Mandarin Affective Speech Corpus

Introduction

This publication contains audio recordings and corresponding transcripts, collected over a three-week period from August to September, 2005.

An emotional speech database of the Chinese language MASC (Mandarin Affective Speech Corpus) is designed and established with two major objectives. First, it is a tool for linguistic and prosodic feature investigation of emotion expression in Chinese. Second, it serves as a source of training and test data essential to support the research in speaker recognition with affective speech.

Authors

Yingchun Yang(yyc@cs.zju.edu.cn); Dongdong Li(lidd@cs.zju.edu.cn); Tian Wu；Zhaohui Wu(wzh@cs.zju.edu.cn)

Data sources

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.

Emotional States

- 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

Speakers

More than 100 speakers were invited to attend the data collection. After being screened, the recordings from 68 speakers are adopted, including 23 females and 45 males. The majorities are at the age of twenties.

The detail information is listed in the file “SpeakerInfo.doc”.

Organization of the corpus

The materials span five distinct emotional categories, neutral (unemotional), panic, anger, sadness and elation and covers all the phonemes in Chinese. For each category, a phonetically representative text of 5 phrases and 20 sentences was compiled. The sentences include 12 semantically neutral ones and 2 emotional ones for each type portraying the four emotional states. The sentences also represented different syntactical types. Each Phrase and sentence is repeated for three times respectively. Additionally, two paragraphs are expressed in neutral. Altogether the database contains 5,100 phrases, 20,400 utterances and 136 paragraphs.

“ContentDetail.pdf” presents the content of the speech and the corresponding slice name.

Description of data

The sample rate for the speech files is 8kHz, and the sample coding is 16-bit (mono waveform data). For the original speech data, .wma files are saved in 44kHz, 32-bit. All the data are collected in a quiet square office room.

Protocol of the database

Two different protocols are defined to evaluate the performance of speaker authentication on varied speech. They differ in the distribution of the training and the test data as can be seen in Fig. 1.

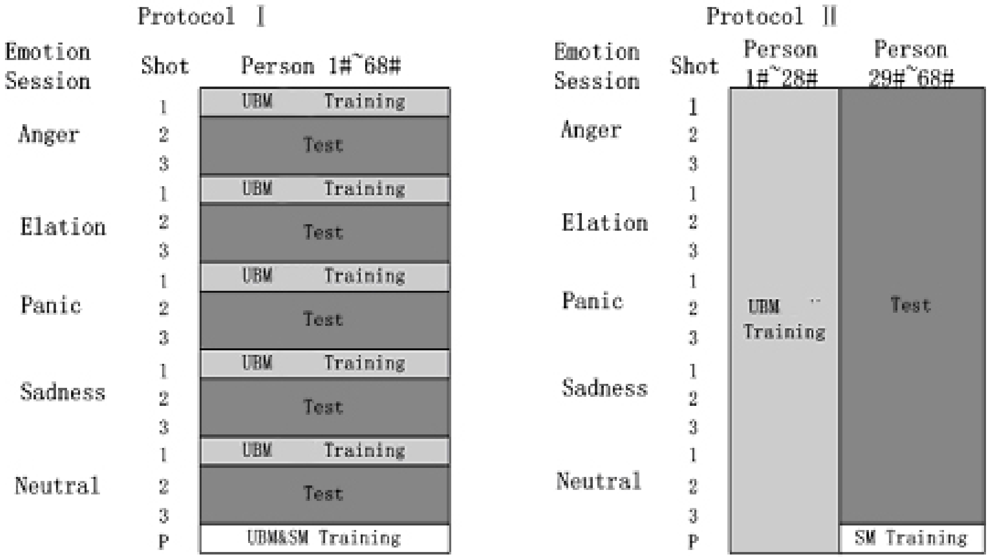

Fig. 1. The diagram shows the partitioning of the MASC according to protocol Ⅰ (left) and Ⅱ (right). Shot ‘P’ of the neutral session stands for the paragraph part of the neutral speech.

The first protocol constructs the UBM with the paragraph part of neutral speech and the first reading of 20 utterances part spanning 5 types of affective speech of all the 68 people. Then five emotional dependent models are adapted with the above 20*68 utterances of each type respectively. The training set for speaker models are the feature from the paragraph part. The remained two readings of 20 utterances each part are used for test.

In the second protocol, we divide the Emotional Mandarin Speech Corpus into two parts. The speech of first 28 people is used to train the UBM. For the rest 40 persons, the features from the paragraph part are used to train speaker models. The utterances part is for testing.

Both the SM training and the pre-build model (like the UBM) training data are drawn from the same speaker sets for Protocol Ⅰ which might lead to good performance on the test set. For Protocol Ⅱ on the other hand, the evaluated parameters are totally separated from the test set of the system. For this description, each subject appears only in one set which is an important requirement to ensure the realistic evaluation of the system. The duration of the materials used to build different models are listed in Table 1.

Table 1. The duration of training and testing speech is listed in Table 1. "Second" is used as measurement. The average time is considered to estimate the duration. For example, the uttances which used to train the UBM in protocol Ⅰ consist of a paragraph and 20 utterances repeated for 5 types of affective speech per person. The paragraph lasts 30 second on the average, while the utterance is 1 second.

	ProtocolⅠ(second)	Protocol Ⅱ(second)
UBM	(30+20*5)*68	(30+20*3*5)*28
SM	30 per person	30 per person
Test	20*2*5*68	20*3*5*40