An Expressive Mandarin Speech Corpus

Jianhua Tao Jian Yu Yongguo Kang

National Laboratory of Pattern Recognition (NLPR),

Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100080

{jhtao, jyu, ygkang}@nlpr.ia.ac.cn

Abstract

The paper introduces an expressive mandarin speech corpus, which is supported by National Hi-tech program (863) and National Science Funding of China (NSFC), for research into expressive speech information processing. The corpus contains emotional speech, dialogue speech, etc. In order to get the subtle acoustic information, the paper also presents the annotation methods with multiple perception results for emotional speech. Furthermore, some acoustic analysis results are also discussed. The corpus has been proved very useful used in our research, on both emotional speech processing and spoken language synthesis/understanding.

1. Introduction

Recently, more and more efforts have been made for the research of expressive speech.In the construction of expressive corpus, people have been made many works, achieving great achievements. Recently, Buckeye has constructed one corpus of spontaneous American English speech [2], a 307,000-word corpus containing the speech of 40 talkers from USA. In the construction of expressive corpus in Chinese, there are also some excellent achievements such as spontaneous monologue corpus CASS and spontaneous conversation corpus CADCC both collected by Chinese Academy of Social Sciences [3] [4].

In the paper, in order to get the more subtle acoustic distribution with various kinds of speech expressiveness, we designed and collected the corpus from different environment and requirement, which includes emotional speech, spontaneous dialogue speech, sentences with special syntactic structure such as questions.

In emotional speech collection, we used two groups of language resources. One group uses the same text for different emotional performance. There are 300 sentences with average length of 10 syllables within this group. Each sentence was recorded by 4 professional actors/actress in six basic emotion states, “neutral,” “happy,” “sad,” “fear,” “angry” and “surprise”. The other group (200 sentences) contains the speech of different emotions with non-parallel text, which is useful for the analysis of the stress shifting and acoustic features of emotional keywords among different emotions.

Lots of the previous analyses have been focused on the definitive emotion states, etc. Though we can, sometimes, perceive the feelings, attitudes, and moods from the emotional speech, emotion is actually not easy to be defined. In Ortony’s work[1], emotions were thought of as zones along an emotional vector. The vector might be a cline of emotions shading into one another, with extremes at either end. In the vector approach, expression would simply be a reflection of intensity in a particular zone. So, we should not just use one group of acoustic parameters, deduced from some basic discrete emotion states, for the emotional speech synthesis, but consider more subtle parameters in various emotion expressiveness as needed.

In order to get the more subtle acoustic distribution in various emotion expressiveness, we labeled the emotional speech based on the ambiguity perception results. In the perception, all of the prepared corpus is labeled by 12 subjects (students) sentence by sentence with five emotion states – “neutral,” “happy,” “sad,” “fear,” and “angry”. One or more states are allowed to be selected for each sentence, and the alignment of multiple selections is also required. Each emotion is classified into three types - “one choice,” “first choice” and “second choice” to simulate the perception ambiguities.

With the perceptually classified emotional speech, we make acoustic analyses on mean F0, F0 range, speaking rate, intensity, voice quality, and stress patterns. Both mean values and standard deviations are analyzed. The results give us more selection of acoustic parameters to synthesize the emotional speech with different expressiveness degree.

The paper also described detailed information of spontaneous dialogue speech corpus which contains 10 hours speeches and was recorded from TV, radio, and office environment with the manual transcription. Furthermore, sentences with special syntactic structure such as questions, were also designed and recorded in professional recording studio. Some acoustic analysis was done on the emotional speech and question sentences to get more detailed acoustic correlation among the speech.

The whole paper is broken down into six major parts. Section 2 introduces the emotional speech corpus, labelling method and acoustic analysis results; section 3 introduces the collection and label of spontaneous dialog speech corpus. Question sentences are specially selected to get more detailed acoustic analysis results. The final conclusion is made in section 4.

2. Emotional speech

2.1 Design and recording

The basic purpose of emotional speech corpus collection is to meet the requirement of acoustic features analysis, such as the distribution of F0, speaking rate distribution, energy, etc. among different emotions. Furthermore, we also wish to find the stress features among them and survey more relationship between linguistics and emotions. With this idea, the text scripts of the corpus are designed with two categories: parallel text and non-parallel text. All mandarin phonemes (initials and finals) are covered in two parts with the balanced context information. Each sentence contains about 10 syllables. After the text script is designed, two actors and two actresses from department of acting in Communication University of China are asked to assist us to complete this corpus. In the situation that the expression of emotion is not perfect, we simulate contextual environment to excite speaker's emotion through talking with he/she.

For the parallel text, the speaker is asked to read the same sentences with different emotional states. Most of them are meaningless sentences to ensure the easiness of reading the sentences with different emotions. For instance, in Table 1, sentence A can be expressed as afraid or angry with certain environments. This part of corpus contains 300 different sentences and 7,200 recording speech with four speakers and six emotions. For another part, the non-parallel text, which contains 1,200 different sentences, is read using with only one determinate emotional state for each utterance. Unlike the first category, this part of corpus contains some distinct words or phrases which might induce the speaker read the sentence by using a certain emotion. Normally, people always read the sentence B, still in Table 1, with the happy emotion, if there is no specific requirement for performers. With this method, the speaker can express the emotion more naturally and distinctly while there are lots of artificial prints in the first category.

Table 1: examples for performed emotional speeches

A:	Example for parallel text
	Eg: 他还没回家 ta1 hai2 mei2 hui2 jia1 Meaning： he didn't go home
B:	Example for non-parallel text
	Eg: 真高兴收到你的礼物? zhen1 gao1 xing4 shou1 dao4 ni3 de5 li3 wu4 Meaning: Glad to receive your gift!

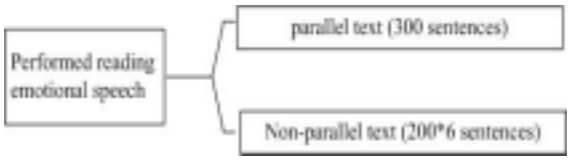

Figure 1: two categories of emotional corpus

2.2 Acoustic analysis based on perceptual results

As we know, one emotional sentence might cause different perception results. Someone think it as "happy", but others might consider it as "neutral". Even for a single person, sometimes it is still hard for him to make sure an emotion state from speech. He might think the sentence could be both "sad" and "fear", even the speaker might claim he/she just use one emotion to speak it. So in order to obtain the ambiguity of emotional speech, in our labelling system we make the labelling tag through one perception experiment.

We excised utterances tokens from the corpus and presented them with a random list to a group of 12 students. They were asked to note the "emotion" they perceived with one or two emotion states from an emotion list "happy, fear, sad, angry, surprise and neutral". To show which the better choice is, the alignment of the multiple selections are also required. The listeners have considerable freedom in their choice of labels. If the listener has strong feeling that the sentence is related to one emotion state, just one state is selected, otherwise, he is asked to select two states and line them with first choice and second choice according to the comparing between two selection results. Different listeners perceive different aspects of this multi-faceted phenomenon and it can be difficult to achieve a consensus on the choice of a single most appropriate label for any given speech utterance. Table 2 shows some response counts from the listeners:

Table 2: examples of our labelling system

utt 18	Happy, neutral
utt 36	fear

In the Table 2, utterance no.18, for example, is rated in "happy" and "neutral". It means the listener think this sentence might be both "happy" and "neutral". But "happy" seems to be stronger than "neutral". Utterance no.36 only has one labeling result "fear". It shows the listener can make sure his decision. Then, we classified the emotion labelling with three types, "one choice", "first choice" and "second choice". "one choice" means there is only one definite selected emotion labeling. "first choice" means there are two selections of the emotion for the sentence due to the perception

ambiguity, but this emotion state is of the first choice. “second choice” denotes the second choice from two selections. Then, all of the corpus was labeled with emotion perception results based on these three types.

With the perception labeling results, we got some prosody distribution on F0 contours, speaking rate, energy, etc. Figure 2 shows the distribution of the F0 mean among different emotion labeling results.

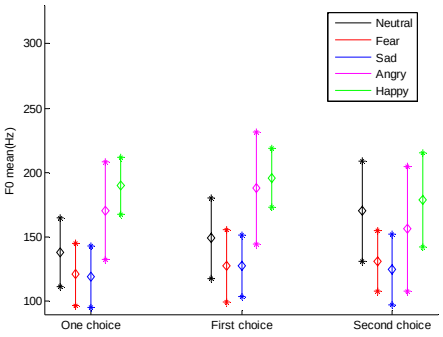

Figure 2: F0 mean distribution among different emotions and perceptual types

Voice quality is another important feature of emotional speech [6]. It is the timbre of speech, and is determined by laryngeal characteristics. There are some kinds of phonation modes, such as breathy, whisper, falsetto, creaky, normal, and so on, which correspond to certain laryngeal characteristics respectively. However, there would be some subtypes within one category. In normal mode, one end of the continuum of subtypes approaching breathy voice, where the laryngeal muscles controlling vocal fold adduction are relatively relaxed. At the other end, tension in the musculature begins to limit the vibration of the folds and voice verges on laryngealized or creaky voice [7, 9].

To be able to assist the research on the voice quality in different emotional speech, we also label the speech with voice source information. Source parameters are estimated in a two-step procedure. The speech signal is inverse-filtered to get the glottal source signal, which is then matched with a source model to estimate the parameters. A general source model is a four-parameter Liljencrants-Fant (LF) model, whose parameters are Ee (the excitation strength), Ra (the measure of the return phase), Rk (the measure of the symmetry/asymmetry of the glottal pulse), and Rg (the measure of the opening branch of the glottal pulse). The familiar parameter, open quotient (Oq), is defined as (1+Rk)/2Rg. It has been found that breathy voice has high Ra, Rk, and Oq values [8]. Figures 3 to 8 show the distribution of the source parameters among different emotion labeling results.

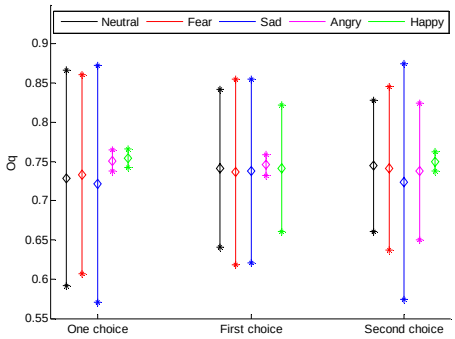

Figure 3, Distribution of “Oq” among different emotions and perceptual types

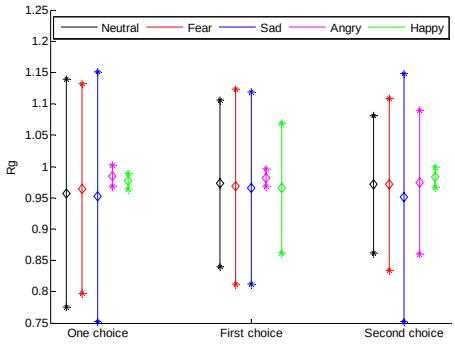

Figure 4, Distribution of “Rg” among perception labeling results

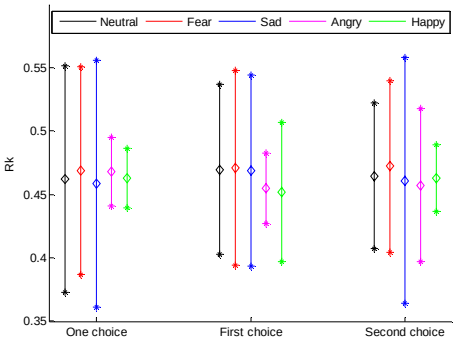

Figure 5, Distribution of “Rk” among perception labeling results

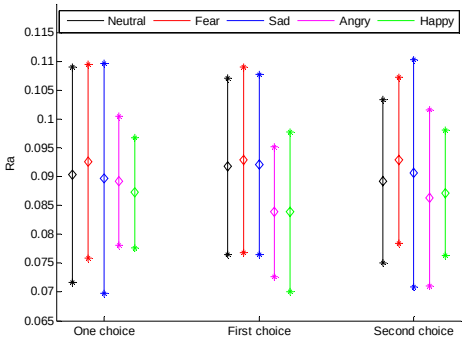

Figure 6, Distribution of “Ra” among perception labeling results

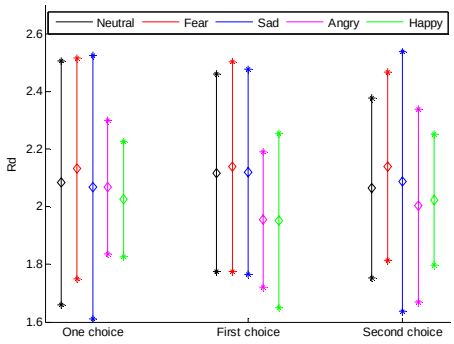

Figure 7, Distribution of "Rd" among perception labeling results

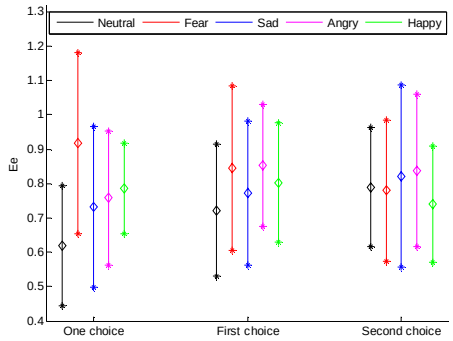

Figure 8, Distribution of "Ee" among perception labeling results

Unlike the traditional acoustic feature analysis, the perceptual labelling gives us fuzzy distribution results. For instance, figure 2 shows that "happy" and "angry" make a very high F0, while "sad" generates lower value than neutral state. We can also find that along the "one choice > first choice > second choice" sequence, the three F0 parameters are less and less distinctive in different emotion states.

From Figures 3 to 8, we can see that the mean values of glottal sources in "sad" utterances' are distinctly less than those of other emotional utterances'. The low Ee of "sad" utterances indicates its overall weak source signal as well as whispery voice. Consistent with Johnstone's experiments [12], extremely low Ra values indicate a sharp instantaneous closure of the glottis. In this regard, "sad" utterances, which are likely to reflect the relatively high degree of laryngeal tension, are similar to tense voice [10]. Thus "sad" utterances have the trend of exhibiting whispery and tense voice qualities. However, compared to higher Sq of tense voice, the mean value of Sq of "sad" utterances is lower too. The wildly suggested association of "angry" emotion with tense voice is not supported in this study. Although the mean values are employed in the current research, it can be indicated that the relation between emotion and voice quality is not one to one and an emotional speech can exhibit several characteristics of different voice qualities. In other words, concluded by Gobl [11], a given quality tends to be associated with a cluster of affective attributes.

Corresponding with their range, it can also be observed from the figures that standard deviations of Sq are the largest and standard deviations of Ra are the smallest among the six glottal parameters. Comparing the five emotional states, "angry" utterances have the largest standard deviation for Ee and Sq, while "sad" utterances have the largest standard deviation for other four parameters. The standard deviation represents the distribution of parameters, so it can be indicated that the variation of spectral intensity is the largest in "angry" utterances while the variation of voice quality is largest in "sad" utterances.

3. Dialog speech

3.1 The collection of spontaneous dialog speech

Except for performed emotional speech corpus, we have also collected the spontaneous dialog speech of which the most distinct character is that this corpus is collected from TV, radio and office environments, rather than being recorded in professional recording studio.

For corpus collection from TV and radio, the dialogs in interviewing programs and entertaining programs such as "Art and Life (CCTV3)" and "super six plus one (CCTV2)" are our first choices, in which the speaking content is not strictly fixed. while in movies and dramas, the actors' lines are strictly designed with few changes. For dialog speeches in office environments, we adopt the following recording method. In the office where we work everyday, there is a microphone connected with computer put in work all the time. Therefore all of the colleagues' talks are recorded by this computer. Then the speeches which are full of para-lingual phenomenon are selected to constitute a part of our spontaneous dialog speeches. Through this way, the most natural dialog speeches without any performing elements are constructed. To date, about 10-hour speeches have been collected and still in preceding now. Figure 9 shows the constitution of this dialog emotional corpus.

Compared with performed emotional speeches, one of the greatest advantages of these dialog speeches is their perfect naturalness. While in performed emotional speeches, there is inevitably some unnaturalness due to speakers' performing elements. In spontaneous dialog speeches collected from office environment, there are not any performing elements, and the quality of dialog speeches from TV and radio is also satisfying. These spontaneous speeches are full of various para-lingual

phenomenons such as murmur and overlap, which is a large experiment material for the research of relevant domain.

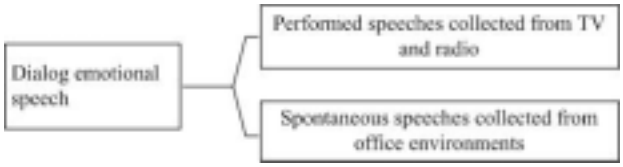

Figure 9: the constitution of dialog emotional corpus

3.2 Labelling

The labelled information for this spontaneous dialog corpus is as follows:

A. First, these speeches must be transcribed orthographically. For the exactness of transcription, five students are asked to do this work respectively. If these five results are not consistent, the transcription that most people make out is considered as most appropriate.

B. Break is perceptually labelled for each syllable. In our corpus, the prosodic boundaries are classified into four layers. They are,

- Break0: syllable boundary.
- Break1: prosodic word boundary, a group of syllables that are uttered closely.
- Break2: prosodic phrase boundary, a group of prosodic words that has a perceptive rhythm break at the end.
- Break3: sentence boundary, the utterance for a whole speech.

C. Because these speeches are uttered with no strict limitation, there are many para-lingual and non-lingual phenomenon which must also be included in the labelled information. The para-lingual and non-lingual phenomenon included in labels are as follows: beep, breathing, crying, coughing, deglutition, hawk, interjection, laughing, lengthening, murmur, noise, overlap, smack, sniffle, sneezes, yawn, etc.

3.2 Question sentences

In order to get more detailed acoustic analysis results from spontaneous speech and improve the expressiveness of the speech synthesis system, we selected some question sentences from recorded spontaneous speech, recorded them again with the help of the speakers in our lab.

In mandarin, question can be segmented into many categories according to their syntactic structure, and the intonation pattern of questions is much dependent on their syntactic structures. So in the designation of question corpus, the syntactic structure of question must be taken into account. According to the categories of question, this part of corpus is segmented in four parts, just like Table 3 shows.

Table 3: the constitution of question corpus

Part A:	"yes/no" question ending with "ma0"
	Eg: 你吃过晚饭了吗? ni3 chi1 guo4 wan3 fan4 le0 ma0 Meaning: do you have supper?
Part B:	"what" question ending with "ne0"
	Eg: 你在做什么呢? ni3 zai4 zuo4 shen1 me0 ne0 Meaning: what are you doing?
Part C:	questions with questioning words other than "ma0" and "ne0"
	Eg: 我的铅笔在那里?/你为什么要走？ wo3 de0 qian1 bi3 zai4 na3 li3?/ni3 wei4 shen1 me0 yao4 zou2 Meaning: where is my pencil?/why do you go?
Part D:	questions without questioning words
	Eg: 你是从日本来的？ ni3 shi4 cong2 ri4 ben3 lai2 de0? Meaning: are you from Japan?

In questions, there are always some questioning words indicating it a question rather than a statement. For the first two kinds of questions, the position of questioning word is in the ending syllable of sentence, that is to say, the syllable 'ma0" and "ne0". For expressing the question mood, the pitch contour begins to upward in the end, while in the end of statement, the pitch contour is trend to downward. Besides, this upward phenomenon is not limited to the last word, sometimes the last two or three syllables also behave the upward trend. For the third kind of sentence, the questioning word is "na3 li3 (where)" and "wei4 shen1 me0 (why)", which is in the middle of sentences. In this situation, the syllables near the questioning word are likely to be stressed, which have higher pitch contour and longer duration.

In all these three kinds of questions there are always certain questioning words such as "ma0", "ne0", "na3 li3" and "wei4 shen1 me0", however, in the last kind of questions, there is no questioning words. And if we delete the interrogation mark in the end of sentence, you can also read it in stating mood. So for making the listener can realize that it is a question rather than a statement, there is a very large upward in pitch contour somewhere, always in the end of sentence. This kind of question is always used to express some surprising mood.

There are 600 sentences in this question corpus in all, which includes 150 sentences for each kind. Two boys and two girls who are fluent in Mandarin are asked to read it.

In the label files of question corpus, except for syllable boundary and pitch information, there is another important and special labelling information, "focus", need to be defined. Normally, the focus in a question sentence bears the intonation accent of the sentence. For example, in part D of Table 3: ni3 shi4 cong2 ri4 ben3 lai2 de0? The focus of the sentence is located at "ri4 ben3 (Japan)". This sentence is used to express that the speaker is uncertain whether his talker is from Japan or not. Figure 3 show the pitch

contour of this sentence, we can see that there is great rise in the pitch contour of “ri4”, which demonstrate that the word “ri4 ben3” is focus of this sentence.

In our labeling procedure, the position of focus is determined through listening test. Like the transcription procedure in the label of spontaneous dialog corpus, five students are required to do this job and the last result is a compromise among their choices if not consistent.

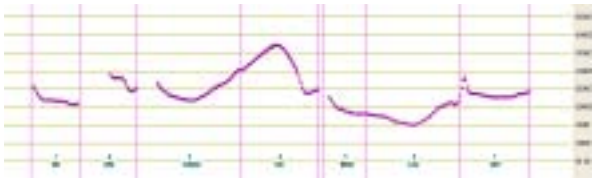

Figure 10: Example for focus in questions: “ni3 shi4 cong2 ri4 ben3 lai2 de0?”

4. Conclusion

The paper introduces an expressive mandarin speech corpus. To get more acoustic features, the paper has described a labeling method which used multiple perception results, and classified the perception results into different degrees of “one choice”, “first choice” and “second choice”. The classification results help us to get more subtle acoustic information. Except for the performed emotional speeches, the paper also introduces another important part, spontaneous dialog speeches which include lots of para-linguistic information. In order to get more detailed acoustic analysis results from spontaneous speech and improve the expressiveness of the speech synthesis system, some question sentences were selected to be recorded again for more detailed analysis. The corpus has been proved very useful used in our research, on both emotional speech processing and spoken language synthesis/understanding.

5. Acknowledgement

The author would like to thank Prof. Aijun Li for her kind advice and some joint discussion. We would also want to thank the students who take part in the perceptual experiments, which results are much helpful for our research.

Reference

[1] Ortony, A. & Turner, T. J. *What's basic about basic emotions?* Psychological Review, 97, 315-331

[2] Mark A. Pitt, Keith Johnson, et al. *The Buckeye corpus of conversational speech: labelling conventions and a test of transcriber reliability.* Speech Communication 2005.

[3] Li Aijun, et al. *Spontaneous Conversation Corpus CADCC.* Oriental COCOSDA’2001, Korea

[4] Li Aijun, Zheng Fang，William Byrne, et al. *Cass: A Phonetically Transcribed Corpus of Mandarin Spontaneous*, ICSLP 2000.

[5] Nick Campbell, *Databases of Expressive Speech*, Oriental COCOSDA’2003

[6] C. Gobl and A. N´ı Chasaide, *The role of voice quality in communicating emotion, mood and attitude*, Speech Communication, vol. 40, pp. 189–212, 2003.

[7] Scherer K.R., *Vocal affect expression: A review and a model for future research*, Psychological Bulletin, vol. 99, pp. 143–165, 1986.

[8] G. Fant, Liljencrants J., and Q. Lin, *A four-arameter model of glottal flow*, STL-QPSR 4, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, pp. 1–13, 1985.

[9] Yildirim, Serdar Bulut, et al, *An acoustic study of emotions expressed in speech*, In INTERSPEECH-2004, 2193-2196.

[10] Helmer Strik, *Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pluses*, Journal of the Acoustical Society of America, vol. 103, no. 5, pp. 2659–2669, 1998.

[11] C. Gobl and A. N´ı Chasaide, *Acoustic charavteristics of voice quality*, Speech Communication, vol. 11, pp. 481–490, 1992.

[12] T. Johnstone and K. R. Scherer, *The effects of emotions on voice quality*, XIVth ICPhS1999.

[13] Nick Campbell, *Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation*, ICSLP2004, Jeju, Oct, 2004.