A version of this paper also appears in the *Proceedings of the*
Sixth International Conference on Spoken Language Processing (ICSLP 2000)

EMOTION RECOGNITION IN SPEECH SIGNAL: EXPERIMENTAL STUDY, DEVELOPMENT, AND APPLICATION

Valery A. Petrushin

Center for Strategic Technology Research (CSTaR)
Accenture
3773 Willow Rd., Northbrook, IL 60062, USA
petr@cstar.accenture.com

ABSTRACT

The paper describes an experimental study on vocal emotion expression and recognition and the development of a computer agent for emotion recognition. The study deals with a corpus of 700 short utterances expressing five emotions: happiness, anger, sadness, fear, and normal (unemotional) state, which were portrayed by thirty subjects. The utterances were evaluated by twenty three subjects, twenty of whom participated in recording. The accuracy of recognition emotions in speech is the following: happiness - 61.4%, anger - 72.2%, sadness - 68.3%, fear - 49.5%, and normal - 66.3%. The human ability to portray emotions is approximately at the same level (happiness - 59.8%, anger - 71.7%, sadness - 68.1%, fear - 49.7%, and normal - 65.1%), but the standard deviation is much larger. The human ability to recognize their own emotions has been also evaluated. It turned out that people are good in recognition anger (98.1%), sadness (80%) and fear (78.8%), but are less confident for normal state (71.9%) and happiness (71.2%). A part of the corpus was used for extracting features and training computer based recognizers. Some statistics of the pitch, the first and second formants, energy and the speaking rate were selected and several types of recognizers were created and compared. The best results were obtained using the ensembles of neural network recognizers, which demonstrated the following accuracy: normal state - 55-75%, happiness - 60-70%, anger - 70-80%, sadness - 75-85%, and fear - 35-55%. The total average accuracy is about 70%. An emotion recognition agent was created that is able to analyze telephone quality speech signal and distinguish between two emotional states --"agitation" and "calm" -- with the accuracy of 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. The architecture of the system is presented and discussed.

1. INTRODUCTION

The first book on expression of emotions in animals and humans was written by Charles Darwin in the nineteenth century [1]. After this milestone work psychologists have gradually accumulated knowledge in the field. A new wave of interest has recently risen attracting both psychologists and artificial intelligence specialists. There are several reasons for this renaissance 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 [2]. As to research on decoding and portraying 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 [3,4]). On the other hand, AI researchers made contributions in the following areas: emotional speech synthesis [5], recognition of emotions [6], and using agents for decoding and expressing emotions [7].

The motivation for our research is to explore the ways how recognition of emotions in speech could be used for business, in particular, in a call center environment. One potential application is the detection of the emotional state in telephone conversations, and providing a 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. One more challenging problem is to use emotional content of the conversation for the operator performance evaluation.

2. EXPERIMENTAL STUDY

Keeping in mind the above motivation, we imposed the following restrictions on our study: we solicited data from people who are not professional actors or actresses; we concentrated on the negative emotions like anger, sadness and fear; we targeted the telephone quality speech, and we relied on voice signal only. The last restriction means that we excluded the speech recognition techniques. There are several reasons to do this. First, in speech recognition emotions are considered as annoying noise that decreases the accuracy of recognition. Second, although it is true that some words and phrases are correlated with particular emotions, the situation usually is much more complex and the same word or phrase can express the whole spectrum of emotions.

The objectives of the experimental study are the following: to learn how well people recognize emotions in speech; to find out which features of speech signal could be useful for emotion recognition; and to explore different mathematical models for creating reliable recognizers.

2.1. Corpus of Emotional Data

To achieve the first objective we had to create and evaluate a corpus of emotional data. We used high quality speech data for the corpus. Thirty subjects recorded the following four short sentences: *"This is not what I expected."; "I'll be right*

there."; "Tomorrow is my birthday."; and "I'm getting married next week." Each sentence was recorded five times; each time, the subject portrayed one of the following emotional states: happiness, anger, sadness, fear and normal (unemotional or neutral). Five subjects have recorded the sentences twice with different recording parameters. Thus, a corpus of 700 utterances was created with 140 utterances per emotional state. Each utterance was recorded using a close-talk microphone; the first 100 utterances were recorded at 22-kHz/8 bit and the rest 600 utterances at 22-kHz/16 bit.

2.2. People Performance

We designed experiments to find the answers to the following questions:

- How well can people without special training portray and recognize emotions in speech?
- How well can people recognize their own emotions that they recorded 6-8 weeks earlier?
- Which kinds of emotions are easier/harder to recognize?

An interactive program was implemented that selected and played back the utterances in random order and allowed a user to classify each utterance according to its emotional content. Twenty-three subjects took part in the evaluation stage, and 20 of whom had participated in the recording stage earlier. Table 1 shows the people performance confusion matrix. The rows and the columns represent true and evaluated categories respectively, for example, second row says that 11.9 % of utterances that were portrayed as happy were evaluated as normal (unemotional), 61.4 % as true happy, 10.1 % as angry, 4.1% as sad, and 12.5 % as afraid. We can also see that the most easily recognizable category is anger (72.2%) and the least easily recognizable category is fear (49.5%). The mean accuracy is 63.5 % that agrees with the results of the other experimental studies [3,4].

Table 1: People Performance Confusion Matrix

Category	Normal	Happy	Angry	Sad	Afraid
Normal	**66.3**	2.5	7.0	18.2	6.0
Happy	11.9	**61.4**	10.1	4.1	12.5
Angry	10.6	5.2	**72.2**	5.6	6.3
Sad	11.8	1.0	4.7	**68.3**	14.3
Afraid	11.8	9.4	5.1	24.2	**49.5**

Table 2 shows statistics for evaluators for each emotional category. We can see that the variance for anger and sadness is significantly less than for the other emotional categories. It means that people better understand how to express/decode anger and sadness than other emotions.

Table 2: Evaluators' statistics

Category	Mean	s.d.	Median	Minimum	Maximum
Normal	66.3	13.7	64.3	29.3	95.7
Happy	61.4	11.8	62.9	31.4	78.6
Angry	72.2	5.3	72.1	62.9	84.3
Sad	68.3	7.8	68.6	50.0	80.0
Afraid	49.5	13.3	51.4	22.1	68.6

Table 3 shows statistics for "actors", i.e. how well subjects portrayed emotions. Speaking more precisely, the numbers in the table show which portion of portrayed emotions of a particular category was recognized as this category by other subjects. It is interesting to see comparing tables 2 and 3 that the ability to portray emotions (total mean is 62.9%) stays approximately at the same level as the ability to recognize emotions (total mean is 63.2%), but the variance for portraying is much larger.

Table 3: Actors' statistics

Category	Mean	s.d.	Median	Minimum	Maximum
Normal	65.1	16.4	68.5	26.1	89.1
Happy	59.8	21.1	66.3	2.2	91.3
Angry	71.7	24.5	78.2	13.0	100.0
Sad	68.1	18.4	72.6	32.6	93.5
Afraid	49.7	18.6	48.9	17.4	88.0

Table 4 shows self-reference statistics, i.e. how well subjects recognize their own portrayals. We can see that people do much better (but not perfect!) in recognizing their own emotions (mean is 80.0%), especially for anger (98.1%), sadness (80.0%) and fear (78.8%). Interestingly, fear was recognized better than happiness. Some subjects failed to recognize their own portrayals for happiness and the normal state.

Table 4: Self-reference statistics

Category	Mean	s.d.	Median	Minimum	Maximum
Normal	71.9	25.3	75.0	0.0	100.0
Happy	71.2	33.0	75.0	0.0	100.0
Angry	98.1	6.1	100.0	75.0	100.0
Sad	80.0	22.0	81.2	25.0	100.0
Afraid	78.8	24.7	87.5	25.0	100.0

From the corpus of 700 utterances we selected five nested data sets, which include utterances that were recognized as portraying the given emotion by at least p per cent of the subjects (p = 70, 80, 90, 95, and 100%). We will refer to these data sets as *s70*, *s80*, *s90*, *s95*, and *s100*. Table 5 shows the number of elements in each data set. We can see that only 7.9% of the utterances of the corpus were recognized by all subjects. And this number lineally increases up to 52.7% for the data set *s70*, which corresponds to the 70%-level of concordance in decoding emotion in speech.

Table 5: p-level concordance data sets

Data set	s70	s80	s90	s95	s100
Size	369	257	149	94	55
	52.7%	36.7%	21.3%	13.4%	7.9%

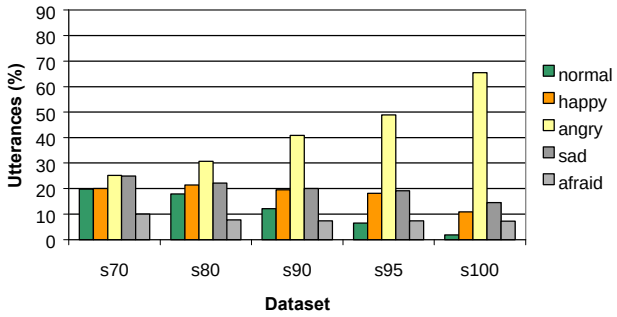

Figure 1: Emotion distributions for the data sets.

Figure 1 presents distributions of utterances among the emotion categories for the data sets. We can notice that it is close to the uniform distribution for *s70* with ~20% for the normal state and happiness, ~25% for anger and sadness, and 10% for fear. But for the data sets with higher level of concordance anger begins to gradually dominate while the proportion of the normal state, happiness and sadness decreases. Interestingly, the proportion of fear stays approximately at the same level (~7-10%) for all data sets. The above analysis suggests that anger is not only easier to portray and recognize but it is also easier to come to a consensus about what anger is.

3. DEVELOPMENT

3.1. Feature extraction

All studies in the field point to the pitch (fundamental frequency F0) as the main vocal cue for emotion recognition. The other acoustic variables contributing to vocal emotion signaling are [8]: vocal energy, frequency spectral features, formants, 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 LPC (linear predictive coding) parameters of signal [7] or features of the smoothed pitch contour and its derivatives [6].

For our study we calculated some descriptive statistics for the following acoustical variables: fundamental frequency F0, energy, speaking rate, first three formants (F1, F2, and F3) and their bandwidths (BW1, BW2, and BW3). Then we ranked them 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, i.e. the line that fits the pitch contour. 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 [9] for feature selection. The top 14 selected 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 (from F0 maximum to speaking rate), the second set extends the first one by two next features (F0 slope and F1 maximum), and the third set includes all 14 top features.

3.2. Computer Performance

To build emotion recognizers we tried the following approaches: *K*-nearest neighbors; neural networks; ensembles of neural network classifiers, set of neural network 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 used 70% of the *s70* data set as a database of cases for comparison and 30% as a test set. The average results for *K* from 1 to 15 and for number of features 8, 10, and 14 were calculated. It turns out that 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. The number of inputs corresponds to the number of features and the number of outputs corresponds to the number of emotional categories. To train and test our algorithms we used the data sets *s70*, *s80* and *s90*. These sets were randomly split into training (70% of utterances) and test (30%) subsets. We created several neural network classifiers trained with different initial weight matrices. This approach applied to the *s70* data set and the 8-feature set 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. Figure 2 shows the average accuracy of recognition for ensembles of 15 neural networks, the *s70* data set, all three sets of features, and both neural network architectures (10 and 20 neurons in the hidden layer). We can see that the accuracy for happiness stays the same (~65%) for the different sets of features and architectures. The accuracy for fear is relatively low (35-53%). The accuracy for anger starts at 73% for the 8-feature set and increases to 81% the 14-feature set. The accuracy for sadness varies from 73% to 83% and achieves its maximum for the 10-feature set. The average total accuracy is about 70%.

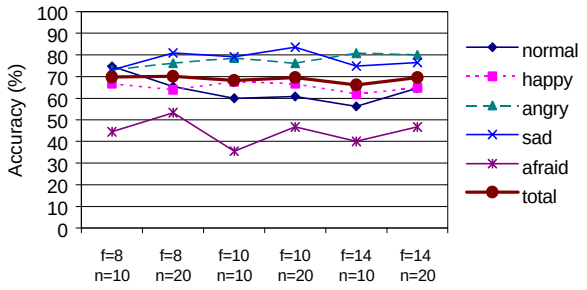

Figure 2: Accuracy of emotion recognition for the *s70* data set.

Set of experts. The last approach uses 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. We also used the same subsets of the *s70* data set as the training and test sets but with only two classes (for example, angry – non-angry). The average accuracy of recognition for the expert neural networks is about 70% 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 rule is to choose the class which expert's value is closest to 1. 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. The first approach gave the total accuracy about 60% for the 10-neuron architecture and about 53% for the 20-neuron architecture. The second approach gave the total accuracy about 63% for both architectures. Thus, it turned out that the accuracy of expert recognizers was not high enough to increase the overall accuracy of recognition. The approach, which is based on ensembles of neural network recognizers, outperformed the others and was chosen for implementation emotion recognition agents at the next stage.

4. APPLICATION

The following pieces of software have been developed: Emotion Recognition Game (*ERG*), Emotion Recognition software for call centers (ER), and a dialog emotion recognition program (*SpeakSoftly)*. 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, which just adds a different user interface to the core of the ER system, demonstrates the real time emotion recognition.

4.1. 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. 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.

4.2. Emotion Recognition Software for Call Centers

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 described above. We created the agents that include ensembles of 15 neural network. The average accuracy is 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 prioritizer 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 emotion content files, which describe the distribution of emotions in the message. The prioritizer is an agent that reads the emotion content files, sorts messages taking into account their emotional content, length and some other criteria, and suggests an assignment of a human agent 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 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.

5. REFERENCES

1. Darwin, Ch. The expression of the emotions in man and animals. Chicago: University of Chicago Press, 1965 (Original work published in 1872).

2. Picard, R. *Affective computing*. The MIT Press. 1997.

3. Bezooijen, R. van *The characteristics and recognizability of vocal expression of emotions.* Drodrecht, The Netherlands:Foris, 1984.

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

5. Murray, I.R. and Arnott, J.L. 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.

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

7. Tosa, N. and Nakatsu, R. Life-like communication agent - emotion sensing character "MIC" and feeling session character "MUSE". *Proceedings of IEEE Conference on Multimedia 96,* 12-19, 1996.

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

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