How well can People and Computers Recognize Emotions in Speech?

Valery A. Petrushin

Center for Strategic Technology Research
Andersen Consulting
3773 Willow Rd., Northbrook, IL 60062, USA
petr@cstar.ac.com

Abstract

The paper describes an experimental study on vocal emotion expression and recognition. Utterances expressing five emotions - happiness, anger, sadness, fear, and normal (unemotional) state - portrayed by thirty non-professional actors were recorded and then replayed to them for evaluation. The results on decoding emotions are consistent with earlier findings. The results on ability of humans to portray emotions and to recognize their own emotions in speech are presented. Computer algorithms for recognizing emotions in speech were developed and tested. Their accuracy is presented.

Introduction

This study explores how well both people and computers in recognizing emotions in speech. Although the first monograph on expression of emotions in animals and humans was written by Charles Darwin in the last century and psychologists have gradually accumulated knowledge in this field, it has attracted a new wave of interest recently by both psychologists and artificial intelligence specialists. There are several reasons for this renewed interest: technological progress in recording, storing, and processing audio and visual information; the development of non-intrusive sensors; the advent of wearable computers; and the urge to enrich human-computer interface from point-and-click to sense-and-feel. 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, 1984; van Bezooijen 1984; Scherer et al., 1991)). On the other hand, AI researchers made contributions in the following areas: emotional speech synthesis (Canh, 1989; Murray and Arnott, 1993), recognition of emotions (Dellaert et al., 1996), and using agents for decoding and expressing emotions (Tosa and Nakatsu, 1996).

Motivation

The 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.

Given this orientation, for this study:

- We solicited data from people who are not professional actors or actresses.
- Our focus is on the negative emotions like anger, sadness and fear.
- We target the telephone quality speech (< 3.4 kHz).
- We rely on voice signal only. This means we exclude the modern speech recognition techniques, which require much better quality of signal and computational power.

Data collecting and evaluating

We have asked thirty of our colleagues to record four short sentences:

- *"This is not what I expected."*
- *"I'll be right there."*
- *"Tomorrow is my birthday."*
- *"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). Five subjects have recorded the sentences twice with different recording parameters. Thus, each subject has recorded 20 or 40 utterances, yielding a corpus containing 700 utterances 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.

After creating the corpus, we designed an experiment 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?

One important result of the experiment was a set of selected most reliable utterances, i.e. utterances that were recognized by the most people. This set we intended to use as training and test data for pattern recognition algorithms running by a computer.

We implemented an interactive program which 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 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 fear. We can also see that the most easily recognizable category is anger (72.2%) and the least easily recognizable category is fear (49.5%). A lot of confusion is going on between sadness and fear, sadness and unemotional state, and happiness and fear. The mean accuracy is 63.5 % that agrees with the results of the other experimental studies (Scherer, 1984; Scherer et al., 1991, van Bezooijen 1984).

Table 1. Performance Confusion Matrix

Category	Normal	Happy	Angry	Sad	Afraid	Total
Normal	66.3	2.5	7.0	18.2	6.0	100 %
Happy	11.9	61.4	10.1	4.1	12.5	100 %
Angry	10.6	5.2	72.2	5.6	6.3	100 %
Sad	11.8	1.0	4.7	68.3	14.3	100 %
Afraid	11.8	9.4	5.1	24.2	49.5	100 %

Table 2 shows statistics for evaluators for each emotional category and for summarized performance that was calculated as the sum of performances for each category. We can see that the variance for anger and sadness is much less then for the other emotional categories.

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
Total	317.7	28.9	314.3	253.6	355.7

Table 3 shows statistics for "actors", i.e. how well subjects portray 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
Total	314.3	52.5	315.2	213.0	445.7

Table 4 shows self-reference statistics, i.e. how subjects were good in recognizing 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
Total	400.0	65.3	412.5	250.0	500.0

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 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.

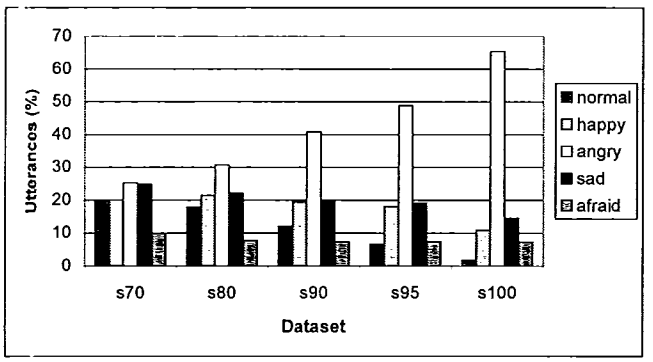

Figure 1. Emotion distributions for the data sets.

Although these results should be considered preliminary, they give us a valuable insight about human performance and can serve as a baseline for comparison to computer performance.

Feature extraction

All studies in the field point to the pitch as the main vocal cue for emotion recognition. Strictly speaking, the pitch is represented by the fundamental frequency (F0), i.e. the main (lowest) frequency of the vibration of the vocal folds. 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, F2) are considered).
- 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 (Tosa and Nakatsu, 1996) or features of the smoothed pitch contour and its derivatives (Dellaert et al., 1996).

For our study we adopted the following strategy. First, we took into account fundamental frequency F0, energy, speaking rate, first three formants (F1, F2, and F3) and their bandwidths (BW1, BW2, and BW3) and calculated for them as many statistics as we can. 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, 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 (Kononenko, 1994) for feature selection. We ran RELIEF-F for the s70 data set varying the number of nearest neighbors from 1 to 12, and ordered features according their sum of ranks. 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 (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.

Computer performance

To recognize emotions in speech we tried two approaches: neural networks and ensembles of classifiers. 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 (67% of utterances) and test (33%) 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 55% with the following distribution for emotional categories: normal state is 40-50%, happiness is 55-65%, anger is 60-80%, sadness is 60-70%, and fear is 20-40%.

For the second approach we used ensembles of 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 (Breiman, 1996) and cross-validated committees (Parmanto, Munro, and Doyle, 1996) 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 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 (~68%) for the different sets of features and architectures. The accuracy for fear is rather low (15-25%). The accuracy for anger is relatively low (40-45%)

for the 8-feature set and improves dramatically (~65%) for the 14-feature set. But the accuracy for sadness is higher for the 8-feature set than for the other sets. The average accuracy is about 55%. The low accuracy for fear confirms the theoretical result which says that if the individual classifiers make uncorrelated errors at rates exceeding 0.5 (it is 0.6-0.8 in our case) then the error rate of the voted ensemble increases (Hansen and Salomon, 1990).

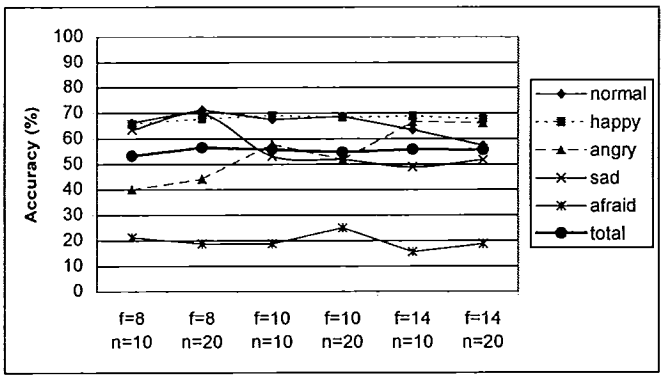

Figure 2. Accuracy of emotion recognition for the s70 data set.

Figure 3 shows results for the s80 data set. We can notice that the accuracy for normal state is low (20-30%). The accuracy for fear changes dramatically from 11% for the 8-feature set and 10-neuron architecture to 53% for the 10-feature and 10-neuron architecture. The accuracy for happiness, anger and sadness is relatively high (68-83%). The average accuracy (~61%) is higher than for the s70 data set.

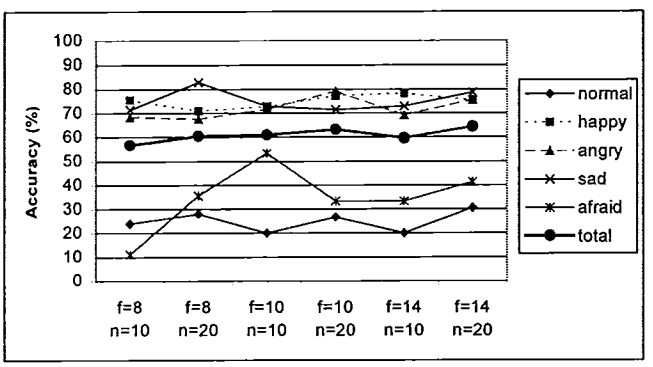

Figure 3. Accuracy of emotion recognition for the s80 data set.

Figure 4 shows results for the s90 data set. We can see that the accuracy of the normal state is very low (0-4%). The accuracy for fear is higher (25-60%) but it follows the same pattern shown for the s80 data set. The accuracy for sadness and anger is very high: 75-100% for anger and 88-93% for sadness. The average accuracy (~62%) is approximately equal to the average accuracy for the s80 data set.

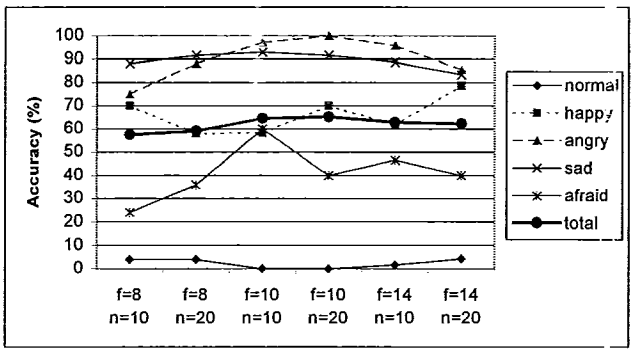

Figure 4. Accuracy of emotion recognition for the s90 data set.

To demonstrate the results of the above research an emotion recognition game has been developed. The program allows a user to compete against the computer or another person to see who can best recognize emotion in recorded speech. One potential practical application of the game is to help autistic people in developing better emotional skills at recognizing emotion in speech.

Future work

In our research we explored how well people and computers recognize emotions in speech. The first obtained results look rather promising, but we still have work to do before we can suggest something useful for business. That is why we plan to explore the other pattern recognition techniques and neural network architectures. We plan to investigate:

- How the quality of speech influences the accuracy of the classifiers.
- How classifiers work for real telephone quality data.

We also plan to develop a real-time version of the emotion recognizer in speech.

Acknowledgements

The author thanks to Anatole Gershman and Larry Birmbaum for fruitful discussions at the early stages of the project; Joe McCarthy and Douglas Bryan for the suggestions on an earlier version of this paper. Thanks also go to the many colleagues who participated in emotional data recording and evaluation. This research is conducted as a part of the Technology Exploration Project at the Center for Strategic Technology Research (CSTaR) at Andersen Consulting.

References

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

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

Breiman, L. Bagging Predictors. *Machine Learning* 24 (2): 123-140, 1996.

Canh, J.E. Generation of Affect in Synthesized Speech. In Proceedings of AVIOS'89, Meeting of the American Voice Input/Output Society, 1989.

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

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

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.

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.

Parmanto, B., Munro, P.W., and Doyle, H.R. Improving committee diagnosis with resampling techniques, In D.S. Touretzky, M.C. Mozer, and M. Hesselmo (eds.) *Advances in Neural Information Processing Systems 8.* Cambridge, Mass.: MIT Press, 882-888, 1996

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

Scherer, K.R. 1984. Speech and emotional states. In: J.K. Darby (ed.) *Speech evaluation in psychiatry.* New York: Grune & Stratton, 1984:189-220.

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

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