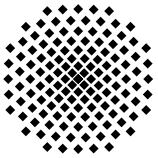

University of Stuttgart
Institute for Signal Processing and System Theory
Professor Dr.-Ing. B. Yang

LAB

Statistical Signal Processing

Pattern Recognition

Author: Yuxin Liu, Shanqi Yang, Qianqian Wei

Date of work begin: Date of work begin TBD
Date of submission: Date of submission TBD

Supervisor: Lukas Mauch

Keywords: Prostate Cancer Segmentation, Speaker Recognition

Abstract: Pattern recognition deals with a wide variety of problems. Methods of pattern recognition are prevalent in applications, such as the facial recognition and speech recognition systems. This pattern recognition lab includes two tasks, prostate cancer segmentation and speaker identification, which cope with medical imaging area and authentication field, respectively. The prostate cancer segmentation aims to detect the cancerous prostate area automatically, which saves cost and time for patients and medical workers. The speaker identification system can be used in several apps for authentication, like online banking, email services. As a consequence, the prostate cancer segmentation achieves an acceptable performance, and speaker identification system performs greatly.

Contents

1. Introduction

1.1. Evaluation

In last section we built up our generation and identification model and in order to get best performance of the model, It is necessary to analysis the hyper-parameters. We applied grid-search method on the parameters including SNR (Signal Noise Ratio), Maximum iterations and covariance types. All of these parameters may have significant impact on the model performance.

1.1.1. Signal Noise Ratio Analysis

In the previous section , we have implemented a voice detector to separate voiced frames and unvoiced frames, its mathematic equation follows . Now the threshold of the voice detector is going to ensured according to cross-validation performance.

As it is seen in the Table.1.1, the model has best Detection Rate when SNR threshold $\gamma = 10$.

1.1.2. Convergence Analysis

In previous section, new model coefficients are generated using EM algorithm. However the model doesn't converge with only one iteration. Then we tuned the maximum iterations of Gaussian Mixture Model , aiming to figure out when convergence achieved and its impact on detection rate. See the Table.1.2, we take expected value among cross validation sets under different maximum iteration. When only one iteration, the GMM model seems to be underfitting because both detection rate and log-PDF are low. With maximum iteration increasing, both detection rate and log-PDF raise. When maximum iteration equals to 3, expected detection rate reaches the peak, and afterward GMM seems to be overfitting that detection rate goes down with log-PDF arising. Due to that GMM api from Sklearn will automatically report convergence, so we know that the model is converged until maximum iteration is close to 100, but it is definitely overfitting.

But only tunned maximum iteration is not sufficient for analyzing convergence and performance, we also put the covariance types into consideration. Table.1.2 implements with FULL

Table 1.1.: Detection Rate with different SNR threshold

	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$	$\gamma = 50$	$\gamma = 100$
Expected Detection Rate	0.994117	0.994705	0.995294	0.993529	0.991764

Table 1.2.: Performance & Convergence Analysis with FULL covariance type

	Expected Detection Rate	Expected Log-PDF	Cross_Valid Time (minutes)
$Max\_iter = 1$	0.995294	-47.724	39.2
$Max\_iter = 2$	0.997058	-52.958	34.5
$Max\_iter = 3$	0.997647	-53.023	38.8
$Max\_iter = 4$	0.997058	-53.031	37.1
$Max\_iter = 5$	0.997058	-53.021	43.2
$Max\_iter = 6$	0.995294	-53.000	46.0
$Max\_iter = 7$	0.994117	-52.966	60.0
$Max\_iter = 8$	0.992352	-52.935	57.1
$Max\_iter = 9$	0.991176	-52.924	59.2
$Max\_iter = 10$	0.991764	-52.930	61.5

Table 1.3.: Performance & Convergence Analysis with DIAGONAL covariance type

	Expected Detection Rate	Expected Log-PDF	Cross_Valid Time (minutes)
$Max\_iter = 1$	0.864117	-36,822	11.4
$Max\_iter = 2$	0.93	-36.606	15.1
$Max\_iter = 3$	0.957058	-36.651	16.7
$Max\_iter = 4$	0.958235	-37.744	17
$Max\_iter = 5$	0.965882	-36.816	18.7
$Max\_iter = 6$	0.97	-36.870	21.2
$Max\_iter = 7$	0.971176	-36.921	21.3
$Max\_iter = 8$	0.968235	-36.969	21.1
$Max\_iter = 9$	0.972352	-37.006	19.8
$Max\_iter = 10$	0.972941	-37.031	20.9

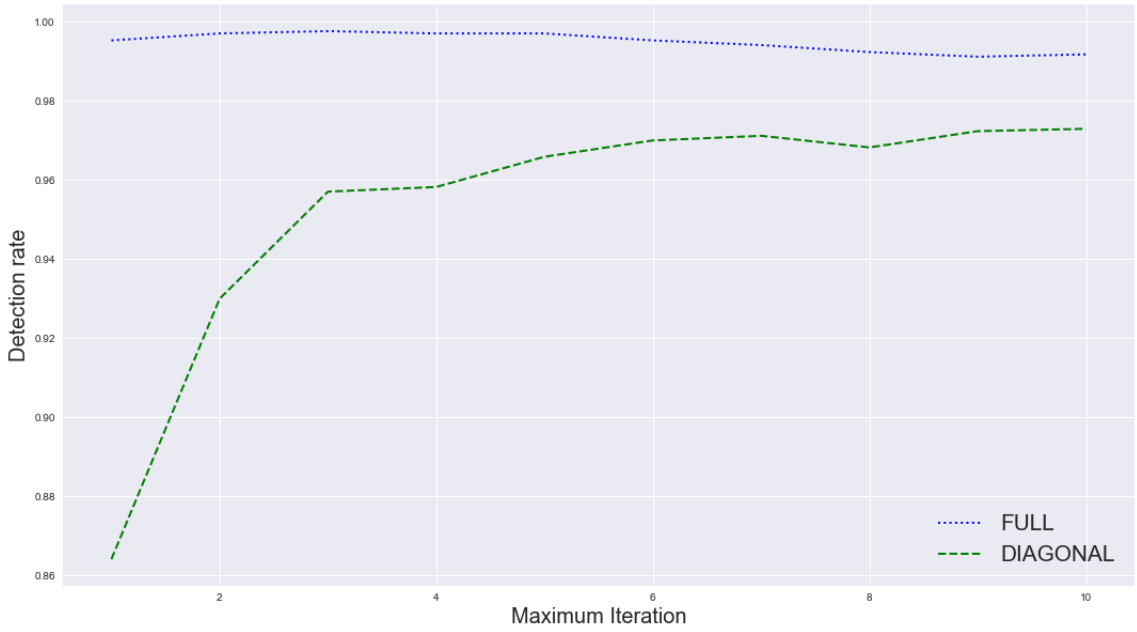

Figure 1.1.: Variation of Detection Rate with different covariance types in models

covariance type and Table.1.3 implements with DIAGONAL covariance. We make a contrast with both covariance types, the calculation time of FUll covariance is 2 or 3 times of DIAGONAL covariance's time. It is easy to think over that FULL covariance has more coefficients than DIAGONAL covariance so it will takes more time when FULL covariance. Furthermore, we plot the variation of detection rates of two covariance types.

Figure.1.1 illustrates that FULL covariance has faster convergence speed and better performance than DIAGONAL covariance does, detection rate of DIAGONAL covariance has a gap 2% between detection rate of FULL covariance when both reach peak plain. To figure out why, Figure.1.2 illustrates covariance types have the different effects on the GMM model distribution.

From the Figure.1.2, FULL covariance type may lead to a sloped ellipse but DIAGONAL covariance lead to an upright ellipse distribution. So our case, few GMM model with FULL covariance type can fit our voice dataset but the same number of models is not sufficient for DIAGONAL covariance type. One sloped ellipse distribution can be made up of several regular and upright ellipses. In conclusion, FULL covariance type fit the irregularly shaped dataset better and lead to faster convergence but it will cost more calculation time than DIAGONAL covariance type dose.

1.1.3. Result Evaluation

Finally, we find out the parameters with best performance lets see how good it is. We plot the confusion matrix of 10-cross-validation of 170 speakers, see the Figure.1.3 , only 4 samples of 1700 samples are misidentified.

Due to the Table.1.4, we found no gender misidentified(one gender misidentified to the other one) and our custom voice (last 2 speakers) are both identified correctly.In addition, We

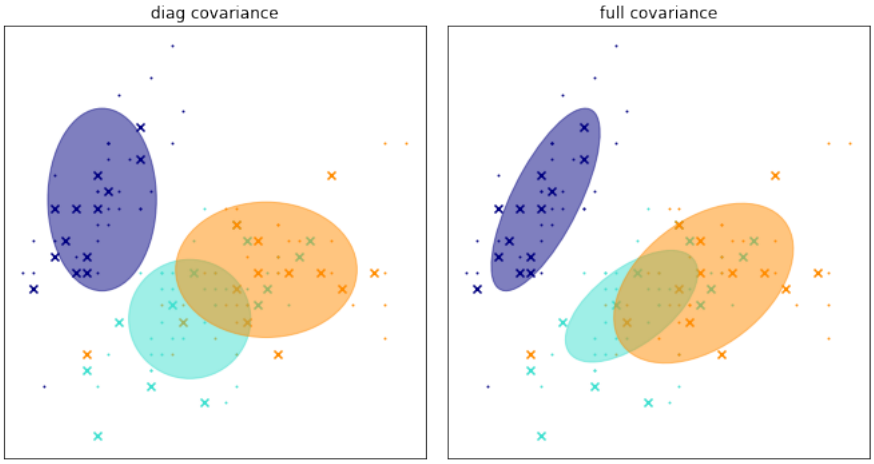

Figure 1.2.: GMM distribution Different Covariance Types

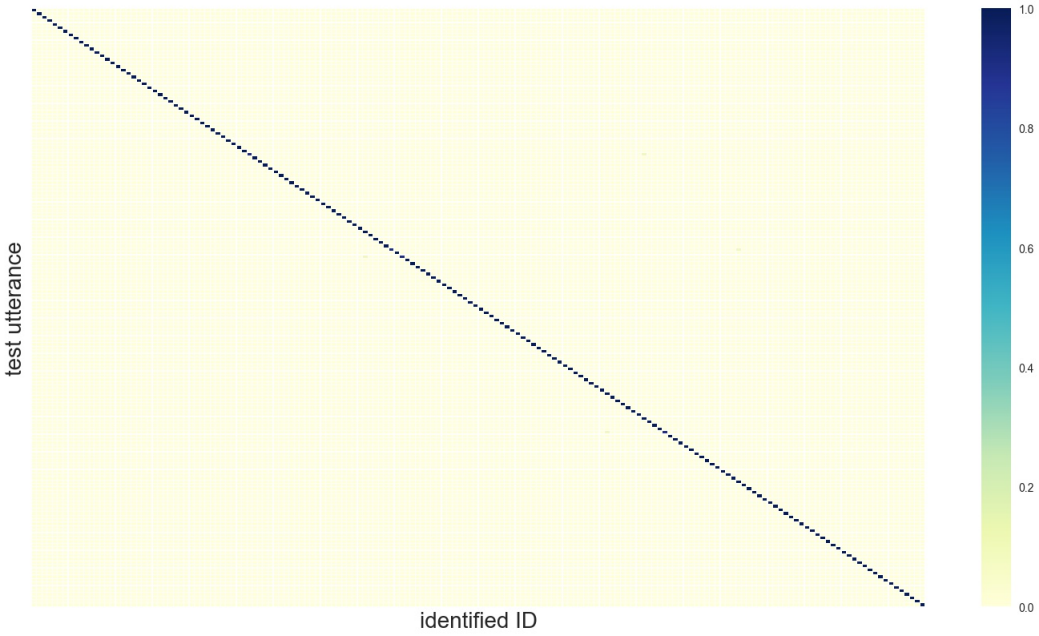

Figure 1.3.: Confusion Matrix Plot of 170 Speakers

Table 1.4.: Details of the misidentification

	True	False
Crossvalidation 4	mbwm0	mklt0
Crossvalidation 4	fjmg0	fadg0
Crossvalidation 5	mrrk0	mcrc0
Crossvalidation 9	fgjd0	fcau0

Table 1.5.: Configuration of the OSTI-SI Dataset

	Female	Male
Registered Speakers	64	106
Non-registered Speakers	66	124

observe from cross validations with different parameters that most misidentified speakers are male, so we suppose that low frequency voice (the voice from most male has lower frequency than from most female) is a little bit difficult to be identified correctly in our model, it can be improved in future work.

1.2. Enhancement

From the content of previous sections, we have achieved high accuracy of Speaker identification among known speakers or we could see speakers registered in our model. However this model is not realistic because if there is an unknown speaker, the model will pair the unknown speaker to one of registered ID and it is obviously incorrect. So unregistered speaker identification is always tough challenge of general Speaker Identification. In order to solve that, we used OSTI-SI (Open-set,Text-independent Speaker Identification), evaluated the error rate and finally optimal the performance of Identification model.

Before start, we selected first 3 files of TIMIT's Training set ($dr1, dr2, dr3$) as unknown speakers' voice set. Table.1.5 shows the people number, gender distribution of selected unknown speakers and in contrast with registered speakers.

The process of the open-set speaker Identification is shown in the Figure.1.4. The most important part is to set up threshold of comparison between registered and unregistered speaker. We introduced the ratio test between registered model and UBM's probability density function(PDF).

$$\frac{P(\underline{b}_{test}|\lambda)}{P(\underline{b}_{test}|\lambda_{UBM})} \underset{Unknown}{\overset{Known}{\gtrless}} \gamma \tag{1.1}$$

As the Equation above, if the expected PDF from the registered model is larger than threshold ratio multiply PDF from UBM model, the speaker is one of known speakers, otherwise not. The reason why choosing UBM model for testing is that UBM model is trained from a large amount of people, so it has more universality than other registered models. Unknown speaker could have higher log-pdf in UBM model. Then we have to set up some indexes in order to evaluate the ratio test. In general, there are 3 types of error will happen in our model:

- a test utterance from one of registered speaker misidentified to another registered speaker, referred to Mislabelling (ML)
- a test utterance from one of registered speaker misidentified to unknown speaker, referred to False Rejection (FR)

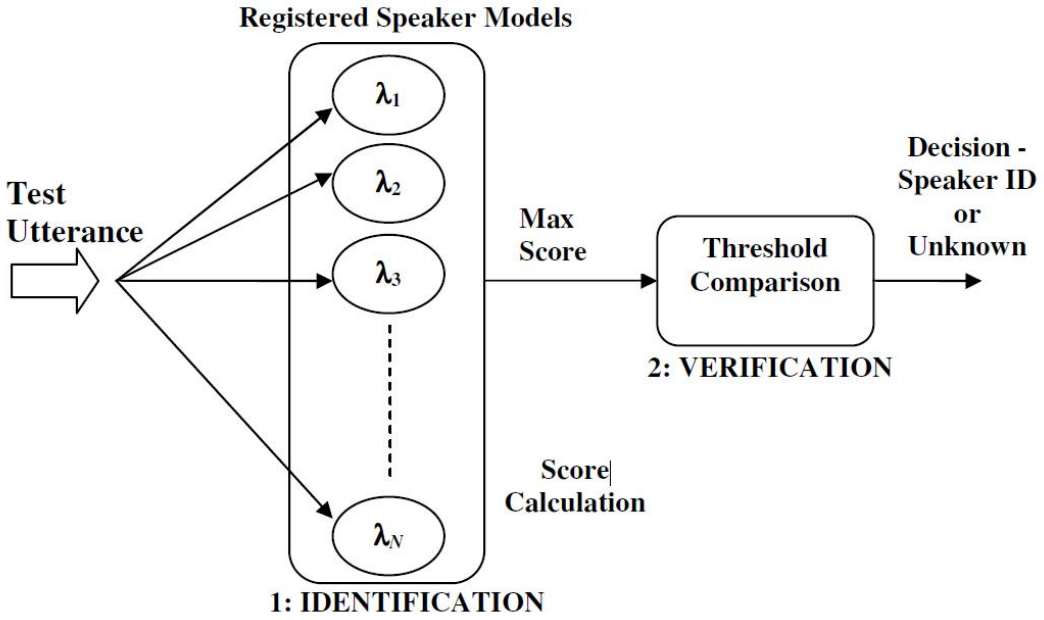

Figure 1.4.: Overview of the open-set, text-independent speaker identification process

- a test utterance from unknown speaker misidentified to one of registered speaker, referred to False Acceptance (FA)

So, the identification problem transfered into tradeoff problem between 3 types of error. In order to obtain the overall tradeoff performance, we set up Accumulative Error Rate (AER), it defines that

$$AER(\varsigma) = 100 * \frac{ML(\varsigma) + FR(\varsigma) + FA(\varsigma)}{T} \tag{1.2}$$

where ς is the threshold ratio and T is the total number of test voice set. Then we applied grid-search setting up several continuous threshold ratio and calculate the Expected ML rate, FR rate, FA rate and AER respectively through cross validation.

From the Figure.1.5, we got conclusion that with the threshold increasing, FA rate increases and FR rate drops. The best threshold is around 0.53 because it has minimum AER which is approximately 42%.

In future work , we think it is promising to replace standard UBM with UBM with score normalization or other form, setting up more benchmarks and find out the best method with lowest error rate.

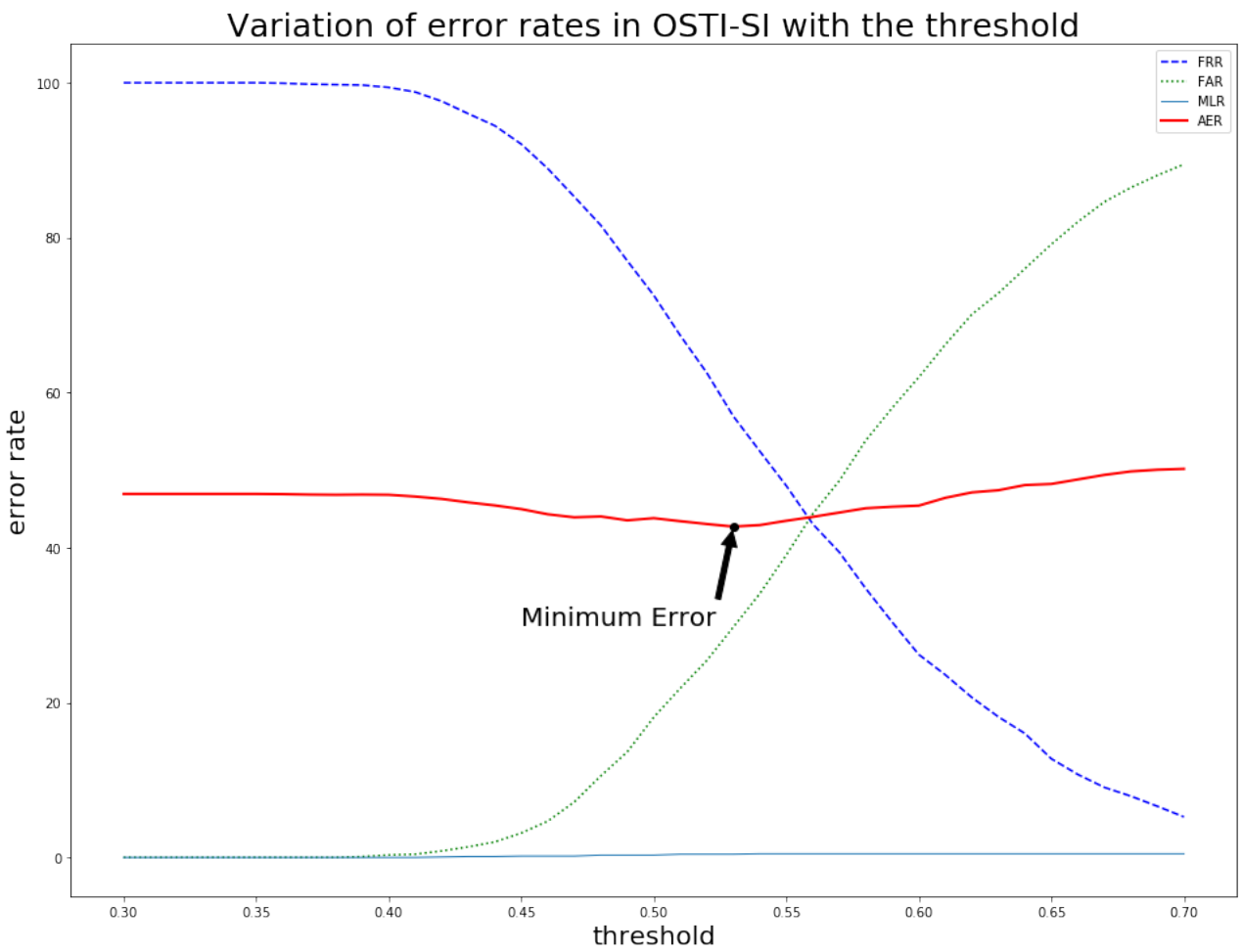

Figure 1.5.: Variation of error rates in OSTI-SI with threshold

A. Additionally

You may do an appendix

List of Figures

List of Tables

Bibliography

Declaration

Herewith, I declare that I have developed and written the enclosed thesis entirely by myself and that I have not used sources or means except those declared.

This thesis has not been submitted to any other authority to achieve an academic grading and has not been published elsewhere.

Stuttgart, TBD Date of sign.

Yuxin Liu, Shanqi Yang, Qianqian Wei