MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition

Tian Wu, Yingchun Yang, Zhaohui Wu and Dongdong Li

CCNT Lab, College of Computer Science and Technology
Zhejiang University, Hangzhou, P.R.CHINA
{wutian, yyc, wzh, lidd}@zju.edu.cn

Abstract

In this paper, a large emotional speech database MASC (Mandarin Affective Speech Corpus) is introduced. The database contains recordings of 68 native speakers (23 female and 45 male) and five kinds of emotional states: neutral, anger, elation, panic and sadness. Each speaker pronounces 5 phrases, 10 sentences for three times for each emotional states and 2 paragraphs only for neutral. These materials covers all the phonemes in Chinese. This corpus is constructed for prosodic and linguistic investigation of emotion expression in Mandarin. It can also be used for recognition of affectively stressed speakers. Furthermore, prosodic feature analysis and speaker recognition baseline experiment are performed on this database.

1. Introduction

Ways of expressing emotions by human and the effect on speech of emotional state changes to speakers have intrigued researchers for a long time. Currently, psychologists have done many experiments and raised a variety of theories [1]. However, collecting large scale affective speech corpus is a very difficult task. Few works are done here. Emotional Prosody Speech and Transcripts (EPST) is an emotional speech database provided by Linguistic Data Consortium (LDC) [2]. This corpus covers 14 emotional states based on Banse & Scherer's selection criteria [3] and is designed to support research in emotional prosody. For speaker-independent emotion recognition, Sony entertainment AIBO is a target scenario to which emotional databases are recorded. These databases simulate different possible situations and comprise all the desired emotions [4]. RUSLANA is a database of emotional utterances and recorded in Russian, aiming for linguistic and speech processing research on communicative and emotive-attitudinal aspects of spoken language [5]. Sixty-one native speakers of standard Russian were recorded for this database.

As mentioned above, academic and applied research activities are stimulated in the area of emotion recognition and analysis. By far, there is still not a large speech database used for affectively speaker recognition. Our motivation of creating an emotional speech corpus arises from the mismatch in automatic speaker recognition. Current speaker verification and identification systems are limited by the effect on speech of transient state changes to speakers. The variability of intra-speaker can cause unacceptably high error rates [6]. Furthermore, in the emotional speech investigation area, the focus has so far been on some major languages as English, German, French and Russian. Very little is known about the vocal correlates of emotion in continuous spoken Mandarin.

Our goal is to provide a large corpora in Chinese designed for emotional speech analysis and affectively speaker recognition purpose, named as MASC@CCNT (Mandarin Affective Speech Corpus at CCNT Lab). Compared with other emotional speech database, MASC concentrates on showing both the characteristics of different emotional states and the intra-speaker variabilities caused by state changes of speakers. In particular, these records are spoken in Mandarin which is a fairly fresh area of emotional speech studies.

The remainder of this paper is organized as follows. The design and collection of MASC is introduced in section 2, including the description of emotional states, the speakers, and the speech materials. In section 3 and 4, the prosodic analysis and baseline of speaker recognition is performed on this corpus. Finally, some discussions and conclusions are extracted , together with proposed future work in section 5.

2. Database Generation

The database is created with two major objectives. On one hand, it is used for prosodic and linguistic investigation of emotion expression in Mandarin. On the other hand, it supplies a training set as well as a test data set for speaker recognition system affected by emotional factors.

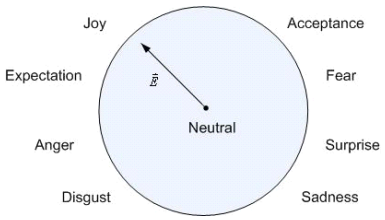

Figure 1: *Schematic diagram of speech production.*

2.1. Emotional States

The selection of emotional states is expected to put the speech on the emotion wheel which has been derived from the Plutchik's work [7]. As Fig. 1 shows, emotion wheel is a model to describe the activation-evaluation-power space of emotion. According to this emotion model, emotional states distribute on a circle which is named Emotion Wheel. The center of this circle stands for the natural origin, a state which gathers all kinds of emotional factors. However, the effects from these emotional factors are so weak that they cannot emerge at the origin. Each emotional state is defined with a unique planar vector $\vec{E}$ that has two parameters, emotional intensity and emotional orientation. Thereinto, emotional intensity indicates the range of $\vec{E}$

and emotional orientation renders the angle of $\vec{E}$. In terms of the emotion wheel, four emotion are selected: anger, elation, panic and sadness, whose descriptions are consulted by Banse & Scherer [3].

- Neutral - Simple statements without any emotion.
- Anger - A strong feeling of displeasure or hostility.
- Elation - Be glad or happy because of praise.
- Panic - A sudden, overpowering terror, often affecting many people at once.
- Sadness - Affected or characterized by sorrow or unhappiness.

2.2. Material

A text of five phrases, fifteen sentences and two paragraphs has been designed to generate the emotional speech corpus, including all the phonemes in Mandarin. Each subject reads these materials of different types portraying the five emotional states: neutral (unemotional), anger, elation, panic and sadness. Altogether this database contains 25,636 utterances (5,100 phrases, 20,400 sentences and 136 paragraphs).

- 5 phrases - "yes", "no" and three nouns as "apple", "train", "tennis ball". In Chinese, these words contain many different basic vowels and consonants.
- 20 sentences - These sentences include all the phonemes and most common consonant clusters in Mandarin. The types of sentences are: simple statements, a declarative sentence with an enumeration, general questions (yes/no question), alternative questions, imperative sentences, exclamatory sentences, special questions (wh-questions).
- 2 paragraphs - They are two readings selected from a famous Chinese novel, stating a normal fact.

These materials cover all the vowels in Mandarin: /a/, /o/, /e/, /i/, /u/, /ü/, /ia/, /ua/, /uo/, /ie/, /eü/, /ai/, /uai/, /ei/, /uei/, /ao/, /iao/, /ou/, /iou/, /an/, /ian/, /uan/, /üan/, /en/, /in/, /uen/, /ün/, /ang/, /iang/, /uang/, /eng/, /ing/, /ueng/, /ong/, /iong/, and all the consonants: /b/, /p/, /m/, /f/, /d/, /t/, /n/, /l/, /g/, /k/, /h/, /j/, /q/, /x/, /zh/, /ch/, /sh/, /r/, /z/, /c/, /s/.

Here we list the detailed content of all materials used in this database. The utterances are marked with Chinese pronunciation and given below along with English translation in brackets.

- Phrases:

  101 是的- *shì de* (Yes.)

  102 不是- *bù shì* (No.)

  103 苹果- *píng guǒ* (Apple.)

  104 火车- *huǒ chē* (Train.)

  105 网球- *wǎng qiú* (Tennis ball.)

- Sentences:

  201 你是个好人- *nǐ shì gè hǎo rén.* (You are a nice person.)

  202 我们那边有网球运动场、餐馆、酒吧和一个面包店- *wǒ mén nà biān yǒu wǎng qiú yùn dòng chǎng cān guǎn jiǔ bā hé yī gè miàn bāo diàn.* (A tennis ball playground, a restaurant, a bar and a baker's shop are located in our area.)

  203 你今天去医院看过病了吗？- *nǐ jīn tiān qù yī yuàn kàn guò bìng le ma* (have you seen the doctor today?)

  204 这个湖是人工的还是自然形成的？- *zhè gè hú shì rén gōng dē hái shì zì rán xíng chéng de* (Is this lake natural or artificial?)

  205 你去把空调打开- *nǐ qù bǎ kōng tiáo dǎ kāi* (Turn on the air-condition.)

  206 他是多么慷慨啊！- *tā shì duō me kāng kǎi a* (How generous he is!)

  207 为什么你不给他看看那本小说呢？- *wèi shén mē nì bù gěi tā kàn kan nà běn xiǎo shuō ne* (Why don't you show him that novel?)

  208 我应该在信里写一些什么呢？- *wǒ yīng gāi zài xìn lǐ xiě yī xiē shén mē ne* (What should I write on the letter?)

  209 我们哪天去欧洲的温莎城堡玩？- *wǒ mén nǎ tiān qù ōu zhōu dē wēn shā chéng bǎo wán* (When shall we go to Windsor Castle in Europe for a trip?)

  210 老翁挖了一个大约五平米的池塘养鱼- *lǎo wēng wā lē yī gè dà yuē wǔ píng mǐ de chí táng yǎng yú* (The old man has dug a fish pond, which is about five square meters.)

  211 小明，你陪外婆去补牙齿吧- *xiǎo míng nǐ péi wài pó qù bǔ yá chǐ ba* (Ming, accompany your grandmother to the dentist's.)

  212 今天晚上会下雨- *jīn tiān wǎn shàng huì xià yǔ* (It will rain tonight.)

  213 巷口来了好多警察！- *xiàng kǒu lái le hěn duō jǐng chá* (A group of policemen appear at the block!)

  214 考试结束时间快到了- *kǎo shì jié shù shí jiān kuài dào l* (Time is over for the examination.)

  215 我们室友总是把寝室弄得很脏- *wǒ mén shì yǒu zǒng shì bǎ qǐn shì nòng dé hěn zāng* (Our roommates always make the dormitory very dirty.)

  216 你抄我的物理作业- *nǐ chāo wǒ de wù lǐ zuò yè* (You copied my physics homework.)

  217 我最要好的朋友要移民去欧洲了- *wǒ zuì yào hǎo de péng yǒu yào yí mín qù ōu zhōu le* (My best friend will emigrate to Europe.)

  218 他们家的小狗死掉了- *tā mén jiā de xiǎo gǒu sǐ diào le* (Their puppy is dead.)

  219 明天要去富春江漂流了- *míng tiān yào qù fù chūn jiāng piāo liú le* (We will drift on Fuchunjiang River tomorrow.)

  220 新桥门的水果摊又开了- *xīn qiáo mén de shuǐ guǒ tān yòu kāi le* (The fruit booth at New Bridge Gate opens again.)

- Paragraphs:

  301 同学们互赠礼物，整理自己的东西；单个照像，集体合影；要好的朋友也纷纷聚在一起照一张留念照。县照像馆干脆专门抽出几个人到中学来为同学们服务。- *tóng xué mén hù zèng lǐ wù zhěng lǐ zì jǐ de dōng xi dān gè zhào xiàng jí tì hé yǐng yào hǎo de péng yǒu yě fēn fēn jù zài yī qǐ zhào yī zhāng liú niàn zhào xiàn zhào xiàng guǎn gān cuì zhuān mén chōu chū jǐ gè rén dào zhōng xué lái wèi tóng xué mēn fú wù* (The students interchanged gifts and packed their things . To commemorate their friendship, they took pictures together. Some bosom friends got together to take a souvenir photo. The photo studio even assigned several photographers to come to the school and serve the students.)

302 他便自然地加入了这个杂乱的军营。找了一块空地方把行李搁下。周围没有人注意他参加到他们的队伍中来。和这些同志比起来，他除了皮肤还不算粗糙外，穿戴和行李没有什么异样的。- *tā biàn zì rán de jiā rù le zhè gè zá luàn de jūn yíng zhǎo le yī kuài kòng dì fāng bǎ xíng lǐ gē xià zhōu wéi méi yǒu rén zhù yì tā cān jiā dào tā mén de duì wǔ zhōng lái hé zhè xié tóng zhì bǐ qǐ lái tā chú le pí fū hāi bù suàn cū cāo wài chuān dài hé xíng li méi yǒu shén me yì yàng de* (He joins the mussy barracks spontaneously, putting his baggage at a space. Nobody notices his enroll to their troop. Compared with his colleagues, he is difference in neither apparel nor luggage except that his skin is not as coarse as veterans.)

2.3. Speakers

Sixty-eight native Chinese speakers were recorded for this corpus: 23 female and 45 male. They were members in CCNT lab and students in Zhejiang University. All of them have lived in Mainland China since their birth and majority were trained to speak in standard Mandarin from early childhood. The average age of speakers is 21.7 years, with the range from 20 to 30 years. The speaker's distribution by age and gender is listed in Table 2 below.

Table 1: *Speakers' distribution by age and gender.*

Age	20	21	22	23	24	25	28	30
Female	0	3	4	10	4	1	1	0
Male	1	7	13	9	8	4	2	1
Total	1	10	17	19	12	5	3	1

2.4. Method

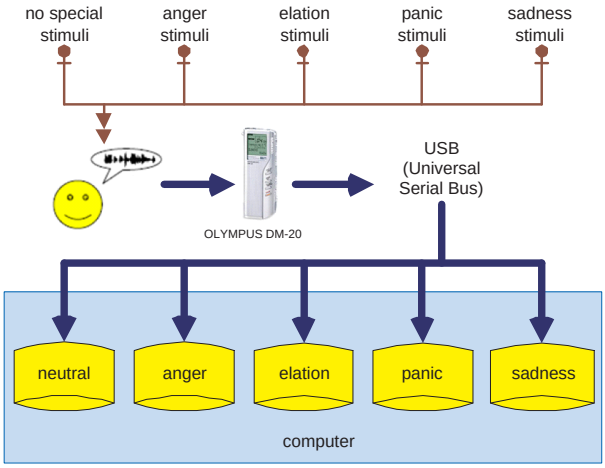

Figure 2: *The recording process for MASC.*

The speech database was recorded by eliciting speakers to express different types of emotional states with certain stimuli. The elicitation is designed as a reading describing a specific scenario such as someone's mistake for eliciting anger, a pleasant trip for elation, a hurry-up scene for panic and a lovely dog's death for sadness. For each emotional state, with the help of the readings, the speaker acted to say something in that mood. All the data were recorded on an OLYMPUS DM-20 digital voice recorder at 22050Hz sampling rate, in a quiet office without disturbance. Afterwards, the recorded voice files were transferred to a personal computer by USB (Universal Serial Bus). The obtained recordings were converted into monophonic Windows PCM format at 8 kHz sampling frequency and 16 bits resolution. Fig. 2 shows the recording process for MASC.

The database has the following structure. The top folder contains a database description file, a speaker information file, a content detail file and 68 folders each of which is related with one single speaker. The description file presents the basic information about the database, including a brief introduction, the explanation of the emotion categories, the refered speakers and the description of data. The speaker information file contains serial number, gender, age and the comments on speaker's performance. The content detail file gives a list of the reading material and its corresponding clip ID in this database. Each speaker folder embodies 5 emotion folders, ANGER, ELATION, NEUTRAL, PANIC and SADNESS. For each emotional state, there are 5 phrases and 20 utterances that are recorded for three times respectively. Specially, 2 paragraphs are expressed only in neutral. All the audio clips are stored as .wav files. A file name has the following template:

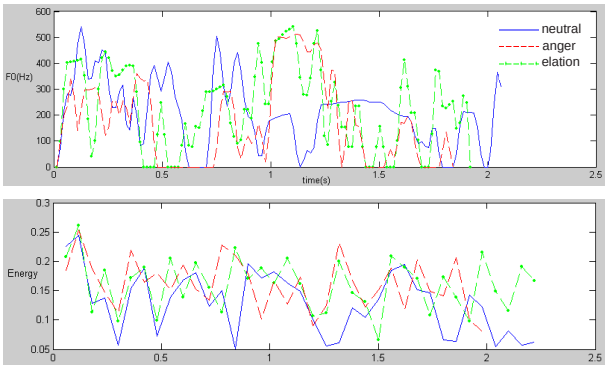

Figure 3: *Prosody of emotional utterances (neutral, anger, and elation.*

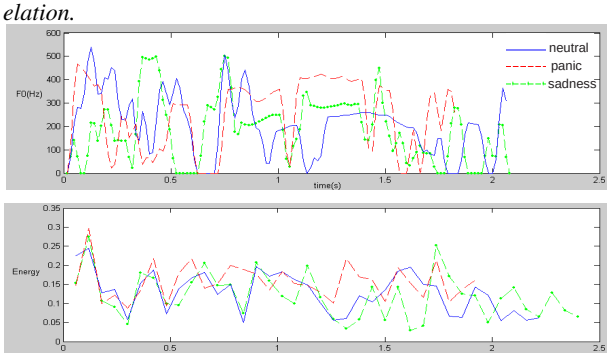

Figure 4: *Prosody of emotional utterances (neutral, panic, and sadness.*

CNNT.wav

where 'C' is one number from the set $1, 2, 3$, where '1' which means 'phrase', '2' which means 'utterance' or '3' which means 'paragraph' and indicates the type of this audio clip; 'NN' is a two digit number that represents the reading content. Because we asked speakers to repeat each utterance for three

times, we use 'T' to distinguish them. For example, '2043.wav' is the third reading (T = 3) of the fourth (NN = 04) sentence (C = 2).

3. Feature Analysis

Two feature analysis methods are employed on our database. One of them shows the contours for prosodic features and the other gives a statistical acoustical analysis.

3.1. Prosody Notes

Figures 3 and 4 display prosodic parameters of a sentence produced by one female speaker. This sentence has a serial number as '2043' in the audio clip list. It is an alternative question having the same meaning as 'Is this lake natural or artifical?'. Figure 3 presents time series for fundamental frequency contours and energy in the neutral sentence, and in the ones that express anger and elation. Figure 4 shows the same data for the states of neutral, panic and sadness.

Many differences in prosodic parameters between the contours can be observed from Figure 3 and 4. Firstly, F0 is much higher when the readings are expressed by a drastic emotional state such as anger and elation. For panic, the average pitch value is higher than the value in neutral and sadness but lower than in anger and elation. All these variants of sentence 2043 have the logical stress on the first phonetic word, after which they follow totally different trends. Secondly, for pitch contour can show the phonemic duration, we observe that the duration varies markedly between different emotional states. Thirdly, the pictures of energy track dynamics illuminate that energy transformation exists when speaking state changes.

3.2. Acoustic Analysis

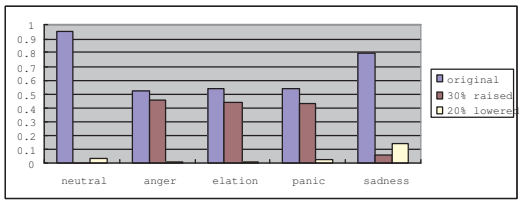

Figure 5: *Average judgments for mean pitch*

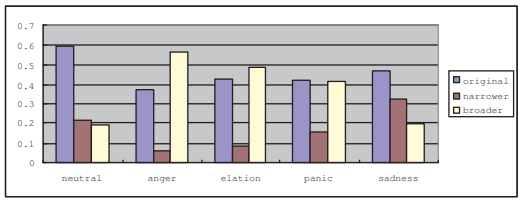

Figure 6: *Average judgments for pitch range*

From earlier studies [3] [10] [11], fundamental frequency (F0) can be used in emotional speech analysis. A series of statistical parameters of pitch, mean pitch, pitch range, pitch variance, pitch skewness, pitch expansion, are employed in our study [12]. An experiment is set up comprising these five features, which vary in the following way:

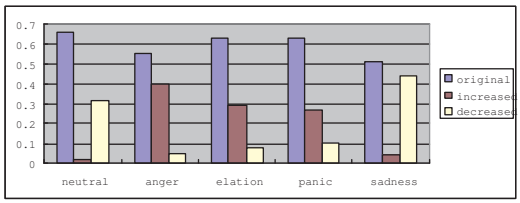

Figure 7: *Average judgments for pitch variance*

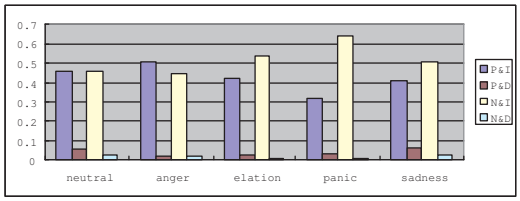

Figure 8: *Average judgments for pitch skewness*

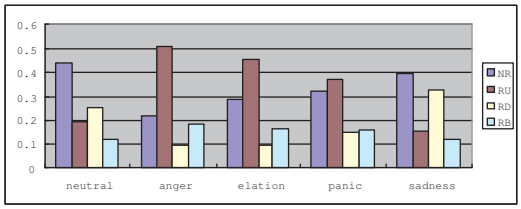

Figure 9: *Average judgments for pitch expansion*

- mean pitch (3 levels): original, 30% raised and 20% lowered. Mean pitch from unemotional speech is chosen to be the original standard.
- pitch range (3 levels): original, 20% narrower and 20% broader. It is aware that a 20% larger does not necessarily result in a 20% increase in variance.
- pitch variance (3 levels): original, 50% increased and 10% decreased. Pitch has wider range of distribution only when the pitch variance is increased.
- pitch skewness (4 types): P&I (Skewness is positive and increased), P&D (Skewness is positive and decreased), N&I (Skewness is negative and increased in absolute value), N&D (Skewness is negative and decreased in absolute value).
- range expansion (4 types): normal range, expansion from the bottom of the range up, expansion from the top of the range down, expansion radiation from the middle of the range outward both directions. [3]

The statistical analysis is executed on MASC corpus in terms of criteria above (Figure 5, 6, 7, 8 and 9). The interpretation of these five emotional states can be concluded as a whole. Sadness gives a similar presentation as neutral. Both of them has a lower mean pitch, a narrower pitch range and a decreased pitch variance. On the other side, anger, elation and panic behave much more drastically.

4. Speaker Recognition Baseline

The baseline strategy employs traditional speaker recognition system on our database. For each speaker, 2 paragraphs in neu-

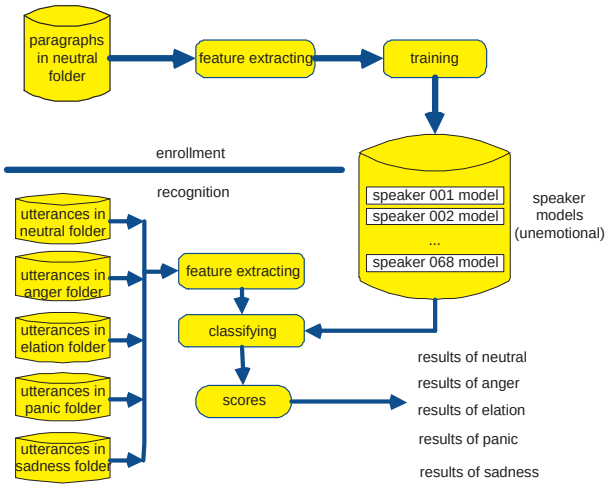

Figure 10: *The block diagram of the ASR system for baseline.*

tral folder are used to train a speaker model while all the sentences in emotion folders are used for testing. The length of the enrollment speech for each speaker is 30 - 40 seconds. For testing, each slice has a 1 - 3 seconds speech and each speaker has 60 test case respectively. The baseline experiment is divided to five parts in terms of the five emotional states. Fig. 10 shows the logical block diagram of the ASR system for this baseline experiment.

The speech signal was segmented into 32 ms with 16 ms overlapped frames, reemphasized and Hamming windowed. 32 MFCCs (Mel Frequency Cepstrum Coefficients) were used to describe the speakers' characters. 32 GMMs (Gaussian Mixture Model) were employed. EER (Equal Error Rate) and IR (Identification Rate) were chosen as the criteria of performance.

Table 2: *Results of baseline.*

	IR (%)	EER (%)
neutral	90.66	10.61
anger	17.28	37.03
elation	16.62	38.30
panic	15.66	38.01
sadness	48.68	23.68

For the baseline, both speaker identification and verification perform very well on neutral. However, the results on four emotional states are not as good as neutral. As mentioned above, the statistical analysis of sadness is more similar as neutral than anger, elation and panic. Therefore, the performance of sadness speech is higher than the other three emotional states. For three drastic states, anger, elation and sadness, their veracities of speaker recognition are extremely low. This result shows that the variability of intra-speaker in these emotional situations can cause unacceptably high error rates.

5. Conclusions and Future Work

The MASC database provides a content labeled, multi-speaker speech corpus for emotion analysis and affectively speaker recognition. By far, it is among the first largest Mandarin database of emotional speech. Unlike the popular databases that are focused on major languages as English, German and Russian, MASC gives a chance to investigate the expressing manners and acoustical features in Mandarin. We also notice that traditional speaker verification and identification systems are limited by their lack of robustness against intra-speaker variability, which the changes of speaking attitude can raise. The MASC corpus can be a useful tool in the study of removing the effects caused by affective speech. It is our hope that the database can serve these research purposes.

As future work, we will carry a series of listening tests out to evaluate this database and to verify that the speech samples contain the intended emotions.

6. Acknowledgements

This work is supported by National Science Fund for Distinguished Young Scholars 60525202, Program for New Century Excellent Talents in University NCET-04-0545 and Key Program of Natural Science Foundation of China 60533040.

7. References

[1] Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck T., Vocal clues in emotion encoding and decoding. Motiv Emotion 1991; 15: 123-148, 1991.

[2] http://www.ldc.upenn.edu/

[3] Banse, R., Scherer, K.R. 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 614-636.

[4] Tato, R., Santos, R., Kompe, R., Pardo, J.M., Emotional Space Improves Emotion Recognition, Proc. ICSLP, 2002.

[5] Makarova, V., Petrushin V.A., RUSLANA: A Database of Russian Emotional Utterances, Proc. ICSLP, 2002.

[6] Scherer, K.R., Johnstone, T., Klasmeyer, G., Banziger, T., Can Automatic Speaker Verification be Improved by Training the Algorithms on Emotional Speech? ICSLP 2000.

[7] Plutchik, R., Emotion: A Psychoevolutionary Synthesis. New York: Harper and Row, 1980.

[8] Suejing, S., Pitch Determination and Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio. IEEE International Conference on Acoustics, Speech, and Signal Processing, May 13-17, 2002

[9] Mustafa, K., Bruce, I.C., Robust Formant Tracking for Continous Speech with Speaker Variability, IEEE Transactions on Speech and Audio Processing, Mar. 2006.

[10] Scherer, K.R., Johnstone, T., Banziger, T.: Verification of emotionally stressed speakers: The problem of individual differences. Proc. of SPECOM98, 1998

[11] Schroder, M., Emotional Speech Synthesis: A Review. EUROSPEECH'01 Volumel, 2001

[12] Wu, T., Yang, Y.C., Wu, Z.H., Improving Speaker Recognition by Training on Emotion-Added Models, Proc. of ACII 2005