Emotion Recognition Agents in Real World

Valery A. Petrushin

Andersen Consulting
3773 Willow Road
Northbrook, IL 60062, USA
petr@cstar.ac.com

Abstract

The paper describes agents for emotion recognition in speech and their application to a real world problem. The agents can recognized five emotional states with the following accuracy: normal or unemotional state - 55-75%, happiness - 60-70%, anger - 70-80%, sadness - 75-85%, and fear - 35-55%. The total average accuracy is about 70%. The agents can be adapted to a particular environment depending on parameters of speech signal and the number of target emotions. For a practical application an agent has been created that is able to analyze telephone quality speech signal and distinguish between two emotional states ("agitation" which includes anger, happiness and fear, and "calm" which includes normal state and sadness) with the accuracy 77%. The agent was used as a part of a decision support system for prioritizing voice messages and assigning a proper human agent to response the message at call center environment.

Introduction

A new wave of interest in doing research on emotion recognition and synthesis in speech 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;
- the development of non-intrusive sensors;
- the advent of wearable computers;
- the urge to enrich human-computer interface from point-and-click to sense-and-feel;
- and the invasion on our computers of lifelike software agents and in our homes robotic animal-like devices like Tiger's Furbies and Sony's Aibo who supposed to be able express, have and understand emotions.

A new field of research in AI known as affective computing has recently been identified (Picard, 1997). As to research on recognizing emotions in speech, on one hand, psychologists have done many experiments and suggested theories (reviews of about 60 years of research can be found in (Scherer et al., 1991)). On the other hand, AI researchers made contributions in the following areas: emotional speech synthesis (Murray and Arnott, 1993), recognition of emotions (Dellaert et al., 1996), and using agents for decoding and expressing emotions (Tosa and Nakatsu, 1996). Our project is motivated by the question of how recognition of emotions in speech could be used for business. One potential application is the detection of the emotional state in telephone call center conversations, and providing feedback to an operator or a supervisor for monitoring purposes. Another application is sorting voice mail messages according to the emotions expressed by the caller.

Recognizer

First we had to develop a methodology for creating emotion recognition engines for speech signal. This methodology is based on our previous research how well can people recognize and portray emotion (Petrushin, 1999). The methodology includes the following steps: collecting and evaluating emotional data, selecting data for machine learning, feature extraction, model selection, training, testing and selecting (a set of) recognizers. Below we describe the results of our research that are laying behind two the most critical steps of our methodology: feature extraction and model selection.

Feature Extraction

All studies in the field point to the pitch (which is represented by the fundamental frequency F0) as the main vocal cue for emotion recognition. The other acoustic variables contributing to vocal emotion signaling are (Banse and Scherer, 1996): vocal energy, frequency spectral features, formants (usually only one or two first formants F1 and F2 are considered), and temporal features (speech rate and pausing). Another approach to feature extraction is to enrich the set of features by considering some derivative features such as LPCC (linear predictive coding cepstrum) parameters of signal (Tosa and Nakatsu, 1996) or features of the smoothed pitch contour and its derivatives (Dellaert et al., 1996). We took into account fundamental frequency F0, energy, speaking rate, first three formants and their bandwidths and calculated some descriptive statistics for them. Then we ranked the statistics using feature selection techniques, and picked a set of most "important" features. The speaking rate was calculated as the inverse of the average length of the voiced part of utterance. For all other parameters we calculated the following statistics: mean, standard deviation, minimum, maximum, and range. Additionally for F0 the slope was calculated as a linear regression for voiced part of speech. We also calculated the relative voiced energy as the proportion of voiced energy to the total energy of utterance. Altogether we have estimated 43 features for each utterance. We used the RELIEF-F algorithm (Kononenko, 1994) for feature selection. The top 14 features are the following: F0 maximum, F0 standard deviation, F0 range, F0 mean, BW1 mean, BW2 mean, energy standard deviation, speaking rate, F0 slope, F1 maximum, energy maximum, energy range, F2 range, and F1 range. To investigate how sets of features influence the accuracy of emotion recognition algorithms we have formed three nested sets of features based on their sum of ranks. The first set includes the top eight features, the second set extends the first one by two next features and the third set includes all 14 top features.

Model Selection

To recognize emotions in speech we tried the following approaches: K-nearest neighbors, neural networks, ensembles of neural network classifiers, set of experts.

K-nearest neighbors. This method estimates the local posterior probability of each class by the weighted average of class membership over the K nearest neighbors. We ran this approach for K from 1 to 15 and for number of features 8, 10, and 14. The best average accuracy of recognition (~55%) can be reached using 8 features, but the average accuracy for anger is much higher (~65%) for 10 and 14-feature sets. All recognizers performed very poor for fear (~13%, ~7%, and ~1% for number of features 8, 10, and 14 correspondingly).

Neural networks. We used a two-layer backpropagation neural network architecture with a 8-, 10- or 14-element input vector, 10 or 20 nodes in the hidden sigmoid layer and five nodes in the output linear layer. This approach gave the average accuracy of about 65% with the following distribution for emotional categories: normal state is 55-65%, happiness is 60-70%, anger is 60-80%, sadness is 60-70%, and fear is 25-50%.

Ensembles of neural network classifiers. An ensemble consists of an odd number of neural network classifiers, which have been trained on different subsets of the training set using the bootstrap aggregation and the cross-validated committees techniques. The ensemble makes decision based on the majority voting principle. We used ensemble sizes from 7 to 15. The average accuracy of recognition for ensembles of 15 neural networks, all three sets of features, and both neural network architectures (10 and 20 neurons in the hidden layer) is about 70% with the following distribution for emotional categories: normal state is 55-75%, happiness is 60-70%, anger is 70-80%, sadness is 70-80%, and fear is 35-53%.

Set of experts. The last approach, which we have tried, is based on the following idea: instead of training a neural network to recognize all emotions, we can build a set of specialists or experts that can recognize only one emotion and then combine their results to classify a given sample. To train the experts we used a two-layer backpropagation neural network architecture with a 8-element input vector, 10 or 20 nodes in the hidden sigmoid layer and one node in the output linear layer. The accuracy of emotion recognition is about 70% for all emotions except fear which is ~44% for the 10-neuron, and ~56% for the 20-neuron architecture. The accuracy of non-emotion (non-angry, non-happy, etc.) is 85-92%. The important question is how to combine opinions of the experts to obtain the class of a given sample. A simple and natural rule is to choose the class which expert's value is closest to 1. This rule gave the total accuracy of 60% for the 10-neuron architecture and about 53% for the 20-neuron architecture. Another approach is to use the outputs of expert recognizers as input vectors for a new neural network. In this case we give a neural network an opportunity to learn itself the most appropriate rule. To explore this approach, we used a two-layer backpropagation neural network architecture with a 5-element input vector, 10 or 20 nodes in the hidden sigmoid layer and five nodes in the output linear layer. We have selected five of the best experts and generated several dozens neural network recognizers. The total accuracy is about 63% and stays the same for both 10-and 20-node architectures. The average accuracy for sadness is rather high ~76%. In general, the approach, which is based on ensembles of neural network recognizers, outperformed the others and was chosen for implementation of emotion recognition agents.

Agents

The following pieces of software have been developed:

- ERG: Emotion Recognition Game.
- ER: Emotion Recognition software for call centers.
- SpeakSoftly: a dialog emotion recognition program.

The first program has been mostly developed to demonstrate the results of the above research. The second software system is a full-fledge prototype of an industrial solution for computerized call centers. The third program just adds a different user interface to the core of the ER system. It has been developed to demonstrate the real time emotion recognition. Below we shall describe briefly the ERG program and give more detail on the ER system.

ERG: Emotion Recognition Game

The program allows a user to compete against the computer or another person to see who can better recognize emotion in recorded speech. After entering his/her name and the number of tasks the user is presented a randomly chosen utterance from a data set of 350 utterances and is asked to recognize what kind of emotions the utterance presents by choosing one of the five basic emotions. The user clicks on a corresponding button and a clown portrays visually the choice. Then the emotion recognition agent presents its decision based on the vector of feature values for the utterance. Both the user's and the agent's decisions are compared to the decision obtained during the evaluation of the utterance. If only one player gave the right answer then he/she/it adds two points to his/her/its score. If both players are right then they add one point to their scores, otherwise no points are assigned. The player has to get the larger score to win. The program serves mostly as a demonstration of the computer's ability to recognize emotions, but one potential practical application of the game is to help autistic people in developing better emotional skills at recognizing emotion in speech.

ERG: Emotion Recognition Game

Goal. The goal of the development of this software was to create an emotion recognition agent that can process telephone quality voice messages (8 kHz/8 bit) in real-time and can be used as a part of a decision support system for prioritizing voice messages and assigning a proper human agent to respond the message.

Agent. It was not a surprise that anger was identified as the most important emotion for call centers. Taking into account the importance of anger and scarcity of data for some other emotions we decided to create an agent that can distinguish between two states: "agitation" which includes anger, happiness and fear, and "calm" which includes normal state and sadness. To create the agent we used a corpus of 56 telephone messages of varying length (from 15 to 90 seconds) expressing mostly normal and angry emotions that were recorded by eighteen non-professional actors. These utterances were automatically split into 1-3 second chunks, which were then evaluated and labeled by people. They were used for creating recognizers using the methodology developed earlier. We created the agents that include ensembles of 15 neural network recognizers for the 8-,10-, and 14-feature inputs and the 10- and 20-node architectures. The average accuracy of this approach lies in the range 73-77% and achieves its maximum ~77% for the 8-feature input and 10-node architecture.

System Structure. The ER system is a part of a new generation computerized call center that integrates databases, decision support systems, and different media such as voice messages, e-mail messages and a WWW server into one information space. The system consists of three processes: the wave file monitor agent, the message prioritizing agent, and the voice mail center. The wave file monitor reads every 10 seconds the contents of voice message directory, compares it to the list of processed messages, and, if a new message is detected, it calls the emotion recognition agent that processes the message and creates a summary file and an emotion description file. The prioritizer is a agent that reads summary files for processed messages, sorts messages taking into account their emotional content, length and some other criteria, and suggests an assignment of human agents to return back the calls. Finally, it generates a web page, which lists all current assignments. The voice mail center is an additional tool that helps operators and supervisors to visualize emotional content of voice messages; sort them by name, date and time, length, and emotional content; and playback the whole message or a part of it.

Summary

A methodology for developing emotion recognition agents in speech has been developed. It extracts some features from the following acoustical variables: fundamental frequency, energy, speaking rate and formant frequencies and bandwidths. It uses an ensemble of neural network classifiers as a recognition engine. The methodology allows to create emotion recognition agents for a particular environment. The agents can be build into variety of applications such as games, tutoring systems, toys, robotic devices, voice-enabled chat rooms on the Web and other socially rich environments. The approach proved to be useful for call centers to improve the quality of customer service.

References

Banse, R. and Scherer, K.R. (1996) Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*. 70: 614-636, 1996.

Dellaert, F., Polzin, Th., and Waibel, A. (1996) Recognizing emotions in speech. *ICSLP 96*.

Hansen, L. and Salomon, P. (1990) Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 12:993-1001, 1990.

Kononenko, I. (1994) Estimating attributes: Analysis and extension of RELIEF. In L. De Raedt and F. Bergadano (eds.) *Proc. European Conf. On Machine Learning*. 171-182, 1994.

Murray, I.R. and Arnott, J.L. (1993) Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotions. *Journal Acoustical society of America*; 93(2): 1097-1108, 1993.

Petrushin, V. A. (1999) Emotion in Speech: Recognition and Application to Call Centers. In C. Dagli, A. Buczak, J. Ghosh, M.J. Embrechts, and O. Ersoy (eds.) *Intelligent Engineering Systems Through Artificial Neural Networks*, vol 9., 1085-1092.

Picard, R. (1997) *Affective computing.* The MIT Press. 1997.

Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck T. (1991) Vocal clues in emotion encoding and decoding. *Motiv Emotion* 1991; 15: 123-148, 1991.

Tosa, N. and Nakatsu, R. (1996) Life-like communication agent-emotion sensing character "MIC" and feeling session character "MUSE". *Proceedings of IEEE Conference on Multimedia 1996*. pp. 12-19