Universidade Federal de Pernambuco

Centro de Informática

Gaussian Mixture Models for Text-Independent Speaker Recognition

Final Term Paper

Eduardo Martins Barros de Albuquerque Tenório

March 3, 2015

Declaration

This paper is a presentation of my research work, as partial fulfillment of the requirement for the degree in Computer Engineering. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

The work was done under the guidance of Prof. Dr. Tsang Ing Ren, at Centro de Informática, Universidade Federal de Pernambuco, Brazil.

Eduardo Martins Barros de Albuquerque Tenório

In my capacity as supervisor of the candidate's paper, I certify that the above statements are true to the best of my knowledge.

Prof. Dr. Tsang Ing Ren

March 3, 2015

Acknowledgements

I am thankful to my family, for the support and patience during the graduation,
To my adviser, Tsang Ing Ren, for the guidance,
To Cleice Souza, for the previous readings and suggestions,
To Sérgio Vieira and James Lyons, for clarify many of my questions.

Live long and prosper

Vulcan salute

Abstract

TODO escrever o abstract após terminar tudo (após a conclusão).

Contents

1. Introduction

The increasing popularity and the intensive usage of computational systems in the everyday of modern life creates the need for easier and less invasive forms of user recognition. While enter a hard to memorize password in a terminal and identify a person using recorded telephone calls are the status quo for respectively authentication and identification, voice biometrics presents itself as a continuing improvement alternative. Passwords can be forgotten and people are biased and unable to be massive trained, but the unique characteristics of a person's voice combined with an automatic speaker recognizer (ASR) outperform any "manual" attempt.

Speech is the most natural way humans have to communicate, being incredibly complex and with numerous specific details related to the speaker [1]. Therefore, it is expected an increasing usage of vocal interfaces to perform actions such as computer login, voice search (e.g., Apple Siri, Google Now and Samsung S Voice) and identification of speakers in a conversation and its content. At present, fingerprint biometrics is adopted in several solutions (e.g., ATMs [2]), authentication through facial recognition comes as built-in software for average computers and iris scan was adopted for a short time by United Kingdom and permanently by United Arab Emirates border controls [3, 4]. These examples indicate a near future where biometrics are common, with people talking to the computer and receiving concise answers, and cash withdrawals allowed via a combination of speaker verification, corrected captcha dictated and other techniques.

Current commercial products based on voice technology (e.g., Dragon Naturally Speaking, KIVOX and VeriSpeak) are usually intended to perform either **speech recognition** (*what* is being said) or **speaker recognition** (*who* is speaking). Voice search applications are designed to determine the content of a speech, usually with no concern about who the speaker is or if there is more than one, while computer login and telephone fraud prevention supplement a memorized personal identification code with speaker verification [5], with no interest on the message spoken. Few applications perform both processes, such as automatic speaker labeling of recorded meetings, that transcribes what each person is saying. To achieve this goal, numerous voice processing techniques have become known in industry and academy, e.g., Natural Language Processing (NLP), Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). Although all of these are interesting state-of-the-art techniques, the subject covered by this paper is the area of speaker recognition and only a small subset of these techniques will be unraveled.

1.1 Speaker Recognition

As stated in [6], speaker recognition may be divided in two subareas. The first is **speaker identification**, aimed to determine the identity of a speaker from a non-unitary set of known speakers. This task is also named speaker identification in **closed set**. In the second, **speaker verification**, the goal is to determine if a speaker is who he or she claims to be, not an imposter. As the set of imposters is unknown, this is an **open set** problem. An intermediate task is **open set identification**, when an "unmatched class" is added in order to categorize all unknown speakers.

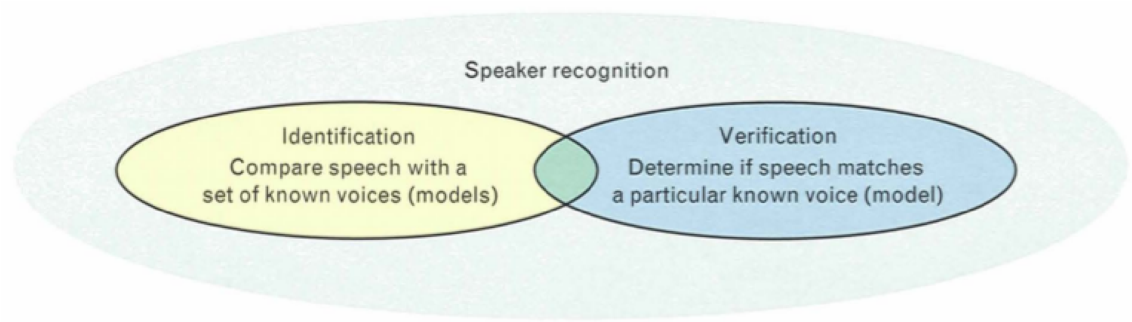

Figure 1.1: Speaker identification and speaker verification are different, but not entirely [5].

The text used may be constrained, such as by type (e.g., digits and letters) and/or by number of words used (e.g., one word or sentences). In **text-dependent** systems the content of the speech is relevant to the evaluation, and the testing texts must belong to the training set (not necessarily be the entire set) [7]. A change in the training text demands an entirely new training section. **Text-independent** systems have no restrictions to the message in both sets, with the non-textual characteristics of the user's voice (e.g., pitch and accent) being the important aspects to the evaluator. These characteristics are presented in different sentences, usage of foreign languages and even gibberish. Between the extremes in constraints falls the **vocabulary-dependent system**, which constrains the speech to come from a limited vocabulary (e.g., digits) from which test words or phrases are selected (e.g., "two" or "one-two-three") [5].

The focus of this paper is in **text-independent speaker recognition** and to achieve that, Gaussian Mixture Models are used.

1.2 Objectives

The objective of this study is to implement an ASR that executes the listed actions:

- From a group of enrolled speakers identify who produced a given speech signal;
- Determine if a speaker is the claimed enrolled speaker or an imposter, given the speech signal produced.

1.3 Document Structure

Chapter 2 contains basic information about voice recognition, as well as the basic architecture of speaker identification and verification systems. The feature extraction process is explained in Chapter 3, from the reasons for its use to the chosen technique (Mel-Frequency Cepstral Coefficient, MFCC). In Chapter 4 the GMM and the UBM are detailed. Experiments are described in Chapter 5, as well as its results. Finally, Chapter 6 concludes the study. Furthermore, this work contains an appendix with the most relevant pieces of the source code (Section A).

2. Speaker Recognition Systems

The process of voice recognition lies on the field of pattern classification, with the speaker's utterance (a speech signal) as input for a classifier and a decision as output. This decision may be, given a speech signal $\boldsymbol{Y}$ produced by a speaker $\mathcal{S}$ and a set $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, ..., \mathcal{S}_S\}$ of known users,

$$\text{classify } \mathcal{S} \text{ as } \mathcal{S}_i \text{ if } i = \arg\max_j P(\mathcal{S}_j|\boldsymbol{Y}). \tag{2.1}$$

This is a case of speaker identification and the output is a $\mathcal{S}_i$ from $\boldsymbol{\mathcal{S}}$. Another type of decision is

$$\text{if } P(\mathcal{S}_i|\boldsymbol{Y}) \begin{cases} \geq \alpha, & \text{accept } \mathcal{S} \text{ as } \mathcal{S}_i, \\ < \alpha, & \text{reject } \mathcal{S} \text{ as } \mathcal{S}_i, \end{cases} \tag{2.2}$$

a speaker verification decision, with a binary output, given a $\mathcal{S}$ who produced $\boldsymbol{Y}$, a claimed identity $\mathcal{S}_i$ from $\boldsymbol{\mathcal{S}}$ and a threshold of acceptance α. This chapter (and indirectly the whole document) is about the type of decision seen in Eq. 2.2.

2.1 Basic Concepts

2.1.1 Utterance

An utterance is a piece of speech produced by a speaker. It may be a word, a statement or any vocal sound. The terms *utterance* and *speech signal* sometimes are used interchangeably, but from herenow speech signal will be associated to an utterance recorded, digitalized and ready to be processed. An example is shown in Fig. 2.1.

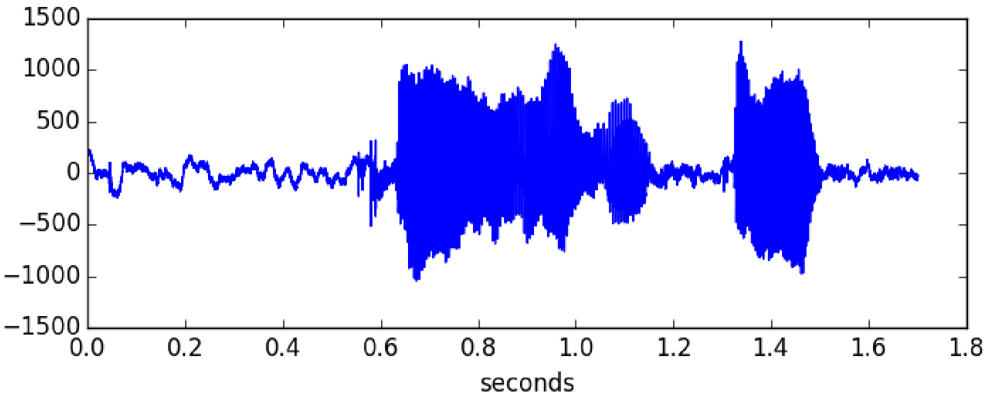

Figure 2.1: Speech signal for utterance “karen livescu”, from the MIT dataset [8].

2.1.2 Features

The raw speech signal is unfit for usage by a recognition system. For a correct processing, the unique features from the speaker's vocal tract are extracted, reducing the number of variables the system needs to deal with (leading to a simpler implementation) and performing a better evaluation (and avoiding the curse of dimensionality). Due to the stationary properties of the speech signal when analyzed in a short period of time, it is divided in overlapping frames of small and predefined length, to avoid "loss of significancy" in the features [9, 10]. This extraction is executed by the MFCC algorithm, explained in details in Chapter 3.

2.1.3 Dependency x Independency

When designing a speaker recognition system, one of the most important aspects to consider is the type of dependency to text it will have. In a text-dependent system the choice of what to say is made at design time, with different degrees of freedom. The testing utterance must be a subset of the training set. A simpler version may require that the same text be spoken during the model's training and testing phases, while a more sophisticated one may allow the speaker to say just a few words from a sentence or even speak them out of order. The most common acoustic model used for this system is the HMM, with the unit modeled and the number of states depending heavily on the application [7].

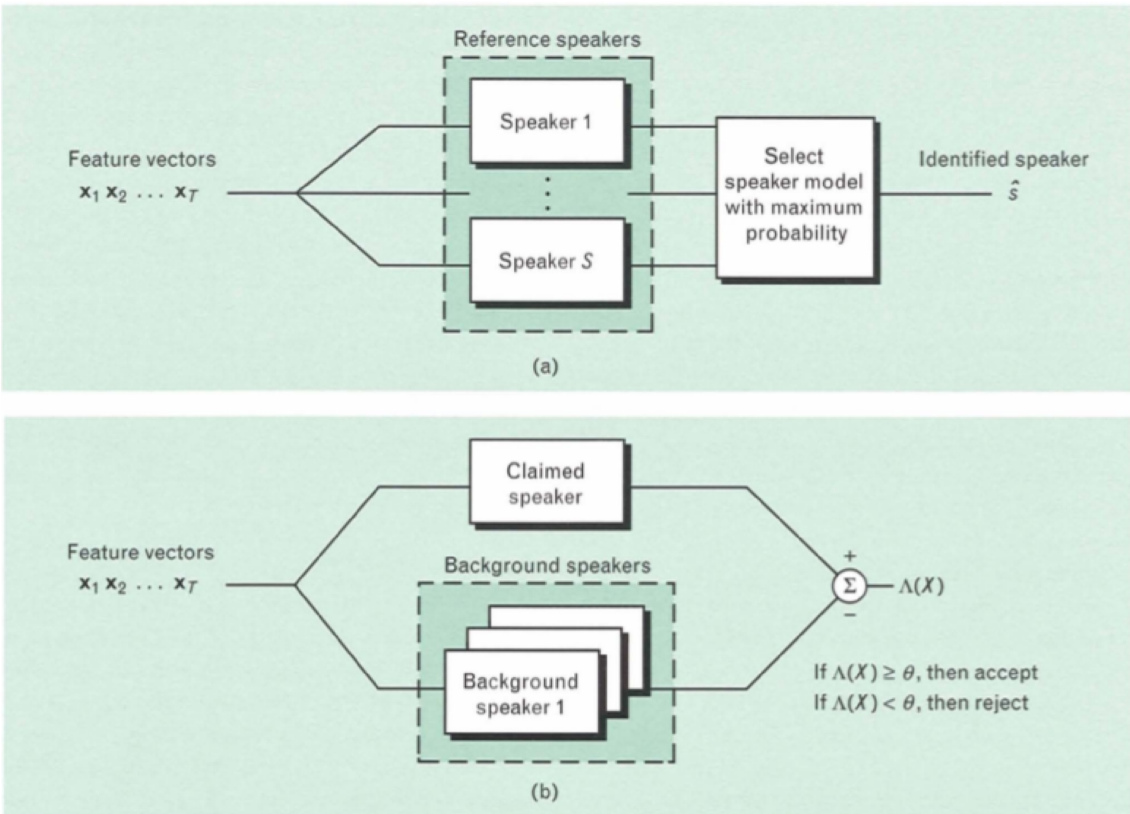

Figure 2.2: Speaker-recognition systems for (a) identification and (b) verification [5].

Text-independent recognition is less problematic than the previous one for several reasons. First, the designer does not need to worry about what the speaker will say, since it is a vocal sound. The recognition is performed over the unique features of each vocal tract, shown when the person speaks. Second, for being free of time constraints, a HMM of single state (i.e. a GMM) fits well for the task [7]. Third, the ability to apply text-independent verification to unconstrained speech encourages the use of audio recorded from a wide variety of sources (e.g., speaker indexing of broadcast audio or forensic matching of law-enforcement microphone recordings) [6].

As stated in Section 1.1, the focus of this paper is in text-independent speaker verification, and due to that it is necessary to understand what is the likelihood ratio test and how the models are trained and tested.

2.2 Basic Speaker Verification Architecture

The architecture of a speaker verification system is pretty basic. Given a speech signal from a speaker $\mathcal{S}$ who claims to be a particular speaker $\mathcal{S}_i$ from a set of enrolled speakers $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, ..., \mathcal{S}_S\}$, the strength of the claim resides on how similar the features $\boldsymbol{X}$, extracted from the speech $\boldsymbol{Y}$ produce by $\mathcal{S}$, are to the features from $\mathcal{S}_i$ "memorized" by the system (see Eq. 2.2). However a subset of enrolled speakers may have vocal similarities, leading to a misclassification of one enrolled speaker as another (a false positive). To reduce the error rate, the system must decide not only if a speech signal came from the claimed speaker, but also if it came from a background composed of all other enrolled speakers. For this, a likelihood ratio test is performed.

2.2.1 Likelihood Ratio Test

Given the speech signal $\boldsymbol{Y}$, and assuming it was produced by only one speaker, the detection task can be restated as a basic test between two hypoteses [11]:

H_0: $\boldsymbol{Y}$ is from the claimed speaker $\mathcal{S}_i$;

H_1: $\boldsymbol{Y}$ is <u>not</u> from the claimed speaker $\mathcal{S}_i$.

The optimum test to decide which hypotesis is valid is the **likelihood ratio test** between both posterior probabilities $P(H_0|\boldsymbol{Y})$ and $P(H_1|\boldsymbol{Y})$,

$$\frac{P(H_0|\boldsymbol{Y})}{P(H_1|\boldsymbol{Y})} \begin{cases} \geq \theta, & \text{accept } H_0, \\ < \theta, & \text{reject } H_0, \end{cases} \tag{2.3}$$

where the decision threshold for accepting or rejecting H_0 is θ. Applying Bayes' rule

$$P(H_i|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|H_i)P(H_i)}{p(\boldsymbol{Y})}, \tag{2.4}$$

and considering all hypoteses equally probable *a priori*, Eq. 2.3 can be simplified to

$$\frac{p(\boldsymbol{Y}|H_0)}{p(\boldsymbol{Y}|H_1)} \begin{cases} \geq \theta, & \text{accept } H_0, \\ < \theta, & \text{reject } H_0, \end{cases} \tag{2.5}$$

where $p(\boldsymbol{Y}|H_i), i = 0, 1$, is the probability density function for the hypothesis H_i evaluated for the observed speech segment $\boldsymbol{Y}$. Fig. 2.3 shows the basic components found in speaker verification systems based on likelihood ratios. The front-end processing module extracts features $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$ (where $\boldsymbol{x}_t$ is the feature indexed at discrete time $t \in [1, 2, ..., T]$) from the speech signal $\boldsymbol{Y}$, and feeds it to the models for the hypotesized speaker and the background. The hypoteses H_0 and H_1 are represented mathematically by models detoned λ_{hyp} and $\lambda_{\overline{hyp}}$, respectively. The likelihood equation from Eq. 2.5 is be better represented as

$$\frac{p(\boldsymbol{X}|\lambda_{hyp})}{p(\boldsymbol{X}|\lambda_{\overline{hyp}})} \begin{cases} \geq \theta, & \text{accept } \mathcal{S} \text{ as } \mathcal{S}_i, \\ < \theta, & \text{reject } \mathcal{S} \text{ as } \mathcal{S}_i. \end{cases} \tag{2.6}$$

The division seen in Eq. 2.6 can be transformed in a subtraction by the application of the logarithm function. Since the logarithm is monotonically increasing, the behavior of the likelihood ratio is maintained, and Eq. 2.6 is replaced by the log-likelihood ratio

$$\Lambda(\boldsymbol{X}) = \log p(\boldsymbol{X}|\lambda_{hyp}) - \log p(\boldsymbol{X}|\lambda_{\overline{hyp}}) \tag{2.7}$$

The more likely $\mathcal{S}$ is of λ_{hyp} and the less likely $\mathcal{S}$ is of $\lambda_{\overline{hyp}}$ easier is to accept $\mathcal{S}$ as the claimed $\mathcal{S}_i$.

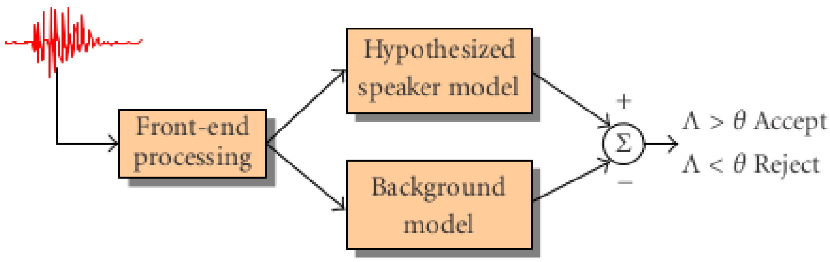

Figure 2.3: Likelihood ratio-based speaker detection system [1].

2.2.2 Training Phase

Once the features are extracted from the speech signal, they are used to train the models λ_{hyp} and $\lambda_{\overline{hyp}}$. A high-level demonstration of the training of λ_{hyp} (mathematical representation of $\mathcal{S}_i$) is shown in Fig. 2.4.

Figure 2.4: The statistical model of $\mathcal{S}$ is created from the features $\boldsymbol{X}$ [1].

Due to λ_{hyp} be a model of $\mathcal{S}_i$, the features used to train it (i.e., estimate $p(\boldsymbol{X}|\lambda_{hyp})$) are extracted from speech signals produced by $\mathcal{S}_i$. For $\lambda_{\overline{hyp}}$ the same process is executed, considering it a representation of another speaker (i.e., the *mirror* of $\mathcal{S}_i$).

The model $\lambda_{\overline{hyp}}$ is not well-defined. It is composed of the features extracted from speech signals from all other speakers except $\mathcal{S}_i$. One popular approach to deal with this difficulty is to define the likelihood of $\lambda_{\overline{hyp}}$ as

$$p(\boldsymbol{X}|\lambda_{\overline{hyp}}) = \mathcal{F}[p(\boldsymbol{X}|\lambda_1), ..., p(\boldsymbol{X}|\lambda_S)], \tag{2.8}$$

where $\mathcal{F}$ is a function of all likelihoods of set $\boldsymbol{\mathcal{S}}$ (except $\mathcal{S}_i$), such as mean or maximum. In various contexts, this set of other speakers has been called likelihood ratio sets, cohorts, and background speakers [12].

Another popular approach is to create only one model for $\lambda_{\overline{hyp}}$, containing features extracted from speeches of all enrolled speakers (even the claimed $\mathcal{S}_i$). The weight of $\mathcal{S}_i$ in $\lambda_{\overline{hyp}}$ is reduced due to the presence of all other speakers from $\boldsymbol{\mathcal{S}}$. This model is named **Universal Background Model** (UBM) and is denoted by λ_{bkg}. Eq. 2.6 is rewritten as

$$\frac{p(\boldsymbol{X}|\lambda_{hyp})}{p(\boldsymbol{X}|\lambda_{bkg})} \begin{cases} \geq \theta, & \text{accept } \mathcal{S} \text{ as } \mathcal{S}_i, \\ < \theta, & \text{reject } \mathcal{S} \text{ as } \mathcal{S}_i. \end{cases} \tag{2.9}$$

The UBM is explained in details in Chapter 4.

2.2.3 Test Phase

As seen in Eq. 2.9, the decision process is based on a function *Score* calculated from the likelihood ratio of $p(\boldsymbol{X}|\lambda_{hyp})$ and $p(\boldsymbol{X}|\lambda_{bkg})$. Being the vector of features $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$, with all $\boldsymbol{x}_t$ independent of the others, the likelihood of a model λ given $\boldsymbol{X}$ can be written as

$$p(\boldsymbol{X}|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda). \tag{2.10}$$

Using the logarithm function, Eq. 2.10 becomes

$$\log p(\boldsymbol{X}|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{x}_t|\lambda), \tag{2.11}$$

where the term $\frac{1}{T}$ is used to normalize the log-likelihood to the duration of the speech signal. That said, the likelihood ratio given by Eq. 2.9 becomes a subtraction

$$\mathit{Score}(\boldsymbol{X}) = \log p(\boldsymbol{X}|\lambda_{hyp}) - \log p(\boldsymbol{X}|\lambda_{bkg}), \tag{2.12}$$

with $\mathit{Score}(\boldsymbol{X})$ being compared to $\log \theta$ and maintaining the same rule from Eq. 2.9.

3. Feature Extraction

As an acoustic wave propagated through space over time, the speech signal is not appropriate to be evaluated by the speaker verification system. In order to deliver decent outcomes, a good parametric representation must be provided to the system. This task is performed by the feature extraction process, which transforms a speech signal into a sequence of characterized measurements, i.e. features. As stated in [9], “the usual objectives in selecting a representation are to compress the speech data by eliminating information not pertinent to the phonetic analysis of the data, and to enhance those aspects of the signal that contribute significantly to the detection of phonetic differences”. According to [13] the ideal features should:

- occur naturally and frequently in normal speech;
- be easily measurable;
- vary highly among speakers and be very consistent for each speaker;
- not change over time nor be affected by the speaker’s health;
- be robust to reasonable background noise and to transmission characteristics;
- be difficult to be artificially produced;
- not be easily modifiable by the speaker.

Features may be categorized based on vocal tract or behavioral aspects, divided in (1) short-time spectral, (2) spectro-temporal, (3) prosodic and (4) high level [14]. Short-time spectral features are usually calculated using millisecond length windows and describe the voice spectral envelope, composed of supralaryngeal properties of the vocal tract, e.g. timbre. Prosodic and spectro-temporal occur over time, e.g. rhythm and intonation, and high level features occur during the conversation, e.g. accents.

The parametric representations evaluated in [9] may be divided into those based on the Fourier spectrum, Mel-Frequency Cepstrum Coefficients (MFCC) and Linear Frequency Cepstrum Coefficients (LFCC), and those based on the Linear Prediction Spectrum, Linear Prediction Coefficients (LPC), Reflection Coefficients (RC) and Linear Prediction Cepstrum Coefficients (LPCC). The better evaluated representation was the MFCC, with minimum and maximum accuracy of 90.2% and 99.4% respectively, leading to its choice as the parametric representation in this work.

3.1 Mel-Frequency Cepstral Coefficient

MFCC is a highly used parametric representation in the area of voice processing, due to its similarity with the mode the human ear operates. Despite the fact the ear is divided in three sections, i.e. outer, middle and inner ears, only the last is mimicked. The mechanical pressure waves produced by the triad hammer-anvil-stirrup are received by the cochlea (Fig. 3.1), a spiral-shaped cavity with a set of inner hair cells attached to a membrane (the basilar membrane) and filled with a liquid. This structure converts motion to neural activity through a non-uniform spectral analysis [10] and passes it to the pattern recognition in the brain.

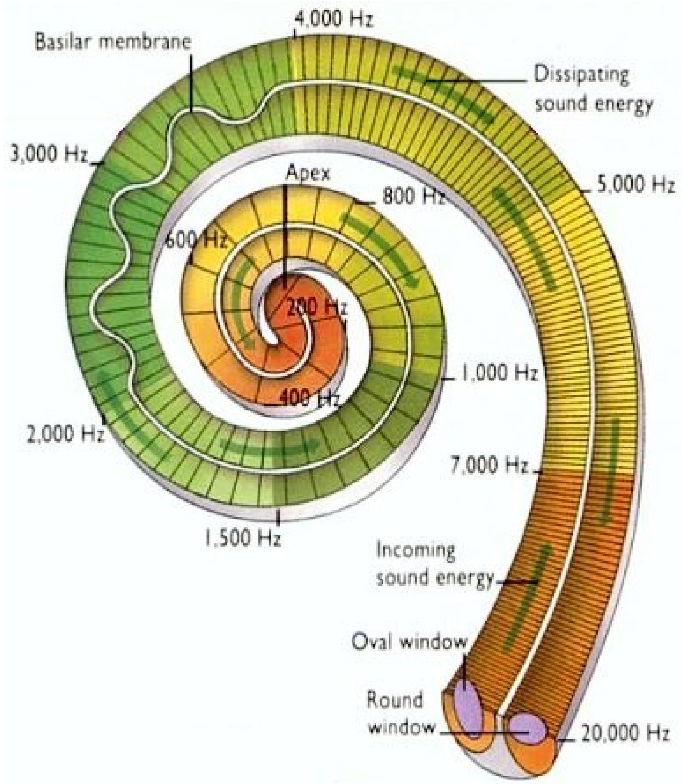

Figure 3.1: Cochlea divided by frequency regions [15].

A key factor in the perception of speech and other sounds is *loudness*, a quality related to the physical property of sound pressure level. Loudness is quantified by relating the actual sound pressure level of a pure tone (in dB relative to a standard reference level) to the perceived loudness of the same tone (in a unit called phons) over the range of human hearing (20 Hz–20 kHz) [10]. As shown in Fig. 3.2, a 100 Hz tone at 60 dB is equal in loudness to a 1000 Hz tone at 50 dB, both having the *loudness level* of 50 phons (by convention).

3.1.1 The Mel Scale

The mel scale is the result of an experiment conducted by Stevens, Volkmann and Newman [17] intended to measure the perception of a pitch and construct a scale based on it. Each observer was asked to listen to two tones, one in the fixed frequencies 125, 200, 300, 400, 700, 1000, 2000, 5000, 8000 and 12000 Hz, and the other free to have its frequency

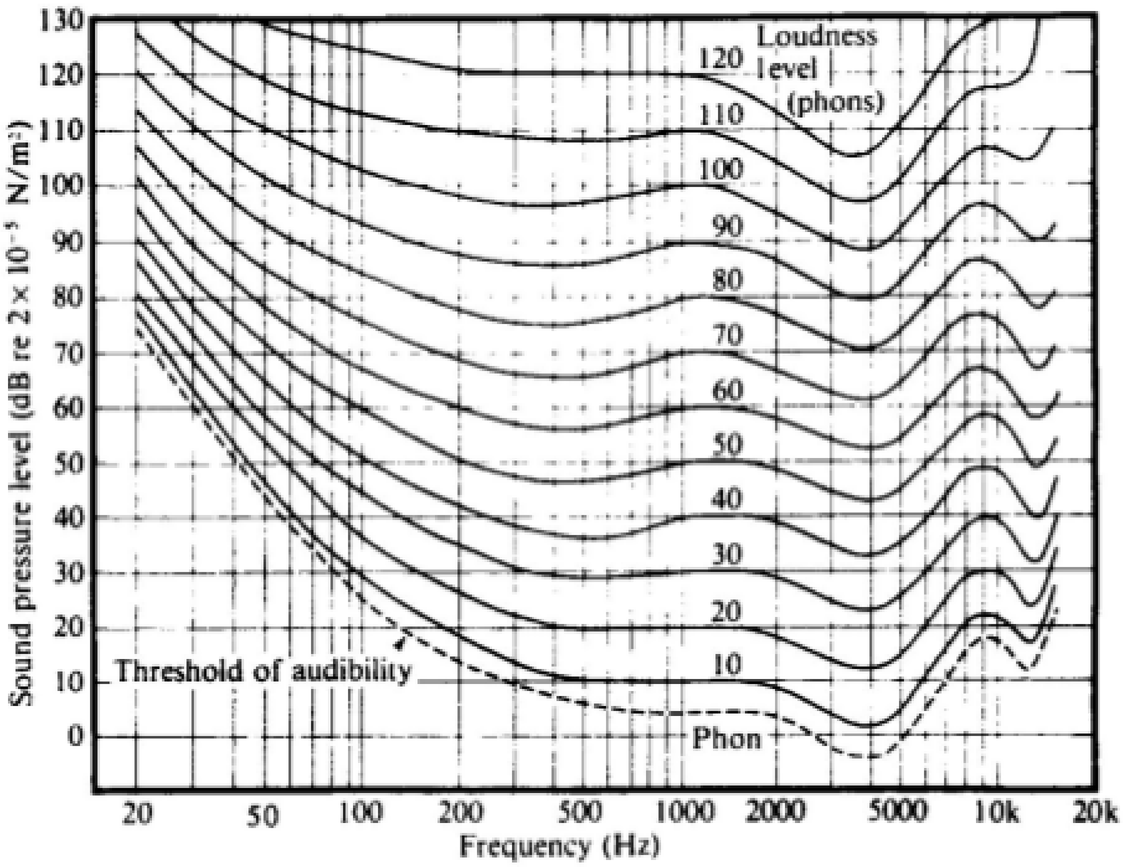

Figure 3.2: Loudness level for human hearing [16].

varied by the observer for each fixed frequency of the first tone. An interval of 2 seconds separated both tones. The observers were instructed to say in which frequency the second tone was “half the loudness” of the first. A geometric mean was taken from the observers’ answers and a measure of 1000 mels was assigned to the frequency of 1000 Hz, 500 mels to the frequency sounding half as high (as determined by Fig. 1 in [17]) and so on.

Decades after the creation of the mel scale, O’Shaughnessy [18] published an equation to convert frequencies in hertz to frequencies in mels.

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

Being logarithmic, the growth of a mel-frequency curve is slow with a linear growth of the frequency in hertz. Eq. 3.1 sometimes is used only for frequencies higher than 1000 Hz while the lower frequencies obey a linear function. In this work all conversions will use Eq. 3.1, as shown by Fig. 3.3.

3.1.2 Cepstrum

3.1.3 Extraction Process

Pre-emphasis

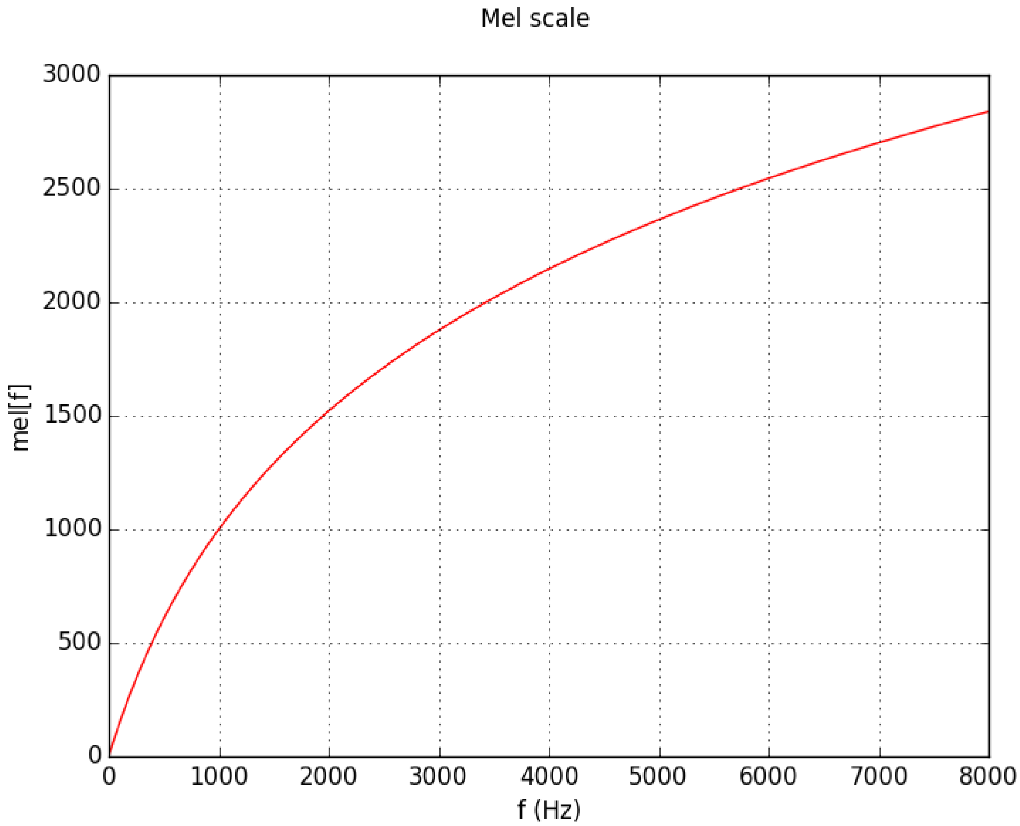

Figure 3.3: The logarithm curve of the mel-frequency.

4. Gaussian Mixture Models

5. Experiments

6. Conclusion

TODO escrever a conclusão após terminar tudo (antes do abstract)

A. Codes

References

[1] Frédéric Bimbot et al. “A Tutorial on text-independent speaker verification”. In: *EURASIP Journal on Applied Signal Processing* 4 (2004), pp. 430–451.

[2] P.T. Wang and S.M. Wu. “Personal fingerprint authentication method of bank card and credit card”. Pat. US Patent App. 09/849,279. 2002. URL: `https://www.google.com/patents/US20020163421`.

[3] M. Angela Sasse. “Red-Eye Blink, Bendy Shuffle, and the Yuck Factor: A User Experience of Biometric Airport Systems”. In: *Security & Privacy, IEEE* 5.3 (2007), pp. 78–81.

[4] Ahmad N. Al-Raisi and Ali M. Al-Khouri. “Iris recognition and the challenge of homeland and border control security in UAE”. In: *Telematics and Informatics* 25.2 (2008), pp. 117–132.

[5] Douglas A. Reynolds. “Automatic Speaker Recognition Using Gaussian Mixture Speaker Models”. In: *The Lincoln Laboratory Journal* 8.2 (1995), pp. 173–192.

[6] Douglas A. Reynolds and William M. Campbell. “Springer Handbook of Speech Processing”. In: ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang. Berlin: Springer, 2008. Chap. Text-Independent Speaker Recognition, pp. 763–780.

[7] Martial Hébert. “Springer Handbook of Speech Processing”. In: ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang. Berlin: Springer, 2008. Chap. Text-Dependent Speaker Recognition, pp. 743–762.

[8] Ram H. Woo, Alex Park, and Timothy J. Hazen. “The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments”. In: *Odyssey 2006: The Speaker and Language Recognition Workshop, San Juan, Puerto Rico, June 28-30, 2006*. IEEE, 2006, pp. 1–6.

[9] Steven B. Davis and Paul Mermelstein. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-28.4 (1980), pp. 357–366.

[10] Lawrence R. Rabiner and Ronald W. Schafer. “Introduction to Digital Speech Processing”. In: *Foundations and Trends in Signal Processing* 1.1-2 (2007), pp. 1–194.

[11] Douglas A. Reynolds. “Speaker identification and verification using Gaussian mixture speaker models”. In: *Speech Communication* 17.1 (1995), pp. 91–108.

[12] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. In: *Digital Signal Processing* 10.1 (2000), pp. 19–41.

[13] Jared J. Wolf. “Efficient acoustic parameters for speaker recognition”. In: *Journal of the Acoustical Society of America* 51 (1972), pp. 2044–2056.

[14] Hector N. B. Pinheiro. *Sistemas de Reconhecimento de Locutor Independente de Texto*. Trabalho de Graduação. Universidade Federal de Pernambuco, 2013.

[15] Ethan. *Don’t you hear that?* May 10, 2010. URL: `http://scienceblogs.com/startswithabang/2010/05/10/dont-you-hear-that/`.

[16] Harvey Fletcher and Wilden A. Munson. “Loudness, Its Definition, Measurement and Calculation”. In: *Bell Telephone Laboratories* 12.4 (1933), pp. 82–108.

[17] Stanley S. Stevens, John Volkmann, and Edwin B. Newman. “A Scale for the Measurement of the Psychological Magnitude Pitch”. In: *The Journal of Acoustical Society of America* 8.3 (1937), pp. 185–190.

[18] Douglas O’Shaughnessy. *Speech Communications: Human and Machine*. Addison-Wesley, 1987.