On the Influences of Feature Extration in Single Emotion Recognition in Naive vs. Acted Speech

Ronald Böck, David Hübner, and Andreas Wendemuth

Otto-von-Guericke-University Magdeburg
Faculty of Electrical Engineering and Information Technology
Institute for Electronics, Signal Processing, and Communications
Chair of Cognitive Systems
ronald.boeck@ovgu.de

Abstract: Generally, in communication several aspects have to be considered: 1) The communicated information itself, 2) the non-verbal information, i.e. poses and gestures, and 3) the emotional part of communication. All parts are necessary if a dialogue shall be successful and effective. Extracting the information from “what is said”, is the issue of the automatic speech recognition and, thus, provides the contents of a dialogue. The non-verbal information is usually faced by image processing and is not object of this paper. The last item is related to both. Hence, in this paper we focus on recognising emotions from speech. Therefore, we investigate the influences of different feature sets on emotion recognition. Moreover, we also compare two approaches of recognition: Hidden Markov Models and Artificial Neural Networks.

1 Introduction

Communicating with the aid of speech is the most intuitive way to submit information. Almost everybody is interested to use this form to handle communication. If we look at discussions between humans, also non-verbal parts of the interaction are important or sometimes more significant than the context, to have a successful dialogue. Every dialogue is, therefore, influenced by the mimics, gestures, and poses of the interacting partners. All are driven by the emotional state.

In this paper we will use the term “emotional state” in the following way: this term summarises all parts of emotional meanings. Each user is influenced by different moods: 1) the physiological one, i.e. the biological reactions of the body, 2) the intuitive moods, and 3) the feelings which are mentally recognised by the user. The second and third reactions could be called moods (according to [1]). Here, we subsume all these meanings under “emotional state”.

The above mentioned circumstances are necessary for a successful dialogue. In particular, we are interested to provide all essential information for this goal by using speech. In fact, we know that it is possible to recognise the uttered expression and, if the domain is more or less fully specified, to interpret the meaning of this utterance. If the domain is not or weak specified or restricted we might get problems (for a possible solution see [5]). So, one can imagine that the recognition of the emotional state from speech is really a hard task. Moreover, the situational state of the user should not be neglected if a dialogue should be led to success, i.e. if one knows the situational state of the counterpart (it does not matter if this is a real user or an artificial system), one could react in several ways. For instance, a user who is well informed about the content of the dialogue should be handled in a different way like one who is displeased with this current situation. In this paper we will concentrate on the emotional state.

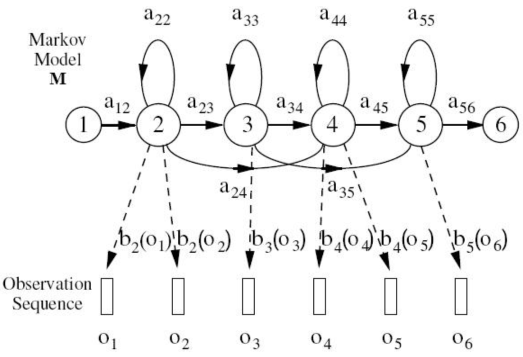


Figure 1 - General structure of an HMM [7]

All aforementioned strategies are really easy for humans; the difficulty is to build artificial systems which have the same or almost the same characteristics like the human counterpart. The more intuitive the system is, the more successful the dialogue can be. It is pointed out, that we focus just on single emotions, i.e. sequential flows of emotional states, e.g. neutral-anger-surprised-neutral, are not considered.

In the context of this paper, we investigate the capabilities of several well known feature sets in the focus of emotion recognition. The goal is to provide and to identify features which are, perhaps, suitable for different emotions. The features we used are Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction cepstral coefficients (PLPs), and Linear Prediction Filter Coefficients (LPCs). Methods for recognising the emotional state are the Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs) with several architectures.

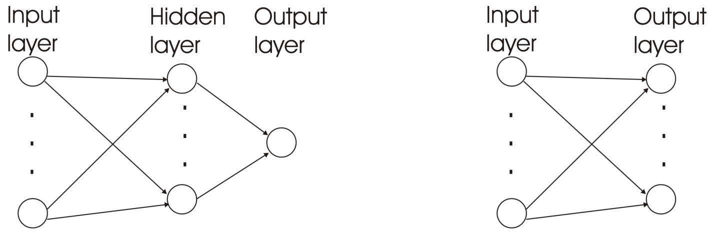
The current paper is organised as follows: Section 2 gives a briefly introduction in the Hidden Markov Model and Section 3 presents the architectures of the used Artificial Neural Networks. Section 4 regards the different feature sets which are the focus of this paper. Finally, Section 5 discusses the results of the experiments grouped by the two approaches.

2 Hidden Markov Models

The Hidden Markov Models are stochastic models which are known in different fields of signal processing. The principle structure of a left-to-right HMM is visualised in figure 1 (notice, that the visualisation is according to the specifications in [7]).Generally, HMMs are powerful in image processing as well as in speech processing and recognition. Such a model is a finite state automata which changes from state i to state j in each time slot, where i, j are elements of the state number set. While traversing the model an observation sequence o_i is produced according to a probability density $b_i(o_i)$. Also the hidden values a_{ij} are probabilistic. The training process of HMMs is done with the aid of the Baum-Welsh-Algorithm and the most likely observation sequence is computed by the Viterbi-Algorithm.

In particular, we are using the HMMs as a recogniser for emotional states, i.e. each model represents exactly one emotional state. According to phoneme-based speech recognition, sequences of different emotional states can be observed or recognised by combining multiple HMMs. On the other hand, the same models can be used to identify single emotions.

In both cases, the observation sequence $E = e_i$ is the most likely sequence of emotional states given a data set which includes emotional speech. For this, the smallest unit which can be recognised is one emotion (cf. phonemes in speech recognition, for instance). Hence, $P(E \mid M)$ is the probability of a whole sequence of emotional states given a set of models $M = \{M_i\}$.



(a) ANN with hidden layer and classification unit

(b) ANN (output: 1-of-N)

Figure 2 - Different ANN architectures used in experiments

Finally, with an emotional word w_e we are able to write

$$P(E \mid w_e) = P(E \mid M_i) \tag{1}$$

In the focus of this paper, as aforementioned, we investigate single emotions and thus e_i is e, i.e. the sequence consists of one emotional state e, only.

3 Artificial Neural Networks

Artificial Neural Networks are biologically inspired mathematical models of neural networks. They are widespread distributed in processing and computational topics with several objectives, e.g. pre- and post-processing of (video-)streams or classification which is the most favourite use case.

In our experiments we are using ANNs to classify the emotional state of an user. This classification bases on speech signal features (including one emotion per audio file).

For this, we use two networks with different architectures which are visualised in figure 2. We differ between two kinds of network realisations: 1) Is a network which has two layers (see figure 2(a)). In this case, the number of hidden neurons is equal to the number of emotions, which varies according to the used database. The output neuron is for interpretation, i.e. this represents the hidden neuron which has the highest activation in the previous layer. 2) In figure 2(b) a corresponding ANN is shown which output is (ideally) an 1-of-N decision. In fact, in applications a threshold is necessary to get a clear decision. For both architectures the input layer is not counted in this context. Moreover, the number of used neurons depends on the observed feature sets, i.e. the number of extracted features. This flexibility is a great advantage of ANNs.

Finally, it is necessary to specify which kind of neurons we have used. Since the input layer is just for getting the information, we decided to use linear neurons. This is also the type of the classification neuron in architecture I. The "real" processing neurons, which are in the hidden layer, are activated by dot product of the weights and the input (see equation 2) and provide an output function as in equation 3 (this is a tuned hyperbolic tangent) or equation 4 - we tested both functions.

$$t_i = \underline{w} \bullet \underline{x} \tag{2}$$

$$y_i = 0.5tanh(t_i) + 0.5 \tag{3}$$

$$y_i = tanh(t_i) \tag{4}$$

4 Feature Sets

In this section we will discuss the feature sets which are used in the different training processes. Due to the various possibilities and advantages of the approaches, we selected several feature sets.

Table 1 - Feature sets related to HMMs

Feature sets 1	Feature sets 2
MFCC_E_D_A	MFCC_0_D_A
PLP_E_D_A	PLP_0_D_A
LPC_E_D_A	- - -[1]

First of all, we concentrate on standard features which are widely used in speech recognition. Namely, these are Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Predictive cepstral coefficients (PLPs), and Linear Perception Filter Coefficients (LPCs). Moreover, for each of the coefficients we added the energy term and, in comparison, the 0th cepstral coefficients. Furthermore, the delta ($+\Delta$) and acceleration ($+\Delta\Delta$) values are computed for each feature set's coefficient which is shown in table 1. So far, for ANN training we concentrate on MFCCs as a basis to generate the features observed. Inspired by [3], we extracted the following statistical meta-features (based on MFCCs):

- mean value m
- (0.1, 0.25, 0.5, 0.75, 0.9)-quantile of F_j
- standard deviation σ of F_j

with $F_j = \{MFCC, \Delta MFCC, \Delta\Delta MFCC\}$ and $j = 1, \ldots, 13$ (13th element is either the Energy term or the Zero coefficient). The dimension of this vector is 273 (= meta-features$\cdot$ F_j = 7$\cdot$39). The classification label size, which depends on the architecture and database used, varies between 1, 6, and 7.

5 Results

In this Section we present the results of our observations. First, we provide information on the experimental setup. Moreover, the results are discussed structured by approach.

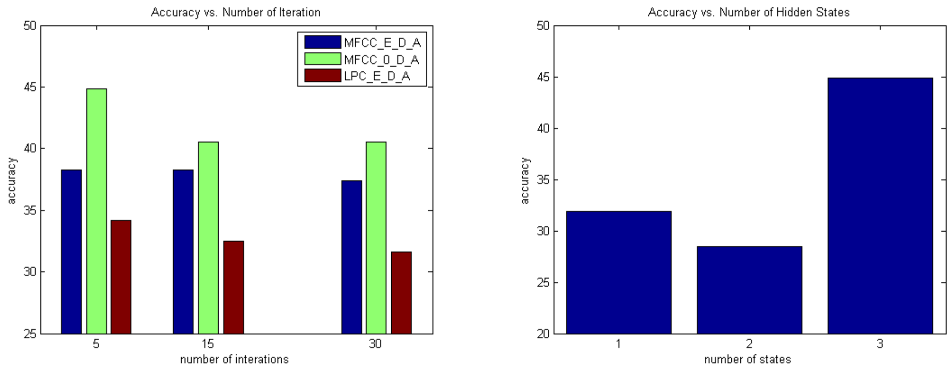
5.1 Experimental Setup

At first, we shortly introduce the databases which we are used for our experiments. Emo-DB [2] (Berlin Database of Emotional Speech by TU Berlin) is a German acted-emotion speech database with 7 emotional classes (anger (ang), boredom (bor), disgust (dis), fear, joy, neutral (neu), and sadness (sad)). The content of the utterances is non-related to the emotional state of the speaker. Each file lasts approximately 2 seconds and these are 493 files. So, we decided to use another database in addition to increase the number of examples. This is the eNTERFACE [6] database which is hosted by Université catholique de Louvain, Belgium. It contains the following emotions: anger, disgust, fear, joy, sadness, and surprise (sur). The 1170 files in this corpus also last 1 to 2 seconds each. The recordings were done during a summer school in 2005, i.e. this is naive material which has an relation between content of the uttered sentence and the emotional state of the user. Both databases contain just files with one emotional state per file which fits our purpose. The available data material is split into a training and test sets (10 % of material is arbitrarily picked as test files).

In the experiments we investigated the Hidden Markov Models and Neural Networks.

The setup of the HMMs is as follows: In the first trial we use an HMM with 3 emitting states (in total 5 states are necessary according to the specifications of HTK [7]) which is widespread

[1] The 0th coefficient is not possible with LPC, so this feature set could not be investigated.



(a) iterations

(b) states

Figure 3 - Recognition accuracy dependent on number of iteration 3(a) and number of states 3(b)

used in speech recognition. Moreover, we varied the number of emitting states and also the number of training iterations. As input the aforementioned feature sets are utilised (see table 1).

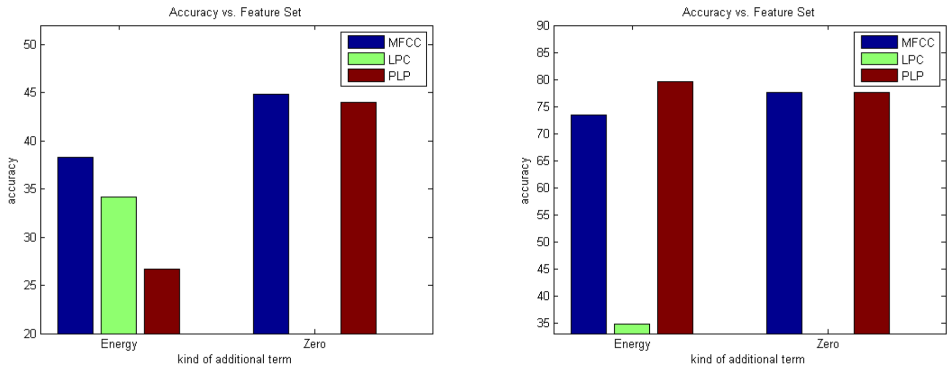
For ANNs we used Torch3 [4] - a C/C++ library - to implement the structure of the networks and also for training and testing. The observed architectures and features are discussed in section 3. More specific, the number of input neurons is 273 and 6 (Emo-DB) or 7 (eNTERFACE) neurons are used as output. While training and testing the optimisation criterion was the Mean Square Error (MSE) and in classification we used the maximum (architecture II) or an interval of ±0.5 (architecture I) to decide which emotional state was recognised.

5.2 Hidden Markov Models

At first, we will discuss the influence of the iteration number in the training process on the recognition accuracy. As it is shown in figure 3(a) the accuracy decreases if the number of training steps is increased. From our point of view, this is due to the over fitting of the models to the training material, i.e. the more the system is trained, the less it can generalises. Almost all experiments showed this behaviour excepting the 2 hidden states LPC_E_D_A which has an enhancement with the number of iterations. Furthermore, the number of emitting states (one, two, or threes) has also an influence regarding the classification accuracy. This is visualised in figure 3(b). Analysing the data which was reached with eNTERFACE we propose an HMM structure of three emitting states. This is also the structure which is commonly used in spoken speech recognition. All further experiments presented in this paper are realised with three state models.

Moreover, we also observed the influence of the different feature sets which are suitable for speech recognition. Figure 4 visualises the accuracies relative to the used feature sets which are previously introduced. The results are grouped with respect to the additional parameter either Energy term or Zero coefficient. As one can see, the Zero coefficient extended feature sets outperform the Energy ones, especially in eNTERFACE case and PLP we achieve a large gain (17.25% absolute). The LPC feature set which is only for Energy feasible has less recognition power: 1) it is outperformed by MFCC and PLP and 2) regarding the confusion matrices no emotional state is clearly recognised. From this, we suggest to skip LPC features.

In the context of naive vs. acted speech and emotional states the differences between the several, remaining features are slightly similar in their behaviour. In both case, one can observe an improvement if Zero coefficient is used. The results of the experiments are shown in figure 4. Due to the better recording conditions and the more expressive utterances Emo-DB with accuracy of 77.55% (PLP_0_D_A) or 79.59% (PLP_E_D_A) outperforms eNTERFACE with 43.97%



(a) feature sets, eNTERFACE

(b) feature sets, Emo-DB

Figure 4 - Recognition accuracy dependent on feature sets. 4(a) visualises the differences for eNTERFACE and in 4(b) the database Emo-DB is used.

Table 2 - Confusion matrices with numbers of recognised emotional states each

(a) PLP_0_D_A, eNTERFACE

	ang	dis	fear	joy	sad	sur
ang	13	0	0	2	3	2
dis	2	3	1	5	4	3
fear	3	1	4	1	8	2
joy	0	3	1	10	1	5
sad	1	0	1	1	15	1
sur	3	2	3	2	4	5

(b) PLP_E_D_A, Emo-DB

	ang	bor	dis	fear	joy	neu	sad
ang	10	0	0	2	1	0	0
bor	0	8	0	0	0	0	0
dis	0	0	3	0	0	0	1
fear	1	1	0	2	1	0	0
joy	1	0	0	1	4	0	0
neu	0	0	0	0	0	8	0
sad	0	1	0	0	0	0	4

(PLP_0_D_A). But, with respect to the noise in the material and that no additional prior knowledge was used, the result is satisfying. Moreover, according to [1] naive speech with emotional states is to prefer to get realistic results.

Finally, we discuss the influences of the feature sets on special emotional states. For this, we calculated the confusion matrices and got the following results: In the naive case sadness and anger are recognised with 84.2% and 75% in worst case using Zero coefficient (see table 2(a)). Generally, the recognition of fear was confused. Joy performs better than fear and can be detected with PLP_0_D_A in an acceptable way. Table 2(b) shows one confusion matrix built for Emo-DB database. As one can see, boredom and neutral are detected absolutely correct and sadness was confounded once. For anger, disgust, and joy the accuracy is greater than 66.7%. Again, fear is outperformed by all other emotional states. Regarding acted speech like Emo-DB we can see that the recognition performance is much better and the detection of an emotional state is really good.

Generally, we ascertain that in several cases the Zero coefficient instead of Energy term should be used to improve the accuracy. Moreover, PLP and MFCC feature sets are investigated and we suggest to apply PLP in most cases because of the outperfoming recognition rate in the different emotional states.

5.3 Artificial Neural Networks

Examining the classification results of the Artificial Neural Networks with architecture II (see figure 2(b)), we see that the numerical values are roughly similar to these of the HMMs. Again,

[2] One data set is a total outlier which is recognised as each emotional state. This is not included in the results.

Table 3 - Confusion matrices with numbers of recognised emotional states using ANNs with 1 hidden layer architecture and applying equation 4

(a) 1 hidden layer, eNTERFACE

	ang	dis	fear	joy	sad	sur
ang	13	0	0	2	3	2
dis[2]	2	3	1	5	4	3
fear	2	1	4	1	8	2
joy	0	3	1	10	1	5
sad	1	0	1	1	15	1
sur	3	2	3	2	4	5

(b) 1 hidden layer, Emo-DB

	ang	bor	dis	fear	joy	neu	sad
ang	12	0	0	0	1	0	0
bor	0	8	0	0	0	0	0
dis	0	1	2	0	0	0	1
fear	1	0	0	4	0	0	0
joy	1	0	1	0	3	1	0
neu	0	2	0	0	0	6	0
sad	0	0	0	0	0	2	3

the acted material performs better than the naive one. On the other hand, some emotional states like joy are significantly better classified (cf. table 3(a)), especially in naive data sets. This is an advantage of the architecture and features as well as of the approach. One has to be aware that we could achieve slightly better results because we used a maximum criterion to get the classification label and in some cases the differences between exact and trained label are just marginal. This observation leads the focus to the fusion aspect. Combining the different results which are the output of the ANNs may improve the accuracy. For this, we have to identify and specify feasible methods and combination functions. This is the aim of our future research.

Furthermore, we investigated the tuned hyperbolic tangent (see equation 3). Using this activation function, we achieved better results especially in disgust and surprise. In our experimental setup, the training processes influences the weight values in such a way that the emotional state joy could not be learned. We assume the weights have to be to large for the other states that the activation function is in the saturation range for joy. In future work we will examine this aspect in more detail. Nevertheless, our results show that we get an improvement in recognition rate for some emotional states if we use different activation functions in the classification process.

At last, we consider architecture I (see figure 2(a)) of ANNs. As aforementioned this network provides a classification unit which selects an element from $\{1,\ldots,6/7\}$ as output. Regarding the results of acted material, only anger and joy is recognised with $\geq 50\%$ correctness by applying an interval of ± 0.5 on the real class label. In case of naive speech it is even worse. Only sadness is detected with almost 50%.

From our point, possible reasons for these results are: 1) The amount of data is to low to train the necessary transitions of the network, especially in acted case. 2) Having strict classification labels could cause a confusion if the emotional states are quite similar (we discussed this effect in architecture I, previously). In the experiments we found that this happened really often generating crisp decisions. Therefore, we suggest to use networks which are more flexible and we prefer architecture II with an 1-of-N classification.

6 Conclusion and Outlook

In this paper, we investigated the influences of several feature sets and approaches, i.e. testing HMMs and ANNs in comparison and inside the methods we tested different architectures. All experiments are applied on the eNTERFACE and Emo-DB database, which provide naive and acted material, respectively. We discussed the results in detail and found the facts as follows: Regarding the HMM approach a three state model is to prefer. Concerning the feature sets in this case PLP and MFCC have the best performance and in most cases we suggest to use Zero coefficient. The ANN results show that architecture II (see figure 2(b)) is the most suitable one.

Selecting a satisfying activation function needs the experiences of the developer. But, selecting suitable functions (we proposed two) may yield in an improvement in recognition accuracy, especially for some emotional states.

Our future work is directed to combine the obtained experiences to achieve a gain in recognition. In fact, all results we get are afflicted with uncertainty. For this, we want to apply several fusion methods, on topmost the Dempster-Shafer Theory, but we are also interested to incorporate other approaches which are based on sensor fusion like Interactive Multiple Models. This will give us the possibility to handle uncertainty without losing any information which is often the case in binary decisions. Moreover, we will create an approach which fuses the information on different levels. The bottom level is dealing with the features and signals directly, whereas the medium level is handling the recognition results and the material which is recognised. The top level is regarding the meaning of the material and incorporates several other information like dialogue history and situational state of the user. The higher the level, the more abstract information is used and fused. Achieving these goals we will focus on modelling the user and his/her emotional/situational state as well as extend common fusion techniques to incorporate these information to improve recognition and dialogue management.

Acknowledgements

We acknowledge support by the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

Literature

[1] BATLINER, A. : Whence and Whither: The Automatic Recognition of Emotion in Speech. In: *Proceedings of the 4th IEEE Tutorial and Research Workshop “Perception and Interactive Technologies for Speech Based Systems”, Kloster Irsee* (2008)

[2] BURKHARDT, F. ; PAESCHKE, A. ; ROLFES, M. ; SENDLMEIER, W. ; WEISS, B. : A Database of German Emotional Speech. In: *Proceedings of the 5th Interspeech* (2005)

[3] CEN, L. ; SER, W. ; YU, Z. L.: Speech Emotion Recognition Using Canonical Correlation Analysis and Probabilistic Neural Network. In: *Proceedings of the 7th International Conference on Machine Learning and Applications, San Diego* (2008)

[4] COLLOBERT, R. ; BENGIO, S. ; MARIETHOZ, J. : Torch: a modular machine learning software library. In: *Technical Report IDIAP-RR 02-46, IDIAP* (2002)

[5] HANNEMANN, M. : Hierarchical Semantic Tagger for Robust Spoken Language Understanding. In: *Masters Thesis, Otto-von-Guericke-University Magdeburg* (2009)

[6] MARTIN, O. ; KOTSIA, I. ; MACQ, B. ; PITAS, I. : The eNTERFACE’05 Audio-Visual Emotion Database. In: *Proceedings of the 22nd International Conference on Data Engineering Workshop* (2006)

[7] YOUNG, S. ; EVERMANN, G. ; GALES, M. ; HAIN, T. ; KERSHAW, D. ; LIU, X. ; MOORE, G. ; ODELL, J. ; OLLASON, D. ; POVEY, D. ; VALTCHEV, V. ; WOODLAND, P. : The HTK Book. In: *Cambridge University Engineering Department* (2006)