Emotion recognition using multilayer perceptron and generalized feed forward neural network

K B Khanchandani* and Moiz A Hussain

Electronics & Telecommunication Department, SSGM College of Engineering, Shegaon 444 203, India

Received 18 July 2008; revised 10 February 2009; accepted 16 February 2009

This paper explores performance analysis of multilayer perceptron neural network (MLPNN) and generalized feed forward neural network (GFFNN) for detection of 7 human emotions (neutral, anger, boredom, disgust, fear, happiness, sadness) using speech signals. Overall accuracy was found as follows: MLPNN, 93%; and GFFNN, 99%.s

Keywords: Berlin Emotional Speech Database, Emotion recognition, Generalized feed forward neural network (GFFNN), Machine intelligence, Multilayer perceptron neural network (MLPNN)

Introduction

Ability to understand human emotions is desirable for computer in applications such as lie detector, developing learning environments, consumer relations, entertainment etc. Experiments have been conducted for designing intelligent human – machine interaction by simulating emotional intelligence of human brain. Machines can recognize “what is said” and “who said it” using speech recognition and speaker identification techniques. If equipped with emotion recognition techniques, machines can also know “how it is said”. In the field of human computer interaction (HCI), apart from facial expressions and gestures, speech is a powerful medium to communicate with emotional intelligence.

This paper explores neural network (NN) methods to recognize human emotions (neutral, anger, boredom disgust, fear, happiness, and sadness) in speech signal.

Emotional Speech Database

This study used Berlin Emotional Speech Database, which contains speech sample from 5 actors and 5 actresses, and 10 different sentences of 7 kinds of emotions (anger, boredom, disgust, fear, happiness, sadness and neutral). Out of 493 speech samples (length of each speech sample, 2-8 s) in this database, 286 samples were of female voice and 207 of male voice. Emotional speech samples were categorized into natural vocal expression, induced emotional expression, and simulated emotional expression. Natural vocal expression was recorded during naturally occurring emotional states of various sorts (TV program, talking shows or interactive game shows). Using psychoactive drugs of games or events generates induced emotions. Simulated emotional expressions are vocal expressions of certain emotions.

Materials and Methods

Emotion Recognition System Design

Emotion recognition system (Fig. 1) consists of 6 modules (speech input, preprocessing, spectral analysis, feature extraction, neural network training, and recognized emotions output). Solutions to emotion recognition depend on type and purpose of emotion.

Computer Simulation Experiment

Speech signals were analyzed to categories emotions by studying prosodic features of speech (formant frequencies, entropy, variance, minima, median, linear prediction corrector (LPC)), which were extracted by spectral analysis and feature extraction algorithms. Then, multilayer perceptron neural networks (MLPNNs) and generalized feed forward neural networks (GFFNNs) trained extracted emotional speech parameters.

*Author for correspondence
E-mail: kkhanchandani@rediffmail.com
(a) Pre-Processing

Speech signal database was filtered from noise and then normalized (Fig. 2).

(b) Feature Extraction

An algorithm was programmed in MATLAB to obtain statistical parameters (formant frequencies, entropy, variance, minima, median, LPC) of a speech signal database (Fig. 3) and dataset for all 493 speech samples was prepared to feed to NN for emotion recognition.

(c) Speech Analysis

Entropy and formant frequency are widely used in speech recognition, verification, and other applications. Energy contours (log entropy, Shannon entropy, threshold entropy, sure entropy, norm entropy) and formant frequency contours (Formant0, Formant1, Formant2, Formant3, and Formant4) were extracted from speech signal. Audible segments were determined by choosing a threshold below maximum energy, and then a contour was produced showing audible/inaudible segments of speech. Speech analysis was performed to find out possible features that can truly represent emotional state regardless of speaker’s gender and context.

Neural Network (NN)

Design architecture of NN depicts 14 input neurons and 7 output neurons (Fig. 4).

Multilayer Perceptron Neural Network (MLPNN)

MLPNNs are layered feed forward (FF) networks typically trained with static back propagation in order to classify static pattern. However, MLPNNs train slowly, and require lots of training data (typically three times more training samples than network weights).
Frequency, Hz

Magnitude, dB

Fig. 3—LPC spectra of speech signal

Fig. 4—Feed forward neural network

Fig. 5—Schematic diagram of GFFNN
Generalized Feed Forward Neural Networks (GFFNN)

GFFNNs are a generalization of MLPNN such that connections can jump over one or more layers (Fig. 5). In theory, a MLPNN can solve any problem that a generalized FF network can solve. In practice, however, generalized FF networks often solve the problem more efficiently. A classic example of this is two-spiral problem. Without describing the problem, it suffices to say that a standard MLPNN requires hundreds of times more training sets than generalized FF network containing the same number of processing elements.

Fig. 6—Comparison of MLPNN and GFFNN for: a) training dataset; b) cross validation dataset
Results and Discussion

MLPNNs and GFFNNs are trained using learning rules [Momentum, Conjugate-Gradient (CG), Quickprop (QP) and Delta Bar Delta (DBD)].

MLPNN for Emotion Recognition System

Extracted features of speech signals are fed randomly to MLPNN and data is trained for different hidden layers. MLPNN with three hidden layer gave better performance. Numbers of processing elements (PEs) in hidden layer were varied. Network was trained and optimal results are obtained when 10 PEs were used in hidden layer. For training MLPNN, data was used as follows: training dataset 90; and testing dataset, 10%. Overall accurate recognition of emotions by MLPNN has been found to be 89.62% (Table 1, Fig. 6).

GFFNN Emotion Recognition System

Extracted features of speech signal was fed randomly to NN and trained for different hidden layers. Transfer function was used for firing of neuron for specified threshold value. Same number of training and testing (cv) data sets of MLPNN were used for GFFNN. After training, GFFNN with five hidden layer gave better performance. GFFNN was trained and optimal results were obtained when 19 PEs were used in hidden layer. Overall accurate recognition of emotions by GFFNN has been found to be 98.08% (Table 1, Fig. 6).

Conclusions

In order to develop computers and robots having multiple intelligence, performance of MLPNN and GFFNN for recognition of emotions was examined. GFFNN (98.08%) recognized emotional test patterns with more accuracy than MLPNN (89.62%).

References