Hidden Markov model-based speech emotion recognition

Björn Schuller, Gerhard Rigoll, Manfred Lang

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Gerhard Rigoll, and Manfred Lang. 2003. "Hidden Markov model-based speech emotion recognition." In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'03: Proceedings, 6-10 April 2003, Hong Kong, China*, II–1. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2003.1202279.

HIDDEN MARKOV MODEL-BASED SPEECH EMOTION RECOGNITION

Björn Schuller, Gerhard Rigoll, and Manfred Lang

Institute for Human-Computer Communication
Technische Universität München
(schuller | rigoll | lang)@ei.tum.de

ABSTRACT

In this contribution we introduce speech emotion recognition by use of continuous hidden Markov models. Two methods are propagated and compared throughout the paper. Within the first method a global statistics framework of an utterance is classified by Gaussian mixture models using derived features of the raw pitch and energy contour of the speech signal. A second method introduces increased temporal complexity applying continuous hidden Markov models considering several states using low-level instantaneous features instead of global statistics. The paper addresses the design of working recognition engines and results achieved with respect to the alluded alternatives. A speech corpus consisting of acted and spontaneous emotion samples in German and English language is described in detail. Both engines have been tested and trained using this equivalent speech corpus. Results in recognition of seven discrete emotions exceeded 86% recognition rate. As a basis of comparison the similar judgment of human deciders classifying the same corpus at 79.8% recognition rate was analyzed.

1. INTRODUCTION

Speech emotion recognition is one of the latest challenges in speech processing. Besides human facial expressions speech has proven as one of the most promising modalities for the automatic recognition of human emotions. Especially in the field of security systems a growing interest can be observed throughout the last year. Besides, the detection of lies, video games and psychiatric aid are often claimed as further scenarios for emotion recognition [1]. Addressing classification in a practical view it has to be considered that a technical approach can only rely on pragmatic decisions about kind, extent and number of emotions suiting the situation. It seems reasonable to adapt and limit this number and kind of recognizable emotions to the requirements given within the application to ensure a robust classification. Yet no standard exists for the classification of emotions in technical recognition. An often favored way is to distinguish between a defined set of discrete emotions. However, as mentioned, no common opinion exists about their number and naming. A recent approach can be found in the MPEG4 standard, which names the six emotions anger, disgust, fear, joy, sadness and surprise. The addition of a neutral state seems reasonable to realize the absence of any of these emotions. This classification is used as a basis for the comparison throughout this work also expecting further comparisons. Most approaches in nowadays speech emotion recognition use global statistics of a phrase as basis [2]. However also first efforts in recognition of instantaneous features exist [3][4]. We present two working engines using both alluded alternatives by use of continuous hidden Markov models, which have evolved as a far spread standard technique in speech processing.

2. EXTRACTION OF THE RAW FEATURE CONTOURS

To estimate a user's emotion by the speech signal one has to carefully select suited features. First of all such have to carry information about the transmitted emotion. However they also need to fit the chosen modeling by means of classification algorithms. Since we aim to use global statistics versus instantaneous features in our solutions, it seems reasonable to select appropriate different feature sets. For better comparability these will be derived by the same underlying contours. The extraction of the raw contours will be described first, since they form the basis for the later derived adapted feature vectors. We chose the analysis of the contours of pitch and energy for their well known capability to carry a large amount of information considering a user's emotion. The selected contours rely rather on broad classes of sounds while spectral characteristics in general seem to depend too strongly on phonemes and therefore on the phonetic content of an utterance. This is a drawback thinking of the premise of independency of the spoken content or even the language. In order to calculate the contours, frames of the speech signal are analyzed every 10ms using a Hamming window function. The values of energy are calculated by the

logarithmic mean energy within a frame. The pitch contour is achieved by use of the average magnitude difference function (AMDF) as can be seen in the equation, where $F_{0,i}$ represents the fundamental frequency in the frame i, $s(k)$ the signal at a discrete time instant k, N stands for the last sample in the frame and f_s is the sampling frequency. The order of the propagated AMDF-function is represented by j.

$$F_{0,i} = \left(\frac{\arg\min_{\kappa} \frac{1}{N} \sum_{k=0}^{N-1} |s_i(k) - s_i(k+\kappa)|^j}{f_s} \right)^{-1}; N = T_w \cdot f_s$$

The AMDF provides a faster alternative to the calculation of the autocorrelation function of a frame. The precondition however is that the AMDF is calculated in first order as used in our calculation. This results in additions instead of multiplications compared to the related auto correlation function, and claims the search of the minimum instead of the maximum to achieve the instantaneous pitch value. As all estimation methods for pitch contour this technique also underlies deviations from the original contour, which could only be measured by glottal measurement. However AMDF proved robust against noise but susceptible to dominant formants.

3. GLOBAL STATISTICS USING GMM'S

Within the first method we derive 20 features of the underlying introduced raw contours. The optimal set of global statistic features is broadly discussed [5]. In general the introduced features have been chosen accepting speaker dependant recognition aiming at optimal results. The features concerning temporal aspects such as the rate of voiced sounds, are approximated with respect to zero levels in pitch contour due to the inharmonic nature of unvoiced sounds. In the following the elements of our feature-vector are listed in detail. The features will be listed according to the amount of their contribution to the emotion analysis as observed in our studies. Features showing the highest discriminance will be listed first.

3.1. Pitch related features

In general the pitch related features showed more potential than the energy related features. However still a clear divergent priority can be observed among them. While the mean duration and standard deviation of duration highly contribute to a robust classification, the introduction of the reversal points added only little support.

- Mean duration of voiced sounds
- Standard deviation of duration
- Average pitch
- Standard deviation of pitch
- Relative pitch maximum
- Relative pitch minimum
- Position of maximum pitch
- Position of minimum pitch
- Maximum of absolute pitch derivation
- Mean distance between reversal points
- Standard deviation of distance between reversal points
- Mean of absolute pitch derivation
- Rate of voiced sounds

3.2. Energy related features

Equivalent to pitch related features a high discrepancy in contribution can be observed in energy related features as well. Once again reversal points showed low discrimination between the classes.

- Relative maximum of derivation of energy
- Position of maximum of derivation of energy
- Average of derivation of energy
- Standard deviation of derivation of energy
- Maximum of absolute second derivation of energy
- Mean distance between reversal points
- Standard deviation of distance between reversal points

3.3. Processing of the derived features

The features are freed of their mean value and normalized to their standard deviation. They are classified by single state HMM's (GMM), which are able to approximate the probability distribution function of each derived feature by means of a mixture of Gaussian distributions. Up to four mixtures have been used. No further gain could be observed using more than these. Each emotion is modeled by one GMM in our approach. The maximum likelihood model will be considered as the recognized emotion at a time throughout the recognition process.

4. USING CHMMS FOR THE RECOGNITION

Within the second method we strive to increase the temporal complexity and use the warping capability of hidden Markov models by introduction of more states in the models. Since global statistics are clearly invalid for this purpose, one has to carefully consider suited features. As mentioned earlier, such have to fit the modeled structure besides that they have to carry emotional information. The short time behavior of human speech shall be respected by direct analysis of low-level contours.

Considering the requirement of independence of the spoken content unvoiced sounds are eliminated in the pitch contour. This results in a loss of temporal information but greatly improves independency. We further use first and second order derivatives of this adjusted pitch contour. Concerning the energy contour without relying on the absolute value of energy we use first and higher order derivatives. To further free the contours of noise they are filtered by use of a symmetrical moving average filter of the width three. The obviously low-pass characteristic impulse-response of the SMA-filter can be seen in its Fourier transform in the next equation where B represents the filter-width:

$$H(f) = \frac{\sin(\pi f B)}{B \sin(\pi f)}$$

The contours were additionally normalized according to their overall standard deviation and freed of their mean value. As a result we achieve a six-dimensional feature vector $\underline{m}_i$ for each frame where F_0 represents instantaneous pitch, E instantaneous energy and i the frame index:

$$\underline{m}_i = (F_{0i}, \frac{dF_{0i}}{dt}, \frac{d^2F_{0i}}{dt^2}, \frac{dE_i}{dt}, \frac{d^2E_i}{dt^2}, \frac{d^3E_i}{dt^3})$$

The continuous HMMs (CHMM's) were trained using Baum Welch re-estimation with a maximum of 10^5 iterations or an abruption criterion of a change in model parameters of $e < 10^{-4}$. Up to four Gaussian mixtures have been used to approximate the emission probability density functions according to the GMM solution. The HMM types were chosen as Left-Right-models, as in usual speech processing, which ideally models advances in time. As a jump constraint the increase in the state index may not exceed two. One model was chosen for each emotional state resulting in seven overall models. The maximum likelihood model was chosen for the assumed emotion.

5. SPEECH CORPUS

The corpus has been collected with use of a dynamic AKG-1000S MK-II microphone in an acoustically isolated room. The phrases were all collected in German and English language. Two different methods have been used to collect speech samples of five speakers resulting in 5250 samples in total: the larger test-set of four speakers consists of acted emotions. The usage of acted emotions seems reasonable to obtain a first impression of recognition performance. To avoid similarities in over-exaggerated pronunciations these utterances have been assembled over a long period of time (about six months per speaker). A further method aimed at the collection of spontaneous emotions motivated by the ability to compare recognition results to those of acted emotions. Test persons were recorded during video gaming in usability tests at our institute. These recordings have been labeled according to the corresponding situation. The utterances vary in length and spoken content throughout the corpus meeting the challenge of ensuring greater independence.

6. RECOGNITION RESULTS

For reasons of comparability the tested engines have been trained with the same samples of five speakers. 100 utterances per emotion and speaker have been used for the training, while a disjunctive set of 50 emotions per speaker and emotion were used throughout the evaluation phase. The tables represent the on average achieved recognition results. Since the introduced approaches tend to strongly depend on the speaker, no cross-speaker evaluation results are presented. A loss of at least 20% in recognition rate could be observed if evaluated with a different speaker. On the other hand no significant difference using spontaneous emotions occurred. However more testing is needed to manifest this effect. In the tables *sur* abbreviates surprise, *joy* joy, *ang* anger, *fea* fear, *dis* disgust, *sad* sadness, and *neu* neutral user emotion. Only results with optimal system parameters are presented.

6.1. Global statistics

In the following the results are listed in detail for the phrase-based approach.

	Recognized emotion						
Labeled emotion	ang	dis	fea	sur	joy	ntl	sad
ang	**91.1**	1.2	0.6	6.3	0.4	0.1	0.3
dis	5.4	**76.8**	6.7	0.1	6.8	3.2	1.0
fea	0.2	6.4	**82.8**	0.6	3.0	0.3	6.7
sur	2.4	2.2	3.0	**87.2**	4.6	0.1	0.5
joy	3.0	0.7	0.8	0.0	**93.2**	0.2	2.1
ntl	0.2	3.4	0.4	0.5	2.7	**89.6**	3.2
sad	0.2	0.1	5.8	3.8	0.4	2.2	**86.6**

Figure (1): Confusion table of global analysis with 86.8% overall recognition rate

The table shows the distribution for prosodic feature analysis using the first feature set, four mixtures and one state. Downwards the acted emotion will be listed, while to the right the recognized emotion can be seen.

6.2. Instantaneous features

The confusion matrix was achieved using 64 states and four mixtures. The effect of the number of states used will be highlighted afterwards.

Labeled emotion	Recognized emotion ang	dis	fea	sur	joy	ntl	sad
ang	**68.5**	12.7	2.6	1.8	2.7	8.4	3.3
dis	12.8	**84.7**	2.1	0.3	0.0	0.1	0.0
fea	1.8	0.1	**95.4**	0.2	2.0	0.4	0.1
sur	6.3	6.7	6.3	**73.5**	6.1	0.9	0.2
joy	10.1	11.8	7.9	1.2	**68.0**	0.5	0.5
ntl	10.4	0.9	1.0	0.1	1.9	**79.6**	6.1
sad	5.9	10.1	2.8	2.1	2.2	1.8	**75.1**

Figure (2): Confusion table instantaneous features continuous HMM with 77.8% overall recognition rate

The next figure shows the effect of increasing the complexity in temporal modeling by usage of more states. A break even was reached at 64 states. Regarding the result with only one state, at 64.7% recognition rate, a direct comparison between the two feature sets can be drawn using the same classification method. The derived features clearly outperform the low-level features in this case.

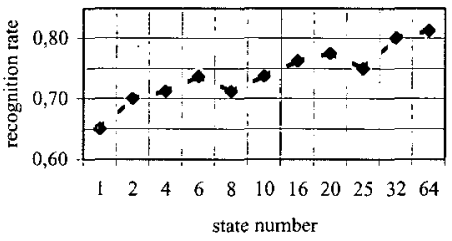

Figure (3): Recognition results depending on the number of states used

As can be seen further, a general increase in performance can be observed using more states. Like this the low-level feature prove their full potential only at a higher number of states used.

6.3. Human judgment

To obtain a benchmark we aimed to let humans analyze the same corpus. The samples were judged by five human deciders. The samples used for evaluation were presented to each test person in random order. Their overall classification rate was on average 81.3% correct assignment of the emotional patterns compared to the emotion intended by the speaker.

7. DISCUSSION

We believe that this contribution shows important results considering emotion recognition with hidden Markov models. The two introduced methods proved both capable of a rather reasonable model for the automatic recognition of human emotions in speech. The confusion tables clearly show that some emotions are often confused with certain others. Furthermore some emotions seem to be recognized more easily. This may be due to the fact that the most test patterns were acted emotions and test-persons have difficulties with feigning certain emotions. Though the same training material and test sets were used, the two proclaimed solutions differ greatly in their behavior. Neither the confusion of emotions nor the performance of recognition of single emotions itself shows significant correlations in the result. While the global phrase statistics outperformed the instantaneous features, still both propagated solutions build a reasonable model. One reason for the better performance can be seen in the loss of information of durations of voiced sounds by eliminating these in the contours as described. The results of both engines reach the abilities of a human decider as described above. In our future work we aim at a hybrid approach combining neural networks and hidden Markov models for the automatic recognition. Also the integration of other modalities such as video based or manual interaction [6] will be investigated further.

8. REFERENCES

[1] R. Cowie, et al.: *"Emotion recognition in human-computer interaction"*, IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] N. Amir: *"Classifying emotions in speech: a comparison of methods"*, Eurospeech 2001, Poster Proceedings, pp. 127-130, Scandinavia, 2001.

[3] A. Nogueiras, A. Moreno, A. Bonafonte, and J. Mariño: *"Speech Emotion Recognition Using Hidden Markov Models"*, Eurospeech 2001, Poster Proceedings, pp. 2679-2682, Scandinavia, 2001

[4] B. Schuller: *"Towards intuitive speech interaction by the integration of emotional aspects"*, SMC 2002, IEEE International Conference on Systems, Man and Cybernetics, Yasmine Hammamet, Tunisia, CD-Rom Proceedings

[5] T. Polzin: *"Verbal and non-verbal cues in the communication of emotions"*, ICASSP 2000, Paper Proc. ID: 3485, Turkey, 2000.

[6] B. Schuller, M. Lang, G. Rigoll: *"Multimodal Emotion Recognition in Audiovisual Communication"*, ICME 2002, 3rd International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, CD-Rom Proceedings