

# Prosody-based classification of emotions in spoken Finnish

*Tapio Seppänen, Eero Väyrynen and Juhani Toivanen*

MediaTeam Oulu, University of Oulu,

P.O. Box 4500, 4SOINFO, FIN-90014 University of Oulu, FINLAND

tapio.seppanen@ee.oulu.fi, eero.vayrynen@ee.oulu.fi, juhani.toivanen@ee.oulu.fi

## Abstract

An emotional speech corpus of Finnish was collected that includes utterances of four emotional states of speakers. More than 40 prosodic features were derived and automatically computed for the speech samples. Statistical classification experiments with kNN classifier and human listening tests indicate that emotion recognition performance comparable to human listeners can be achieved.

## 1. Introduction

The scientific study of the vocal expression of emotion is now reaching a level of maturity where the focus is on important applications, for example, those involving human-computer interaction [1] and audio search machines [2]. The vocal parameters of emotions have been extensively researched; see [3] for a general review of the literature. As it is now understood that affective computing may have an important industry potential, automatic recognition of emotions in speech has become a more attractive area of research. For example, the automatic recognition and classification of emotions and affect in speech, based on prosodic/acoustic features, could open up exciting new possibilities for content-based information retrieval from radio play databases.

The methods to extract prosodic data automatically, especially fundamental frequency (F0), are still lacking, although many algorithms have been proposed in the literature [4]. There has already been some research on the automatic recognition/classification of emotions in speech, mainly for such major languages as English and German [5], [6], [7]. Systematic data mining was used by [6], and some of the features utilized were: mean F0, max F0, min F0, range between max and min, variance of F0 and intensity distributions, and of F0 rising segments. Research groups aiming to develop automatic recognition of emotions for colloquial speech include [7] and [8]. In [7], three classes (approval/attentional bids/prohibition) were recognized. Speech was analyzed using three classes: F0 (variance, slope, range, mean), formant transitions and energy variations. A speaker-dependent classifier worked best, with an accuracy of 66%.

A general conclusion suggested in the literature is that, for four basic emotions, speaker-independent recognition rates can reach 60% or so at best. In research of this kind, the immense speaker variability with respect to the vocal expression of emotions is clearly one of the most difficult problems: the way in which vocal cues correlate with affect may be to some extent speaker-dependent. Another problem is that systematically collected emotional speech databases are not usually publicly available – hence the lack of any standard database in Finnish. It is difficult to get authentic data due to copyright restrictions, see for example [9] and [10]. In some

experiments, the researchers have even behaved in an arrogant and offensive manner in order to induce real emotions to the subjects [11]. For spoken Finnish, there has been very little research on vocal correlates of emotions, mainly for short syllables [12].

Our aim is to develop content based information retrieval methods for spoken Finnish, utilising the MediaTeam emotional speech corpus, a first large Finnish emotional speech database. In addition to acoustic measurements, we use subjective listening tests in order to determine how well basic emotions can be differentiated automatically vs. perceptually. In this paper, the focus is on the technical aspects of the speech analysis algorithms and classification procedure.

## 2. Methods

Before computing prosodic features of speech, signal is first processed for detecting voiced segments and measuring F0 curves. Then 43 features are computed.

### 2.1. Voiced/unvoiced segmentation

A digital audio signal is first partitioned into 60ms overlapping segments in 10ms steps. Through the guidelines described below, a cepstrum is computed for each segment, and voiced/unvoiced (V/UV) classification is performed by combining information from consecutive segments. Cepstrum peaks are estimated in two stages: first rough estimates are computed, and then more accurate values are achieved.

For computing rough estimates, a cepstrum is computed for each segment. An amplitude correction by linear weighing is performed on the cepstrum in range of  $1/700 - 1/40$  quefrenciesation  $e$  in order to compensate for F0 variation within a segment [13]. This operation enables F0-independent global thresholding for peak detection, to be described next. It also enables better F0 estimation at the end points of voiced segments. To emphasize the F0 peaks of a noisy cepstrum and thus make peak detection more reliable, a running average liftering over the cepstrum is performed. Medians of the cepstrum peak amplitudes and segment root-mean square (RMS) energies over the speech recording are next calculated.

These are used in the second stage as thresholds to find the cepstral peak locations of voiced segments. If multiple peaks are present within a segment the one lowest in quefrency is selected. The peak detection operation is embedded in a F0 tracking routine that uses a 2ms tolerance window for locating the next expected pulse peak in the signal segment. This function is designed to enhance the processing of trailing voiced segments [14].

A common problem that makes F0 estimation difficult is the frequency doubling caused by higher formants of speech. This problem was solved by applying a nonlinear function developed in this work to the signal amplitude prior to cepstrum calculations. By flattening the spectrum it reduces

the predominance of higher formants [4]. This algorithm also improves glottal pulse peak determination of creaky voiced segments that are common in Finnish speech by stabilizing the amplitude of consecutive cycles.

Finally, a segment is classified voiced if it and the segment immediately before it have RMS energy and cepstral peak amplitude higher than the corresponding median-based thresholds. Consecutive voiced segments and segments with only one unvoiced segment between them are then joined to form the final V/UV segmentation data [14].

## 2.2. F0-contour estimation

A waveform-matching algorithm is used to estimate the F0 pitch contours for each voiced segment [15]. Accurate F0 estimation is required in order to estimate features like jitter and shimmer.

First, a finite impulse response (FIR) band-pass filter is adapted to the F0 distribution obtained from the rough pitch information during V/UV segmentation from cycle period information of cepstrum peak locations. A zero-crossing calculation is then used to construct rough cycle boundaries for the waveform-matching algorithm.

Every consecutive pair of roughly marked cycles of 1kHz FIR low-pass pre-filtered data is then screened for maximum or minimum peak match using least squared error method with quadratic peak interpolation. The raw cycle peak frequency contour is then screened for simple errors and fitted with a cubic smooth contour for prosodic parameter calculations [4].

## 2.3. Prosodic feature computation

From the V/UV segmentation and F0 contour data a total of 43 prosodic features are calculated automatically, see Table 5. Features include F0 frequency, segment energy, voiced/unvoiced/silent temporal and spectral derivatives as well as other high level correlates.

## 2.4. Feature selection

A feature selection was performed using the method of unexplained variance [16] in which, at each step, the prosodic feature that minimizes the sum of the unexplained variation between groups is selected in the feature vector. The threshold for adding/removing a feature was set so that 10 best prosodic features were selected. The utilization of more correlates was found not to improve the performance of the classifiers used in this study. The small amount of data also suggested that longer vectors would not be advantageous due to over learning of data.

## 2.5. Statistical classification

Classification was performed using k Nearest Neighbor classifier (kNN) and Fisher's linear classifier [17]. These classifiers were chosen as they are commonly used in the classification [6]. Both classifiers were tested using leave-one-out classification method to maximize the utilization of data; therefore no separate training data was used. In kNN, prior to testing a sample against the data, all samples of that person were removed to ensure that no match would occur due to similarity of voice rather than emotion. However, when tested, this had only slight effect on the results.

# 3. Experimental results

## 3.1. Data

The MediaTeam emotional speech corpus includes 56 monologues reflecting basic emotions, each about one minute in length. To collect speech material, professional actors (eight men and six women) were recruited to simulate basic emotions in Finnish speech. The age of the speakers varied between 25 and 50. First each speaker was asked to read out a phonetically rich Finnish passage of some 120 words in a neutral or natural tone of voice. Each one-minute monologue signal was divided into five consecutive segments for feature extraction and classification purposes.

An attempt was made to find a text that would be semantically as neutral as possible; the text dealt with the nutritional value of the Finnish crowberry. Then the speakers were to read out the text simulating the following emotions: happiness/joy, sadness and anger. The actors were encouraged to take their time to prepare for each emotional state and they could retake the reading (as many times as they wanted) if they were not satisfied with the first version. The data was collected in fourteen consecutive sessions within a period of two months. All the speech material was digitally recorded with DAT in an anechoic studio to produce a 48 kHz, 16-bit recording. The data was stored in a PC as wav format files.

For comparison purposes, a performance test for human emotion recognition has been performed. To investigate the perceptual adequacy of the emotional speech samples, listening tests were used: eighteen test subjects, university students of adult education (thirteen women, five men, aged between 19 and 32), were recruited to listen to the data. The test subjects heard the speech samples in random order; in the forced choice test, the emotional labels were the same as those actually expressed by the speakers. The test subjects heard the speech data in eight consecutive sessions within a period of two months in connection with the regular lectures they were attending.

The listening tests were arranged in a classroom where the test subjects heard the speech data from two computer speakers. The test subjects were instructed to listen to the "tone of voice" only and bear in mind that the lexical content of each speech sample was the same. It was also made clear that the emotional labels to choose between were limited to the intended emotions, not containing any distracters. The results for human performance (classification accuracy in the units of %) are presented in Table 1. The average classification accuracy was 76.9%. The tests also indicated that the quality of emotional content in the utterances varied significantly from actor to actor; a range of 57-93% classification accuracy was calculated from the listening experiment.

Table 1. Confusion matrix of listening tests

Human	Neutral	Sad	Angry	Happy
Neutral	78.4	16.9	2.6	2.1
Sad	14.9	85.3	1.0	0.8
Angry	14.9	2.9	76.9	5.3
Happy	24.3	5.4	3.3	67.0

Average accuracy: 76.9%

### 3.2. Application scenarios

Three scenarios of recognizing speaker emotions in a speech-driven computer UI were designed to test classification performance in situations of varying difficultness. In scenario 1, the speaker has been recognized earlier and the new emotional speech samples were compared with the speech of the same speaker in the database. In scenario 2, it was assumed that the speaker acted as a user of a computer trained to automatically recognize the emotional content of the speech of 14 persons, without knowing the identity of the speaker. In scenario 3, a speaker-independent application was assumed, in which the emotional speech samples of the current speaker were not included in the database.

### 3.3. Results

The performance of our classifier was measured using kNN with k values of 1, 3, 5 and 7, and the best results are reported below. As expected, the classification accuracy decreases as the recognition situation becomes more difficult. The first two scenarios show similar levels of performance as the human listeners. Scenario 3 represents speaker-independent model and yields a decreased performance.

Table 2. Confusion matrix for Scenario 1 (k=1)

	Neutral	Sad	Angry	Happy
Neutral	92.9	0	1	7.1
Sad	2.9	95.7	0	1.4
Angry	8.6	2.9	70.0	18.5
Happy	18.6	2.9	14.3	64.2

Average accuracy: 80.7%

Table 3. Confusion matrix for Scenario 2 (k=3)

	Neutral	Sad	Angry	Happy
Neutral	84.2	2.9	0	12.9
Sad	10.0	88.6	0	1.4
Angry	10.0	1.4	71.4	17.2
Happy	25.7	2.9	14.3	57.1

Average accuracy: 75.4%

Table 4. Confusion matrix for Scenario 3 (k=5)

	Neutral	Sad	Angry	Happy
Neutral	64.3	12.9	8.6	14.3
Sad	14.3	75.7	1.4	8.6
Angry	8.6	0	47.1	44.3
Happy	24.2	4.3	18.6	52.9

Average accuracy: 60.0%

## 4. Conclusions

The results indicate two things. Firstly, in spoken Finnish, basic emotions can be perceived accurately. Compared with the results reported in the literature, the results of the recognition scores in the listening tests are quite good (76 %): for example, according to Scherer et al. [18], in the western cultural context, basic emotions can be recognized on the basis of prosodic cues with an accuracy of about 66 %.

Secondly, the automatic classification was also encouraging (60-80%) in comparison with previous results: Bosch [19] concludes that 60 % correct classification is an attainable goal for automatic systems aiming to (speaker-independent) limited happiness/joy, anger, sadness/grief discrimination. According to Whiteside [20], the automatic recognition of a speaker's three principal emotions – joy, sadness and anger – is possible, with an average classification rate varying between 60% and 80%. McGilloway et al. [21] report a classification level of 55% for the automatic recognition of fear, anger, joy, sadness, and the neutral emotion.

It should be pointed out that emotional content of our samples were relatively pure and intense: as professional actors produced the emotional speech material, effective vocal portrayals of emotions could be expected. The quality was not perfect, however, as only 57-93% classification accuracy was achieved with human listeners.

It should be pointed out that as the size of the data is small, strong conclusions cannot be drawn. In the future research, more data will be needed, in terms of both speakers and listeners. Furthermore, more authentic speech data, reflecting genuine emotions more reliably, should be used.

## 5. Acknowledgements

The Technology Development Centre (Tekes) and the Finnish Academy are gratefully acknowledged.

## 6. References

- [1] E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33-60, 2003.
- [2] F. Yu, E. Chang, Y.-Q. Xu and H.-Y. Shum, "Emotion detection from speech to enrich multimedia content," *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia*, pp. 550-557, Peking, 2001.
- [3] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms", *Speech Communication*, 40, 227-256, 2003.

[4] W. Hess, Pitch Determination of Speech Signals: Algorithms and Devices. Berlin, Springer-Verlag, Germany 1983.

[5] G. Klasmeyer, "An automatic description tool for time-contours and long-term average voice features in large emotional speech databases", *Proceedings of ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, 66-71, 2000.

[6] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk & S. Stroeve, "Approaching automatic recognition of emotion from voice: a rough benchmark", *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, 200-205, 2000.

[7] M. Slaney & G. McRoberts, "Baby Ears: A Recognition System for Affective Vocalization", *Proceedings of ICASSP 1998*.

[8] C. Breazal, *Sociable Machines: Expressive Social Exchange Between Humans and Robots*, PhD Thesis, MIT AI Lab, 2000.

[9] P. Roach, R. Stibbard, J. Osborne, S. Arnfield and J. Setter, "Transcription of prosodic and paralinguistic features of emotional speech," *Journal of the International Phonetic Association*, vol. 28, pp. 83-94, 1998.

[10] E. Douglas-Cowie, R. Cowie and M. Schroeder, "A new emotion database: considerations, sources and scope," *Proceedings of the ISCA ITRW on Speech and Emotion*, pp. 39-44, Belfast, 2000.

[11] G. Stemmler, M. Heldmann, C.A. Pauls and T. Scherer, "Constraints for emotion specificity in fear and anger: The context counts," *Psychophysiology*, vol. 38, pp. 275-291, 2001.

[12] A.-M. Laukkanen, E. Vilkmann, P. Alku & H. Oksanen, 1996, "Physical variations related to stress and emotional state: a preliminary study", *Journal of Phonetics* 24: 313-335.

[13] A.M. Noll, "Cepstrum peak determination", *Journal of Acoust. Soc. Am.* 41, 293-309, 1967.

[14] S. Ahmadi, and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transaction on Speech and Audio Processing*, Vol. 7, NO. 3, 333-338, May 1999.

[15] I. R. Titze and L. Haixiang, "Comparison of F0 extraction methods for high-precision voice perturbation measurements," *Journal of Speech and Hearing Research*, Vol. 36, 1120-1133, Dec 1993.

[16] SPSS Inc, SPSS 7.5 Statistical Algorithms. SPSS Inc., 1997.

[17] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Recognition*, 2nd edition. New York, John Wiley & Sons Inc., 2001.

[18] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across language and cultures," *Journal of Cross-Cultural Psychology*, 32, 76-92, 2001.

[19] L. Bosch, "Emotions: what is possible in the ASR framework," in *ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[20] Whiteside S., 1998, "Simulated Emotions: An Acoustic Study of Voice and Perturbation Measures", *Proceedings of ICSLP 1998*, p. 699-703.

[21] McGilloway S., Cowie R., Douglas-Cowie E., Gielen S., Westerdijk M. & Stroeve S., 2000, "Approaching automatic recognition of emotion from voice: a rough

benchmark", *Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, 2000*, p. 200-205.

Table 5. List of prosodic features.

Mean F0 frequency (Hz)
Median F0 frequency (Hz)
Maximum F0 frequency (Hz)
Minimum F0 frequency (Hz)
F0 frequency range (Hz)
95% value of F0 frequency (Hz)
5% value of F0 frequency (Hz)
5%->95% F0 frequency range (Hz)
Average F0 rise during cont. voiced segment (Hz)
Average F0 fall during cont. voiced segment (Hz)
Average F0 rise steepness (Hz/cycle)
Average F0 fall steepness (Hz/cycle)
Max rise during cont. voiced segment (Hz)
Max rise during cont. voiced segment (Hz)
Max steepness of F0 rise (Hz/cycle)
Max steepness of F0 fall (Hz/cycle)
Normalized segment frequency distribution width variation
F0 variation
Trend corrected mean proportional random F0 perturbation
Mean RMS intensity
Median RMS intensity
Max RMS intensity
Min RMS intensity
Intensity range
95% value of intensity
5% value of intensity
5%->95% intensity range
Normalized segment intensity distribution width variation
Intensity variation
Trend corrected mean proportional random int perturbation
Average length of voiced runs
Average length of nonvoiced segments shorter than 500ms
Average length of silence segments shorter than 400ms
Average length of nonvoiced segments longer than 500ms
Average length of silence segments longer than 400ms
Max length of voiced segments
Max length of nonvoiced segments
Max length of silence segments
Ratio of speech against long nonvoiced pauses
Ratio of voicing against pauses
Ratio of silence against speech
Proportion of Low Frequency Energy under 500Hz
Proportion of Low Frequency Energy under 1000Hz