Conditioned Hidden Markov Model Fusion for Multimodal Classification

Michael Glodek, Stefan Scherer, Friedhelm Schwenker

Institute for Neural Information Processing, University of Ulm, 89075, Germany

firstname.lastname@uni-ulm.de

10.21437/Interspeech.2011-603

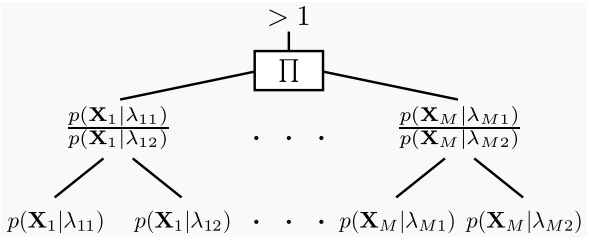
Abstract

Classification using hidden Markov models (HMM) is in general done by comparing the model likelihoods and choosing the class more likely to have generated the data. This work investigates a conditioned HMM which additionally provides a probability for a class label and compares different fusion strategies. The notion is two-fold: on the one hand applications in affective computing might pass their uncertainty of the classification to the next processing unit, on the other hand different streams might be fused to increase the performance. The data set studied incorporates two modalities and is based on a naturalistic multiparty dialogue. The goal is to discriminate between laughter and utterances. It turned out that the conditioned HMM outperforms classical HMM using different late fusion approaches while additionally providing a certainty about class decision.

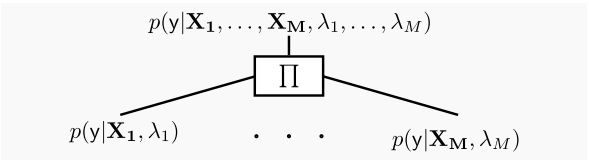
1. Introduction

Nowadays the classical hidden Markov model (HMM) is still widely used to classify sequential data [1]. However, a HMM renders only a probability about how likely the model has generated the data but does not provide a probability distribution over classes. Morency et al. proposed the latent-dynamic conditional random field (LDCRF) based on undirected graphs which inherits the structure from classical HMM but has additional random variables representing a class label connected to all hidden random variables [4]. Within this work its Bayesian pendant is presented and studied. We set the focus on lateral classifier fusion as illustrated in Figure 1. Figure 1a shows the classical approach to discriminate two classes using the probability $p(\mathbf{X}_i|\lambda_{iy})$ of the HMM where λ_{iy} represents the model for feature i and class y and $\mathbf{X}_i$ is the sequence to be classified. Classifier fusion is realized by multiplying the ratios of each feature decision $\frac{p(\mathbf{X}_i|\lambda_{i1})}{p(\mathbf{X}_i|\lambda_{i2})}$ (in general done in log-space). It is not possible to use different complex models for one feature, nor to intuitively weight the classifiers. Furthermore, an uncertainty of the decision summing up to one is not given. Lateral fusion using conditioned HMM as illustrated in Figure 1b is able to overcome these drawbacks by providing a probability $p(\mathrm{y}|\mathbf{X}_i, \lambda_i)$ where y denotes the random variable over classes. Within this work different fusion approaches, namely decision averaging and naive Bayes, have been compared using the application of laughter detection.

Section 2 gives a brief introduction to conditioned HMM. The data set used to compare classical HMM with conditioned HMM (CHMM) is described in Section 3. The experiments conducted are presented in Section 4 and summarized in Section 5 which also gives an overview of future work.



(a)



(b)

Figure 1: **Different approaches to fuse sequential classifiers using M different streams.** (a) Fusion architecture using classical HMM. For each feature the ratio of two HMM, each one adapted for a class, is computed. The ratio of two models is then fused by multiplication. (b) Fusion architecture using conditional HMM. For each feature the distribution over the classes can be fused using decision averaging or, as illustrated, by using naive Bayes which is also denoted by the product.

2. Model Description

Let $\{\mathbf{X}^{(n)}\}_{n=1}^{N}$ be a set of sequential data with corresponding sequential labels $\{\mathbf{y}^{(n)}\}_{n=1}^{N}$. The observations and the label sequences are given by $\mathbf{X}^{(n)} = (\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_T^{(n)})$ and $\mathbf{y}^{(n)} = (y_1^{(n)}, \ldots, y_T^{(n)})$ where T denotes the number of time steps. Figure 2 shows the model which is a simple extension of the classical HMM. At each time step the latent variable is influenced by the corresponding class label. In order to keep the number of parameters small, the probability of a label to produce a hidden state is assumed to be identical for all time steps. The model likelihood of the sequence $\mathbf{X}$ and the class labels $\mathbf{y}$

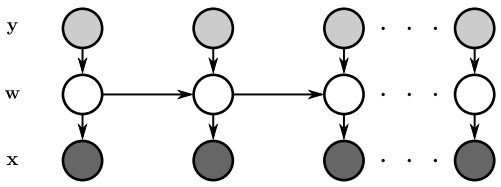


Figure 2: **Graphical model of the CHMM.** The gray nodes are required for learning the parameters of the model but only the dark grey nodes will be used testing a new sequence.

Copyright © 2011 ISCA

is given by

$$p(\mathbf{X}, \mathbf{w}|\mathbf{y}\lambda) = \sum_{\mathbf{w}\in\mathbf{w}} p(\mathsf{w}_1 = w_1|\mathsf{y}_1 = y_1, \boldsymbol{\pi}) \cdot \Big(\prod_{t=2}^{T} p(\mathsf{w}_t = w_t|\mathsf{w}_{t-1} = w_{t-1}, \mathsf{y}_t = y_t, \mathbf{A})\Big) \cdot \Big(\prod_{t=1}^{T} p(\mathbf{x}_t = \mathbf{x}_t|\mathsf{w}_t = w_t, \theta)\Big)$$

where $\mathbf{w} = (\mathsf{w}_1, \ldots, \mathsf{w}_T)$ are the hidden random variables for each time step t and $\lambda = \{\boldsymbol{\pi}, \mathbf{A}, \theta\}$ is the set of parameters. The probability of the initial state is given by the matrix $\boldsymbol{\pi} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{y}|}$ while the label depended transition probabilities are given by the matrix $\mathbf{A} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{w}|\times|\mathsf{y}|}$. Within this study the emission probability $p(\mathbf{x}_t = \mathbf{x}_t|\mathsf{w}_t = j, \theta)$ are given by a Gaussian mixture model (GMM) such that $\theta_j = \{\{\boldsymbol{\phi}_{j,k}\}_k^K, \{\boldsymbol{\mu}_{j,k}\}_k^K, \{\boldsymbol{\Sigma}_{j,k}\}_k^K\}$. To maximize the complete data log-likelihood

$$Q(\lambda, \lambda^{\text{old}}) = \sum_{\mathbf{w}\in\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \lambda^{\text{old}}) \ln p(\mathbf{X}, \mathbf{w}|\mathbf{y}, \lambda)$$

$$= \sum_{j\in\mathsf{w}_1} \gamma_{1,\mathbf{y}}(j) \ln \pi_{j,y_1} \quad (1)$$

$$+ \sum_{t=2}^{T} \sum_{i\in\mathsf{w}_{t-1}} \sum_{j\in\mathsf{w}_t} \xi_{t-1,t,\mathbf{y}}(i,j) \ln A_{i,j,y_t} \quad (2)$$

$$+ \sum_{t=1}^{T} \sum_{j\in\mathsf{w}_t} \gamma_{t,\mathbf{y}}(j) \ln p(\mathbf{x}_t|\mathsf{w}_t = j, \theta_j) \quad (3)$$

the state belief $\gamma_{t,\mathbf{y}}(j)$ and transition beliefs $\xi_{t-1,t,\mathbf{y}}(i,j)$ need to estimated. They are obtained by recursively evaluating the forward and backward variable

$$\alpha_{t,\mathbf{y}}(j) = p(\mathbf{x}_t|\mathsf{w}_t = j) \sum_{i\in\mathsf{w}_{t-1}} \alpha_{t-1,\mathbf{y}}(i) p(\mathsf{w}_t = j|\mathsf{w}_{t-1} = i, y_t) \quad (4)$$

$$\beta_{t-1,\mathbf{y}}(j) = \sum_{i\in\mathsf{w}_t} \beta_{t,\mathbf{y}}(i) p(\mathbf{x}_t|\mathsf{w}_t = i) p(\mathsf{w}_t = i|\mathsf{w}_{t-1} = j, y_t). \quad (5)$$

The probability for a observation sequence given a label sequence is then given by

$$p(\mathbf{X}|\mathbf{y}) = \sum_{i\in\mathsf{w}_T} \alpha_{T,\mathbf{y}}(i). \quad (6)$$

The joint probability is determined by $p(\mathbf{X}, \mathbf{y}) = p(\mathbf{X}|\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is the prior over all label sequences. The conditioned distribution on the class labels is the given by

$$p(\mathbf{y}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{y})}{\sum_{\mathbf{y}\in\mathbf{y}} p(\mathbf{X}, \mathbf{y})} \quad (7)$$

As inference over all possible class label sequences $\mathbf{y}$ is expensive, we assume furthermore that the labels are identical for all time steps. During maximization, unbalanced data sets are problematic. A class holding the majority will dominate the parameters of the CHMM. In order to solve the inequality of the labels, the priors of the data set need to be estimated and removed during maximization. Since sequences are given, two different priors can be determined: the initial state data set prior $\nu_y = 1/N \sum_n \delta(y_1^{(n)}, y)$ and the overall data set prior $\omega_y = 1/N \sum_n 1/T^{(n)} \sum_t \delta(y_t^{(n)}, y)$ where $\delta(\cdot,\cdot)$ denotes the Kronecker delta and n the sample index. The initial state prior normalizes the initial state probability $p(\mathsf{w}_1|\mathsf{y}_1)$ by being multiplied to the state belief $\gamma_{t,\mathbf{y}}(j)$. The overall data set prior normalizes the transition and the emission probability by multiplying the quantity to the transition belief $\xi_{t-1,t,\mathbf{y}}(i,j)$, or the state belief $\gamma_{t,\mathbf{y}}(j)$ respectively.

To keep the number of parameter small, an additional assumption has been made. We assumed that the next hidden state is independently influenced by the predecessor and the class label such that we obtain the matrix $\mathbf{A} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{w}|}$ and the compatibility matrix $\mathbf{C} \in \mathbb{R}^{|\mathsf{w}|\times|\mathsf{y}|}$. Initialization can be done by dividing the number of hidden states into as many partitions as labels. Then for each label a classical HMM is trained for a few iterations using the corresponding subset of hidden states. The obtained parameters $\tilde{\mathbf{a}}_{\mathsf{y}=y}$, $\tilde{\mathbf{A}}_{\mathsf{y}=y}$ and $\tilde{\theta}_{\mathsf{y}=y}$ can then be transferred directly into the new model. Setting the undetermined parameters to zero as shown in Figure 3 will reduce the conditioned HMM to almost the classical HMM setting, with the exception that classification is still possible by using the probability $p(\mathbf{y}|\mathbf{X})$ and not the model likelihood. Setting a subset or all undetermined parameters to random values will increase the model complexity but might also increase performance.

Figure 3: **Initialization of the state transition matrix A.** The transition matrices $\tilde{\mathbf{A}}_{\mathsf{y}=y}$ are trained by classical HMM and used to initialize the CHMM. Setting the unknown parameters to zero will result in a behaviour close to classical HMM.

$\tilde{\mathbf{A}}_{\mathsf{y}=1}$	$\mathbf{0}$
$\mathbf{0}$	$\tilde{\mathbf{A}}_{\mathsf{y}=2}$

3. Laughter Detection in Naturalistic Multiparty Conversations

The Freetalk data set comprises two separate 90 minute long recordings of four people sitting around a table involved in a lively unconstrained natural conversation. All participants are recorded by an unobtrusive 360° spherical camera, capturing coarse head, body and hand movements, and a microphone placed on the middle of the table. From the audio channel the features perceptual linear prediction (PLP) and the modulation spectrum (MOD) have been extracted due to their capability to appropriately represent sequential audio characteristics [2, 3]. The faces have been tracked in the video stream and for each face a coarse feature containing the relative head movement was calculated (MOV). Both recordings are annotated for *utterances* and *laughter* using both available modalities, resulting in a high annotation accuracy. Based on the first recording a pre-segmented data set with 256 laughter and 2270 utterance samples has been created for this study. The Freetalk data set is available at `http://www.speech-data.jp/` and a more detailed description of the data set is given in [7, 6].

4. Experiment and Results

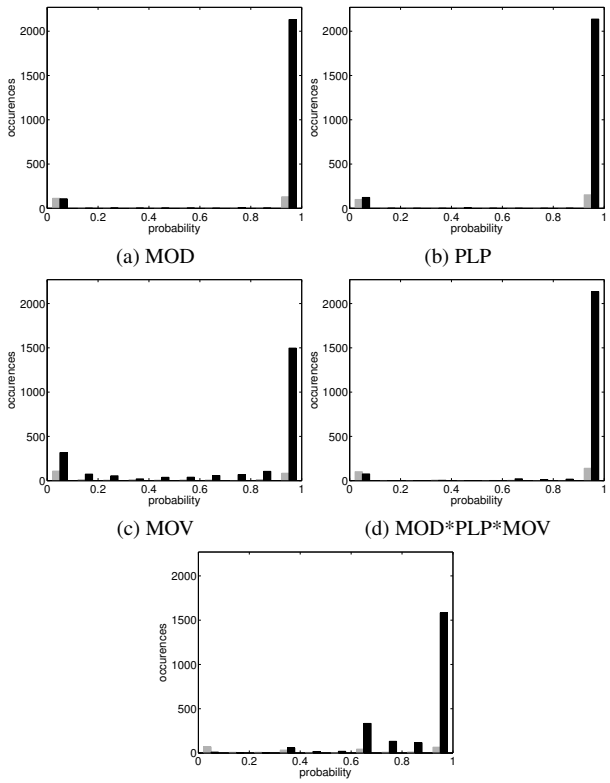
Experiments have been conducted in order to give insight in the model behaviour. All models investigated are restricted to observation emitted by Gaussian mixture models (GMM) having full co-variance matrices. For all models a prior according to the number of occurrences in the pre-segmented data set has been used.

The first experiment reflects the performance of classical HMM classifier fusion as illustrated in Figure 1a. The classifier decision of each feature is based on the ratio of the HMM for each class. The ratios of all features are multiplied to give a fused overall decision (calculations were performed in log-space). The rates and standard deviation (given in brackets) using 10-fold cross-validation using an eight state HMM with full transition matrix are shown in Table 1. The classifier per-

formance based on a single feature is given in the first three rows. The fusion based on different combinations of features can be found in the consecutive rows. The best result is achieved by fusing all streams and modalities and reaches an error rate of 7.3%. The F_1-measure of 0.96 for the utterances is very high, while the F_1-measure of laughter reaches only 0.62. Hence, the utterances is detected very reliably but the laughter is often confused with utterance.

An equivalent model to the first experiment is a 16-state conditioned HMM in which the transition matrix is set to zero except of the elements initialized by the classical HMM. Due to the usage of the EM-algorithm it is guaranteed that the elements that are set to zero will remain zero during training. Therefore, both models will evolve independently during training which gives an analogy to the previous approach. The distribution over the random variable y allows different fusion strategies to be applied. We investigated naive Bayes and averaging is denoted by * and + respectively. The rates and standard deviation using 10-fold cross-validation are shown in Table 2. While classifiers MOD and PLP achieve slightly worse results, the classifier based on the MOV features shows an improvement of around 4%. The fusion strategy based on naive Bayes performs close to the fusion of the classical HMM (nota bene that the multiplied quantities of both approaches are fundamentally different). The fusion based on decision averaging, which corresponds to fuzzy voting, performs as well close to the classical HMM approach. However, fusing MOD+PLP outperforms all former results and reaches an error rate of 5.8% with an F_1-measure of 0.97 for utterances and 0.68 for laughter. All these three quantities are statistically significantly better than the best fusion result of the HMM reference experiment (MOD*PLP*MOV) as shown using a paired t-test.

Figure 4a, 4b and 4c illustrate the distribution of the classifier outputs per class. The black bars correspond to the class utterances with a high occurrence of samples, while the grey bars indicate the decisions for the class laughter. The decisions of the two audio features are very similar and almost all decisions have a high degree of certainty. The less precise video modality, however, shows a more spread distribution explaining its bad performance but also shows the advantage of the model. Figure 4d and 4e illustrate the fused distributions using naive Bayes and averaging. While naive Bayes enforces the distinctness of the classes, the averaging regards the diversity of the classifiers and, therefore, is more affected by false decisions due to the MOV features.



(a) MOD (b) PLP

(c) MOV (d) MOD*PLP*MOV

(e) MOD+PLP+MOV

Figure 4: **Distribution of the class probability of conditioned HMM.** The decisions for the classes labeled to be utterances are shown by the black bars, while the decisions for the class laughter is shown by grey bars. Model settings are: 16 hidden states, transition matrix with zeros on the not initialized elements, 6 mixture components.

Table 1: **Classifier error rate in % and F_1 measure of the HMM.** The model settings are: 8 hidden states, full transition matrix and 6 mixture components.

Feature Set	Error Rate	F_1(Laughter)	F_1(Utterance)
MOD	8.44 (1.88)	0.558 (0.11)	0.953 (0.01)
PLP	8.87 (1.99)	0.585 (0.09)	0.950 (0.01)
MOV	29.26 (5.78)	0.234 (0.04)	0.818 (0.04)
MOD*MOV	8.95 (1.82)	0.530 (0.11)	0.950 (0.01)
MOD*PLP	7.45 (1.83)	0.615 (0.10)	0.959 (0.01)
PLP*MOV	8.83 (2.27)	0.583 (0.10)	0.951 (0.01)
MOD*PLP*MOV	7.29 (1.84)	0.621 (0.10)	0.960 (0.01)

Table 2: **Classifier error rate in % and F_1 measure of the conditioned HMM.** Statistically significantly better results compared to the best performing HMM result are marked by $\star$. The model settings are: 16 hidden states, transition matrix with zeros on the not initialized elements and 6 mixture components.

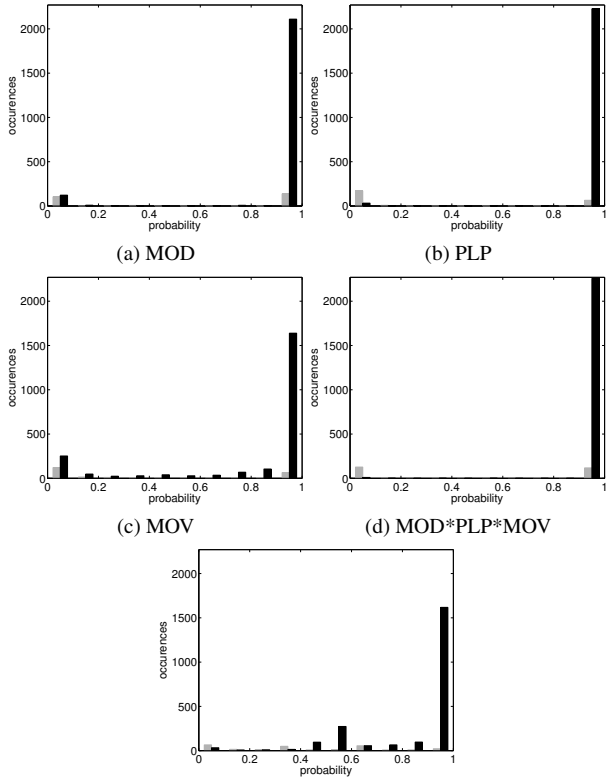
Feature Set	Error Rate	F_1(Laughter)	F_1(Utterance)
MOD	9.62 (1.08)	0.519 (0.07)	0.947 (0.01)
PLP	8.94 (1.46)	0.575 (0.08)	0.950 (0.01)
MOV	25.43 (5.11)	0.265 (0.05)	0.845 (0.04)
MOD*MOV	9.10 (1.20)	0.533 (0.07)	0.950 (0.01)
MOD*PLP	7.44 (1.73)	0.618 (0.10)	0.959 (0.01)
PLP*MOV	8.87 (1.62)	0.573 (0.09)	0.951 (0.01)
MOD*PLP*MOV	7.37 (1.80)	0.619 (0.10)	0.959 (0.01)
MOD+MOV	9.05 (1.24)	0.536 (0.07)	0.950 (0.01)
MOD+PLP	5.77 (1.09)*	0.677 (0.08)*	0.968 (0.01)*
PLP+MOV	8.76 (1.60)	0.577 (0.09)	0.951 (0.01)
MOD+PLP+MOV	8.24 (1.52)	0.565 (0.08)	0.954 (0.01)

The third experiment uses a fully occupied transition matrix with 16 hidden states. The results are shown in Table 3. Again the averaging of MOD and PLP results in very low error rate of 5.7% with an F_1-measure of 0.97 for utterance and 0.63 for laughter. Furthermore, the naive Bayes approach of MOD and PLP as well as MOD, MOV and PLP leads to error rates of around 5.7%. All three fusion strategies have an error rate and F_1 measure for utterances which are statistically significantly better than the best performing HMM fusion result (MOD*PLP*MOV) based on a paired t-test.

The distribution of these classifier decisions are presented in Figure 5. The single classifiers are shown in Figure 5a, 5b and 5e. While MOD and PLP decisions are confident (also in case a confusions occur), the classifier based on the feature MOV is considerably less confident about the class decision.

Table 3: **Classifier error rate in % and F_1 measure of the conditioned HMM.** Statistically significantly better results compared to the best performing HMM result are marked by $\star$. Model settings are: 16 hidden states, transition matrix with random values on the not initialized elements and 6 mixture components.

Feature Set	Error Rate	F_1 (Laughter)	F_1 (Utterance)
MOD	10.10 (1.84)	0.524 (0.073)	0.943 (0.01)
PLP	8.67 (1.21)	0.388 (0.139)	0.953 (0.01)
MOV	21.94 (5.12)	0.249 (0.068)	0.870 (0.04)
MOD*MOV	9.98 (1.47)	0.511 (0.068)	0.944 (0.01)
MOD*PLP	5.70 (1.04)**	0.626 (0.087)	0.969 (0.01)**
PLP*MOV	8.43 (0.80)	0.399 (0.100)	0.955 (0.00)
MOD*PLP*MOV	5.74 (0.99)**	0.624 (0.086)	0.969 (0.01)**
MOD+MOV	9.96 (1.53)	0.512 (0.069)	0.944 (0.01)
MOD+PLP	5.62 (1.07)**	0.625 (0.088)	0.970 (0.01)**
PLP+MOV	8.30 (0.70)	0.403 (0.098)	0.955 (0.00)
MOD+PLP+MOV	7.84 (1.36)	0.513 (0.093)	0.957 (0.01)



(a) MOD (b) PLP

(c) MOV (d) MOD*PLP*MOV

(e) MOD+PLP+MOV

Figure 5: **Distribution of the class probability of conditioned HMM.** The decisions for the classes labeled to be utterances are shown by the black bars, while the decisions for the class laughter is shown by grey bars. Model settings are: 16 hidden states, transition matrix with random values on the not initialized elements and 6 mixture components.

Again the naive Bayes approach reduces the error rate while the the decisions remain confident. The decision averaging is still strongly influenced by the MOV classifier and therefore does not reaches a comparable performance.

5. Conclusion

Within this work CHMM have been applied to detect laughter in naturalistic multiparty dialogue and were compared to classical HMM. Additionally, a focus was set on late classifier fusion. While classical HMM can only be fused based on the probability ratios of the likelihood to have produced a given sequence, the CHMM comes along with the advantage of returning a probability over class labels. The setup of the study aimed at a transparent comparison and thus the setting of the examined CHMM were chosen according to the reference experiment based on the classical HMM. In order to investigate the performance of late classifier fusion, three different streams where fused using naive Bayes and decision averaging. It turned out that the performance of CHMM is in general close to the one of classical HMM. However, fusing the probabilistic outputs of the audio features of the CHMM outperform the classical HMM approach clearly. It is important to emphasize that, due to the restricted setup of the study, the capabilities of the CHMM and fusion architecture have by far not been driven to their limits. For instance, it is possible to combine different models having a variable number of hidden states, mixture components and transition matrices while the fusion process remains transparent. Furthermore, weighting the classifier decisions based on the classifier performance can be realized. Future work will investigate the fusion performance using different models and a reweighing of classifiers in greater detail. Furthermore, a layered CHMM architecture will be studied in the near future.

Acknowledgment 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). Special thanks to Prof. Nick Campbell for making the data available.

6. References

[1] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)

[2] Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition 2, 578–589 (1994)

[3] Maganti, H.K., Scherer, S., Palm, G.: 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). pp. 710–711. Springer (2007)

[4] Morency, L., Quattoni, A., Darrell, T.: Latent-Dynamic Discriminative Models for Continuous Gesture Recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)

[5] Quattoni, A., Collins, M., Darrell, T.: Conditional Random Fields for Object Recognition. In: Advances in Neural Information Processing Systems (2004)

[6] Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisul data. ACM Transactions on Interactive Intelligent Systems: Special Issue on Affective Interaction in Natural Environments (under review)

[7] Scherer, S., Schwenker, F., Campbell, W.N., Palm, G.: Multimodal laughter detection in natural discourses. In: Ritter, H., Sagerer, G., Dillmann, R., Buss, M. (eds.) Proceedings of 3rd International Workshop on Human-Centered Robotic Systems (HCRS'09). pp. 111–121. Cognitive Systems Monographs (2009)