and 3.32% over CMVN.

Compared with CMVN, the proposed method not only performs better at SNRs 0–20 dB but also holds great advantage at the clean set and SNR −5 dB training sets. The relative improvement is 0.01% at clean set and 3.71% at SNR −5 dB. For FM and LI, although the experimental results at SNR −5 dB appear to be less competitive, it has to be noted that SNR −5 dB is also too noisy for human ear to work efficiently. Besides, in terms of the overall average recognition rate, the warped 2D filter is also the best, 73.6524%, which is better than 72.4350% of FM, 69.4774% of LI and 71.7350% of CMVN. The relative improvement is 1.68%, 6.01% and 2.67%, respectively.

5.3. Comparison with original 2D

Further comparison is made against the original 2D psychoacoustic filter, introduced in (Dai and Soon, 2009). The



(a) Recognition Results (b) Relative Improvement

Fig. 13. Comparison of warped 2D filter vs. FM, LI, FM + LI and CMVN.

experimental results are shown in Table 6 and Fig. 14. It is clear that the warped 2D psychoacoustic filter holds advantage at all SNRs.

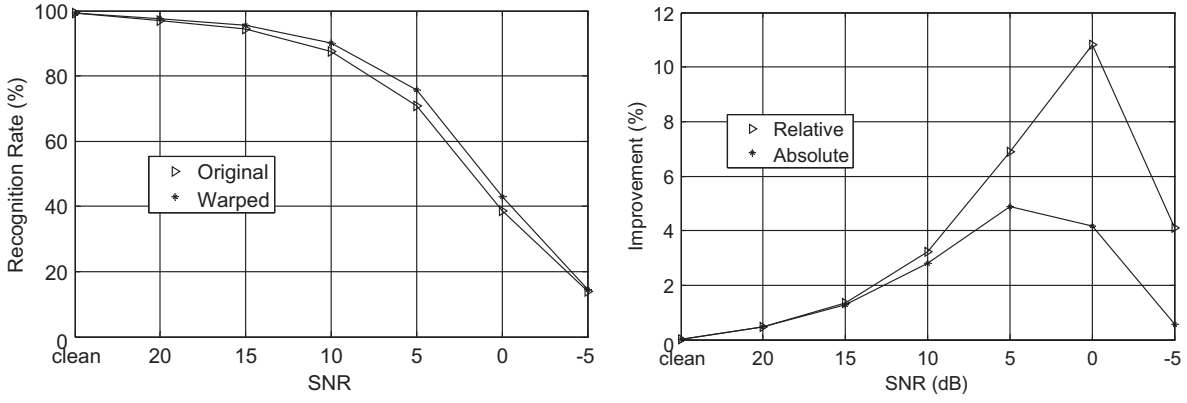
Fig. 14(a) gives the detailed recognition rates. Fig. 14(b) shows the improvement from the proposed method over the original 2D filter. The increase in the recognition rate is very impressive. In the clean training set, the warped 2D filter appears better with a relative improvement of 0.01%. Besides, for noisy conditions the relative improvement reaches 0.48% at SNR 20 dB, 1.35% at SNR 15 dB, 3.21% at SNR 10 dB, 6.89% at SNR 5 dB, 10.81% at SNR 0 dB and 4.11% at SNR −5 dB. From Fig. 14(b) it can be found that the warped 2D psychoacoustic filter greatly improved the speech recognition performance under noisy conditions especially from SNRs 20 dB to 0 dB.

After the discussion of the recognition results from different SNRs, the overall performance of the two methods is compared using two averages, Avg 0–20 and Avg All. In terms of Avg 0–20, the advantage of the proposed method is quite obvious. The recognition rate increases from 77.64% to 80.36%, yielding an relative improvement of 3.50%. Besides, the relative improvement in overall average also reaches 2.83%.

5.4. Comparison with relative spectra filter (RASTA)

The last set of comparison is made against the RASTA filter. It is a famous method which utilizes a FIR filter in the cepstral domain. The RASTA filter is applied with CMVN and is denoted as RASTA in Table 6. And the detailed comparison is shown in Fig. 15. The recognition results are given in Fig. 15(a). And the analysis of improvement is shown in Fig. 15(b).

Firstly, for the clean test set, the proposed method gives very good performance, 99.33%, which is much better than 99.08% of the RASTA filter. From Fig. 15(b), it can be easily seen that the warped 2D psychoacoustic filter performs quite well in SNRs 20–0 dB. The relative improvement is very encouraging, 1.05% at SNR 20 dB, 2.23% at SNR 15 dB, 4.40% at SNR 10 dB, 7.27% at SNR 5 dB and 0.16% at SNR 0 dB. The average of recognition results over SNR 20–0 dB is 80.36%, better than 77.94% of the RASTA filter. Although the recognition results at SNR



(a) Recognition Results (b) Relative Improvement

Fig. 14. Comparison of the warped 2D filter vs. the original 2D filter.

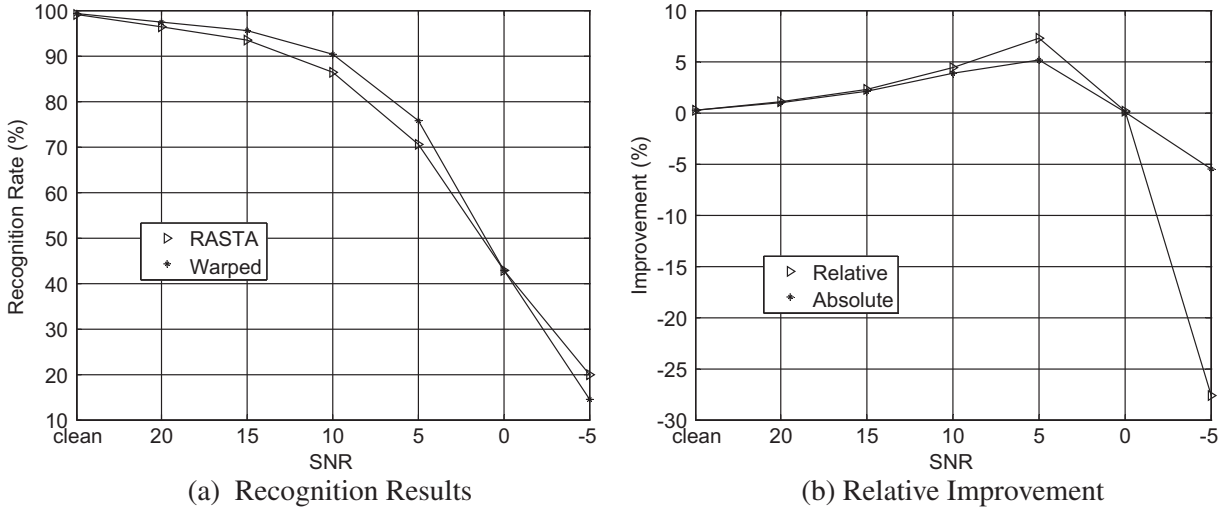


Fig. 15. Comparison of the warped 2D filter vs. the RASTA filter.

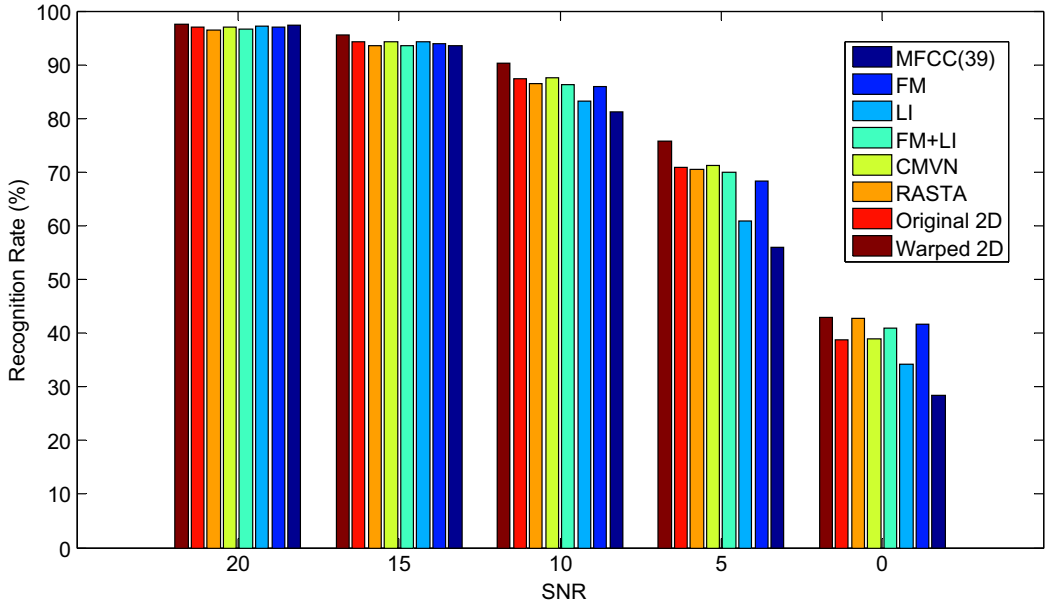


Fig. 16. Comparison of recognition rate (%) of warped 2D filter vs. MFCC(39), FM, LI, FM + LI, CMVN, RASTA filter, and the original 2D filter.

−5 dB appear to be less competitive, the overall performance of the proposed method is still promising. The overall average recognition rate of the proposed method is 73.65%, about 1.77% better than the RASTA filter.

5.5. Summary

Fig. 16 shows the recognition results from the above sections. It can be seen that the warped 2D psychoacoustic filter holds great advantage especially at SNRs 20–0 dB. The improvement obtained by the warped 2D psychoacoustic filter has some theoretical backing.

Both speech perception and the response of individual auditory nerves are extremely sensitive to the frequency position of local spectral peaks (Strope and Alwan, 1997). Therefore, sharpening cepstral peaks becomes one typical approach for the improvement of speech recognition system, which usually involves the utilization of high pass filters.

Lateral inhibition is one of the peak sharpening approach. As stated in Section 2.1, lateral inhibition filter is efficient in removing white noise which have a flat spectrum. The proposed 2D psychoacoustic filter is a high emphasis filter, which is designed to incorporate both lateral inhibition and forward masking. After this process both spectral peaks and valleys are emphasized.

6. Conclusions

In this paper, MFCC integrated with the warped 2D psychoacoustic filter have been proposed and applied to automatic speech recognition system. We evaluated this

front-end algorithm with a series of recognition experiments based on HMM on standard AURORA2 database. Comparison has been made against ordinary MFCCs, forward masking (FM), lateral inhibition (LI), cepstral mean and variance normalization (CMVN), RASTA filter, and the original 2D filter. The obtained results verify that our algorithm effectively improves the recognition rate under noisy environments. Also, because the proposed method works in the feature extraction part, it is easy to be combined with other noise-robust methods.

References

Allen, J., 1994. How do humans process and recognize speech. IEEE Trans. Speech Audio Process., 567–577.

Ambikairajah, E., Davis, A.G., Wong, W.T.K., 1997. Auditory masking and MPEG-1 audio compression. Electron. Comm. Eng. J. 9, 165–175.

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

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

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

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

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

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

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

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

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.

Kvale, M.N., Schreiner, C.E., 2004. Short term adaptation of auditory receptive fields to dynamic stimuli. J. Neurophysiol. 91, 604–612.

Leonard, R.G., 1984. A database for speaker independent digit recognition. In: Proc. ICASSP, Vol. 3, pp. 42–53.

Lois, L.E., 1962. Backward and forward masking of probe tones of different frequencies. J. Acoust. Soc. Amer. 34 (8), 1116–1117.

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

Nghia, P.T., Cuong, D.D., Binh, P.V., 2008. A new wavelet-based wideband speech coder. In: Proc. ICATC, pp. 349–352.

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

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

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.

Strope, B., Alwan, A., 1997. A model of dynamic auditory perception and its application to robust word recognition. IEEE Trans. Speech Audio Process. 5, 451–464.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7, 126–137.