Chapter 8.20
Voice Driven Emotion Recognizer Mobile Phone:
Proposal and Evaluations

Aishah Abdul Razak
Multimedia University, Malaysia

Mohamad Izani Zainal Abidin
Multimedia University, Malaysia

Ryoichi Komiya
Multimedia University, Malaysia

ABSTRACT

This article proposes an application of emotion recognizer system in telecommunications entitled voice driven emotion recognizer mobile phone (VDERM). The design implements a voice-to-image conversion scheme through a voice-to-image converter that extracts emotion features in the voice, recognizes them, and selects the corresponding facial expression images from image bank. Since it only requires audio transmission, it can support video communication at a much lower bit rate than the conventional videophone. The first prototype of VDERM system has been implemented into a personal computer. The coder, voice-to-image converter, image database, and system interface are preinstalled in the personal computer. In this article, we present and discuss some evaluations that have been conducted in supporting this proposed prototype. The results have shown that both voice and image are important for people to correctly recognize emotion in telecommunications and the proposed solution can provide an alternative to videophone systems. The future works list some modifications that can be done to the proposed prototype in order to make it more practical for mobile applications.

INTRODUCTION AND MOTIVATION

Nonverbal communication plays a very important role in human communications (Komiya, Mohd Arif, Ramliy, Gowri, & Mokhtar, 1999). However,

in telephone systems, only audio information can be exchanged. Thus, using telephony, the transmission of nonverbal information such as one's emotion would depend mostly on the user's conversation skills. Although the importance of nonverbal aspects of communication has been recognized, until now most research on nonverbal information concentrated on image transmission such as transmission of facial expression and gesture using video signal. This has contributed to the emergence of a videophone system, which is one of the most preferred ways to exchange more information in communication. Such services, however, require a wide bandwidth in order to provide real time video that is adequate for a natural conversation. This is often either very expensive to provide or difficult to implement. Besides, in a videophone system, the user has to be fixed in front of the camera at the correct position during the conversation, so that the user's image can be captured and transmitted correctly. This limitation does not happen in the normal telephone system.

Another approach is to use model-based coding (Kidani, 1999). In this approach, instead of transmitting video signals containing an image of the user, only the human action data such as the facial expressions, movement of the mouth, and so on acquired using a microphone, a keypad, and other input devices, are transmitted over the network. When these data are received by the receiver, the polygon coordinate data for each facial feature is recalculated in accordance with the displacement rules and the person's expression is synthesized.

Our approach is similar to the second approach in a sense that a synthesize image is used for the facial expression reconstruction at the receiver side. However, the difference is that only voice is transmitted and the emotion data is extracted from the received voice tone at the receiving side. This is based on the idea that, voice, besides for communication, it is also an indicator of the psychological and physiological state of a speaker. The identification of the pertinent features in the speech signal may therefore allow the evaluation of a person's emotional state. In other words, by extracting the emotion information from the voice of the speaker, it is possible to reconstruct the facial expression of that speaker. Thus, based on this voice-to-image conversion scheme, we propose a new system known as voice driven emotion recognizer mobile phone (VDERM), as seen in Figure 1. This system uses a voice-to-image converter system at the receiver side that identifies the emotional state of the received voice signal and selects the corresponding facial expression of that particular emotion from the image bank to be displayed. Using this approach, only audio transmission is required. Therefore, the existing second generation (2G) mobile phone infrastructures can be used. Another advantage is that the user does not need to be fixed in front of the camera during the conversation because there is no need for image transmission.

VOICE TO IMAGE CONVERSION

Referring to Figure 1, the voice-to-image conversion for this system is done at the receiving side. The conversion scheme can be divided into two parts: the emotion recognition and facial expression reconstructor. These two processes are done by the voice-to-image converter.

Emotion Recognition

Before we come out with the emotion recognizer design, first we have to deal with these three issues:

1. What kind of emotion to be recognized?

How many and what types of emotional states should be recognized by our system is an interesting yet difficult issue. Besides, there is no widely accepted definition and taxonomy of emotion; it