PAPER

Noise Robust Feature Scheme for Automatic Speech Recognition Based on Auditory Perceptual Mechanisms

Shang CAI^{†a)}, Yeming XIAO[†], Jieli PAN^{†b)}, *Nonmembers*, Qingwei ZHAO[†], *Member*,
and Yonghong YAN^{†c)}, *Nonmember*

SUMMARY Mel Frequency Cepstral Coefficients (MFCC) are the most popular acoustic features used in automatic speech recognition (ASR), mainly because the coefficients capture the most useful information of the speech and fit well with the assumptions used in hidden Markov models. As is well known, MFCCs already employ several principles which have known counterparts in the peripheral properties of human hearing: decoupling across frequency, mel-warpage of the frequency axis, log-compression of energy, etc. It is natural to introduce more mechanisms in the auditory periphery to improve the noise robustness of MFCC. In this paper, a k -nearest neighbors based frequency masking filter is proposed to reduce the audibility of spectra valleys which are sensitive to noise. Besides, Moore and Glasberg's critical band equivalent rectangular bandwidth (ERB) expression is utilized to determine the filter bandwidth. Furthermore, a new bandpass infinite impulse response (IIR) filter is proposed to imitate the temporal masking phenomenon of the human auditory system. These three auditory perceptual mechanisms are combined with the standard MFCC algorithm in order to investigate their effects on ASR performance, and a revised MFCC extraction scheme is presented. Recognition performances with the standard MFCC, RASTA perceptual linear prediction (RASTA-PLP) and the proposed feature extraction scheme are evaluated on a medium-vocabulary isolated-word recognition task and a more complex large vocabulary continuous speech recognition (LVCSR) task. Experimental results show that consistent robustness against background noise is achieved on these two tasks, and the proposed method outperforms both the standard MFCC and RASTA-PLP.

key words: automatic speech recognition, noise robustness, critical bandwidth, frequency masking, temporal masking

1. Introduction

The development of automatic speech recognition (ASR) technology in recent years has made it practical for commercial applications. The state-of-the-art ASR systems have obtained satisfactory performance in controlled laboratory environments. Unfortunately, the recognition accuracy often degrades dramatically if there is a mismatch between the trained speech models and the real-world speech to be recognized, due to factors like background noise, channel distortion, etc. Robustness against noisy conditions is an essential challenge for practical ASR systems since they often need to carry out recognition in a variety of everyday acoustic environments in many applications. The environmental

acoustic conditions are highly variable and unpredictable, so they are difficult to account for during the ASR system training.

Several techniques [1]–[3] have been developed to address the aforementioned problem. These methods mainly focused on minimization of the acoustic mismatch caused by noise and can be constructed either in the model domain or in the feature domain. Some of them, such as model adaptation [4] and parallel model combination [5], belong to the above category of model domain technologies, which are trying to adjust the acoustic models to fit the noise conditions better using noisy speech. Some other researchers try to normalize the features for ASR. Cepstral mean subtraction (CMS) is a simple but effective way to remove the dc component of features. Several other well-known normalization methods for feature domain have been proposed, such as cumulative histogram used in histogram equalization (HEQ) [6] and cepstral shape normalization (CSN) [7]. On the other hand, it is observed that human auditory system is remarkably more robust than the state-of-the-art ASR systems in the presence of variable noises. This fact has motivated researchers to find more useful processing strategies in the human auditory system in noise suppression and to apply them to ASR. The most widely used feature extraction scheme MFCC already utilized several principles which have known counterparts in the peripheral properties of human hearing: decoupling across frequency, mel-warpage of the frequency axis and log-compression of energy. Therefore, it seems natural to consider more properties from the psychoacoustics to design acoustic features. Researchers have previously tried to represent the speech signal using new robust features motivated by the auditory system in order to reduce the effect of noise. Ensemble interval histograms (EIH) are probably the most well-known auditory-based features [8]. In [9], a novel feature set called sub-band spectral centroid histograms (SSCH) integrates dominant-frequency information with sub-band power information. Another type of feature widely used in current ASR systems is perceptual linear prediction (PLP) [10]. It is developed based on psycho-physical findings and simulates several well-known properties of hearing, such as equal-loudness pre-emphasis and intensity-loudness power law. In [11], RASTA processing [12], which is a bandpass modulation filtering, is introduced into the PLP analysis. And a new speech analysis technique RASTA-PLP is presented.

In addition to the methods mentioned above, an alter-

Manuscript received October 26, 2011.

Manuscript revised January 15, 2012.

[†]The authors are with Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, 21 Beisihuan XiLu, Beijing, R.P. China.

a) E-mail: caishang@hcl.ioa.ac.cn

b) E-mail: panjielin@hcl.ioa.ac.cn

c) E-mail: yanyonghong@hcl.ioa.ac.cn

DOI: 10.1587/transinf.E95.D.1610

native approach to designing new features is to modify the standard feature extraction routine. To improve noise robustness, some mechanisms based on auditory evidence—frequency masking, critical bandwidth from human psychoacoustic and temporal masking—are investigated and a feature extraction front-end modified from standard MFCC schedule is proposed in this paper. The masking effect suppresses unimportant signals, and may be helpful for ASR. We propose a k-nearest neighbors based frequency masking filter, and describe a new third-order infinite impulse response (IIR) filter to model the temporal masking effect. Furthermore, the auditory bandwidth of filter-bank is employed. This schema does not assume any additional parameters that require learning from data and we believe it is arguably more robust than data-driven techniques since environmental acoustic condition is unpredictable. Moreover, the computational cost is small and it can be effectively combined with some other methods to further increase the noise robustness.

The organization of this paper is as follows. Section 2 introduces the framework of the proposed feature extraction scheme and reviews the used auditory perceptual mechanisms. Then we specify the key elements of the processing in some detail in Sect. 3. In Sect. 4, the experiments and results are presented. Finally, we summarize our results and give conclusions in Sect. 5.

2. The Structure of the Proposed Feature Scheme

The purpose of the feature extraction scheme is to represent the acoustic speech signal with a succinct form, which retains the information needed for speech recognition as much as possible and abandons content irrelevant information. In order to improve the ASR performance, we employ some auditory perceptual characteristics here. There are two reasons for us to follow this paradigm. Firstly, human listeners outperform today's best ASR system [13], especially in noise conditions. Techniques stemming from auditory properties may bring some recognition accuracy gain. Secondly, as a communication medium, human speech has evolved to fit the characters of human auditory system. When researchers found something beneficial for ASR, it could be eventually explained by some previously ignored properties of human auditory perception [14]. In order to achieve better performance, how to use obtained auditory knowledge felicitously is the key issue.

In this work, some auditory perceptual mechanisms are investigated and our purpose is to find whether they are helpful for robust speech recognition. These mechanisms are introduced into the standard MFCC scheme and evaluated. The MFCC algorithm can be summarized as follows: a speech signal is pre-emphasized, windowed, Fourier transformed to the frequency domain, scaled by a bank of triangular filters, and the output of each filter is log-compressed and transformed via the discrete cosine transform (DCT) to cepstral coefficients. Its structure is shown in the left-hand side of Fig. 1. And the right-hand side is the proposed fea-

Fig. 1 Comparison between the standard MFCC feature extraction and the proposed MFCC discussed in this paper.

ture extraction scheme discussed in this paper.

As shown in Fig. 1, the proposed feature extraction algorithm is similar to the traditional MFCC. Three shadowed blocks show the improvements.

- Firstly, a filter mimicking the phenomenon of frequency masking is introduced. It suppresses the weak signal components when there is a spectrum peak occurring at the same time.
- Secondly, the bandwidth of the triangle filter bank is designed according to critical bandwidth determined by equivalent rectangular bandwidth (ERB).
- Finally, we propose an IIR bandpass filter to model the temporal masking phenomenon, considering both effects of synaptic adaptation and temporal integration.

The three mechanisms above are presented in detail in Sect. 3. The corresponding modules can be selectively incorporated into the standard MFCC feature extraction.

3. Three Auditory Perceptual Mechanisms Used for Robust ASR

3.1 Frequency Masking

One of the useful perceptual mechanisms which has been seldom utilized is the “masking” effect of human auditory system. In the masking phenomenon, one acoustic stimulus cannot be perceived if another acoustic stimulus with sufficiently high intensity appears closely. The low intensity level signals are usually unimportant for speech recognition

and may be noises to be suppressed. So the nature of this phenomenon may improve the performance of ASR. There are two kinds of masking, frequency masking and temporal masking. In this section we focus on frequency masking, and temporal masking will be discussed in Sect. 3.3.

Frequency masking means the masked signal is suppressed by the masker with a close frequency which presents simultaneously[†]. Lateral inhibition may be one reason for frequency masking. It is a common phenomenon involved in sensory reception of biological systems and has been discovered through physiological experiments [15]. This phenomenon is caused by interconnected neurons and makes them more accurate to stimulus. The essence of frequency masking is to enhance the dominant signal components, such as spectral peaks, and to suppress the adjacent components. Furthermore, the spectra between spectral peaks are sensitive to noise. Noise contributes a lot of variabilities to spectra area with low energy and spectral valleys may be buried by noise, while the spectral peaks with higher energy are less affected. When such noise presents, it could be masked or its audibility is reduced in the human perception system. We can utilize the same mechanism to speech spectra in order to generate robust features for speech recognition.

There are already some reports in the literature using frequency masking model for robust speech recognition [16], [17]. In [17], frequency masking threshold is modeled as a triangle. And an improved version of this method is reported in [18]. It introduces a bi-directional non-linear filter to the standard MFCC. The noise-caused variability is diminished and those spectral valleys are masked by frequency masking. Here we improve this scheme and adopt it into our robust feature extraction routine. A non-linear filter is applied to the original power spectrum so that the valleys in the spectra are masked by the neighboring spectral peaks. Unlike [18], a k -nearest neighbors based filter is proposed to mimic this phenomenon. It is more faithful to the mechanism of frequency masking, since the spectral valleys are also masked by the bins which have sufficiently high intensity, even though they are not the nearest neighbor. We find this method works better in our experiments.

The filter is described as follows:

$$y_i = \max(x_i, t_{L_i}, t_{H_i}) \quad (1)$$

where y_i is the output of the frequency masking filter and x_i is the original power spectrum at frequency index i . t_{L_i} is the auditory perception threshold generated from lower frequency bins and similarly, t_{H_i} is generated from higher frequency bins. The two thresholds are calculated as:

$$t_{L_i} = \max(\alpha_{i-m}^m x_{i-m}); \quad \max(1, i-k) \leq m \leq i \quad (2)$$

$$t_{H_i} = \max(\beta_{m-i}^m x_{i+1}); \quad i \leq m \leq \min(i+k, N) \quad (3)$$

where N is the total number of FFT-bins. α_i and β_i are two sets of parameters controlling the masking thresholds. They are determined by a simple liner interpolation along with the

Fig. 2 Schematic representation of frequency masking. In this example, x_i will be masked by t_{L_i} .

Fig. 3 The effect of the frequency masking filter on a randomly selected frame of speech. The original spectra is compared with its corresponding output of the proposed filter.

frequency axis. α_i is calculated as an interpolation of 0.3 and 0.5, while β_i is interpolated by 0.6 and 0.8. k controls the number of neighbors used to calculate the corresponding threshold. Since farther distance leads to smaller influence, we chose the value of $k = 3$.

Figure 2 is the schematic representation of this mechanism. And Fig. 3 shows an example of frequency masking. In Fig. 3, the solid line curves the output of the filter and the dotted line is the original spectra. We can find that deep valleys in the speech spectra are masked.

3.2 Auditory Bandwidth of Filter-Bank

Human auditory processing relies on a bank of overlapping band-pass filters. In the standard MFCC extraction procedure, a bank of triangular filters are used, and the centers of the filters are determined equally according to the mel-warped frequency axis, which is inspired by the human auditory system. Meanwhile, the critical bandwidth of human hearing is related to the auditory filter at a given center frequency. However, the bandwidth of each filter in MFCC is arbitrarily set by fixing the base of each triangular filter by the center frequencies of the neighboring filters.

The notion of ERB can be used to quantify the band-

[†]Frequency masking may well be named spectral masking or simultaneous masking in the literature.

width of each filter. Moore and Glasberg suggested an approximation of critical bandwidth measured in ERB [19].

$$ERB_i = 6.23f_{c_i}^2 + 93.39f_{c_i} + 28.52\text{Hz} \quad (4)$$

where f_{c_i} is the center of the i -th filter. It has been noted previously that the classical MFCC may be outperformed in speech recognition by making the bandwidth of the underlying mel-filter-bank a free design parameter. In [20], [21], Moore and Glasberg's critical band ERB is utilized to determine the bandwidth of MFCC's filters. While some researchers find Gammatone filters or Gabor filters may be a good choice, we focus only on the bandwidth of filters in this study and use triangular filters as the standard MFCC schedule. The filter placing is equidistant in mel scale and the mel frequency $\hat{f}$ is related to linear frequency f by

$$\hat{f} = 1127 \ln(1 + \frac{f}{700}) \quad (5)$$

Let f_{l_i} and f_{h_i} be the low and high frequencies of the i -th filter respectively. Center frequencies are equally spaced in mel frequency

$$\hat{f}_{c_i} = (\hat{f}_{l_i} + \hat{f}_{h_i})/2 \quad (6)$$

Denote $|H(f)|$ as the amplitude of a filter transfer function. For triangular pass-band filters,

$$ERB = \frac{\int |H(f)|^2 df}{|H(f_c)|^2} = \frac{f_h - f_l}{2} \quad (7)$$

Given the frequency range of the entire filter bank, we can determine the center frequencies f_{c_i} using Eqs. (4), (5), (6) and (7). Combination of Eqs. (5) and Eqs. (6) yields

$$(700 + f_{c_i})^2 = (700 + f_{l_i})(700 + f_{h_i}) \quad (8)$$

Then lower and upper frequencies for each filter can be calculated by:

$$f_{l_i} = \sqrt{(700 + ERB_i)^2 + f_{c_i}(f_{c_i} + 1400) - 1400ERB_i} - (700 + ERB_i) \quad (9)$$

$$f_{h_i} = f_{l_i} + 2ERB_i \quad (10)$$

When the center, lower and upper frequencies are determined, the filter-bank can be constructed by connecting straight lines. The triangle has zero height at each end and unity height at the center. Although some researchers obtained even better performance by scaling the bandwidth [20], we find that it may be task-related in our study.

3.3 Temporal Masking

As described in Sect. 3.1, there is another kind of masking called temporal masking. Similar to frequency masking, temporal masking is a phenomenon that one sound cannot be perceived if another sound with sufficiently high intensity appears closely. Unlike frequency masking where

masked sound and masker are present simultaneously, temporal masking occurs when the masker just precedes or just follows the masked sound in time. The actual cause of temporal masking is not clear. Synaptic adaptation and temporal integration are the possible mechanisms of the temporal masking. These two mechanisms are described below, and in consideration of both of them, a new bandpass IIR filter is utilized to imitate the temporal masking phenomenon.

3.3.1 Synaptic Adaptation

Synaptic adaptation is a dynamic action in the human peripheral auditory system. In an inner hair cell there is a "readily releasable pool" (RRP) of synaptic vesicles near each synapse. When an acoustic stimulus occurs, the neurotransmitters in the RRP will be released into the synaptic cleft. Then the pool will be refilled. As the rate of refill is lower than the initial vesicle fusion rate, the vesicles may deplete with time. So at the beginning of an acoustic stimulus, plenty of vesicles are available to fuse and the auditory nerve has a strong initial response. After that the nerve activity is rapidly depressed. This mechanism can enhance sound detection and improve the immunity to stationary noise. In [22], synaptic adaptation is implemented using a first-order IIR filter and it has been shown outperforming the RASTA processing. The transfer function of this filter is

$$H_s(z) = \frac{2f_s\tau - 2f_s\tau z^{-1}}{1 + 2f_s\tau + (1 - 2f_s\tau)z^{-1}} \quad (11)$$

where f_s is the frame rate and τ is the adaptation time constant, which is chosen to be 240 ms. It is a high-pass filter and its magnitude response is shown in Fig. 4.

3.3.2 Temporal Integration

The hearing threshold of human auditory system is inversely proportional to the durations of sounds less than 200 ms [23]. When a sound is transmitted into the ear, the

Fig. 4 The frequency response of the synaptic adaptation filter.

Fig. 5 The frequency response of the temporal integration filter.

temporal integrator needs some time to accumulate intensity. While the sound stops, the accumulated intensity needs some time to decay. In [16], temporal integration is modeled as

$$y[n] = x[n] + A \sum_{k=1}^{\infty} \alpha^{-k} x[n-k] - B \sum_{k=1}^{\infty} \beta^{-k} x[n-k] \quad (12)$$

where A and B are scaling constants, and α, β are time constants of integration. The second term on the right side of Eqs. (12) represents accumulation and the third term denotes the masking integration. Its transition function is

$$H_t(z) = \frac{1 - [(1-A)\alpha + (1+B)\beta]z^{-1} + (1-A+B)\alpha\beta z^{-2}}{1 - (\alpha + \beta)z^{-1} + \alpha\beta z^{-2}} \quad (13)$$

Figure 5 shows its magnitude response. One can see that it is a bandpass filter which de-emphasizes the signal components varying too fast or too slow.

3.3.3 Implementation of Temporal Masking

In [24], both synaptic adaptation and temporal integration achieve better performance than the standard MFCC, and combining both of them yields further performance gain. This combination is a straight-forward implementation. The signal is filtered by Eqs. (11) and Eqs. (13) respectively, then both outputs of the two filters are directly added together. If these two mechanisms are the actual causes of temporal masking, it is not reasonable that the input signal passes through these two filters just linearly independently. The effect of Eqs. (11) and Eqs. (13) could be simulated with a third-order filter. Considering both synaptic adaptation and temporal integration, we propose a new filter which may model temporal masking appropriately,

$$H(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}}{b_0 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3}} \quad (14)$$

The parameters, a_k and b_k , are optimized jointly by experiments. They are tuned on an isolated-word task in clean and

Fig. 6 The frequency response of the proposed temporal masking filter.

supermarket noisy conditions. We find they also give better performance in other tests. In this study, we use $a_0 = 5.0$, $a_1 = -12.3$, $a_2 = 10.0$, $a_3 = -2.7$, $b_1 = 4.8$, $b_1 = -11.5$, $b_2 = 8.7$ and $b_3 = -2.0$ for all tasks and all noisy conditions in this paper. Figure 6 is the magnitude response of Eqs. (14). It has a narrower pass-band compared to the temporal integration filter.

Our model of temporal masking is implemented in the logarithmic mel spectra domain. A temporally filtered version of the spectra is summed with the original spectra. In this work $x[n]$ denotes the logarithmic mel spectra at frame n . Note that $x[n]$ is a function of the k -th component of the mel filter-bank, but an additional index for k is not introduced in order to keep the equations more readable. The following routine for temporal masking is applied.

1. Subtract the first frame of logarithmic mel spectra from all of the frames to avoid an initial transient

$$x'[n] = x[n] - x[1], n \in [1, N] \quad (15)$$

where N is the total frame number.

2. Apply the proposed filter Eqs. (14) to $x'[n]$. It is implemented as

$$\sum_{i=0}^3 b_i y[n-i] = \sum_{j=0}^3 a_j x'[n-j] \quad (16)$$

3. Modify the original logarithmic mel spectra by summing with the output of the filter

$$x_{tm}[n] = x[n] + y[n] \quad (17)$$

where $x_{tm}[n]$ is the temporally masked logarithmic mel spectra.

4. Experimental Evaluation

In this section, we present experimental results to demonstrate that the introduced auditory perceptual mechanisms are helpful for robust ASR. We begin in Sect. 4.1 with an overview of the system setup. We provide some results

on a medium-vocabulary isolated-word recognition task in Sect. 4.2. The frequency masking filter and the temporal masking filter presented in Sect. 3 are evaluated. Then we assess the contributions of various mechanisms. In Sect. 4.3, a more complex large vocabulary continuous speech recognition (LVCSR) task is conducted for evaluation.

The term of relative improvements (RI) is defined to evaluate the benefit that the new methods bring in comparison to the baseline. It is calculated as

$$RI = \frac{NewAccuracy - Baseline}{100 - Baseline} \times 100\% \quad (18)$$

where *NewAccuracy*, *Baseline* are recognition accuracies for proposed methods and reference baseline methods respectively.

4.1 System Setup

Both the isolated-word recognition task and the LVCSR task use the same acoustic models. The recognition results presented in Sect. 4.2 and Sect. 4.3 are based on the experimental configuration described here.

4.1.1 Training Data and Noise Source

The training data used in our experiments is a Mandarin reading speech corpus about 90 hours, which is published by the Chinese governmental research program 863, with a total of about 86 k sentences [25]. This training set is gender balanced. All training files are transformed into 8 k-Hz 16-bit PCM form.

In order to evaluate the superiority of these auditory perceptual mechanisms for robust ASR, we added five different noise types to corresponding test set at the required signal noise ratio (SNR), namely CarRoad, Exhibition, Music, Park and Supermarket. The background music was obtained from a musical database. Randomly selected musical segments constituted the noise file. All the other noises were recorded by us in the corresponding environments.

4.1.2 Acoustic Model Training

The plain, standard MFCC produced by Cambridge HMM Toolkit (HTK) [26] is used as the baseline in this work. A reduced bandwidth analysis, 60 Hz–3400 Hz, is employed to generate features. Every utterance is analyzed with 25 ms frames at 100 frames/s, and a pre-emphasis filter of the form $x'_n = x_n - 0.97x_{n-1}$ is applied. Each frame is Hamming windowed. 24 sub-band filters are used. Finally we calculate 13 cepstral coefficients (MFCCs, C1-C12 plus the log-energy). For each feature used in this work, the delta and the acceleration derivatives are appended to the corresponding static feature vector to form a 39-dimensional feature vector. The window length used to calculate both delta and acceleration derivatives is 5 frames. As a matter of fact, the delta calculation can be regarded as a filtering process for modulation frequency which has a selective frequency response [12]. While the temporal masking filter proposed in

this work has a broader pass-band as shown in Fig. 6.

PLP is another type of feature widely used in today's ASR systems. It has been derived independently of the MFCC technique and employs some auditory perceptual mechanisms. The well-known RASTA methodology can make PLP more robust. In this work the effort of RASTA-PLP is evaluated for comparison. We use a publicly available implementation of RASTA-PLP [27].

Constructing an ASR system with excellent performance for Mandarin is a challenging task. In China, the accent problem is very crucial. Almost every province has its own dialect. When speaking Mandarin, a person's dialect usually brings a heavy accent to his/her speech [28]. Another problem is the Chinese homophones ambiguity. Many different Chinese characters have the same pronunciation. As Mandarin is a tonal language, the phone set for HMM modeling consists of 29 initials and 150 finals with tone markers for our work. The final HMMs are gender independent crossword triphone models with 3-state left-to-right topology. A robust state clustering with two-level phonetic decision trees are used [29]. About 6 k shared states are empirically determined with 16-component Gaussian mixture output densities per state.

4.2 Recognition Results on an Isolated-Word Task

A medium-vocabulary isolated-word recognition task is carried out in this section. The isolated-word recognition is a relatively simple task without the influence from the language model. So it is beneficial to compare different kinds of features. The test corpus BJX we used here is self-collected, which consists of 4004 queries. The content of each query is a people's full name spoken in Mandarin. The grammar which we adopt for decoding consists of a name list of 501 entries. And all testing queries are in-grammar test case.

Firstly, we evaluate the k-nearest neighbors based frequency masking filter. The recognition accuracies in different acoustic conditions are shown in Table 1. Frequency masking brings considerable performance improvements, and the average relative performance gain upon the standard MFCC is 11.11%. We can also find that the proposed filter works slightly better than the method presented in [18], and the average relative improvement is 1.33%. Although

Table 1 Recognition accuracies obtained using frequency masking filter. In this table, ref FM corresponds to using the method presented in [18], new FM corresponds to using the filter introduced in Sect. 3.1, RI1 and RI2 are the relative improvements compared with MFCC and ref FM, respectively. The noise types are added to the test set with an SNR of 10 db.

noise	MFCC	ref FM	new FM	RI1	RI2
Clean	85.14	86.59	87.19	13.78	4.46
CarRoad	68.13	69.48	67.98	-0.47	-4.91
Music	56.69	57.36	57.49	1.85	0.29
Park	49.38	62.81	64.49	29.85	4.50
Exhibition	59.37	60.46	60.34	2.40	0.32
Supermarket	46.00	51.32	53.37	13.64	4.21
Average	60.79	64.67	65.14	11.11	1.33

Table 3 Recognition accuracies on the isolated-word task. In this table, the baseline corresponds to the standard MFCC, FM corresponds to the frequency masking filter, ABW corresponds to using auditory bandwidth filter-bank, and TM corresponds to the temporal masking filter. The above three auditory perceptual mechanisms are added incrementally.

noise	feature	clean	20 db	15 db	10 db	5 db	0 db	average	RI
CarRoad	baseline	85.14	82.59	78.32	68.13	37.46	9.11	60.13	~
	+FM	87.19	81.89	76.62	67.98	40.48	11.49	60.94	2.05
	+FM+ABW	86.79	84.72	80.82	73.15	44.06	12.41	63.66	8.85
	+FM+ABW+TM	87.21	85.24	81.84	77.60	55.72	18.68	67.72	19.03
	RASTA-PLP	81.12	80.62	78.05	76.52	60.02	29.25	67.59	18.73
Music	baseline	85.14	79.40	70.33	56.69	30.39	8.62	55.09	~
	+FM	87.19	80.64	71.25	57.49	30.24	9.39	56.06	2.09
	+FM+ABW	86.79	81.17	73.10	59.62	32.64	8.97	57.05	4.35
	+FM+ABW+TM	87.21	81.29	73.43	63.76	37.11	12.64	59.24	9.23
	RASTA-PLP	81.12	75.47	67.81	56.72	34.34	14.01	54.91	-0.41
Park	baseline	85.14	74.18	63.99	49.38	24.45	5.37	50.42	~
	+FM	87.19	80.69	73.60	64.49	39.79	12.06	59.64	18.59
	+FM+ABW	86.79	84.09	80.79	76.80	56.07	20.58	67.52	34.49
	+FM+ABW+TM	87.21	85.64	81.94	79.05	58.07	23.35	69.21	37.90
	RASTA-PLP	81.12	80.29	77.47	73.35	57.72	27.72	66.28	31.99
Exhibition	baseline	85.14	82.22	75.75	59.37	24.45	3.60	55.09	~
	+FM	87.19	81.27	73.40	60.34	28.02	4.52	55.79	1.57
	+FM+ABW	86.79	83.17	77.87	64.74	31.57	5.29	58.24	7.02
	+FM+ABW+TM	87.21	84.82	80.50	70.63	37.81	6.87	61.31	13.85
	RASTA-PLP	81.12	79.92	74.73	66.08	37.19	9.64	58.11	6.74
Supermarket	baseline	85.14	76.87	66.21	46.00	13.04	1.20	48.08	~
	+FM	87.19	79.07	69.03	53.37	20.98	2.50	52.02	7.60
	+FM+ABW	86.79	80.54	73.20	62.26	28.35	3.27	55.74	14.75
	+FM+ABW+TM	87.21	83.64	77.02	66.66	30.09	3.70	58.05	19.22
	RASTA-PLP	81.12	75.95	66.53	50.80	21.70	3.65	49.96	3.62
Average	baseline	85.14	79.05	70.92	55.91	25.96	5.58	53.76	~
	+FM	87.19	80.71	72.78	60.73	31.90	7.99	56.89	6.76
	+FM+ABW	86.79	82.74	77.16	67.31	38.54	10.10	60.44	14.44
	+FM+ABW+TM	87.21	84.13	78.95	71.54	43.76	13.05	63.11	20.21
	RASTA-PLP	81.12	78.45	72.92	64.70	42.19	16.85	59.37	12.13

Table 2 Recognition accuracies obtained using temporal masking scheme. In this table, ref TM corresponds to using the method presented in [24], new TM corresponds to using the filter designed in Sect. 3.3.3, RI1 and RI2 are the relative improvements compared with MFCC and ref TM, respectively. The noise types are added to the test set with an SNR of 10 db.

noise	MFCC	ref TM	new TM	RI1	RI2
Clean	85.14	87.09	87.46	15.63	2.87
CarRoad	68.13	73.38	74.60	20.30	4.58
Music	56.69	64.84	65.28	19.84	1.25
Park	49.38	51.90	52.80	6.76	1.87
Exhibition	59.37	67.51	68.63	22.80	3.45
Supermarket	46.00	51.00	52.87	12.72	3.82
Average	60.79	65.95	66.94	15.70	2.91

the gain from taking k-nearest bins is somewhat small, it is more faithful to the mechanism of frequency masking.

Then the temporal masking filter proposed in this work is evaluated. It is similar to the well-known RASTA filter. But Eqs. (11), which is designed to imitate synaptic adaptation, is demonstrated to work better than RASTA [22]. In addition, both synaptic adaptation and temporal integration achieve good performance in [24], and combining both of them achieves further performance gain. The method proposed in [24] is filtering the signal by Eqs. (11) and Eqs. (13) respectively, then adding together these two outputs. Table 2 summarizes the experimental results in dif-

ferent acoustic conditions. The temporal masking filter described in Sect. 3.3.3 provides better recognition results than the standard MFCC, and the average relative improvement is 15.70%. Furthermore, the proposed filter works consistently better than the method recommended in [24], and 2.91% average relative improvement is achieved.

Finally we evaluate the proposed feature scheme and assess the contributions of various mechanisms introduced in Sect. 3. The three auditory perceptual mechanisms—frequency masking, new auditory filter bandwidth and temporal masking—are added to the standard MFCC scheme incrementally. Experiments are conducted in more different noise conditions. The accuracies of the recognition results are shown in Table 3. From this table we can see that the frequency masking is helpful in improving accuracy, particularly for speech distorted by relatively stationary Park noise. The auditory filter bandwidth provides a further improvement, and the temporal masking brings a substantial recognition accuracy gain in various types of noise. The average relative improvement upon the standard MFCC is 20.21%. Another phenomenon is that relatively smaller improvements are obtained in the background music noise, since it is a more difficult transitory environment with clear interfering human speech.

Moreover, the performance of RASTA-PLP is also

Table 4 Recognition accuracies (100%-WER) on the LVCSR task. In this table, pMFCC corresponds to the revised MFCC with the three auditory perceptual mechanisms described in Sect. 3, RI1 and RI2 are the relative improvements compared with the standard MFCC and RASTA-PLP, respectively. The noise types are added to the test set with an SNR of 10 db.

noise	MFCC	RASTA-PLP	pMFCC	RI1	RI2
Clean	80.5	75.5	81.4	4.62	24.08
CarRoad	41.9	52.3	51.4	16.35	-1.89
Music	32.0	25.6	33.9	2.79	11.12
Park	33.5	44.6	52.1	27.97	13.54
Exhibition	34.6	37.7	41.7	10.86	6.42
Supermarket	30.2	25.3	36.2	8.60	14.59
Average	42.12	43.5	49.45	12.67	10.53

shown in Table 3. It outperforms the standard MFCC and the average relative improvement is 12.13%. The revised MFCC works better than RASTA-PLP. Its relative improvement upon RASTA-PLP is $(63.11 - 59.37)/(100 - 59.37) = 9.21\%$. Looking at the clean and high SNR conditions individually, we see that the revised MFCC improve the recognition accuracies, while RASTA-PLP worsens them.

4.3 Recognition Results on an LVCSR Task

To confirm the generalization of the findings, we evaluate the revised MFCC on another more complex LVCSR task. The test corpus is published by Chinese governmental research program 863. It is about 11 hours, including 9692 utterances from fourteen speakers (7 females and 7 males). A tri-gram language model built with SRILM tools [30] is used in decoding. The vocabulary contains about 43 k words.

Detailed recognition accuracies are shown in Table 4. The difference between the revised MFCC and the standard MFCC is the same as the previous isolated-word recognition task, and 12.67% average relative improvement is achieved, although the gap is diminished somewhat. This is commonly observed according to our experience as we move to a more complex task. Once again, we can find that the revised MFCC outperforms RASTA-PLP, and the average relative improvement is 10.53%.

5. Conclusions

This paper presents our recent work in employing some auditory perceptual mechanisms to improve the robustness of feature extraction for ASR. We demonstrate the positive benefits of these mechanisms. In this work, a more faithful k-nearest neighbors based frequency masking filter is introduced and a third-order IIR filter is proposed for temporal masking. As shown in Table 1 and in Table 2, both of them bring consistent performance gain over the baseline, and outperform their reference methods respectively. Moreover, three mechanisms—the frequency masking filter, auditory bandwidth of filter-bank, and the temporal masking filter — are incorporated into the standard MFCC routine, and a new feature extraction scheme is proposed. Experiments are carried out on a medium-vocabulary isolated-word recognition task and on a more complex LVCSR task

to verify the generalization. The contributions of these three auditory perceptual mechanisms are assessed. The experimental results show that the new feature is considerably more robust and performance gains are achieved in all acoustic conditions. Namely, average relative improvements of 20.21% and 12.67% are obtained respectively on the isolated-word recognition task and the LVCSR task compared with MFCC; while 9.21% and 10.53% average relative improvements are obtained compared with RASTA-PLP. Thus, the proposed feature scheme is effective for robust ASR. Our future work will focus on finding more helpful auditory perceptual mechanisms, using these mechanisms in a more elegant way, and combining them with other robust techniques, such as model adaptation.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

References

- [1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol.16, pp.261–291, 1995.
- [2] X. Huang and H.W. Hon, *Spoken language processing: A guide to theory, algorithm and system development*, Prentice Hall PTR, 2001.
- [3] P. Moreno, *Speech recognition in noisy environments*, Ph.D. Thesis, Carnegie Mellon University, 1996.
- [4] M.J.F. Gales, "The generation and use of regression class trees for MLLR adaptation," Univ. Cambridge, Tech. Rep. CUED/FINFENG/TR263, 1996.
- [5] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, pp.845–848, 1990.
- [6] A. De La Torre, J.C. Segura, C. Benitez, A.M. Peinado, and A.J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," *Proc. ICASSP*, pp.401–404, 2002.
- [7] J. Du and R.H. Wang, "Cepstral shape normalization (CSN) for robust speech recognition," *Proc. ICASSP*, pp.4389–4392, 2008.
- [8] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. Phonetics*, vol.16, pp.109–123, 1988.
- [9] B. Gajic and K.K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Trans. Audio Speech Language Process.*, vol.14, no.2, pp.600–608, 2006.
- [10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol.87, pp.1738–1752, 1990.
- [11] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," *Proc. ICASSP*, vol.1, pp.121–124, 1992.
- [12] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process.*, vol.22, no.1, pp.578–589, 1994.
- [13] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol.22, no.1, pp.1–15, 1997.
- [14] H. Hermansky, "Should recognizers have ears?," *Speech Commun.*, vol.25, no.1-3, pp.3–27, 1998.
- [15] S.A. Shamma, "Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *J. Acoust. Soc. Am.*, vol.78, no.5, pp.1622–1632, Nov. 1985.
- [16] Park, K.Y. and S.Y. Lee, "An engineering model of the masking for the noise-robust speech recognition," *Neurocomputing*, vol.52,

- pp.615–620, 2003.
- [17] T. Usagawa, M. Iwata, and M. Ebata, “Speech parameter extraction in noisy environment using a masking model,” *Proc. ICASSP*, vol.2, pp.81–84, Adelaide, 1994.
 - [18] W. Zhu and D. O’Shaughnessy, “Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm,” *Proc. 7th International Conf. on Signal Processing*, vol.1, pp.617–620, 2004.
 - [19] B.C.J. Moore and B.R. Glasberg, “Suggested formulae for calculating auditory filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, vol.74, pp.750–753, 1983.
 - [20] M.D. Skowronski and J.G. Harris, “Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition,” *J. Acoust. Soc. Am.*, vol.116, pp.1774–1780, 2004.
 - [21] D. von Zeddelmann, F. Kurth, and M. Muller, “Perceptual audio features for unsupervised key-phrase detection,” *Proc. ICASSP*, pp.257–260, 2010.
 - [22] M. Holmberg, D. Gelbart, and W. Hemmert, “Automatic speech recognition with an adaptation model motivated by auditory processing,” *IEEE Trans. Audio Speech Language Process.*, vol.14, no.1, pp.43–49, 2006.
 - [23] B. Moore, *An introduction to the psychology of hearing*, Academic press, 2003.
 - [24] S.C. Chiou and C.P. Chen, “Noise-robust feature extraction based on forward masking,” *Proc. Interspeech 2009*, pp.2159–2162, 2009.
 - [25] Q. Zhang, J. Pan, Y. Lin, J. Shao, and Y. Yan, “Development of a Mandarin-English bilingual speech recognition system for real world music retrieval,” *IEICE Trans. Inf. & Syst.*, vol.E91-D, no.3, pp.514–521, March 2008.
 - [26] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, “The HTK book version 3.4,” 2006.
 - [27] SPRACHcore, <ftp://ftp.icsi.berkeley.edu/pub/real/dpwe/>, 2004.
 - [28] C. Huang, T. Chen, and E. Chang, “Accent issues in large vocabulary continuous speech recognition,” *Int. J. Speech Technology*, vol.7, no.2, pp.141–153, 2004.
 - [29] C. Liu and Y. Yan, “Robust state clustering using phonetic decision trees,” *Speech Commun.*, vol.42, no.3, pp.391–408, 2004.
 - [30] SRILM, <http://www.speech.sri.com/projects/srilm/>, 2007.

Jieli Pan received his B.S. from Peking University in 1986 and his Master degree in Electronic Engineering from Tsinghua University in 1989. From Jan. 2000 to July 2001 he was working in Intel China Research Center as a senior researcher and speech recognition group manager. In Dec. 2002 he joined Key Laboratory of Speech Acoustics and Content Understanding as a professor. His current research includes: LVCSR, speech analysis, acoustic model and search algorithm.

Qingwei Zhao received his Ph.D. in Electronic Engineering from Tsinghua University in 1999. Currently he is an Associate Professor at Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS. Before joining CAS, he was with Intel as a senior researcher. His research interests include voice search, spontaneous speech recognition and spoken term detection.

Yonghong Yan received his B.E. from Tsinghua University in 1990 and his Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998–2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Now he is a professor and

director of Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS. His research interests include speech processing and recognition, language/speaker recognition and human computer interface. He has published more than 100 papers and holds 40 patents.

Shang Cai received his B.S. degree from University of Science and Technology of China (USTC) in 2007. Now he is a M.S. & Ph.D. candidate in Key Laboratory of Speech Acoustics and Content Understanding at Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research interests include speech signal processing, noise robust speech recognition and acoustic model.

Yeming Xiao received his B.E. degree from the Beihang University, Beijing, China in 2008. Now he is a M.S. & Ph.D. candidate in Key Laboratory of Speech Acoustics and Content Understanding at IOA, CAS. His research interests include automatic speech recognition, deep learning and neural network.