Application of HMMs for the Recognition of Emotional Sequences in the Valence-Arousal Space

David Hübner, Ronald Böck and Andreas Wendemuth

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

Abstract: This paper will show how models can be generated, which are capable of recognizing sequences of emotional states from speech. For this purpose Hidden Markov Models (HMMs) are introduced, which are trained on spontaneous, non-acted emotions. Unlike other publications in this area, whose main focus is often on the classification in one of the basic classes introduced by Ekman or Plutchik, we will generate 2-dimensional representations of the user's emotion in the Valence-Arousal space. Hence, not only the basic emotions are recognized, but also an additional parameter, the word frequency, is extracted from the speech signal. We trained two gender specific models and one combined model and tested these afterwards on unknown data. The evaluation of the robustness is done by using two cross-validation methods.

1 Introduction

In today's society the interaction between men and machines using language is becoming more and more self-evident. Machines still lack of many human abilities which would considerably simplify communication. Beside the pure speech recognition and the derived semantics, the recognition of emotions from speech is an important criterion, as emotions influence our perception and decision making process. Using information about the emotional state of a user, machines will be able to react in a more appropriate and individual way to the wishes of a user, and they will also provide possibilities to adapt their dialog strategies, depending on the user's emotions. The papers' focus is on recognizing sequences of emotional speech. Hidden Markov Models (HMMs) are already established in the field of speech recognition, and will here be used to recognize sequences of emotions. Since we later want to be able to recognize emotions in realistic scenarios, we chose the SmartKom database [1] which provides non-acted emotional dialogs generated in Wizard-of-Oz experiments. Hence the data can be considered as natural. The original labeling consists of 7 emotion classes: anger/irritation, helplessness, pondering, neutral, surprise, joy/gratitude and unidentifiable emotions. Further all classes but neutral are optionally weighted as *weak* and *strong*. Taking the amount of training material into account we decided to reduce the classes in order to gain better classification results. According to [2] we merged the classes into a 4-class scenario. The final classes are neutral, joy, helplessness and anger whereas Table 1 shows which of the above mentioned classes were combined.

The paper is organized as follows: Initially we describe the structure and training of the HMMs. Then the Valence-Arousal space is introduced and we describe how the results are mapped into this space. Further we show how the word frequency is determined and discuss the problems of the current approach. In Section 5 we show and discuss the results. The last section gives an outlook on our ongoing research and how we want to refine our models.

Table 1 - Reduction of emotion classes

Original classes	**New class**
strong/weak joy/gratitude + strong/weak surprise	joy
strong/weak pondering + strong/weak helplessness	helplessness
strong/weak anger/irritation	anger
neutral + unidentifiable	neutral

2 Preparation of the Data

We selected 1887 files from the data basis which contain different sequences of the 4 emotions. The sequence lengths vary between 1 and 3 emotions/file, which means a maximum of 2 transitions between emotions in an utterance. In order to create gender dependent models the data set was again split manually into female and male speakers. The amount of data from female speakers outnumbers the male data by a factor of ca. 1.5 (compare Table 2).

Table 2 - Distribution of the data with respect to the emotional transitions and the gender

Transitions	**Female**	**Male**	Σ
0	729	511	1240
1	255	172	427
2	140	80	220
Σ	1124	763	1887

The 4 emotions are not uniformly distributed over the data. For instance, looking at the static emotions, which build a group consisting of 1240 files nearly 84% represents a neutral emotion, while the remaining emotions occur with 5-6% less frequently.

Since our aim is to identify sequences of emotions, we analyzed the data with respect to files which contain up to 2 emotional changes. For such sequences consisting of 2 or 3 emotions the transition matrix in Table 3 summarizes the frequency of occurrence of each pairwise transition. Transitions from helplessness to neutral (and vice versa) can be observed most frequently while quite a few transitions occur between anger and joy. This distribution reflects the situation in everyday dialogs quite well since direct changes between complementary emotions, like anger and joy, are also to be expected rather infrequently and transitions mostly include/pass a neutral state. The 4 most frequent sequences with 2 emotional changes are given as follows: 1) neutral-helplessness-neutral (n=93), 2) neutral-joy-neutral (n=28), 3) helplessness-neutral-helplessness (N=25) and neutral-anger-neutral (n=12).

Table 3 - Number of pairwise transitions between the emotions

	neutral	**anger**	**joy**	**helplessness**
neutral	0	53	108	207
anger	46	0	9	11
joy	62	2	0	15
helplessness	330	8	16	0

3 The HMM-Models

The whole model was build up with the Hidden Markov Toolkit (HTK) [5]. From the original wav-files we extracted the Mel Frequency Cepstral Coefficients (MFCCs). This results in 39 features per sample: the zero coefficient and 12 MFCCs as well as the corresponding delta and acceleration values. As also common in speech recognition we used HMMs consisting of 3 emitting states. To be conform to HTK standard the HMM definition provides an additional state at the beginning and at the end, so that we end up with 5 states in total. Each of the 4 emotions is represented by its own HMM. We trained 3 models, one for male speakers, one for female speakers and a combined one.

3.1 Training the Models

The single HMMs are initializend using the *HInit* method of the HTK and data which represents the corresponding emotion, for instance the Anger-HMM is initialized using data that contains the static emotion anger. Afterwards 5 iterations of training the model applying the *HERest* tool follow. Here the training data consists of files containing up to 2 emotional changes. The training data is only a subset of the complete data set, as some unseen data is reserved for testing. We tried several numbers of training iterations and 5 showed the best generalization capability. Increasing the iteration number leads to over fitting of the model and the recognition results generated during the tests decrease.

3.2 Removing Silence

Our first recognition rates were quite frustrating, barely better than guessing. A closer look at the provided data showed mainly two problems: 1) The data is quite noisy, for instance traffic-noise including car horns contaminates some files. 2) Approximately 50% of each file consists of silence, as often some seconds between the dialog turns among the user and the machine pass by. This would not be a problem if silence was annotated as such, but here the silent parts contribute to the true emotions on the level of annotation.

To achieve a gain in recognition we applied a self-written tool which analyzes the normalized energy values of each sample and creates a new file by leaving out all samples which are below a certain threshold. In our case we chose a threshold of 0.42, which seems quite high as the normalized values are in the range of [0;1] but this is necessary as the basic noise is quite high as well. Due to the different volumes of the files, it sometimes happens that both the noise and the speech signal are below the threshold and the resulting file would be empty. In order to keep all training data these files remain unchanged. We are currently working on an adaptive threshold, which will solve this problem. But still this first approach resulted in an enormous gain in performance and the percentage of correctly classified emotions increased by 15-25% per class.

4 The Valence-Arousal (VA) Space

One of our future aims is to provide not only discrete emotion classification, but also to represent continuous models. Being able to determine a point in the VA space results in a much more flexible and complex recognition system, as it is the case when having discrete classes. Starting with the basic emotions, Plutchik [3] suggested a circular representation in the VA-Space (compare Figure 1).

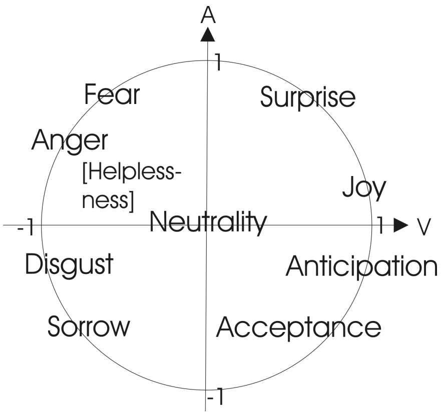

Figure 1 - The basic emotions in Plutchik's "wheel".

4.1 Transferring the Results and Determining the Word Frequency

In a first approach we mapped our discrete emotion classes to fix points on the valence axis. Since helplessness is not a basic emotion we rooted it between anger and neutral. An advantage of the VA space is that some acoustic measures, like the word frequency, are directly correlated with the 2 dimensions. Hence we decided that the arousal dimension shall be represented by the normalized word frequency. Since we found no higher word frequencies than 4.5 words/second the normalized value of 1 represents a frequency of 5 words/second.
To measure how fast a user speaks is a quite difficult task. In a realistic system one would need a speech recognizer in addition, which counts the number of recognized words and the time for each word, as well as the time for the complete utterance, which of course may contain pauses. On basis of these data the average word frequency can be computed. Since training and testing of an speech recognition system is a time consuming task, we chose a different way. The Smartkom material is emotional annotated as well as the transcriptions of "what is said" are provided. By counting the number of words and determining the length of the corresponding file a rough measure for the word frequency can be computed. Applying this straight forward approach on some material shows some drawbacks: 1) Since our silence remover works with an static threshold many files still contain too much silence. On the other hand silent parts which are characteristically for emotion like helplessness may also be removed. 2) The transcriptions do not clearly provide machine readable information about who is speaking. Hence both silence and utterances from the system often make it nearly impossible to get an accurate value for the word frequency of a user. In most cases it is too low. Further approaches will be tested in future research.

5 Results

5.1 Validation

In order to test our models we chose 2 cross-validation methods. The first one split the data randomly in 90% training data and 10% test data and will be called 90-10 Cross-Validation. We repeated this experiment 50 times and computed the confusion matrixes, which show how well the emotions are recognized and how often they are mixed up. Table 4 shows the confusion matrix, which is the combination of all 50 experiments for the gender dependent model. The number of training files in the female model is 1011, while 113 are used for testing. In case of the male model 687 files are used for training and 76 for testing. One can see that the neutral emotion is with more than 80% best recognized for both genders, while the male model performs slightly better. Helplessness is also recognized quite well in case of females, while males perform 5% worse. A huge difference between the genders can be observed in the recognition of joy: here the male model performs much better than the female one. For anger we received the lowest recognition rates with only 50% for both genders.

Table 4 - Confusion matrix for the gender dependent models (up to 3 emotion transitions)

Emotion ♀/♂	**anger**	**neutral**	**joy**	**helplessness**	Σ	**correct [%]**
anger	257/83	88/35	85/20	78/20	508/158	**50.6/52.5**
neutral	216/132	4046/2927	328/243	440/132	5030/3434	**80.4/85.2**
joy	112/45	114/61	247/242	154/45	627/393	**39.4/61.6**
helplessness	117/120	70/35	182/156	1065/711	1434/1022	**74.3/69.6**

For the combined model we used 1699 files for training and 188 for testing. The results are quite similar to the ones for the gender dependent model. For anger the recognition rate even increased, while he other rates are a litlle bit reduced (compare Table 5). This results seem to contradict former results from [4] who analyzed the same database before and showed the the gender dependent models perform better than the mixed one. The difference between their and our approach is in the selected features and methods used for training: we apply only MFCC features while [4] used further features as well as an optimization algorithm to collect only the important features for a Bayes classifier. However, in our models which are purely based on the MFCC features the difference between the gender dependent and the gender independent model is much smaller.

Table 5 - Confusion matrix for the combined model (up to 3 emotion transitions)

Emotion ⚥	**anger**	**neutral**	**joy**	**helplessness**	Σ	**correct [%]**
anger	409	122	111	133	775	**52.8**
neutral	429	6449	622	847	8347	**77.3**
joy	181	138	604	112	1035	**58.4**
helplessness	289	127	253	1690	2359	**71.6**

The second cross-validation method is called Leave-One-Speaker-Out. The difference to the former method is that it excludes one speaker completely from training, who is later used for testing. So the models can be tested for robustness in an optimal way. Since we have 85 female speakers we could train and test 85 versions of the female model. In case of the male model only 61 different versions could be generated. Tables 6 and 7 summarize the results for each of the three models. The results differ only marginally from the former cross validation method.

Table 6 - Confusion matrix for the gender dependent models (up to 3 emotion transitions)

Emotion ♀/♂	**anger**	**neutral**	**joy**	**helplessness**	Σ	**correct [%]**
anger	54/18	23/8	15/6	19/3	111/35	**48.6/51.4**
neutral	36/20	815/594	70/52	86/26	1007/692	**80.9/85.8**
joy	19/8	25/12	47/56	28/9	119/85	**39.5/65.9**
helplessness	28/27	11/8	36/33	208/130	283/198	**73.5/65.6**

Table 7 - Confusion matrix for the combined model (up to 3 emotion transitions)

Emotion ⚥	**anger**	**neutral**	**joy**	**helplessness**	Σ	**correct [%]**
anger	77	21	21	28	147	**52.4**
neutral	85	1295	139	164	1683	**76.9**
joy	32	27	120	25	204	**58.8**
helplessness	61	26	59	341	487	**70.0**

5.2 Word Frequency

We combined the results from the emotion recognition with those from the word frequency analysis in order to get a 2-dimensional representation in the Valence-Arousal space. It must be mentioned that the in Section 4 described problems only allow us to show results on some manually selected files. Figure 2 visualizes typical word frequencies for the 4 emotions based on 8 utterances. Since we are only able to measure the word frequency over the complete utterance, we consider only single emotions. One sees that joy has the highest word frequency while helplessness has the lowest. The minimum, the maximum and the average word frequency are shown for each emotion. We observed word frequencies between 1.5 and 4.5 words per second and normalized them to one, which represents a frequency of 5 words/second.

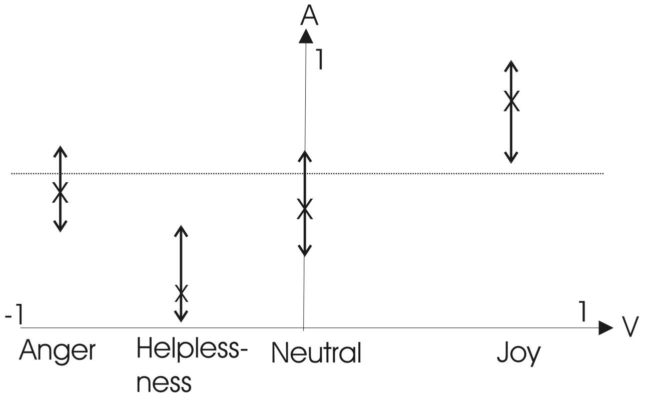

Figure 2 - Results mapped into the VA space

6 Conclusion / Outlook

We showed how HMMs can be applied in order to train models which are able to recognize sequences of emotions. Using the SmartKom data we found no strong evidence for the use of gender dependent models. In opposite to results in speech recognition, the gender of a person seems to be of lower importance in emotion recognition. The word frequency can be used to produce results in the Valence-Arousal space. Since different emotions provide different word frequencies this information can be used to support or discard the hypothesis made by the emotion recognizer.

In future research we want to refine our models mainly in two directions. Currently we are working on a history model which benefits from the the fact that changes from one emotion to another do not occur with the same probabilities. For instance a change from neutral to helplessness is much more probable than a change from anger to joy. On the basis of a set of example sequence we are able to create a history model which reflects these probabilities and can afterwards be used by the emotion recognizer to support the acoustic evidence by preffering changes to more probable emotions. This approach is already widely spread in the area of speech recognition, where such models are denoted as language models. Further we want to use the log-likelihoods provided by the emotion recognizer to ensure a continuous classification in the Valence-Arousal space. So the single emotions will no longer have a fix value on the valence axis, but are allowed to apodt values between the standard emotions, too.

Acknowlegdements

We acknowledge support within the Neuroscience Excellence Project ”Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems” (NIMITEK) funded by the Federal State of Sachsen-Anhalt (2005-2010), FKZ XN3621H/1005, cosponsored by the European Union. This work was further supported by the Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

References

[1] SMARTKOM: Dialog-based Human-Technology Interaction Multi-Modal Database, University of Munich, 2004

[2] BATLINER, A. ; ZEISSLER, V. ; FRANK, C. ; ADELHARDT, J. ; SHI, R. ; NOETH, E. : We are not amused - but how do you know? User states in a multi-modal dialogue system. In: *Proceedings of EUROSPEECH*, 2003, S. 733 – 736

[3] PLUTCHIK, R. : The Nature of Emotions. In: *American Scientist*, 2001

[4] VOGT, T. ; ANDRE, E. : Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In: *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006

[5] YOUNG, S. ; EVERMANN, G. ; GALES, M. ; HAIN, T. ; KERSHAW, D. ; LIU, X. ; MOORE, G. ; ODELL, J. ; OLLASON, D. ; POVEY, D. ; VALTCHEV, V. ; WOODLAND, P. : In: *The HTK Book (for HTK Version 3.4)*, 2006