Automatic Spoken Affect Analysis and Classification

Deb Roy and Alex Pentland
MIT Media Laboratory
Perceptual Computing Group
20 Ames St.
Cambridge, MA 02129 USA
dkroy, sandy@media.mit.edu

Abstract

This paper reports results from early experiments on automatic classification of spoken affect. The task was to classify short spoken sentences into one of two affect classes: approving or disapproving. Using an optimal combination of six acoustic measurements our classifier achieved an accuracy of 65% to 88% for speaker dependent, text-independent classification. The results suggest that pitch and energy measurements may be used to automatically classify spoken affect but more research will be necessary to understand individual variations and how to broaden the range of affect classes which can be recognized. In a second experiment we compared human performance in classifying the same speech samples. We found similarities between human and automatic classification results.

1 Introduction

Spoken language carries many parallel channels of information which may roughly be divided into three categories: what was said, who said it, and how it was said. In the computer speech community the first two categories and the associated tasks of speech recognition and speaker identification have received a great deal of attention whereas the third category has received relatively little. We are studying a component of the third category; we are interested in automatically classifying affect through speech analysis. Although there has been much research to identify acoustic correlates of affect [14, 17, 3, 4, 12, 8], the authors are not aware of any previous work which attempts automatic classification of affect by explicitly modeling these acoustic features.

In this paper we report on initial experiments to determine useful acoustic features for automatic affect classification of speech. The task was to classify short sentences spoken with either approving or disapproving affect. Our experimental data consisted of recordings of three adult speakers who were asked to speak a set of sentences as if they were speaking to a young child. We extracted several acoustic features from the speech recordings which we expected would be correlated with affect. Standard pattern classification techniques were applied to measure the accuracy of automatic classification based on these features.

In a second experiment we conducted a human listening test on the same collected speech samples to compare human classification accuracy against the automatic analysis.

The motivation for this work is to move towards a speech interface which pays attention to all information in the speech stream. We are interested in exploring new types of interfaces which can detect and react to the emotion of the user [10]. We believe that an interface which listens to the user must detect all three types of information in the speech stream: not only what was said and who said it, but also how it was said.

2 Background

For the purposes of this paper we divide the “how it was said” channel of speech into two further categories. The first includes the prosodic effects which the speaker uses to communicate grammatical structure and lexical stress. Experiments suggests that this information is mainly carried in the fundamental frequency (F0) contour, the energy (loud/quiet) contour and phone durations [11, 9].

The second factor which effects the “how it was said” channel is the emotional or affective state of the speaker. Some of the commonly identified correlates of affect include average pitch, pitch range, pitch changes, energy (or intensity) contour, speaking rate, voice quality, and articulation [4, 8]. For example Williams and Stevens found correlations between anger, fear and sorrow with the F0 contour, the average speech spectrum and other temporal characteristics [17]. Scherer reports that the basic emotions can be communicated by pitch level and variation, energy level and variation, and speaking rate [14].

We note that the acoustic correlates which communicate affect overlap significantly with the acoustic correlates which communicate grammatical structure and lexical stress. Thus in principle it is impossible to completely separate the analysis of the two sources of variation, however in this paper we will treat effects due to affect in isolation.

Although the literature is consistent in which acoustic features are correlated with affect, the manner in which a given feature is adjusted to communicate a specific emotional state seems to be less clearly understood. Scherer found that although voice quality and F0 level convey affective information independent of verbal content, F0 contours could only be correctly interpreted by human listeners when the verbal content was also available [15]. This suggests that the F0 contour cannot be used for affect classification unless the text and grammatical structure of the spoken utterance is available. Streeter et. al. studied the pitch level of two speakers during the build up to a stressful event (the speakers were the systems operators on duty at the time of the 1977 New York blackout) [16]. Streeter found that as the situational stress increased one speaker's pitch level increased while the other speaker's pitch level decreased. This indicates that the manner in which acoustic features are correlated with affect is speaker dependent.

3 Data Collection

We made speech recordings of three adult native English speakers. The speakers were asked to imagine that they were speaking to young child and speak a set of sentences which were grouped into approving and disapproving sets. The subjects were given a printed list of sentences which they were asked to speak with pauses between each sentence. The sentence prompts were designed to convey a message of approval or disapproval without referring to any specific topic. Examples of approval include "That was very good" and "Keep up the good work". Examples of disapproval include "You shouldn't have done that" and "That's enough". There were 12 unique sentences for each affect class (a total of 24 sentences). The sentences were all short in duration (two to six words) since it is difficult for speakers to sustain consistent affect in long spoken utterances [2].

Each subject recorded 180 sentences (90 from each affect class). Each subject read 30 sentences from one affect class, then 30 from the other and repeated this cycle three times for a total of 180 sentences. Approximately one third of the samples were systematically discarded to remove effects of switching from one class to the other during recording. A few samples containing hesitations, coughing, and laughter were also removed. There were a total of 303 recordings from the three subjects in the final data set.

Recordings were made in a sound proof room with a Sony model TCD-D7 DAT recorder and an Audio Technica model AT822 stereo microphone. The DAT recordings were made in 16-bit 48 kHz sampled stereo. The automatic gain control in the DAT recorder was enabled during all recordings to ensure full dynamic range. The audio was then transferred to a workstation and converted to a 16-bit single channel 16 kHz signal.

4 Experiment 1: Automatic Affect Classification

4.1 Analysis

The acoustic features which we considered for performing automatic affect classification are summarized in Table 1. The features were chosen to be independent of the verbal content and sentence level prosodic structure.

Feature	Method
F0 (mean, variance)	Autocorrelation function
Energy (variance, derivative)	Short-time energy
Open Quotient	First and second Harmonic amplitude ratio
Spectral Tilt	Ratio of first harmonic to third formant

Table 1: Acoustic features used in the classification experiment.

All features are computed on 32ms frames of the signal. Adjacent frames overlap by 21ms.

The first two features are the mean and variance of the fundamental frequency (F0). The F0 is found by locating the peak of the autocorrelation function of the speech signal over each 32ms window [13]. The peak value is required to meet range constraints (70 to 450 Hz) and is smoothed using a median filter.

The third and fourth features are the variance and derivative of the short-time energy of the signal, computed over 32ms windows [13].

The fifth feature is the open quotient which is the ratio of the time the vocal folds are open to the total pitch period. This feature is estimated by the ratio of the amplitudes of the first two harmonics [6, 7].

The sixth feature is the spectral tilt which is estimated by the ratio of the amplitude of the first harmonic to the amplitude of the third formant [6, 7].

We chose not to use the absolute energy level as a feature since it is dependent on the exact recording configuration, and has been shown to contain little information about affect in perceptual tests.

The current implementation of our pitch tracker occasionally doubles or halves its F0 estimates which leads to very noisy time derivative measures. For this reason we have not included F0 changes as a feature.

With the exception energy, all features are computed only on voiced portions of the speech recordings. The voiced/unvoiced decision is made by computing the ratio of energy in the high and low frequency bands and multiplying this ratio by the short-time energy of the signal. By thresholding this measure we can detect segments of the signal with high energy and a high proportion of that energy present in the lower frequencies which are the characteristics of voiced speech spoken in a quiet environment.

4.2 Classification

As a first step we were interested in learning the discrimination ability of each feature in isolation. We built Gaussian probability models based on each feature for each affect class and used a likelihood ratio test with equal priors to classify test data.

Since the data set consisting of 303 sentences is relatively small we used cross-validation (also known as the hold-one-out method) for all classification experiments [1]. Cross-validation is performed by holding out a subset of data, building a classifier with the remaining labeled training data, and then testing the classifier on the held-out test set. A new set of test data is then held-out and the train-and-test cycle is repeated. This process proceeds until all data has been held out once. Errors are accumulated across all test sets.

In our case each held-out test data set consisted of all recordings corresponding to one text sentence. Thus the training data did not contain any occurrences of the sentence which the classifier was tested on. This assured that the classification results reflected text-independent performance.

Once we had tested classification performance using each feature in isolation, we used the Fisher linear discriminant method [5] to find an optimal combination of all six features. The Fisher method finds a linear projection of the 6-dimensional training data onto a one dimensional line which maximizes the interclass separation of the training data. We computed Gaussian statistics on the projected values of data for each class and used a likelihood ratio test with equal priors to classify test data.

4.3 Results

Table 2 presents the results of the classification experiments. The first three columns show classification accuracy for each speaker (F1 is female, M1 and M2 are male). The fourth column "All" is the pooled data from all three speakers. The first six rows show classification accuracy using each feature in isolation as described above. Note that random classification will lead to an expected accuracy of 50% for large data sets since this is a two class problem.

These results show that the most discriminative feature for each speaker is different. Speaker F1 uses large variations in speaking intensity in her disapproving speech and "smoother" speech to express approval. The result is high discrimination using the energy range feature. In contrast M1 relies most on changing the range of F0; when he speaks approvingly he uses a smaller F0 range than when he speaks disapprovingly. M2 relies most on the average pitch level; for approving utterances his average F0 is significantly higher than disapproving sentences in which he has a relatively low average F0.

In the combined case where data from all speakers were pooled we found that the best features were the average F0 and the open quotient measure. These two features are modified most consistently by all three speakers.

Feature	F1	M1	M2	All
Average F0	45	53	**82**	**64**
F0 Range	58	**67**	53	51
Energy range	**88**	57	67	61
Energy change	83	60	58	62
Open quotient	76	63	58	**64**
Spectral tilt	58	62	61	56
All features	**88**	**65**	**84**	**65**

Table 2: Classification results (percent correct) using Gaussian probability models for each feature in isolation (top six rows) and using an optimal combination of all six features (bottom row).

5 Experiment 2: Human Classification

In a second experiment we were interested to learn how well humans would perform in the affect classification task given the data we collected for the first experiment. In particular we were interested to learn if there would be a significant drop in accuracy for speaker M1 similar to the results from the automatic system. To do this we randomly selected one example of each approval and disapproval sentence from each speaker for a total of 70 sentences[1]. These recordings were grouped by speaker but the sequence within each speaker set was randomized (i.e. approval and disapproval sentences were randomly ordered for each speaker).

The recordings were played in reverse as a means of masking the verbal content of the sentences while retaining the voice quality and level and range characteristics of the pitch and energy. In effect many of the cues which a human would rely on such as the F0 contour and verbal content (which we did not use in the automatic classifier) were masked to make the human task comparable to the automatic task. However, all of the features which we did use in the automatic task were still present in the reversed audio. This masking technique has been used in similar perceptual tasks by Scherer [15]. Scherer observed that "the most serious potential artifact of reversed speech is the creation of a new intonation contour." [15]. Indeed some subjects in our experiment commented that they found themselves trying to interpret the F0 contour even though they knew it was reversed speech. We note that alternate methods of masking such as low-pass filtering and random splicing also have serious artifacts [15].

A simple graphical interface was created to present the reversed speech samples to subjects in sequence. The subject was asked to make a binary approval/disapproval classification for each sample. The subject was given control to go back and change the classification of samples as often as desired until he/she was satisfied.

Figure 1 shows the classification accuracy (percent correct) for seven native English speaking subjects. The accuracies averaged across all seven subjects for the three speakers F1, M1 and M2 were 76%, 69% and 74% respectively.

Note that the relative ordering of accuracy for the three speakers is consistent in the human and automatic experiments. This might mean that speaker M1's speech does not contain as many indicators of his affective state.

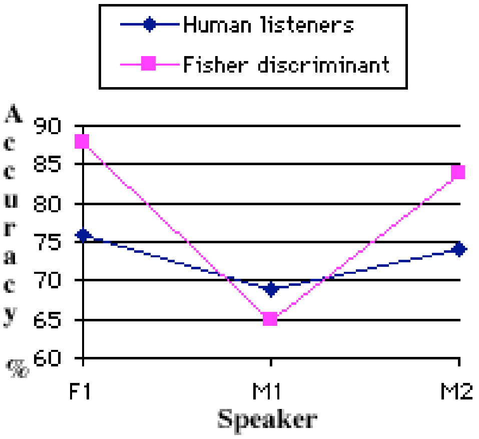

Figure 1: Human listening experiment results compared to automatic classification results (taken from Table 2).

Human listeners performed somewhat worse than the automatic classifier. One reason for this is that the automatic classifier was supplied with

[1] Due to an error during the design of this experiment only 22 sentences (rather than 24) from M1 were used. Thus the total number of sentences presented to each subject was 70 rather than 72.

labeled training data of each speaker. We expected that human listeners would be able to apply prior knowledge of acoustic correlates of emotion to do the classification task without training data. Artifacts of reversing the playback of recordings (noted earlier) probably contributed to the errors.

6 Conclusions and Future Work

The results of these initial experiments are promising. We achieved classification accuracies ranging from 65% to 88% (where random choice would lead to 50% accuracy) for three speakers. In a related experiment human listeners achieved classification accuracies ranging from 69% to 76%. The human listeners and the automatic classifier both had higher errors for the same speakers.

There are several short-time features which we can add to the present framework to potentially improve performance including speaking rate estimation, long term spectrum [12], and rate of change of F0.

Due to the limited scope of the experiments we cannot draw strong conclusions but the data suggests that energy and F0 statistics may be effectively used for automatic affect classification. This is in accord with previous findings in the psycholingistic community. However it is not clear how to deal with variations in individual speaking styles. We plan to collect data from more speakers to see if there are natural clusters of speaking styles in which case an affect classifier could first decide which cluster a speaker belongs to, and then apply the appropriate decision criteria.

It is likely that high accuracy in spoken affect classification will not be achieved without analysis of verbal content and sentence level prosodic cues such as the F0 contour. Our limited human verification task suggest that without this information, humans are not able to perform the task well either. Possible extensions of this work include integration with other sources of information which are obtained by speech recognition, speaker identification, and visual face analysis.

Acknowledgments

We thank Janet Cahn for countless helpful discussions and references, and to all the subjects who freely volunteered their time.

References

[1] Bishop, C.M. Neural networks for pattern recognition. Oxford University Press, 1995, 372-375.

[2] Cahn, J. Personal communication, 1996.

[3] Cahn, J.E. Generating expression in synthesized speech. Masters thesis, MIT Media Laboratory, May 1989.

[4] Cahn, J.E. Generation of affect in synthesized speech. Proc. of the 1989 conference of AVIOS, 251-256.

[5] Duda, R.O., and Hart, P.E. Pattern classification and scene analysis. John Wiley & Sons Inc., 1973.

[6] Hanson, H. Glottal characteristics of female speakers. Ph.D. thesis, Harvard University, Division of Applied Sciences, May 1995.

[7] Klatt, D.H., and Klatt, L.C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87 (2), February 1990, 820-857.

[8] Murray, I.R., and Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Am. 93 (2), February 1993, 1097-1108.

[9] O'Shaughnessy, D. Speech communication in human and machine. Addison-Wesley, 1987.

[10] Picard, R.W. Affective Computing. MIT Media Laboratory Perceptual Computing Section Technical Report No. 321.

[11] Pierrehumbert, J.B. The phonology and phonetics of English intonation. Ph.D. thesis, MIT, September, 1980.

[12] Pittam, J., Gallois, C., and Callan, V. The long-term spectrum and perceived emotion. Speech Communication 9 (1990) 177-187.

[13] Rabiner, L.R., and Schafer, R.W. Digital processing of speech signals. Prentice-Hall, 1978.

[14] Scherer, K.R., Koivumaki, J., Rosenthal, R. Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech. J. of Psycholinguistic Research, 1972, 269-285.

[15] Scherer, K. R., Ladd, R., and Silverman, K. Vocal cues to speaker affect: Testing two models. J. Acoust. Soc. Am., Vol. 76, No. 5, November 1984, 1346-1355.

[16] Streeter, L.A., Macdonald, N.H., Apple, W., Krauss, R.M., and Galotti, K.M. Acoustic and perceptual indicators of emotional Stress. J. Acoust. Soc. Am. 73 (4), April 1983, 1354-1360.

[17] Williams, W.E., and Stevens, K.N. Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Am. 52 (4) 1972, 1238-1250.