Recognition of Emotions in Interactive Voice Response Systems

Sherif Yacoub, Steve Simske, Xiaofan Lin, John Burns
Hewlett-Packard Laboratories, 1501 Page Mill Rd. MS 1126
Palo Alto, CA 94304
{sherif.yacoub,steven.simske,xiaofan.lin,john.burns}@hp.com

Abstract

This paper reports emotion recognition results from speech signals, with particular focus on extracting emotion features from the short utterances typical of Interactive Voice Response (IVR) applications. We focus on distinguishing anger versus neutral speech, which is salient to call center applications. We report on classification of other types of emotions such as sadness, boredom, happy, and cold anger. We compare results from using neural networks, Support Vector Machines (SVM), K-Nearest Neighbors, and decision trees. We use a database from the Linguistic Data Consortium at University of Pennsylvania, which is recorded by 8 actors expressing 15 emotions. Results indicate that hot anger and neutral utterances can be distinguished with over 90% accuracy. We show results from recognizing other emotions. We also illustrate which emotions can be clustered together using the selected prosodic features.

1. Introduction

The recognition of emotion in human speech has gained increasing attention in recent years due to the wide variety of applications that benefit from such technology. Although human emotions are hard to characterize and categorize [9], research on machine understanding of human emotions is rapidly advancing.

Emotion recognition solutions depend on which emotions we want a machine to recognize and for what purpose. Emotion recognition has applications in talking toys, video and computer games, and call centers. We are particularly interested in the application of emotion recognition technologies for Interactive Voice Response (IVR) systems with specific application to call centers. An obvious example is the automatic call routing of angry customer to agents (customer representatives) and the automatic quality monitoring of agents performance. These systems are conversational and hence utterances are usually short.

In this paper we present results for emotion recognition in IVR applications. Our studies can be characterized by the following properties. First, *speaker-independence*, we test the proposed emotion classifiers using unseen samples from unseen speakers to ensure speaker independence. Second, *no transcription required*, meaning emotion recognition occurs before automatic speech recognition output is obtained. Third, *IVR specific*, meaning we calculate emotion features from the short utterances typical of IVR systems. Fourth, *prosodic features*, meaning in addition to pitch contour and energy contour features, we use features from the audible/inaudible contour as explained later.

2. Emotion Recognition

Automatic emotion recognition of speech can be viewed as a pattern recognition problem [2,10]. The results produced by different experiments is characterized by: a) the *features* that are believed to be correlated with the speaker's emotional state, b) the type of *emotions* that we are interested in; c) the *database* used for training and testing the classifier; and d) the type of *classifier* used in the experiments. To compare classification results, we must use the same dataset and agree on the set of emotions. The purpose of this section is not to compare results reported in earlier research but instead to review briefly techniques used in emotion recognition.

Dellaert *et al.* [4] compared three classifiers: the maximum likelihood Bayes classification, kernel regression, and k-nearest neighbor (K-NN) methods with particular interest in sadness, anger, happiness, and fear. They used features from the pitch contour. An accuracy of 60%-65% was achieved. Lee *et.al.* [6] used linear discrimination, k-NN classifiers, and support vector machines (SVM) to distinguish two emotions: negative and non-negative emotions where they reached a maximum accuracy of 75%. Petrushin [10] developed a real-time emotion recognizer using neural networks for call center applications, and achieved 77% classification accuracy in two emotions ("agitation" and "calm") using eight features chosen by a feature selection algorithm.

Tato *et.al.* [13] discussed techniques that exploit emotional dimension other than prosody. Their experiments showed how "quality features" (based on formant analysis) are used in addition to "prosody features" (pitch and energy) to improve the classification of multiple emotions. The quality features were mostly speaker-dependent and hence cannot be used in IVRs. Yu *et.al.* [15] used SVMs for emotion detection. They built classifiers for four emotions: anger, happy, sadness, and neutral. Since SVMs are binary classifiers, their recognizers worked on detecting one emotion versus the rest. An average accuracy of 73% was reported.

3. Database

The performance of an emotion classifier relies heavily on the quality of the database used for training and testing and its similarity to real world samples (generalization). Speech data used for testing emotion recognition can be grouped under three categories depending on the way the speech signal was captured. The first method uses actors to record utterances, where each utterance is spoken with multiple feigned emotions. The actors are usually given the time to imagine themselves into a specific situation before speaking. The second method called Wizard-Of-Oz (WOZ) uses a program that interacts with the actor and drives him into a specific emotion situation and then records his responses. The third method, which is hard to obtain, is actual real-world recording of utterances that express emotions.

In our experiments, we used a database from the Linguistic Data Consortium, University of Pennsylvania [14]. The original data set has 9 hours of English recordings in sphere format and their transcripts. The dataset is encoded in

2-channel interleaved 16-bit PCM. Each speech file is a continuous recording of several emotions from one speaker.

We developed a splitter component that takes these recordings and the associated transcripts and emits separate utterances and their transcripts. Each utterance file represents one utterance by one actor expressing one emotion. As a result we obtain a set of 2433 utterances roughly distributed over fifteen emotions: neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, and contempt. These are short utterances of 3-4 words each 16-bit PCM, 22.05 KHz, and one channel. The utterances are spoken by 8 actors, mostly in the mid-20s with five females and three males.

4. Feature Extraction

In IVRs, utterances spoken by the caller are often short since they are responses to specific system prompts or selections from available menu options. Therefore, we focus on utterance-level features as opposed to word-level features. Global acoustic features calculated at the whole utterance level seem to have the favor of many recent studies [4,11]. We also perform emotion recognition at the signal level regardless of any information obtained from a speech recognizer.

In our experiments, each utterance is split into frames of size 384 samples and a window step of 192 samples where the sampling rate is 22.05 KHz. We calculate 37 prosody features related to pitch (fundamental frequency), loudness (energy), and segments (audible durational) as follows.

4.1. Fundamental frequency features

We obtain the pitch contour for the input utterance; the pitch values of each frame as a function of time excluding unvoiced frames. We calculate a total of 12 pitch features.

- *The pitch contour.* We obtain the minimum, maximum, mean, and standard deviation.
- *The first derivative of the pitch contour.* We obtain the minimum, maximum, mean, and standard deviation.
- *The jitter.* Jitter is defined in [5] as " *the small frequency changes and modulation of a signal."* To calculate jitter we use a linear filter system similar to the one used in [7]. In this system, the sequence of pitch values are provided as input to a filter g and then normalized by the frame pitch value. We use the following filter g= ¼ {-1, 3, -3, 1}. From the jitter contour, we calculate the mean, minimum, maximum, and standard deviation.

4.2. Energy features

We obtain the energy contour for input utterance; the energy for each frame in the utterance as a function of time. We calculate a total of 12 energy features.

- *The energy contour.* We obtain the minimum, maximum, mean, and standard deviation.
- *The first derivative of the energy contour.* We obtain the minimum, maximum, mean, and standard deviation.
- *The shimmer.* Shimmer is similar to jitter but based on the energy contour rather than the pitch contour. We calculate the mean, minimum, maximum, and standard deviation for the shimmer contour.

4.3. Audible durational features

We obtain a set of features related to audible segments per utterance. To identify audible segments in a speech utterance, we first obtain the maximum frame energy in the utterance and then consider any frame whose energy level is below a threshold (percentage of the maximum energy) as an inaudible frame; else it is audible. We picked the threshold to be 1%. The audible/inaudible contour is then filtered using a low pass filter to smooth the segments. As a result, an audible and inaudible segments contour is produced. Inaudible segments are related to pauses in the speaker utterance. From this contour, we obtain the following 13 features:

- Minimum, maximum, mean, and standard deviation duration of audible segments.
- Minimum, maximum, mean, and standard deviation duration of inaudible segments.
- Ratio of total duration of audible segments to total duration of inaudible segments.
- Ratio of the duration of audible segments to the total duration of an utterance.
- Ratio of the duration of inaudible segments to the total duration of an utterance.
- Ratio of average duration of audible segment to average duration of inaudible segments.
- The number of audible frames divided by the number of audible segments.

5. Classification and Results

Classifiers: We use: neural networks, SVMs (in case of binary classifications), 3-nearest neighbors, and in some cases the C4.5 decision tree [12]. The Weka toolkit (http://www.cs.waikato.ac.nz/~ml/weka/) was used for the experiments reported in this paper.

Training and testing data: We split the data into training and validation set and an unseen test set. The unseen test set is used to report accuracy measures. The LDC database was produced by 8 actors. We split the data such that unseen test set is actor-based; i.e. utterance from an actor used in training cannot be used in testing. In most of the experiment, we used 7 actor utterances for training and validation and eighth actor utterances for testing. In cases where we have enough training samples, we used 6 actor utterances for training and validation and the other two actors for testing. We believe that this data split gives us an indication of how the classifier is able to generalize (speaker independence). We call the first split 7/1 actor-split and the second split 6/2 actor-split.

Training and validation. We use 10-folds cross validation technique where the training data is randomly split into ten sets; nine of which are used in training and the tenth for validations then another nine is picked and so forth.

5.1. Recognizing hot anger and neutral

In the first set of experiments we are concerned with distinguishing anger from neutral speech. We use four classifiers: neural networks with the number of hidden layers equal to half the sum of the input and output nodes; SVM; 3-Nearest Neighbors; and C4.5 decision trees. We also experiment with two types of datasets 7/1 actor-split and 6-2 actor-split as mentioned earlier. In the 7/1 split, the training to testing ratio is 87% to 13% and in the 6/2 split the ratio is 70% to 30%. Table 1 summarizes the recognition accuracy when all 37 features are used. Table 2 illustrates the precision and recall of each of the classifiers.

From table 1, we find that the neural network classifier is better in the presence of sufficient training data while the SVM outperforms other classifiers in the scarcity of training data. From table 2 we conclude that when a classifier outperforms another classifier it usually does so in both precision and recall accuracy and in both emotions.

Table 1 Accuracy in recognizing hot anger versus neutral speech using 37 features

Classifier	7/1 actor-split	6/2 actor-split
Neural Net	94.00%	86.90%
SVM	90.90%	90.79%
3NN	87.88%	81.60%
C4.5	63.65%	76.32%

We then perform another experiments to identify the features that are most significant in the classification (feature selection). We used the *forward selection* feature selection algorithm to rank the features (using the training data and 10 fold cross validation). We then select the top significant features (19 features) which included: maximum minimum and mean of the pitch contour first derivative, maximum minimum and mean of jitter, and maximum energy among others. The above experiments are then repeated using just the 19 features. Table 3 illustrates the results on the testing set. The neural network classifier performance deteriorated by 3%, the SVM performance remained the same, and the 3-NN performance deteriorated by 11%.

Table 2 Precision and recall statistics

	Classifier		7/1 split	6/2 split
Hot Anger	*Precision*	Neural	0.875	0.917
		SVM	0.867	0.971
		3-NN	0.813	0.861
	Recall	Neural	1	0.825
		SVM	0.929	0.85
		3-NN	0.929	0.775
Neutral Emotion	*Precision*	Neural	1	0.825
		SVM	0.944	0.854
		3-NN	0.941	0.775
	Recall	Neural	0.895	0.917
		SVM	0.895	0.972
		3-NN	0.842	0.861

Table 3 Accuracy in recognizing hot anger versus neutral speech using the most significant 19 features.

Classifier	7/1 actor-split
Neural Net	91.00%
SVM	90.91%
3NN	76.15%
C4.5	72.73%

5.2. Recognizing hot/cold anger versus neutral/sadness

We created two emotion groups; the first contained hot anger and cold anger utterances and the second contained neutral and sadness utterances. We used the 6/2 actor split which yields a training/validation to testing ratio of 72% to 28%. Using the 37 features, table 4 summarizes the performance of three classifiers on the unseen test set. The best performance is obtained from SVM with 87% accuracy.

One interpretation of this still high classification accuracy between the two groups is due to the fact that the cold and hot anger are close to each other in the prosodic feature space (the 37 features we extracted). To confirm this, we conducted a binary classification experiment to distinguish cold anger from hot anger. We used the 7/1 actor split and SVM. The accuracy achieved in this case was 58%, which is close to random. Hence, the features we extracted in the previous section are not suitable for distinguishing cold and hot anger. We also conducted a binary classification experiment to distinguish neutral from sadness. We used the 7/1 actor split and SVM. The accuracy was 50%, which is a random guess. Hence, the features we extracted in the previous section are not suitable for distinguishing sadness and neutral emotion.

Table 4 Accuracy in recognizing (hot and cold anger) versus (neutral and sadness) speech using 37 features

Classifier	6/2 actor-split
Neural Net	82.7%
SVM	87.0%
3-NN	67.3%

We then used the forward selection algorithm to order the features relevance and picked the top five features which are: standard deviation of pitch, minimum and standard deviation for jitter, ratio of audible duration to inaudible duration, and the maximum energy. Table 5 summarizes the accuracy obtained using just the five features.

Table 5 Accuracy in recognizing group 1 (hot and cold anger) versus group 2 (neutral and sadness) speech using 5 features

Classifier	6/2 actor-split
Neural Net	82.7%
SVM	78.4%
3-NN	78.4%

For the neural network and 3-NN classifiers, it appears that the five features are sufficient. However, for the SVM, there is performance degradation associated with the features reduction.

5.3. Recognizing hot anger, sadness/neutral, and happiness

In this experiment we study the recognition of: hot anger, neutral and sad, and happy emotions, in a three-class classification problem. We use the 6/2 actor split which yielded a training and testing ratio of 72% to 28%. We used two classifiers: a neural network and a 3-NN.

Using a neural network classifier, the overall accuracy was 57% where the accuracy on happy is 47%, on hot anger 50%, and on neutral 77.5%. Using a 3-NN classifier, the overall accuracy was 58.8% where the accuracy on happy is 47%, on hot anger 30%, and on neutral 82.5%. Hence, we deduce that when happy signals are introduced the recognition of hot anger decreases significantly, which drives us to the following section.

5.4. What emotions are prosodically close to each other?

We run this experiment to understand what emotions are close to each other in the 37 features space. We select five emotions: hot anger, happiness, sadness, boredom, and neutral emotion. We use a C4.5 decision tree classifier. Table 6 illustrates the confusion matrix we obtained.

From this confusion matrix, we conclude that: sadness is mostly confused for boredom, boredom is mostly confused for sadness, happy is mostly confused for hot anger, hot anger is mostly confused for happy, and neutral is mostly confused with sadness (as reported in our previous experiments).

Table 6 Confusion matrix for five emotions: sadness, boredom, happy, hot anger, and neutral.

Sad	Bore	Happy	Anger	Neutral	<- as
0.42	**0.30**	0.16	0.04	0.09	Sad
0.28	0.36	0.18	0.05	0.13	Bore
0.15	0.12	0.44	**0.17**	0.12	Happy
0,02	0.08	**0.19**	0.70	0.01	Anger
0.16	**0.25**	0.20	0.02	0.37	Neutral

Hence, in classifying multiple emotions, one can group (happy and hot anger) and (sadness and boredom) and use separate features/classifiers across groups and within the same group. Similar results were reported in [13]. To support our results, we conducted an experiment where we grouped happiness with hot anger in one group and sadness with boredom in another group. We then ran binary classification experiments using SVM and C4.5 decision trees with a 6/2 actor split. For SVM, we obtained an overall accuracy of 77% with 71% accuracy on the (sadness and boredom) group and 82% on the (happy and anger) group. Using decision trees, we obtain an overall accuracy of 81.8% with 65.6% on the (sadness and boredom) group and 99% on the (happy and anger) group. Apparently, it seems easier to mistake a sadness/boredom motion for a happy/anger and not the other way around. We also used a forward selection algorithm to determine which features are most significant. We found that the following features are most significant in distinguishing the two groups: the maximum and mean of pitch contour first derivative, the maximum jitter, the minimum duration of inaudible segments, and ratio between audible segments duration and total duration of utterance.

5.5. Recognizing all 15 emotions

In this experiment, we considered the database in its entirety; i.e. all 15 emotions and all utterances. We use a neural network classifier and a 6/2 actor split which yielded a training to testing ratio of 73% to 27%. The accuracy obtained from such classification is 8.7%, above random 6.7%. It is a challenge (even to humans) to identify all the types of emotions defined in the selected database. . This indicates an IVR system should focus on the more readily identified emotions, such as anger and boredom.

6. Conclusions

In this paper, it is reported that using features extracted from pitch contour, energy contour, and audible segment duration contours we can achieve a high degree of accuracy in distinguishing certain emotions. We focused on utterance level features and short utterances, which is typical of IVR applications. A database from LDC consortium is used. We also compared different classifiers when applicable. We summarize our findings as follows:

- We are able to recognize hot anger versus neutral with accuracy exceeding 90%.
- For the given data, when a classifier has better accuracy than others it was notable that classifier also performs better in terms of both precision and recall accuracies.
- Hot and cold anger are not easily distinguished using the prosodic features discussed earlier. Similarly, sadness and neutral emotions are not easily distinguished.
- Five features (see section 5.2) are sufficient for distinguishing anger (hot and cold) from neutral and sad emotions with accuracy of 83%.
- The accuracy of classifying multiple emotions at the same time using the prosodic features discussed in this paper is low but still above random. Hierarchical classification and grouping of emotions is thus desirable.
- Anger and happiness are close to each other in the prosodic dimensions and hence classifiers often confuse one for the other. This also applies to sadness and boredom.

7. References

[1] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke A., "Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", in *Proc. of ICSLP-2002*, Denver, Colorado, Sept. 2002.

[2] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Noth, E., "Desperately Seeking Emotions: Actors, Wizards, and Human Beings", in *Proc. of the ISCA Workshop on Speech and Emotion*, the Queen's university of Belfast, Northern Ireland, Sept. 5-7, 2000.

[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G., "Emotion Recognition in Human-Computer Interaction", *IEEE Signal Processing Magazine*, 18(1), pp. 32-80, Jan 2001.

[4] Dellaert, F., Polzin, T., and Waibel, A., "Recognizing Emotion in Speech", in *Proc. of ICSLP 1996*, Philadelphia, PA, pp. 1970 -1973, 1996.

[5] Hess, W.: *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin 1983.

[6] Lee, C., Narayanan, S., and Pieraccini, R., "Classifying Emotions in Human-Machine Spoken Dialogs", in *Proc. of International Conference on Multimedia and Expo,* Lausanne, Switzerland, August 2002.

[7] Levit, M., Huber, R., Batliner, A., and Noeth, E., "Use of Prosodic Speech Characteristics for Automated Detection of Alcohol Intoxication", *ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. Red Bank, NJ October 22-24, 2001

[8] McGilloway *et.al.*, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", *ISCA Workshop on Speech and Emotion*, Belfast 2000.

[9] Ortony, A., Clore, G.L., and Collins, A.: *The Cognitive Structure of Emotions*, Cambridge Univ. Press, 1988.

[10] Petrushin, V., "Emotion in Speech: Recognition and Application to Call Centers", in *Proc. of Artificial Neural Networks in Engineering*, pp. 7-10, Nov. 1999.

[11] Petrushin, V., "Emotion Recognition in Speech Signal: Experimental Study, Development, and Application", in *Proc. of International Conference on Spoken Language Processing*, ICSLP 2000.

[12] Quinlan, J.R.: *C4.5: Programs for Machine Learning*, Morgan Kauffman, 1993.

[13] Tato, R., Santos, R., Kompe, R., Pardo, J.M., "Emotional Space Improves Emotion Recognition", in *Proc. of ICSLP-2002*, Denver, Colorado, September 2002.

[14] Linguistic Data Consortium, "Emotional Prosody Speech",www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28. *University of Pennsylvania*.

[15] Yu, F., Chang, E., Xu, Y.Q., and Shum. H.Y., "Emotion Detection From Speech To Enrich Multimedia Content", in the *Second IEEE Pacific-Rim Conference on Multimedia,* October 24-26, 2001, Beijing, China.