Research on Robustness of Emotion Recognition under Environmental Noise Conditions

Yongming Huang¹, Jing Xiao¹, Kexin Tian¹, Ao Wu¹, Guobao Zhang¹
¹Laboratory of Measurement and Control of Complex Systems of Engineering; Ministry of Education, School of Automation, Southeast University

Corresponding author: Yongming Huang (huang_ym@163.com).

This work was supported by Natural Science Foundation of Jiangsu Province (no.BE2016805), National Natural Science Foundation of China (no.61503081, no.61473079) and Fund of Basic scientific research operating expenses of central universities.

ABSTRACT Noise is an unneglectable problem in emotion recognition if we want to put it into practice. First, aiming at the problem of noise in speech, we design a new acoustic feature, Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC), for better robustness. To extract LW-WPCC feature, first the best wavelet packet basis is constructed. On the basis of this, a robust wavelet packet Cepstral Coefficient is extracted by combining short time frame analysis with long time frame analysis. After that, we introduce a sub-band spectral center-of-mass parameter with good robustness to additive noise and propose an extraction algorithm of LW-WPCC. Through experiments on speech emotion recognition of different SNR levels, it is shown that our proposed method shows better noise robustness and performance on speech emotion recognition. What’s more, as facial expressions will not be affected by noise, we do bio-modal emotion recognition based on audio-visual data to improve robustness by making a decision-level fusion. Experiments based on audio-visual data are conducted to evaluate efficiency of our method. Results show that bio-modal emotion recognition based on audio-visual data can improve robustness and achieve better performance by benefiting from different kinds of emotion data.

INDEX TERMS Robust noise; Speech emotion recognition; LW-WPCC feature; Feature extraction algorithm; Bio-modal emotion recognition.

I. INTRODUCTION
At present, most of the researches on speech emotion recognition are limited to the recorded emotional corpus. However, the acquisition of voice data in real-life scenarios usually does not have the conditions in the laboratory to reduce noise influence. The voice data collected in realistic scenes is always accompanied by background noise. The existence of background noise destroys real acoustic characteristics and model parameters of speech signal, making it easier to confuse different speech signals and having a negative effect on extraction of acoustic features, thus affecting the reliability of results of speech emotion recognition.

So far, only a few studies on speech emotion recognition have considered speech signals under environmental noise conditions. Aiming at the problem of noise, most of current studies focus on the stage of pretreatment for signal denoising. For example, Schuller et al[1] studied the influence of additive noise of different signal-to-noise levels on the performance of speech emotion recognition. In Chinese Academy of Acoustics, Professor Yan’s team [2] proposed a method for emotion recognition under environmental noise conditions.

In addition, it is an important direction to extract robust speech emotion characteristics of the environment. Iliev et al[3] studied the validity and noise robustness of the acoustic flow signal in speech emotion recognition. It was shown that the glimpse airflow feature shows better noise robustness through experiments on speech emotion recognition with White Gaussian Noise. Based on previous research results of robust automatic speech recognition, Weninger F et al[4] used Non-negative Matrix Factorization (NMF) for the extraction of speech emotion characteristics, proving good noise robustness of NMF-based features in speech emotion recognition.
For the interference of sound source noise in speech emotion recognition, some researchers have tried to circumvent this problem and have improved the robustness by multi-modal recognition. Poria et al [5] studied feature extraction of multi-modal data and proposed a framework of multi-modal emotion recognition. Yan et al [6] proposed a new SKRRR fusion method for bio-modal emotion recognition based on audio-visual data. Zhang et al [7] used CNN for bio-modal emotion recognition based on audio-visual data. Somandepalli et al [8] designed a framework of multi-modal emotion recognition for real-time emotion recognition. Bio-modal emotion recognition was used for human-robot interaction of social robots by Fernando et al. Laura Boccanfuso’s [9] team used infrared thermography tracking thermal changes in five areas of the face, and used machine learning techniques for classifying thermal data by emotion state. Vikramjit Mitra’s [10] team compared the effect of noise and reverberation on depression prediction using standard Mel-frequency cepstral coefficients (MFCCs), and features designed for noise robustness, damped oscillator cepstral coefficients (DOCCs). Yuan Gong’s team [11] focused on end-to-end deep learning, a method that can learn directly from raw audio data instead of using hand-engineered features.

Although great progress on methods of improving performance of speech emotion recognition under environmental noise conditions have been made in recent years, speech emotion recognition in realistic scenes has not yet reached the point of large-scale application. It is the key point of speech recognition research under environmental noise conditions to extract acoustic features which can effectively express emotion information and show good noise robustness. Wavelet packet analysis can provide good time-frequency localization analysis for signals containing a large amount of medium and high frequency information. Since different wavelet packet bases have different capabilities for time-frequency localization, we can select suitable wavelet bases with great flexibility according to different signals.

How to extract features with noise robustness which can effectively express emotional information is our focus in this paper. Additionally, we introduce facial expression to improve our robustness. We do bio-modal emotion recognition based on audio-visual data by making a decision-level fusion, which improves robustness and takes advantages of different kinds of information.

In this paper, first the long-term frame noise is estimated by the sub-band of the optimal wavelet packet and the new sub-band spectrum is combined with the wavelet packet cepstral coefficient to construct a new acoustic feature with good noise robustness. On this basis, a robust speech emotion feature extraction algorithm based on long time frame analysis weighted wavelet packet analysis is presented. Through experiments on speech emotion recognition of different SNR levels, LW-WPCC feature proposed in this paper is analyzed and evaluated. Also, we combine speech emotion recognition and facial expression recognition and evaluate its performance by experiments based on audio-visual data of different SNR levels.

II. Sub-band noise compensation based on wavelet packet basis

Compared with fixed structure of triangular band-pass filter bank [16] used in extraction of MFCCs, wavelet packet analysis is more flexible and shows better performance on spectrum analysis. Therefore, wavelet packet analysis can be considered as an extension of the Mel-frequency triangular band-pass filter bank. Based on that, we propose a new speech spectrum characteristic, Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC), which is extracted based on noise estimation and compensation on the divided sub-bands based on optimal wavelet packet basis.

For wavelet analysis, optimal wavelet packet basis needs to be constructed. In this paper, we use the bottom-to-top optimal wavelet packet-base search algorithm described in [12] to construct the optimal wavelet packet basis. Details about wavelet packet analysis are not stated in this paper. Now we will start from introducing details about noise compensation.

A. Asymmetric Noise Suppression

Speech signals are usually divided into speech frames of duration between 10ms to 30ms for analysis according to most studies on speech emotion recognition. However, many researchers have found that longer speech frames are more effective for noise estimation [17] because energy associated with environmental noise changes more slowly than speech-related energy. Hermansky and other scholars believe that accuracy of speech emotion recognition can be improved by using the information describing the sub-band long time frame envelope in the TRAPS algorithm and FDLP algorithm [19].

In this paper, the long time frame energy \( \hat{Q}[x,i] \) is obtained by calculating the dynamic averaging of the energy component \( E[x,i] \) of the sub-band, which is shown in (1).

\[
\hat{Q}[x,i] = \frac{1}{2M+1} \sum_{x'M=-M}^{x+M} E[x',i]
\]

(1)

Where \( x \) represents the speech frame number, and \( i \) represents the sub-band sequence number. In this paper, we choose \( M = 2 \).
Energy of speech usually changes more rapidly than that of environmental noise in the same sub-band. In other words, speech component usually has a higher frequency spectrum than the noise component. More or less, many algorithms for noise suppression such as RASTA-PLP process use a high-pass filter or a band-pass filter [20]. The simplest method is to perform high-pass filtering [21] on each band to suppress low-variability components in the signal, which are more likely to be noise rather than valid speech components. One issue that we can’t have failed to notice is that the output of a conventional high-pass filter [22] in energy field may be negative. Thus, speech signal can be processed by taking logarithm of energy component first and filtering later, which is similar to signal processing in the extraction of MFCCs.

Asymmetric noise suppression (ANS) method is often adopted for noise compensation. Besides, spectral subtraction is another useful way for noise suppression. Hence, in this paper, an asymmetric filter is applied to obtain the time-varying noise threshold. Then, the noise threshold is subtracted from the instantaneous energy for noise compensation.

Asymmetric noise suppression and transient masking process are shown in Figure 1. Asymmetric suppression process for input \( \hat{Q}_{in}[x,i] \) and output \( \hat{Q}_{out}[x,i] \) is expressed as:

\[
\hat{Q}_{out}[x,i] = F_{\lambda_a, \lambda_b} \left[ \hat{Q}_{in}[x,i] \right]
\]

where \( \lambda_a = 0.999 \) and \( \lambda_b = 0.5 \) [23]. For more convenient expression, Equation (3) is used to simplify the asymmetric suppression process expressed by (2), which means filtering on each sub-band \( i \) of speech frame \( x \).

\[
\hat{Q}_{as}[x,i] = \lambda_a \hat{Q}_{as}[x-1,i] + (1-\lambda_a) \hat{Q}_{as}[x,i] \quad \hat{Q}_{as}[x,i] \geq \hat{Q}_{as}[x-1,i] \\
\lambda_b \hat{Q}_{as}[x-1,i] + (1-\lambda_b) \hat{Q}_{as}[x,i] \quad \hat{Q}_{as}[x,i] < \hat{Q}_{as}[x-1,i]
\]

It’s proved that we can get best results when \( \lambda_a = 0.999 \) and \( \lambda_b = 0.5 \) [23]. For more convenient expression, Equation (3) is used to simplify the asymmetric suppression process expressed by (2), which means filtering on each sub-band \( i \) of speech frame \( x \).

\[
\hat{Q}_{as}[x,i] = F_{\lambda_a, \lambda_b} \left[ \hat{Q}_{as}[x,i] \right]
\]

The basic calculation steps of (3) are expressed explicitly as follows:

1. First, the lower envelope \( \hat{Q}_{le}[x,i] \) of \( \hat{Q}[x,i] \) is calculated according to (4).

\[
\hat{Q}_{le}[0,i] = 0.9 \hat{Q}[0,i] \quad \text{where } \hat{Q}_{le}[0,i] \text{ is initialized to } 0.9 \hat{Q}[0,i]
\]

2. Secondly, \( \hat{Q}_{le}[x,i] \) is obtained through an ideal linear half-wave rectification on \( \hat{Q}[x,i]-\hat{Q}_{as}[x,i] \):

\[
\hat{Q}_{le}[x,i] = \begin{cases} 
\hat{Q}[x,i]-\hat{Q}_{as}[x,i] & \hat{Q}[x,i]-\hat{Q}_{as}[x,i] > 0 \\
0 & \hat{Q}[x,i]-\hat{Q}_{as}[x,i] \leq 0
\end{cases}
\]
(3) Thirdly, the lower envelope $\tilde{Q}_i[x,i]$ of the rectified output $\tilde{Q}_i[x,i]$ is obtained by the ANS process:

$$\tilde{Q}_i[x,i] = F_{2999,0.5}[\tilde{Q}_i[x,i]]$$  \hspace{1cm} (6)

(4) Fourthly, the temporal masking output $\tilde{Q}_m[x,i]$ is limited by using $\tilde{Q}_i[x,i]$ obtained by the previous step as threshold ($\tilde{Q}_m[x,i]$ will be described in section II-B):

$$\tilde{Q}_m[x,i] = \max(\tilde{Q}_m[x,i], \tilde{Q}_i[x,i])$$  \hspace{1cm} (7)

(5) Last, the output $\tilde{R}[x,i]$ of the ANS module is calculated according to (8). Where it shows best performance on speech emotion recognition under environmental noise conditions when $c = 2$.

$$\tilde{R}[x,i] = \left\{ \begin{array}{ll} \tilde{Q}_i[x,i] & \tilde{Q}_i[x,i] \geq c\tilde{Q}_m[x,i] \\ \tilde{Q}_m[x,i] & \tilde{Q}_i[x,i] < c\tilde{Q}_m[x,i] \end{array} \right. $$  \hspace{1cm} (8)

B. Temporal masking

There is a temporal masking in the human auditory system. In this section, a simple method to solve the problem of time domain masking will be introduced.

First, the real-time peak energy $\tilde{Q}_s[x,i]$ for each sub-band is calculated according to (9).

$$\tilde{Q}_s[x,i] = \max(\lambda\tilde{Q}_s[x-1,i],\tilde{Q}_s[x,i])$$  \hspace{1cm} (9)

Where $\lambda$ is the real-time peak forgetting factor, $x$ and $i$ correspond to the speech sequence number and the sub-band number respectively.

Then, the temporal masking value of the speech segment is calculated by:

$$\tilde{R}_0[x,i] = \left\{ \begin{array}{ll} \tilde{Q}_i[x,i] & \tilde{Q}_i[x,i] \geq \lambda\tilde{Q}_s[x-1,i] \\ \tilde{Q}_s[x,i] & \tilde{Q}_i[x,i] < \lambda\tilde{Q}_s[x-1,i] \end{array} \right. $$  \hspace{1cm} (10)

Studies have shown that it shows best performance on speech emotion recognition under environmental noise conditions when $\lambda = 0.85$ and $\mu_t = 0.2$.

C. Spectral Weight Smoothing

A few studies [21] have found that smoothing of sub-bands is a good way for noise compensation and improving performance on speech emotion recognition under environmental noise conditions.

Because of different thresholds for different sub-bands, we combine the asymmetric noise suppression method and the temporal masking solution to form a conversion function $\tilde{R}[x,i]/\tilde{Q}[x,i]$, which is smoothed by calculating the dynamic average of $\tilde{R}[x,i]/\tilde{Q}[x,i]$ on sub-band $i$.

Time-frequency average conversion function is expressed in (11).

$$\tilde{S}[x,i] = \frac{1}{i_t - i_t + 1} \sum_{i=1}^{i_t} \tilde{R}[x,i]$$  \hspace{1cm} (11)

Where $i_t = \min(i + N, I)$, $i_t = \max(i - N, 1)$, $I$ is the total number of sub-bands. $\tilde{S}[x,i]$ is used to modulate the initial short-term frame energy $E[x,i]$ . $T[x,i]$ is the time-frequency normalized value that combines short-term frame and long-term frame speech energy.

$$T[x,i] = E[x,i] \tilde{S}[x,i]$$  \hspace{1cm} (12)

D. Spectral centroid

Traditional acoustic features such as prosodic features, voice quality features and cepstral coefficient features are sensitive to additive noise. The existence of noise will affect the extraction of these acoustic features. As a consequence, the accuracy and robustness of speech emotion recognition system under noise conditions will be reduced. However, Sub-band Spectral Centroid (SSC) is relatively less sensitive to noise, so it shows better noise robustness. Sub-band spectral centroid reflects the distribution of the signal energy in the sub-band, thus it has a good complementarity with cepstral features which can reflect energy of frequency band of the signal. Therefore we can combine the two to construct a new feature to improve performance.

The frequency band $[0, f_1]$ of the original signal is divided into a series of sub-bands by wavelet packet decomposition. The set of leaf nodes in wavelet packet $\mathcal{T}$ is given as $\mathcal{L}(\mathcal{T}) = \{(j_i, p_i)_{1 \leq i \leq I}\}$, and the signal energy of node $(j_i, p_i)$ is centrally distributed within the following range:

$$[f_{\ell_1}^{\ell_0}, f_{\ell_2}^{\ell_0}] = \{q_i \frac{\ell_0}{2^{j_i}}, (q_i + 1) \frac{\ell_0}{2^{j_i}} \} \hspace{1cm} 1 \leq i \leq I$$  \hspace{1cm} (13)

Where $q_i = G[p_i]$ is the inverse of the Gray code.

The power spectrum $P(f)$ of the signal in the $i$th sub-band $[f^{(i)}_{\ell_1}, f^{(i)}_{\ell_2}]$ is estimated by the autoregressive spectral estimation method [16]. According to $P(f)$ in the sub-band, we can calculate the corresponding sub-band Spectral Centroid:

$$C_i = \frac{\int_{f^{(i)}_{\ell_1}}^{f^{(i)}_{\ell_2}} f \cdot P_i(f) df}{\int_{f^{(i)}_{\ell_1}}^{f^{(i)}_{\ell_2}} P_i(f) df}$$  \hspace{1cm} (14)

Where $\lambda$ is the control factor of dynamic range of power spectrum and $\lambda < 1$. Previous studies show that the best performance of amplitude compensation can be obtained when the value of $\lambda$ is 0.5[27].
Because of complementary benefits of the wavelet packet cepstral coefficient and sub-band spectral centroid, in this paper, a weighted method is adopted to combine the wavelet packet cepstral coefficient and the sub-band spectral centroid. After that, it is combined with noise estimation and compensation technique of long-time frame to obtain our proposed acoustic feature—Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC). The process of the extraction of LW-WPCC feature can be represented by a block diagram as shown in Figure 2.

**FIGURE 2.** The calculation block diagram of noise robust Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient

Steps of Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC) extraction algorithm can be described as follows:

1. Endpoint detection, pre-emphasis, framing, windowing, and other pre-processing on the original speech signal $s$.
2. Wavelet packet decomposition on each speech frame $x$ to obtain the wavelet packet set $\{d_{j,i}\}_{i,j \in I}$ and frequency sorting on $I$ leaf nodes of the wavelet packet tree $T$ to obtain $\{d_i\}_{i \in I}$ in ascending order of its frequency range.
3. Calculation of the energy component $E_i$ of each sub-band of speech frame $x$ and $T[x,i]$ according to Section II

$$E_i = \sum_{n=-\infty}^{\infty} |d_{i,n}|^2, \quad 1 \leq i \leq I$$

(15)

4. Taking Logarithm of $T[x,i]$ of every sub-band of speech frame $x$.
5. Autoregressive spectral estimation to estimate $P_i(f)$ of each sub-band of speech frame $x$ and calculation of $C_i$ of each sub-band of speech frame $x$.
6. Calculation of weights $w_i, i = 1, 2, \ldots, I$ according to (16) and weighted sub-band energy $WL_i$ according to (17).

$$w_i = \frac{C_i - f_i^{(v)}}{f_i^{(H)} - f_i^{(L)}}$$

(16)
Yongming Huang, Jing Xiao, Kexin Tian, Ao Wo, Guobao Zhang: Research on Robustness of Emotion Recognition under Environmental Noise Conditions

\[
WL_x = w_i \cdot \lg(T[x,i])
\]  
(17)

(7) Discrete cosine transform (DCT) for weighted sub-band energy to obtain LW-WPCC of L order, which is expressed as (18). In this paper, \( L = 12 \).

\[
\text{LW-WPCC}_i = \sum_{t=1}^{L} \cos[(i-0.5)\cdot\frac{\pi}{L}]\cdot WT_{li}, \quad 1 \leq i \leq L
\]  
(18)

(8) Taking logarithm of energy of speech frame as 0-order LW-WPCC coefficient to form our feature vector of \( L+1 \) dimensions with LW-WPCC of L order.

IV. Experiments on Speech Emotion Recognition under Environmental Noise Conditions

A. Experiment Setup

The overall process of speech emotion recognition based on Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient is shown in Figure 3, which can be divided into three stages: construction of optimal wavelet packet base, training and classification. Emotional sample sets in three stages are independent of each other. Therefore, before performing experiments on speech emotion recognition under environmental noise conditions, synthetic samples should be divided into three independent sets, which are used for wavelet packet base constructing, classifier training and testing.

Taking everything into consideration, we use 20% of the emotional speech samples to construct the wavelet packet base, and the remaining are used for classifier training and testing. We apply 5-fold cross validation method for training and testing (20% testing samples, 80% training samples) to guarantee objectivity of our experimental results. Preprocessing is the same as our previous work on speech emotion recognition [13]. The frame length is 16ms and the frame is shifted to 10ms in this paper.

Experiments on speech emotion recognition under environmental noise conditions are performed on Berlin emotional corpus. For inadequate number of samples of disgust and 79.6% accuracy rate on humans’ judgements on samples of disgust which is much lower than other emotional categories in Berlin emotional corpus, samples of disgust are not included in our experiments. Samples of other 6 emotional categories (489 sentences) are considered in our experiments. And speech with background noise is synthesized by noise in the NOISEX-92 and speech in the Berlin emotional corpus. The NOISEX-92 is synthesized by white noise, such as noise from cars, restaurants, factory workshops and pink noise.

FIGURE 3. The overall process of speech emotion recognition based on Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient

Daubechies wavelet filter is applied for wavelet decomposition in experiments. According to its disappearance of the different values, Daubechies wavelet can be divided into different orders. The order of common used Daubechies wavelet filter is 4 to 40, which can be recorded as Db4 to Db40. Taking bandpass filtering characteristic of the wavelet packet basis and the computational complexity of feature extraction into account, Db40 is adopted in our experiments.

Psychoacoustic studies [28] show that humans’ auditory system can be modeled by a set of multi-channel filters. Each sub-band of the multi-channel filter bank, namely critical band, corresponds to a physical location of the cochlear basement membrane [26]. Humans’ auditory perception region (20Hz ~ 16kHz) is usually divided into 24 critical bands [29]. Auditory masking may occur in the same critical band. Therefore, the number of divided sub-bands in our work is limited to a small range around the number of critical bands. Since speech frequency range of Berlin corpus is [0, 8000 Hz] and there are about 21 critical bands in this

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Yongming Huang, Jing Xiao, Kexin Tian, Ao Wo, Guobao Zhang: Research on Robustness of Emotion Recognition under Environmental Noise Conditions

frequency range, the number of sub-bands divided by wavelet packet is 15 to 25 in this paper. Experiments of different number of sub-bands on speech emotion recognition under environmental noise conditions will be performed to evaluate the influence of different numbers of sub-bands based on wavelet packet base.

The Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC) described in Section II is a short-term feature, so it is used as an acoustic low-level descriptor. It’s statistically calculated according to the statistical characteristics listed. 39 low-level descriptors consist of 13 dimensional spectral centroid-weighted noise robust wavelet packet cepstral coefficients and their first-order, second-order statistical feature parameters of differences. Therefore, the dimension of acoustic feature vectors is 13*3*11 = 429. So it is necessary to reduce the dimensionality of features. The method of feature reduction is based on the feature selection method of speech segment trajectory model to select features which contain rich emotional information.

As can be seen from section II, one optimal wavelet packet base corresponds to each division number of sub-bands, which represents a way of sub-band division. Therefore, different numbers of sub-bands is to choose different methods of frequency band division. Figure 4 shows the results of accuracies of speech emotion recognition under environmental noise conditions with different numbers of sub-bands. The signal-to-noise ratio (SNR) level of synthetic signal is 30db. The wavelet packet base is Daubechies wavelet basis with Db40 and feature vectors are original vectors without dimensionality reduction. We apply 5-fold cross-validation for training and testing.

As Figure 4 is shown, number of sub-bands has influence on the accuracy of speech emotion recognition under noise conditions, which means different divisions of frequency bands show different abilities for emotion recognition. Frequency domain distribution of speech energy is affected by emotional state in the speech. When extracting LW-WPCC, band division is a concrete description of speech by emotional state in the speech. When extracting LW-WPCC, band division is a concrete description of speech energy distribution. Therefore, method of band division with higher accuracy can describe the frequency domain distribution of emotion information in speech signal better, which also shows that emotional states do have an effect on the distribution of the speech energy.

What’s more, it can be seen from Figure 4 that different sub-bands show different abilities for emotion representation. At the same time, it’s reflected that accuracy of speech emotion recognition has certain relation with number of sub-bands. When the number of sub-band division is small, which means a rough division, the accuracy is low. With the increase of the number of sub-bands, accuracy is also improved accordingly. However, when the number of sub-bands increases to some extent, more detailed division of frequency band is no longer helpful, even the accuracy will decrease. These experiments show that we have better performance when the number of sub-bands is 19 to 21, especially we have highest accuracy of 72.12% when the number of sub-bands is 21.

B. The effect of frequency band division on Speech Emotion Recognition under environmental noise conditions

The effect of sub-band division on emotion recognition under noise condition is evaluated by experiments in this section.

For classification, we use support vector machine with the weight weighting of the radial basis function as classifier. For six categories of emotional classification, we need to train 6 * (6-1) / 2 = 15 "one to one" SVM classifiers. Lib-SVM open source library which is developed by Professor Lin Zhiren and other researchers in the Taiwan University is used to achieve that [30]. Experiments are performed on MATLAB R2015a on Windows 7.

C. Comparison of LW-WPCC Feature and Traditional Acoustic Features in Emotion Recognition under Environmental Noise conditions

Comparison of our proposed noise robust Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient and
traditional acoustic features is our focus in this section. We use commonly used traditional acoustic features as baselines to evaluate our proposed feature LW-WPCC. Three chosen sets of traditional acoustic features (prosodic, sound quality and spectral characteristics) are shown in Table II. These acoustics features are short-time so preprocessing in extraction of them is the same as that of wavelet packet cepstral coefficients.

### Table II

**The Traditional Acoustic Characteristics Used for Comparison**

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Low Order Descriptor (LLD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic</td>
<td>Fundamental frequency</td>
</tr>
<tr>
<td></td>
<td>Logarithmic frame energy</td>
</tr>
<tr>
<td></td>
<td>First formant</td>
</tr>
<tr>
<td>Sound Quality</td>
<td>Second formant</td>
</tr>
<tr>
<td></td>
<td>Third formant</td>
</tr>
<tr>
<td>Spectral Characteristics</td>
<td>WPCC</td>
</tr>
</tbody>
</table>

In our experiments, the noise is added considering several signal to noise ratio (SNR) levels of a range from 30 to 0 dB.

The preprocessing of speech signal is the same as that in section IV-A, and the number of optimal wavelet packet baseband divisions used to calculate WPCC and LW-WPCC is 21 because we have best performance on speech emotion recognition when the number of sub-bands is 21, which is shown in Figure 4. The extracted WPCC and LW-WPCC and three sets of traditional acoustic features are used as low-level descriptors. According to the statistical features in Table I, the feature selection method of global statistical features is same as the feature selection method in our previous experiment. The calculated statement-level statistical features are used for training and testing of the support vector machine. In this section, noise robustness and accuracy of various characteristics are evaluated and compared on speech emotion recognition under different SNR levels. Also, 5-fold cross validation method is applied. The results are shown in Figure 5.

As is shown in Figure 5, for all the features, the lower SNR is, the lower accuracy of speech emotion recognition is. It’s to be noticed that the robustness of WPCC is worst in the 5 features. When SNR decreases, accuracy of speech emotion recognition of WPCC decreases sharply, which is the most obvious among features used in our experiments, and accuracy of WPCC is always the lowest when SNR is less than 15db. Also, it’s obvious that noise sensitivity of spectral features is higher than that of prosodic and sound quality characteristics.

![Figure 5. Five-fold cross validation test results on different SNR levels.](image)

However, accuracy of spectral characteristics outweighs that of prosodic and sound quality characteristics. In the spectral characteristics, the LW-WPCC feature has better noise robustness than the WPCC feature.

When SNR of speech signal is higher than 25db, compared with WPCC characteristic, performance of LW-WPCC characteristic does not show obvious advantages. However, with the reduction of SNR, especially when SNR is between 5db and 15db, we can see that LW-WPCC feature shows better noise robustness. Therefore, it is proved that the wavelet packet feature extraction algorithm based on the long time frame and short time frame analysis proposed in this paper does improve the robustness of speech emotion recognition to some extent.

In summary, although LW-WPCC feature is not prominent in the speech emotion recognition without noise and increased computational load for the noise compensation steps, LW-WPCC feature has better performance than WPCC on noise robustness.

## V. Bio-modal emotion recognition based on audio-visual data

The way humans express their emotion is usually multimodal: visual, audio, textual modalities and so on. And visual and audio data convey 90% of humans’ emotion, which are useful clues for emotion recognition. In real life, we get better understanding of a speaker’s emotion when we see his/her facial expressions while he/she is speaking. It’s obvious that together audio and visual signals provide more information than they provide alone. What’s more, audio signals are often disturbed by background noise inevitably and visual signals are not. Therefore, audio-visual emotion recognition is a good way to increase the accuracy and reliability of emotion recognition.

### A. Framework

The framework of bio-modal emotion recognition based on audio-visual data in this paper is shown in Figure 6. First,
speech emotion recognition and facial expression recognition are done exclusively. Then, we make a decision-level fusion.

For speech emotion recognition, our proposed acoustic feature LW-WPCC and other traditional acoustic features are extracted, which are explained in detail in Section V. Facial feature extraction algorithm is our focus in this Section.

B. Facial feature Extraction

Our research on audio-visual emotion recognition is based on data in IEMOCAP [31]. Facial data in IEMOCAP are trajectories of the facial markers’ positions while speaking. Facial data corresponding to a whole sentence is a matrix consisting of many frames. The data of 3D position of every marker is recorded in every frame. 53 markers are attached to the face to capture facial expression information, which are placed according to the feature points defined in MPEG-4 standard [31]. The layout of facial markers is shown in Figure 7.

For facial expression recognition on IEMOCAP, framework is shown in Figure 8.

Emotion expression is a continuous process. However, not all the facial expressions while speaking convey current emotional state as Figure 9 is shown. It’s unnecessary to extract facial features from all the facial expression images in a process for unnecessary computational overhead. Therefore, information from emotional start frame to emotional peak frame, which contains most of emotional information, is used in our paper.
In order to discover emotional apex automatically, Slow Feature Analysis (SFA) algorithm [32] is adopted. SFA is an unsupervised learning method for learning invariant or slowly varying features from temporal varying input signals. The slowest output is the most slowly varying feature which can be regarded as the most important feature. Therefore, the frame with maximum amplitude of the slowest output is selected as the emotional apex. Then we find out the start of emotional state according to Figure 10. Figure 11 and Figure 12 show selected emotional onset and apex of a happy sample.

After that, we need to normalize frame numbers (10 in this paper) by linear interpolation since dimension of feature vectors of samples need to be the same. There are three-dimensional coordinates of 55 facial feature points in a frame. Therefore, the dimension of facial feature vectors is 55*3*10 = 1650.

For feature learning, the stacked auto-encoder is used. Recently, deep learning has aroused wide concern. Many studies and practical successes demonstrate the advantage of using deep architectures to learn more complex features and achieve good classification performance.

An auto-encoder is one type of neural networks, structurally defined by three layers: input layer, hidden layer, and output layer, as Figure 13 is shown. The aim of the auto-encoder is to learn representation of the input by minimizing the reconstruction error between the input and the reconstructed one from the learned representation.

The basic auto-encoder is trained by minimizing the loss function according to (19):

$$J_\theta = \frac{1}{m} \sum_{i=1}^{m} (\hat{x}^{(i)} - x^{(i)})^2 + \lambda \sum_{j=1}^{2} ||W^{(j)}||$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\hat{x}^{(i)} - x^{(i)})^2 + \lambda \sum_{j=1}^{2} ||W^{(j)}||$$

(19)

Where, $m\sum_{i=1}^{m} (\hat{x}^{(i)} - x^{(i)})^2$ denotes the reconstruction error, $\lambda \sum_{j=1}^{2} ||W^{(j)}||$ is integrated into the loss function to prevent over-fitting.
Several auto-encoders can be stacked to construct a deep learning model, which is the stacked auto-encoder, to learn hierarchical features for the original input, as is shown in Figure 14 and Figure 15. The stacked auto-encoder is trained by two stages: pre-training and fine-tuning. As shown in Figure 14, in the pre-training stage, unsupervised greedy layer-wise pre-training from bottom to top is completed. After pre-training, the parameters are set to the initial parameters of the stacked auto-encoder. As presented in Figure 15, in the fine-tuning stage, some labeled samples are used to fine-tune the whole network.

C. Decision-level Fusion

Hierarchical features learnt from the stacked auto-encoder are the inputs of SVM classifiers. After getting classification results of speech emotion recognition and facial expression recognition alone, decision-level fusion is needed.

For decision-level fusion, first we need to get posterior probabilities. The output of SVM classifier is normalized to the range of 0 and 1 according to (20).

\[
p(y = 1 | f(x)) = \frac{1}{1 + e^{W(x) + B}}
\]

(20)

Where \( f(x) \) is the output of SVM classifier, \( p(y = 1 | f(x)) \) corresponds to the probability of correct classification when the output is \( f(x) \), A and B are parameters we need to settle. A and B are determined by minimizing negative log likelihood loss, which is shown in (21).

\[
-\sum_{i=1}^{n} \left[ t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \right]
\]

(21)

Where

\[
p_i = \frac{1}{1 + e^{W(x) + B}}
\]

(22)

\[
t_i = \frac{y_i + 1}{2}
\]

(23)

Suppose there are \( E \) categories and \( L \) modals for sample \( x \). Then we can get vector of posterior probabilities of sample \( x \), as (24) is shown.

\[
\{p_{e}(x), l = 1, 2, ... , L; e = 1, 2, ... , E\}
\]

(24)

Then the weighted method is adopted to get probabilities of \( E \) categories of the sample according to (25) and (26). Where \( \alpha_l \) is the weight of the \( l^{th} \) modal.

\[
p_{e}(x) = \sum_{l=1}^{L} \alpha_l p_{e}(x)
\]

(25)

\[
\sum_{l=1}^{L} \alpha_l = 1
\]

(26)

So the final category is determined by (27).
\[ f(x) = \arg \max_{\alpha} (p_\alpha(x)) \] (27)

VI. Emotion recognition based on audio-visual data under Environmental Noise Conditions

A. Experiment setup
For experiments based on audio-visual data, improvised data in IEMOCAP is adopted. 739 sentences which are classified in four categories: anger, happiness, neutral state and sadness (happiness is combined by happiness and excitement) are selected as samples for our study. Speech with noise is synthesized by noise in NOISEX-92 and speech in IEMOCAP.

The framework of bio-modal emotion recognition based on audio-visual data in this paper is shown in Figure 6. First, speech emotion recognition and facial expression recognition are done exclusively. Then, we make a decision-level fusion.

For speech emotion recognition on IEMOCAP, in general, experimental methods and settings on IEMOCAP are the same as those on BEO, which is introduced in section IV-A. What’s different is that \(4 \times (4-1)/2 = 6\) “one to one” SVM classifiers need to be trained for 4 emotional categories in IEMOCAP.

B. Comparison of LW-WPCC Feature and Traditional Acoustic Features in Emotion Recognition under Environmental Noise conditions
In this section, comparison of LW-WPCC and traditional acoustic characteristics of speech emotion recognition under environmental noise conditions on IEMOCAP is focus. Traditional acoustic features and experimental methods are the same as those in section IV-C.

The number of sub-bands used to calculate WPCC and LW-WPCC on IEMOCAP is determined to be 24 according to experiments of different numbers of sub-bands on IEMOCAP. Other experimental settings are the same as those in section IV-C. 5-fold cross validation method is also applied in this section.

As Figure 16 is shown, we can get the same results in section IV-C. Our proposed LW-WPCC feature shows better performance than WPCC in terms of noise robustness, especially under 15-5dB SNR levels.

C. Comparison of Single-modal and Bio-modal Emotion Recognition under Environmental Noise conditions
In this section, bio-modal emotion recognition based on audio-visual data is mainly discussed. First facial expression recognition and speech emotion recognition are done exclusively according to method and experimental setting expressed in section V-A. Then a decision-level fusion is applied according to section V-A.

Weights of modals for decision-level fusion are determined according to (27). Where, \(\text{Accuracy}_f\) is accuracy of facial expression recognition and \(\text{Accuracy}_v\) is accuracy of speech emotion recognition.

\[ \alpha_f = \frac{\text{Accuracy}_f}{\text{Accuracy}_v} \alpha_f + \alpha_v = 1 \] (27)

Table III-V show results on IEMOCAP (without noise). We can observe that accuracy of sadness is not satisfying in facial expression recognition while recognition of sadness in speech emotion recognition shows good result. Facial expression recognition shows better performance on other emotions. Therefore, emotion recognition based on audio and visual data is a good way to improve accuracy. Besides, facial expressions are not affected by noise. Therefore, it’s feasible to improve accuracy under noise conditions by audio-visual emotion recognition.

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>CONFUSION MATRIX OF FER ON IEMOCAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral State</td>
</tr>
<tr>
<td>Neutral State</td>
<td>0.8966</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.2083</td>
</tr>
<tr>
<td>Anger</td>
<td>0.1500</td>
</tr>
<tr>
<td>Happiness</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV</th>
<th>CONFUSION MATRIX OF SER ON IEMOCAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral State</td>
</tr>
<tr>
<td>Neutral State</td>
<td>0.4828</td>
</tr>
<tr>
<td>Sadness</td>
<td>0</td>
</tr>
<tr>
<td>Anger</td>
<td>0.1000</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.1455</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE V</th>
<th>CONFUSION MATRIX OF AUDIO-VISUAL EMOTION RECOGNITION ON IEMOCAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral State</td>
</tr>
<tr>
<td>Neutral State</td>
<td>0.8621</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.0208</td>
</tr>
<tr>
<td>Anger</td>
<td>0.0500</td>
</tr>
<tr>
<td>Happiness</td>
<td>0</td>
</tr>
</tbody>
</table>
Results of single-modal and bio-modal emotion recognition under noise conditions are given in Figure 17.

![Graph showing emotion recognition accuracy under different SNR levels](image)

**FIGURE 17.** Results of single-modal and bio-modal emotion recognition

It’s shown that emotion recognition based on audio-visual data shows better performance than speech emotion recognition at any time because more information is made full use of under the premise of ideal conditions for FER. Accuracy of FER is easy to be affected by lights and positions in real life, influences of which are avoided in IEMOCAP. Also, we can see that emotion recognition based on audio-visual data shows better performance under environmental noise conditions since facial expressions are not affected by noise. However, as SNR decreases, accuracy of audio-visual emotion recognition declines for unreliable acoustic features under noise conditions. But we can observe that when SNR is 0-25db, accuracy of audio-visual emotion recognition is around 15% higher than the accuracy of speech emotion recognition, which is around 10% higher under conditions without background noise. Therefore, audio-visual emotion recognition is a good choice for better accuracy and robustness.

**VII. Conclusions**

This paper focuses on speech emotion recognition under environmental noise conditions. Aiming at background noise in speech data, firstly, the optimal wavelet packet base is constructed. On this basis, a robust wavelet packet cepstral coefficient is extracted by combining short time frame analysis and long time frame analysis. Secondly, we introduce a sub-band spectral center-of-mass parameter with good robustness to additive noise and propose a noise robust Long time frame Analysis Weighted Wavelet Packet Cepstral Coefficient (LW-WPCC) extraction algorithm. Besides, emotion recognition based on audio-visual data is preliminarily discussed in this paper. Facial expression is introduced to achieve better noise robustness.

Experiments on speech emotion recognition of different SNR levels show that LW-WPCC feature proposed in this paper shows better noise robustness and higher accuracy of recognition compared with traditional acoustic characteristics under environmental noise conditions. And when SNR level is lower than 15dB, it’s better to extract LW-WPCC for recognition for better robustness. Also, experiments prove that multi-modal emotion recognition is also a useful way to improve accuracy. Emotion recognition based on audio-visual data shows better performance than speech emotion recognition because more information is made full use of without considering light and position influences. Therefore, audio-visual emotion recognition is a good way for better performance under environmental noise conditions. But when position and light are not appropriate for facial expression recognition, we may get worse results.

For further study, we need to explore more on multi-modal emotion recognition, which is preliminarily discussed in this paper. Also, more limited realistic conditions about emotion recognition should be considered in our future study.

**REFERENCES**


[12] Yongming Huang*, Wu Ao, Guobao Zhang, Novel Sub-band Spectral Centroid Weighted Wavelet Packet Features with


[16] Li Yue. Research on speech emotion recognition based on optimal wavelet packet decomposition[D]. Southeast University, 2014.


Yongming Huang, Jing Xiao, Kexin Tian, Ao Wu, Guobao Zhang: Research on Robustness of Emotion Recognition under Environmental Noise Conditions

Yongming Huang was born in April 1982. He received the Ph.D. degree in Pattern Recognition and Artificial Intelligence from Southeast University in 2012. He has been an associate professor in College of Automation in Southeast University since 2012. His research interests include speech recognition, speech emotion recognition, affective computing and robots. He has published 3 SCI papers and 15 EI retrieval papers as first author.

Jing Xiao received the B.S. degree in automation from Southeast University. She is currently pursuing the M.Sc. degree in automation at Southeast University. Her research interests include emotion recognition and artificial intelligence.

Kexin Tian received the B.S. degree in automation from Southeast University. She is currently pursuing the M.Sc. degree in automation at Southeast University. Her research interests include speech emotion recognition and artificial intelligence.

Ao Wu received the M.Sc. degree in automation from Southeast University. His research interests include speech emotion recognition and machine learning.

Guobao Zhang was born in 1965. He received the B.S. degree from Zhejiang University in 1986. He received the M.Sc. degree and the Ph.D. degree in Automation from Southeast University in 1991 and 1998. He was a visiting scholar in Hong Kong Polytechnic University from 1999 to 2001. He has been a professor in College of Automation in Southeast University since 1991. His research interests include complex networks and intelligent control. He has published more than 90 papers in Chinese and international journals, including 6 SCI papers and 22 EI retrieval papers.