Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 387–396

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An improved model of masking effects for robust speech recognition system

Peng Dai *, Ing Yann Soon

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

Received 19 December 2011; received in revised form 10 December 2012; accepted 10 December 2012
Available online 25 December 2012

Abstract

Performance of an automatic speech recognition system drops dramatically in the presence of background noise unlike the human auditory system which is more adept at noisy speech recognition. This paper proposes a novel auditory modeling algorithm which is integrated into the feature extraction front-end for Hidden Markov Model (HMM). The proposed algorithm is named LTFC which simulates properties of the human auditory system and applies it to the speech recognition system to enhance its robustness. It integrates simultaneous masking, temporal masking and cepstral mean and variance normalization into ordinary mel-frequency cepstral coefficients (MFCC) feature extraction algorithm for robust speech recognition. The proposed method sharpens the power spectrum of the signal in both the frequency domain and the time domain. Evaluation tests are carried out on the AURORA2 database. Experimental results show that the word recognition rate using our proposed feature extraction method has been effectively increased.

Keywords: Automatic speech recognition; Auditory modeling; Simultaneous masking; Temporal masking; AURORA2

1. Introduction

One of the biggest obstacles to widespread use of automatic speech recognition technology is robustness to those mismatches between training data and testing data, which include environment noise, channel distortion and speaker variability. On the other hand, the human auditory system is extremely adept at all these mentioned situations (Allen, 1994; Hermansky, 1998). Therefore, it is natural to study the auditory model and apply it to an automatic speech recognition (ASR) system. In this paper, a feature extraction front-end with noise-robust characteristics based on auditory evidence is proposed to improve the performance of an automatic speech recognition system.

Psychoacoustics is the science of quantifying the human perception of sounds, which comprises relationships between sound pressure level and loudness, human response to different frequencies, and a number of masking effects (Gold and Morgan, 2000). Understanding of these neural responses has led to auditory modeling of these nonlinear perceptions by computational models. For example, MFCC, one of the most popular feature extraction algorithms for ASR, has employed an auditory-based warping of the frequency axis called critical-band, for modeling the frequency sensitivity of human hearing, and a first order pre-emphasis filter is applied in the time domain to boost the high frequencies (Milner, 2002). This paper aims to integrate masking effects into the MFCC auditory model and to study the resulting effects when applied to automatic speech recognition.

Masking effects may be classified as simultaneous or temporal according to the occurrence of the signals. Masking effect between any two signals which occur at the same time is called simultaneous masking (frequency masking). Signals can also be masked by the preceding sound, called forward masking, or by the sound after it,

* Corresponding author. Tel.: +65 86 93746815.
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 © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.12.005

called backward masking. These masking effects are caused by the principal mechanism of neuronal signal processing both in time domain and frequency domain (Shamma, 1985a,b). Strope presented detail experimental results and mathematical model about temporal masking in (Strope and Alwan, 1997). Haque proposed a spectral subtraction based psychoacoustic algorithm and achieved very promising results in (Haque and Togneri, 2010). Recently, Zhang published a series of psychoacoustic results about a new kind of masking effects, and implemented in an ASR system (Zhang et al., 2012). All of the above mentioned algorithms are trying to implement psychoacoustics in speech processing so as to make the system more robust to noise (Strope and Alwan, 1997; Haque and Togneri, 2010; Zhang et al., 2012; Palomäki and Brown, 2011).

In our previous work (Dai and Soon, 2010, 2012), a 2D psychoacoustic filter was proposed to implement masking effects based on MFCC feature extraction algorithm. Promising results were reported (Dai and Soon, 2010, 2012). The advantage of the 2D psychoacoustic filter is that it manages to implement both temporal masking and simultaneous masking with a simple 2D mask. However, in order to combine the two masking effects, temporal masking parameters have to be re-estimated in time frequency domain, which makes the model inaccurate. Dai tried to overcome the inaccuracy by mathematical compensation in (Dai et al., 2009). This paper intends to directly implement temporal masking and simultaneous masking. The novelty of this paper lies in two parts. Firstly, the proposed algorithm integrates both simultaneous and temporal masking in MFCC feature extraction algorithm. Secondly, the computational load is nearly negligible.

Verification tests are carried out on the AURORA2 database. It is a widely used standard English database, and contains isolated digits as well as digit serials. Extensive comparison is made against MFCCs, forward masking (FM), lateral inhibition (LI), cepstral mean and variance normalization (CMVN), RASTA filter, MVA, and the temporal frequency warped (TFW) 2D filter Dai and Soon, 2010, 2012; Chen and Bilmes, 2007; Hermansky and Morgan, 1994. Significant improvements are obtained.

The rest of the paper is organized as follows. Section 2 provides an overview of simultaneous and temporal masking. The actual model of simultaneous masking and temporal masking used in the proposed algorithm is detailed in Section 3 together with a description of the application of the proposed algorithm to an automatic speech recognition system. Section 4 details the performance evaluation based on Hidden Markov Model (HMM) on AURORA2 database and discussions while Section 5 concludes the paper.

2. Masking effects

2.1. Simultaneous masking

Simultaneous masking is more popular in speech enhancement than in speech recognition. A number of papers report that simultaneous masking works for speech enhancement (Chen et al., 2007; Virag, 1999; Akbari et al., 1995). Lateral inhibition is one approach of simultaneous masking, which has been discovered through physiological, psychophysical and neurophysiological experiments (Shamma, 1985b). Cheng's paper (Cheng and O'Shaughnessy, 1991) shows that lateral inhibition can give a significant reduction of noise for noisy speech. An isolated word recognition technique based on a combination of instantaneous and dynamic features of the speech spectrum is proposed by Furui (1986). Lu et al. (2000) suggests a feature extraction front-end that consists of both MFCC and lateral inhibition functions, but the lateral inhibition masker is a separate module after the MFCC as opposed to our proposed approach whereby the lateral inhibition masker is built into the MFCC module itself.

Lateral inhibition models the sensory reception process of biological systems by which neurons are able to determine more precisely the origin of a stimulus. This reception is caused by inter-connected neurons. When the skin is touched by an object, several neurons other than the exact one will be stimulated. In order to determine the origin of the stimulus, the biological system will tend to suppress the neighboring neurons to ensure only the center neurons fire. In other words, the general function of lateral inhibition is to sharpen input changes. In this paper, we only discuss the effects of lateral inhibition in the spectral domain.

Fig. 1 is a sketch from a set of psychophysical data (Houtgast, 1972), which shows the lateral inhibition effect of a tone on another tone (with a constant intensity) as frequency varies. It is clear to conclude from Fig. 1, the subjective intensity (I) of the second tone received by human auditory system varies with its frequency. The frequency domain is mel-scaled. When we apply such a masker in the frequency domain to speech signals, it helps to distinguish the central signal from others by enhancing the spectral peaks.

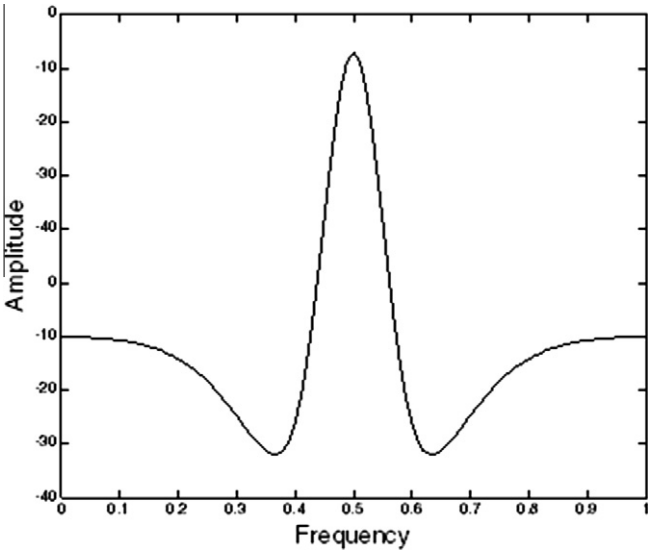

Fig. 1. Characteristic curve of lateral inhibition.

Fig. 1 shows the characteristic curve of lateral inhibition (LI). A simple proof is given below to show the masking effect of lateral inhibition on the spectral domain. Assume that a speech signal, $s(t)$, is corrupted by noise, $n(t)$, resulting in a noisy speech, $x(t)$. The relationship is given by:

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

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

In time frequency domain, Eq. (1) can be written as:

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

where f is the frequency index, $X(f,t)$, $S(f,t)$ and $N(f,t)$ are the power spectral density of $x(t)$, $s(t)$ and $n(t)$, respectively.

Let $M_{\mathrm{LI}}(f)$ represent the lateral inhibition masker, the lateral inhibition masker is modeled to satisfy the following constraint (Cheng and O'Shaughnessy, 1991),

$$\int_{-\infty}^{\infty} M_{\mathrm{LI}}(f)df = 0 \tag{3}$$

As proposed in Cheng's paper (Cheng and O'Shaughnessy, 1991), the LI masker was implemented in the speech power spectrum domain. However, in this paper it is directly applied to the complex noisy speech spectrum. Similarly,

$$\begin{aligned}\hat{X}(f,t) &= \int_{-\infty}^{\infty} X(f,t)M_{\mathrm{LI}}(f)df \\ &= \int_{-\infty}^{\infty} S(f,t)M_{\mathrm{LI}}(f)df + \int_{-\infty}^{\infty} N(f,t)M_{\mathrm{LI}}(f)df \\ &= \hat{S}(f,t) + \int_{-\infty}^{\infty} N(f,t)M_{\mathrm{LI}}(f)df\end{aligned} \tag{4}$$

If noise is stationary,

$$\int_{-\infty}^{\infty} N(f,t)M_{\mathrm{LI}}(f)df \approx 0 \tag{5}$$

Therefore,

$$\hat{X}(f,t) = \hat{S}(f,t) + \int_{-\infty}^{\infty} N(f,t)M_{\mathrm{LI}}(f)df \approx \hat{S}(f,t) \tag{6}$$

The advantage of the above mentioned algorithm lies in two parts. Firstly, from Eqs. (4) and (6), the algorithm not only sharpens the spectrum of the input signal but also removes noise. Furthermore, because the LI masker is implemented in time frequency domain, the proposed approach is able to make use of phase information.

$$\begin{aligned}\hat{X}(f,t) &= \int_{-\infty}^{\infty} X(f,t)M_{\mathrm{LI}}(f)df \\ &= \int_{-\infty}^{\infty} X_r(f,t)M_{\mathrm{LI}}(f)df + j \\ &\quad \times \int_{-\infty}^{\infty} X_i(f,t)M_{\mathrm{LI}}(f)df\end{aligned} \tag{7}$$

where $X_r(f,t)$ and $X_i(f,t)$ are the real component and imaginary components of $X(f,t)$; j is the imaginary unit.

2.2. Temporal masking

Temporal masking comprises of both forward masking and backward masking. Forward masking describes how a clearly audible sound is affected by the sound before it, while backward masking shows how it is affected by the sound after it. Because forward masking is much more effective than backward masking, only forward masking is taken into consideration in later discussion. Forward masking reveals that over short durations, the usable dynamic range of the human auditory system depends on the spectral characteristics of the previous stimuli. According to the results of pure tone forward masking experiments from Jesteadt et al. (1982), a probe following a masker is less audible than a probe following silence. Therefore, the amount of forward masking can be viewed as a consequence of auditory adaptation to the masker which can be expressed as (Zhang et al., 2012),

$$P = M_{\mathrm{FM}}(1-m)(1-b^d)a^u \tag{8}$$

where P is masking threshold, u is the duration between masker and signal, d is the masker duration, M_{FM} is the magnitude of a forward masking effect masker at certain frequency and m, a, b are frequency dependent constants.

An example of the behavior of forward masking is given in Fig. 2. The input is a series of lower intensity impulses following a higher intensity square pulse. The output shows that the signal starts to adapt at the rising edge of the square pulse, followed by the recovery process. The adaption of the lower intensity impulses to the square pulse rides on top of the square pulse's recovery. After adapting to the square pulse, the impulses take time to readjust and hence recovery time is necessitated before the relatively less intense probe becomes audible. The duration is also a function of the duration of the masker, thereby reflecting the time required for the auditory system to adapt completely to the masker.

Models of adaptation have been successfully applied in ASR (Zhang et al., 2012; Tchorz and Kollmeier, 1999; Oxenham, 2001). Adaptation and the popular RASTA techniques have obvious similarities to each other (Hermansky and Morgan, 1994). However, forward masking has a recovery step after adaptation as reported by Oxenham (2001). Strope and Alwan (1997) and Zhang et al. (2012) have demonstrated forward masking and its application to ASR. Tchorz and Kollmeier (1999) tested various adaptation parameters in ASR system. Recently, a simple adaptation model has been developed and evaluated on the AURORA2 database and AURORA3 database (Holmberg et al., 2006).

Most of the studies have been done either on simultaneous masking or temporal masking (Shamma, 1985a; Cheng and O'Shaughnessy, 1991; Park and Lee, 2003), while only some preliminary research in the complete auditory model has been done so far. In Park and Lee's paper (Park and Lee, 2003), a Mexican-hat convolutional filter is used to simulate a lateral inhibition masker which neglects

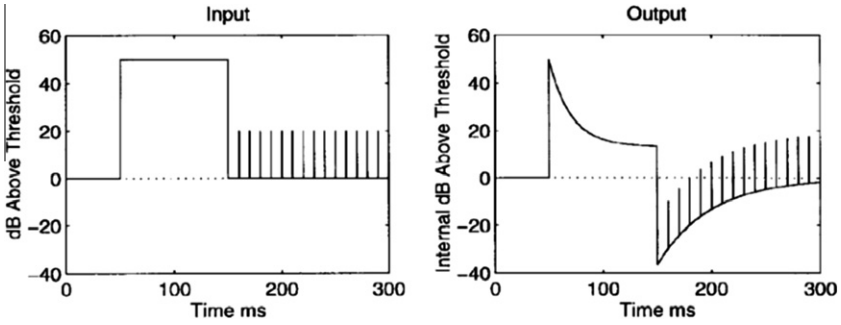

Fig. 2. Example of recovery process for forward masking.

the frequency asymmetry of the masking effect, rendering it less accurate. Regarding temporal masking, they only implemented forward masking. In addition, they only use a simple nearest-neighbor classifier as the recognizer in the experiments. A simple recognizer can register a better percentage improvement in performance given its lower base value reading compared to a sophisticated recognizer which can yield a higher base value reading to start with. In comparison, we propose a feature extraction front-end integrated with a complete auditory characteristic model with a computational load that is nearly negligible.

3. Algorithm description

The proposed front-end feature extractor is modified from the MFCC model provided by Voicebox (Brookes, 0000) by integrating the lateral inhibition masking, temporal spectral averaging and forward masking. The whole proposed new MFCC model with all three masking effects is given on the bottom of Fig. 3, compared to the traditional MFCC model shown on the top of Fig. 3.

3.1. An efficient and simple lateral inhibition model (LI)

The continuous lateral inhibition model based on mel-frequency scale requires a large amount of computation to be applied to a feature extractor (Park and Lee, 2003). For a practical and feasible system, this paper proposes a simplified lateral inhibition model for front-end feature extraction. Although lateral inhibition is frequency dependent, the frequency discrimination of the human auditory system is approximately logarithmic in frequency. This means that by using mel-frequency scale, lateral inhibition can be approximated to a mel-frequency independent masker. Such an approximation therefore simplifies the lateral inhibition filter by reducing the number of variables. Fig. 4 shows our simplified lateral inhibition filter as a function of mel-frequency abscissa and the magnitude of its parameters, constrained by zero mean condition.

According to Cheng's paper (Cheng and O'Shaughnessy, 1991), $\Delta f = f_2 - f_1 = f_3 - f_2 = f_4 - f_3 = 1$, $m_1 = -0.6$, $m_2 = 1$ and $m_3 = -0.4$ is a set of optimal values, which corresponds to the asymmetrical characteristic of lateral inhibition in the frequency domain.

$$[M_{-2} \quad M_{-1} \quad M_0 \quad M_1 \quad M_2] = [-0.6 \quad 0 \quad 1 \quad 0 \quad -0.4] \tag{9}$$

However, by implementing this set of masking value directly onto AURORA2 database, a poorer recognition result is obtained than the baseline due to the over attenuation of some coefficients with speech information. In order to reduce the over attenuation effects, a combination of the original signals and the lateral inhibition outputs is used as the input signal to a recognizer.

Let $P(f,t)$ represent the power spectral density of the resultant signal of a weighted combination of the original signal and the lateral inhibition masker output. Let $P_x(f,t)$ represent the power spectral density of a speech signal $x(t)$, and M_k be the power spectral density of a lateral inhibition masker $m(k)$. k is the index shown as subscript in Eq. (9).

$$\begin{aligned} P(f,t) &= (1-\beta)P_x(f,t) + \beta\sum_{j=-N}^{N} M_j P_x(f+j,t) \\ &= \beta M_{-2}P_x(f-2,t) + \beta M_{-1}P_x(f-1,t) \\ &\quad + (1-\beta+\beta M_0)P_x(f,t) + \beta M_1 P_x(f+1,t) \\ &\quad + \beta M_2 P_x(f+2,t) \end{aligned} \tag{10}$$

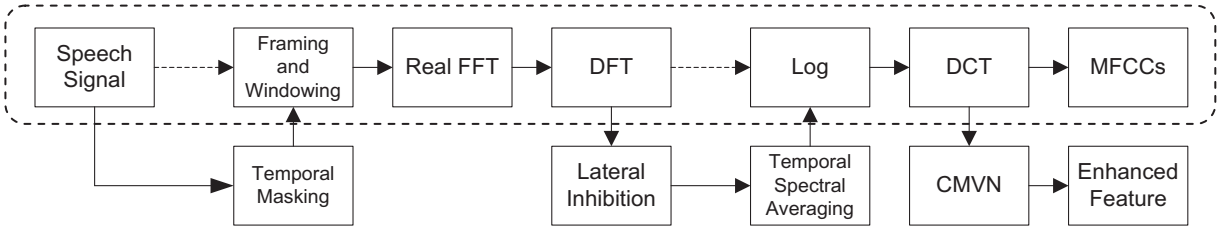

Fig. 3. Diagram of the proposed algorithm.

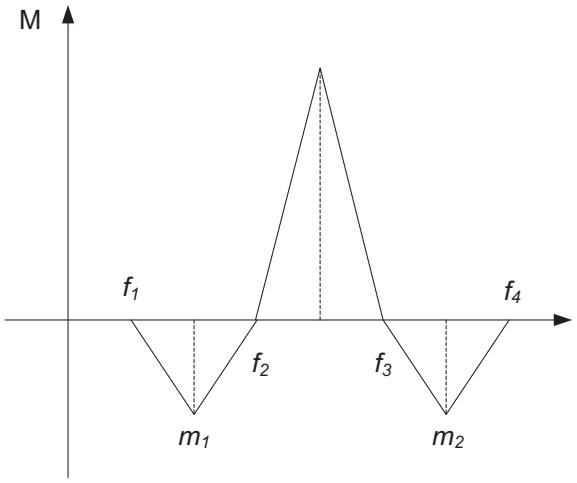

Fig. 4. Simplified lateral inhibition model.

where j indicates different frequency components of a masker. $j = [-N, N]$, where $N = \frac{L_{mas\,ker}-1}{2}$. β is a percentage of the lateral inhibited signal out of the combination signal.

The parameters, N and β, are tuned by maintaining the masker value given in Cheng's paper (Cheng and O'Shaughnessy, 1991). After a series of experiments, the optimal values for the recognition results are obtained, $N = 2$ and $\beta = 0.1$. Therefore, the final lateral inhibition filter is

$$[M_{-2} \quad M_{-1} \quad M_0 \quad M_1 \quad M_2] = [-0.06 \quad 0 \quad 1 \quad 0 \quad -0.04] \tag{11}$$

The outputs after lateral inhibition masking may have some negative values, which means all of them are inhibited fully and they are below the audible threshold. Since we are proposing a masking effect based on human auditory system, any signal below audible threshold is meaningless to the recognition process. Hence a rectifier is used after the lateral inhibition masking to set all negative outputs to zero.

3.2. Temporal spectral averaging

Lateral inhibition is very sensitive to spectral changes and it will enhance the peaks in both speech signals and noise signals. Noise signals which do not have fairly stationary segments in the power spectrum, such as subway noise, will be enhanced together with speech signals thereby greatly degrading the recognition results. Hence, a method is needed to alleviate this effect and that is to employ temporal spectral averaging in spectral domain:

$$\bar{P}(f,t) = \frac{1}{2N_{tsa}+1} \sum_{m=-N_{tsa}}^{N_{tsa}} Z_m P(f, t+m) \tag{12}$$

In our present implementation, the parameters are set as reported in Cheng's paper (Cheng and O'Shaughnessy, 1991), $N_{tsa} = 2$ and $[lllllZ_{-2} \quad Z_{-1} \quad Z_0 \quad Z_1 \quad Z_2] = [lllll0.4 \quad 1.3 \quad 1.6 \quad 1.3 \quad 0.4]$.

Table 1
Temporal masking parameters.

Frequency	Parameter values		
	a	b	m
250	0.864	0.474	0.19
500	0.854	0.510	0.20
1000	0.816	0.543	0.26
2000	0.851	0.525	0.29
4000	0.858	0.507	0.34

3.3. Temporal masking

According to Strope's paper (Zhang et al., 2012), the parameters for temporal masking are frequency dependent and are listed in Table 1. For simplicity, only the 2 kHz parameters are used for system implementation.

4. Design of automatic speech recognition experiments

4.1. Database overview

Performance of the proposed front-end is evaluated on the AURORA2 database (Hirsch and Pearce, 2000) which contains speech data sampled at 8 kHz. There are two training conditions in AURORA2. The clean training set in this database has no noise added and it consists of 8,440 utterances recorded from 55 male and 55 female adults. 4,004 utterances from 52 male and 52 female speakers are split equally into 4 subsets with 1,001 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, four types of noises (subway, babble, car, and exhibition hall) are added to the four different clean data subsets at SNRs from −5 to 20 dB with 5 dB step size. So there are 28,028 utterances in this test set. In test set B, 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. In test set C, subway and street noises filtered by a MIRS (modified Intermediate Reference System) filter which simulates the frequency characteristics of a telecommunication terminal are added to the subsets of clean data. 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. Feature extraction front end

4.2.1. Baseline system

We used VoiceBox Toolkit reference code for MFCC to calculate 13 cepstral coefficients including log energy term (logE). Furthermore, delta coefficients and acceleration coefficients are computed using

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

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)

We also tried cepstral mean normalization on our system, since it is a commonly used feature enhancement method for robust speech recognition (Mokbel et al., 1996; Zhu and O'Shaughnessy, 2005). We applied the Hidden Markov Model Toolkit (HTK) to calculate the ceptral mean normalization in our experiment. The mean of MFCCs across frames is calculated by

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

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.

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

This technique is very effective in practice where it compensates for long term spectral effects such as communication channel distortion.

4.3. Recognizer back end

The same recognizer is used for both the proposed front-end feature extraction algorithm and the comparison targets for a meaningful comparison. Each digit is modeled by a simple left-to-right 18 states (including two non-emitting states) HMM model. There are three Gaussian mixtures per state. Two pause models are defined. One is "sil". It has 3 HMM states and models the pauses before and after each utterance. The other one is "sp", which is

Table 2
Experimental results for LTFC (%).

	SNR	Set A				Set B				Set C		Avg
		Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train	Subway	Street	
Clean training	Clean	99.08	98.97	99.19	99.20	99.08	98.97	99.19	99.20	98.96	98.88	99.07
	20 dB	97.85	98.13	98.33	97.44	98.34	97.43	98.39	97.93	96.96	96.98	97.78
	15 dB	95.92	96.92	96.75	94.63	97.48	96.10	97.32	96.85	94.47	94.89	96.13
	10 dB	92.66	94.11	93.38	89.48	94.50	91.72	95.35	93.27	89.93	89.42	92.38
	5 dB	84.43	84.22	83.63	78.19	86.67	81.80	86.46	84.70	78.42	77.33	82.59
	0 dB	66.23	58.89	59.23	56.77	64.05	60.61	62.66	61.71	55.82	55.53	60.15
	−5 dB	35.43	28.96	27.05	32.98	33.71	31.44	29.85	29.22	24.56	26.78	30.00
	Avg 0–20	87.42	86.45	86.26	83.30	88.21	85.53	88.04	86.89	83.12	82.83	85.81
Multi training	Clean	98.25	98.19	98.15	98.46	98.25	98.19	98.15	98.46	98.28	98.28	98.27
	20 dB	98.53	98.61	98.51	98.43	98.68	98.10	98.66	98.98	98.25	98.28	98.51
	15 dB	97.67	98.19	98.00	97.38	98.4	97.76	98.15	98.12	97.36	97.82	97.90
	10 dB	96.35	97.22	96.51	94.82	97.39	96.10	97.2	96.51	95.79	95.53	96.34
	5 dB	93.15	92.62	92.54	88.77	92.97	91.44	93.2	92.66	91.31	90.15	91.88
	0 dB	82.19	75.21	80.17	73.90	77.31	78.14	80.52	79.08	76.94	75.79	77.93
	−5 dB	55.54	43.56	48.46	51.00	48.17	50.06	48.76	51.25	47.96	46.25	49.10
	Avg 0–20	93.58	92.37	93.15	90.66	92.95	92.31	93.55	93.07	91.93	91.51	92.51

Table 3
Clean training condition.

SNR/dB	Clean	20	15	10	5	0	−5	Avg 0–20
MFCC(39)	99.36	97.37	93.51	81.16	56.02	28.39	13.04	71.29
FM	99.03	97.02	93.91	85.89	68.24	41.65	21.30	77.34
LI	99.42	97.19	94.23	83.29	60.92	34.21	17.07	73.97
TSA	99.38	97.48	94.13	83.86	61.41	33.96	17.58	74.17
FM + LI + TSA	99.06	96.63	93.59	86.26	69.97	40.94	16.47	77.48
CMVN	99.32	96.97	94.32	87.59	71.20	38.84	13.90	77.78
RASTA	99.08	96.45	93.51	86.41	70.57	42.78	19.93	77.94
MVA	99.21	97.33	95.28	90.55	80.01	57.59	27.38	84.15
AFE	99.20	97.19	94.98	89.56	77.07	52.33	24.77	82.23
TFW 2D	99.38	97.76	95.94	91.61	80.42	56.26	24.37	84.40
LTFC	99.07	97.78	96.13	92.38	82.59	60.15	30.00	85.81

a single state model (tied with the middle state of "sil") and models the pauses among words.

5. Results and discussion

5.1. Evaluation on the AURORA2 database

Table 2 presents the experimental results for the proposed LTFC (LI + TSA + FM + CMVN) algorithm. LI stands for lateral inhibition. TSA stands for temporal spectral averaging. FM means forward masking. CMVN refers to Cepstral Mean and Variance Normalization. The performance metric used is word recognition rate in percentage. Results are captured for both the clean and multi-condition training conditions. For the latter, the result is averaged over the three noisy test sets with SNRs ranging from 0 to 20 dB denoted as Avg 0–20.

The first set of comparison is made against MFCC, FM, LI, FM + LI + TSA, and CMVN. The second set of comparison is made against three state of the art front-end noise removal algorithms, RASTA (Hermansky and Morgan, 1994), MVA (Chen and Bilmes, 2007), and the temporal frequency warped 2D psychoacoustic filter (Dai and Soon, 2012). Tables 3 and 4 give the detailed experimental results for clean training condition and multi training condition, respectively. Since the recognition results for clean subset are all about 99%, it is meaningless to discuss improvement at such level. Only the experimental results for noisy condition are used for further discussion.

5.1.1. Comparison of different parts of LTFC

The first set of comparison is made against different parts of the proposed LTFC algorithm. From the experimental results in Tables 3 and 4, it can be seen that FM, LI, TSA, and CMVN all help to improve the performance of speech recognition system. All of them manage to obtain better results than the baseline MFCC(39) system. The relative improvements of the proposed LTFC over the other algorithms are given in Table 5. The relative improvements are calculated in terms of word recognition rate. It can be found out that for all noisy conditions the proposed LTFC algorithm performs better than all the above mentioned algorithms.

Compared with the baseline MFCC system, the proposed algorithm achieves very impressive results. The relative improvements for clean training condition are 0.42% at SNR 20 dB, 2.80% at SNR 15 dB, 13.82% at SNR 10 dB, 47.43% at SNR 5 dB, 111.87% at SNR 0 dB, and 130.06% at SNR −5 dB. For multi training condition, the relative improvements are 0.34%, 0.31%, 0.86%, 4.87%,

Table 4
Multi condition training.

SNR/dB	Clean	20	15	10	5	0	−5	Avg 0–20
MFCC(39)	99.11	98.18	97.60	95.52	87.61	60.37	26.83	87.85
FM	98.74	98.16	97.47	95.25	87.19	59.32	25.46	87.48
LI	99.13	98.19	97.62	95.53	88.06	61.93	26.59	88.26
TSA	99.24	98.38	97.82	95.50	88.03	61.20	26.93	88.19
FM + LI + TSA	98.59	98.11	97.43	95.06	87.90	64.16	28.56	88.53
CMVN	98.94	98.51	97.89	96.27	91.06	74.81	42.63	91.71
RASTA	98.60	98.41	97.57	95.86	90.23	74.31	44.75	91.27
MVA	99.04	98.49	97.87	96.15	91.13	77.49	49.04	92.22
AFE	99.14	98.50	97.89	96.31	91.52	74.94	42.33	91.83
TFW 2D	98.87	98.35	97.80	96.09	91.09	75.73	43.86	91.81
LTFC	98.27	98.51	97.90	96.34	91.88	77.93	49.10	92.51

Table 5
Relative improvements of LTFC (%).

	SNR/dB	20	15	10	5	0	−5	Avg 0–20
Clean training condition	MFCC(39)	0.42	2.80	13.82	47.43	111.87	130.06	20.37
	FM	0.78	2.36	7.56	21.03	44.42	40.85	10.95
	LI	0.61	2.02	10.91	35.57	75.83	75.75	16.01
	TSA	0.31	2.12	10.16	34.49	77.12	70.65	15.69
	FM + LI + TSA	1.19	2.71	7.09	18.04	46.92	82.15	10.75
	CMVN	0.84	1.92	5.47	16.00	54.87	115.83	10.32
Multi training condition	MFCC(39)	0.34	0.31	0.86	4.87	29.09	83.00	5.30
	FM	0.36	0.44	1.14	5.38	31.37	92.85	5.75
	Li	0.33	0.29	0.85	4.34	25.84	84.66	4.82
	TSA	0.13	0.08	0.88	4.37	27.34	82.32	4.90
	FM + LI + TSA	0.41	0.48	1.35	4.53	21.46	71.92	4.50
	CMVN	0	0.01	0.06	0.88	4.13	15.15	0.84

Table 6
Relative improvements of LTFC (%).

	SNR/dB	20	15	10	5	0	−5	Avg 0–20
Clean	RASTA	1.38	2.80	6.91	17.03	40.60	50.53	10.10
	MVA	0.46	0.89	2.02	3.22	4.45	9.57	1.97
	AFE	0.61	1.21	3.15	7.16	14.94	21.11	4.35
	TFW 2D	0.02	0.20	0.84	2.70	6.91	23.10	1.67
Multi	RASTA	0.10	0.34	0.50	1.83	4.87	9.72	1.36
	MVA	0.02	0.03	0.20	0.82	0.57	0.12	0.31
	AFE	0.01	0.01	0.03	0.39	3.99	15.99	0.74
	TFW 2D	0.16	0.10	0.26	0.87	2.91	11.95	0.76

Table 7
Relative improvements of LTFC for selected noise types (%).

SNR/dB	Noise	20	15	10	5	0	−5	Avg 0–20
RASTA	Babble	0.81	2.80	6.92	17.40	47.82	57.31	10.38
MVA		0.59	1.23	2.61	6.34	8.59	18.69	3.31
AFE	Restaurant	0.75	1.44	3.77	10.87	15.26	25.04	5.38
TFW 2D		0.01	0.47	1.54	5.80	11.2	17.29	2.98

29.09%, and 83.00% for 20B, 15B, 10, 5, 0, and −5 dB, respectively.

In clean training condition results, the proposed algorithm manages to obtain significant improvements. At SNR 20 and 15 dB, all the above mentioned algorithm get about 98% recognition rate. The relative improvements are about 1% at SNR 20 dB over FM, LI, TSA, FM + LI + TSA, and CMVN. At SNR 15 dB the relative improvements are about 2% compared with the above mentioned comparison targets. At SNR 10 dB, the relative improvements become very impressive, 2.36% over FM, 10.91% over LI, 10.16% over TSA, 2.71 over FM + LI + TSA, and 5.47 over CMVN. As SNR further drops, the relative improvements become quite large. At SNR 5 dB, the proposed relative improvements range from 16% (CMVN) to 35.57% (LI). At SNR 0 and −5 dB, the relative improvements becomes over 50%. For CMVN, the improvement even reaches 115.83%.

For multi training condition results, the recognition rate for SNR 20, 15, 10, and 5 dB are all above 90%. Therefore, the relative improvements appear to be small. It has to be noted that CMVN performs very well in multi training condition. Except SNR 0 and −5 dB, the relative improvements are less than 1%. However, the proposed algorithm still performs better, especially at low SNR level. The relative improvements are 4.13% at SNR 0 dB and 15.15% at SNR −5 dB.

For FM, LI, TSA, and FM + LI + TSA, the relative improvements for SNR 20 and 15 dB are all less than 1%. For SNR 10 dB, the relative improvements are around 1%. Some are larger than 1%, such as FM 1.14% and FM + LI + TSA 1.35%. At SNR 5 dB, the relative improvements range from 4.34% to 5.38%. At SNR 0 and −5 dB, the relative improvements become very impressive. The relative improvements are all over 20% at 0 dB and over 70% at SNR −5 dB.

5.1.2. Comparison with other algorithms

The second set of comparison is made against three state of the art front-end noise removal algorithms. Experimental results are given in Tables 3 and 4. The relative improvements are given in Table 6. It can be easily found out that the proposed algorithm successfully improves the performance of speech recognition system. As SNR decreases the improvement becomes more and more significant.

For the clean training condition results, the advantage of the proposed LTFC algorithm is very obvious. At SNR 20 dB the recognition rate is 97.78%. The relative improvements over RASTA, TFW 2D, AFE and MVA are 0.10%, 0.02%, 0.61% and 0.46% respectively. For SNR 15 dB, the relative improvements are 2.80%, 0.20%, 1.21% and 0.89%. As SNR further drops down to 10 dB, the relative improvements become 6.91%, 0.84%, 3.15 and 2.02% for RASTA, TFW 2D, AFE and MVA respectively. For SNR 5 and 0 dB, the relative improvements are 17.03% and 40.60% for RASTA, 2.70% and 6.91% for TFW 2D, 7.16% and 14.94% for AFE as well as 3.22% and 4.45% for MVA. The largest improvements come at SNR −5 dB. The relative improvements are 10.10% for RASTA, 1.97% for MVA, 4.35% for AFE and 1.67% for TFW 2D.

For multi training condition, as discussed in previous sections at SNR 20–5 dB the recognition rates are all over 90%. Therefore, the relative improvements are small at high SNR levels. However, there are still improvements. At SNR 20 dB the relative improvements are 0.10% over RASTA, 0.16% over TFW 2D, 0.01% over AFE and 0.02% over MVA. At SNR 15 dB, the relative improvements are 0.34% over RASTA, 0.10% over TFW 2D, 0.01% over AFE and 0.03% over MVA. For SNR 10 dB, the relative improvements are 0.50% over RASTA, 0.26% over TFW 2D, 0.03% over AFE and 0.20% over MVA. For SNR 5 dB, the relative improvements are 1.83% over RASTA, 0.87% over TFW 2D, 0.39% over AFE and 0.82% over

MVA. When it comes to SNR 0 dB, the relative improvements become more obvious. The relative improvements are 4.87% over RASTA, 2.91% over TFW 2D, 3.99% over AFE and 0.57% over MVA. At SNR −5 dB, the largest improvements are observed, 9.72% over RASTA, 11.95% over TFW 2D, 15.99% over AFE and 0.31% over MVA.

The performance of noise robust algorithms varies for different noise types. For example, the proposed LTFC algorithm performs better for restaurant as well as airport noise and relatively worse at exhibition noise, for high SNR level (>0 dB). Therefore, bigger improvements can be obtained for certain noise types. For MVA and RASTA, the largest improvements are for babble noise, given in Table 7. As for AFE and TFW 2D, the relative improvements for restaurant noise are better. The relative improvements are significant for speech recognition systems under noisy environment, hence vindicating that our algorithm has made the traditional recognizer more robust to the noisy condition.

To illustrate the negligible computational load of the proposed method, let us take lateral inhibition masker as an example. Regardless of programming language, to implement masking is a convolution process. For the proposed lateral inhibition masking, the convolution can be implemented using only two multiplications per frequency point i (see Eq. (10)), since the center weight is one and the other two weights are zero which therefore do not require any multiplications. Since there are 24 mel-filter coefficients, only 44 multiplications are required per 30 ms frame after taking endpoint conditions into account. Similarly, there are 104 multiplications for TSA and 64 multiplications for TM per 30 ms frame. Table 8 gives an example of the processing time of the same set of data. The sample data are taken from the AURORA2 database. Clean1 and Clean2 are two clean test sets from Set A. The processing time is measured as elapsed time using Matlab 2008b, on Microsoft Windows XP system (Intel(R) Core(TM)2 Quad CPU @ 2.66 Hz). It can be seen that LTFC is more computationally efficient.

5.2. Reverberant speech

Reverberation is the persistence of sound in a particular space after the original sound is produced (Valente et al., 2000). It is created when a sound is produced in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the walls and air (Valente et al., 2000). As time passes, the volume of the many echoes is reduced until the echoes cannot be heard at all. Reverberation is one of the most important contamination sources of speech. Therefore, it is necessary to evaluate the performance of the proposed algorithm in reverberant speech conditions.

Table 8
Processing time for different algorithms (s).

	Clean1	Clean2
RASTA	12.23	12.68
MVA	15.46	15.78
AFE	13.66	14.02
TFW 2D	13.70	13.94
LTFC	11.86	12.09

Table 9
Experimental results for reverberant speech (%).

	Clean	Avg 0–20	−5
MFCC(39)	97.17	69.34	16.13
RASTA	94.65	71.18	18.32
MVA	96.73	78.04	22.05
AFE	96.73	75.70	20.48
TFW 2D	98.49	78.81	21.70
LTFC	98.23	81.41	26.29

Table 10
Experimental results (%).

	Office	Fan
MFCC(39)	85	55
LTFC	90	85

The reverberant speech is generated using the algorithm proposed in McGovern's paper (McGovern, 0000). It uses a simple image method to calculate a room impulse response for reverberant speech (McGovern, 0000). All the speech files are processed by the algorithm to generate the corresponding reverberant speech. The HMMs are trained based on the original training material of the AURORA2 database (no reverberation). In our current implementation the reverberation time (RT) is about 300 ms. Table 9 gives the clean training recognition results for the reverberation version (processed by McGovern's algorithm) of AURORA2 test sets. It can be seen that the proposed algorithm achieves very promising results. For Avg 0–20, the relative improvements are 17.41% over MFCC, 14.37% over RASTA, 4.32% over MVA, 7.54% over AFE and 3.30% over TFW 2D. When it comes to SNR −5 dB, the relative improvements are more significant, 62.99% over MFCC, 43.50% over RASTA, 19.23% over MVA, 28.37% AFE and 21.15% over TFW 2D.

5.3. Evaluation on real life noisy speech

The noisy speech in the AURORA2 database is generated by artificially adding real life noise to the clean speech. Therefore, it is very beneficial to test the speech on speech in real noise. Two non-native English speaking adults are chosen to read a series of materials based on the contents of the AURORA2 database. Two different types of background are chosen. The first set is in a normal office, SNR ≈ 15 dB. The second one is in the presence of a noisy electric fan, SNR ≈ 6 dB. Each person is required to read 10 sentences, each consisting of a number of connected digits. Therefore, the total test material consists of 20 sentences.

Table 10 gives the experimental results of the above mentioned test data. The HMMs are trained using clean

training condition materials of the AURORA2 database. The table gives the percentage of correctly recognized sentences. It can be seen that the proposed algorithm works much better than the baseline system in real life noise.

6. Conclusion

In this paper, a MFCC feature extraction front-end integrated with lateral inhibition masking, temporal spectral averaging, forward masking and cepstral mean normalization have been proposed and applied to automatic speech recognition system. The proposed algorithm is evaluated with a series of classification experiments based on HMM using standard AURORA2 database. The obtained results verify that the proposed algorithm effectively improves the recognition rate under noisy environments. It has to be noted that the proposed method is negligible in terms of computational overhead. The idea can be extended in future work by implementing more psychoacoustic effects to make the frontend algorithm closer to the human auditory system. Such effects can include the absolute threshold of hearing, asymmetry of masking etc.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2012.12.005.

References

Akbari, A., Azirani, Bouquin, L., Jeannes, R., Faucon, G., 1995. Optimizing speech enhancement by exploiting masking properties of the human ear. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 800–803.

Allen, J., 1994. How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2, 567–577.

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

Chen, C.P., Bilmes, J.A., 2007. MVA processing of speech features. IEEE Trans. Audio Speech Lang. Process. 15, 257–270.

Chen, J., Benesty, J., Huang, Y., 2007. On the optimal linear filtering techniques for noise reduction. Speech Comm. 49, 305–316.

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., 2010. A temporal warped 2D psychoacoustic modeling for robust speech recognition system. Speech Comm. 53, 229–241.

Dai, P., Soon, I.Y., 2012. A temporal frequency warped (TFW) 2D psychoacoustic filter for robust speech recognition system. Speech Comm. 54, 402–413.

Dai, P., Soon, I.Y., Yeo, C.K., 2009. 2D psychoacoustic filtering for robust speech recognition. In: Proc. ICICS, Macau, pp. 1–5.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Process. 34, 52–59.

Gold, B., Morgan, N., 2000. Speech and Audio Signal Processing–Processing and Perception of Speech and Music. John Wiley and Sons Inc..

Haque, S., Togneri, R., 2010. A psychoacoustic spectral subtraction method for noise suppression in automatic speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1618–1621.

Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25, 3–27.

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

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluations of speech recognition system under noisy conditions. In: Proc. ISCA ITRW ASR, pp. 17–21.

Holmberg, M., Gelbart, D., Hemmert, W., 2006. Automatic speech recognition with an adaptation model motivated by auditory processing. IEEE Trans. Speech Audio Process. 14, 43–49.

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

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

Lu, X., Li, G., Wang, L., 2000. Lateral inhibition mechanism in computational auditory model and its application in robust speech recognition. In: Proc. IEEE Signal Processing Society Workshop, vol. 2, pp. 785–794.

McGovern, S., A Model for Room Acoustics. Available: <http://sgm-audio.com/research/rir/rir.html>.

Milner, B., 2002. A comparison of front-end configuration for robust speech recognition. In: Proc. ICASSP, pp. 797–800.

Mokbel, C., Jouvet, D., Monne, J., 1996. Deconvolution of telephone line effects for speech recognition. Speech Comm. 19, 185–196.

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

Palomäki, K.J., Brown, G.J., 2011. A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise. Speech Comm. 53, 924–940.

Park, K.Y., Lee, S.Y., 2003. An engineering model of the masking for the noise-robust speech recognition. Neurocomputing 52–54, 615–620.

Shamma, S.A., 1985a. Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. J. Acoust. Soc. Amer. 78, 1612–1621.

Shamma, S.A., 1985b. 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, 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.

Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automation speech recognition. J. Acoust. Soc. Amer. 106, 2040–2050.

Valente, M., Hosford-Dunn, H., Roeser, R.J., 2000. Audiology: Diagnosis, Treatment and Practice Management. Thieme Medical Publishers, New York.

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

Zhang, B., Long, Z.C., Zhang, Y.P., 2012. Effect of familiar masking background on target speech recognition. Acta Acustica United with Acustica 98, 328–333.

Zhu, W., O'Shaughnessy, D., 2005. Log-energy dynamic range normalization for robust speech recognition. In: Proc. ICASSP, pp. 245–248.