GMM-HMM System for Text-dependent Speaker Verification

Name of author

Address - Line 1
Address - Line 2
Address - Line 3
email@address

Abstract

The GMM is a widely popular approach for modeling the acoustic information and HMM is popular for temporal dynamics modeling. We used to GMM-HMM model to represent and characterize the speaker and temporal information in the text-dependent speaker verification task. The entire system is developed on Kaldi.

1. System Description

In the text-dependent speaker verification has been popular in recent years due to its applicability in smart devices. Earlier temporal modeling approaches such as dynamic time warping (DTW), a hierarchical multi-layer acoustic model (HiLAM)is proposed based on Gaussian mixture model (GMM)-hidden Markov model (HMM) architecture, i-vector/HMM and unsupervised HMM-universal background model (UBM), joint speaker-utterance model with GMM-HMM, DNN-HMM, etc. These techniques uses speaker modeling and temporal modeling techniques together for text-dependent SV task [1–8]. In this work, we are using GMM-HMM model for text-dependent speaker verification.

1.1. Feature Extraction

The short term processing is performed on the entire database with a frame size of 25 ms and a shift of 10 ms. 60-dimensional mel frequency cepstral coefficient (MFCC) features including delta coefficient are extracted for every frame considering 23 logarithmically placed mel filters. There is no voice activity detection applied as the SV framework used in this work as we can ignore the silence frames in likelihood computation. We have not used any other data except Task1 SdSV data [9]. Also, we have not used any data augmentation schemes. We have only submitted a single system.

1.2. Training and Enrollment

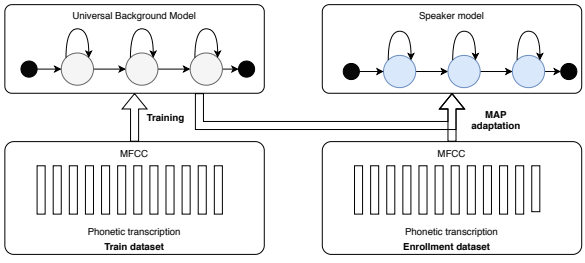

Figure 1: *Training and Enrollment for GMM-HMM speaker verification.*

The lexical information in a speech can be realized in terms of phonetic sequence and speaker information is captured in MFCC features. We used the phonetic transcription provided by the organizers. We train the monophone HMM models using the entire train set and build a Universal background model (λ_{UBM}). During Enrollment, we adapt the speaker-utterance model using the pair of speaker and utterance ids using MAP adaptation. In the experiment, we use the smoothing constant τ in the MAP adaptation script of Kaldi as 15. We call it a speaker-utterance model, i.e., ($\lambda_{spk-utt}$). Figure 1 shows the training and enrollment procedure.

1.3. Testing

During the testing phase, we use the testing features X and the claimed speaker-utterance model $\lambda_{spk-utt}$. We use transcription $\mathcal{W}$ from the claimed speaker-utterance model, i.e., utterance id. To compute the background likelihood scores, we use the universal background model λ_{UBM}. Thus, the final likelihood score can be computed as follows:

$$\mathcal{S}_X^{\mathcal{W}} = \log P(X|\lambda_{spk-utt}, \mathcal{W}) - \log P(X|\lambda_{UBM}, \mathcal{W}) \quad (1)$$

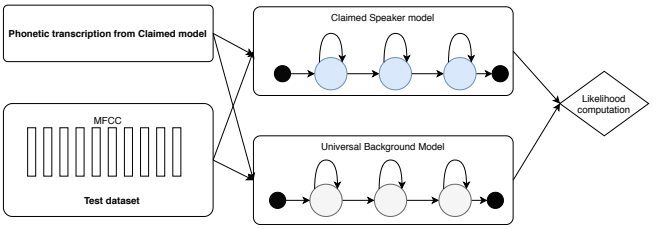

Figure 2: *Testing for GMM-HMM speaker verification.*

The likelihood computation procedure in the testing phase is shown in Figure 2. In testing score submission, we have not used any score normalization approaches. We neglected silence as it does not contribute to the speaker verification task.

2. Results

There are two baseline systems provided by the organizers, namely, i-vector/HMM [7] and X-vector [10]. These approaches are widely adopted in both text-dependent and text-independent speaker verification systems. Results are analyses in different cases. Experimental results are tabulated in Table 1 and Figure 3 shows the DET curve for two baseline systems and our system. It can be observed that our system outperforms the x-vector system in the majority of the test condition and operating threshold on DET curve but it does not outperform

Table 1: *Experimental Results*

Trial conditions	EER			min DCF		
	X-vector	I-vector/HMM	Our system	X-vector	I-vector/HMM	Our system
progress	9.05	3.47	6.96	0.529	0.1472	0.3146
progress-male	7.9	2.53	6.54	0.4927	0.1302	0.319
progress-female	9.54	4.05	7.17	0.5348	0.1569	0.3004
progress-EN	9.33	3.56	6.79	0.4645	0.12	0.2535
progress-FA	8.9	3.43	6.9	0.534	0.1659	0.3313
progress-TC-vs-IC	4.47	3.07	5.69	0.1987	0.1021	0.2451
progress-TC-vs-TW	19	4.92	10.17	0.7719	0.24	0.452
progress-EN-male	7.99	2.53	6	0.4193	0.1052	0.2466
progress-EN-female	9.95	4.12	7.2	0.4742	0.1285	0.2559
progress-FA-male	7.98	2.47	6.41	0.5066	0.1422	0.3229
progress-FA-female	9.21	3.97	6.94	0.5382	0.1795	0.3156
evaluation	9.05	3.49	7.01	0.5287	0.1464	0.3163
evaluation-male	7.77	2.41	6.49	0.4919	0.1277	0.3171
evaluation-female	9.62	4.08	7.25	0.5364	0.157	0.3042
evaluation-EN	9.32	3.55	6.91	0.4611	0.1204	0.255
evaluation-FA	8.87	3.43	6.91	0.5388	0.1642	0.3333
evaluation-TC-vs-IC	4.44	3.04	5.75	0.1984	0.1019	0.2456
evaluation-TC-vs-TW	19.02	4.86	10.2	0.7747	0.2387	0.4527
evaluation-EN-male	7.9	2.48	6.09	0.4167	0.1041	0.2456
evaluation-EN-female	9.97	4.18	7.27	0.4694	0.1299	0.2586
evaluation-FA-male	7.8	2.36	6.29	0.5058	0.1394	0.3182
evaluation-FA-female	9.31	4.03	7.07	0.5441	0.1775	0.3214

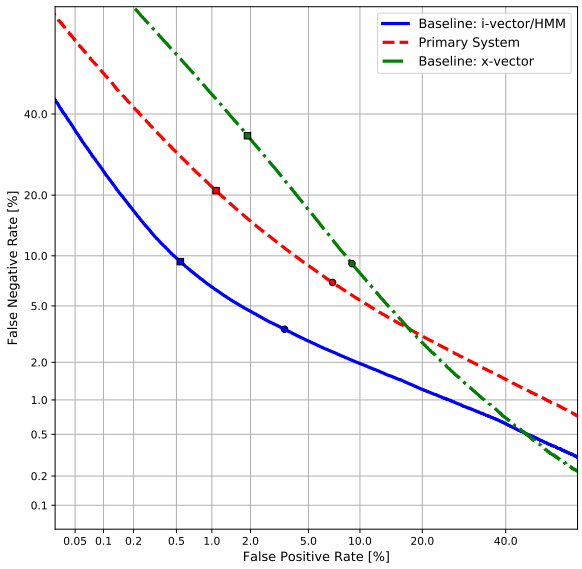

Figure 3: *DET curve.*

i-vector/HMM for all the cases. Overall results indicate that lesser speaker discrimination performance (TC-vs-IC). It is better than the x-vector system for TC-vs-TW trial categories.

3. References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[3] H. Zeinali, H. Sameti, L. Burget, J. ernock, N. Maghsoodi, and P. Matjka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge," in *Interspeech 2016*, 2016, pp. 440–444.

[4] Y. Liu, L. He, Y. Tian, Z. Chen, J. Liu, and M. T. Johnson, "Comparison of multiple features and modeling methods for text-dependent speaker verification," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*. IEEE, 2017, pp. 629–636.

[5] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using un-supervised hmm-ubm and temporal gmm-ubm," in *Interspeech 2016*, 2016, pp. 425–429.

[6] R. K. Das, M. Madhavi, and H. Li, "Compensating utterance information in fixed phrase speaker verification," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1708–1712.

[7] H. Zeinali, H. Sameti, and L. Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1435, 2017.

[8] H. Zeinali, L. Burget, H. Sameti, O. Glembek, and O. Plchot, "Deep neural networks and hidden markov models in i-vector-based text-dependent speaker verification," in *Odyssey 2016: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 21-24, 2016*, L. J. Rodríguez-Fuentes and E. Lleida, Eds. ISCA, 2016, pp. 24–30.

[9] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2020:the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.