

Soutenance de thèse

Investigating the speaker information carried on by the fundamental frequency

Presented by:
Shalini Priya

Supervisors:
Pierre CHAMPION
Denis JOUVET
Hubert NOURTEL

September 6th, 2022

Contents

1. Introduction
2. Speaker Recognition System
3. Model Architecture
4. Experiments and Results
5. Conclusion and Future Work

Contents

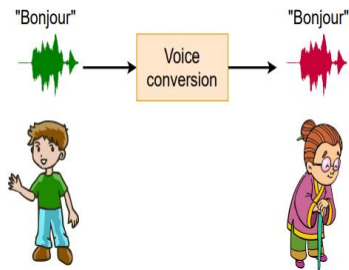
- 1 Introduction
- 2 Speaker Recognition System
- 3 Model Architecture
- 4 Experiments and Results
- 5 Conclusion and Future Work

Sensitivity of speech data

- spoken words
- the speaker's identity
- speaker attributes (age, gender, accent, etc)
- emotions
- etc.

Context

- Widespread use of speech data in Speech technologies
- **ANR DEEP- PRIVACY** - aims to **anonymize** speech data for easier sharing
- **Example - Voice Conversion System** - modifies a speech signal to sound as if pronounced by another speaker



Objective

We aim to find the answer of the following question in this thesis:

Is speaker information contained in the fundamental frequency?

Speech Production Mechanism

- Sound production requires two things:
 - power / energy source - **airflow from the lungs**
 - vibrating element - **vocal folds**
- It is the activity of the vocal folds that determines the state as **"voiced"**, **"unvoiced"** and **"breathing"**

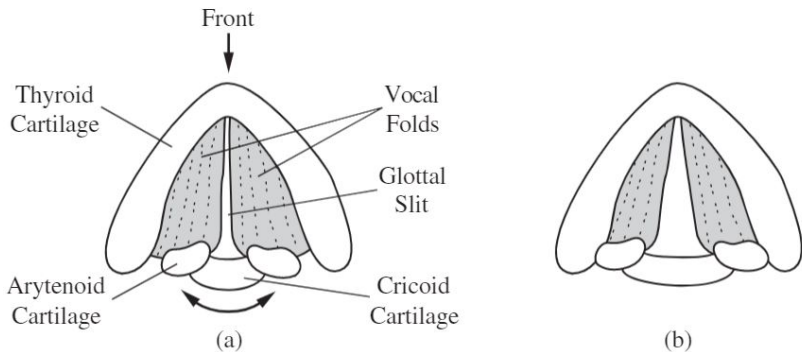
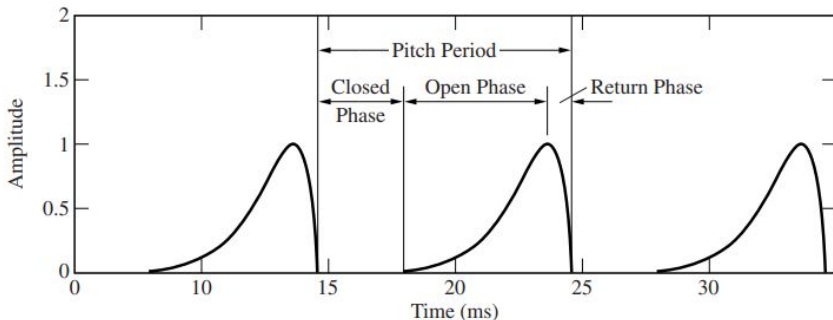


Figure shows the downward looking view of human larynx (a) voiced (b) breathing

Fundamental Frequency

- If one were to measure the airflow velocity as a function of time, obtained waveform will be similar to the figure:

- **Closed phase** - folds are closed and no airflow occurs
- **Open phase** - folds are open and the flow increases upto the maximum
- **Return phase** - time interval from the maximum air flow until the glottal closure



- Time duration of one glottal cycle is referred to as **Pitch Period**
- Reciprocal of pitch period is referred to as **Pitch**, also called as **Fundamental Frequency (F0)**

Fundamental Frequency (cont.)

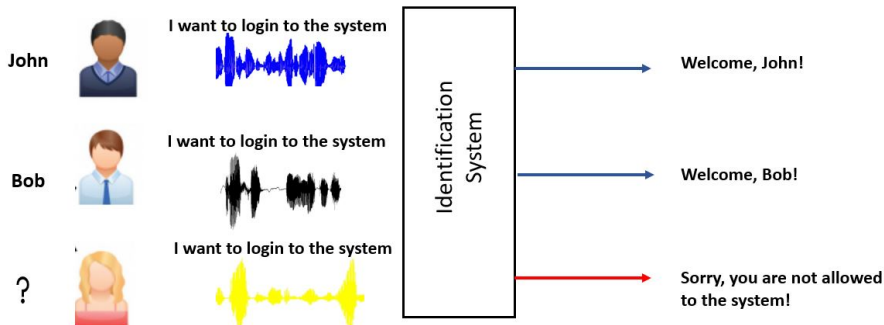
- F0 is determined by multiple factors:
 - It increases with vocal folds length
 - F0 inversely varies with mass and thickness of the vibrating part of the vocal folds
 - Tension in the vocal folds

Contents

- 1 Introduction
- 2 Speaker Recognition System
- 3 Model Architecture
- 4 Experiments and Results
- 5 Conclusion and Future Work

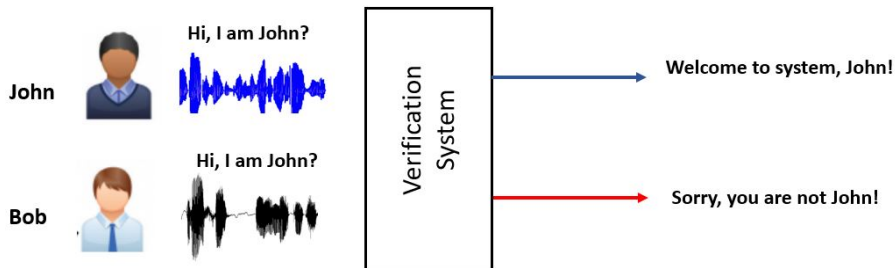
Speaker Recognition System

- **Speaker Identification** - “Which speaker (out of a known group) is this?”



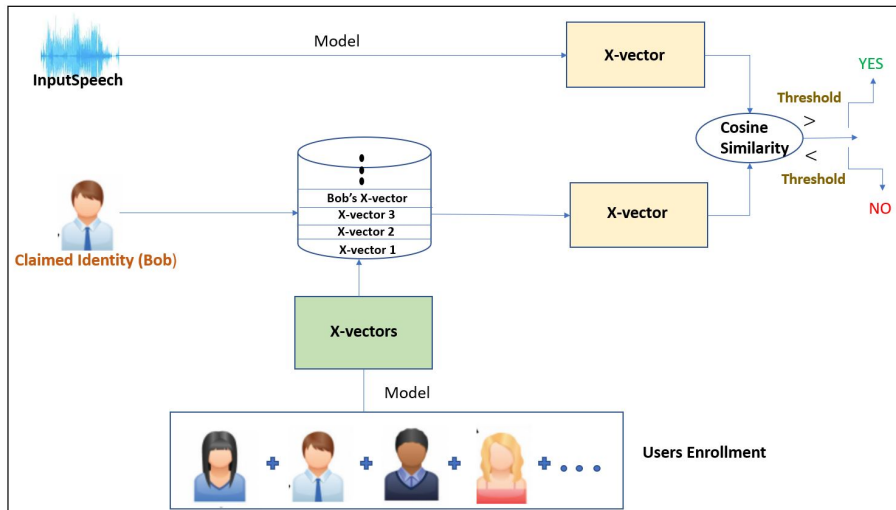
Speaker Recognition System

- **Speaker Verification** - “Is the speaker who they claim to be?”



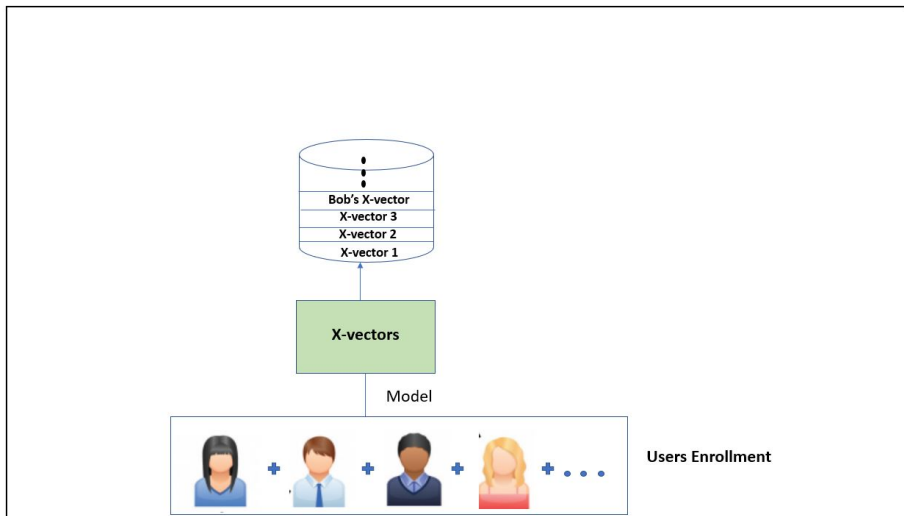
Applications - Call centres, banking sectors, authentication of smart devices etc,

Speaker Verification Pipeline



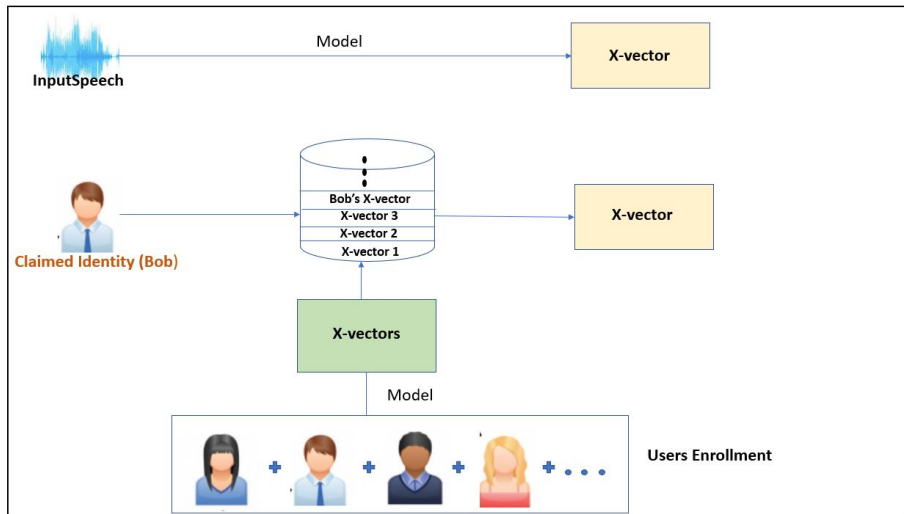
- **Toolkit** - SIDEKIT (Speaker IDEentification toolKIT)
- **Evaluation metrics** - Equal Error Rate

Speaker Verification Pipeline



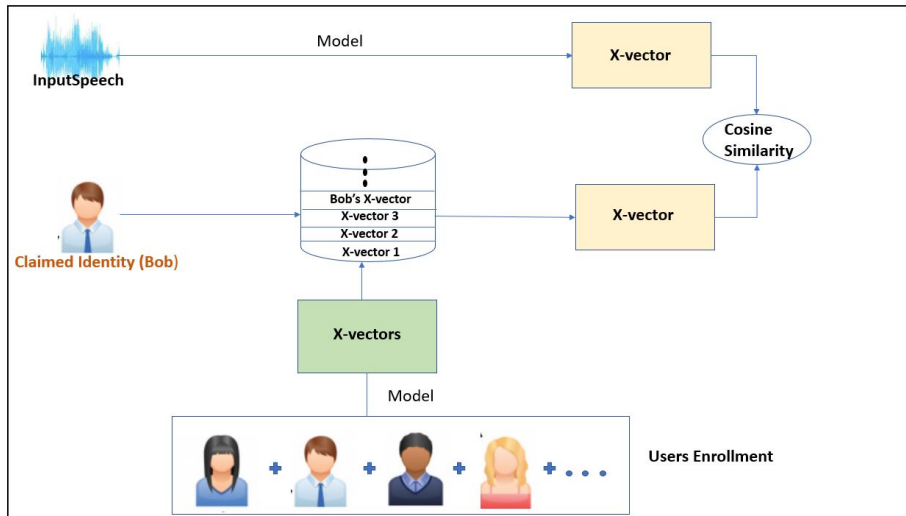
- Toolkit - SIDEKIT (Speaker IDEentification toolKIT)
- Evaluation metrics - Equal Error Rate

Speaker Verification Pipeline



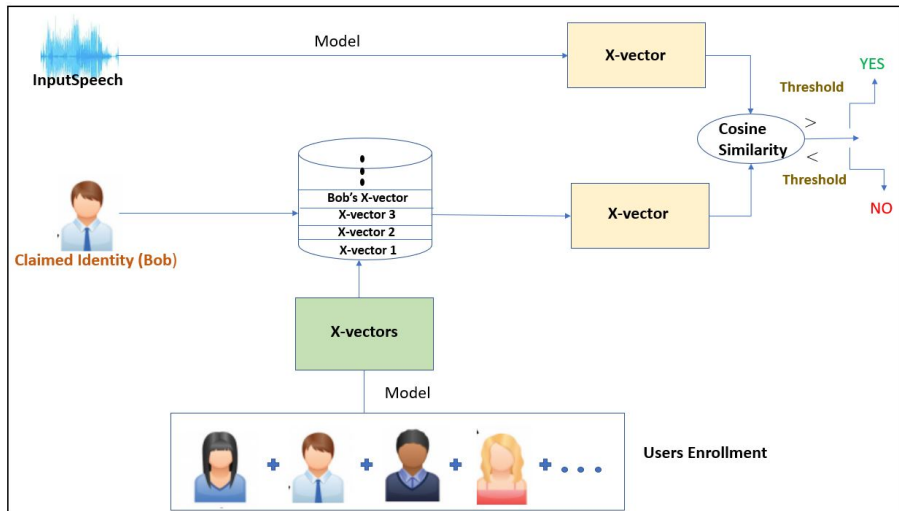
- **Toolkit** - SIDEKIT (Speaker IDEentification toolKIT)
- **Evaluation metrics** - Equal Error Rate

Speaker Verification Pipeline



- **Toolkit** - SIDEKIT (Speaker IDEntification toolKIT)
- **Evaluation metrics** - Equal Error Rate

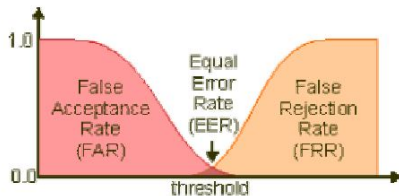
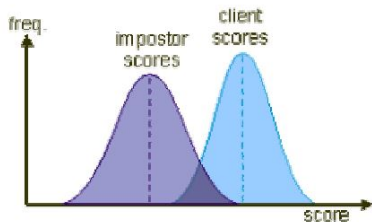
Speaker Verification Pipeline



- **Toolkit** - SIDEKIT (Speaker IDEentification toolKIT)
- **Evaluation metrics** - Equal Error Rate

Equal Error Rate (EER)

- Test data consists of both imposter and genuine speakers
- Scores are normalized and decision is made.
- Two types of errors occur, **False Acceptance Rate** (falsely accepts an utterance that is not from the claimed speaker) and **False Rejection Rate** (falsely rejects an utterance from a claimed speaker)



Feature Extraction

It is a process to extract speaker personal feature traits. It transforms speech to a set of feature vectors with reduced dimensions, to enhance speaker specific information and to suppress redundant information

■ Choice of Methodology

- Mel Frequency Cepstrum Coefficients (MFCC)
- Fundamental Frequency (F0)

MFCC (Mel Frequency Cepstral Coefficients)

Most of the Speaker Verification System uses MFCC as the feature vectors because of its performance.

- **80** features are generated for each frame

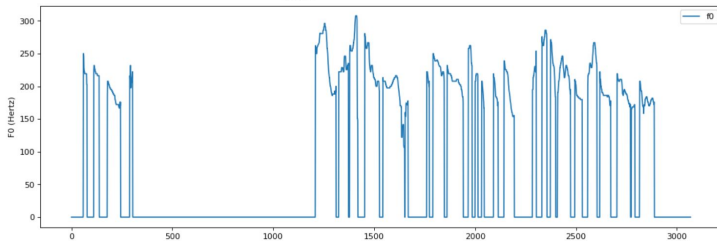
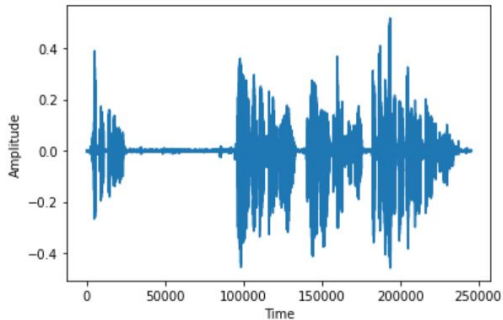
Fundamental Frequency

- Contains both voiced and unvoiced frames
- Unvoiced or silence value is 0 while voiced has some value
- Measured in Hertz
- One feature per frame as an input

Package - YAAPT

Fundamental Frequency

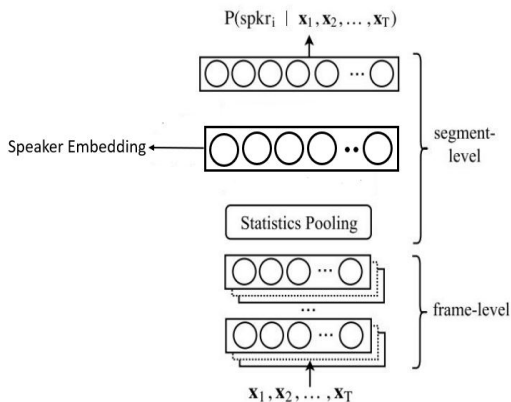
Sound Wave



Contents

- 1 Introduction
- 2 Speaker Recognition System
- 3 Model Architecture**
- 4 Experiments and Results
- 5 Conclusion and Future Work

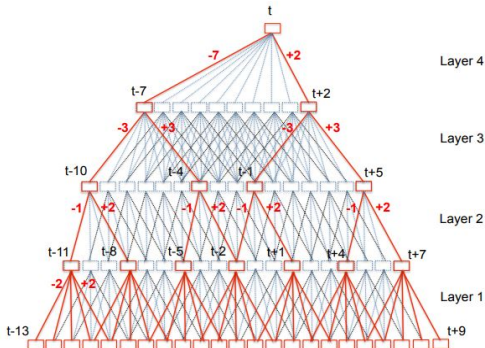
Network Architecture



- The input represented here $x_1, x_2, x_3, \dots, x_T$ is the sequence of frame MFCC vectors, or frame F0 values
- First two layers in the figure extracts frame level characteristics from utterances
- Segment level comprises of Statistic pooling layer and Softmax layer

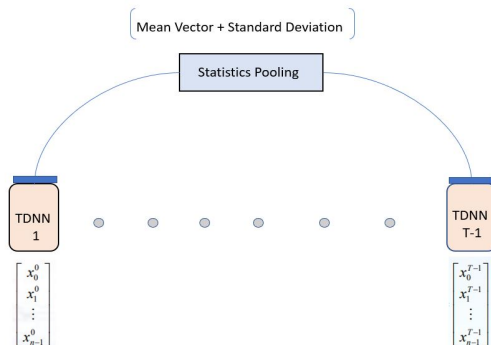
Time Delay Neural Network (TDNN) - Frame level

- ▶ Time Delay Neural Network is based on **Context Windowing** and **Sub-Sampling**
- ▶ Five TDNN layers are used in sequence followed by a ReLU and Batch Normalization layer.



Statistic Pooling -Segment level

- **Mean** (aggregates the frame level features)
- **Standard Deviation** (helps in capturing temporal variability over long contexts)



Softmax/Cross-Entropy - Segment level

- Network is trained using cross-entropy
- Softmax layer is discarded and the layer before that, is used as the speaker representation (X-vector)

Contents

- 1 Introduction
- 2 Speaker Recognition System
- 3 Model Architecture
- 4 Experiments and Results**
- 5 Conclusion and Future Work

LibriSpeech Dataset

Collection of read English speech corpus derived from audiobooks

- 921 speakers (482 Male and 439 Female) **Training**
- 40 speakers (20 Male and 20 Female) **Testing**

Experiment One

- We have used segment duration of 3 seconds for both features (MFCC and F0)
- We want to achieve EER as low as 0
- According to the gender, Male, Female and Both were studied independently

Experiment One

1. Speaker Verification for MFCC and F0 features

Feature	Both EER	Male EER	Female EER
MFCC	4.3±0.5	1.1±0.5	5.6 ±1.0
F0	20.0±1.2	17.8±1.7	21.8 ±1.7

- EER of F0 is higher than MFCC
- Female EER is higher than Male (also seen in the state of the art for the same dataset)

Experiment Two

2. Fundamental Frequencies for different segment durations

Segment Durations (secs)	Both EER	Male EER	Female EER
2	24.3±1.0	23.6±1.7	25.0±1.5
3	20.8±1.2	17.7±1.9	23.0±1.6
4	18.5±1.0	16.5±1.8	19.5±1.5
5	19.6±1.3	18.5±1.8	20.6±1.6

- This experiment was only performed on training dataset
- Studied the impact of segment duration
- EER of 4 second segment duration is less, better performance

Experiment Three

- In our next experiments, we would want to have an algorithm capable of modifying F0 value such that the EER is higher with least perturbing system and achieves privacy
- F0 modification is based on anonymization schemes - noise and normalization
- EER should not hit maximum of 50%

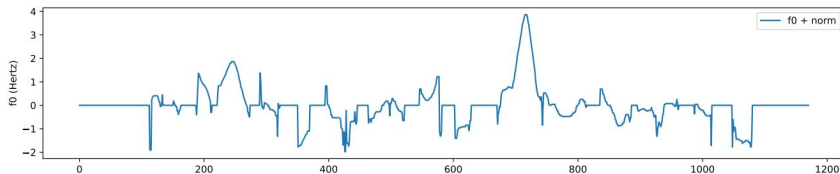
Experiment Three

3. Privacy preserving F0 transformation (both training and testing)

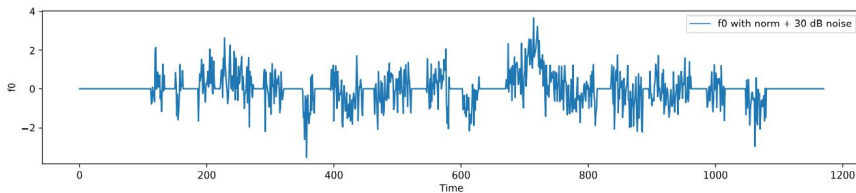
Type of Anonymization	Both EER	Male EER	Female EER
Original F0	20.0±1.2	17.8±1.7	21.8 ±1.7
Noise(15dB)	30.9±1.4	29.2±2.2	32.0 ±2.0
Noise(30dB)	36.0±1.5	32.6±2.2	37.9 ±2.1
Norm	19.9±1.1	18.3±1.8	21.1 ±1.5
Norm with Noise(15dB)	44.2±9.1	43.3±11.8	46.3 ±8.2
Norm with Noise(30dB)	46.0±7.3	44.3±10.6	47.5 ±6.8

- Huge difference with noise 15dB and 30dB when compared with original F0
- F0 normalization alone may not improve privacy
- Norm with Noise (30dB) performed better
- Privacy is more in case of Female as on original F0

Graphical Representation



(a) F0 with normalization



(b) F0 with noise (30dB) and normalization

Experiment Four

4. EER for voiced segments of speech

Feature	Both EER	Male EER	Female EER
Voiced	44.8±3.8	44.2±2.5	41.5 ±2.4

- Assigning the values of voiced frames of speech to 1, while unvoiced is equal 0
- This experiment has not been presented as a privacy transformation but more as a specific evaluation
- EER rate is higher which is a good goal
- EER is quite close to 50% which means that it does contain very little speaker information

Contents

- 1 Introduction
- 2 Speaker Recognition System
- 3 Model Architecture
- 4 Experiments and Results
- 5 Conclusion and Future Work

Conclusion

- EER of MFCC is less than that of F0. The original **F0 retains information** about the speaker
- Performance of Speaker Verification systems also **depends on the segment duration**
- Experiments shows that **modifying F0** (by adding noise and normalization together) **conceals** speaker information and allows better privacy
- **Female** performed better than Male in terms of Privacy
- Experiment where the **voiced frames** of F0 were flattened, showed little **presence of speaker information** (reduced quality)

We conclude that F0 transformation can be used in an anonymization algorithm which can then be used for improving speech technologies

Future Work

- Perform further experiments to check the utility by calculating Word Error Rate (WER) on anonymized speech
- We would like to add additional information like phonetic along with F0 value to test the privacy of the system

Thank You!

MFCC (Mel Frequency Cepstral Coefficients)

Most of the Speaker Verification System uses MFCC as the feature vectors because of its performance.

- **Pre-emphasis, framing and windowing** - Pre-emphasis filter is applied to amplify the high frequencies. The signal is then split into short frames. Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame
- **Spectrum** is obtained through FFT (Fast Fourier Transform, which converts each frame from time domain into frequency domain)
- The above spectrum is passed to **Mel filter bank** to obtain the **Mel spectrum**
- **Reverse spectral analysis** is performed on the Mel spectrum (first take the logarithm of the energy in each Mel filter, then do the inverse transformation, the actual inverse transformation is generally achieved by the DCT discrete cosine transformation)
- Mel Frequency Cepstral Coefficient MFCC is obtained. The system results in 80 real-valued MFCC features.

Normalization

Normalization is done based on the equation below:







$$\hat{x}_t = \frac{1}{\sigma_x} (x_t - \mu_x) \quad (1)$$

x_t = the source F0



\hat{x}_t = the normalized F0

μ_x and σ_x : the mean and standard deviation statistics

Bibliography

-  Fang, Fuming et al. 'Speaker Anonymization Using X-vector and Neural Waveform Models'. In: Sept. 2019, pp. 155–160. DOI: [10.21437/SSW.2019-28](https://doi.org/10.21437/SSW.2019-28).
-  Hanzlíček, Zdeněk and Jindřich Matoušek. 'F0 transformation within the voice conversion framework'. In: vol. 1. Aug. 2007, pp. 1961–1964. DOI: [10.21437/Interspeech.2007-549](https://doi.org/10.21437/Interspeech.2007-549).
-  Highered, Pearson.
Production and Classification of Speech Sounds 3.1 Introduction.
<https://documents.pub/document/production-and-classication-of-speech-sounds-pearson-and-classication-of.html?page=5>.
-  Mingote, Victoria et al. 'Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems'. In: Sept. 2019, pp. 2903–2907. DOI: [10.21437/Interspeech.2019-2550](https://doi.org/10.21437/Interspeech.2019-2550).
-  Reynolds, Douglas et al. 'A Tutorial on Text-Independent Speaker Verification'. In: [EURASIP Journal on Advances in Signal Processing](https://doi.org/10.1155/S1110865704310024) 2004 (Apr. 2004). DOI: [10.1155/S1110865704310024](https://doi.org/10.1155/S1110865704310024).
-  Snyder, David et al. 'Deep Neural Network Embeddings for Text-Independent Speaker Verification'. In: [INTERSPEECH](https://doi.org/10.21437/Interspeech.2017). 2017.

Bibliography (cont.)

-  Srivastava, Brij Mohan Lal et al. 'Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers'. In: [IEEE ICASSP](#). 2020.
-  Tiwari, Vibha Tiwari. 'MFCC and its applications in speaker recognition'. In: [Int. J. Emerg. Technol.](#) 1 (Jan. 2010).