AN IMPLEMENTATION OF ANGER DETECTION IN SPEECH SIGNALS

A. A. Mohamoud, M. Maris

TNO Defence, Security and Safety,
Business Unit Observation systems
Waalsdorperweg 63
2509 JG The Hague, The Netherlands
E-mail: {ali.mohamoud, marinus.maris}@tno.nl
Fax: +31 70 374 06 53

Keywords: Emotion in speech, Ambient Intelligence, Speech signal.

Abstract

In this paper, an emotion classification system based on speech signals is presented. The classifier can identify the most common emotions, namely *anger*, *neutral*, *happiness* and *fear*. The algorithm computes a number of acoustic features which are fed into the classifier based on a pattern recognition approach. The classification system is of potential benefit for ambient intelligence in which the emotional and physical states of a person should be known to the intelligence of the environment. Using such information, the environment can better support humans in their daily activities in accordance with their preferences.

1. Introduction

Within the scope of Ambient Intelligence, an intelligent environment supports its human users in their daily activities. In order to do so, the environment monitors the humans and has a notion of their preferences. In the type of intelligent environments we are interested in, the human user is assisted in maintaining optimal well-being. Hence, it is important to monitor the emotional and physical conditions of the user, as they provide indications about the user's health state. It is known that stress may have a negative influence on our notion of well-being [1],[2]. It has also been studied that in some cases, certain human emotions can lead to stress [3]. Particular *anger* plays a major role with respect to this [4].

One approach to thwart the effects of negative influences on ones' well-being is to use an intelligent environment which monitors the user and make suggestions to undertake certain activities or even to actively control the surroundings. For example, the intelligent environment can try to relieve stress factors by creating a relaxing atmosphere with appropriate environmental conditions. This may include the use of proper music, lights and projecting suitable images on the wall [5].
Our research group has a strong focus on studying principles of ambient intelligence. Much work is performed within a large project called Smart Surroundings. Within the scope of this project, a person-identification system has been developed, as well as a stress detector (in cooperation with other laboratories). What failed was an emotional state detector to better react to the preferences of the user of the intelligent environment. We chose to develop such an emotional state classifier based on speech analysis. Hence, the idea is that when the users utter words while being in the monitoring range of the intelligent environment, the classifier is capable of identifying emotional state of the users.

Speech analysis is a large research field with three main topic areas, namely speech synthesis, speech coding and speech recognition. The area of speech recognition includes human machine interfacing where emotion recognition plays an important role. Therefore, research on human speech emotion has recently been attracting much attention within the scientific community. The major force behind this increase is interest is to improve upon current human-machine interfacing capabilities. In [16], for example, robots learn to interact with humans and recognize human emotions. There are also methods to teach robotic pets to understand not only spoken commands but also other information like the emotional state of its human commander and modify their behaviour accordingly. A major problem however encountered in speech recognition is the variability of human speech. There have been numerous efforts to enhance speech recognition algorithms tackling this problem [10],[11].

The work in this paper builds on this previous work and discusses an innovative application of this technology. Concretely, an emotional state recognition system is presented based on speech analysis. The algorithm enables automatic classification of speech utterances into a predefined set of emotional states. This classifier will be

deployed in an intelligent smart environment we are currently developing.

2. Emotion database

For most viable emotion detection applications, a comprehensive emotional database is of paramount importance. To this end we used the online available *Berlin Emotional Database*. This database contains a set of so-called 'big four' emotions (*anger*, *fear*, *happiness* and *sadness*). It also contains *neutral*, *disgust* and *boredom* emotions. The database was created by ten actors (5 male and 5 female) who simulated the emotions producing ten German utterances. The 16-bit recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz. The reason behind this down sampling is to reduce the computational power needed for speech processing [6]. The online Berlin Emotional Database contains 494 utterances and was evaluated by 20 human listeners. For our emotion classifier system, we used utterances pertained to the above mentioned four emotions. These utterances – 323 in total – have a perception test result of 80% and higher.

Based on expectations related to the Smart Surroundings project, we selected three of the above-mentioned 'big four' emotions, namely *anger*, *happiness* and *fear*. We also choose to distinguish *neutral* as we anticipate that this will be a very common emotional state in Smart Surroundings applications. In the database, a total of 323 utterances were found, divided over the four emotional states as shown in Table 1.

Table 1 Number of utterances available per emotion type

	Anger	*Neutral*	*Happiness*	*Fear*
Number of utterances	127	79	63	54

In our system, classification of speech utterances takes two steps. In the first step, a number of relevant acoustic features are extracted. In the second step, these features are fed into the classifier. This is explained in the next section.

3. Acoustic features

Feature selection is the most challenging task in speech emotion detection. As stated in [12] and [13] the feature choice appears to be data-dependent. As long as there is no common agreement on the best features to use, we chose to select the set of features reported in [8] for our emotion recognition system. These are the Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients and the energy and pitch of the utterance.

A block diagram detailing our feature extraction architecture is depicted in Figure 1. The first stage of the analysis process is to find the endpoints of each speech utterance. This is followed by calculating the amount of energy in the higher frequencies of the signal. Namely, in the spectrum of voiced speech, the amount of energy in higher frequencies is much lower than in lower frequencies. This co-called spectral tilt is compensated in the *Pre-emphasis* stage.

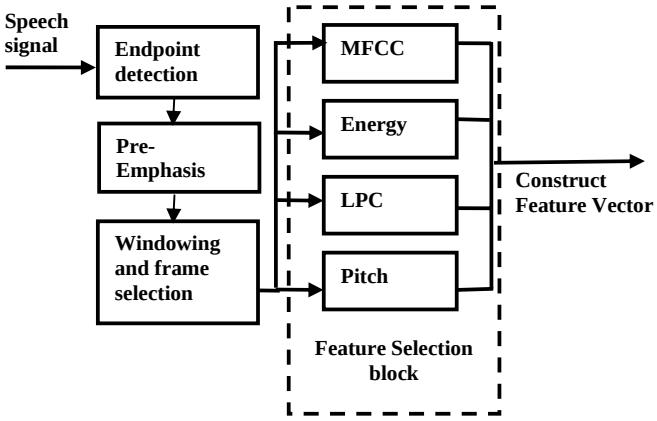

Figure 1 Block diagram of feature extraction and selection.

After these two first processing steps, each incoming speech signal is divided into overlapping frames of 16 msec. long. The overlapping duration is set to 8 msec. (see figure 2). Short frame length and sufficient overlap were chosen to in order to account for rapid speech signal changes.

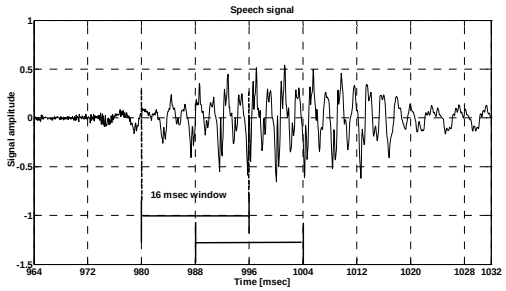

Figure 2 Window selection of a speech signal

The resulting train of frames is put through the feature extraction and selection routine, enclosed by the dotted line shown in figure 1. Here, the four selected features are shown. Note that for the MFCC and energy, we also generated the first and second derivatives in order to increase the feature set.

We used the sequential forward feature selection (SFFS) method [12] to identify the number of coefficients for the MFCC. Starting with an empty feature set, this algorithm finds the best number of coefficients that satisfies some criterion function in each step.

The generated feature vector as a result of the sequential feature (coefficient) selection method for each frame

consists of 12 MFCCs, 12 first derivatives of MFCCs and 12 second derivatives of MFCCs. Apart from the MFCC's, the generated feature vector consists of energy, the first derivative of the energy, the second derivative of the energy, 10 LPCs and the pitch. The number of MFCC and LPC coefficients is a result of the feature selection algorithm.

4. Classification

For the generation of the resulting feature vector (see Figure 1) we performed vector quantization based on the Linde-Buzo-Gray (LBG) k-means (where k = 16) clustering algorithm [9]. The features of all frames in one emotion utterance are averaged to form a feature vector. The next step in generating a complete feature vector for each emotion is to cluster these features into one emotion class by using LBG algorithm. This method of vector quantization is very powerful in reducing feature vectors without losing useful information and has a low computation complexity [15]. These are two important requirements in our implementation.

Speech emotion recognition is considered a pattern recognition task [7]. Feature extraction, feature selection, classifier choice training and testing are the main stages in the pattern recognition cycle. Applying this approach we cycle through these stages as sketched in figure 3.

As emotion classifier we used a K-NN (K-Nearest Neighbour with K being 8) method because it is well-proven and computationally inexpensive. For this method, we need to train a classifier model. This model actually generates the classification result as a pattern. Training is performed by the Linde-Buzo-Gray algorithm, which yields the trained vector quantization. Hence, in the pattern recognition process three different levels can be identified, namely, feature extraction and selection, classifier modelling and training.

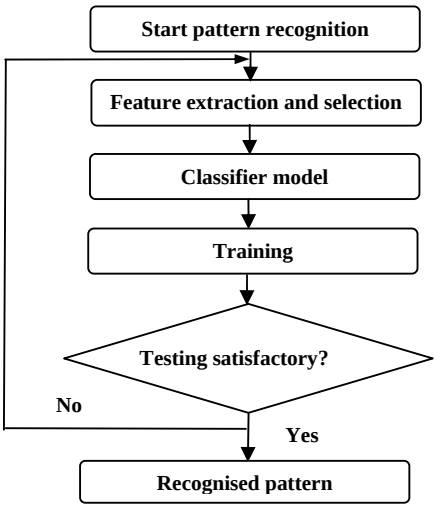

Figure 1 Classification of emotion in speech: a pattern recognition approach.

5. Experiments

The emotion detection and classification system has been realized in Matlab. A nice aspect of this programming environment is that the algorithms as well as the graphical user interface are developed in the same environment.

For the evaluation, we randomly selected ten speech utterances (40 in total) from the database for the four emotional states we presented in table 1 (*anger*, *neutral*, *happiness* and *fear*). This is done in order to test and validate algorithms pertained to the implemented emotion classifier. To easily evaluate and test our algorithm, we developed an application to validate the four real life emotions. The GUI of this application is depicted in figure 4.

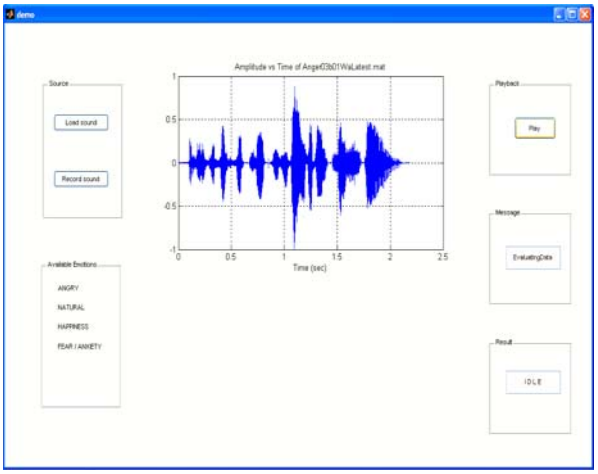

Figure 2 Emotion classifier GUI

6. Classification results

The results obtained with the classifier described in the previous section are given in table 2 where classification results are given in percents. The diagonal values give the correct classified percentage of the emotion in question. A high recognition rate is found for anger and neutral utterances which are very promising in our Smart Surroundings applications.

Due to the lack of more test data, we clustered the features resulting from each utterance by using LBG k-means (where k = 8) algorithm. This test data is put through the k-nn classifier (where k = 8). In contrary to the majority vote decision rule which is mostly utilized, we used the ratio of correctly recognized emotion versus the all calculated nearest neighbours. These classifier results are depicted in table 2.

Table 2 Classifier result using the Berlin Emotional Database

Emotion	Anger	Neutral	Happiness	Fear
Anger	80.10	1.25	11.15	7.50
Neutral	6.85	84.60	1.05	7.50
Happiness	24.88	6.70	64.30	4.12
Fear	23.13	6.88	4.59	65.4

7. Conclusion

In this paper, we discussed an emotional state analysis system. We plan to use this system in TNO's ambient environment system which is currently capable of identifying persons. It will be augmented with a stress measurement system in cooperation with another laboratory which will provide stress level measurements using a heartbeat monitor. While assessing the stress level of humans, it was mentioned that the emotional state *anger* provides an important clue as it may be a source for undesired stress. Hence, anger monitoring will be part of the ambient intelligent system we have in mind to monitor the stress level of humans. Therefore, the results for the *anger* classification are the most interesting one in this study. As shown in Table 2, our system is capable of detecting *anger* of persons (utterances extracted from the database) with a reliability of 80.1%. In other words, with this system we are capable of differentiating particularly anger from other emotions.

Such information is crucial for an intelligent ambient system which supports the well-being of the users. Future work will include detection of additional emotional states such as *fear* or *happiness*.

8. Acknowledgements

This work was carried out within the BSIK project Smart Surroundings funded by the Ministry of Economic affairs. We thank the TNO laboratory and the project team for their support.

9. References

[1] Maslach C, Schaufeli W, Leiter MP. Job burnout. Annu Rev Psychol. 2001;52:397-422.

[2] Maslach C. Burnout: A Multidimensional Perspective. In: Schaufeli Maslach Ch, Marek T, eds. Professional Burnout:Recent Developments Theory and Research. New York.: Taylor & Francis.; 1993:19-32.

[3] Zapf D., Vogt C., Seifert C, Mertini H. and Isic A. Emotion work as a source of stress: the concept and developement of an instrument. European Journal of Work and Organizational Psychology 1999;8:371-400.

[4] Johnson, E.H. The Deadly Emotions: The Role of Anger, Hostility, and Aggression in Health and Emotional Well-Being. Praeger, New York, 1990.

[5] Weber W, Rabaey J.M and Aarts E. Ambient Intelligence. Springer Berlin Heidelberg, 2005. ISBN: 978-3-540-23867-6

[6] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W and Weiss, B. A Database of German Emotional Speech, *Proc. Interspeech* 2005.

[7] Sidorova, J. Speech Emotion Recognition DEA report doctoral program Ciència Cognitive I Llengiatge

[8] Yun-Mao Cheng, e.a. Using Recognition of Emotions in Speech to Better Understand Brand Slogans, Multimedia Signal Processing, 2006 IEEE 8th on, pages 238-242

[9] Linde, Y. Buzo, A. Gray, R. An Algorithm for vector quantization design. IEEE Transactions on Communications, vol. 28, pp.84-95, Jan 1980

[10] Iain, R. Arnott, J. Synthesizing emotions in speech. Fourth International Proceedings on Spoken Language, 3-6 Oct 1996.

[11] Gerhard, D. Audio Signal Classification: History and Current Techniques. Technical Report. ISBN 0 7731 0456 9

[12] Ververidis, D. Kotropoulos, C. Sequential forward feature selection with low computational cost. Department of Informatics, Aristotle university of Thessaloniki.

[13] Devillers, L., Vidrascu, L., Lamel, L., Challenges in real-life emotion annotation and machine learning based detection, Journal of Neural Networks 2005, 18/4, "Emotion and Brain"

[14] Juslin, P.N., & Laukka, P., "Communication of emotions in vocal expression and music performance: different channels, same code?." Psychological Bulletin: 129 (5), 2003, p 770–814

[15] Shanbehzadeh, J. Ogunbona, P.O. On the computational complexity of the LBG and PNN algorithms. IEEE Transactions on Image Processing. Volume 6, Issue 4, Apr 1997 Page(s):614 – 616

[16] Kanda T., Iwase K., Shiomi M., Ishiguro H. A tension-moderating mechanism for speech-based human-robot interaction. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005), Edmonton, Alberta, Canada, August 2-6, 2005.