# TOWARDS AN AUTOMATIC CLASSIFICATION OF EMOTIONS IN SPEECH

*N. Amir, S. Ron*

Communications Engineering Department
Center for Technological Education Holon
52 Golomb st., Holon 58102, Israel.
noamoto@wine.cteh.ac.il

## ABSTRACT

In this paper we discuss a method for extracting emotional state from the speech signal. We describe a methodology for obtaining emotive speech, and a method for identifying the emotion present in the signal. This method is based on analysis of the signal over sliding windows, and extracting a representative parameter set. A set of basic emotions is defined, and for each such emotion a reference point is computed. At each instant the distance of the measured parameter set from the reference points is calculated, and used to compute a fuzzy membership index for each emotion, which we term the "emotional index". Preliminary results are presented, which demonstrate the discrimative abilities of this method when applied to a number of speakers.

## 1. INTRODUCTION

Although the main role of the voice is to communicate, voice is also an indicator of the psychological and physiological state of the speaker. The identification of the pertinent features in the speech signal may therefore allow the evaluation of the person's emotional state.

Among the emotional states, the most noticeable one identified long ago as reflected in subject's speech was stress. Previous studies concentrated on specific stress related correlates such as provoked anxiety [1], workload demand [2] and lie detection [3]. It was reported for example, that the fundamental frequency (pitch), and vocal intensity (loudness) increased significantly with workload demands and were robust in showing differences for individual subjects.

In a recent review on human vocal emotion, Murray and Arnott [5] suggested that it possible to build up from the fragmented reports an acoustical picture of the human voice during the expression of a number of emotions. The emotions most commonly accepted as "basic" (happiness, sadness, anger, fear and disgust), also considered to be the primary emotions, are the emotions most studied. As one variable may exhibit the same behaviour for more than one emotion, it is clear that only an interaction of several variables can discriminate reliably between emotions. It is only recently [6][7] that quantitative attempts were made to analyze the speech signal for indications of emotion.

The overall aim of the present study was to develop a speech analysis algorithm that can analyze the subject's emotions in real time, so an indication can be continuously obtained regarding the degree to which the subject is experiencing each of the five emotional states. Fulfilling this objective requires finding the answers to several basic questions, namely: Is there a set of acoustical parameters that reflects a subjects emotion reliably and consistently? Do these parameters behave consistently for a single subject?

To this end, we define two procedures: a data collection procedure, and an automatic classification procedure. In this paper we describe the method applied to collecting a database of emotional utterances in a manner reliable as possible. A certain degree of validation is obtained by measuring several physiological variables having a known correlation to emotion. We present an algorithm used to analyze the speech signals and classify their emotional content. The approach is different to those previously reported in the literature, in that the classification is not binary (i.e. angry/not angry, sad/not sad, etc.). Each emotion is graded by an "emotional index" between 0 and 1, so that different degrees of emotion can be detected. Finally we present a set of results obtained by this method.

## 2. DATA COLLECTION - MATERIALS AND METHODS

### 2.1. Subjects

Twenty four (12 males and 12 females) healthy students (age 23 to 31, mean 25.5) served in this study as subjects. The subjects were explained the experimental procedure and gave informed consent for the experiment. None of the subjects took any drug in the last three days preceding the experiment.

### 2.2. Physiological examination

The following physiological measures were assessed: Forehead electromyography, heart rate (HR) and Galvanic Skin Resistance (GSR).

### 2.3. Pretest training

The subjects were told that the aim of the experiment was to recall an emotional event and experience the same feelings that they felt when they participated in this event. They were asked to try to keep their eyes closed and recount an event while the physiological measurements were used as indicators to examine whether the subjects were indeed experiencing the described emotion. If the subject

did not show such physiological activity, he was asked to try again and get more involved in the recalled event. If a subject did not evoke the expected physiological response he was excused and did not participate in the experiment.

### 2.4. Procedure

The psychophysiologic assessment was initiated with the recording of baseline measures, followed by five consecutive script trials. They were asked to recall an event that they took part in and evoked one of the five emotions: happiness, anger, sorrow, fear or disgust.

Subjects were asked to keep their eyes closed and talk about the event, participate emotionally and feel the same feelings that they felt when it occurred.

## 3. ACOUSTICAL ANALYSIS

A careful review of the literature reveals that there is no agreed method of analyzing the speech data to give a reliable indication of the emotional state. A number of papers state some qualitative guidelines suggesting that the emotional state is indicated by pitch, intensity, pitch range, and less rigorously defined terms such as inflection, articulation, and speech rate. Only recently has there been work carried out in an attempt to quantify this type of analysis [6][7]. Though both of these papers attempt to characterize each of the primary emotions quantitatively, the classification is binary: either the emotion is present or not. The approach taken in this paper seems to parallel events more closely, in that each emotion is characterized by an emotional "index" that can occupy a certain well defined range, say between 0 and 1. In any case, the quality of the analysis clearly depends on the correct choice of parameters and the accuracy with which they are evaluated.

### 3.1. Examination of Speech parameters

Most of the parameters appearing in the literature are based on either the pitch or the intensity of the speech signal. Therefore these two are extracted from the signal as a fundamental starting point. The speech signal was sampled at 8 Khz, with 8 bits per sample. The pitch detection program was set up to provide 40 pitch values per second. Pitch analysis was performed on windows of 400 samples, with an overlap of 200 samples. Intensity was also calculated, and normalized for pitch period length.

### 3.2. The parameters used

The speech parameters used were based on [5], with certain variations, since the work in [5] presented only a qualitative description of the parameters. To enable continuous analysis, they were computed over sliding windows of between 2 to 6 seconds. These parameters are:

1. Pitch average: The average pitch in a window.
2. Pitch variance: pitch variance was used as a measure of pitch range in a window. This was found to be more useful than simpler measure of maximum pitch minus minimum pitch, since it gave much less weight to any spurious results in the pitch curve.
3. Jitter: This is a measure of tremor in the speech signal. Jitter was obtained by counting the number of changes in sign of the pitch derivative in a window.
4. Speech rate: Speech rate was estimated by the frequency of occurence of unvoiced periods, also inferred from the pitch signal.
5. Intensity: the relative intensity averaged over a window was calculated by summing the squares of the speech samples, for voiced segments only. The result was normalized to account for the relative length of voiced speech in the analysis window.
6. Intensity variance: as for pitch variance, over voiced segments only.
7. Tremor: this is a measure of tremor in the intensity rather than in the pitch. Calculated similary to jitter, but over the Intensity curve.

Each of the above parameters is initially computed on an absolute scale. In order to make comparisons between different speakers, it is necessary to normalize it with respect to a baseline value for each specific speaker. For this purpose each speaker must be recorded at a "neutral" emotional level, taken as the baseline. The measured values are then normalized as follows:

$$(normalized\ value) = \frac{(measured\ value) - (baseline\ value)}{(baseline\ value)} \quad (1)$$

### 3.3. Relating speech data to emotions

Based on reviews such as [5], the measured parameters are all correlated to a certain extent to the emotional state of the speaker. This correlation does not determine how all of the parameters together can be combined to obtain a single measure for each emotion, which we term an emotional index.

In [6] the authors adopted a binary approach to emotion classification, i.e. each utterance was defined as belonging or not belonging to a certain emotional class. In reality the emotional state is not necessarily a binary concept - an individual can be in one or several emotional states at the same time, and to a different degree for each emotion. Therefore, in this work we took the approach of measuring the degree of each emotion as an index on some arbitrary scale. Furthermore, In this work the classification was performed at a constant *rate* of 20 times per second, rather than on complete utterances.

#### *3.3.1. The DM method*

This method has a certain resemblance to that used in [6]. N-tuples (with N=7) in this case are calculated constantly at a rate of 20 per second from the speech signal. The distances of each N-tuple from 5 reference N-tuples (one for each emotion) are then calculated, and processed to obtain the emotional indices. It remains to determine:

1. how to find the reference points;
2. which distance measure gives the best results;
3. how to calculate the indices themselves.

The reference points were found by averaging the values for all of the N-tuples in a file or series of files judged to represent each emotion. This process introduces a certain inaccuracy, since it is not clear that the speakers are in the required emotional state during the entire utterance. Furthermore, in this method the reference point is judged to be the "ultimate" degree of each emotion; if for instance during a test utterance the speaker is *more* angry than in the reference utterance, the distance from the reference anger point will increase instead of decreasing.

The distance measure is obtained by first normalizing each parameter with respect to its variance and computing a Mahalanobis distance to each reference point. This is defined by:

$$d_M(X,Y) = \left[(X-Y)^T C_X^{-1}(X-Y)\right]^{1/2} \qquad (2)$$

where $X$ is the reference vector, $Y$ is the measurement vector, and $C_X$ is the covariance matrix of the measurements for the emotion. The emotional index is then calculated using the fuzzy membership index:

$$u_{ij} = \frac{\frac{1}{d_M^2(X_j,V_i)}}{\sum_j \frac{1}{d_M^2(X_j,V_i)}} \qquad (3)$$

where $X_j$ is the reference vector for the $j$'th emotion, $V_i$ is a measurement vector, and $u_{ij}$ is the $j$'th emotional index. This algorithm performs normalization with respect to the variance and takes correlation between the parameters into account. A further advantage to this method is that if we wish to decide which emotion is strongest at a given moment - choosing the emotion with the highest emotional index corresponds exactly to maximum likelihood estimation.

## 4. RESULTS

### 4.1. Data collection

Before any analysis was performed, the physiological data for each recording was examined. Four subjects (16 percent) exhibited no physiologically correlated activity during the event recount, and were discarded at the outset. In addition, eight subjects (16 percent) did not show changes in at least two event recounts. Based on peer judgement, our general impression was that even for some of the recordings which did show the expected physiological activity, in an unformal listening test, the speech signal was not always judged as reflecting the appropriate emotion.

All of this shows that the problem of obtaining reliable test data is not to be taken lightly. Some other studies use utterances produced by professional actors [8], usually without physiological verification. It remains to be seen which method is most reliable.

### 4.2. Quantitative results

In this stage, the entire analysis was performed on a single speaker at a time. In other words, the recordings of one speaker were first analyzed for pitch and intensity, from which the acoustic features were extracted. The reference points for each emotion were taken as their average

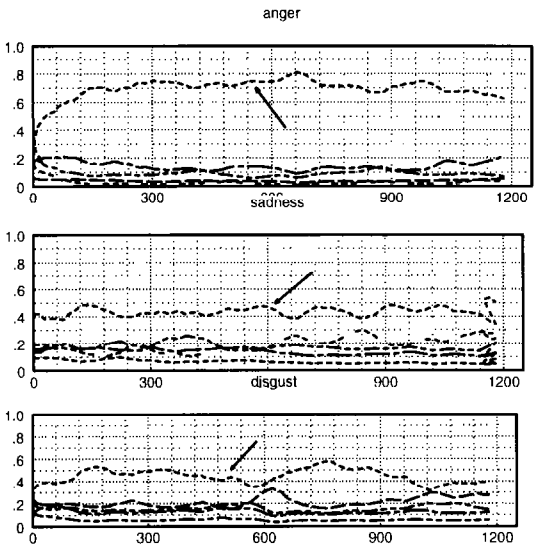


Figure 1: 3 examples of emotional indices

values over the appropriate file. The covariance matrices were also calculated from these files. Next, the emotional indices were computed as a function of time over the *same* files. The objective of this analysis was to examine the degree to which the acoustic features were consistent for each emotion, for a single speaker. The results for one speaker is presented in figures 1, where all the speech files examined were approximately 1 minute long. A temporal plot is shown for each utterance. In the plot, five lines are indicated, representing the continuous changes of each emotional index.

### 4.3. Emotion discrimination

One possibility is to use the emotional indices to obtain a global decision for each file, relating it to a specific emotion, as done by [6]A number of methods are possible: using the average value (AV) of these indices over the entire file, choosing the emotion to be the one who's index is largest. another possibility is measuring the percentage of time (TP) for which a specific index is $x$ points above the rest, where $x$ can be set arbitrarily, and it's units are the fuzzy membership units, between 0 and 1. Table 1 summarizes AV for two speakers:

High values on the diagonal indicate good discrimination. For each speaker the TP measure was performed three times, with $x$ chosen as 0.01, 0.1, 0.2. The results are summarized in table 2 for the same two speakers:

For these two speakers, AV proved to be a good overall discriminator. Using TP, the threshold $x$ determines the result to a large extent. The minimum threshold for which a visual decision would agree with the computed decision is around 10%. We do not attempt at this stage to determine any specific threshold and time percentage to be used as decisive values.

### 4.4. Discussion

This study, together with [6] demonstrate that automatic classification of emotions is possible. A review of the liter-

| | | | speaker 1 | | |
|---|---|---|---|---|---|
| utterance: → | anger | joy | sadness | fear | disgust |
| em. index ↓ | | | | | |
| ANGER | 56 | 20 | 7 | 9 | 8 |
| JOY | 11 | 53 | 11 | 14 | 10 |
| SADNESS | 4 | 13 | 49 | 11 | 23 |
| FEAR | 6 | 15 | 13 | 56 | 10 |
| DISGUST | 4 | 10 | 22 | 9 | 55 |
| | | | speaker 2 | | |
| utterance: → | anger | joy | sadness | fear | disgust |
| em. index ↓ | | | | | |
| ANGER | 34 | 19 | 11 | 19 | 17 |
| JOY | 19 | 38 | 15 | 15 | 12 |
| SADNESS | 12 | 15 | 43 | 17 | 13 |
| FEAR | 15 | 19 | 14 | 37 | 15 |
| DISGUST | 13 | 13 | 15 | 12 | 46 |

Table 1: AV for all utterances of two speakers

ature on the problem reveals that these two are nearly the only quantitative studies published on the subject. The current work brings to light a number of problems.

First, the automatic classification can only be as good as the reference data. This is a weak link, in the sense that it involves peer judgement. This makes it difficult to compare several different methods carried out by different researchers, on different databases. As in fields such as image or audio compression, a universal agreed database widely available would be very helpful in this case.

Second, creating a mapping between human judgement of emotion and an absolute scale is problematic. Psychologists have defined the basic "space" of emotions as composed of 4 or 5 basic emotions, but it is clear that there are no "pure" emotions and emotions can overlap or appear in various degrees.

Due to lack of data on these issues, a rather "engineering" approach was adopted, where we made a number of ad hoc assumptions that facilitated the analysis. The most successful method used in this work was based on two assumptions: The sum of the indices for all of the emotions is always one, due to the form of the fuzzy distance measure; and the reference points for each emotion represent the highest degree of that emotion.

The first assumption necessarily means that when an

| | speaker 1 | | |
|---|---|---|---|
| $x$: | .01 | .1 | .2 |
| ANGER | 100 | 100 | 97 |
| JOY | 100 | 99 | 95 |
| SADNESS | 99 | 95 | 84 |
| FEAR | 100 | 99 | 98 |
| DISGUST | 100 | 99 | 96 |
| | speaker 2 | | |
| $x$: | .01 | .1 | .2 |
| ANGER | 90 | 55 | 22 |
| JOY | 94 | 91 | 48 |
| SADNESS | 100 | 95 | 66 |
| FEAR | 100 | 88 | 26 |
| DISGUST | 100 | 98 | 87 |

Table 2: TP for all utterances of two speakers

index for one emotion grows, some others must shrink, which implies a negative type of coupling between emotions. This has not been verified or disproved on the psychological level, and it definitely merits further exploration. From the practical point of view it means that any of the emotional indices can receive a large value only when the parameters for that emotion correspond closely only to the reference for that emotion. If there exists a state in which an individual is both very disgusted and very afraid, at the most both indices can have an index of 0.5 at the same time.

The second assumption is actually related to the fact mentioned above that it is difficult to define an objective scale for subjective phenomena. Across a wide range of subjects. For this reason we believe the analysis performed well on an individual basis for each speaker, yet still proved problematic if applied across several subjects.

Despite the above assumptions and reservations, the bottom line is that the automatic classification agreed well with the human classification, over the range of emotions studied. Further work will probably be carried out in improving the set of speech parameters being used, in tuning the distance measure defining the speech indices, in reaching a better understanding of how emotions are expressed from a psychological point of view, and compiling a good database of recordings.

## 5. REFERENCES


[1] Fuller, B. F., Horii Y., Conner A. (1992). Validity and reliability of nonverbal voice measures as indicators of stressor-provoked anxiety. Res. Nurs. Health, 1992, 15, 379-389.

[2] Brenner, M. , Doherty, T., Shipp, S. (1994). Speech Measures Indicating Workload Demand. Aviat. Space Environ. Med., 65, 21-26.

[3] Hollien, H. Geison, L., Hicks, J. (1987). Voice stress evaluators and lie detection. 32, 405-418.

[4] Sherer, K. R., Vocal indicators of stress. In: Speech Evaluation in Psychiatry, J. Darby, Ed., Grune and Stratton, New York, pp. 171-187.

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

[6] Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotion in speech. ICSLP '96.

[7] Cowie, R., Douglas-Cowie, E.(1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. ICSLP '96.

[8] Engberg, I., Hansen A., Andersen O., Dalsgaard P., (1996). Design, recording and verification of a danish emotional speech database. ICSLP '96.