Using a new Discretization of the Fourier Transform to Discriminate Voiced From Unvoiced Speech

Antonio Camarena-Ibarrola and Edgar Chavez

{camarena,elchavez}@umich.mx

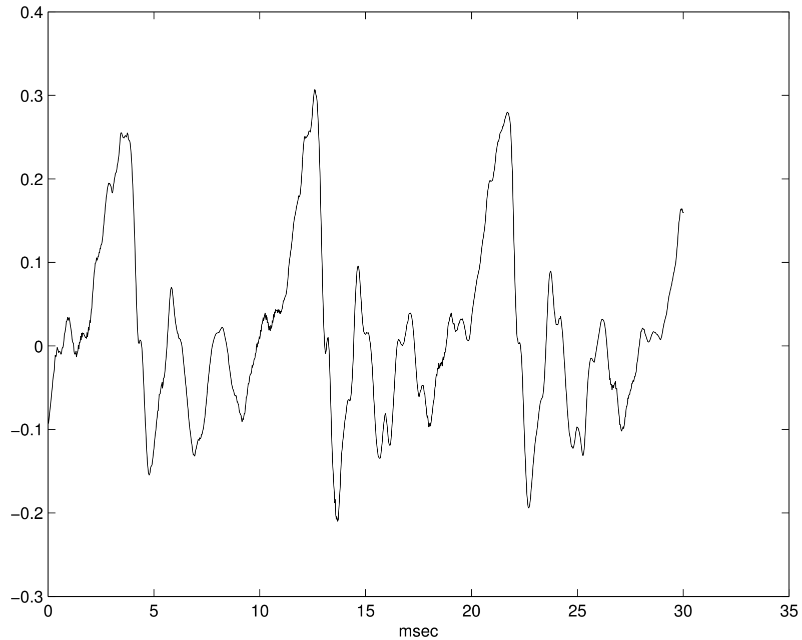
Universidad Michoacana de Sán Nicolás de Hidalgo Morelia, Michoacán. México

The talk

- To show a recent signal procesing analisys technique that we adapted to extract features from the speech signal.
- It was applied to a simple especific, yet relevant problem.
- The results encourage its use in robust ASR or speaker ID.
- Although there are several feature extractors DFT, MFCC, LPC, wavelets, it is always worthed to search for more robust feature extraction techniques.

Voiced Speech

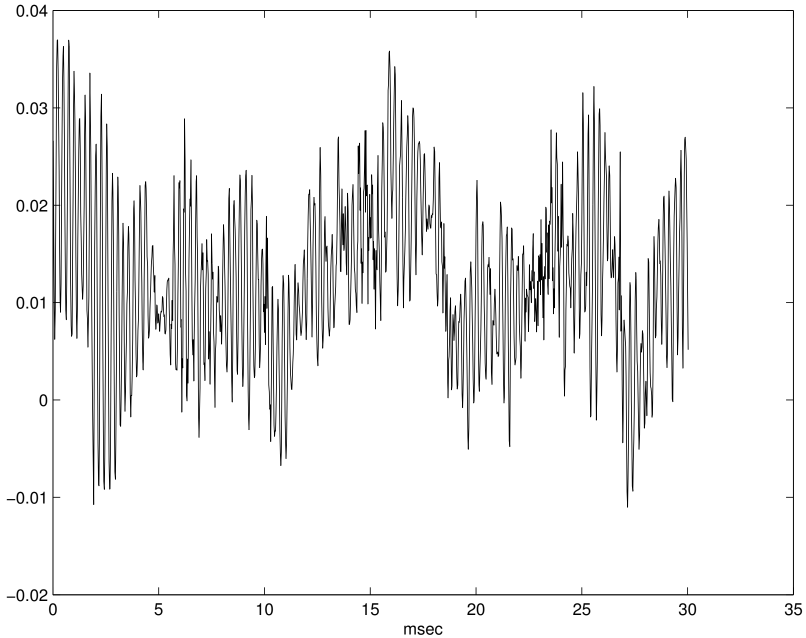
A segment of the speech signal is known as voiced if the vocal folds vibrate during its production. This vibration introduces periodicity in the signal



30 milliseconds of voiced speech. Vowel e

Unvoiced speech

Its statistical properties are predictable (Stationary signal)



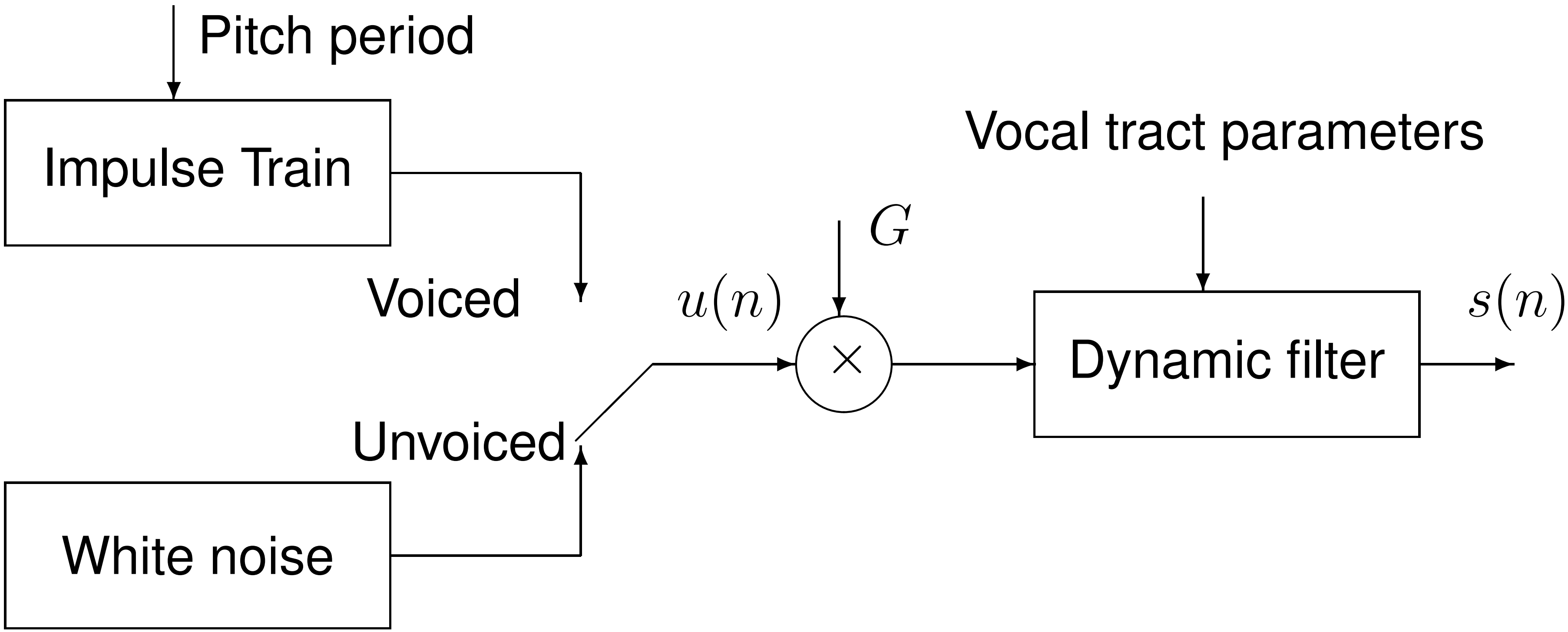
30 miliseconds of unvoiced speech. Sound of the “s”

Importance

Discriminating voiced from Unvoiced Speech

- Automatic Speech Recognition
- Detection of Laringeal Diseases
- Speech synthesis

LPC Based Speech Synthesis



The CFT and the DFT

The CFT ($X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$) has an infinite kernel

The DFT ($X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}$) has a kernel of the same than the lenght of the signal

$$\begin{bmatrix} X[0] \\ X[1] \\ X[2] \\ \vdots \\ X[N] \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)^2} \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ x[2] \\ \vdots \\ x[N] \end{bmatrix}$$

(1)

Where $W_N = e^{-j2\pi/N}$

The CFT with a finite kernel!!!

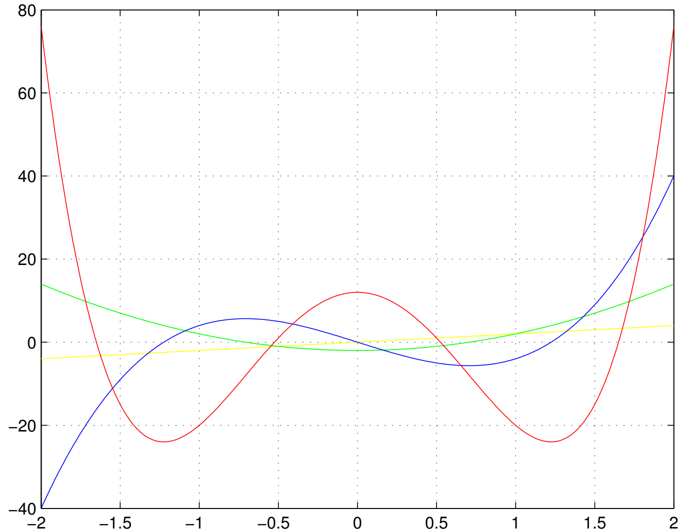
$$F_{i,j} = \frac{\pi}{\sqrt{2n}} \sqrt[4]{\frac{4n+3-x_j^2}{4n+3-x_i^2}} [cos(x_i x_j) + j sin(x_i x_j)]$$

- x is a vector with the roots of Hermite's Polynomial of degree n
- $f(x)$ is the signal sampled at points x (not necesarily vailable)
- $g = Ff(x)$ is a vector with the CFT evaluated at the roots of the Hermite's Polinomial

Hermite's Polynomials

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{x^2}$$

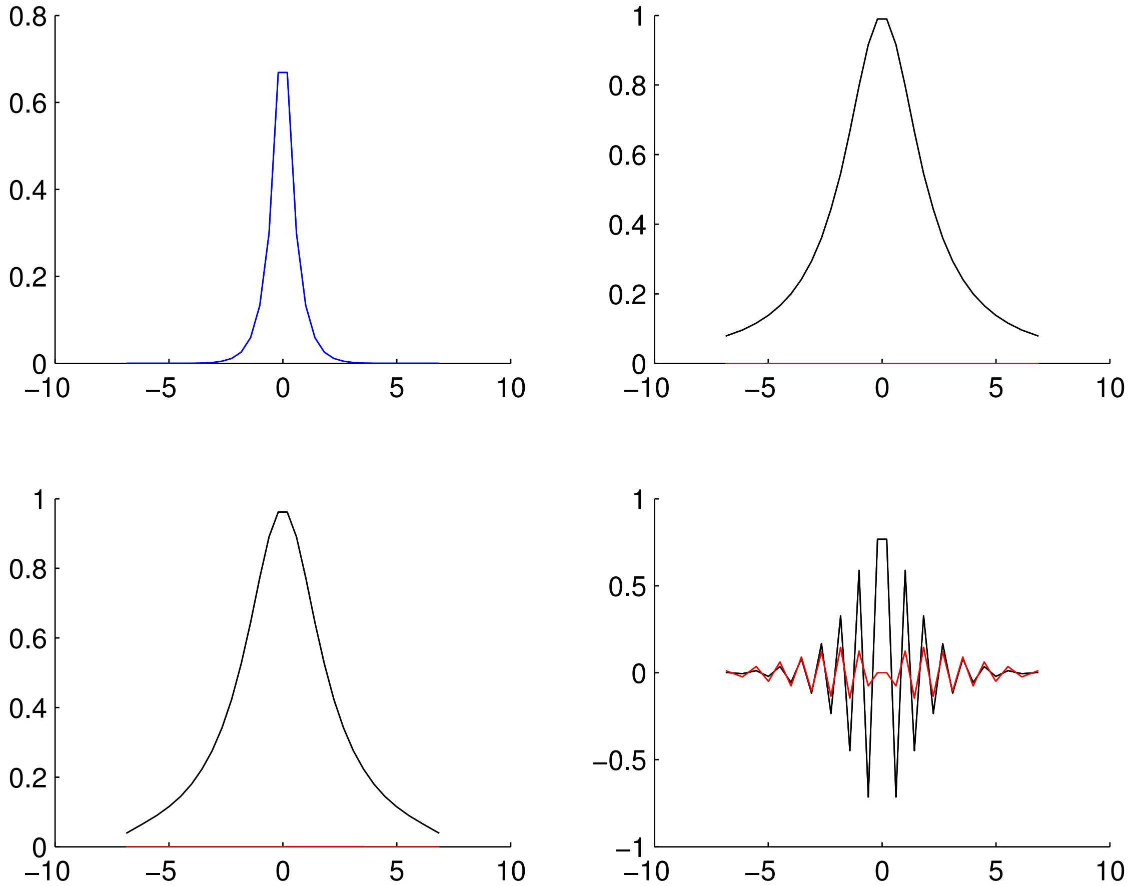
$\int_{-\infty}^{\infty} H_n(x) H_m(x) dx = 0$ for any n, m



$$H_1(x) = 2x$$
$$H_2(x) = 4x^2 - 2$$
$$H_3(x) = 8x^3 - 12x$$
$$H_4(x) = 16x^4 - 48x^2 + 12$$
$$H_5(x) = 32x^5 - 160x^3 + 120x$$

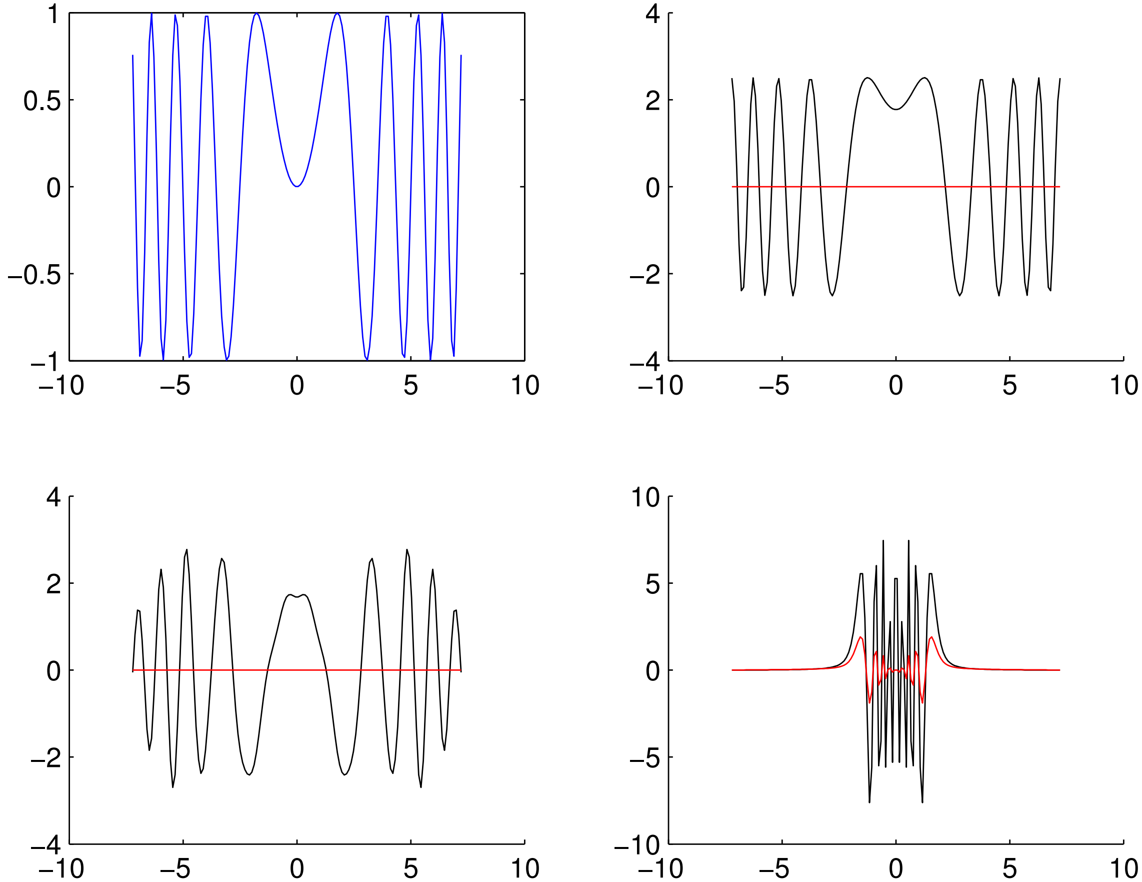
An example of the CFT

If $x(t) = e^{-2|t|}$ then $X(j\omega) = \int_{-\infty}^{\infty} e^{-2|t|} e^{-j\omega t} dt$=$X(j\omega) = \frac{4}{\omega^2+4}$



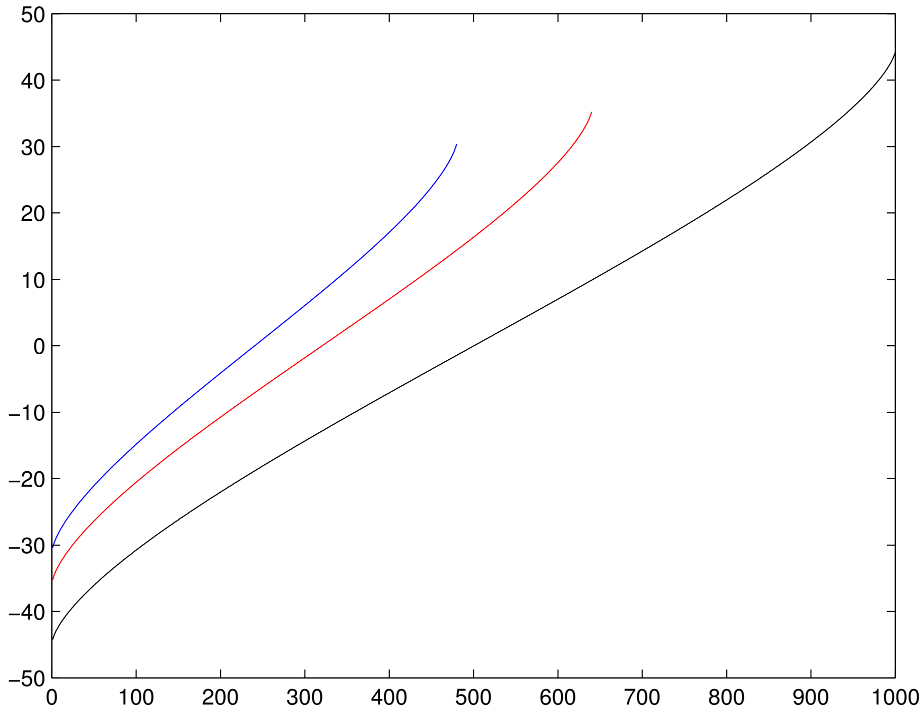
Another example of the CFT

If $x(t) = sin(\frac{t^2}{2})$ then $X(j\omega) = \pi[cos(\frac{\omega^2}{2}) + sin(\frac{\omega^2}{2})]$



Zeros of Hermite's Polynomials

In the Linear range, the zeros are equidistant
blue: 480 Zeros Red: 640 Zeros Black 1000 Zeros



We use the 140 roots of least magnitud of the Hermite's polynomial of degree 480

Short Time CFT for speech

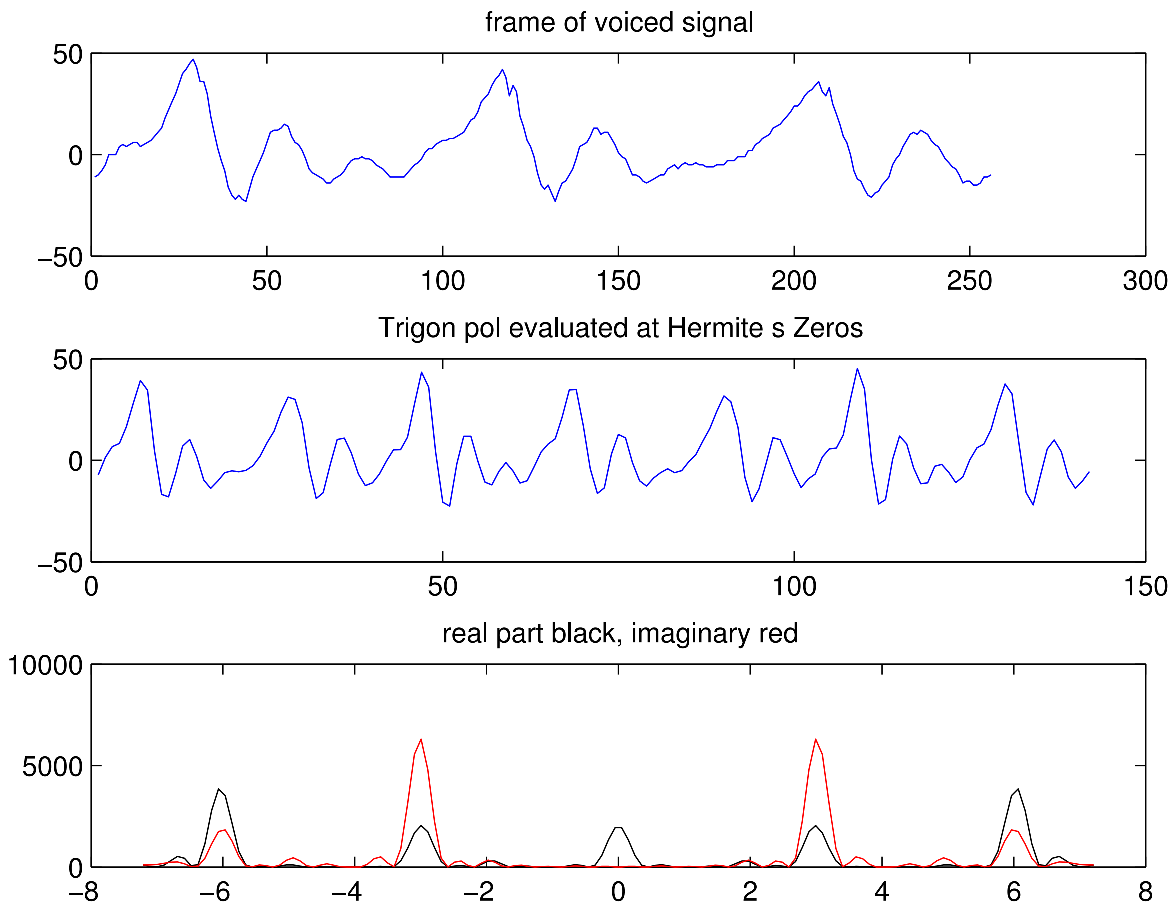
1: Fill vector $\hat{x}$ with the roots of the Hermite's polynomial of degree n
2: Construct the finite kernel matrix F using $\hat{x}$
3: **while** there are more frames **do**
4: Adjust the speech waveform to a trigonometric polynomial $p(x)$
5: Evaluate the trigonometric polynomial in the hermite's zeros $\hat{x}$, call this vector f ($f = p(\hat{x})$)
6: Compute $g = Ff$ (CFT for this frame)
7: **end while**

Some parameters and details

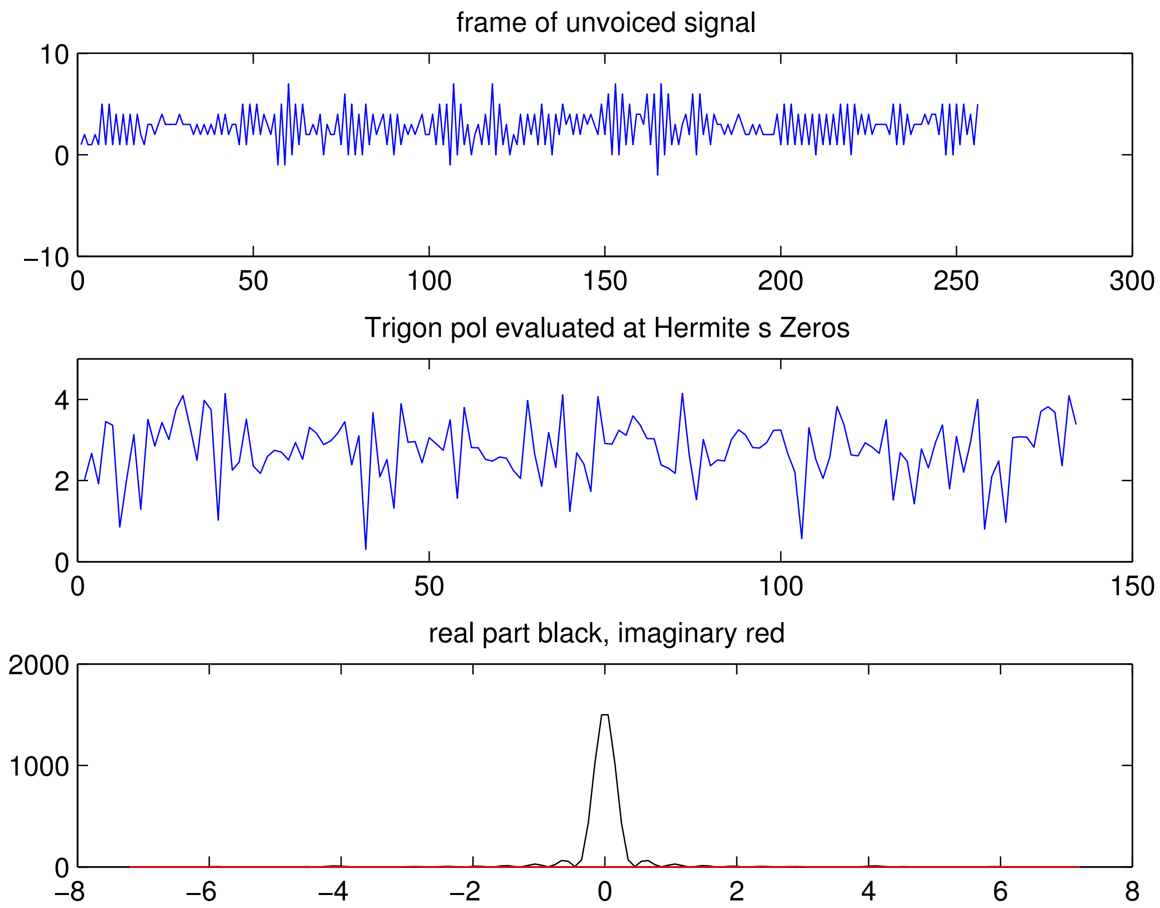
- Frames of 30 ms at a Sample rate of 8kHz (250 samples)
- overlapping of 50 % samplesize of 8 bits
- The signal is adjusted to the following trigonometric polynomial [1]:
  $\frac{a_0}{2} + \sum_{j=1}^{M} [a_j cos(jx) + b_j sin(jx)]$ Where:
  $a_j = \frac{2}{N} \sum_{k=1}^{N} [f(x_k) cos(jx_k)] \quad \forall \quad j = 0, 1, \ldots, M$
  $b_j = \frac{2}{N} \sum_{k=1}^{N} [f(x_k) sin(jx_k)] \quad \forall \quad j = 1, 2, \ldots, M$
  where $M < N/2$ (We used M=100)
- 140 zeros of least magnitude of Hermite's polynomial of degree 480

[1] Mathews, Jhon H & Fink, Kurtis D. *Numerical Methods with Matlab 3rd Edition* PH

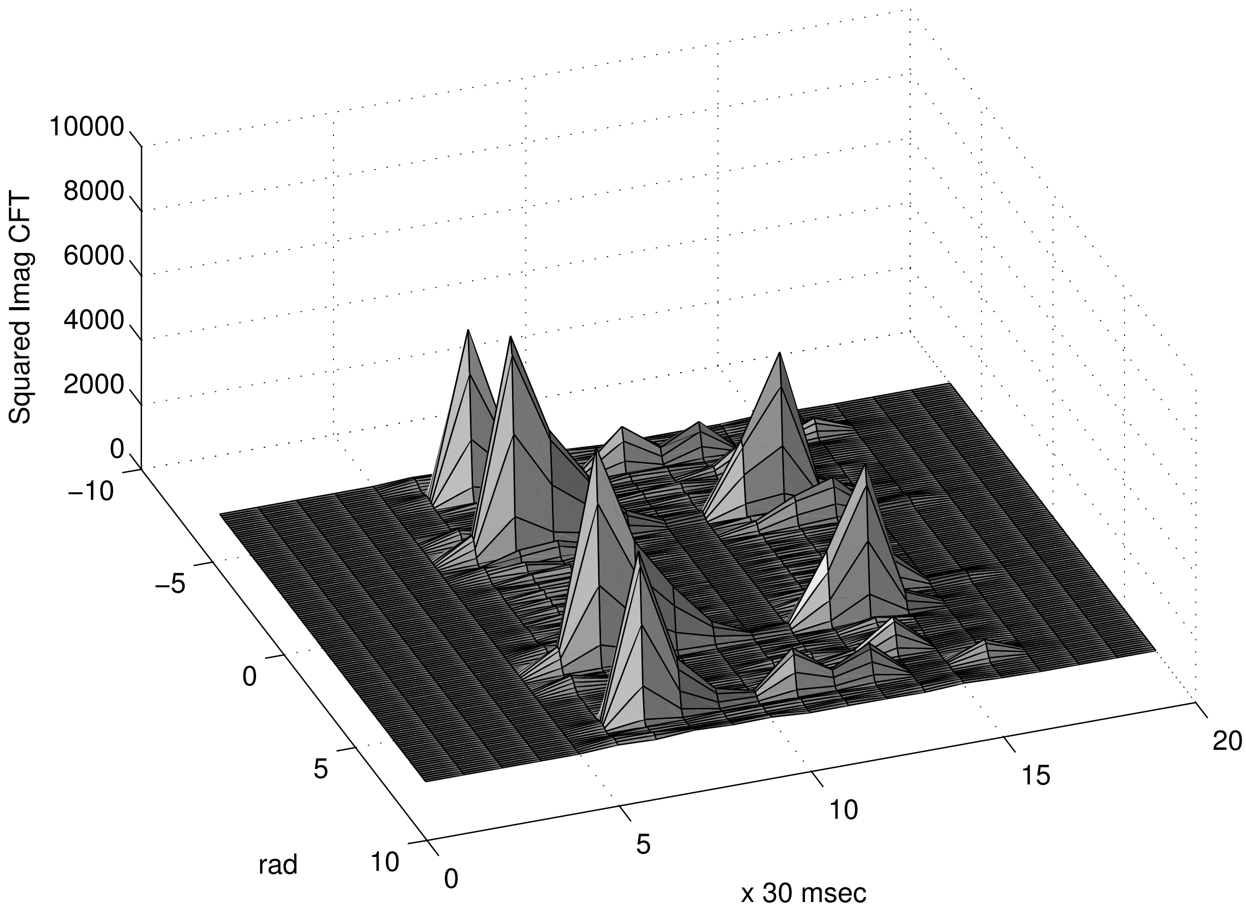
CFT for Voiced speech



CFT for Unvoiced speech

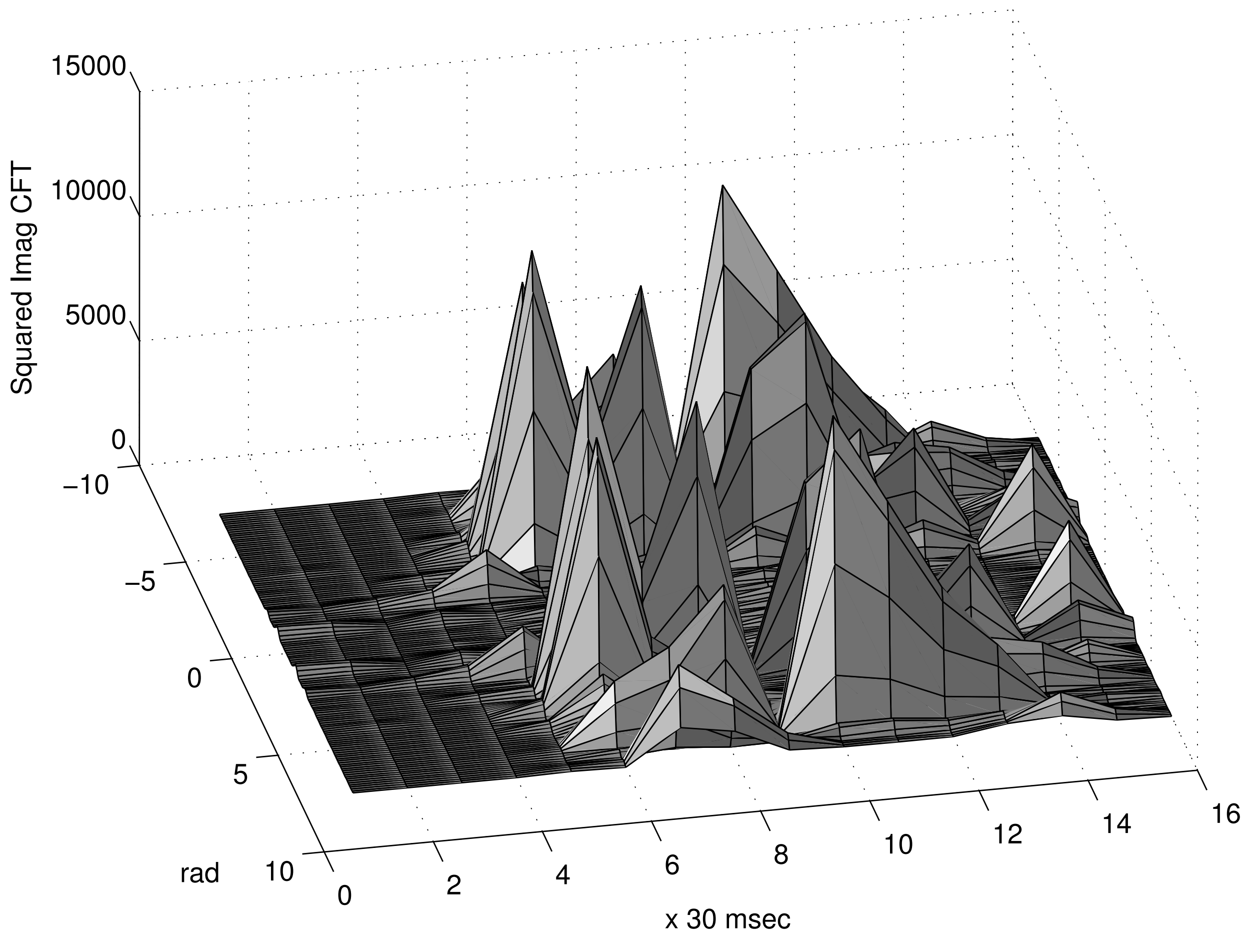


Square of the Imaginary part of the CFT



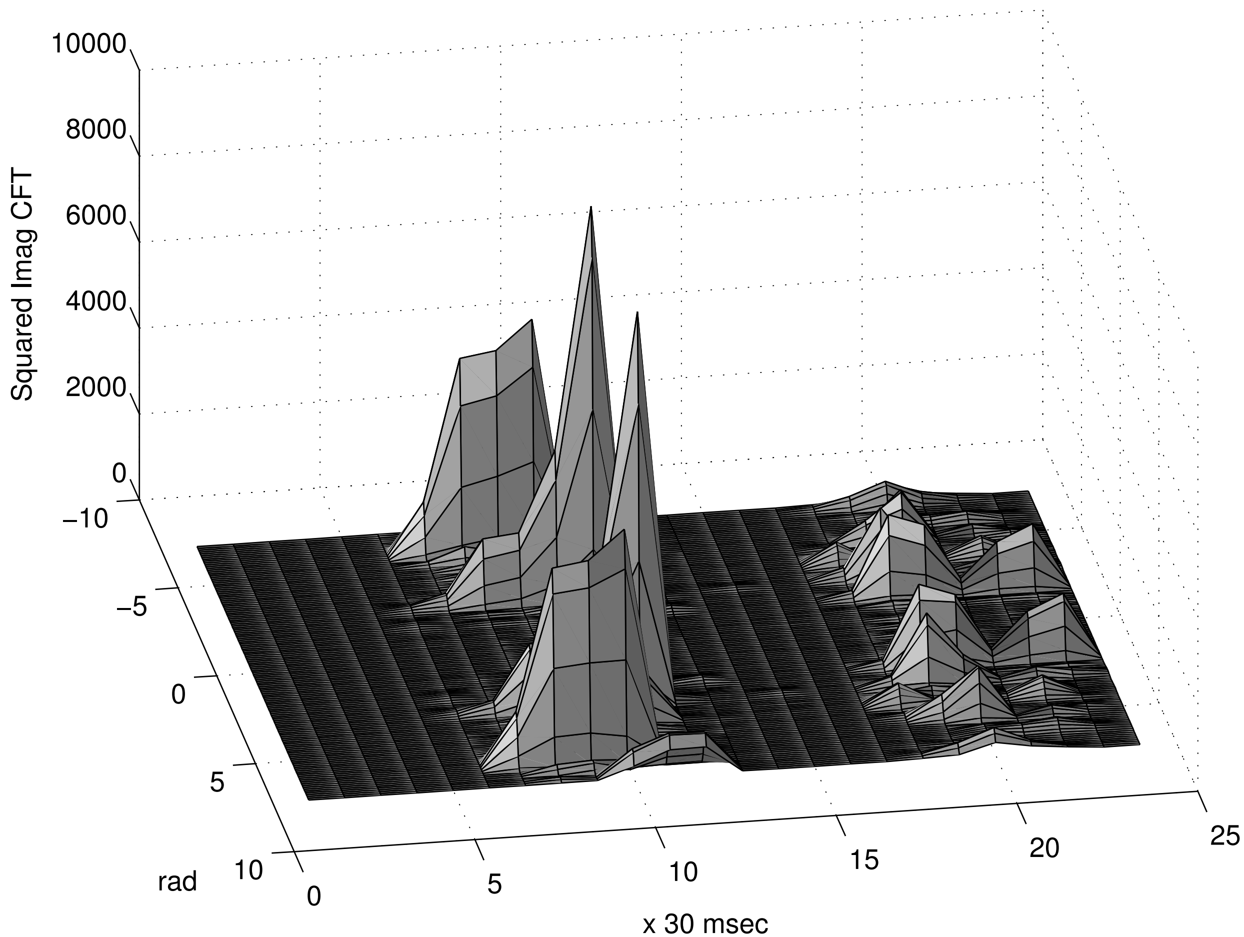
spanish word “seis”

Square of the Imaginary part of the CFT



Word “feo”

Square of the Imaginary part of the CFT



“ceja” (mexican pronuntiation)

Experiments

- Ten digits pronunced
- Classes considered: Voiced, unvoiced, silence
- VOICED.-Enough energy both in real and imaginary parts of the CFT
- UNVOICED.- Too little energy in the imaginary part but enough energy in the real part of the CFT
- SILENCE.- Too little energy both in real and imaginary parts of the CFT

	ZCR	E_n	Cepstrum	STAF	MSTAF	CFT
FAR	0.07	0.18	0.106	0.016	0	0
FRR	0.06	0.16	0.136	0.02	0	0

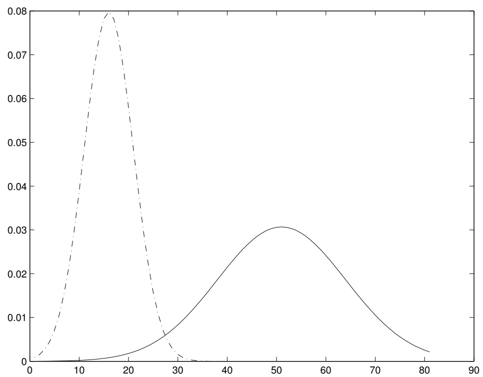
Conclusions and future work

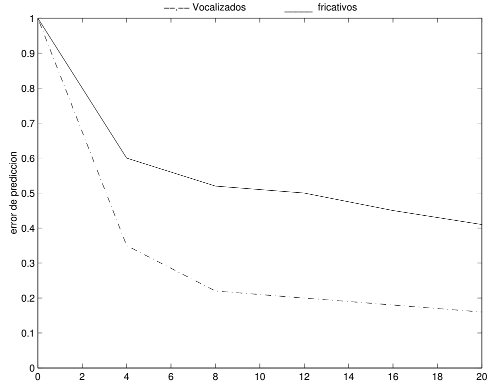
- CFT is as accurate as MSTAF
- Unlike MSTAF, CFT does not need data from the next frame of the signal
- In the future we intend to use the discretization of the CFT for Automatic Speech Recognition and for Individual Identification.
- compare with LPC and MFCC

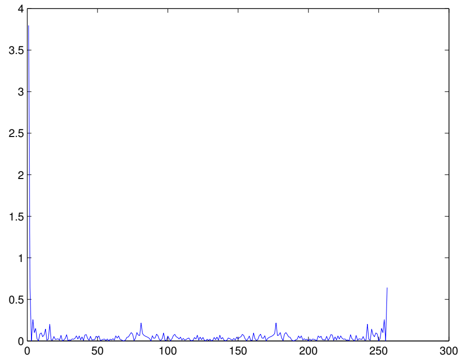
QUESTIONS?

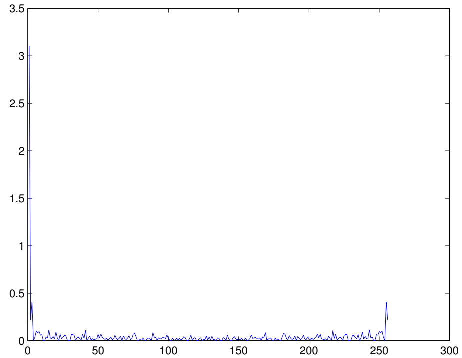
camarena@umich.mx MSc Antonio Camarena Ibarrola
elchavez@umich.mx Phd Edgar Chavez

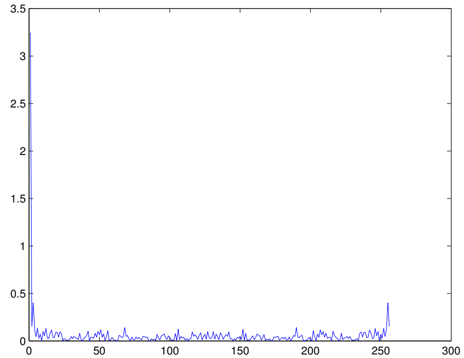
Related Work

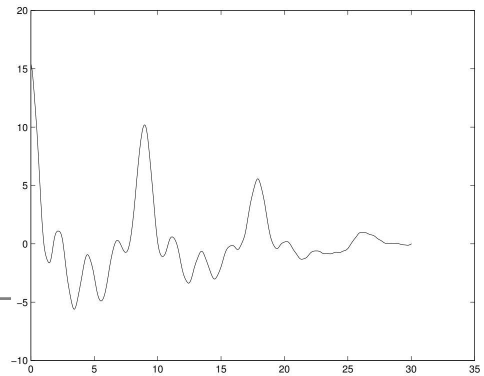


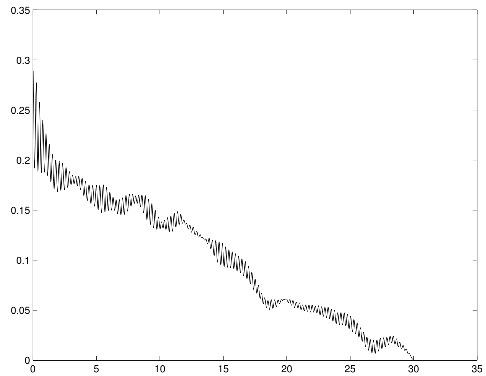




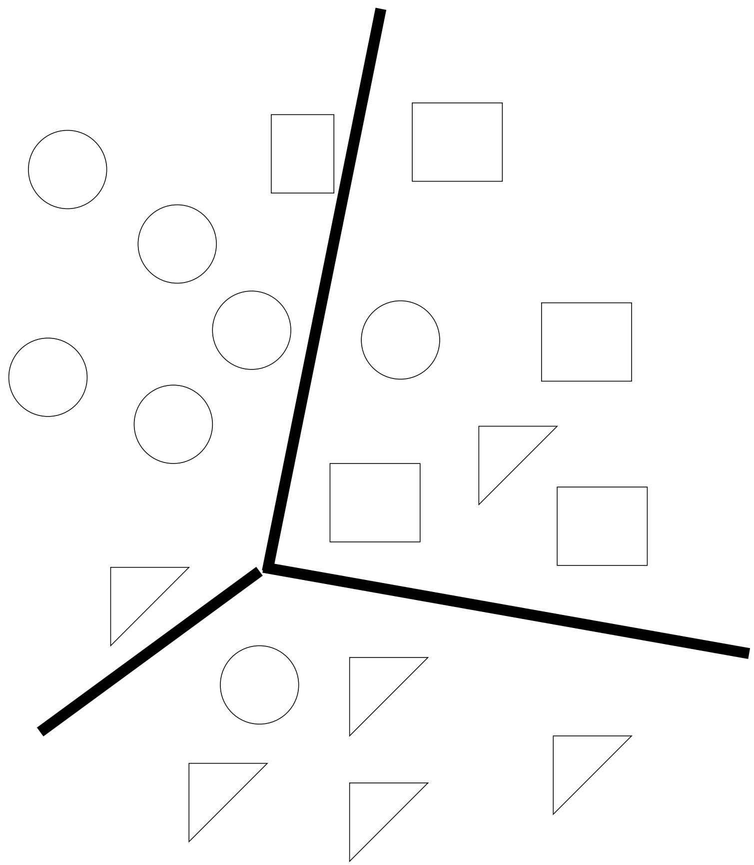








False Aceptance and False Rejection



$$FAR = \frac{2/11+2/13+1/12}{3} = 0.1397$$

$$FRR = \frac{2/7+1/5+2/6}{3} = 0.273$$

THANKS!!