Improved Emotion Recognition with Large Set of Statistical Features

Vladimir Hozjan, Zdravko Kačič

University of Maribor
Faculty of Electrical Engineering and Computer Science, Maribor
vladimir.hozjan@uni-mb.si, kacic@uni-mb.si

Abstract
This paper presents and discusses the speaker dependent emotion recognition with large set of statistical features. The speaker dependent emotion recognition gains in present the best accuracy performance. Recognition was performed on English, Slovenian, Spanish, and French InterFace emotional speech databases. All databases include 9 speakers. The InterFace databases include neutral speaking style and six emotions: disgust, surprise, joy, fear, anger and sadness. Speech features for emotion recognition were determined in two steps. In the first step, acoustical features were defined and in the second statistical features were calculated from acoustical features. Acoustical features are composed from pitch, derivative of pitch, energy, derivative of energy, duration of speech segments, jitter, and shimmer. Statistical features are statistical presentations of acoustical features. In previous study feature vector was composed from 26 elements. In this study the feature vector was composed from 144 elements. The new feature set was called large set of statistical features. Significant improvement was achieved for all speakers except for Slovenian male and second English male speaker were the improvement was about 2%. Large set of statistical features improve the accuracy of recognised emotion in average for about 18%.

1. Introduction
Today, the need for emotion modelling arises because of the need for a spoken and multimodal communication system. For example, various software agents need emotion models to communicate with humans in the most natural way. These software agents need emotion models for emotion recognition using the face, body movements, and speech, as well as for synthesis of emotions in the animation of a software agent’s face, animation of body movements, and in synthesised speech [1].

The most widely accepted concept in emotion research is that certain emotions are primary, and others are secondary. Secondary emotions can be explained as a combination of primary emotions. This idea can be traced back to Descartes [2]. There is no definitive list of basic emotions. However, a general agreement exists on a basic list of emotions: anger, disgust, fear, happiness, sadness, and surprise [3].

The aim of speech synthesis systems is to produce speech that sounds naturally emotional. This is the same goal that actors pursue [5]. Recording of a collection of natural speech that contains naturally expressed emotion is a complex problem. Some researchers have done an analysis of naturally expressed or spontaneous emotion [6].

Many researchers have analysed simulated and spontaneous emotions in emotional speech for primary emotions such as anger, joy, fear, and sadness. They found that joy, anger, and fear have higher mean F0 and amplitude, with greater F0 and amplitude variations, than neutral speech. Sadness has an opposite effect on F0 and amplitude than the emotions anger, joy and fear. Sadness has low mean F0 and amplitude, and low variations of F0 and amplitude [7]. Speech rate also varies among emotions. The highest speech rate is for neutral and angry speech and the slowest for sad speech [6].

Emotion recognition is developing in different directions. The developers of emotion recognisers use different types of databases, features, and methods for recognition. Databases are usually composed from one language and have a small number of recordings or number of emotions. Most developers use long-time features. Dilemma over the usage of short-time or long-time features has always existed. The long time features are determined from whole utterance or recording of speech and short-time features are determined in smaller time window that has usually duration from 20 to 100ms. Researchers on this topic have noted that long term-features identify emotion better than short time features [8]. Artificial neural networks and hidden Markov models are the most used methods. Noquieras et. al. [9] used short-term features and hidden Markov models for emotion recognition using seven emotions and achieved 82.5% emotion recognition accuracy. Noam [10] used long-term features, neural networks and distance measurement for the emotion recognition of four emotions. Emotion recognition using neural networks achieved from 55% to 98% accuracy and with distance measurement from 61% to 83%. These two authors did not use comparable databases. Therefore, these results are not directly comparable.

The idea of large set of statistical features comes from Oudeyer [16] that achieved the accuracy of speaker dependent emotion recognition over 90%. He used only 2 speakers and 3 emotional states and one neutral state. He used 2000 sentences for emotion recognition. Oudeyer’s feature set was composed from 200 features derived from pitch, energy and spectral components. Oudeyer showed insignificant contribution of spectral components in the emotion recognition.

The aim of this paper is to present the speaker dependent emotion recognition with large set of statistical features. We will expand Oudeyer set with the duration features, with the sum of absolute pitch and energy difference that are a part of prosody and with shimmer and jitter, which are part of speech quality. Database used in research included 9 speakers and about 23000 sentences. Therefore, this database has more speakers and sentences than the Oudeyer’s database. The
result of speaker dependent emotion recognition with new large set of statistical features was compared with results from previous research presented in [17].

2. Methods

Emotion recognition needs different kind of methods for feature extraction, detection of emotions and evaluation of results.

2.1. Feature extraction

2.1.1. Acoustical features

It is well established that emotions affect the pitch, energy, and duration of speech segments. The pitch contour (F0) of the speech signal, duration of phonemes and pauses, energy of the speech signal, the derivative of the speech signal pitch contour (ΔF0), the derivative of the speech signal energy contour, the energy of high frequency and energy of low frequency, the sum of absolute difference of F0, and the sum of the absolute difference of energy, the jitter and the shimmer were defined as acoustical features.

The values of pitch contour were calculated only on speech segments that were voiced. The values of the pitch contour were set to 0 for the speech segments that were unvoiced. This was also a problem for determining the derivative of pitch contour. The derivative was determined only on the voiced segments of speech. The values for the unvoiced segments of speech were set to 0.

The normalised cross-correlation function [11] was used to calculate F0 from the speech signal. The method calculated one value per 10ms. The energy of the speech signal was calculated using the RMS (Root Mean Square) method. RMS calculates energy in a square window that is 10ms long at a frame rate of 200Hz. The acoustical features were converted into statistical features for the emotion recognition so that the different frame rates of the F0 and the energy calculation would not have any effect on emotion recognition.

The energy of low frequencies was determined from energy below 250Hz and energy of high frequencies above 250Hz. This border of 250Hz between low and high frequencies was proposed by Oudeyer [16].

Phoneme segmentation was accomplished automatically using the HTK toolkit [12]. Here, the speech signal was first pre-emphasised with a factor of 0.95. Twelve mel-cepstral coefficients and the energy, as well as the first and second derivatives of the mel-cepstral coefficients, were used as a feature vector. The cepstral coefficients were calculated using a Hamming window (5 ms long) at a frame rate of 400Hz. Normalisation of the cepstral mean value was used. A 3-state left-right HMM was used. The emission probabilities were modelled with continuous Gaussian mixture densities. The models for all phonemes, for pauses between words and the start and end points of sentences were used for acoustic modelling.

HMM phoneme models were developed for each speaker separately. We wanted to make segmentation as accurate as possible. Therefore, the training database consisted of all the sentences for each speaker for all the emotional states. On initialisation the global mean value and variances were estimated. HMM parameters were re-estimated using six iterations of the Baum-Welch algorithm. In the next stage, the number of Gaussian mixture densities was increased from one to eight in steps of two. Two iterations of HMM parameters were performed, after each increase in the number of Gaussian mixture densities. Segmentation was done on all the sentences of each speaker for all emotional states.

From the segmentation six duration series were determined; phoneme duration, syllable duration, vocal duration, fricative duration, consonant duration, and affricate duration.

The sum of absolute difference is well known measure in prosody. The sum of absolute difference of pitch and energy were often used to determine pitch and energy accent of syllables and words in speech signal [19], [20].

Jitter and shimmer are quantified by the pitch perturbation quotient and the energy perturbation quotient using the perturbation quotient presented in [19]. The jitter and shimmer are two parameters that are classified as parameters of speech quality.

2.1.2. Statistical features

Statistical presentations of acoustical features are defined as statistical features. First set of features is composed from 26 features and is presented in [17], [22]. The statistical features present information about the prosody of the speaker over the entire sentence. Statistical features contain information about intonation, tempo, and loudness.

Statistical features derived from F0 were: mean value of F0, standard deviation of F0, minimum and maximum values of F0, and the F0 range. Statistical features derived from ΔF0 were mean value of ΔF0, standard deviation of ΔF0, minimum and maximum values of ΔF0, and the ΔF0 range.

The second group encompasses features calculated from the energy contour and its derivative: mean value of energy contour, standard deviation of energy contour, energy of words, energy of syllables, mean value of derivative energy contour, and standard deviation of derivative energy contour.

The third group consists of features extracted from the phoneme duration: duration of words, duration of syllables, duration of vowels, duration of fricatives, duration of affricates, duration of sonorants, duration of plosives, and duration of pauses.

2.1.3. Large set of statistical feature (LSSF)

LSSF was determined from sixteen acoustical features described in previous section 2.1.1. For each acoustical feature nine statistical feature were determined. Statistical features are: mean, standard deviation, maximum, minimum, range, median, first quartile, third quartile, and inter quartile. For each sentence in the database 144 features were determined.

2.2. Emotion recognition

Recognition of emotions was accomplished using artificial neural networks. Neural networks were defined by a SNNS toolkit (Stuttgart Neural Network Simulator) version 4.2 [13].

2.2.1. Topology of neural networks

The emotion recogniser was defined using a multilayer perceptron (MLP). The size of the input layer depends on the size of the input vector. The size of the input layer in the MLP was equal to the number of used features. The size of the
output layer depends on the size of the target vector. The number of elements in the target vector was equal to the number of emotions. All six emotions and neutral styles were used in all InterFace databases [14], [22].

The topology of artificial neural networks for 26 features is presented in [17]. An artificial neural network for emotion recognition with LSSF has 144 input neuron, 7 output neurons, and 49 neurons in hidden layer.

All neurons have a hyperbolic tangens activation function and an output identity function for output function [13]. The "r-prop" procedure was used as a training algorithm, with a random initialisation function and a topological order update function. The "r-prop" stands for "Resilient backpropagation" and is a local adaptive learning scheme, performing supervised batch learning in the MLP.

2.2.2. Data preparation

Features must be converted into the appropriate form for the topologies of the neural networks described above. The data were normalised with mean value and standard deviation of neutral speech style.

The learning procedure requires sets of input vectors and target vectors. A target vector is the desired output of the MLP. The value of the target vector depends on the input vector. The output vector also depends on the input vector and is an actual output of the MLP. The input vector and the target vector are combined together into a pattern. All patterns are stored in a pattern file, which is in an appropriate format for the SNNS toolkit.

The target vector was binary decoded with values -1 and 1 and includes seven element. Each element of the target vector represents a specific emotion or neutral speaking style. For example, the first element of the target vector denotes anger. If the target vector represents anger, then the first element contains value 1 and all other elements contain the value -1.

Eighty percent of all sentences from the database were randomly selected for the training of the neural network, and the remaining twenty percent were used for testing.

2.2.3. Methods for result evaluation

Result evaluation was made using the well known evaluation method "max-correct" [10]. It evaluates the correctness of the entire output vector. One output vector denotes the emotion of one sentence. Therefore, the max-correct evaluates the accuracy of emotion recognition. The max-correct method searches for those elements in the output and target vectors that have maximum values. If the element with the maximum value in the target vector and the element with the highest value in the output vector denote the same emotion or the same speaking style, then the vector is defined as the correct vector.

3. Emotional speech database

English, Slovenian, Spanish, and French InterFace databases [14] were used in our experiment. The databases were recorded in the framework of the InterFace1 project. All the InterFace databases consist of actors’ recordings. In the English InterFace database two male speakers (speaker M1 and speaker M2) and one female speaker were recorded. In the Slovenian, Spanish and French database one male and one female speaker in each language were recorded.

3.1. Speech materials

InterFace corpora were recorded over six basic emotions, and neutral speaking style. The recorded emotions are anger (A), disgust (D), fear (F), joy (J), surprise (S), sadness (T), and neutral (N). These emotions are defined as “basic” emotions [3].

<table>
<thead>
<tr>
<th>Emotions</th>
<th>26 features</th>
<th>LSSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>2,78%</td>
<td>76,38%</td>
</tr>
<tr>
<td>Disgust</td>
<td>4,17%</td>
<td>86,30%</td>
</tr>
<tr>
<td>Fear</td>
<td>0,00%</td>
<td>90,41%</td>
</tr>
<tr>
<td>Joy</td>
<td>86,11%</td>
<td>64,38%</td>
</tr>
<tr>
<td>Neutral</td>
<td>69,86%</td>
<td>91,78%</td>
</tr>
<tr>
<td>Surprise</td>
<td>48,61%</td>
<td>87,50%</td>
</tr>
<tr>
<td>Sadness</td>
<td>64,29%</td>
<td>93,15%</td>
</tr>
</tbody>
</table>

Table 2: The accuracy for each emotion of Spanish male speaker emotion recognition with 26 features and with LSSF.

<table>
<thead>
<tr>
<th>Emotions</th>
<th>26 features</th>
<th>LSSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>69,74%</td>
<td>75,00%</td>
</tr>
<tr>
<td>Disgust</td>
<td>52,63%</td>
<td>56,57%</td>
</tr>
<tr>
<td>Fear</td>
<td>59,21%</td>
<td>68,42%</td>
</tr>
<tr>
<td>Joy</td>
<td>32,89%</td>
<td>42,10%</td>
</tr>
<tr>
<td>Neutral</td>
<td>67,11%</td>
<td>65,78%</td>
</tr>
<tr>
<td>Surprise</td>
<td>36,84%</td>
<td>50,00%</td>
</tr>
<tr>
<td>Sadness</td>
<td>80,26%</td>
<td>64,47%</td>
</tr>
</tbody>
</table>

Table 3: The accuracy for each emotion of Slovenian male speaker emotion recognition with 26 features and with LSSF.

4. Results

The results of the speaker dependent emotion recognition are presented in table 1. Also shown the results with 26 features [17] are shown in table.

Accuracy of speaker dependent emotion recognition showed improvement for all speakers. The improvement was insignificant for English male2 speaker and for Slovenian male speaker. The improvement is for both speakers less than...
2%. The improvement is biggest for Spanish male and is 44.99%. In average the achieved improvement is 18.07%.

The tables 2 and 3 show the results for each emotion of Spanish male and Slovenian male emotion recognition with 26 and large set of statistical features. Slovenian male speakers had the worst accuracy and Spanish male speaker had the best improvement in accuracy.

5. Discussion

The improvement of emotion recognition accuracy shows that it was reasonable to expand feature vector with additional acoustical and statistic features.

The absolute values of speaker dependent emotion recognition differ from speaker to speaker. The accuracy of speaker dependent emotion recognition with LSSF is indirect pointer of quality of the speakers' ability to express emotions.

Speakers expressed their emotion best they could. We are not able to measure the intensity of expressed emotions in speech signal so there will be always a doubt about the quality of expressed emotions. Our opinion is that Slovenian male speaker and English male speaker expressed emotions with lower intensity as others did. We attribute insignificant improvement for English male2 and Slovenian male speaker to lower degree of emotion expression intensity. The results also show the improvement of accuracy for most emotions in emotion recognition with LSSF.

26 features used in emotion recognition described in [17] were not able to capture all properties of emotion expression. This case was worst for Spanish male speaker where improvement with LSSF was the biggest. We assume that LSSF capture most of the prosody properties of emotional speech and some properties of speech quality, therefore, speaker dependent emotion recognition improves.

6. Conclusions

In this paper the speaker dependent emotion recognition with LSSF and neural networks is presented. LSSF were expanded with additional acoustical and statistical features. The results of emotion recognition were compared with the emotion recognition described in [17]. Speaker dependent emotion recognition improved in average for about 18%.

7. References