# Searching for a Voice "DNA"

Eduardo R. Silva, Manfredo H. Tabacniks
edurs@if.usp.br, tabacniks@if.usp.br

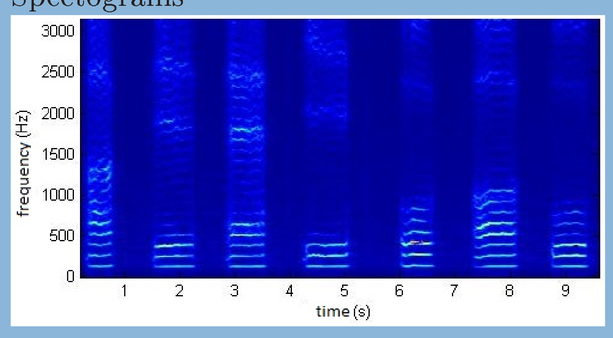
Instituto de Física
Universidade de São Paulo

## Introduction

In the forensic sciences, a topic of paramount importance and yet to be fulfilled (if ever it will be) is the **positive speaker identification** by automatic means with high levels of trustworthiness. As a multidisciplinary task, it involves linguistics, phonetics, probability theory, information theory, pattern recognition, acoustic (and psychoacoustic), physics etc [1]. It is also an area where fiction and state-of-art reality differ completely, misleading the general public's expectation: the "CSI Effect" leads to the feeling that a software that attains "perfect" speaker identification is easily available, a conviction actually perceived in court rooms where judges and jury members come with high expectations about the evidences presented in court [2].

## Speech Signal Representations

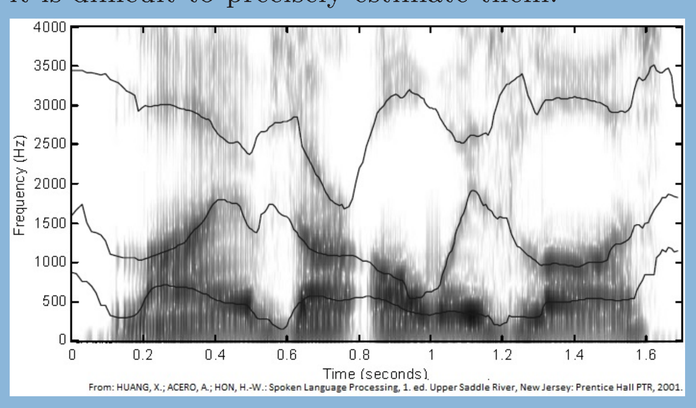
- Short-time Fourier Analysis
  - Spectograms




  - Pitch-Synchronous Analysis
- Acoustical Model of Speech Production: analyzes the laws of physics envolved with propagation of sound in the vocal tract. It connsiders: three-dimensional wave propagation, vocal tract shape, losses to heat conduction and viscous friction at the vocal tract walls, nasal coupling etc.
- Linear Predictive Coding (LPC): assumes that speech signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds (sibilants and plosive sounds). LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz.
- Perceptually Motivated Representations: set of methods motivated by the behavior of the human auditory system
  - The Bilinear Transform

$$s = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \text{ for } 0 < \alpha < 1$$

    is a mapping in the complex plane that maps the unit circle onto itself. It is similar to the Bark and mel scale for an appropriate choice of $\alpha$.
  - Cepstral Processing and **Mel-frequency Cepstrum**: see box on the right.
  - Perceptual Linear Prediction (PLP)
- Statistical Formant Tracking: Formant frequencies are the resonances in the vocal tract. They are very useful features for speech recognition, but it is difficult to precisely estimate them.

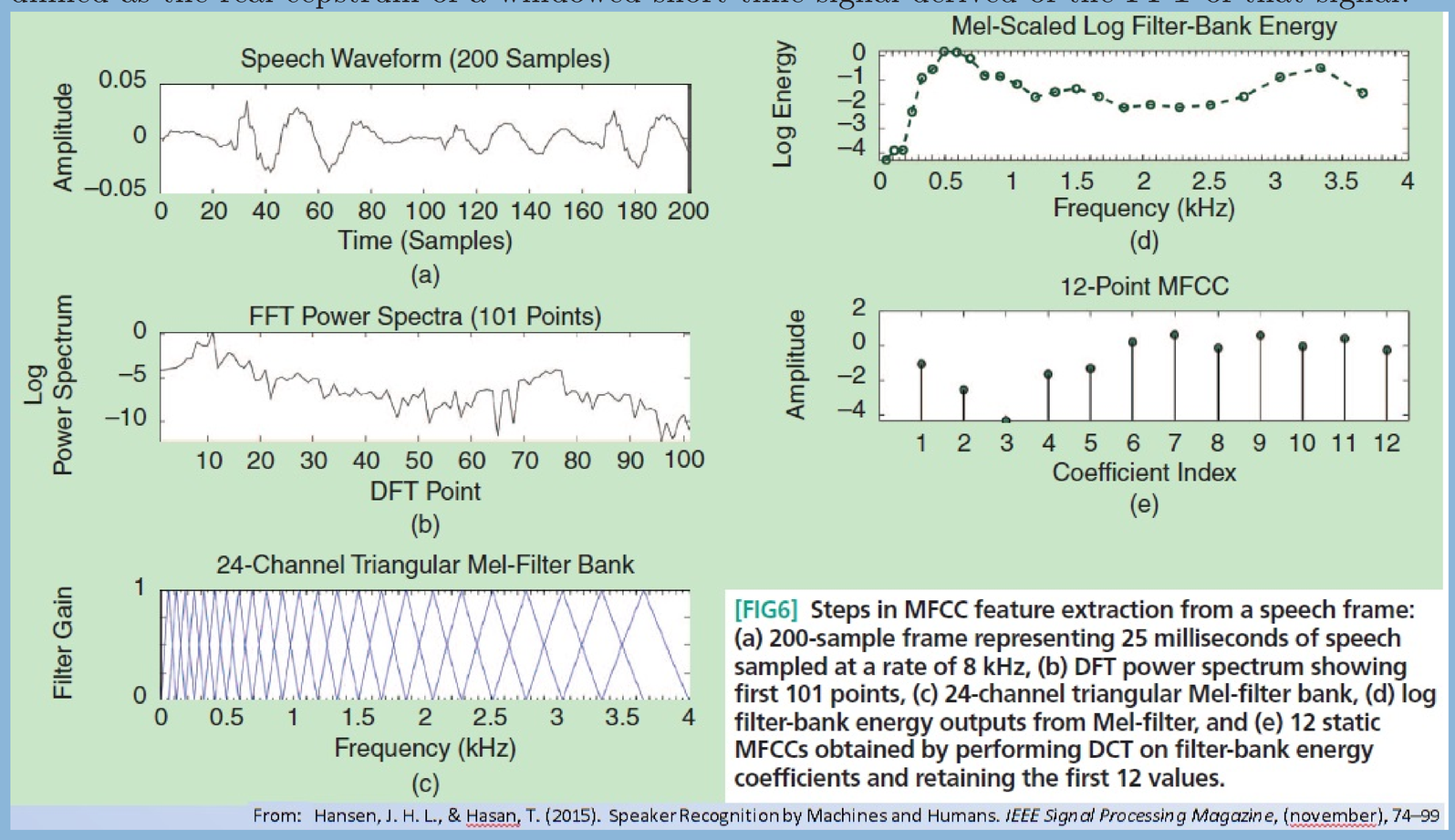



## MFCC: Mel-Frequency Cepstrum Coefficients

*Cepstrum* was first defined [4] as:

$$\text{Power Cepstrum of a Signal} = \left|\mathscr{F}^{-1}\left\{log(|\mathscr{F}\left\{f(t)\right\}|^2)\right\}\right|^2$$

The Mel-Frequency Cepstrum Coefficients (MFCC) is a perceptually motivated signal representation difined as the real cepstrum of a windowed short-time signal derived of the FFT of that signal.

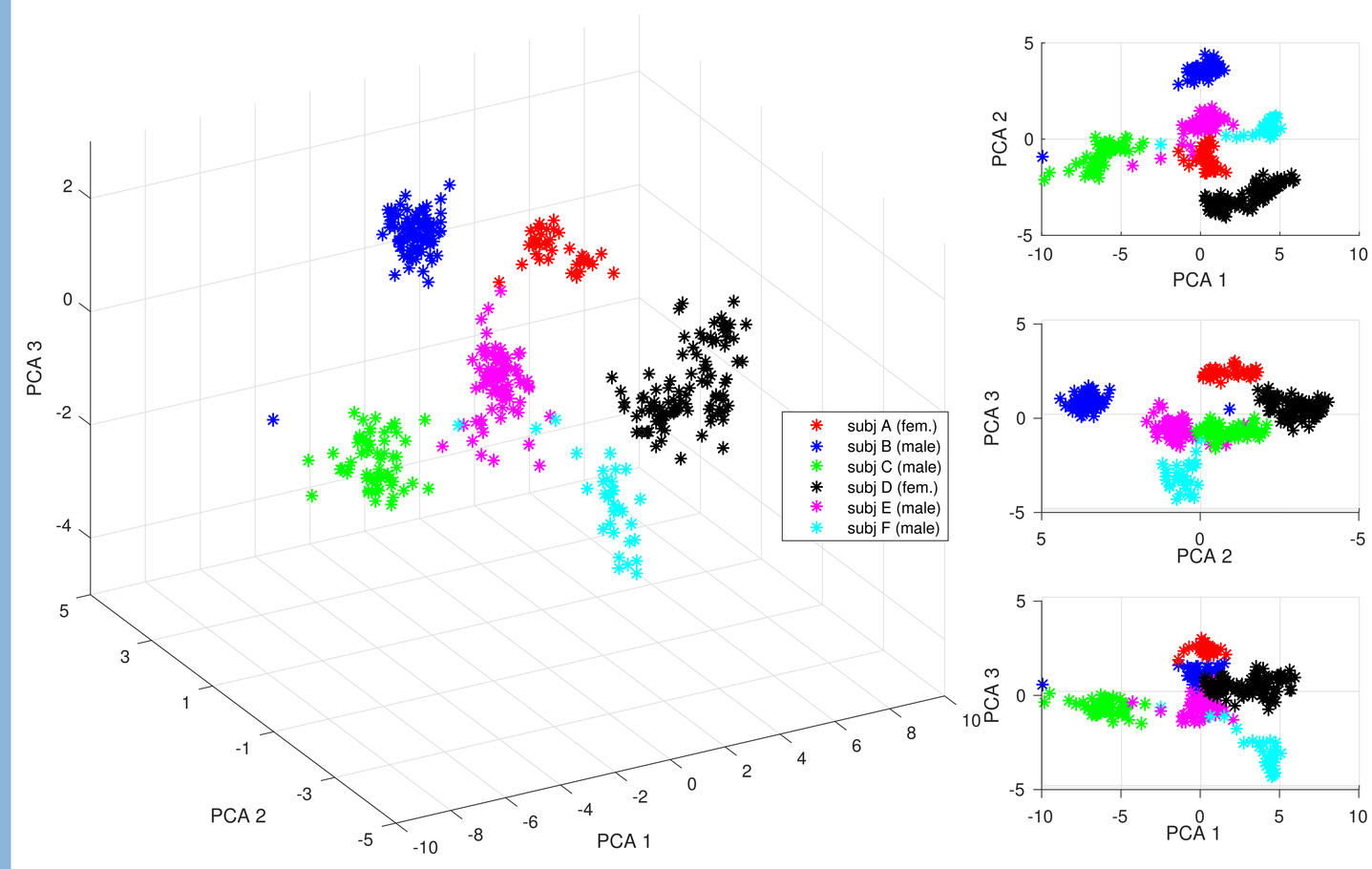



[FIG6] Steps in MFCC feature extraction from a speech frame: (a) 200-sample frame representing 25 milliseconds of speech sampled at a rate of 8 kHz, (b) DFT power spectrum showing first 101 points, (c) 24-channel triangular Mel-filter bank, (d) log filter-bank energy outputs from Mel-filter, and (e) 12 static MFCCs obtained by performing DCT on filter-bank energy coefficients and retaining the first 12 values.

From: Hansen, J. H. L., & Hasan, T. (2015). Speaker Recognition by Machines and Humans. *IEEE Signal Processing Magazine*, (november), 74–99

## Preliminary Results

This work aims the development and identification of sound processing routines to improve the selection and positive identification of speakers by reviewing current signal processing methods.
As an example, MFCC extraction procedure was applied to a set of recordings of the vocal /a/ spoken by six subjects – four males and two females – where principal component analysis (PCA) was employed to identify clusters. As can be seen in the figure below, despite the distinguishable "clouds" of MFCC for each subject (the first three principal components explain 81% of the variance), there is a considerable "grey area".
Next step of this research is applying relevance learning as a pre-processing tool to evidence distinction in sets of voice parameters.




## References

[1] HUANG, X.; ACERO, A.; HON, H.-W.: **Spoken Language Processing**, 1. ed. Upper Saddle River, New Jersey: Prentice Hall PTR, 2001.

[2] MAHER, R. C.: *Audio Forensic Examination: authenticity, enhancement, and interpretation*, **IEEE Signal Processing Magazine**, p. 84-94, mar. 2009.

[3] HANSEN, John H. L.; HASAN, Taufiq: *Speaker Recognition by Machines and Humans*, **IEEE Signal Processing Magazine**, p. 74-99, nov. 2015.

[4] BOGERT, B. P.; HEALY, M. J. R.; TURKEY, J. W.: *The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking*, **Proceedings of the Symposium on Time Series Analysis**, (M. Rosenblatt, Ed) Chapter 15, 209-243. New York: Wiley, 1963.