www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **38**, No. 4, pp. 457–463 (2013)

DOI: 10.2478/aoa-2013-0054

Speech Emotion Recognition under White Noise

Chengwei HUANG[(1)], Guoming CHEN[(1)], Hua YU[(1)], Yongqiang BAO[(2)], Li ZHAO[(1)]

[(1)] *School of Information Science and Engineering*
Southeast University
2# Sipailou, Nanjing 210096, Jiangsu Prov., China; e-mail: huang.chengwei1@gmail.com

[(2)] *School of Communication Engineering*
Nanjing Institute of Technology
1# Hongjing Ave., Nanjing 211167, Jiangsu Prov., China

(*received April 30, 2012; accepted February 6, 2013*)

Speaker's emotional states are recognized from speech signal with Additive white Gaussian noise (AWGN). The influence of white noise on a typical emotion recogniztion system is studied. The emotion classifier is implemented with Gaussian mixture model (GMM). A Chinese speech emotion database is used for training and testing, which includes nine emotion classes (e.g. happiness, sadness, anger, surprise, fear, anxiety, hesitation, confidence and neutral state). Two speech enhancement algorithms are introduced for improved emotion classification. In the experiments, the Gaussian mixture model is trained on the clean speech data, while tested under AWGN with various signal to noise ratios (SNRs). The emotion class model and the dimension space model are both adopted for the evaluation of the emotion recognition system. Regarding the emotion class model, the nine emotion classes are classified. Considering the dimension space model, the arousal dimension and the valence dimension are classified into positive regions or negative regions. The experimental results show that the speech enhancement algorithms constantly improve the performance of our emotion recognition system under various SNRs, and the positive emotions are more likely to be miss-classified as negative emotions under white noise environment.

Keywords: speech emotion recognition; speech enhancement; emotion model; Gaussian mixture model.

Notations

$x(t)$ – clean speech signal,
$n(t)$ – noise signal,
$y(t)$ – speech signal contained noise,
ω – Fourier frequency,
$X(\omega)$ – Fourier transformation of $x(t)$,
$Y(\omega)$ – Fourier transformation of $y(t)$,
$N(\omega)$ – Fourier transformation of $n(t)$,
$P_x(\omega)$ – power spectral density of $x(t)$,
$P_y(\omega)$ – power spectral density of $y(t)$,
$P_n(\omega)$ – power spectral density of $n(t)$,
m – index of speech signal frame,
k – discrete Fourier frequency,
$X(m,k)$ – discrete Fourier Transformation of the m-th frame,
$\widehat{X}(m,k)$ – enhanced speech from $X(m,k)$,
G_{SP} – transfer function,
$\xi_{m|m'}$ – priori SNR,
γ_m – posterior SNR,
$T(m,k)$ – masking threshold,
$\alpha(m,k)$ – parameter in the second speech enhancement algorithm (10),
M – symbol for simplicity of presentation,
σ_s^2 – speech signal power,
σ_n^2 – noise signal power,
B – symbol for simplicity of presentation,
C – symbol for simplicity of presentation,
$\mathbf{S}$ – feature vector of the input sample,
$\boldsymbol{\lambda}$ – parameters of GMM,
q – index of the Gaussian mixtures,
Q – the mixture number,
a_i – the mixture weight,
b_i – Gaussian distribution function,
j^* – index of the target emotion,
N – number of emotions,
j – index of emotions,
μ_q – mean of q-th Gaussian distribution,
$\boldsymbol{\Sigma}_q$ – covariance matrix of q-th Gaussian distribution,
K – class number in the K-mean clustering.

1. Introduction

Emotions in vocal communication are very important for understanding speaker's intention, mood, and attitude. Unlike linguistic information, affective information is expressed even without the notice of the

speaker (Johnstone *et al.*, 2005; Scherer, 2003). These emotions are naturally expressed and uneasy to disguise or control, which makes speech emotion recognition very important in natural human-computer interaction.

There are many challenges in the real world applications. In the in-car environment, driver's emotional stability is a crucial problem for driving safety. C.M. Jones and I. Jonson studied an in-car emotion recognition system which may help the driver to respond appropriately (Jones, Jonson, 2005). C. Clavel *et al.* studied the fear-type emotions in an audio-based surveillance system (Clavel *et al.*, 2008). Fear-type emotions may be a sign of potential threats. J. Ang *et al.* used the prosodic features to detect frustration and annoyance in natural human-computer dialog (Ang *et al.*, 2002).

Oriented to the real world applications, the noise problem is considered in our research. Noise is an important factor which affects the performance of most speech recognition systems (Varga, Steeneken, 1993). However, it has rarely been studied in speech emotion recognition, since most of the researches were carried out in an ideal lab environment (Huang *et al.*, 2011; Huang *et al.*, 2009; Johnstone *et al.*, 2005; Neiberg, 2006; Truong, 2009). Schuller *et al.* first studied the noise problem in automatic speech emotion recognition (Schuller *et al.*, 2006). Tawari *et al.* then proposed a framework to improve the speech emotion recognition under noisy environment (Tawari, Trivedi, 2010). In their framework the noise cancellation was based on the adaptive thresholding in wavelet domain. However there are still more speech enhancement methods to be explored in order to work properly with the emotion recognition module.

Emotion model is another important problem in emotion recognition. Emotion class model was used in the most attempts to classify emotions in the early researches (Ayadia *et al.*, 2010; Scherer, 2003; Zeng *et al.*, 2009). Basic human emotions like happiness, sadness, fear, anger, disgust, surprise were detected from speech signals under controlled conditions. However, in the real world applications we need to deal with various emotions which may not be recognized using the pre-trained emotion class models. M. Wöllmer *et al.* suggested abandoning the emotion classes (Wöllmer *et al.*, 2008). They proposed to detect the arousal dimension and the valence dimension instead. In this paper we adopt both the emotion class model and the dimension space model to evaluate our speech emotion recognition system under a noisy environment.

2. The database

A Chinese speech emotion database built in our lab is adopted in this paper, which includes two data sources, the acted speech and the induced speech, as shown in Table 1. The acted speech data contains six emotions, which are fear, surprise, anger, happiness, sadness and neutral (Cai, 2005). Six professional male actors and six professional female actresses were required to simulate the emotions. Subjects who didn't participate in the recording were asked to carry out a listening test to verify the emotional data. A majority vote method was used for selecting the utterances with good quality.

Table 1. Emotion types and data collection method.

Emotions	Collecting method	Number of speakers
fear, surprise, anger, happiness, sadness and neutral	Acted	12
anxiety, hesitation and confidence	Induced	1

The induced speech data contains three emotions, anxiety, hesitation and confidence. These emotions were induced in a cognitive task (Zou, 2011). One male subject was required to work on a series of math calculations and report the answers orally. Negative emotions are not easy to induce in a lab environment, and the subject is generally more cooperative to express his or her positive emotions, like happiness, confidence, etc. Therefore noise stimulations and sleep deprivation were used for inducing the negative emotions (anxiety and hesitation). The subject was required to wear a headset and heavy noise recorded from construction sites and other real world environment is played for inducing the negative emotions. Sleep deprivation is a common method in emotion eliciting and cognitive related experiments, which was also used in our experiment. The subject was required to stay up in a separate room for 36 hours. After the recording, a listening test was carried out to verify the emotional data.

3. Speech enhancement

In real world applications, such as mobile phones, call-centers and interactive toys, speech signals are often corrupted by acoustic background noise. In these applications, speech enhancement is a necessary module for the emotion recognition system. In this section we present a basic spectral subtraction method and an advanced method based on masking properties.

3.1. Speech enhancement based on spectral subtraction

Spectral subtraction is a widely used speech enhancement algorithm first proposed by Boll (1979). Let $x(t)$ be the clean speech signal, $n(t)$ be the noise

signal following a zero-mean Gaussian distribution, and $y(t)$ be the speech signal with noise:

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

where $Y(\omega)$, $X(\omega)$, and $N(\omega)$ are the Fourier transformations of $y(t)$, $x(t)$, and $n(t)$, thus we have:

$$Y(\omega) = X(\omega) + N(\omega) \tag{2}$$

and the power spectral density:

$$\begin{aligned} |Y(\omega)|^2 \; &= \; |X(\omega)|^2 + |N(\omega)|^2 \\ &+X^*(\omega)N(\omega) + X(\omega)N^*(\omega). \end{aligned} \tag{3}$$

Suppose the speech signal and noise signal are independent, we have:

$$|Y(\omega)|^2 = |X(\omega)|^2 + |N(\omega)|^2. \tag{4}$$

Let $P_y(\omega)$, $P_x(\omega)$, and $P_n(\omega)$ be the power spectral density of $y(t)$, $x(t)$ and $n(t)$:

$$P_y(\omega) = P_x(\omega) + P_n(\omega). \tag{5}$$

The estimation of noise power spectral $P_n(\omega)$ is achieved from the silent duration:

$$P_x(\omega) = P_y(\omega) - P_n(\omega). \tag{6}$$

To ensure the non-negativity, when $P_y(\omega) < P_n(\omega)$, let $P_x(\omega) = 0$:

$$P_x(\omega) = \begin{cases} P_y(\omega) - P_n(\omega) & P_y(\omega) \geq P_n(\omega), \\ 0 & P_y(\omega) < P_n(\omega). \end{cases} \tag{7}$$

In the spectral subtraction method, the phase information for IFFT to recover speech signal in time domain is directly obtained from the original speech signal with noise, since human listening perception is not sensitive to phase changes.

3.2. Speech enhancement based on masking properties

The spectral subtraction method is a low computational complexity algorithm, and it may effectively improve the signal-to-noise ratio (SNR). However, the speech signal after spectral subtraction enhancement usually contains musical noise, which may affect the speech quality. Therefore we adopt a more sophisticated speech enhancement algorithm proposed by Chen *et al.* (2007), which is based on the masking properties and short-time spectral amplitude estimation. Masking properties of human auditory system were first introduced by Johnston (1988) and later used in the speech enhancement by Tsoukalas *et al.* (1997) and Virag (1999). Generally speaking the speech signal is the stronger signal than the background noise is the weaker signal. The frequency domain masking can be modeled by a noise masking threshold, below which all components are inaudible. Therefore when the residual noise after speech enhancement is below the noise masking threshold, it cannot be perceived by human auditory system.

The enhanced speech signal should satisfy:

$$\widehat{X}(m,k) = \arg\min_{\widehat{X}} E\left\{ d\left| X(m,k), \widehat{X(m,k)} \right| \, |Y^{m'} \right\}, \tag{8}$$

where m stands for the frame index, k stands for the discrete frequency and $Y^{m'}$ is the Fourier transform of the m'-th frame of the speech signal, and $d\left|X(m,k), \widehat{X(m,k)}\right|$ is the distance measurement between the original speech signal $X(m,k)$ and the enhanced speech signal $\widehat{X}(m,k)$.

Let G_{SP} denotes a transfer function, we have:

$$\begin{aligned} \widehat{X}_m \; &= G_{SP}(\xi_{m|m'}, \gamma_m) Y_m \\ &= \sqrt{\frac{\xi_{m|m'}}{1+\xi_{m|m'}} \left(\frac{1}{\gamma_m} + \frac{\xi_{m|m'}}{1+\xi_{m|m'}} \right)} Y_m, \end{aligned} \tag{9}$$

where $\xi_{m|m'}$ is the priori SNR and γ_m is the posterior SNR, details can be found in (Cohen, 2005).

We propose a parameterized spectral estimation of the speech signal in the following form (Chen *et al.*, 2007):

$$\widehat{X}_m = \sqrt{\frac{\xi_{m|m-1}}{a*} \left(1 + \frac{\xi_{m|m-1}\gamma_m}{a*} \right)} Y_m, \tag{10}$$

where $a* = \alpha(m,k) + \xi_{m|m-1}$.

Let $T(m,k)$ denotes the masking threshold, considering the masking property we have (Virag, 1999):

$$E\left\{ \left| X^2(m,k) - \widehat{X}^2(m,k) \right| \right\} \leq T(m,k). \tag{11}$$

Let $M = \dfrac{\xi_{m|m-1}}{\alpha(m,k) + \xi_{m|m-1}}$, σ_s^2 denotes the speech signal power, and σ_n^2 denotes the noise power. Subject (10) to (11). Notice $E\left\{X^2(m,k)\right\} = \sigma_s^2$ and $E\left\{N^2(m,k)\right\} = \sigma_n^2$, we have:

$$\begin{aligned} \sigma_s^2 - T(m,k) \; &\leq \; M(1+M\gamma_m)(\sigma_s^2 + \sigma_n^2) \\ &\leq \sigma_s^2 + T(m,k). \end{aligned} \tag{12}$$

When the speech signal power is below the masking threshold ($\sigma_s^2 - T(m,k) \leq 0$) let $\alpha(m,k) = 1$. Otherwise we have:

$$\begin{aligned} \frac{2\gamma_m \xi_{m|m-1}}{-1 + \sqrt{4C\gamma_m}} \; - \; \xi_{m|m-1} \leq \alpha(m,k) \\ \leq \frac{2\gamma_m \xi_{m|m-1}}{-1 + \sqrt{4B\gamma_m}} - \xi_{m|m-1}, \end{aligned} \tag{13}$$

where

$$B = \frac{\sigma_s^2 - T(m,k)}{\sigma_s^2 + \sigma_n^2}$$

and

$$C = \frac{\sigma_s^2 + T(m,k)}{\sigma_s^2 + \sigma_n^2}.$$

Therefore the parameter $\alpha(m,k)$ can be determined by the auditory masking threshold, the estimated speech power spectral and the noise power spectral (Chen *et al.*, 2007). It may dynamically adjust the transfer function, and an optimized tradeoff among the reduction of noise, the speech distortion and the level of musical residual noise may be achieved. This speech enhancement method based on masking properties of human auditory system may be suitable for emotional speech, in the experimental section we will carry out a set of emotion recognition tests using the two different speech enhancement methods for comparison.

4. Recognition methodology

4.1. Emotional feature extraction

Various acoustic features have been studied for speech emotion recognition. The prosodic features may be related to arousal dimension and the voice quality features may be related to valence dimension (Gobl, Chasaide, 2003; Johnstone *et al.*, 2005). Both temporal features and static features can be used for speech emotion recognition. Typically the temporal features may be used with Hidden Markov Model (HMM) while the static features may be used with GMM. Since the static features are considered less dependent on phoneme information, we adopt the static features including maximum, minimum, mean, standard deviation and range for the construction of the emotional features. A total of 372 features are generated, as shown in Table 2. Basic Linear Discriminant Analysis (LDA) is then adopted for feature dimension reduction.

When searching for the emotional features, it may be better to exclude the influence of the text variations. In a good emotion data set, the text should be well designed so that its proportion among various emotion classes is balanced. However the uncontrolled naturalistic data is often unbalanced in text, consequently the selected emotional features may be influenced by the phonetic information.

The utterances are categorized according to their time durations, since time duration is an important character of emotional expression in speech. For a balanced data set, we compared the statistics on the time duration and selected the training samples to reduce the variations among different emotion classes, as shown in Table 3.

Table 2. Feature extraction ("dev" is short for deviation; "MFCC" stands for Mel-Frequency Cepstral Coefficients and "BBE" stands for Bark Band Energy).

Feature Index	Feature Description
1–10	max, min, mean, std, range of pitch and dev pitch
10–11	Jitter, Shimmer
12–52	max, min, mean, std, range of F1 to F4 and dev of F1 to F4
52–62	max, min, mean, std, range of intensity and dev intensity
62–192	max, min, mean, std, range of MFCC1 to MFCC13 and dev of MFCC1 to MFCC13
192–372	max, min, mean, std, range of BBE1 to BBE18 and dev of BBE1 to BBE18

Table 3. The text length balance of each emotion class (number of characters in each utterance).

Emotion class	Max of duration	Min of duration	Mean of duration
Happiness	13	2	6.3
Sadness	11	2	7.1
Anger	13	2	6.8
Surprise	12	2	6.9
Fear	14	2	7.0
Neutral	13	2	6.8
Anxiety	13	2	6.8
Hesitation	14	2	7.0
Confidence	11	2	6.7

4.2. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is successfully applied to speaker and language identification. And recently GMM has shown its promising performance in speech emotion recognition (Kockmann *et al.*, 2011). It can be seen as a HMM of one state. The probability density function of an m-order GMM is consist of weighted summation of m Gaussian probability density function, which can be expressed as (Reynolds *et al.* 1995; Reynolds, 1997):

$$p(\mathbf{S}\,|\boldsymbol{\lambda}) = \sum_{q=1}^{Q} a_q b_q(\mathbf{S}), \tag{14}$$

where $\mathbf{S}$ is the feature vector of the input sample, $\boldsymbol{\lambda}$ denotes the parameters of GMM, q is the index of the Gaussian mixtures, Q stands for the mixture number, a_q is the mixture weight and b_q stands for the Gaussian distribution function.

Bayes method is used in the identification of emotion. Among the N unknown models, the emotion class whose corresponding model gets the maximum likelihood probability is the target emotion:

$$j^* = \arg\max_{1\le j\le N} \log P(\mathbf{S}|\boldsymbol{\lambda}_j), \tag{15}$$

where j^* denotes the index of the target emotion.

We adopt the EM (expectation-maximization) algorithm in GMM parameter estimation. GMM parameters can be presented as:

$$\boldsymbol{\lambda} = \{a_q, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\}, \qquad q = 1, 2, ..., Q, \tag{16}$$

where Q is the mixture number, $\boldsymbol{\mu}_q$ is the mean of each Gaussian distribution, and $\boldsymbol{\Sigma}_q$ is the covariance matrix.

K-mean clustering is used for initialization where K equals to the GMM mixture number, and EM procedure is used for parameter estimation. The EM equations for training a GMM can be found in (Reynolds *et al.* 1995; Reynolds, 1997).

5. Experimental results

Based on two types of emotion model theories (the basic emotion theory and the dimension space theory), we adopted two types of classification tasks for the evaluation of the classification system: i) the emotion class classification, and ii) the arousal-valence dimension region classification.

In the training and testing stages, 400 utterances of each emotion class were used for training and 100 utterances of each emotion class were used for testing, including nine emotion types. For the dimension region classification, both the arousal dimension and the valence dimension were classified into positive and negative. We took four training sets for training the positive and the negative model in arousal dimension and valence dimension respectively, each training set contained 800 samples. We also took four testing set, each contained 200 samples. The arousal classifier and the valence classifier were trained separately. Both classifiers classify the input sample into positive dimension or negative dimension.

5.1. Parameter settings

To study the noise influence on speech emotion recognition, we adopted the clean condition training. The training dataset contains clean speech while the noise levels (SNR) of the testing dataset are different. The clean speech was mixed with AWGN at various signal-to-noise ratios (15 dB, 10 dB and 5 dB). Before testing, we apply two types of speech enhancement algorithms to the noisy speech.

The sampling rate was 11.025 kHz, the digitalizing bit was 16 bit. Hamming window was used on the speech data, the frame length was 256, with an overlap of 128.

The GMM mixture number was set to 32 for emotion classification, and 64 for dimension region classification. The maximum iteration in the EM algorithm was set to 50. K-mean cluster algorithm was used for the initialization in the GMM parameter estimation, and k equals to GMM mixture number.

5.2. Classification results

The classification rates under various noise levels are shown in Fig. 1 through Fig. 2. Two speech enhancement algorithms were evaluated separately on both emotion-class classification task and arousal-valence dimension classification task. As the SNR drops from 15 dB to 5 dB, the classification rates decrease subsequently through all emotion classes and both valence and arousal dimension.

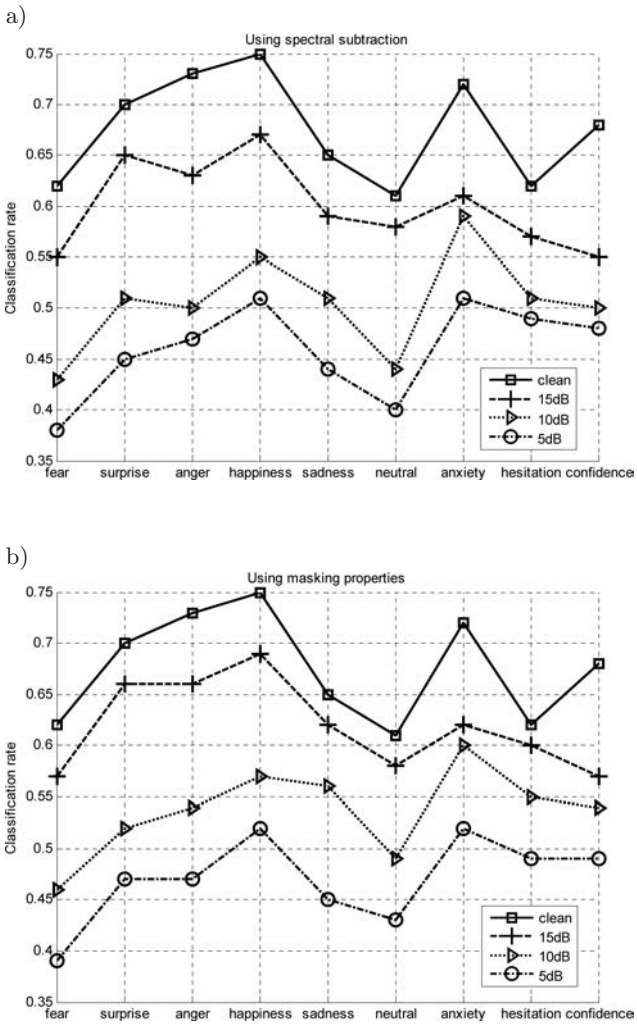
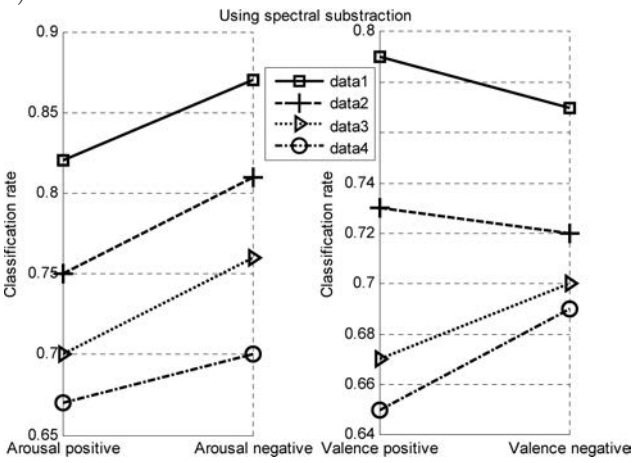


Fig. 1. Emotion-class classification rate under various noise levels: a) using speech enhancement algorithm based on spectral subtraction; b) using speech enhancement algorithm based on masking properties.

a)



b)

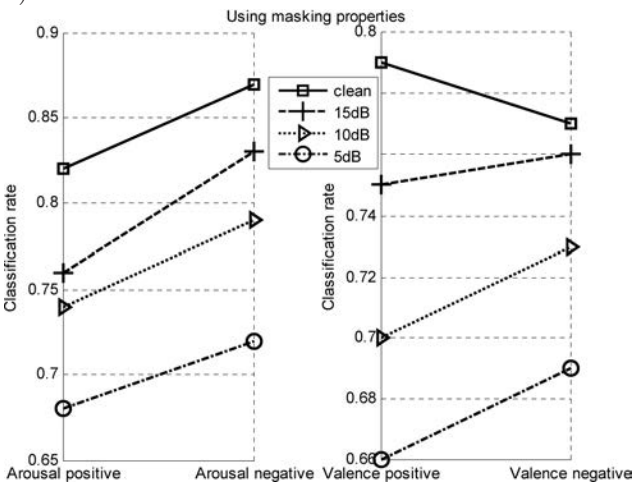


Fig. 2. Arousal-Valence region classification rate under various noise levels: a) speech enhancement based on spectral subtraction; b) speech enhancement based on masking properties.

The advantage of the second algorithm (masking properties based speech enhancement) is obvious, it constantly over-perform the first algorithm (basic spectral subtraction method). The second algorithm takes advantage of human auditory properties, which provides a better tradeoff between the amount of noise reduction and the emotion feature distortion. By an automatic adaptation based on human perception criterions it may be more suitable for emotion recognition tasks.

In the emotion class classification experiment, when tested with clean speech "happiness" is the highest recognized emotion type. However when the SNR drops to 10 dB, "anxiety" becomes better detected than other emotions. As shown in Fig. 1. This accuracy shift is caused by noise influence, and it may be classifier dependent. Similar results are observed in the dimension classification experiment, the classification rate of negative emotions becomes higher than the classification rate of positive emotions as the noise level increase, as shown in Fig. 2. Since "anxiety" and other negative emotions are most related to valence dimension, the voice quality features may be distorted by the noise, and caused the miss-classification of the positive emotions.

6. Conclusions

In this paper we evaluated the speech emotion recognition system from two points of views, the emotion class view and the arousal-valence dimensional view. From the former one we built GMM based models for each individual emotion class, from the later one we classified the positive and the negative regions of the arousal-valence space also using GMM based models.

The noise influence is an important factor to many of the automatic speech recognition systems, especially when it comes to real world applications. In our study we investigated the speech emotion recognition problem under various white noise conditions. To deal with the AWGN we applied two existing speech enhancement algorithms, the spectral subtraction based algorithm and the algorithm based on masking properties.

The experimental results show that the second algorithm is better than the first algorithm when applied to the speech emotion recognition problem. Speech enhancement is a necessary procedure for speech emotion recognition systems working in a noisy field environment. When increasing the noise level, the overall classification rate dropped, and the positive emotions were more likely to be miss-classified as negative emotions (in valence dimension).

In our study on the speech enhancement, we only compared two existing algorithms, and considered only under AWGN condition. Verifying our emotion recognition system on different databases and various noise types other than white noise may be an interesting future topic. In the feature selection stage we used the same feature set constantly, it is also interesting to select noise robust features for future practical systems.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions on improving this paper. This work is supported by the Natural Science Foundation of China (No. 60872073; No. 60975017; No. 51075068), the Doctoral Fund of Ministry of Education of China (No. 20110092130004), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 10KJB510005).

References

1. Ang J., Dhillon R., Krupski A., Shriberg E., Stolcke A. (2002), *Prosody-based automatic detection of annoyance and frustration in human-computer dialog*, 7th International Conference on Spoken Language Processing, pp. 2037–2040, Denver, Colorado, USA.
2. Ayadia M.E., Kamelb M.S., Karray F. (2010), *Survey on speech emotion recognition: Features, classification schemes, and databases*, Pattern Recognition, **44**, 3, 572–587.
3. Boll S. (1979), *Suppression of acoustic noise in speech using spectral subtraction*, IEEE Transactions on Acoustics, Speech and Signal Processing, **27**, 2, 113–120.
4. Cai L. (2005), *Speech emotion analysis and recognition based on data fusion*, Master Thesis, Department of Radio Engineering, Southeast University, China.
5. Chen G., Zhao L., Zhou C. (2007), *Speech Enhancement Based on Masking Properties and Short-Time Spectral Amplitude Estimation*, Journal of Electronics & Information Technology, **29**, 4, 863–866.
6. Clavel C., Vasilescu I., Devillers L., Richard G., Ehrette T. (2008), *Fear-type emotion recognition for future audio-based surveillance systems*, Speech Communication, **50**, 487–503.
7. Cohen I. (2005), *Relaxed statistical model for speech enhancement and a priori SNR estimation*, IEEE Transactions on Speech and Audio Processing, **13**, 5, 870–881.
8. Gobl C., Chasaide A.N. (2003), *The role of voice quality in communicating emotion, mood and attitude*, Speech Communication, **40**, 189–212.
9. Huang C., Jin Y., Zhao Y., Yu Y., Zhao L. (2009), *Speech emotion recognition based on re-composition of two-class classifiers*, 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–3, Amsterdam, Netherlands.
10. Huang C., Zhao Y., Jin Y., Yu Y., Zhao L. (2011), *A Study on Feature Analysis and Recognition for Practical Speech Emotion*, Journal of Electronics & Information Technology, **33**, 1, 112–116.
11. Johnston J.D. (1988), *Transform coding of audio signals using perceptual noise criteria*, IEEE Journal on Selected Areas in Communications, **6**, 2, 314–323.
12. Johnstone T., van Reekum C.M., Hird K., Kirsner K., Scherer K.R. (2005), *Affective speech elicited with a computer game*, Emotion, **5**, 4, 513–518.
13. Jones M.C., Jonson I.M. (2005), *Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses*, Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, Canberra, Australia.
14. Kockmann M., Burget L., Cernocky J.H. (2011), *Application of speaker- and language identification state-of-the-art techniques for emotion recognition*, Speech Communication, **53**, 1172–1185.
15. Neiberg D., Elenius K., Laskowski K. (2006), *Emotion recognition in spontaneous speech using GMMs*, International Conference on Spoken Language Process, pp. 809–902, Pittsburgh, Pennsylvania, USA.
16. Reynolds D.A., Rose R.C. (1995), *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE Transactions on Speech Audio Process, **3**, 72–83.
17. Reynolds D.A. (1997), *Comparison of background normalization methods for text-independent speaker verification*, European Conference on Speech Communication and Technology, pp. 963–966, Rhodes, Greece.
18. Schuller B., Arsic D., Wallhoff F., Rigoll G. (2006), *Emotion recognition in the noise applying large acoustic feature sets*, 3rd International Conference on Speech Prosody, Dresden, Germany.
19. Scherer K.R. (2003), *Vocal communication of emotion: A review of research paradigms*, Speech Communication **40**, 227–256.
20. Tawari A., Trivedi M. (2010), *Speech emotion analysis in noisy real-world environment*, International Conference on Pattern Recognition, pp. 4605–4609, Istanbul, Turkey.
21. Truong K. (2009), *How does real affect affect affect recognition in speech?* Ph.D. Thesis, Department of Electrical Engineering, Mathematics and Computer Science, University of Twente.
22. Tsoukalas D.E., Mourjopoulos J.N., Kokkinakis G. (1997), *Speech enhancement based on audible noise suppression*, IEEE Transactions on Speech and Audio Processing, **5**, 6, 497–514.
23. Varga A., Steeneken H.J.M. (1993), *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems*, Speech Communication, **12**, 3, 247–251.
24. Virag N. (1999), *Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System*, IEEE Transactions on Speech and Audio Processing, **7**, 2, 126-137.
25. Wöllmer M., Eyben F., Reiter S., Schuller B., Cox C., Douglas-Cowie E., Cowie R. (2008), *Abandoning emotion classes – Towards continuous emotion recognition with modeling of long-range dependencies*, 9th Annual Conference of the International Speech Communication Association, pp. 597–601, Brisbane, Australia.
26. Zeng Z., Pantic M., Roisman G.I., Huang T. (2009), *A survey of affect recognition methods: audio, visual and spontaneous expressions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **31**, 1, 39–58.
27. Zou C., Huang C., Han D., Zhao L. (2011), *Detecting practical speech emotion in a cognitive task*, 20th International Conference on Computer Communications and Networks, Maui, Hawaii, USA.