Construction and evaluation of a speaker identification and verification system using the YOHO database

Introduction

This report describes the construction and evaluation of a system for text-independent speaker identification and verification on a set of 10 assigned speakers of the YOHO database [1]. Already in 1995 D. Reynolds described a high-performance speaker identification and verification system based on Gaussian Mixture models (GMM), capable of a robust, statistically based representation of a speaker identity, which served as a model for this report [2]. For identification a maximum likelihood classifier was used and for verification a likelihood ratio hypothesis tester utilizing both the speaker model and a universal background model (UBM) was used.

Front-End of the system

At first, a front-end for the identification and verification systems was built. The front-end loads the training and test data needed and performs the feature extraction which transforms the raw signal into MFCC feature vectors, where speaker specific properties are emphasized, and statistical redundancies are suppressed. One MFCC feature vector is created from a 20ms window of samples of the original signal, progressing in steps of 10ms. Furthermore, a voice activity detector (VAD) was used to discard frames that contain no speech. For the MFCC feature extraction as well as the voice activity detection the Voicebox toolbox was used [3]. For each speaker a training set with over 34,000 and a test set with over 14,000 MFCC feature vectors was obtained. Therefore, no cross-validation was used, since the main reasons for using cross-validation is that there is not enough data available to partition it into separate training and test sets without losing significant modelling or testing capability [4].

Gaussian Mixture Models for speaker identification

Next, a straight forward speaker identification system for the assigned core-set of speakers (Appendix A) based on GMM without an UBM was built. To train the GMM on the gathered training data, the EM algorithm was used with a runtime of five iterations, which is considered as a sufficient number for this application scenario [5]. For initialisation of the EM algorithm the simpler K-Means algorithm was executed beforehand to improve convergence. Also, each component has its own full-rank covariance matrix determining the shape of the gaussian component. Moreover, to make the EM algorithm work robustly, regularisation of the covariance matrices was used. With this, a small value is added to the diagonal entries of all covariance matrices: $\sigma_{reg} = \sigma + \epsilon I$, with $\epsilon = 0.0001$. This ensures that the covariance matrix has a low condition number, which makes the computation of the inverse more stable [6]. For the implementation, I used code from another assignment I did for the lecture “Machine Learning” of the Computer Vision Group of the RWTH Aachen University [7].

Another crucial aspect for the successful application of GMM for machine learning tasks is the number of gaussian distributions. Thus, I conducted a performance test for different numbers of distributions for the GMM, which is illustrated in figure 1. The X-axis shows the number of gaussian distributions used for the EM algorithm and the Y-axis shows the accumulated Log-Likelihoods for the test data of the trained model. From this, a correlation between an increasing number of gaussian

components and better performance can be derived. The performance enhances with an increasing number of gaussians in a logarithmic scale. Consequently, for the further experiments a 64 component GMM was used.

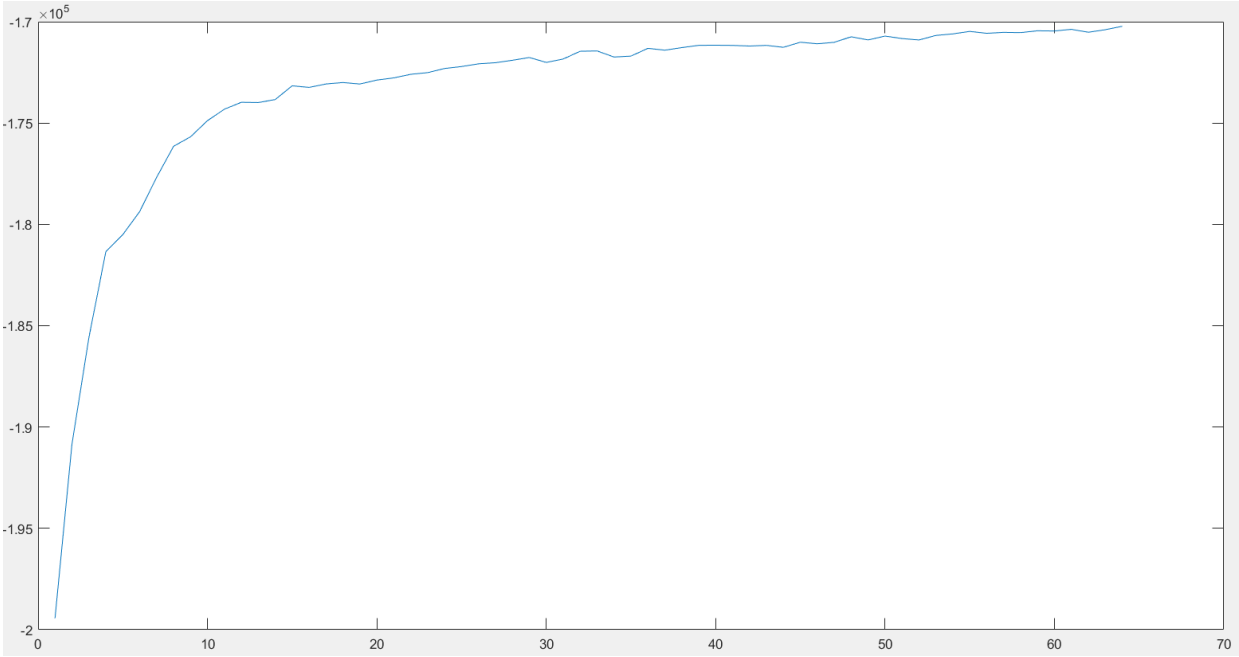

Figure 1: Accumulated Log-Likelihoods of the test-set vectors using GMMs with ascending number of gaussian distributions

GMM system evaluation

After the training of the GMM for each speaker in the identification system, the performance of it was tested with a maximum likelihood classifier for all utterances in the test set of each speaker. An utterance is represented by a set of approximately 360 MFCC feature vectors. In figure 2 the confusion matrix for the core-set of speakers is illustrated. The Y-axis shows the actual speaker (1-10) and the X-axis shows the prediction by the speaker identification system. The confusion matrix suggests that the system gives an accurate prediction with 201 misclassifications in 2957 utterances total, which is equivalent to an accuracy of 93.2%. But as other systems e.g. for speaker verification require even higher accuracy, there is still need for improvement.

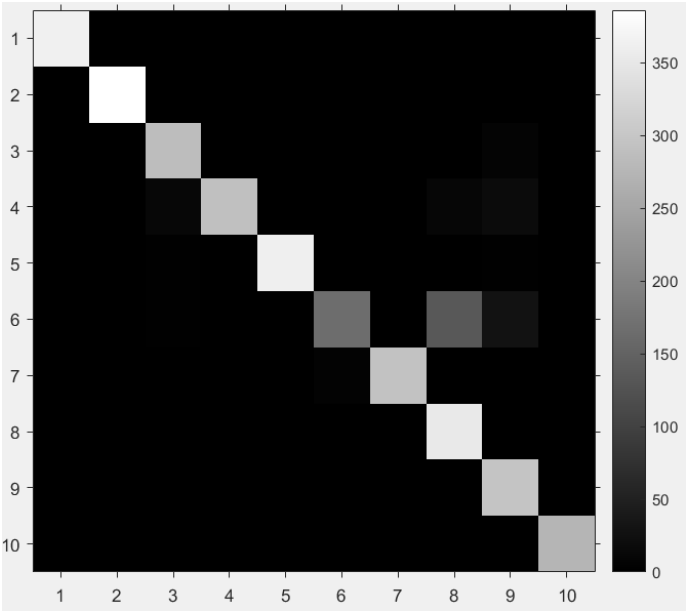

Figure 2: Confusion matrix for the 64 component GMMs, each trained on one of the core-set speakers

Universal Background Model

In the next step, a UBM was implemented with a larger set of speakers among the YOHO database. D. Reynolds describes that often for these background models, systems relied on speaker models that are similar to the target speaker, to emphasize the unique features of a speaker out of similar voices. But in practice it was discovered that this can have issues for speakers with very dissimilar voices, as neither the GMM nor the UBM is modelling the dissimilar voice well [2]. Therefore, a wide variety of speaker accents and a similar gender distribution as in the core-set of assigned speakers was chosen for the UBM to cover both similar and dissimilar voices for each speaker. A full list of the 47 speakers in the UBM dataset can be found in Appendix B. With this set, the UBM was trained in the same way as the other GMM previously.

GMM-UBM for speaker identification and verification

Next, the GMM-UBM system was created which was used for both speaker identification and speaker verification. With the UBM given, it is possible to adapt specific speaker models on the training data, without having to compute them from scratch with the EM algorithm as previously. This method gives a better estimate on where the speaker stands out of the background speakers, i.e. which features of the speaker are unique, and therefore leads to better performance.

The new GMM models for the GMM-UBM system were computed with the Maximum A Posteriori (MAP) Parameter Estimation, which consists of two steps. The first step is identical to the "Expectation" Step in the EM algorithm, where estimates of the sufficient statistics of the training data are computed for each mixture in the prior model. In the second step, these "new" sufficient statistic estimates are then combined with the "old" sufficient statistics from the prior mixture parameters using a data-dependent mixing coefficient for adaptation. The data-dependent mixing coefficient is designed so that mixtures with high counts of new data rely more on the new sufficient statistics for final parameter estimation and mixtures with low counts of new data rely more on the old sufficient statistics for final parameter estimation [8]. The mixing coefficient contains a constant relevance factor r. Researchers suggest that the relevance factor is dependent on the data and application, but for GMM based speaker recognition a factor of 5-10 is recommended [9; 10]. Thus, the relevance factor was set to $r = 9$. Furthermore, in speaker recognition applications it is common to only adapt certain GMM parameters, such as only the mean vectors. Hence, for the GMM-UBM

system only the mean vectors were adapted. For the implementation of this, the Netlab toolbox was used, which provides an implementation of the MAP estimation [11].

For the verification system a likelihood ratio hypothesis tester was used, utilizing both the GMM of the speaker λ_{target} and the UBM λ_{UBM}. When a speaker is attempting to verify his identity, his utterance is analysed with the sum of the MFCC vectors x_t of the utterance. The speaker is accepted, if the Log-Likelihood Ratio of $LLR = \sum log\, p(x_t|\lambda_{target}) - logp(x_t|\lambda_{UBM})$ is greater than a certain threshold θ [5]. The threshold can be determined in different ways. Researchers suggest among others the estimation of θ with the relation-based approach, which seeks a threshold that satisfies a specific relationship between FAR (false acceptance rate) and FRR (false rejection rate) curves [12]. This approach was also used here by evaluating the minimum likelihood of an utterance in the training data of a speaker and the maximum likelihood in the training data of the UBM predicted by the GMM-UBM system and setting the threshold θ to the average of these two variables.

GMM-UBM system evaluation

On the previously introduced speaker identification task, the GMM-UBM system did better than the previous system. In fact, the accuracy of the system could be increased substantially. The confusion matrix in figure 3 illustrates this enhancement to a high-precision prediction model with only 6 misclassifications in 2957 utterances total, which is equivalent to an accuracy of 99.8%.

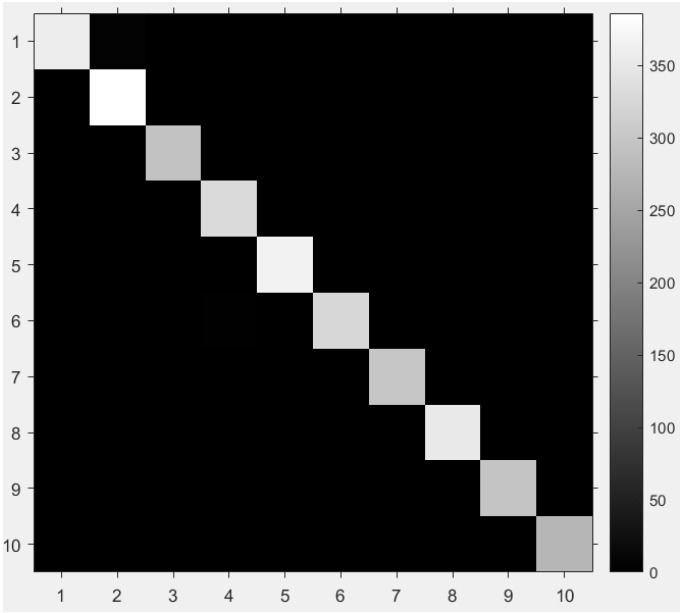

Figure 3: Confusion matrix for the 64 component GMM-UBM systems, each trained on one of the core-set speakers

For assessing the performance of the speaker verification system built from the GMM-UBM system, a different testing approach named impostor trials was used. Impostor trials are used to assess the reliability of a verification system. Unseen speakers (impostors) get selected and together with unseen data samples from the target speaker attempt to get accepted by the verification system [2; 5]. As impostors, a set of 12 unused speakers from the YOHO database was utilized (Appendix C), consisting of four similar speakers (same accent and sex distribution as in core-set speakers), and 8 dissimilar speakers, as a mixture of different speakers provides the best evaluation of the system like previously discussed in the UBM section.

Speaker	1	2	3	4	5	6	7	8	9	10	**Average**
FRR (%)	18.73	0.26	8.19	13.98	8.77	75.38	1.33	10.23	6.73	18.91	16.25
FAR (%)	1.54	4.27	0.15	0.0005	0.0002	0	0	0	0.4	0.0005	0.65

The table above shows the performance of the GMM-UBM models of speakers in the core-set. Firstly, one can see that the average FRR is quite high with 16.25% while the average FAR is very low with 0.65%. This suggests that the threshold θ of the GMM-UBM models could be set to a slightly lower level. However, the FAR could increase therewith, which is not desirable for a verification system. Furthermore, the verification system performed very differently among the core-set of speakers on the designed impostor trials showing FRRs of between 75.38% and 0.26% and FARs of between 4.2% and 0%, which might be due to the fact that for each speaker a individual threshold θ was set, although the same method to estimate θ was used.

Discussion

In this report, two systems for text-independent speaker identification based on GMM were developed and evaluated. Using the probabilistic speaker representation of the GMM, the systems were defined by a maximum likelihood classification. Both systems achieved a high accuracy on the dataset (>93%) but only the GMM-UBM system provided a satisfactory accuracy for verification purposes (99.8%). Therefore, with this system a speaker verification system was built, using likelihood-ratio hypothesis testing. This system performed not as accurate as the system for the identification task, but still comes to a fair result on the designed testing approach. The FAR, which is crucial for a verification system, is on a low level with 0.65% in average. However, the FRR is quite high for some of the speakers in the core-set, which could limit the usability of the system. Further improvement could be done by using a different technique to estimate the threshold θ for the speaker verification system as well as parameter tuning for the training of both the UBM as well as the GMM-UBM system.

Conclusively, the results from the paper of D. Reynolds on the YOHO dataset were mostly confirmed. On the speaker identification task, D. Reynolds reports an accuracy of 99.5% and the system described in this report achieved a similar, even higher accuracy of 99.8%. Also, on the speaker verification task, the FAR of the system described in this report is close to the described rate of Reynolds with an average of 0.65% for this system and an average of 0.1% for Reynold's system. However, the FRR of the speaker verification system of this report is substantially higher (16.25%) than in the Reynolds paper (0.65%). This could be based on the fact that Reynold determines the threshold θ after the conducting the impostor trials, so that FAR and FRR are minimized, whereas this system used training data of the speaker and the UBM to set the individual threshold for the speaker, which was done to provide a fully independent and practical testing approach.

References

1. Campbell, J. P. (1995, May). Testing with the YOHO CD-ROM voice verification corpus. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (Vol. 1, pp. 341-344). IEEE.
2. Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech communication*, *17*(1), 91-108.
3. Brookes, M. (1997): "Voicebox: Speech processing toolbox for matlab". *Software, available [Jan. 2018] from* http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.
4. Ertel, W. (2009). Grundkurs Künstliche Intelligenz. *Auflage, Wiesbaden*.
5. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12-40.
6. Christopher M. Bishop (2006): "Pattern Recognition and Machine Learning", Springer.

7. Computer Vision Group, RWTH Aachen University, https://www.vision.rwth-aachen.de/course/19/.
8. Reynolds, D. (2000): “Gaussian Mixture Models”, MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.
9. You, C., Li, H., Ma, B., & Lee, K. A. (2012). Effect of Relevance Factor of Maximum a posteriori Adaptation for GMM-SVM in Speaker and Language Recognition. In *INTERSPEECH* (pp. 2065-2068).
10. Marc, F., Koichi, S., & Sadaoki, F. (2010). Inter-speaker weighted MAP adaptation for GMM-supervector speaker recognition. *研究報告音声言語情報処理 (SLP)*, *2010*(12), 1-4.
11. Nabney, Ian T. (2004): “Netlab Toolbox 3.3”. http://www.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/, accessed at 15/01/2018.
12. R. Diaz-Amador, E. Castillo-Guerra and J. Cardenas-Barrera, "Adaptive management of speaker verification thresholds," *2008 Canadian Conference on Electrical and Computer Engineering*, Niagara Falls, ON, 2008, pp. 001763-001766.

Appendix

A. Core-set of assigned speakers

ID	Gender	Accent
167	F	new york
172	F	new york
229	M	new york
233	M	new york
238	M	new york
239	M	new york
241	M	new york
242	M	new york
243	M	new york
244	M	new york

B. Set of speakers for UBM

ID	Gender	Accent
124	M	
146	F	
147	M	estonia
140	M	hong kong
110	M	midwest
115	M	midwest
120	M	midwest
138	M	midwest
133	M	N.Y.,LongIsland
104	M	new jersey
105	F	new jersey
107	M	new jersey
108	M	new jersey
109	M	new jersey
116	M	new jersey

122	F	new jersey
142	M	new jersey
143	M	new jersey
101	M	new york
103	M	new york
106	M	new york
111	M	new york
112	M	new york
113	M	new york
114	M	new york
117	M	new york
118	M	new york
119	M	new york
121	M	new york
125	M	new york
126	M	New York
127	M	new york
130	M	new york
131	M	new york
132	M	new york
134	M	new york
135	F	new york
137	F	new york
141	F	new york
148	M	new york
150	M	new york
139	M	northeast
128	M	ny-brooklyn
136	M	pittsburgh
102	M	south
144	M	south
145	M	upstate NY

C. Set of impostor speakers

ID	Gender	Accent
188	M	canada/NE
198	M	chinese
181	M	long island, NY
192	F	maryland
152	F	new jersey
159	M	new jersey
163	F	new jersey
164	M	new jersey
156	M	new york
157	M	new york
158	M	new york
161	F	new york