Speech Recognition for Emotions with Neural Network: A Design Approach

Shubhangi Giripunje[1] and Ashish Panat[2]

[1]Lecturer, G. H. Raisoni college of Engineering, Nagpur, India
shubha_giripunje@yahoo.com
[2]Senior Lecturer, College of Engineering, Bandera, India
ashishpanat@rediffmail.com

Abstract. Worldwide research is going on to judge the emotional state of a speaker just from the quality of human voice. This paper explores use of supervised neural network to design a classifier that can discriminate between several emotions like happiness, anger, fear, sadness & unemotional state in speech. The results found to be are significant, both in cognitive science and in speech technology. In the current paper, statistics of the pitch like, first and second formants, and Energy and speaking rate are used as relevant features. Different neural network based recognizers are created. Ensembles of such recognizers are used as an important part of decision support system for prioritizing voice messages and assigning a proper agent to response the message. The developed intelligent system can be enhanced to automatically predict and adapt to detect people's emotional states and also to design emotional robot or computer system.

1 Introduction

The ability to express and recognize emotions or attitudes through the modulation of the intonation of the voice is fundamental to human communication. A new wave of interest has recently risen attracting both psychologists and artificial intelligence specialists. There are several reasons for this renewed interest such as: technological progress in recording, storing, and processing audio and visual information. A new field of research in AI known as affective computing has recently been identified [1]. As to research on recognizing emotions in speech, on one hand, psychologists have done many experiments and suggested theories. AI researchers had made contributions in the areas of emotional speech synthesis [2-3], recognition of emotions [4], and using agents for decoding and expressing emotions [5]. In current paper, an attempt has been made in the area of speech recognition. The proposed system can be used in application such as telephone call center and to develop emotional robot and computer system.

2 Emotions in Speech

Human emotional states can only be identified indirectly. Different emotional states affect the speech production mechanism of a speaker in different ways, and lead to

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 640–645, 2004.

acoustical changes in their speech. Listeners can perceive these changes as being due to emotion. Generally, *emotion* refers to short-term states. Emotions have some mechanical effects on physiology, like heart rate modulation or dryness in the mouth, which in turn have effects on the intonation of the voice. This is why it is possible in principle to predict some emotional information from the prosody of a sentence. The various emotions in speech about their characteristics can be expressed as follows.

Anger: Anger generally seems to be characterized by increase in mean F0 (fundamental frequency), F0 (fundamental frequency) variability and mean energy. Further anger effects include increases in high frequency energy and downward directed F0 (fundamental frequency) contours.

Sadness: A decrease in mean F0 (fundamental frequency), F0 range and mean energy is usually found, as are downward directed F0 (fundamental frequency) contours. There is evidence that high frequency energy.

Happiness: Findings converge on increases in mean F0 (fundamental frequency), F0 (fundamental frequency) range, F0 (fundamental frequency) variability and mean energy. There is some evidence for an increase in high frequency energy.

To recognize emotions in speech, the existing approaches are K-nearest neighbors, and set of experts etc. In this paper, the possibility of use of ANN is explored to detect various types of emotions in speech. While designing classifiers for emotions, it is essential to learn how well people recognize emotions in speech, to find out which features of speech signal could be useful for emotion recognition, and explore different mathematical models for creating reliable recognizers.

3 Features Extraction Technique

All studies in this field point to the pitch (fundamental frequency) as the main vocal cue for emotion recognition. The other acoustic variables contributing to vocal emotion signaling are: vocal energy [6], frequency spectral features and formants. Usually only one or two first formants (F1, F2) and temporal features (speech rate and pausing) are considered. Another approach to feature extraction is to enrich the set of features by considering some derivative features such as LPC (linear predictive coding) parameters of signal or features of the smoothed pitch contour and its derivatives. For our study authors estimated the acoustical variables like fundamental frequency F0 (fundamental frequency), energy, speaking rate, first three formants (F1, F2, and F3) and their bandwidths (BW1, BW2, and BW3), and calculated some descriptive statistics for them. Then ranking has been done for the statistics using feature selection techniques, and picked a set of most “important” features. All other parameters, which have been calculated, are mean, standard deviation, minimum, maximum, and range. The pitch of an utterance offers a both meaningful and reasonably reliably detectable representation. The scheme outlined above yielded *F0 (fundamental frequency)* values consistent with the perceived pitch. In most analyses,