Towards intuitive speech interaction by the integration of emotional aspects

Björn Schuller

Technical University of Munich

Institute for Human-Machine-Interaction

Arcisstraße 16, D-80290 Munich, Germany

schuller@ei.tum.de

ABSTRACT

In this contribution new results in instantaneous recognition of emotion in non-verbal speech shall be presented. As classification method dynamic programming with Dynamic Time Warp or Bakis-Hidden-Markov-Models with Vector Quantization or Gaussian mixtures are used to analyze the pitch and energy contour of a speech signal. As emotional states joy, anger, fear, sadness, disgust, irritation, and an additional neutral user state have been evaluated. As rather unusual innovative user states the influences of tiredness and alcohol consumption of a speaker on his speech have been analyzed by use of the same methods. One of the main goals of the presented work was to keep the models applicable for new users and recognition simple for real-time evaluation. Finally observed results are presented and discussed.

Keywords: **Automatic emotion recognition, social competence in the human-computer-interaction, speech processing**

I. INTRODUCTION

A growing interest in the recognition of human emotions can be observed. In our daily live emotions play an important role. For example in dialogs between humans they help to interpret the counterpart's intention with the emotional background of the speaker. The communication between a human and a machine on the other hand has been greatly improved in the last decades. More powerful and smaller processors helped to enable complex algorithms to e.g. understand human speech. With further new modalities like gesture recognition or gaze tracking the interaction resembled more and more the communication between human beings. This encourages users to also show their emotions but makes them also expect a very natural system to understand e.g. their ironic manner of talking. Realizing a user's need of aid at a surprised or helpless user emotion might be another use of automatic emotion recognition. Also safety routines could be activated if users seem tired, or influenced by alcohol. If the user tends to be dissatisfied a system could either apologize or check where an error might have been made. Finally, if a user resembles satisfied it could learn without external supervision. Fields of application reach from medical psychological analysis over the detection of lies to video games and many more.

A. *Emotional classification in psychological theory*

If we want to classify emotions in a reasonable way, we have to reflect the meaning of an emotional state in the given context. In psychology research four general views exist according to Cornelius [1]. They all establish their own model of the origin and nature of the human feelings.

Darwin [2] claims that emotions developed in the evolution of mankind and form an essential factor to survive. According to his theory certain behavior patterns are linked to emotional sentiments. Like Descartes, who initially followed this idea, the supporters of Darwin believe in underlying basic emotions which number is strongly discussed. Ekman claims six, Plutchik eight and Izard ten. It is believed that these fundamental emotions possess similar cross-cultural meaning. Shaver found a high correlation between six analyzed emotional reactions in Italy, China and the USA: love, joy, surprise, anger, sadness, and fear. The previously mentioned models likewise postulate similar emotional states.

James [3] on the other hand defines emotions as the perception of the body's reaction to certain events. This means that an emotion arises from the stimulus of an organ of sense. By afferent impulses that reach the cortex an object is noticed. The inner organs and muscles are stimulated by efferent impulses. Finally their afferent pulses lead back to the cerebrum cortex. Perception of this corporeal change is defined as emotion. Like this emotional sentiment without a precedent physical reaction is not possible.

The cognitive approach represented by Arnold differs from the latter in the description of a judgment before the body's reaction. The emotion is experienced according to this valuation. A change in consideration of the context therefore results in a change of emotion.

Defenders of the recent socially constructivist perspective support the opinion that emotion is the result of trained social rules. Among others Averill [4] and Harré [5] regard the culture as the most important factor for the contextual appraisal leading to emotional sentiment. Causes of e.g. anger differ greatly among cultures but also among people. To interpret an emotion correctly it is therefore regarded most essential to analyze the underlying cultural background.

Nowadays none of the above approaches is in general valued as the only right model. Mixed approaches to emotional classification can also be found. However unresolved problems exist, as Cowie et al. [6] describe: No set of standard emotions exists, and also criteria to define such differ greatly. Another contradiction is the belief that emotion is a product of evolution in view of the cultural differences.

Like this some syndromes are clearly classified as emotion in some cultures, while others neglect these.

B. Significance for the automatic recognition of emotions

A technical approach can only rely on pragmatic decisions about kind, extent and number of emotions suiting to the situation. It seems reasonable to adapt and limit the number and kind of recognizable emotions to the requirements given in the application to ensure a robust classification. Yet no standard exists for the classification of emotions in technical recognition. The most often persuaded way is to distinguish between a defined set of discrete emotions. However, as mentioned, no unity exists about their number and naming. A first approach can be found in the MPEG4 standard, which names the six emotions anger, disgust, fear, joy, sadness and surprise. A neutral emotional state is often added. Another idea is to assume an emotional hyper-sphere. In such a sphere emotions can be assigned to spots. Mostly the orthogonal axes positivity and activity span such a space. As an advantage one has not to limit classification to concrete emotions saving flexibility. Furthermore the temporal change in emotion can be followed more exactly. Cowie et al. [6] regret the loss of information at the reduction in complexity by the transformation of the high-dimensionally emotion sphere into a two-dimensional plain. Pereira tries to avoid this by definition of a third dimension [7]. The additional axis permits measurement of the affect in this work. Other research groups use the extent of an emotion as third axis.

C. Classification in our research

A very first step was a linear classification to distinguish between a satisfied or dissatisfied user. In this one-dimensional space the positivity was estimated. In a next approach we also intended to achieve a measurement for surprise or tiredness of the user, what leads to the former described two-dimensional sphere spanned by the axes positivity and activeness as can be seen in figure 1. We classified only discrete emotion points $e[p, a]^T$.

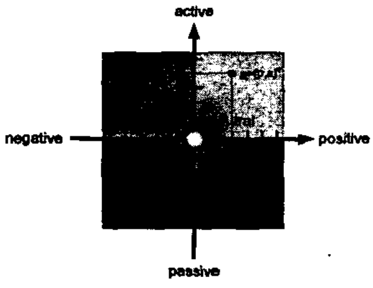

Figure 1: Two-dimensional emotion plain

Due to the high computational effort and the limited reliability this method proved less suitable in our researches. It seems a possible but still hard task for a human decider to estimate the degree of positivity or activeness. In a technical solution a quantization of the axes leads to a quadratic or cubic total number of different states according to the number of quantization steps. To achieve first reasonable results the complexity was reduced by use of maximal ten different emotions. The number was decreased if only less emotional information was used by the application to keep recognition performance stable on a high level. Using only such discrete emotions however, the latter described confidence in an assumed emotion allows for interpretation of the extent. This means that a higher trust in an emotion is correlated with a greater extent, which can be seen controversially. Used states were the mentioned states defined in MPEG4 plus additional states as a neutral user state. As innovative mental speaker states also the influences of alcohol and tiredness were integrated in the analysis.

II. Speech Corpus

A. Speech as modality

While there are several approaches in recognition by video-based features like mimic expressions [8][6], physiologic features and manual interaction [9] also in speech clues can be found. A system can either actively ask a user about his state [8] or try to collect itself data. An interrogation of the user has the advantage that it is up to the user how much he wants to let the system know, while a disadvantage is the interruption in communication for the emotional data collection.

B. Data collection

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 language and are throughout acted emotions. This has been widely discussed due to the fact that acted emotions are not spontaneous and tend to be exaggerated. To avoid similarities in over-exaggerated pronunciations the samples have been assembled over a very long period of time (about six months per speaker). The three speakers were in average 25 years old and male. Per emotion and speaker around 160 samples could be collected what results in overall more than 3500 samples. However the results achieved can be seen as an upper benchmark if we assume this collection as idealized.

III. Feature Extraction

Speech as one of the most natural communication forms among humans seems very suitable to recognize emotion. Besides verbal clues [9][10][11] also prosodic non-verbal clues carry information about the emotion. Prosody is the entirety of attributes as accent, intonation, quantity and breaks in speech. In general they are related to units larger than single phonemes. The analysis of speech-rate, rhythm and pauses also belong to paralinguistic prosodic features [12]. Global statistics of feature contours are often used to assign emotions: among others mean values, standard deviations or quartiles of fundamental frequency, energy and jitter [13], tremor [11] or temporal changes in spectral coefficients [14]. Others use instantaneous analysis of the contours as pitch and energy [15]. This method is also favored in the presented work. Further features are based on durations, e.g. of voiced or unvoiced sounds, the articulation or affect bursts [16]. In recent approaches more contextual knowledge integration can be observed as in our former publications [17]. Besides semantic information also the social background is highlighted in the evaluation process. To model the probability density functions Gaussian mixtures of first or higher orders are very popular [11][18]. Cross-cultural or multilingual interpretation is at its beginning. The classification methods reach from linear algorithms as Euclidean distance metric [19] to Hidden-Markov-Models (HMM) [15] or artificial neuronal networks [19][20]. In this work we present results achieved by use of Dynamic-Time-Warp (DTW) and HMMs. Even rule-based approaches obtained good classification. In the sound system Halliday [21] a neutral passage is correlated with a consistent decrease in pitch, while a question or an objection to the uttered content shall be recognized by an increase. Uncertainty can be seen by an increase-decrease-increase progression or an increase very shortly before the end of a phrase. Former works also analyzed global features by linear classification [10]. One of our main goals was an easily adaptable system without the need of too many training utterances for online user adaptation. Therefore we regarded discrete and continuous Hidden-Markov-Models.

Frames are analyzed every 10ms and a Hamming window function is used. The energy is calculated by the mean energy within a frame. The pitch contour is achieved by use of the average magnitude difference function (AMDF) [22] as can be seen in the equation, where F_0 represents the fundamental frequency, $s(k)$ 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 i was set to one in our calculation.

$$F_0 = \left(\frac{\arg\max_{\kappa} \sum_{k=0}^{N-1+\kappa} \left| s(k) - s(k+\kappa) \right|^{i}}{f_s} \right)^{-1}$$

The principle of the AMDF is a maximum detection in the auto-correlation function of the speech signal. 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 is robust against noise but susceptible to dominant formants. Further more the algorithm is fast due to the restriction to additions if it is calculated in first order.

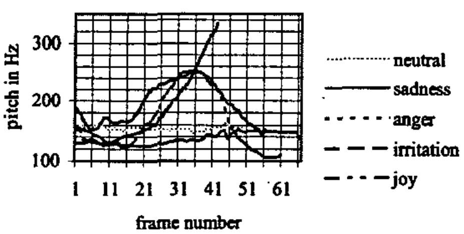

Figure 2: Examples of pitch contour for the word "no"

To eliminate further high frequent noise additions low-pass filtering of the pitch and energy contours is achieved by smoothening with a symmetrical moving average (SMA) filter. This method seems suited to deliver a smooth contour and is often applied in digital signal processing [23]. In other works a median filter is used instead [15] which is optimal to correct single outlying events. However in our contours almost none such events could be found. The equation shows the recursive algorithm for the SMA-filter where B represents the odd broadness of the filter window, and x stands for the signal.

$$\bar{x}_k = \frac{1}{B} \sum_{i=k-\frac{B-1}{2}}^{k+\frac{B-1}{2}} x_i; \quad k = \frac{B-1}{2}, \cdots, N - \frac{B-1}{2}$$

The first values have to be calculated with increasing width, while the last values are respectively filtered with decreasing filter width.

$$\begin{array}{ll} \bar{x}_0 = x_0 & B^* = 0 \\ \bar{x}_1 = \frac{1}{3}(x_0 + x_1 + x_2) & B^* = 1 \\ \vdots & \textit{while } B^* < \frac{B-1}{2} \end{array}$$

The obviously low-pass characteristic impulse-response of the SMA-filter can be seen in the next equation:

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

We also use first and second order deviations of these contours. 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 $\underline{m}$ where F_0 represents the pitch and E the energy contour:

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

Unvoiced parts are cut away according to the fundamental frequency that has to exceed a given threshold. This method neglects most temporal effects but improves independence of the spoken words. In our research of instantaneous recognition this proceeding surpassed the mentioned loss in information.

IV. AUTOMATIC CLASSIFICATION

The classification is realized with methods of Dynamic Programming. In a first step a DTW-algorithm with local Itakura [24] constraints and Euclidean distance metric was used. For each emotion a set of references was created. The minimum distance to a representative of each class was calculated to achieve a score for each emotion as described later in this chapter. In a second advance we used HMM with Vector Quantization (VQ). In the HMM approach each emotion was represented by a HMM. We used 128 codebook entries and the codebook was optimized with use of the Linde-Buzo-Gray (LBG) [25] algorithm. The feature vector was reduced by the first and second order deviations to the dimension two. This reduction avoided further loss at quantization. Finally we used continuous HMMs with Gaussian Mixtures. The HMMs are trained using Baum Welch re-estimation [26] with 10^5 iterations or an abruption criterion of a change in model parameters $\varepsilon < 10^{-4}$. We use one up to four mixtures, but analysis of the data suggests use of only one Gaussian function for complexity reasons. This result could be verified in the real classification tests. The HMM types were chosen as Left-Right-models, as in usual speech processing. As a jump constraint the increase in the index may not exceed two.

To obtain a confidence measure in a standardized form one has to consider the relative difference to concurring hypotheses as well as the absolute probability of a hypothesis. The normalization used to compare results partly neglects these aspects for reasons of simplification in view of real-time capability. We normalized the HMM-evaluation probabilities to their sum, while the DTW-distances d_i are each normalized to the maximum occurring distance d_{max} to achieve a pseudo-confidence measurement c_i according to the following equation:

$$c_i = 1 - \frac{d_i}{d_{max}}$$

This solution allows direct comparison or integration of the different results.

V. User Profiling

To adapt the recognition results to a new speaker the system can be trained online supervised. A simple user profile helps to enhance recognition results. The probabilities integrated in the user model are [17]:

- $P(E_k^{(t)}|U^{(t)})$, the conditional probability of the occurrence of emotion E_k under the condition of user U at time t.
- $P(E_k^{(t)}|U^{(t)},M_1^{(t)},...,M_N^{(t)})$, the conditional probability of an emotion under the condition of the model M_i if different models are used.
- $P(E_k^{(t)}|U^{(t)},E^{(t-1)})$, to further integrate knowledge of the precedent emotion
- $P(E_k^{(t)}|U^{(t)},X^{(t)})$, for integration of external influences and finally
- $P(E_k^{(t)}|U^{(t)},S^{(t)})$, if the state of an application is given.

This leads to the integrated expectation of a user emotion E_E:

$$E_e^{(t)} = \arg\max_k \left(\frac{\sum_{i=1}^{|M|} \left(P(E_k \mid M_i^{(t)}) \cdot P(M_i^{(t)}) \right)}{\sum_{i=1}^{|M|} P(M_i^{(t)})} \cdot P(E_k^{(t)} \mid U^{(t)}, E^{(t-1)}, \bar{X}^{(t)}, S^{(t)}) \right)$$

The only way to obtain these models is a playfully interrogation of the user at the moment.

VI. Results

In the following recognition results can be found. They base on the selection of the maximum-likelihood model. A human decider with 82% recognition rate in classifying the four emotions joy, anger, sadness and fear can be seen as a benchmark for a comparison. The best guess as further comparison would be 25% recognition rate. First a DTW-algorithm was evaluated with use of several reference vectors per emotion. 40 samples per speaker have been used in a first performance measurement. The overall recognition rate was 62.5% correct assignment of the intended emotion with the four emotions: anger, irritation, joy and neutral state. While the neutral user state could be recognized with almost absolute certainty, this method could not distinguish reliably enough between joy and anger, which in general seems a difficult task. In a next step HMMs were used for classification and tested with between 50 and 80 samples per emotion and speaker. If not specially announced, the models were trained with all available samples as described in chapter III. Figure 3 shows the overall recognition results for four emotions achieved with a discrete HMM with Vector Quantization (VQ) with less preprocessing of the signal as very fast solution in training and recognition compared with a continuous HMM solution with preprocessing as described in chapter IV. The recognition results vary greatly with the number of states used for the models. A general problem remains the lack of sufficient spontaneous training-data.

In the following figures these abbreviations are used:
ang: anger; *irr*: irritation, *joy*: joy; *ntl*: neutral user state; *dis*: disgust; *fea*: fear; *sad*: sadness; *alc*: alcohol influence; *tir*: tiredness; *all*: overall performance.

State number	Rec. results discrete HMM 1)	Rec. results continuous HMM 2)
1	0.50	0.65
2	0.61	0.70
4	0.73	0.71
6	0.65	0.74
8	0.68	0.71
10	0.68	0.74
16	0.58	0.76
20	0.48	0.78
24	0.54	0.75
32	0.58	0.80
64	0.51	0.81

Figure 3: Recognition results emotions ang, irr, joy, ntl: 1) Discrete HMM without SMA-filter and normalization 2) Continuous HMM, 100 training samples per emotion, with SMA-filter (B=3) and global normalization

Figure 4 visualizes the trend of increasing recognition performance with raising number of states used. The use of more than 64 states did not result in a noteworthy further gain. However, the trend measured is not strictly monotone. Especially the fact of an observed maximum at four states with use of discrete models seems to come from a data-problem. Especially with only one state a clear difference between the approaches with discrete or continuous models catches one's eye. The continuous models outline the discrete by far in performance besides at four states, where the discrete models once exceed the continuous models in our measurements. But one has to keep in mind that the signal has not been filtered and normalized in the first solution to save computation time for a very fast estimation.

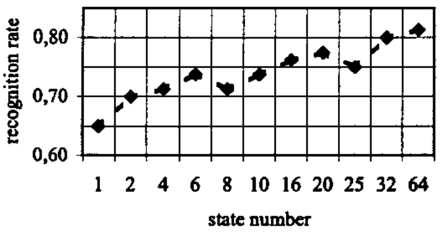

Figure 4: Recognition results emotions ang, irr, joy, ntl: continuous HMM, variable state number, with SMA-filter (B=3) and global normalization

In a real application it might be useful to train the model with the actual user data as mentioned in chapter VI. To obtain an impression how much data is at least needed the recognition results as a function of training samples used is shown in figure 5.

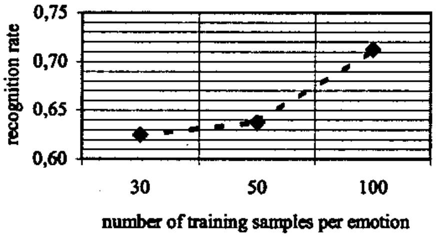

Figure 5: Recognition results emotions ang, irr, joy, ntl: continuous HMM, variable training samples per emotion, without SMA-filter, local normalization

As a test of conformity the models were reclassified with optimal HMM configuration and the complete training set. Figure 6 shows results with four emotions and variable state number. Furthermore the results for each emotion can be seen. Appreciable is the almost perfect reclassification of neutral utterances. Already with only one state remarkable results can be observed. This resembles global statistics of an utterance with Gaussian mixtures distributions to approximate the original probability density function.

sta	ang	irr	joy	ntl	all
1	0.60	0.49	0.74	1.00	**0.71**
2	0.70	0.79	0.74	1.00	**0.81**
4	0.82	0.77	0.77	1.00	**0.84**
8	0.84	0.83	0.82	0.99	**0.87**
16	0.85	0.84	0.81	0.98	**0.87**
32	0.90	0.89	0.88	0.98	**0.91**
64	0.92	0.91	0.89	0.98	**0.93**

Figure 6: reclassification rates, continuous HMM

Figure 7 shows the confusion matrix obtained with real tests with optimal configuration and maximum training set.

Intended emotion	Recognized emotion: ang	dis	fea	irr	joy	ntl	sad
ang	**0.40**	0,45	0,02	0,02	0,02	0,08	0,02
dis	0.13	**0.85**	0.02	0.00	0.00	0.00	0.00
fea	0.02	0.00	**0.96**	0.00	0.02	0.00	0.00
irr	0.12	0.10	0.06	**0.67**	0.06	0.00	0.00
joy	0.13	0.13	0.08	0.00	**0.65**	0.00	0.00
ntl	0.15	0.00	0.00	0.00	0.00	**0.79**	0.06
sad	0.12	0.10	0.00	0.00	0.02	0.02	**0.75**

Figure 7: Confusion matrix continuous HMM: 64 states, SMA-filter, global normalization, 0.73 overall rec. rate

Three further emotions have been added, what results in a loss in performance. Obviously the key-problem was the distinction between anger and disgust. In the next figure a completely different set of mental user states has been evaluated. These are the influences of alcohol consume and fatigues compared with a neutral state.

Intended emotion	Recognized emotion: alc	ntl	tir
alc	**0.36**	0.49	0.15
ntl	0.11	**0.88**	0.02
tir	0.00	0.07	**0.93**

Figure 8: Confusion matrix continuous HMM: one mixture, 64 states, SMA-filter, global normalization, 0.72 overall rec. rates

The recognition of tiredness seems satisfyingly solved with this approach, while alcohol influences tend to be confused with neutral user state. On the other hand a neutral state is recognized reliably.

VII. Discussion

In our work it proved more reasonable to classify discrete emotions than spots in a two-dimensional sphere [27]. The use of HMMs [29] excelled the use of DTW in recognition results as expected, but nonetheless a rate of 62,5% correct emotion detection could be achieved with DTW compared with 68% using discrete HMMs and 81% maximum recognition rate with continuous HMMs for four emotions. With classification of seven emotions a performance of 73% correct assignment could be reached. It could also be shown that it is possible to detect tiredness with the same approach at 93% recognition rate. While the features seem less suited to detect alcohol influences of the speaker in one utterance at least only few confusions of neutral state with alcohol influence took place. This still allows usage of the mentioned features, but one of the major problems is the negligence of stretched phoneme occurrences. Their integration might contribute to a better recognition of this effect. In a next study we also aim to investigate pain influences and exhaustion on the speaker's voice. These ideas might be interesting in high-risk tasks where it is essential to realize if someone in a reliable position is obviously drunk, sleepy or hurt.

No spectral characteristics have been integrated in the feature vector yet. The energy below 250Hz seems very promising regarding the results of McGilloway [13]. Scherer et al. use the energy below 635 Hz instead [28]. In general the use of more than one Gaussian mixture only slightly enhanced the performance provided enough sample utterances.

The integration of contextual knowledge could not be valued yet according to a lack of test-data. More studies have to take place to manifest an improvement by such knowledge integration.

VIII. Acknowledgments

The work presented in this paper has been supported by the FERMUS project, a cooperation of BMW Group, DaimlerChrysler, SiemensVDO and the institute of Human-Machine Communication at the TUM. The project stands for error-robust multimodal speech dialogues.

The contents discussed largely benefit from the collaboration with the student assistants Stephan Reiter and Ronald Müller.

IX. References

1. R. Cornelius, *„Theoretical approaches to emotion,"* ISCA Workshop on Speech and Emotion, Belfast 2000.
2. C. Darwin, *"The Expression of the emotions in man and animals,"* Chicago: University of Chicago Press 1872 / 1965
3. W. James, *"What is an emotion?,"* Mind, 19, pp. 188-205, 1884
4. J. R. Averill, *"A constructivist view of emotion,"* published in R. Plutchik & H. Kellerman (Eds.) "Emotion: Theory, research and experience, vol. 1," pp. 305-339, New York: Academic Press, 1980
5. R. Harré (Ed.), *"The social costruction of emotions,"* Oxford: Basil Blackwell, 1986
6. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, *„Emotion recognition in human-computer interaction,"* IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32-80, Januar 2001.
7. C. Pereira, *"Dimensions of emotional meaning in speech,"* ISCA 2000, Canada, 2000
8. R. W. Picard, *"Toward computers that recognize and respond to user emotion,"* IBM Systems Journal, Vol. 39, NOS 3&4, S. 705-719, 2000
9. B. Schuller, M. Lang, G. Rigoll, *"Multimodal Emotion Recognition in Audiovisual Communication,"* ICME 2002, IEEE, Lausanne, CH
10. B. Schuller, M. Lang, G. Rigoll, *"Automatic Emotion Recognition by the speech-signal,"*, SCI 2002, IIIS, Orlando, Florida, USA
11. T. S.. Polzin, *„Verbal and non-verbal cues in the communication of emotions,"* ICASSP 2000, Paper Proc. ID: 3485, Turkey, 2000
12. H. Bußmann, *"Lexikon der Sprachwissenschaft,"* Alfred Kröner Verlag, Stuttgart, 2. Ausg. 1990
13. S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, *"Approaching automatic recognition of emotion from voice: a rough benchmark,"* Proceedings of the ISCA workshop on Speech and Emotion, pp. 207-212, Newcastle 2000
14. M. Kienast, W. F. Sendlmeier, *"Acoustical analysis of spectral and temporal changes in emotional speech,"* ISCA 2000, Canada, 2000
15. A. Nogueiras, A. Moreno, A. Bonafonte, J. Mariño, *"Speech Emotion Recognition Using Hidden Markov Models,"* Eurospeech 2001, Poster Proceedings, pp. 2679-2682, Scandinavia, 2001
16. M. Schröder, *„Experimental study of affect bursts,"* ISCA 2000, Canada, 2000
17. B. Schuller, M. Lang, *"Integrative rapid-prototyping for multimodal user interfaces,"* USEWARE 2002, VDI/VDE #1678, Darmstadt, BRD, pp. 279-284
18. L. ten Bosch, *„Emotions: What is possible in the ASR framework,"* ISCA 2000, Canada, 2000
19. N. Amir, *„Classifying emotions in speech: a comparison of methods,"* Eurospeech 2001, Poster Proceedings, pp. 127-130, Scandinavia, 2001
20. R. Cowie, E. Douglas-Cowie, B. Appoloni, J. Taylor, A. Romano and W. Fellenz, *„What a neural net needs to know about emotion words,"* CSCC Proceedings, pp. 5311-5316, CSCC 1999
21. B. Eppinger, E. Herter, "Sprachverarbeitung," Carl Hanser Verlag, München Wien, 1993
22. G.S. Ying, L.H. Jamieson, and C.D. Mitchell, *"A Probabilistic Approach to AMDF Pitch Detection,"* published in "Proceedings of the Fourth International Conference on Spoken Language Processing," Philadelphia, PA, October 1996
23. S. W. Smith, *"The Scientist and Engineer's Guide to Digital Signal Processing,"* Second Edition, California Technical Publishing, San Diego, CA, 1999
24. F. Itakura, *"Distance Measure for Speech Recognition Based on the Smoothes Group Delay Spectrum,"* Proc. of the ICASSP 87, Vol. 3, pp. 1257-1260, 1987
25. Y. Linde, A. Buzo and R.M. Gray, *"An algorithm for vector quantizer design,"* IEEE Transactions on Communications, vol. 28, pp. 84-95, January 1980.
26. L. E. Baum, *"An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes,"* Inequalities, Vol. 1, pp. 1-8, 1972
27. R. Cowie, E. Douglas-Cowie, S. Savvidou, E McMahon, M. Sawey and M. Schröder, *"'Feeltrace': an instrument for recording perceived emotion in real time,"* ISCA 2000, Canada, 2000
28. K. R. Scherer, R. Banse, H. G. Wallbott, T. Goldbeck, *„Vocal cues in emotion encoding and decoding,"* Motivation and Emotion 15, pp. 123-148, 1991
29. L. R. Rabiner, *„A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,"* Proceedings of the IEEE, Vol. 77, No. 2, February 1989