

FINAL PROJECT SPEAKER IDENTIFICATION

Speech and Audio Processing
Master in Multimedia and Communications
University Carlos III of Madrid

Students: Félix Jiménez Monzón (100290742)
Javier Carnerero Cano (100292885)

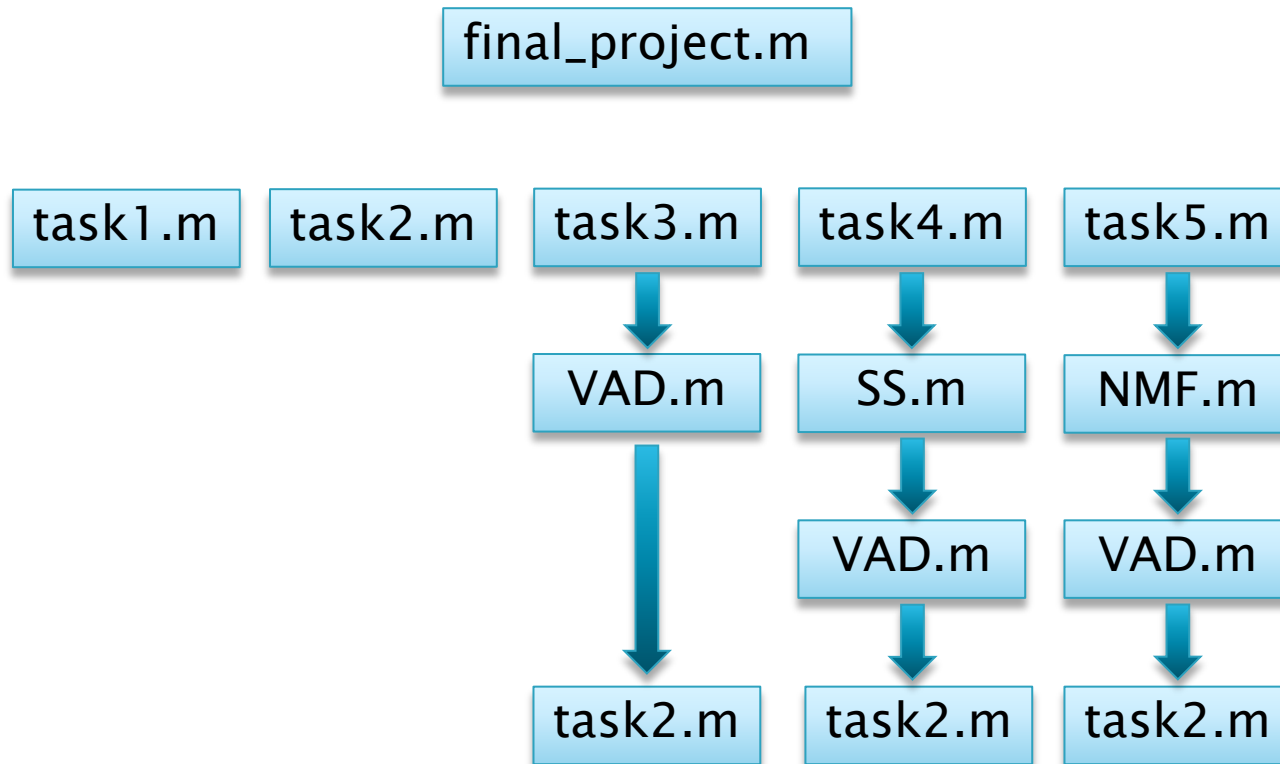


Objective

- ▶ The objective of this project is to implement a text-independent Speaker Identification (SI) system based on Gaussian Mixture Models (GMM). This system takes a speech utterance from an unknown speaker and provides the name or identification code of that speaker.

Methodology

- ▶ The main function is final_project.m
- ▶ Program scheme:

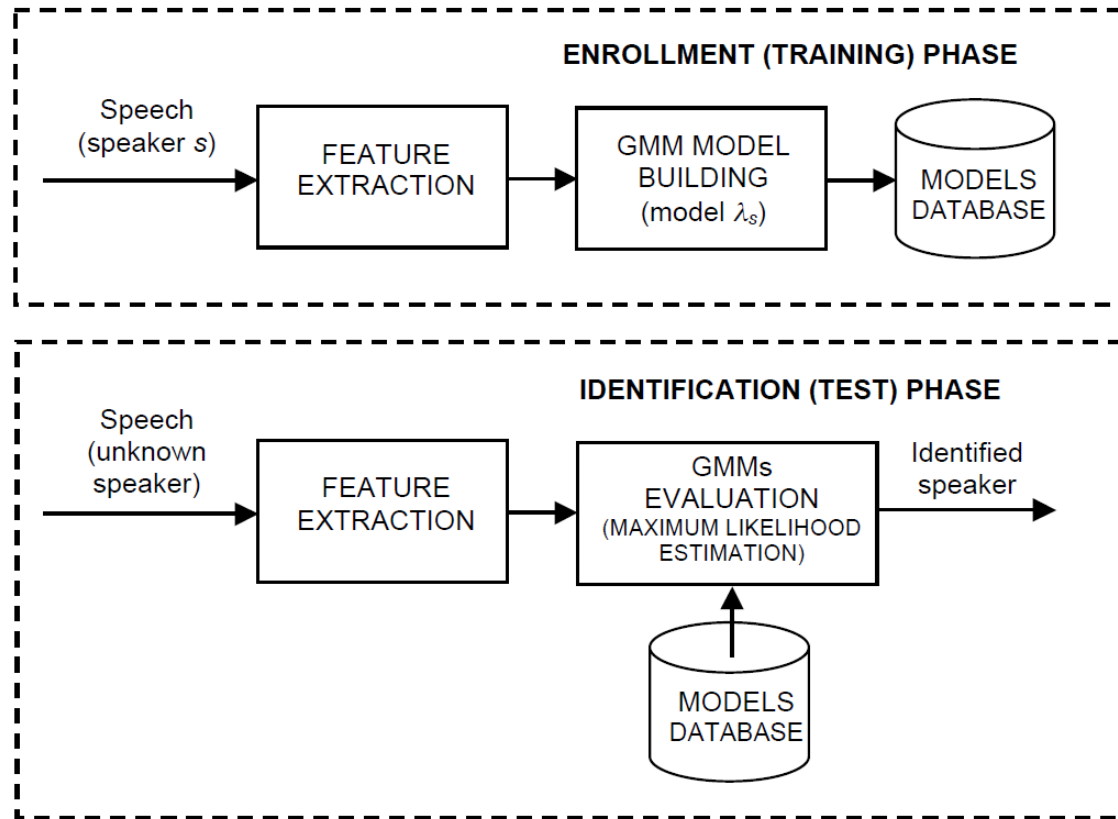


TASK 1

- ▶ The principal objective is to obtain the GMM model of the audio samples.
- ▶ We build one set for the training audios, and two sets for the test audios: clean and noisy.
- ▶ The main steps are:
 1. Feature extraction: we extract the MFCC coefficients for each train and test audio files.
 2. Speaker GMM models building: we generate the distribution model from the train coefficients for each speaker.
 3. Testing each of the test files with the models: we obtain the log-likelihoods of the test audio files for each speaker GMM model.
 4. Selection of the identified speaker: the speaker whose model produces the maximum log-likelihood is chosen.

TASK 1

- ▶ Train and test processes:



TASK 2

- ▶ Evaluation of the baseline system implemented in Task 1.
- ▶ The performance of the speaker identification system is measured in terms of the identification accuracy.
- ▶ The obtained identification accuracy is:
 - Clean audio: 99,375%
 - Noisy audio: 71,250%

TASK 3

- ▶ The objective of this task is to study the influence of the Voice Activity Detection (VAD) algorithm on the train and test speech files, in order to observe the performance of the whole system.
 - In our previous implementation, we filled with zeros at the end of the vector output, in order to obtain a number of samples that is multiple of the period frame.
 - Thus, we had to delete the zero-elements in order to obtain the MFCC coefficients.
- ▶ The obtained identification accuracy is:

• Clean audio:	• Noisy audio:
99,375%	73,750%

TASK 4

- ▶ The objective of this task is to study the influence of the Spectral Subtraction algorithm (SS) on the performance of the whole system.
 - On our previous implementation, we filled with zeros at the end of the vector output, in order to obtain a number of samples that is multiple of the period frame.
- ▶ First, we apply the VAD algorithm and after that we apply the SS algorithm.
- ▶ The obtained identification accuracy is:
 - Clean audio: 98,750%
 - Noisy audio: 83,750%

TASK 5

- ▶ We propose the use of the NMF algorithm for improving the stages of the developed speaker identification system.
- ▶ Two stages: training + denoising.
- ▶ The training steps are:
 1. To divide both, each train and noise ('factory.wav') audio files, into overlapped frames and to apply the FFT to each frame in order to create the \mathbf{V}_s and \mathbf{V}_n matrices.
 2. To apply the NMF algorithm both to \mathbf{V}_s and \mathbf{V}_n in order to obtain the \mathbf{W}_s and \mathbf{W}_n matrices.
 3. To concatenate the \mathbf{W} matrices and to obtain the \mathbf{W}_r matrix.

TASK 5

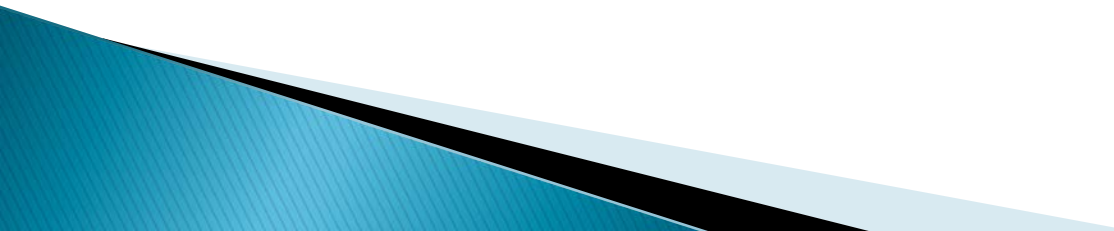
- ▶ The denoising steps are:
 1. To divide each test audio file into overlapped frames and to apply the FFT to each frames in order to create the \mathbf{V}_r matrix.
 2. To apply the NMF algorithm to \mathbf{V}_r while keeping constant the \mathbf{W}_r matrix obtained in the train stage. This way, we compute the \mathbf{H}_r matrix.
 3. To take the $r=64$ first rows from \mathbf{H}_r and to create the \mathbf{H}_s matrix.
 4. To obtain the \mathbf{V}'_r matrix by multiplying the \mathbf{W}_s matrix (computed in the training stage) and the \mathbf{H}_s matrix.
 5. Finally, to determine the denoised audios by applying the IFFT to \mathbf{V}'_r (multiplied by its phase), and then by overlapping the last one.
- ▶ The obtained identification accuracy is:
 - Clean audio: 98,750%
 - Noisy audio: 91,250%

Identification Accuracy comparison

Identification Accuracy (%):		
	clean	noisy
Base	99.375000	71.250000
VAD	99.375000	73.750000
SS	98.750000	83.750000
NMF	98.750000	91.250000

- ▶ As we can see, the best result for noisy audio is obtained with the NMF algorithm.

Difficulties

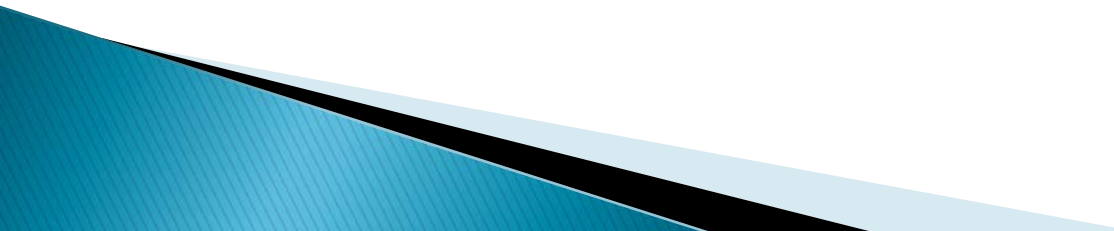
- ▶ We had to change the 'load_train_data.m' function into 'load_test_data.m' for loading the test files, because we had to arrange each test audio file independently, in order to generate the MFCC coefficients.
 - ▶ We have observed that the final identification accuracy is a little bit different for each program execution. That may be caused by the randomness of the Gaussian model generation. Nevertheless, this change is not critical for the general performance.
- 

Conclusions

- ▶ When we apply the three different algorithms to the **clean** audio files, we can see that the identification accuracy is similar or even decreases slightly. This is due to the fact that those algorithms are oriented to reduce the noise.
- ▶ When we apply the three different algorithms to the **noisy** audio files, we can see that the identification accuracy increases as follows:

Baseline < VAD < SS+VAD < NMF+VAD

Conclusions

- ▶ As we can see, we obtain good results. We can also check, when we listen to the denoised audio files, that the sound quality is better.
 - ▶ As a future work line, it could be interesting to optimize the NMF and SS performance (trade-off) in order to generalize them to possible clean files.
- 

References

- ▶ “An overview of text-independent speaker recognition: From features to supervectors”, T. Kinnunen, H. Li, Speech Communication, vol. 52, pp. 12–40, 2010. URL: <http://www.sciencedirect.com/science/article/pii/S0167639309001289>.
- ▶ “Reduction of Non-stationary Noise for a Robotic Living Assistant using Sparse Non-negative Matrix Factorization”, B. Cauchi, S. Goetze and S. Doclo, Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments (SMIAE '12), pp. 28–33, 2012. URL: <http://dl.acm.org/citation.cfm?id=2392871&picked=formats&prelayout=tabs>
- ▶ “Speech denoising using nonnegative matrix factorization with priors”, K. W. Wilson, B. Raj, P. Smaragdis and A. Divakaran, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008). URL: <http://web.engr.illinois.edu/~paris/pubs/wilson-icassp08.pdf>
- ▶ “Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition”, C. Joder and B. Schuller, 10 ITG Symposium Speech Communication, Braunschweig, Germany, pp. 1–4, 2012. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6309609>