INVESTIGATION OF HIERARCHICAL CLASSIFICATION FOR SIMULTANEOUS GENDER AND AGE RECOGNITION

Ingo Siegert, Ronald Böck, David Philippou-Hübner, Andreas Wendemuth

Chair of Cognitive Systems, Otto von Guericke University Magdeburg
ingo.siegert@ovgu.de

Abstract: For a successful speech-controled human-machine-interaction individualized models are needed. If the system is designed to run with many users for short times each, a complete user adaptation is not useful. A possible solution would be to use user-group pre-adapted models and recognize the group the actual speaker belongs to in the very first beginning of the interaction. In this paper we present investigate different methods to recognize age and gender groups with an hierarchical model to improve the recognition rate. We could prove, that our method could get adequate results on a four class problem compared with classical approaches.

1 Introduction

Research on user dependent Human-Computer-Interaction (HCI) has gathered much effort in recent years. For a successful dialogue individualized models for speech and emotion recognition are needed. One solution would be to adopt the models during dialogue using speech samples, another solution would be to use partly-individualized models beforehand. This partly individualized models are trained on specific acoustic characteristics of different speaker groups, e.g. age, gender. Using adapted or partly-individualized models in dialogue systems require that the models can come to a accurate and robust decision with only a few speech samples.

In speech recognition, independent age or gender detection has already been implemented (c.f.[2]). In this paper we concentrate on the question, whether an hierarchical setting using prior knowledge can gain accuracy. Therefore, we investigate different strategies to recognize age and gender using data form a realistic HCI experiment. First we implement a four-class problem as baseline, to distinguish male, female and also young and old speakers at once. Second we use hierarchical classifiers and answer the question which order of classifiers give the best performance: distinguish a) first the gender and then the age, or b) first the age and then the gender of a speaker.

The remainder of the paper is structured as follows: In the first section we describe the used data. Afterwards, we will present the used methods and the hierarchical classifiers. In the third section we describe the experiments and present the results and in in the next section we will discuss them. At the end we will conclude the experiments and give an outlook for further research.

2 Used Data

For our investigation we utilize the LAST MINUTE corpus. This corpus contains multimodal recordings from a WOZ experiment that allows to investigate how users interact with a companion system in a mundane situation with the need for planning, re-planning and strategy change. It represents a naturalistic human-machine-interaction with nearly balanced groups of native

German speakers, namely young or old male and young or old female speakers. The distribution of the two age groups is as follows: 18-28 years for the young group and over 60 years for the elder group.

The design of this experiment is described in [3], first results on affect recognition can be found in [1]. The corpus includes 136 speakers with nearly 56 hours recorded material. It contains recordings form four high quality video-cameras, two directional microphones and one headset microphone. Also biopsychological signals (skin reductance, heartbeat, respiration) where recorded for some participants.

We used a selected subset of 104 speakers, see Table 1. Furthermore we extracted all speaker utterances from the experiments. We ignored utterance parts where speaker and wizard talked the same time and also omitted laughter and other paralingual speech parts. In this way, we could utilize about 7 hours as training data. For test data, we use only one utterance "Ja" or "Nein" from each speaker.

Table 1 - Distribution of speaker groups

	Male	**Female**	∑
Young	27	28	55
Old	18	31	49
∑	45	59	104

3 Utilized Methods

To answer the question, whether a hierarchical pre-knowledge setting outperforms the classical approach, we used a Gaussian-Mixture-Model (GMM) utilizing the Hidden Markov Toolkit [4] with generally accepted Features: 12 Mel Frequency Cepstral Coefficients (MFCCs), their Deltas and Accelerations, the zero mean static coefficient, and the zeroth Cepstral Coefficient and Energy. Furthermore we used a hamming window with a length of 25ms and a step-size of 10ms. We also applied a preemphasis with 0.97 as parameter.

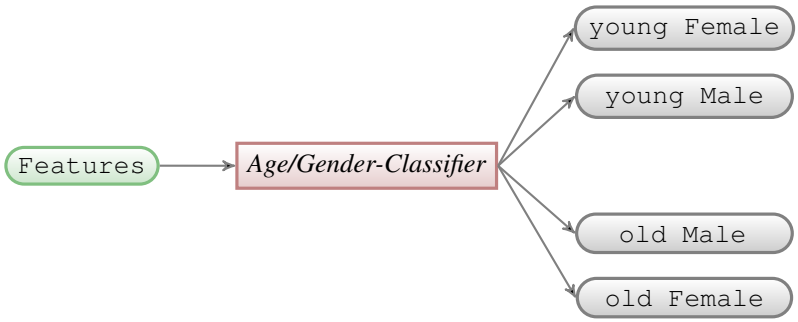


Figure 1 - Scheme of utilized 1x4-class classifier, using one age/gender classifier.

For comparison we trained a 1x4-class classifier as a baseline that is able to distinguish between alls four classes at the same time (see Figure 1). Furthermore, we utilized two hierarchical classifier sets. Both have in common, that they classify in a first step a shared attribute: age or gender respectively. Afterwards in a second step one of the four classes is recognized by two different classifiers, which only have to distinguish between young male and old male (having a gender classifier in the first step) or young male and young female (having an age classifier in the first step). The complete setting of both hierarchical classifiers is illustrated in the Figures 2

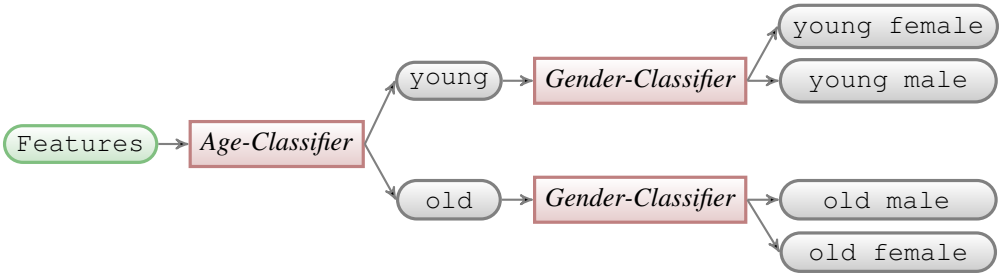


Figure 2 - Scheme of our age-gender-classifier, utilizing an age classifier in the first and a gender classifier in the second step.

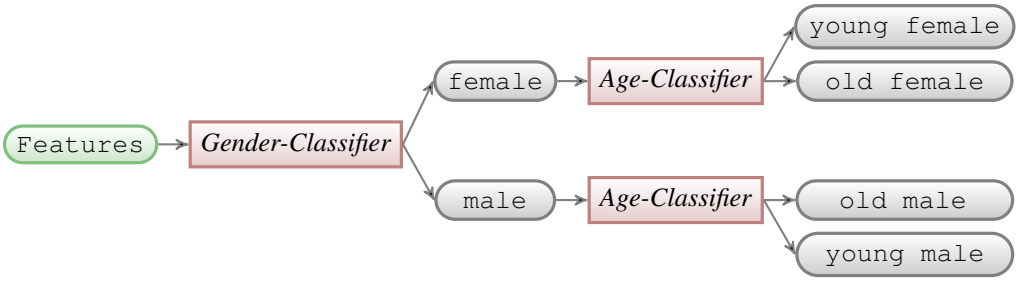


Figure 3 - Scheme of our gender-age-classifier, utilizing a gender classifier in the first and an age classifier in the second step.

and 3. Hence, within the following considerations we will simply use the term age-gender-classifier for the approach using an age-classifier before the final gender-classifier while term gender-age-classifier addresses the other classifier
Therefore, seven different classifiers where generated, one for the 1x4-class setting and three for both hierarchical 2x2-class settings. We further did a Leave-Speaker-Out validation, where we test the classifiers on speakers that where not used during training. Additionally we applied a ten-fold validation, to avoid influences of the training material.
To investigate question, with order of hierarchical classifiers gives the best accuracy and which recognition rate can be reached, we pursued different experiments and compared the combined recognition rates of our hierarchical setting against the recognition rate of our 1x4-classifier. To also measure the influence of different features and parameters of the model, we tested different settings utilizing MFCCs, different numbers of mixtures and training techniques.

4 Experiments

4.1 Classical 1x4-classifier Approach

First, we used different feature and model parameter combinations, to get an optimal baseline for our 1x4-classifier approach. The best combination is then used for our hierarchical approach. Our one-state GMM is trained with 80, 100, 120 and 140 Gaussian mixtures. The following feature combinations are used:

1. MFCCs with Deltas and Accelerations (MFCC_D_A)
2. MFCC_D_A and Energy (MFCC_E_D_A)
3. MFCC_D_A and zero Mean static Coefficient (MFCC_0_D_A)

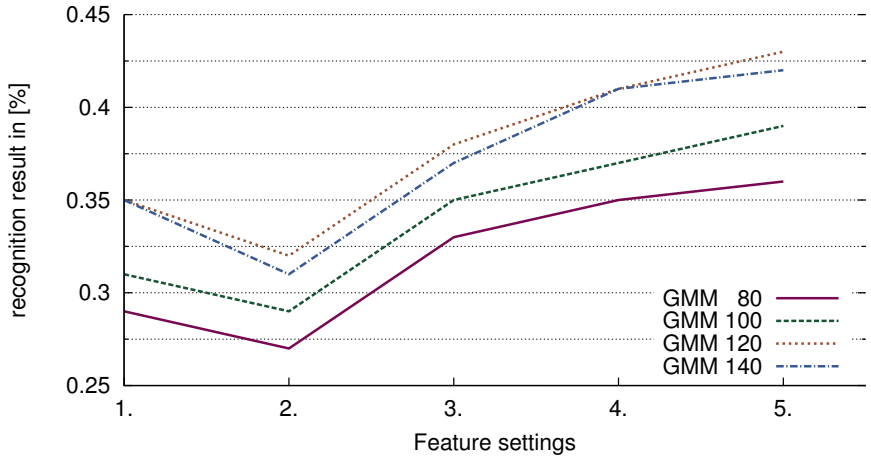


Figure 4 - Comparison of weighted average recogniton rates for different feature sets and model parameters of the 1x4-classifier.

4. MFCC_D_A and zeroth Cepstral Coefficient (MFCC_D_A_Z)
5. MFCC_D_A, zero Mean static Coefficient and zeroth Cepstral Coefficient (MFCC_0_D_A_Z)

The results for the recognition rates with different feature sets and model parameter can be found in Figure 4. This Figure gives a good impression, about useful recognition rates. A GMM with 120 mixture components gives the best result. When increasing the number of components up to 140 the recognition rate decreases. In terms of used features, we could gain the best results, when utilizing the zero Mean static Coefficient and zeroth Cepstral Coefficient with MFCCs (MFCC_0_D_A_Z). Utilizing the Energy instead the recognition rate fall down drastically. This is based on the used corpus, as this is gathered during a natural human-machine-interact the recording amplitude vary a lot. Which also affects the energy directly. The worst recognition rate with about 27% is gathered with a 80 mixture components GMM with MFCC_E_D_A, the best using 120 mixture components and MFCC_0_D_A_Z as features.

Table 2 - Confusion matrix of 1x4 classifier

		predicted				CORR
		y_F	**o_F**	**y_M**	**o_M**	%
label	**y_F**	**34**	1	11	-	0.739
	o_F	25	**13**	7	9	0.241
	y_M	14	4	**22**	4	0.50
	o_M	8	2	12	**5**	0.19

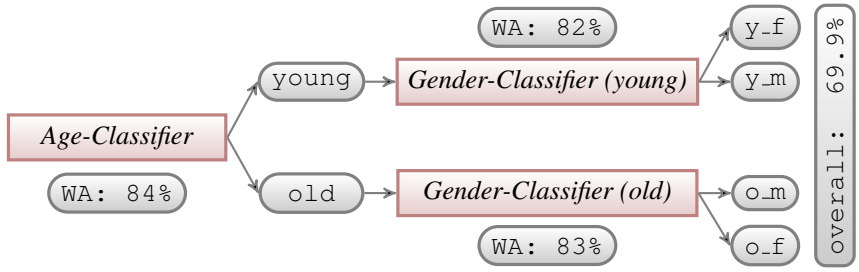
To get a better feeling, which characteristics are miss-recognized we also investigated the right and false recognitions. As for each individual validation set only a small amount of test samples is used, we summarized them over all ten sets. The resulting confusion matrix can be found in Table 2. When considering the confusion matrix we noticed two effects. The first one is, that for the young speakers we have a high confusion in the gender. 11 young Females are recognized as young Male, but only 1 as an elder Female and also 14 young Males where recognized as young Females, here we have 4 recognized as elder Female and also 4 as elder Male. Secondly,

for the elder speakers we notice a big confusion within the same gender. Here we also have the problem, that more speakers are miss-recognized: 13 elder Females are recognized correct, but also 25 are recognized as young Females. The same situation can be noticed for elder Males, where 12 are recognized as young and only 5 where correctly assigned. For both elder groups the overall variation also is increased, in addition to already mentioned miss-classifications, we also have 16 false-classifications (Male) for Female speakers and 10 for Male speakers. Therefore, it is likely that a age-classification as first step of our hierarchical classifier promise better recognition results.

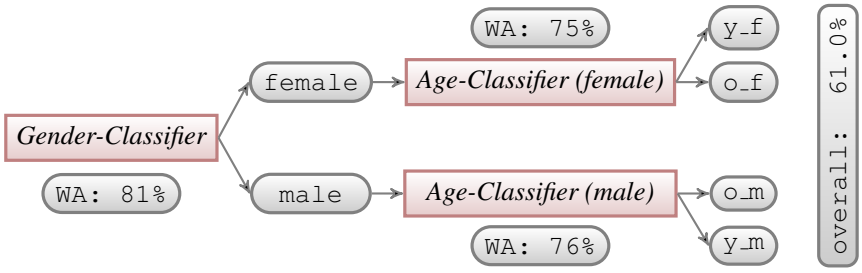
4.2 Hierachical 2x2-classifier Approach

As the investigations for the 1x4-classifier have shown the best results can be obtained, when using MFCCs with Deltas, Accelerations, zero Mean static Coefficient and Zero Mean static Coefficient together with 120 Gaussian mixture components. This setting will be also used for the hierarchical classifiers. Additionally the following model and training parameters are used: Gaussian Mixture Model with 120 mixture components, a hamming window, pre-emphasis with $\alpha = 0.97$.

To compare our both approaches, we trained several classifiers with subsets of utilized training material, each classifier now only have to distinguish between two groups, like male and female. As test-set we also use a specific amount of speakers that are not used for training as test set from the specific groups. Also here a ten-fold validation is used. We could reach a overall weighted average recognition rate of 69.9% for the age-gender classifier and 0.61% for the gender-age classifier. A detailed overview about separated recognition results in given in Figure 5.



(a) Age-gender-classifier



(b) Gender-age-classifier

Figure 5 - Individual recognition rates and combined recognition for both hierarchical classifier settings.

The results for both hierarchical classifier settings prove, that a two-step classification outperforms the classical one-step classification. This is somehow obvious, as we split the task into two independent problems. With our settings we can increase the recognition performance from 0.42% for the baseline classification system up to 69.9% for the best hierarchical one. As we

already assumed in the analysis of results from the 1x4 classifier setting, using a age-gender-classifier give the best result for the hierarchical setting.

For comparison we also analysed the single confusion matrices for all classifiers of both hierarchical settings. The Tables 3 and 4 present the matrices for two different hierarchical settings. It can be seen, that the confusion for all the both first step classifiers is much better than the direct confusion in the single-step approach. In decreasing the amount of data to discriminate, this also improves the recognition rates for the second step classifiers.
In [2], they are using 4 age groups: Child (7-14), Young (15-24), Adult (25-54) and Senior (55-80) toghether with gender discrimination whereby the children are not gender discriminated. They utilize five different corpora and get a weighted average of 48.1% for gender and 89.5% for gender recognition. But therefore they utilize five different classifier frontends and a fusion backend. Their combined result for age-gender detection is about 43%.

Table 3 - Confusion matrices of the age-gender-classifier

First level (age)				CORR
		predicted		
		y	**o**	%
label	**y**	**92**	18	0.836
	o	14	**76**	0.844

2nd level (gender young)				CORR
		predicted		
		y_F	**y_M**	%
label	**y_F**	**41**	9	0.820
	y_M	9	**41**	0.820

2nd level (gender old)				CORR
		predicted		
		o_F	**o_M**	%
label	**o_F**	**49**	11	0.816
	o_M	5	**25**	0.833

Table 4 - Confusion matrices of the gender-age-classifier

First level (gender)				CORR
		predicted		
		f	**m**	%
label	**f**	**89**	21	0.809
	m	17	**73**	0.811

2nd level (age male)				CORR
		predicted		
		y_M	**o_M**	%
label	**y_M**	**38**	12	0.760
	o_M	8	**22**	0.733

2nd level (age female)				CORR
		predicted		
		y_F	**o_F**	%
label	**y_F**	**45**	15	0.750
	o_F	12	**38**	0.76

5 Discussion

With our results we can state, that the hierarchical 2x2 setting, can outperform the single-step 1x4 setting. So splitting the classification task into sub-problems leads to a better recognition. Utilizing our approach, we could reach the following recognition rates: 41.8% weighted average recall rates for 1x4 classifier setting, 69.9% for age-gender setting, and a 61.1% for gender-age setting. This improvement of about 28% in totel is directly affected by the fact, that we split the problem into sub-problems. Both age and gender affect the same features in a similar way. A comparison with other research in that field, shows that our approach leads to comparable results.
Our approach can be further improved, by using an two-step Universal Background Model for example. Further improvement can be expected, when including context. This could be either long term prosodic features or linguistic. Until now we do not distinguish the content neither for training nor for testing. Especially for testing the prosodic context between “Ja” and “Nein” respective “Yes” and “No” is very large and can influence the recognition result.

6 Conclusion

We showed, that a hierarchical classification strategy could outperform classical strategies, where no prior information for age or gender is available and the classifiers have to discriminate

age and gender simultaneously. With our method we could reach a 69,9% weighted average classification rate, when using our two step age-gender-classifier and a 61% weighted average classification rate for gender-age classification.
This method is also applicable for similar problems, where two different properties are coded by the same features. Both age and gender change the voice in similar ways, thereby classify both simultaneously increases the confusion. Our next steps are to include also prosodic informations.

7 Acknowledgement

This research was supported by grants from the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

References

[1] FROMMER, J., B. MICHAELIS, D. RÖSNER, A. WENDEMUTH, R. FRIESEN, M. HAASE, M. KUNZE, R. ANDRICH, J. LANGE, A. PANNING and I. SIEGERT: Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus. In Proceedings of the Eigth International Language Resources and Evaluation (LREC’12), Istanbul, Turkey, 2012. accepted.

[2] MEINEDO, H. and I. TRANCOSO: Age and gender detection in the I-DASH project. ACM Trans. Speech Lang. Process., 7(4):13:1–13:16, Aug. 2011.

[3] RÖSNER, D., R. FRIESEN, M. OTTO, J. LANGE, M. HAASE and J. FROMMER: Intentionality in interacting with companion systems: an empirical approach. In Proceedings of the 14th international conference on Human-computer interaction: towards mobile and intelligent interaction environments - Volume Part III, HCII’11, pp. 593–602. Springer-Verlag, Berlin, Heidelberg, 2011.

[4] YOUNG, S., G. EVERMANN, M. GALES, T. HAIN, D. KERSHAW, X. LIU, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV and P. WOODLAND: The HTK book (for HTK Version 3.4). No. July 2000. Cambridge University Press, Cambridge, UK, 2006.