Speaker Recognition
Using the GMM-UBM Framework in Python

João António Fernandes da Costa - bio12046@fe.up.pt

June 25, 2018

Laboratory Projects
Master in Bioengineering - Biomedical Engineering
Faculty of Engineering of University of Porto

Introduction

Voice

Different physiology of larynx and vocal cords,

Voice

Speech mannerisms, rhythm, intonation, vocabulary,...

Voice

Our voice makes us unique.

Speaker Recognition

- Naïve Speaker Recognition

Speaker Recognition

- Naïve Speaker Recognition
- Forensic Speaker Recognition

Speaker Recognition

- Naïve Speaker Recognition
- Forensic Speaker Recognition
- Automatic Speaker Recognition

Speaker Recognition

- Naïve Speaker Recognition
- Forensic Speaker Recognition
- Automatic Speaker Recognition $\Rightarrow$ Biometric Recognition

Speech Machine Learning Tasks

Overview

When are people speaking?

Who is speaking?

Can it be an imitation of said person?

Who is speaking when?

What is the person saying?

What are the main challenges when dealing with speech signals?

Speaker Recognition

Speaker Identification

Identify the speaker from an unknown utterance from a set of known speakers

Multi-class classification task

Speaker Verification

Determine the validity of an identity claim from an unknown utterance

Binary classification task

Text dependent or text independent

Closed or open set

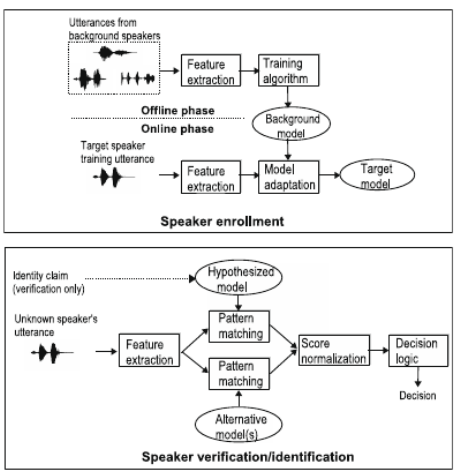

Figure 1: Speaker Recognition System[1]

Speech-Related Challenges

Voice, unlike other biometric factors, is highly variable and dependent on several internal and external factors.

Speech-Related Challenges

Voice, unlike other biometric factors, is highly variable and dependent on several internal and external factors.

Variability sources:

Speaker	stress, emotion, disease, intraspeaker variability,...
Conversation	dialogues, monologues, written text,...
Technology	quality of stream and recording devices, background noise,...

Speech-Related Challenges

Voice, unlike other biometric factors, is highly variable and dependent on several internal and external factors.

Variability sources:

Speaker	stress, emotion, disease, intraspeaker variability,...
Conversation	dialogues, monologues, written text,...
Technology	quality of stream and recording devices, background noise,...

Some can be mitigated with adequate mathematical modeling.

Methods

Overview

Text-independent, closed set Speaker Identification System implemented in Python using:

- VoxCeleb Dataset,
- Mel-Frequency Cepstrum Coefficients (MFCC) features,
- Gaussian Mixture Models with Universal Background Model (GMM-UBM) modeling with Maximum *a posteriori* (MAP) adaptation.

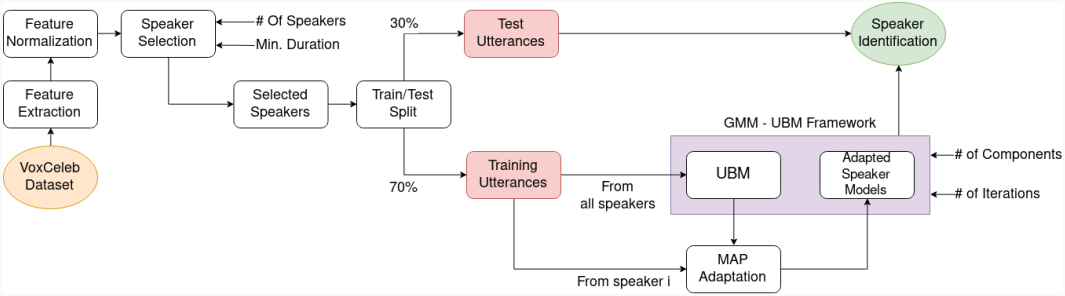

Figure 2: Implemented Speaker Identification Pipeline

VoxCeleb Dataset [2]

Dataset collected from YouTube videos of 1251 People of Interest (POI).

Unconstrained settings: interviews, speeches, background noise, overlapping speakers,...

Average of 116 utterances per user, with an average of 8 seconds of duration.

Feature Extraction - MFCCs [4]

Short-term Spectral Features used in Speech and Speaker Recognition systems.

Implemented in SpeechPy module [3].

Algorithm:

1. Pre-emphasis
2. Windowing
3. Discrete Fourier Transform
4. Mel-scale Log Filterbank Energy
5. Discrete Cosine Transform

Pre-emphasis

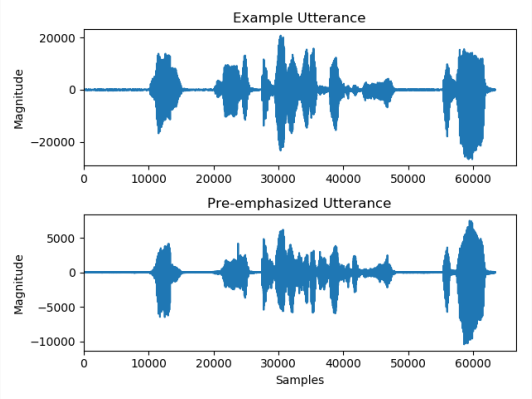

Figure 3: Example of utterance before and after processing

Utterance by Neil deGrasse Tyson

"Then...there's a cause and effect here, about..."

Audio vector is pre-emphasized with a first order high-pass filter:

- Removes DC component,
- Relatively increases the magnitude of higher frequencies (these contain a considerable amount of speech information).

Windowing

0.02s, non-overlapping, rectangular windows are applied to the audio signal.

Decomposition of the signal into short time frames.

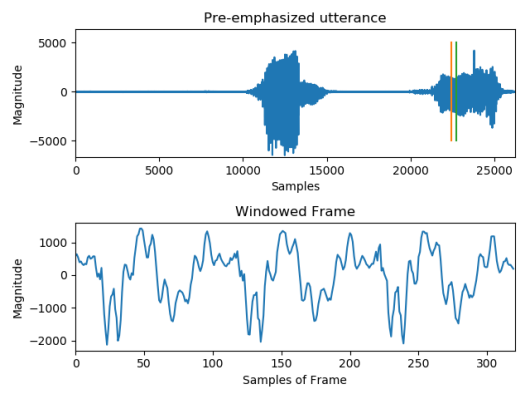

Figure 4: Example of frame extraction from pre-emphasized signal. The top signal is zoomed in to the ”Then... there’s” section of speech

Discrete Fourier Transform

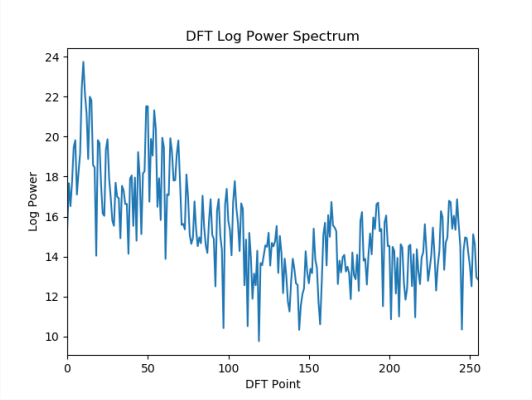

Figure 5: Log Power Spectrum of example frame obtained from DFT

DFT of size 512 is applied to each frame (zero-padded).

Power spectrum of DFT is considered, $E_k = X_k \cdot X_k^*$.

Mel-scale

The mel-scale is a psychoacoustic perceptually uniform frequency space [5].

Reference point: 1000Hz = 1000 mel.

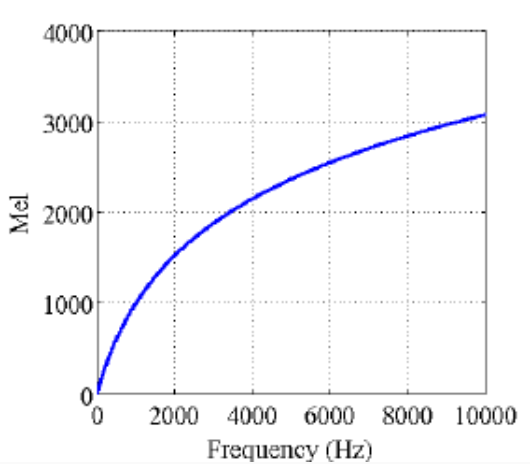

Figure 6: Pitch in Mels as a function of frequency

Log Mel-scale Filterbank Energy

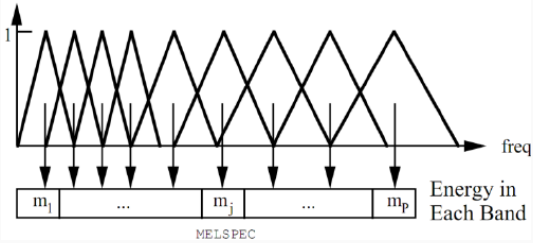

Figure 7: Schematic of Mel-scaled filterbank and energy output

Filterbank of 40 triangular overlapping filters with linearly spaced central frequencies in the mel scale.

Logarithm of filter responses is computed, resulting in log-energy output of the filterbank.

Discrete Cosine Transform

DCT (real-valued DFT) is applied to the log-energy output of the filterbanks.

13 lowest coefficients are kept. ⇒ MFCCs

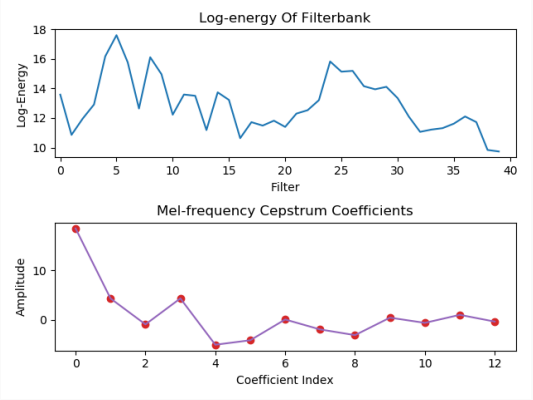

Figure 8: Log-energy filterbank response and extracted MFCC for the considered frame

ΔMFCCs and ΔΔMFCCs

A speaker utterance contains hundreds of frames and corresponding MFCCs.

ΔMFCCs (1st derivative) and ΔΔMFCCs (2nd derivative) are used to capture the dynamics of spectro-temporal changes.

13 MFCC + 13 ΔMFCC + 13 ΔΔMFCC = 39 features per frame

Global Cepstral Mean and Variance Normalization

Feature Normalization mitigates noise effects and improves general performance of speaker recognition systems.

Global Cepstral Mean and Variance Normalization

Utterance-level feature mean and variance is calculated.

Feature vectors are normalized to $\mu = 0, \sigma = 1$.

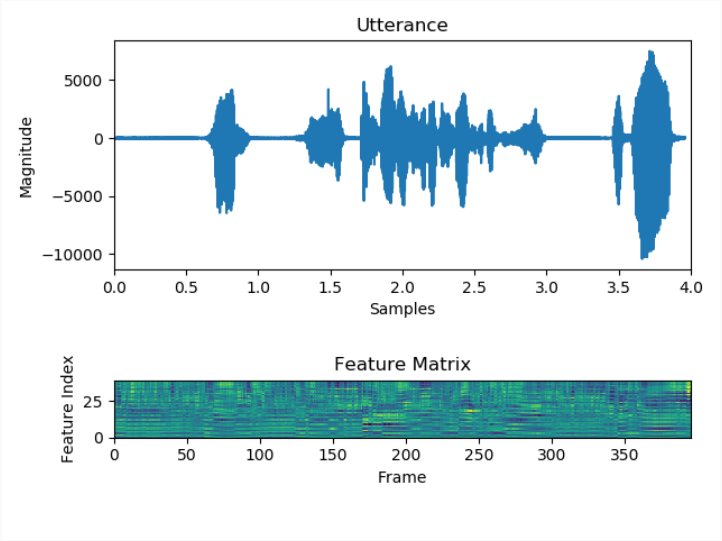

Figure 9: Utterance and resulting collection of normalized feature vectors (vertical). 0-12: MFCCs, 13-26: ΔMFCCs, 27-39:ΔΔMFCCs

Speaker Selection

Table 1: Selected speakers and total duration of utterances in minutes

Adrianne Curry	5.18
Alex Pettyfer	5.09
Andrew Dice Clay	5.19
Caroline Rhea	5.20
Hye-kyo Song	5.21
Joan Cusack	5.01
Keeley Hawes	5.01
Lacey Turner	5.09
Simon Baker	5.20
Thomas Jane	5.19

10 speakers were selected

Minimum of 5 minutes of summed utterance durations.

Train/Test Split

Splitting was done on an utterance level:

70% of a speaker's utterances are used for training, the rest are reserved for testing.

Table 2: Statistics for Train and Test Data per Speaker

	Avg. # of utt.	Avg utt. length (s)	Min. utt. length (s)	Max. utt. length (s)
Train	37	12.4	7.9	35.3
Test	13	12.9	7.9	40.2

Speaker Modeling

Natural variability of speech in a single utterance can lead to poor results.

Necessity to create mathematical models:

- Describe in general the feature characteristics of a selected speaker,
- Compensate for intra-speaker and intra-utterance variability.

Speaker Models

- Gaussian Mixture Models (GMM),
- GMM supervectors,
- Joint Factor Analysis,
- i-vectors,
- Deep Learning approaches.

Gaussian Mixture Models - GMM

Standard of speaker modeling for speaker recognition. Gaussian Mixture Models (GMM) are probabilistic models that assume data points are generated from a finite sum of Gaussian distributions.
GMM implemented in *scikit-learn* in the *GaussianMixtureModel* class.

Gaussian Mixture Models - GMM

Standard of speaker modeling for speaker recognition. Gaussian Mixture Models (GMM) are probabilistic models that assume data points are generated from a finite sum of Gaussian distributions. GMM implemented in *scikit-learn* in the *GaussianMixtureModel* class.

Probability density function of GMM λ:

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{C} \omega_k \cdot \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \quad (1)$$

$\mathbf{x}$ - F-dimensional feature vector;
C - total number of components of the GMM;
ω_k - mixing weight of k-th Gaussian component;
$\sum_{k=1}^{C} \omega_k = 1$;
$\mathcal{N}(\mu_k, \Sigma_k)$ - k-th Gaussian Component, with means μ_k and covariance matrix Σ_k;
$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ - probability of $\mathbf{x}$ being generated by component k.

Universal Background Model - UBM

Speech applications frequently require a model that translates the speaker-independent variability.

World or Universal Background Model:

- Out-of-set systems require a background model to determine if an utterance is from a speaker present in the training group,
- Creation of speaker-specific models from the adaptation of UBM parameters can help reduce variability due to intra-speaker or inter-channel variability.

UBM Training

All training utterances from all speakers are used to model the UBM (speaker-independent) with the corresponding feature vectors:

$$[\text{MFCCs}, \Delta\text{MFCCs}, \Delta\Delta\text{MFCCs}].$$

The number of components (C) and training iterations (I) is varied to evaluate the influence of these parameters:

C	1, 4, 16, 64, 256
I	5, 10, 20

Covariance matrices are considered diagonal for computational efficiency.
Training is performed using the Expectation-Maximization (EM) algorithm.

Maximum *a posteriori* (MAP) Adaptation

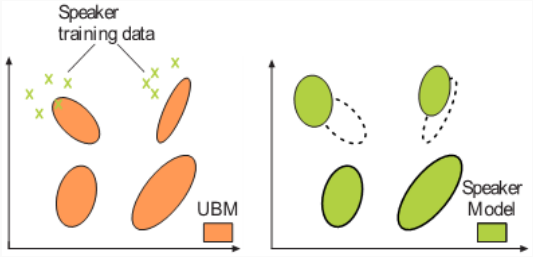

Figure 10: MAP adaptation of a UBM to create a speaker model [6].

The UBM is adapted during speaker enrollment with MAP adaptation.

Given a trained UBM and a set of speaker-specific training data, sufficient statistics are calculated to adjust model parameters (weights, means and covariances), creating a speaker-specific adapted model.

Log-likelihood of an Unknown Utterance

Given an unknown speaker utterance X, consisting of T feature vectors (T frames) $X = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, the likelihood of observing X with GMM λ is:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t|\lambda) \Leftrightarrow \log p(X|\lambda) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda) \tag{2}$$

The log-likelihood ratio score of X being produced by speaker S is:

$$\Lambda(X|S) = \log p(X|\lambda_S) - \log P(X|\lambda_{UBM}) \tag{3}$$

The predicted speaker S_p in a set of N speakers will have $\max \{\Lambda(X|S_i)\}_{i=1}^{N}$

Results and Discussion

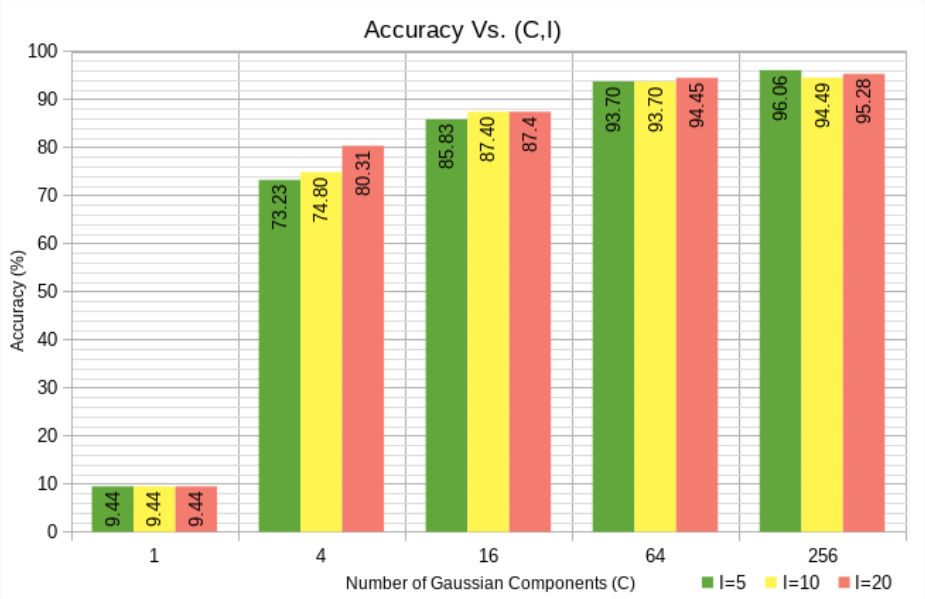

Figure 11: Accuracy results obtained for multi-class classification of test utterances

Accuracy Results

- Similar accuracy scores are obtained comparing to a CNN framework in [2], but the CNN system is trained for the full VoxCeleb dataset.

Accuracy Results

- Similar accuracy scores are obtained comparing to a CNN framework in [2], but the CNN system is trained for the full VoxCeleb dataset.
- Small number of components is not sufficient to model variability in training data.

Accuracy Results

- Similar accuracy scores are obtained comparing to a CNN framework in [2], but the CNN system is trained for the full VoxCeleb dataset.
- Small number of components is not sufficient to model variability in training data.
- Performance reaches a plateau for large numbers of C: further increase of C results in the addition of redundant or irrelevant components.

Accuracy Results

- Similar accuracy scores are obtained comparing to a CNN framework in [2], but the CNN system is trained for the full VoxCeleb dataset.
- Small number of components is not sufficient to model variability in training data.
- Performance reaches a plateau for large numbers of C: further increase of C results in the addition of redundant or irrelevant components.
- GMM-UBM systems usually use 1024 or 2048 components, but 64 components were sufficient to capture 10 speakers variability.

Accuracy Results

- Similar accuracy scores are obtained comparing to a CNN framework in [2], but the CNN system is trained for the full VoxCeleb dataset.
- Small number of components is not sufficient to model variability in training data.
- Performance reaches a plateau for large numbers of C: further increase of C results in the addition of redundant or irrelevant components.
- GMM-UBM systems usually use 1024 or 2048 components, but 64 components were sufficient to capture 10 speakers variability.
- Increasing the number of iterations did not cause substantial increase of accuracy: fully converged models are not necessary to model variability.

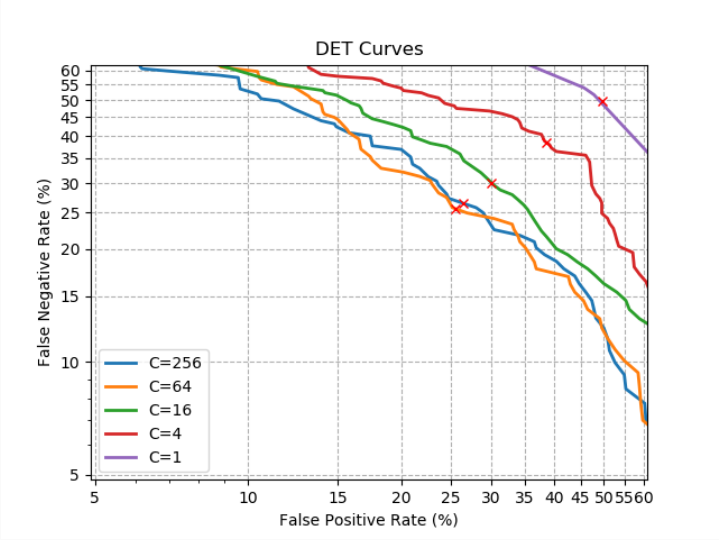

Figure 12: DET curves for GMM-UBM framework parameter pairs (C,I=10). Equal Error Rate points are marked with a red cross.

DET Curve Results

- Increasing the number of components increases the overall performance shown by the system's DET curve.

DET Curve Results

- Increasing the number of components increases the overall performance shown by the system's DET curve.
- The best Equal Error Rates (EER) obtained were of 26%. [2] reached EER of 10% with the full dataset.

Conclusions and Future Work

A GMM-UBM framework was implemented in Python, achieving accuracies of 96% using a 10 speaker subset of the VoxCeleb database, showing the capabilities of modeling of the GMM-UBM framework and the descriptive capabilities of MFCCs.

To achieve more comparable results with the VoxCeleb database, a full-set model must be created and evaluated.

Speaker verification and out-of-set testing can be performed in the future with the same framework to evaluate its capability with outsiders.

More recent speaker modeling techniques can be applied, such as GMM supervectors and i-vectors, to better illustrate the evolution of speaker recognition systems.

Thank you for your attention!

Questions?

References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010, ISSN: 01676393. DOI: 10.1016/j.specom.2009.08.009. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2009.08.009.

References ii

[2] A. Nagraniy, J. S. Chungy, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 2616–2620, 2017, ISSN: 19909772. DOI: 10.21437/Interspeech.2017-950.

[3] A. Torfi, *SpeechPy: Speech recognition and feature extraction*, Aug. 2017. DOI: 10.5281/zenodo.840395. [Online]. Available: https://doi.org/10.5281/zenodo.840395.

References iii

[4] M. Sahidullah and G. Saha, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,” *Speech Communication*, vol. 54, no. 4, pp. 543–565, May 2012, ISSN: 01676393. DOI: 10.1016/j.specom.2011.11.004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639311001622.

[5] S. S. Stevens, J. Volkmann, and E. B. Newman, “A Scale for the Measurement of the Psychological Magnitude Pitch,” *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, Jan. 1937, ISSN: 0001-4966. DOI: 10.1121/1.1915893. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1915893.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing: A Review Journal*, vol. 10, no. 1, pp. 19–41, 2000, ISSN: 10512004. DOI: 10.1006/dspr.1999.0361.