Multiple Classifier Systems for the Classification of Audio-Visual Emotional States

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, and Friedhelm Schwenker

Ulm University, Institute of Neural Information Processing, 89081 Ulm, Germany

Abstract. Research activities in the field of human-computer interaction increasingly addressed the aspect of integrating some type of emotional intelligence. Human emotions are expressed through different modalities such as speech, facial expressions, hand or body gestures, and therefore the classification of human emotions should be considered as a multimodal pattern recognition problem. The aim of our paper is to investigate multiple classifier systems utilizing audio and visual features to classify human emotional states. For that a variety of features have been derived. From the audio signal the fundamental frequency, LPC- and MFCC coefficients, and RASTA-PLP have been used. In addition to that two types of visual features have been computed, namely form and motion features of intermediate complexity. The numerical evaluation has been performed on the four emotional labels *Arousal, Expectancy, Power, Valence* as defined in the AVEC data set. As classifier architectures multiple classifier systems are applied, these have been proven to be accurate and robust against missing and noisy data.

1 Introduction

Research in affective computing aim to provide simpler and more natural interfaces for human-computer interaction applications. In this kind of applications detecting and recognizing the emotional status of an user is important in order to develop efficient and productive human-computer interaction interfaces [3]. Analysis of human emotions and processing recorded data, for instance the speech, facial expressions, hand gestures, body movements, etc. is a multidisciplinary field that has been emerging as a rich area of research in recent times [5,11,20,24,21,27]. In this paper multiple classifier systems for the classification of audio-visual features have been investigated, the numerical evaluation of the proposed emotion recognition systems has been carried out on the data sets of the AVEC challenge [23]. Combining classifiers is a promising approach to improve the overall classifier performance [25,19]. Such a team of classifiers should be accurate and diverse [9]. While the requirement to the classifiers to be as accurate as possible is obvious, diversity roughly means that classifiers should not agree on the set of misclassified data. Various feature views on the audio and visual data are utilized to achieve such a set of diverse and accurate classifiers.

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 359–368, 2011.

2 Features

In the following section we briefly describe the features we extracted within the audio and visual channel, that served as input for the successive classification architecture.

2.1 Audio Features

For the audio analysis we extracted a variety of standard features:

Fundamental Frequency (f_0) and Energy. From each speech segments the f_0 values are extracted, using the f_0 tracker available in the ESPS/*waves+*[1] software package. This f_0 track as well as the energy of the plain wave signal is extracted from 32 ms frames with an offset of 16 ms. Both one dimensional signals are combined with the eight dimensional LPC signal to form a ten dimensional feature vector.

Linear Predictive Coding Coefficients (LPC). Linear predictive coding (LPC) is a popular method to represent the spectral envelope, which corresponds to a curve fitting tightly around the peaks of the short time log magnitude spectrum of a signal, in a highly compressed manner. As described in [14] the main reasons for the popularity of LPC are, that especially for steady state voiced regions of speech, such as vowels with constant vocal source pressure and tract, LPC provides a good approximation of the spectral envelope of the signal. However, the main flaws of LPC are in the representation of unvoiced regions of an utterance. Furthermore, the straightforward and computationally cheap method to extract the LPC coefficients outperforms most of the other methods to approximate the spectral envelope.

Mel Frequency Cepstral Coefficients (MFCC). The Mel frequency cepstral coefficient (MFCC) representation is motivated by biological factors, as the known perceptual variations in the human ear are modeled using a filter bank with filters linearly spaced in lower frequencies and logarithmically in higher frequencies in order to capture the phonetically important characteristics of speech [4]. The MFCC are extracted, following [28]:

1. The speech signal is divided into windowed frames of 25 ms in size with an offset of 10 ms. A Hamming window is applied to reduce discontinuities in the spectrum at the end of the frame.
2. For each frame calculate amplitude spectrum using short-term fast Fourier transform (FFT).
3. Apply a filter bank of triangular filters that are equally spaced in the Mel scale.
4. Take the log energy of every filter output.
5. Take discrete cosine transform (DCT) yielding de-correlated cepstral coefficients for each frame.

[1] `http://www.speech.kth.se/software/`

MFCC are quite commonly used features based on short-time spectrum in speech recognition tasks, since they are compact representations of the speech and its spectral envelope with the characteristic to retain most of the phonetically significant acoustic information [4]. As mentioned in [4] the key features of the MFCC include the following important points:

- Parameters such as MFCC derived from the short-term Fourier spectrum preserve acoustic information to a larger extent than those relying on LPC.
- MFCC allow an improved suppression of higher frequencies that are less relevant parts of the spectrum for speech applications.
- MFCC allow a very compact representation of the acoustic signal since only a few coefficients suffice for the most relevant data (mostly 8-24 coefficients).

Relative Spectral Perceptual Linear Predictive Coding (RASTA-PLP). The perceptual linear predictive (PLP) analysis is based on two perceptually and biologically motivated concepts, namely the critical bands, and the equal loudness curves [6,17]. Frequencies below 1 kHz need higher sound pressure levels than the reference, and sounds between 2 - 5 kHz need less pressure, following the human perception.

The critical band filtering is analogous to the MFCC triangular filtering, apart from the fact, that the filters are equally spaced in the Bark scale (not the Mel scale) and the shape of the filters is not triangular, but rather trapezoidal.

After the critical band analysis and equal loudness conversion, the subsequent steps required for the relative spectral (RASTA) processing extension, follow the implementation recommendations in [8]. After transforming the spectrum to the logarithmic domain and the application of RASTA filtering, the signal is transformed back using the exponential function.

The last steps are according to the estimation of the LPC coefficients. The LPC coefficients are calculated over the critical band energies of a single frame, which is followed by the transformation of the LPC coefficients to cepstral values. In [7], PLP speech analysis was first introduced as a method to represent speech signals with respect to the human perception and with as few parameters as possible. However, PLP was, as most of the other analysis techniques, sensitive towards steady-state spectral factors caused by transmission channels, such as telephone recordings or the usage of different microphones [8]. Therefore, [8] introduced the RASTA methodology for PLP rendering it more robust towards these channel distortions and reducing error rates in several speech recognition experiments with only a slightly more computationally expensive extraction method. In this study 21 critical bands are analyzed in frames of 25 ms in length and with a 10 ms offset.

2.2 Visual Features

Initial processing is organized along two mainly independent pathways, each specialized for the processing of form and shape as well as the processing of motion, respectively. The organization of both pathways is hierarchical, i.e. along

a series of processing stages with increasingly larger scales of interaction [18]. Form processing is mainly orientation selective and combines activities to build representations of localized features and shape configurations. Motion processing, on the other hand, is direction selective and combines activities to build representations of flow discontinuities as well as motion patterns. We assume that the hierarchical processing in both pathways is organized in a similar fashion and, thus, make use of a generic processing architecture for neural feature extraction as shown in Figure 1. The architecture is a modified variant of the object-recognition model proposed by [10,16,26].

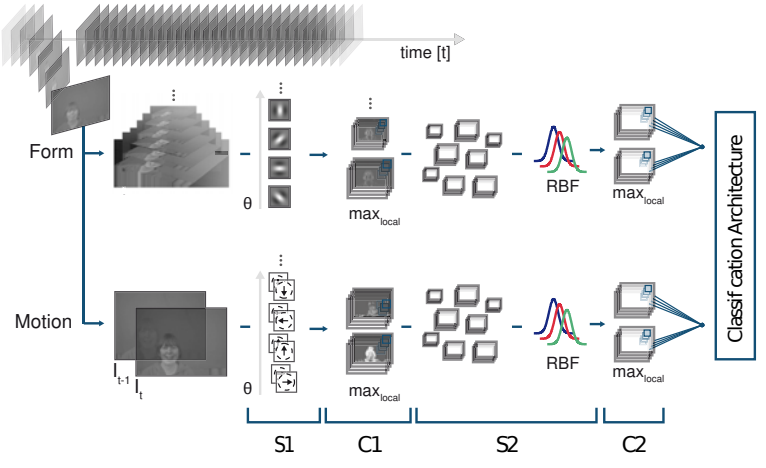

Fig. 1. Visual feature extraction. Motion and form features are processed along two separate pathways, composed of alternating layers of filtering (S) and non-linear pooling (C) stages. In layer S1, different scale representations of the input image are convolved with 2D Gabor filters of different orientations (form path) and a spatio-temporal correlation detector is used to build a discrete velocity space representation (motion path). Layer C1 cells pool the activities of S1 cells of the same orientation (direction) over a small local neighborhood and two neighboring scales and speeds, respectively. The layer S2 is created by a simple template matching of patches of C1 activities against a number of prototype patches. These prototypes are randomly selected during the learning stage (for details, see [10]). In the final layer C2, the S2 prototype responses are again pooled over a limited neighborhood and combined into a single feature vector which serves as input to the successive classification stage.

Initial Feature Detection for Form and Motion Processing. The model architecture consists of a series of stages consisting of alternating levels of filtering and pooling steps (S- and C-layers, respectively). These stages operate at different scales of spatial neighborhood.

– Oriented contrast detection. For the generation of initial contrast representation, an input image is transformed into a pyramid of 9 different spatial scales, with a downscaling factor of $2^{\frac{1}{4}}$ (using bicubic interpolation). Each scale is convolved with a bank of 2D Gabor filters given by

$$G(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right)\cdot\cos\left(\frac{2\pi}{\lambda}X\right), \quad (1)$$

where $X = x\cos\theta - y\sin\theta$ and $Y = x\sin\theta + y\cos\theta$ (for performance reasons only the even part is used). The variables x and y range from -5 to 5 and θ varies between 0 and π in steps of $\pi/6$. The aspect ratio γ, the effective width σ, and the wavelength λ are set to 0.3, 4.5, and 5.6, respectively (in accordance with [10]. This results in six orientation maps at each scale.

– Directional motion detection. The initial motion representation is generated by utilizing a spatio-temporal correlation detector which is quantized into 12 movement directions and two different speeds to build a discrete velocity space representation [1].

These maps represent the initial S1 layer for the form and motion path.

The activations are subsequently pooled over a small 10×10 spatial neighborhood and integrated over two neighboring scales by a local maximum operation. This operation forms the C1 layer representations in the form and motion pathway. The non-linear pooling operation by max-selection achieves an input pattern invariance against variations in position and size. The distributed activity maps are subsequently pruned by a combined thresholding and normalization step. Minimum and maximum activities, R_{min} and R_{max}, are determined at each location of S1 or C1 responses. Responses R of an S1/C1-unit are set to zero if $R < R_{min} + h(R_{max} - R_{min})$ where $h = 0.5$ denotes the inhibition level.

Intermediate-Level Feature Processing. In the successive processing stage, intermediate level features are learned by selecting the most descriptive and discriminative prototypes among an exhaustive number of response patches. This is achieved by randomly sampling the C1 responses. The resulting S2-prototype patterns P denote filters with complex feature selectivities topographically organized around the spatial locations of their most likely spacial occurrence. The response of an $n\times n$ patch of C1 units X to a particular S2-prototype P is calculated by

$$R(X,P) = \exp\left(-\frac{\|X-P\|^2}{2\sigma^2\alpha}\right), \quad (2)$$

implementing a Gaussian radial basis function to weight the allowed degree of dis-similarity w.r.t. prototypical patterns in the shape or motion domain, respectively. To further increase generalization, only the dominant activities at each spatial location of a patch P is taken into account. The standard deviation of the Gaussian σ is set to 1. A patch of C1 units X as well as the S2-prototype P have dimensions $n\times n\times 6$ in the form path, while $n\times n\times 12$ in the motion path ($n\in\{4,8,12,16\}$). To obtain a larger degree of similarity in higher dimensional space. The normalization constant is set to $\alpha = (n/4)^2$ in the case of $n > 4$.

At the final stage, responses from all prototypical complex filters as generated in the S2 representation are again pooled over a limited spatial neighborhood. This process selectively operates on the different spatial scales. The responses

are combined into a single C2 feature vector which serves as input to the classification architecture described in Section 3. For the implementation of the visual feature extraction architecture, we used the *CNS: Cortical Network Simulator* as described in [13].

3 Classifier Architectures

In this section the proposed multi classifier systems for the different sub-challenges are described.

3.1 Audio Sub-Challenge

For each label dimension and and for each audio feature a bag of hidden Markov models (HMM) have been trained [2,15]. The hidden states and the number of mixture components of the HMM have been optimized using a parameter search resulting in the selection of three hidden states and two mixture components in the Gaussian mixture model (GMM) having full covariance matrices.

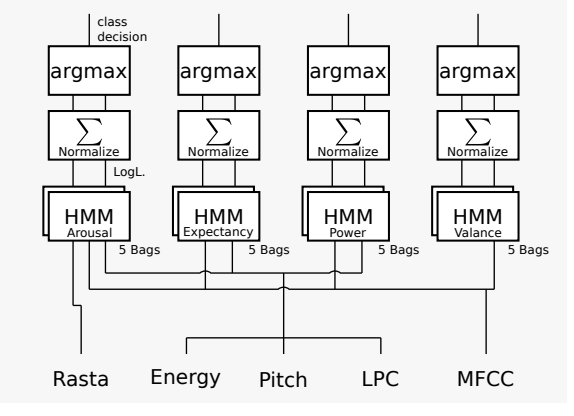

Fig. 2. Audio classifier system. For each label a bag of HMM have been trained on selected features sets.

The evaluation of the optimization process further inferred that some features appear to be inappropriate to detect certain labels. It turned out, that only the label *Arousal* can draw information from all features, *Expectancy* and *Power* performed better using only the energy, fundamental frequency and the MFCC features. The label *Valance* favored only the MFCC features. For each label the log likelihoods of every HMM trained on the features are summed. To obtain more robust models, we decided to additionally use five times as many models per class and summed the outcome as well.

Furthermore, the assumption was made that the labels are changing only slowly over time. We therefore conducted the classification on turn basis by

collecting the detections within one turn and multiplied the likelihoods to obtain more robust detections. A schema visualizing the applied fusion architecture is shown in Figure 2. The results of this approach are reported in Table 1.

3.2 Video Sub-Challenge

Within the video challenge the ν-SVM[2] was employed as the central classifier [22]. We concatenated 300 form and 300 motion feature and used them to train a ν-SVM having a linear kernel and probabilistic outputs according to Platt [12]. Due to memory constraints only 10.000 randomly drawn samples were used. Again a parameter search was applied to obtain suitable parameters,

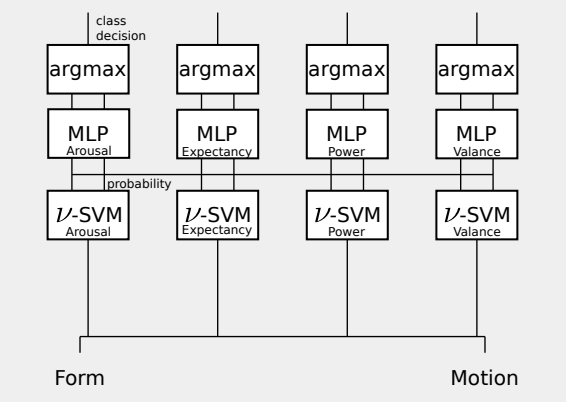

Fig. 3. Video classifier system. The form and motion features are concatenated and used to train ν-SVM for each label dimension. The outputs of the classifiers are used to train a intermediate fusion layer realized by multi layer perceptrons.

resulting in setting ν to 0.3 for *Arousal* and *Power* and 0.7 for *Expectancy* and *Valence*. Based on the results of all label dimensions an intermediate fusion was conducted using a multi layer perceptron (MLP) to obtain the final prediction. A schema illustrating the architecture used is shown in Figure 3. The results are reported in Table 1.

3.3 Audio-Visual Sub-Challenge

Considering the audio-visual challenge, we used the same approach for each modality as described in the earlier sections but omitted the last layer in which the class decision was performed. The probabilistic outputs of the video stream are collected using averaging and multiplication with a subsequent normalization such that the decision are on word level. The HMM log likelihood of the label

[2] The implementation was taken from the well-known *libsvm* repository.

dimensions are transformed and normalized such that they are ranging between zero and one. By concatenating the results of all label dimensions, a new 12 dimensional feature vector is obtained. The new features are then used to train an intermediate fusion layer based on a MLP. Like in the audio challenge, the final decision is done on a turn basis by collecting the outputs within one turn and fusing them using multiplication. Figure 4 shows the audio-visual classifier system, while the results are given in Table 1.

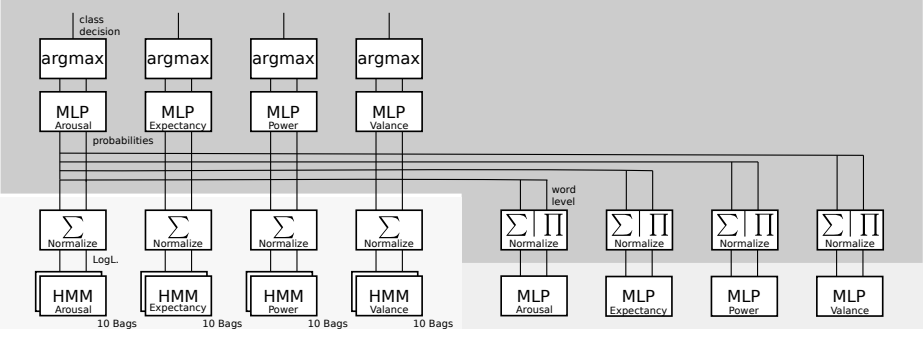

Fig. 4. Audio-visual classifier system. The output of all modalities are collected on word level and used to train a multi layer perceptron for each label dimension.

Table 1. Classification results. The weighted accuracy (WA) corresponds to the correctly detected samples divided by the total number of samples. The unweighted accuracy (UA) is given by the averaged recall of the two classes of a label dimension.

	Arousal		Expectancy		Power		Valence	
Data set	WA	UA	WA	UA	WA	UA	WA	UA
Audio sub-challenge								
Develop	66.9	67.5	62.9	58.5	63.2	58.4	65.7	63.3
Test	63.5	65.7	41.1	41.4	43.3	29.9	65.4	65.4
Video sub-challenge								
Develop	58.2	53.5	53.5	53.2	53.7	53.8	53.2	49.8
Test	56.9	57.2	47.5	47.8	47.3	47.2	55.5	55.5
Audiovisual sub-challenge								
Devel	69.3	70.6	61.7	60.1	61.3	59.1	68.8	66.4
Test	54.2	54.3	58.5	57.8	42.7	40.0	44.8	35.9

4 Discussion

The results presented in Table 1 are preliminary and must be evaluated in several directions: 1) Feature extraction techniques as described in the previous sections have been successfully applied to the recognition of Ekman's six basic emotions for benchmark data sets consisting of acted emotional data. In these data sets emotions shown by the actors are usually over-expressed and different from the

emotional states that can be observed in the AVEC data set. 2) The classifier architecture is based on the so-called late fusion paradigm. This is a widely used fusion scheme that can be implemented easily just by integrating results of the pre-trained classifier ensemble by fixed or trainable fusion mappings, but more complex spatio-temporal patterns on an intermediate feature level can not be modeled by such decision level fusion scheme. 3) Emotional states of the AVEC data set are encode by crisp binary labels that difficult to get from human annotators. They have usually problems to assign a confident crisp label to an emotional scene (e.g. single spoken word or a few video frames), and thus dealing with fuzzy labels or labels together with a confidence value during annotation and classifier training phase could improve the overall recognition performance.

Acknowledgments. This work was been supported by a grant from the Transregional Collaborative Research Center SFB/TRR62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Bayerl, P., Neumann, H.: A fast biologically inspired algorithm for recurrent motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(2), 246–260 (2007)
2. Breiman, L.: Bagging predictors. Machine learning 24(2), 123–140 (1996)
3. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. Signal Processing Magazine 18(1), 32–80 (2001)
4. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Transactions on Acoustics, Speech and Signal Processing 28(4), 357–366 (1980)
5. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18(4), 407–422 (2005)
6. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87(4), 1738–1752 (1990)
7. Hermansky, H., Hanson, B., Wakita, H.: Perceptually based linear predictive analysis of speech. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 10, pp. 509–512. IEEE, Los Alamitos (1985)
8. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: RASTA-PLP speech analysis technique. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 121–124. IEEE, Los Alamitos (1992)
9. Kuncheva, L., Whitaker, C.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51(2), 181–207 (2003)
10. Mutch, J., Lowe, D.: Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision 80(1), 45–57 (2008)
11. Oudeyer, P.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies 59(1-2), 157–183 (2003)

12. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74 (1999)
13. Poggio, T., Knoblich, U., Mutch, J.: CNS: a GPU-based framework for simulating cortically-organized networks. MIT-CSAIL-TR-2010-013/CBCL-286 (2010)
14. Rabiner, L., Juang, B.: Fundamentals of speech recognition. Prentice-Hall Signal Processing Series (1993)
15. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. IEEE 77(2), 257–286 (1989)
16. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience 2, 1019–1025 (1999)
17. Robinson, D.W., Dadson, R.: A re-determination of the equal-loudness relations for pure tones. British Journal of Applied Physics 7, 166–181 (1956)
18. Rolls, E.: Brain mechanisms for invariant visual recognition and learning. Behavioural Processes 33(1-2), 113–138 (1994)
19. Schels, M., Schwenker, F.: A multiple classifier system approach for facial expressions in image sequences utilizing GMM supervectors. In: International Conference on Pattern Recognition (ICPR), pp. 4251–4254 (2010)
20. Scherer, S., Schwenker, F., Palm, G.: Classifier fusion for emotion recognition from speech. In: Advanced Intelligent Environments, pp. 95–117 (2009)
21. Schmidt, M., Schels, M., Schwenker, F.: A hidden markov model based approach for facial expression recognition in image sequences. In: Schwenker, F., El Gayar, N. (eds.) ANNPR 2010. LNCS(LNAI), vol. 5998, pp. 149–160. Springer, Heidelberg (2010)
22. Schölkopf, B., Smola, A.J., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Computation 12(5), 1207–1245 (2000)
23. Schuller, B., Valsta, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: The first international audio/visual emotion challenge and workshop (AVEC 2011). In: D´Mello, S., et al. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 415–424. Springer, Heidelberg (2011)
24. Schwenker, F., Scherer, S., Magdi, Y.M., Palm, G.: The GMM-SVM supervector approach for the recognition of the emotional status from speech. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part I. LNCS, vol. 5768, pp. 894–903. Springer, Heidelberg (2009)
25. Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M.: Multiple classifier systems for the recogonition of human emotions. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 315–324. Springer, Heidelberg (2010)
26. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 994–1000 (2005)
27. Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal emotion classification in naturalistic user behavior. In: Jacko, J.A. (ed.) HCI International 2011, Part III. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)
28. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. Journal of Computer Science and Technology 16(6), 582–589 (2001)