Speaker Reconition

Dhaval Parmar

Artificial Intelligence
EECS Division, Indian Institute of Science
dhavalparmar@iisc.ac.in

Abstract

Speaker recognition is the task of automatically recognizing who is speaking using speaker specific features. It is one of the fundamental and important task in the domain of speech information processing. The class of problems can further be classified as speaker verification and speaker identification. It is being studied from number of decades due to various practical applications in number of fields like customer services, security verification, forensic, ome automation and many more. This paper talks about two approaches, using Gaussian Mixture Model and Universal Background Model (GMM-UBM) based approach [1] and more advanced technique using lower dimensional i-vector [2][3]. Results are compared using standard TIMIT data set.

Index Terms: speaker Recognition, GMM-UBM, i-vector, Total Variability Space

1. Introduction

Speaker recognition is well studied task from number of decades due to various important application. Approach being used is continuously improved and still it is an active area of research.

Initially for decades Gaussian Mixture Model (GMM) was widely used for text-independent speaker recognition application [4][5]. GMM-UBM based approach uses universal background model and speaker specific model represented by GMM for speaker recognition using likelihood probability. Then joint factor analysis (JFA) [6][7] shown great improvement for text-independent task. JFA provide procedure to model inter-speaker variability and to compensate channel variability using GMM. Application of support vector machines (SVMs) on supervector space [8] has also shown promising results. Kernel function represents linear approximation of Kullbuck-Leibler(KL) distance between UBM and speaker GMM. We can take benefit of both of this method. Using factor analysis we can convert speech observation vector to low dimension total variability space. This extracted feature are known as i-vector. Cosine kernel is used to get score for target and test speaker. Linear Discriminant Analysis (LDA) can be used to increase inter speaker variability and reduce intra-speaker variability.

This paper describes and compare both GMM-UBM and i-vector based approaches for speaker verification. I have also tried improving performance by redefining scoring function using various kernel functions.

2. Approach Description

2.1. GMM-UBM

Approach mainly based on likelihood ratio of observation sequence for general universal background model and speaker specific model. For text independent speaker recognition task where we do not have knowledge of spoken words Gaussian mixture model (GMM) is better way to represent likelihood of utterances. UBM can be represented by GMM representing likelihood of speaker specific features of whole speaker corpus. Parameters of GMM can be found using Expectation maximization (EM) algorithm. Speaker model can be built by updating UBM parameters using speech features belongs to particular speaker using [eq. 11-13][1].

For speaker verification from test features X we can find likelihood ratio P(X/λ_{spk})/p(X/λ_{ubm}) were λ represents parameter of GMM model. Using certain threshold level from likelihood ratio we can decide speaker is same as speaker or not.

2.2. I-vector

This approach is inspired from joint factor analysis (JFA) [6][7]. In JFA supervector is extracted considering two different space for speaker dependent and channel dependent supervector. We can consider speaker and channel dependent supervector as M = m + Tw where m is speaker and channel-independent supervector, T is total variability matrix and w is low dimensional intermediate vector(i-vector). So there is no distinction between speaker and channel variability.

Initially we can built UBM as GMM using all the utterances of different speakers. Baum-Welch zeroth, first and second order statistic can be found using UBM [2][3]. Hidden variable w can be define as posterior distribution condition on Baum-welch statistic. This posterior distribution is gaussian distribution and mean of this distribution represents i-vector. Parameter of this distribution can be found using T matrix whic is initialised randomly and covariance matrix Σ which is initialised using UBM covariance matrix. We can update this parameters using EM algorithm [2,3]. In expectation step parameters of posterior distribution of w are calculated for each utterances using current estimate of T and Σ. In maximization step T and Σ are updated by solving set of linear equations.

For speaker verification w vector are found for target and test utterances using final T and Σ parameters. Generally cosine kernel is used to find similarity score between utterances. As speaker and channel variability is not distinct here, linear discriminant analysis (LDA) is used to compensate effect of channel in i-vector space. From this score using certain threshold we can identify test speaker is same as target speaker or not. To improve scoring I tried using various kernel functions and compared the results.

3. Experiment

I have experimented for speaker verification task using both methods. I am using standard TIMIT Speech Corpus having 630 speakers and 10 utterance per speaker. Due to computation and memory limitations I am using 10 speaker data to build UBM.I am using 10 utterances of each speaker as tar-

Figure 1: *Average Likelihood Ratio*

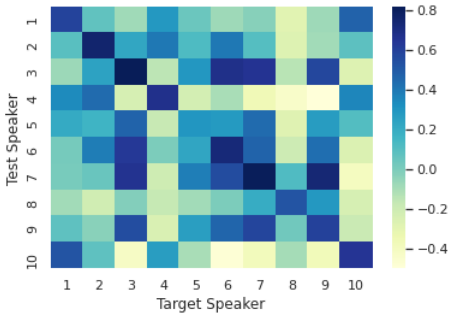

Figure 2: *Average Similarity Score*

Table 1: *Results 10 test speaker and 500 utterances*

Approach	FAR	FRR	Accuracy(%)
GMM-UBM	0.142	0.058	78
I-vector	0.027	0.094	83

get and 5 utterance of each speaker as test set(not same as target). For gaussian mixture moddel 64 number of clusters are being used.I am using 39 dimensional speaker specific feature which is concatenation of 13 dimension mel-frequency cepstrum coefficients (MFCC) feature, delta vector and double delta vector. To extract MFCC feature I am using 25 ms hamming window with step of 10ms. For verification I ave kept threshold of 1 for likelihood ratio in GMM-UBM approach and 0.6 for similarity score in i-vector approach. To improve scoring I have tried different kernel functions like Multiquadric kernel and Rational Quadratic Kernel. Results for that are shown in Table 2. As we can see results are deteriorating for other kernel functions. Code for implementation can be found at https://github.com/DhavalParmar61/Speech-Information-Processing.git

4. Results

Here to compare the results I am using False Acceptance Rate (FAR) and False Rejection Rate (FRR) as metric. FAR is percentage of accepted un-authorised person and FRR is percentage of rejected authorised person. Results for both the method are shown in table. As we can see from the results FAR is more compare to FRR for GMM-UBM while for I-vector base approach FAR is better compare to FRR. As it is more important to have lower FAR for application like security authentication I-vector approach is better compare to GMM-UBM method. Average Likelihood ratio and similarity for various target and test speaker are shown in figure 1 and 2. I-vector approach can identify speaker with better confidence because as we can see difference in likelihood ratio for same speaker and different speaker is less compare to difference in similarity score. Also accuracy of speaker verification that is percentage of correct verification is more for i-vector base approach compare to GMM-UBM approach.

Table 2: *Results with various kernel function for i-vector approach*

Kernel	FAR	FRR	Accuracy(%)
Multiquadric	0.156	0.092	75.2
Rational Quadratic	0.266	0.082	65.2

5. Conclusion

This paper describes two fundamental and important technique for speaker recognition GMM-UBM and I-vector based approach. I have compared results on TIMIT data set and shown that I-vector based approach can do better compare to GMM-UBM based approach. Due to recent advancement in Deep Learning field we can learn better embedding features compares to I-vector wich is known x-vector. Various recent research shows [9] that x-vector based approach outperforms I-vector based speaker recognition task. My future work will focus on speaker verification using x-vector features.

6. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models digital signal processing," *Digital Signal Processing*, vol. 10, pp. 19–341, 2000.

[2] K. Patrick, B. Gilles, and D. Pierre, "Eigenvoice modeling with sparse training data," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 13, no. 3, 2015.

[3] D. Najim, K. Patrick, D. Reda, D. Pierre, and O. Pierre, "Eigenvoice modeling with sparse training data," *IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING*, 2009.

[4] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 293–296, 1990.

[5] D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," *Ph.D. thesis, Georgia Institute of Technology*, 1992.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, *Digital Signal Processing*.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, *IEEE Transaction on Audio, Speech and Language*.

[8] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," *IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse*, vol. 1, pp. 97–100, 2006.

[9] S. David, G. Daniel, S. Gregory, and P. D. K. Sanjeev, "Generative x-vectors for text-independent speaker verification," *IEEE Spoken Language Technology Workshop (SLT)*, 2018.