Enhanced Robustness in Speech Emotion Recognition Combining Acoustic and Semantic Analyses

Ronald Müller, Björn Schuller, Gerhard Rigoll
Technische Universität München, Institute for Human Machine-Communication
Arcisstr. 21, D-80333 München, Germany
{mueller, schuller, rigoll}@mmk.ei.tum.de

Abstract

In this contribution we like to give a very short introduction to a system allowing for enhanced robustness in emotion recognition from speech. The underlying emotion set consists in seven discrete states derived from the MPEG-4 standard: Anger, disgust, fear, joy, neutral, sadness, and surprise. Within here several novel approaches and investigations to acoustic feature sets, semantic analysis of spoken content, and the combination of those two information streams in a soft decision fusion are presented in short. Firstly statistic features on prosody, extracted from the speech signal, are ranked by their quantitative contribution to the estimation of an emotion. After investigation of various classification methods Support Vector Machines performed out within this task. Secondly an approach to emotion recognition by the spoken content is introduced applying Bayesian Network based spotting for emotional key-phrases. Finally the two information sources will be integrated in a soft decision fusion by using a MLP, which eventually leads to a remarkable improvement in overall recognition results.

1. System overview

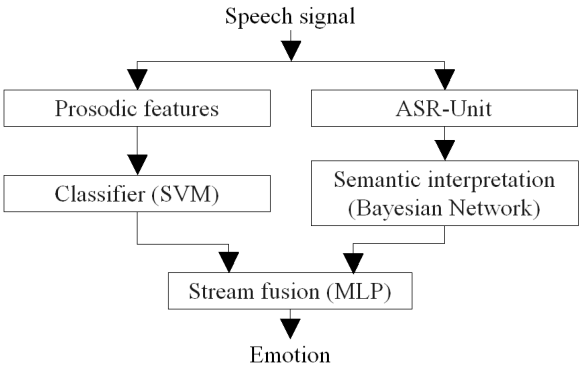

Figure 1. System architecture

Figure 1 shows the proposed system architecture allowing for robust emotion recognition as cutout of an automotive infotainment application. Spoken utterances are recorded and subsequently processed by units for feature extraction and Automatic Speech Recognition (ASR), which leads to two information streams: One focusing on acoustic properties, the other addressing the linguistic information contained. At the end of both streams we obtain overall 14 confidences, i.e. one per emotion and stream to allow for an entirely probabilistic post-processing within the stream fusion. Thereby a Multi-Layer-Perceptron (MLP) eventually performs a soft decision fusion providing one confidence for each of the seven discrete emotions as output. Via maximum likelihood the final hard decision may take place at this point if desired.

2. Emotional Speech Corpus

The emotional speech corpus has been collected in the framework of the FERMUS III project [1], dealing with emotion recognition in an automotive environment. The corpus consists of 2828 acted emotional samples from 13 speakers used for training and evaluation in the prosodic and semantic analysis. The samples were recorded over a period of one year to avoid anticipation effects of the actors. While these acted emotions tend to form a reasonable basis for a first impression of the obtainable performance, the use of spontaneous emotions seems to offer more realistic results, especially in view of the spoken content. A second set consists of 700 selected utterances in automotive infotainment speech interaction dialogs recorded for the evaluation of the fusion. In this project disgust and sadness were of minor interest. Therefore these have been provoked in additional usability test-setups to ensure equal distribution among the emotions in the data set.

3. Acoustic Analysis

Unlike former works [1], which compared static and dynamic feature sets for the prosodic analysis, we focus on derived static features herein. Initially the raw contours of pitch and energy are calculated because they rather rely on broad classes of sounds. Spectral characteristics on the other hand seem to depend too strongly on the phonetic content of an utterance. Therefore only spectral energy below 250Hz and 650Hz is used considering spectral information. The values of signal energy resemble the logarithmic mean energy within a frame. The Average Magnitude Difference Function (AMDF) provides the pitch contour. This method proves robust against noise but susceptible to dominant

formants. A low-pass filtering applying a symmetrical moving average filter of the filter-width three smoothens the raw contours prior to the statistical analysis. In a next step higher-level features are derived out of the contours, freed of their mean value and normalized to their standard deviation. The temporal aspects of voiced sounds are approximated with respect to zero levels in the pitch contour due to the inharmonic nature of unvoiced sounds. As the optimal set of global static features is broadly discussed [2][3], we considered an initially large set of more than 200 features. Via LDA a set of 33 most discriminating features could be selected. For example the best three features are: Maximum pitch gradient, relative position of the pitch maximum, and the standard deviation in pitch. In a direct comparison a combination of all pitch related features lead to 69.8% correct recognition rate, compared to 36.6% correct recognition rate for the use of all energy related features. Finally the combination of the complete set of 33 features and using Support-Vector-Machines (SVM) as Classifier results in a recognition rate of 74.2% . Hereby seven SVM's are trained for the emotional states, each "one-against-all", as SVM's are designed for two-class problems. Thus after scaling and standardization we obtain a pseudo-confidence for each emotion and refer it to the subsequent stream fusion for a soft decision.[3]

4. Semantic Analysis

Apart from the acoustic properties in many cases spoken utterances also contain emotional information lying in the choice of words, which, as shown in the following, should not be ignored in order to improve recognition performance significantly. Hereby it is assumed, that the speaker's expression of his emotion consists in usage of certain phrases that are likely to be mixed with other meaningful statements [4][5]. Therefore an approach with abilities in spotting for emotional relevant information is needed. Furthermore as an Automatic Speech Recognition (ASR) unit provides probably incomplete and uncertain data to work on, the processing interpretation algorithm should not only be able to deal with but use such knowledge. Hence, as mathematical background for the semantic analysis of spoken utterances we chose Belief Networks (BN) for their capabilities in spotting and handling uncertain and incomplete information on the input level as well as providing real recognition confidences at the output, which is valuable in regard to a subsequent late fusion with results from prosodic analysis. The aim here is to make the Belief Network maximize the probability of the root node modeling the specific emotion expressed by a speaker via his choice of words and phrases.

The approach presented here is to be based on integration and abstraction of semantically similar units to higher leveled units in several layers. This method proved to be applicable for Natural Language Interpretation in several restricted domains, like natural language man-machine dialogues for intuitive application controlling, with considerable success.

Training and evaluation procedures ran on the data corpus described before. Thereby 12% of items are free of any emotional content and are therefore assigned to "neutral". The contained emotion of a number of utterances recorded and transcribed in the database can unambiguously be identified only via their prosodic properties as from the spoken content alone a distinctive assignment appeared to be impossible even for a human mind. Hence, the average recognition rate at 59.6% of this semantic interpretation approach left on its own seems to be rather poor compared to methods based on acoustics. Nevertheless this additional information appears quite valuable, as shown in the following.

5. Stream Fusion

As introduced in Fig. 1 the two outputs of the acoustic and the semantic information streams, each comprising one confidence for each of the seven emotions, are combined. For this several methods have been applied [4]. We investigated the pairwise maximum mean value as well as a maximum likelihood decision on the output of a MLP build up at the evaluative best with 14 input-, 100 hidden-, and 7 output-neurons. Table 2 shows an overview of the average recognition rates with the system parameters described in this contribution.

Model	Acoustic Information	Language Information	Fusion by means	Fusion by MLP
Rate, %	74.2	59.6	83.1	92.0

Table 2. Performance gain means-based and by MLP

6. References

[1] B. Schuller, M. Lang, G. Rigoll: "Multimodal Emotion Recognition in Audiovisual Communication," *ICME 2002*, CD-Rom Proceedings, Lausanne, Switzerland.

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

[3] B. Schuller, G. Rigoll, M. Lang: "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture," *ICASSP 2004*, Montreal, Canada.

[4] C. M. Lee, R. Pieraccini: "Combining acoustic and language information for emotion recognition," *ICSLP 2002*, Denver, CO, USA.

[5] L. Devillers, L. Lamel: "Emotion Detection in Task-Oriented Dialogs," *ICME 2003*, IEEE, Multimedia Human-Machine Interface and Interaction I, Vol. III, pp. 549-552, Baltimore, MD, USA.