AUTOMATIC EMOTION CLASSIFICATION VS. HUMAN PERCEPTION : COMPARING MACHINE PERFORMANCE TO THE HUMAN BENCHMARK

Jose Esparza[1], *Stefan Scherer*[1,3], *André Brechmann*[2], *Friedhelm Schwenker*[1]

[1] University of Ulm, Institute of Neural Information Processing, Germany
[2] Leibniz Institute for Neurobiology, Magdeburg, Germany
[3] USC Institute for Creative Technologies, Playa Vista, CA, USA

ABSTRACT

Emotion classification is performed by humans in all kinds of situations. However, perception tests, in the literature, show that humans do not perform perfectly even when classifying prototypically performed expressions. The fact that humans commit errors in this task, demonstrates that higher performance accuracies do not directly imply more realistic comprehension of the emotions' nature. Therefore, in this study we do not aim for perfect recognition performances, but rather analyze the influence of the derived features to the overall classification results and compare results to human perception tests. For this purpose, our experimental results, achieved with multi-classifier multi-class support vector machines, combining eight separate feature sets, are based on standard datasets. The results are compared, using confusion matrices, with the human perception capabilities, yielding similar accuracies.

Index Terms: human perception of emotion, human benchmark, automatic emotion classification

1. INTRODUCTION

"The measure of all things is man" (Protagoras; 490-411 BC), means in the context of emotion classification, that the results of the automatic recognition should always be compared with human perception capabilities, and of course their errors. Therefore, it is crucial to find features, and classification approaches comparable to the human perception, because otherwise it might happen that the machine recognizes expressions based on artifacts and not on actual modulation caused by humans' affective state.

Traditional approaches for emotion classification tend to rely on widely-accepted-to-perform-well sets of features, which are often immensely large and later pruned to perform especially well for automatic classification, but hardly ever to optimally fit human perception abilities [1].

While searching for suitable classifiers and discriminative feature sets, several different approaches towards the classification of acted emotional expressions from audio have been conducted in the literature especially on the standard Database of German Emotional Speech (EmoDB), also used as a reference approach in this study.

In [2], an analysis using the SVM-GMM supervector approach, utilizing PLP and ModSpec features, was conducted. The approach yields a classification performance of 79.0% on the seven available emotion categories, which is comparable to the result of the human perception test reported in Table 1 with 84% accuracy.

For example in [3] frame- and turn-level classification approaches have been compared with each other. Frame-level in this case means, that short audio segments (25ms in length) of the recorded utterances are used to estimate the portrayed emotions using Gaussian mixture models (GMMs). However, the single frame-wise decisions are again combined to form a turn-level or utterance based decision summing the likelihoods of each frame-wise decision. Using this majority vote-like approach and mel frequency cepstral coefficients (MFCC) features an accuracy of 77.1% is reached after identifying the optimal number of mixtures for the GMM beforehand. The turn-level approach, utilizing manifold features comprising pitch, energy, MFCC, formants etc. and several statistical functionals of them, such as mean, maximum, and standard deviation, performs slightly worse and yields 74.9% accuracy. After speaker normalization and feature selection the result improves to 83.2%. Further the combination of both frame- and turn-level approaches reaches an accuracy of 89.9%. However, as [3] remain unclear about the usage of validation sets for the optimization of the number of mixtures, parameters of speaker normalization, and feature selection the reported performance might be overestimated.

On the same set including all seven emotional categories in EmoDB, Wagner et al. in [4] thoroughly analyze and compare hidden Markov models (HMM) and support vector machines (SVM) using MFCC and energy based features. It is shown that HMM clearly outperform SVM on a word level analysis (36.6% accuracy for SVM vs. 48.6% accuracy for HMM), whereas on an utterance basis it is the other way around, with 73.3% accuracy for SVM and 69.5% accuracy for HMM. Slightly improved results, i.e. 77.4% accuracy, using similar features with additional pitch statistics and a correlation based feature selection, as in [5], are achieved in [6].

In this work, we propose a purpose-oriented choice of feature sets. There often exist features that do not perform well as a whole, but prove to be extremely powerful in very specific decisions. By analyzing the data from the point of view of standard features, it is possible to spot the

978-1-4673-0382-8/12/$31.00 ©2012 IEEE

Table 1. *Confusion matrix of the human performance test generated from the available labels for each of the utterances listed in the Database of German Emotional Speech.*

	F	D	H	N	S	A	B
Fear	**.85**	.04	.03	.03	.01	.04	.00
Disgust	.03	**.79**	.01	.04	.08	.02	.02
Happiness	.02	.02	**.83**	.06	.01	.05	.06
Neutral	.00	.00	.01	**.87**	.04	.02	.06
Sadness	.06	.02	.00	.06	**.78**	.00	.08
Anger	.01	.01	.01	.01	.00	**.96**	.00
Boredom	.00	.01	.00	.00	.11	.03	**.85**

strongest cross-class confusions. With this information, a further study of the human perception for solving these confusions leads to the design of a completely new feature set. The newly designed feature sets, in combination with the standard ones, render a better overall classification accuracy.

The remainder of the paper is organized as follows: Section 2 introduces the used datasets and the human perception abilities, reported as confusion matrices. Section 3 then describes the employed feature sets, as well as the encoding of sequential sets. The experimental setup is briefly described in Section 4 and the results are reported in Section 5. The automatic classification performances are then compared with the human perception in Section 6, and Section 7 concludes the paper.

2. DATASET DESCRIPTION

2.1. WaSeP Corpus

The experiments in this work are based on the "Corpus of spoken words for studies of auditory speech and emotional prosody processing" (WaSeP©) [7], which consists of two main parts: a collection of German nouns and a collection of phonetically balanced pseudo words, which correspond to the phonetical rules of German language, such as "hebof", "kebil", or "sepau". For this study the pseudo words have been chosen as the basis. This pseudo word set consists of 222 words, repeatedly uttered by a male and a female actor in six different emotional prosodies: neutral, joy, sadness, anger, fear, and disgust. The average duration of the speech signals depends on the specific emotion, ranging from 0.75 sec. in the case of the "neutral" prosody, to 1.70 sec. in the case of "disgust". The data was recorded using a Sony TCD-D7 DAT-recorder and the Sennheiser MD 425 microphone in an acoustic chamber with a 44.1 kHz sample rate and later down-sampled to 16 kHz with a 16 bit resolution. Furthermore, a perception test has been conducted with 74 native German listeners, who were asked to rate and name the category or prosody that they were just listening to, resulting in an overall accuracy of 78.53%. Table 2 shows the confusion matrix of the human perception test. It was also observed that the most confused emotion is "disgust", which is conform with the assumptions of [8].

Table 2. *Confusion matrix of the human performance test generated from the available labels for each of the utterances listed in the WaSeP database,* [12]

	F	D	H	N	S	A
Fear	**.77**	.01	.08	.03	.10	.01
Disgust	.05	**.72**	.06	.03	.07	.07
Happiness	.01	.00	**.75**	.22	.02	.00
Neutral	.01	.02	.05	**.79**	.00	.13
Sadness	.05	.01	.04	.13	**.76**	.01
Anger	.01	.03	.00	.01	.01	**.94**

2.2. Berlin Database of Emotional Speech (EmoDB)

As a reference approach we used the well known Database of German Emotional Speech (EmoDB) recorded in an anechoic chamber at the Technische Universität Berlin, Technical Acoustics Department. The audio was recorded at 48 kHz using a Sennheiser MKH40 P48 microphone and a Tascam DA-P1 portable DAT recorder and later down-sampled to 16 kHz [9]. The database comprises recordings of ten actors equally distributed with respect to gender. The actors were asked to speak ten selected sentences, both short and long sentences, in seven full blown emotions, comprising the same emotions as for the WaSeP corpus and boredom. The database has been the basis of many analysis [10, 11, 3, 4].

In order to ensure the quality of the recordings a human perception test with 20 subjects was performed to benchmark the quality of the recorded data. The subjects listened to the utterances in a random order and were allowed to listen to the samples only once. Out of more than 800 utterances around 500 fulfilling the requirements of 80% recognition and 60% naturalness were chosen for further analysis. The perception test overall yielded a mean accuracy of around 84% [9]. We derived the confusion matrix for all six emotions, which is as shown in Table 1. The values in the table are listed in percent of the utterances in one emotion.

3. FEATURES

3.1. Feature Selection

In similar work, different combinations of audio features are said to perform well in classification of audio data [13]. Given the characteristics of the used data set, the chosen features for this work are the following:

1. *MFCC / ΔMFCC*: based on the human perceptual scale of pitches. For the *MFCC* extraction a window length of 25 ms and a shift time of 10 ms is used, with a total of 20 cepstral coefficients, as well as their derivatives [14].

2. *modSpec*: implemented in an attempt to measure the *modulation* of the spectral coefficients. This is a way of accounting how much and how fast the features vary over time [15, 16].

3. *Voice Quality*: the dynamic use of voice qualities in spoken language can reveal useful information

on a speakers attitude, mood and affective states. The exact set of the utilized features is described in detail in [17].

4. f_0: it is possible to obtain different values of f_0 over time. From the f_0 trail different statistics are calculated: mean, standard deviation, maximum and quartile values, forming the feature set [14].

5. *Energy*: the frame average energy is calculated using a window size of 32 ms with an overlap of 16 ms. Similar statistics to those of f_0 are used for this [14].

6. *PLP*: perceptual linear predictive (PLP) analysis is based on perceptually and biologically motivated concepts, the critical bands, and the equal loudness curves, as described in [18].

7. *Periodicity*: As introduced in Section 3.2.

3.2. Periodicity

Previous analysis of the utilized dataset showed that some of the emotional expressions can be discriminated by using the segment lengths as a feature [12]. Since our aim is to find features that can be obtained in any context, the use of this property would result in unfair improvement of our results. Therefore, we developed a new feature set that can partly resemble that information.

Considering the number of syllables per second we do not incur in the use of unfair measures as it can be estimated from the signal directly and therefore in future applications. Syllable detectors may be implemented in different ways as shown in [19, 20, 21]. The utilized approach in this study is described in the following:

Let us assume that each syllable contains at least one vowel. If we consider the high periodicity that characterizes vowels in contrast to consonants, detecting speech segments with a high periodicity would give us markers of the syllables. A straight forward approach for obtaining a periodicity value is to compute the auto-covariance function in smaller sub-segments of the original speech. Once this step is completed, the sub-segments can be grouped according to their periodicity score as periodic or non-periodic. In order to achieve this, we designed a double-threshold system to resemble a hysteresis cycle [22]. This system marks the beginning of a periodic zone when a value over 80% of the maximum is found. In a similar way, the start of a non-periodic zone will be detected by the presence of a lower value than 30% of the maximum.

Additionally, for detecting syllables we consider energy variations over time. While assuming lower energy segments within the syllable boundaries, one may use an envelope detector and, once again, the double-threshold system to spot the syllables.

With the initial speech segment divided into periodic and non-periodic subparts, as well as, low and high energy parts, we calculate: the lengths of the parts, the largest difference in width, and the energy ratio of the parts to the total length. The combination of those features resembles the full periodicity feature set.

3.3. Feature alignment

Emotion classification from speech data proves to be a challenging problem due to the sequential nature of the data. Therefore, dynamic features extracted on short segments of speech (usually around 32ms windows) are useful for the classification of expressive clips. However, in order to be able to compare these sequential features with static features it is necessary to encode them into vectors of a fixed length. There exist different approaches for dealing with this type of situations. In this work, the use of HMM, as in [23], are used to encode the sequential data to a new representation space, where every sequence can be represented in terms of a fixed number of dimensions.

Let us assume a reference set of sequential observations $\mathcal{R} = \{O_1, ..., O_R\}$, where $O_i, \forall i = 1, ..., R$, is an observation without restriction of length. It is possible to train, for each of the reference observations, an HMM that best represents the hidden model that produced it. Therefore, we can easily obtain a set $\lambda = \{\lambda_1, ..., \lambda_R\}$ of HMM where $\lambda_i, \forall i = 1, ..., R$, represents the HMM trained from the observed sequence O_i. Given an unseen sequential observation $\tilde{O}$, its representation as a single point in the R-dimensional encoding space is obtained by calculating its log-likelihood with respect to the trained reference HMM:

$$D_R(\tilde{O}) = \frac{1}{T} \times \begin{pmatrix} \log P(\tilde{O} \mid \lambda_1) \\ \log P(\tilde{O} \mid \lambda_2) \\ \vdots \\ \log P(\tilde{O} \mid \lambda_R) \end{pmatrix}, \qquad (1)$$

where T represents the length of the given sequence and $P(\tilde{O} \mid \lambda_i)$ is the likelihood of the i-th HMM, given the observation $\tilde{O}$. With this transformation, we create a new space of dimension R, where every observed sequence is represented as one single vector. In this Euclidean space, any standard techniques for classification (supervised learning, unsupervised learning, clustering, etc.) may be applied.

Additionally, since the feature spaces are usually very heterogeneous, data normalization is performed. During the training of the system, mean and standard deviation (μ_{train} and σ_{train}) are calculated in each feature domain and for each class, prior to the HMM training. To remove the effect of outliers, all values above and below the 95% and 5% percentiles, respectively, are discarded. With the normalized data, the HMM are trained and the same normalization values (μ_{train} and σ_{train}) are later used to normalize the unseen data in the test step, before calculating their likelihood values.

4. EXPERIMENTAL SETUP

The experiments carried out for this work were for male and female speakers, with only one speaker per gender.

From the initial set of audio segments, a subset is randomly chosen for training the HMM. For each different class c, $\forall c = 1, ..., C$, where C represents the number of classes, 2 HMM are trained. Since the used data set contains 6 different classes, the final number of HMM is $R = 12$. Every HMM is initialized with 2 states and 2 GMM per state. Each of the models is trained with 5 different audio segments for a number of iterations no longer than 30. Experiments with a higher number of states and mixture components proved not to give better results and, on the contrary, required much more computation time.

The trained HMM are used to convert all the remaining audio segments into vectors of a similar length, obtaining a different representation of every segment for each of the feature sets. After the conversion of the features to the new spaces is completed, training of the SVM is to be performed. In this work, one against one SVM are used, with a total of $\frac{n \cdot (n-1)}{2}$ SVM for every different feature, where n is the number of classes. With the trained SVM, a fuzzy classification is performed on the test vectors. Combining of the different fuzzy outputs obtained at each different feature space is performed by multiplication and normalization. For each gender set, ten fold cross validations with a 90% training and 10% test data-set split for the evaluation of SVM were conducted.

5. RESULTS

Confusion matrices have been computed to analyse decisions. Every row sums up to one, showing how much data from one class is classified by the system as belonging to any of the possible ones. The columns (which do not necessarily sum up to one) show how much data from all classes is classified as part of a given one.

5.1. Gender dependent

Tests were initially conducted on a gender dependent basis: one single male and female speaker for training and evaluation. The average accuracy for the male speaker over all classes is 87% and 84% for the female.

For the male, most classes perform above 80% (or even 90%) accuracy, except for happiness. In this case, the biggest confusion is with fear and neutral, which together take over 20% of the misclassified samples .

The results for female differ slightly with those obtained with the male test. Happiness shows still one of the lowest performances but this time second to fear. In the case of fear, most of the erratic decisions are misclassified as happiness. For happiness, the biggest confusion is with neutral.

5.2. Gender Independent

This experiment was conducted with data from both male and female speakers. In order to let the data be well modelled by the HMM, a larger number of these is used. In particular, for each feature set and class, 2 HMM are trained with male-only data and another 2 with female-only data, which doubled the dimensionality of the features. To train the SVM, also equal amount of data from male and female speakers was used. Results were calculated without considering whether the test samples were produced by a male or female speaker.

Table 3. *Confusion matrix for the gender-independent automatic classification experiments, conducted with the WaSeP dataset.*

	F	D	H	N	S	A
Fear	**.80**	.03	.08	.01	.02	.06
Disgust	.01	**.88**	.05	.00	.04	.03
Happiness	.08	.02	**.71**	.12	.04	.03
Neutral	.00	.01	.16	**.82**	.01	.00
Sadness	.02	.00	.03	.01	**.95**	.00
Anger	.01	.07	.02	.03	.00	**.86**

Table 4. *Accuracy rate & F_1 measures for the gender-independent automatic classification experiments, conducted with the WaSeP dataset. Only those for* MFCC, periodicity, f_0 *and the* fusion *of all the sets described in Section 3 are shown.*

(acc./F_1)	**MFCC**	**periodicity**	$\mathbf{f_0}$	**fusion**
Fear	.80/.81	.40/.44	.57/.55	.80/.83
Disgust	.71/.67	.66/.61	.52/.52	.88/.85
Happiness	.62/.62	.28/.31	.46/.51	.71/.70
Neutral	.81/.80	.65/.60	.53/.52	.82/.83
Sadness	.90/.90	.76/.71	.69/.65	.95/.92
Anger	.81/.84	.74/.75	.50/.52	.86/.87

The classification accuracy in the gender-independent test is only slightly lower than in the previous cases and remains at 84% accuracy in average (see Table 3). Once again, happiness produces the lowest number of hits, being highly confused with fear and neutral. Observing Table 4, it is seen that the fusion clearly improves the results. A paired t-test shows a highly statistically significant improvement for the fusion over the single best feature set, namely MFCC ($p < .001$). For example, in the case of disgust or happiness (the categories with the lowest accuracy), an increase of 0.08 in F_1 measure is achieved.

The F_1 measure is calculated using the following expressions: $F_1 = \frac{2 \cdot p \cdot r}{p+r}$, $p = \frac{tp}{tp+fp}$, $r = \frac{tp}{tp+fn}$, where *p*: precision, *r*: recall, *tp*: true possitives, *fp*: false possitives, *fn*: false negatives.

5.3. Results for EmoDB

For comparison of the results, tests are also carried out with the EmoDB dataset. Evaluation of the proposed system on this dataset shows a 77% accuracy, slightly lower than the 84% obtained for the WaSeP dataset. The confusion matrix for this test can be seen in Table 5.

6. DISCUSSION

The confusion matrices provided in Section 5 provide a good basis for the comparison of human and machine capabilities and errors. A first glance at them shows that human and machine performances are quite similar on an

Table 5. *Confusion matrix of the gender-independent automatic classification experiments, conducted with the EmoDB dataset.*

	F	D	H	N	S	A	B
Fear	**.77**	.01	.10	.01	.06	.00	.05
Disgust	.10	**.69**	.04	.04	.07	.01	.05
Happiness	.08	.02	**.53**	.00	.03	.00	.33
Neutral	.00	.01	.00	**.80**	.16	.03	.00
Sadness	.01	.01	.00	.06	**.92**	.00	.00
Anger	.00	.00	.00	.03	.07	**.89**	.00
Boredom	.04	.01	.14	.00	.00	.00	**.81**

overall scale. With the WaSeP dataset, the 84% accuracy rate obtained is exactly the same as that from humans in average. In the case of EmoDB, a 77% accuracy rate performed by automatic recognition compares to the 84.7% achieved by humans. The average scores for both datasets are very similar, however, there are a few patterns that seem to diverge quite strongly. First of all, with respect to the WaSeP dataset (compare tables 2 and 3) a lot of human perception errors are due to votes for the class neutral, which leads to the assumption that humans tend to vote for a class that does not provide clear evidence for the intended emotion. The machine does not vote for neutral that frequently, partly of course due to the fact that the word neutral does not bare any meaning to it.

Further, it is seen that happiness is badly recognized in both automatic classification experiments. For the human perception tests this does not hold. Even though, there are evidences in the literature that happiness is often difficult to recognize as reported in [24], where happiness only reached an accuracy of 48%. In [12] it is also reported that the recognition performance of humans with respect to the male recordings of happiness is at 66%, with the main confusions towards neutral speech, which is confirmed by the automatic classification in Table 3.

Anger is in both human perception experiments the best recognized emotion, whereas the automatic classification does not reach such high levels of accuracy. The classifier on the other hand reaches great recognition performances for the sad expressions, which humans seem to confuse quite frequently for both datasets. In [12] once more a gender difference is reported: the female sad expressions are only recognized at an accuracy of 65% and again mostly confused with neutral.

With the lowest human accuracy in the WaSeP experiments, there is disgust, which is in accordance to [25, 26]. It is also among the worst in EmoDB, only second to sadness. This effect was initially present in our experiments with the standard features. The design of a new feature set, as explained in Section 3.2, with a high F_1 value and accuracy for disgust permits an improvement on the fusion rates of about 20%. This effect can be observed in Table 4.

7. CONCLUSIONS

In the task of emotion classification, there is documented prove that humans perform with higher error rates than in other recognition tasks. In this work, we studied the errors in order to better comprehend the real emotional information in the communication process. Experiments for automatic classification were conducted on standard emotion datasets and the results were compared with human perceptual accuracies. However, there is still room for additional investigations of the single feature sets as well as the type of fusion performed, which could not be included in this paper due to space restrictions.

Further, as the utilized datasets were composed of acted speech segments, future work should include the study of natural data as well as a deeper knowledge of the representative characteristics of each different emotion.

8. ACKNOWLEGDEMENTS

The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

9. REFERENCES

[1] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[2] F. Schwenker, S. Scherer, Y. Magdi, and G. Palm, "The GMM-SVM supervector approach for the recognition of the emotional status from speech," in *19th International Conference on Artificial Neural Networks 2009 Part I*, C. Alippi, Ed., Berlin Heidelberg, 2009, vol. LNCS 5768, pp. 894–903, Springer.

[3] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, Berlin, Heidelberg, 2007, pp. 139–147, Springer-Verlag.

[4] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech," in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*, Berlin, Heidelberg, 2007, pp. 114–125, Springer-Verlag.

[5] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," in *Proceedings of the 21st Australian Computer Science Conference*. 1998, pp. 181–191, Springer.

[6] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proceedings of IEEE*

International Conference on Multimedia and Expo (ICME'05). 2005, pp. 474 –477, IEEE.

[7] B. Wendt and H. Scheich, "The "Magdeburger Prosodie Korpus" - a spoken language corpus for fMRI-studies," in *Speech Prosody 2002*. 2002, pp. 699–701, SProSIG.

[8] K. R. Scherer, T. Johnstone, and G. Klasmeyer, *Handbook of Affective Sciences - Vocal expression of emotion*, chapter 23, pp. 433–456, Affective Science. Oxford University Press, 2003.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Interspeech 2005*. 2005, pp. 1517–1520, ISCA.

[10] S. Scherer, M. Oubbati, F. Schwenker, and G. Palm, "Real-time emotion recognition from speech using echo state networks," in *Proceedings of the 3rd IAPR workshop on Artificial Neural Networks in Pattern Recognition (ANNPR'08)*, Berlin, Heidelberg, 2008, pp. 205–216, Springer.

[11] S. Scherer, F. Schwenker, and G. Palm, "Classifier fusion for emotion recognition from speech," in *3rd IET International Conference on Intelligent Environments 2007 (IE07)*. Sept. 2007, pp. 152–155, IEEE.

[12] B. Wendt, *Analysen Emotionaler Prosodie*, vol. 20 of *Hallesche Schriften zur Sprechwissenschaft und Phonetik*, Peter Lang Internationaler Verlag der Wissenschaften, 2007.

[13] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533 – 544, 2001.

[14] L. R. Rabiner, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[15] H. K. Maganti, S. Scherer, and G. Palm, "A novel feature for emotion recognition in voice based applications," in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction (ACII'07)*. 2007, pp. 710–711, Springer.

[16] H. Hermansky, "The modulation spectrum in automatic recognition of speech," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. 1997, pp. 140–147, IEEE.

[17] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *IEEE Transactions on Audio, Speech and Language Processing*, under review.

[18] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition*, vol. 2, pp. 578–589, 1994.

[19] H.R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proceedings of The 4th International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, 1996, vol. 2, pp. 1261–1264.

[20] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of The Acoustical Society of America*, vol. 88, pp. 101–112, 1990.

[21] H. J. Cedergren and H. Perreault, "Speech rate and syllable timing in spontaneous speech," in *Third International Conference on Spoken Language Processing (ICSLP 94)*. 1994, pp. 1087–1090, IEEE.

[22] I. Mayergoyz, *Mathematical Models of Hysteresis and their Applications*, Springer, 2003.

[23] M. Bicego, V. Murino, and M. Figueiredo, "Similarity-based clustering of sequences using hidden markov models," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner and A. Rosenfeld, Eds., vol. 2734, pp. 95–104. Springer, 2003.

[24] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, pp. 76–92, 2001.

[25] R. Van Bezooyen, *Characteristics and Recognizability of Vocal Expressions of Emotion*, Foris Pubns USA, 1984.

[26] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and Emotion*, vol. 15, pp. 123–148, 1991.