Preparation of Verbal and Nonverbal Information of Speech in Spontaneous Conversation Database

Klára Vicsi

Department of Telecommunication and Media Informatics of Budapest University of Technology and Economics
vicsi@tmit.bme.hu

Two channels have been distinguished in human interaction. One conveys messages with a specific semantic content (verbal channel); the other (the non-verbal channel) conveys information related to both the image content of a message and to the general feeling and emotional state of the speaker.

Enormous efforts have been made in the past to understand the verbal channel, and the huge number of databases was prepared, whereas the role of the non-verbal channel is less well understood. There are some results about emotion characterisation in speech and emotional recognition in the literature, but those results were obtained in clear lab speech. Most of them usually used simulated emotional speech databases, more frequently produced by artists.

Group of emotional categorization is the one commonly used in psychology, linguistics and speech technology, also described in the MPEG-4 standard: happiness, sadness, anger, surprise, and scorn/disgust. But the real word data differ much from acted speech, and in the application of speech technology, real word data processing is necessary. And these 5 emotion categories do not mask the emotions in the real conversation. In the last years some works were published dealing with examination and recognition of emotion in spontaneous everyday conversations.

This paper describes database construction techniques where the linguistic content (verbal channel) and emotion state of the speaker (nonverbal channel) are processed parallel on a corpus of spontaneous everyday conversations between telephone dispatchers and customers, through telephone line. This technique is advised for general use to process spontaneous everyday conversations.

On the base of this database not only the acoustical parameters of emotions were examined and classified, but word and word connection statistics of different emotional text is planned to prepare on the base of the corpus, and word spotting, as well.

The following questions would be nice to discuss in the workshop:

Which kind of the emotions are *distinguishable in the everyday speech?*

Nowadays, in automatic speech recognition we have to collect different databases in different acoustical environments. Can we do something to reduce the influence of the acoustical environments?

Language independent automatic annotation tools are needed. How we can help?