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

Antonio Camarena-Ibarrola and Edgar Chávez
Universidad Michoacana de Sán Nicolás de Hidalgo
Ciudad Universitara, Edif ”B”, primer piso, CP 58004
Morelia, Mich., México
{camarena,elchavez}@umich.mx

Abstract

In Automatic Speech Recognition, Voice Synthesis, Speaker Identification and identifying laringeal diseases, it is critical to classify speech segments as voiced or unvoiced. Several techniques have been proposed for this issue during the last twenty years, unfortunately, they either have especial cases where the result is unreliable or need to use not only the present segment of speech but the next one as well, this fact limits its applications (i.e Continuos Speech recognition). In this paper we present an alternative to voiced/unvoiced classification using a Discretization of the Continuos Fourier Transform.

1 Introduction

A segment of the voice signal is known as voiced if the vocal cords vibrate during its production. This vibration introduces periodicity in the signal as you may observe in figure 1. If no such vibration exists the signal looks more like noise as seen in figure 2, nevertheless, its statistical properties are predictable (i.e It is a stationary signal).

The need for voiced/unvoiced discrimination emerges in almost any area where the speech signal is analyzed, for example the classical technique of speech synthesis uses an all pole dynamic model of the vocal tract whose parameters are the well known Linear Prediction Coefficients (LPC) [8], a set of these parameters are good for the production of a no longer than 30 ms segment of speech. However, as we can see in the production model depicted in figure 3, three additional parameters must be determined: The pitch, the Gain and of course the nature of the signal (voiced or unvoiced).

For the issue of finding out the nature of the signal, several approaches exist which make use of: the Zero Crossings Rate, the prediction error, the cepstrum, the short time autocorrelation function and the Modified Short time auto-

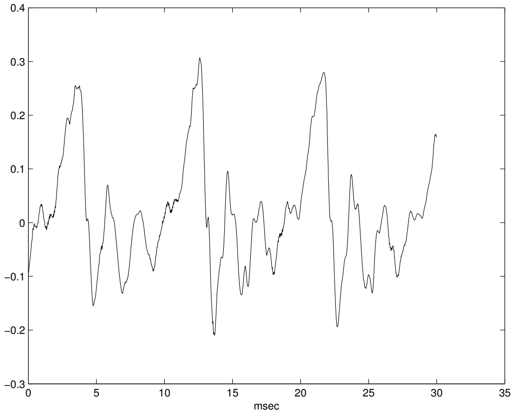

Figure 1. 30 milliseconds of voiced speech. Sound of the vowel “e” in spanish

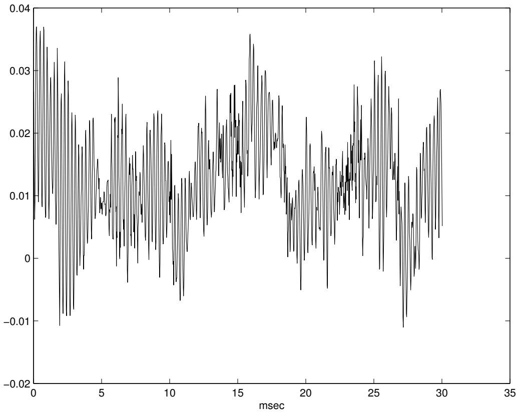

Figure 2. 30 miliseconds of unvoiced speech. Sound of the “s” in mexican pronunciation

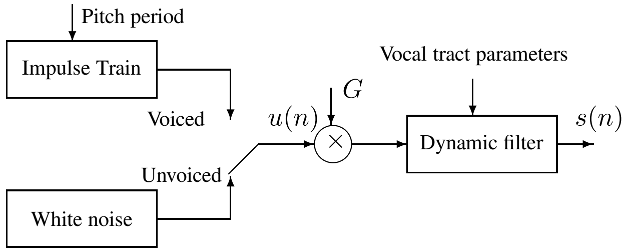

Figure 3. LPC based voice production model

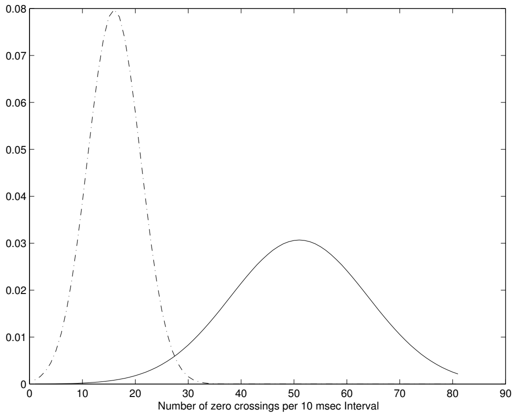

Figure 4. Distribution of zero crossings for unvoiced (continuos) and voiced (dashed) [8]

correlation function.

1.1 The Use of the Zero Crossings Rate (ZCR)

By looking at figures 1 and 2, one should thing that checking the number of times the speech signal crosses by zero should be enough to decide wether the speech is voiced or unvoiced. Unfortunately, as frequently dealing with the speech signal, things are not that easy. Figure 4 shows the distribution of both voiced and unvoiced sounds in relation to ZCR and we can clearly see that there is an overlapping region where we would not be certain about the nature of the signal. At the intersection (about 27 crossings by zero) the uncertainty would be of fifty percent!.

1.2 The Use of the Prediction Error

The short time prediction error E_n seems like a very easy way to make a decision on wether the segment is voiced or not. The idea behind using E_n is based on the fact that the LPC based model of the vocal tract is a very good one for voiced sounds and a poor one for fricatives (unvoiced) [8]. Figure 5 shows E_n Vs. the predictor's order p for a signal sampled at a rate of 10 KHz. E_n decreases as p increases both in voiced and unvoiced speech, but what matter to us is that E_n is always higher for fricatives than it is for voiced sounds. Once fixed the predictor's order one could simply establish a threshold and assume that if the error is below that, then we must be dealing with a voiced sound and unvoiced otherwise. E_n is defined in (1),(2) and (3) as the sum of the squares of the differences between the synthesized signal and the original signal samples.

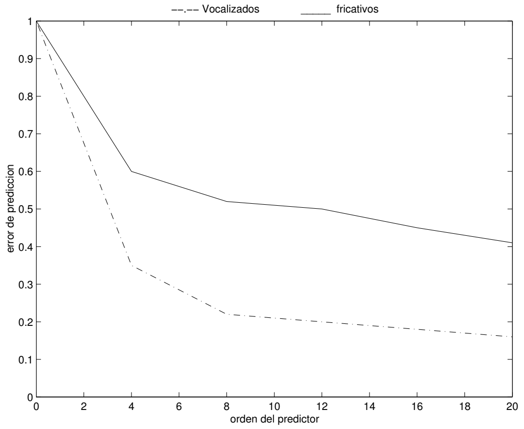

Figure 5. Prediction error E_n Vs. the predictor's order p for a signal sampled at 10 KHz [8]

$$E_n = \sum_m e_n^2(m) \tag{1}$$

$$E_n = \sum_m (s_n(m) - \hat{s}_n(m))^2 \tag{2}$$

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right]^2 \tag{3}$$

Where α_k is the k-th LPC coefficient.

In practice however, the LPC based synthesizer works surprisingly well for some unvoiced speech, figure 5 shows the average of the prediction error, but it does not give us any information about the variance. The idea of using the prediction error to make a decision on the nature of the speech signal by this means may not be a very good idea.

1.3 Use of the Cepstrum

The cepstrum is defined as the Inverse Discrete Fourier Transform (IDFT) of the magnitud logarithm of the Discrete Fourier Transform (DFT) of the signal. This kind of analysis is known as homomorphic and is considered the third option when one is choosing between time domain and frequency domain methods, in fact, the word "cepstrum" is intentionally an anagram of the word "spectrum".

For voiced speech there is a peak in the cepstrum at the fundamental period of the speech segment. No such peak appears in the cepstrum for unvoiced speech segments [8]. This property of the cepstrum can be used as a basis for determining wether a speech segment is voiced or unvoiced and for estimating the fundamental period of voiced speech [1]. The cepstrum peak is searched for a peak near the vicinity of the expected pitch period. If the cepstrum peak is above a pre-set threshold, the speech is likely to be voiced and the position of the peak is a good estimation of the pitch period. If the peak does not exceed the threshold the speech segment is likely to be unvoiced.

The presence of a strong peak in the cepstrum in the range 3-20 msec is a very strong indication that the input speech segment is voiced. However, the absence of a peak or the existence of a low level peak is not necessarily a strong indication that the input speech segment is unvoiced. That is, the strength and even the existence of a cepstral peak for voiced speech depends on a variety of factors including the length of the window applied to the input signal as well as the relative position of the window and the speech signal [8]. In conclusion, this method will leave us with uncertainty about the nature of the speech segment.

1.4 The use of the Short Time Autocorrelation and Modified Autocorrelation Functions

The autocorrelation function is defined as in (4) where $s(m)$ stands for the m-th sample of signal s. Note that $R(0)$ is the energy of the signal, that is why the global maxima of the autocorrelation function will always be at cero. The autocorrelation can be seen as a periodicity estimator. For a perfectly periodic signal, the period can be estimated by finding the location of the first maximum in the autocorrelation function. Of course speech is not strictly periodic, not even voiced segments.

To analyze a short segment of speech, the short time autocorrelation function $R_n(k)$ defined in (5) and (6) is used where w is the window function that selects the specific segment of speech in consideration and also attenuates both ends (except of course the rectangular window), s_n is the segment of signal already selected by the window and N is the size of the window which should be as small as posible but include at least two periods of the waveform.

$$R(k) = \sum_{m=-\infty}^{\infty} s(m)s(m+k) \tag{4}$$

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m+k)w(n-k-m) \tag{5}$$

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \tag{6}$$

The global maximum of $R_n(k)$, discarding $R_n(0)$ of course, should be the best estimation for the pitch period, as depicted in figure 6 where the autocorrelation of the voiced segment of speech depicted in figure 1 is shown, here the maximum is found at t=9 msec in agreement with the period of the waveform shown in figure 1. For unvoiced segments there are no strong autocorrelation periodicity peaks thus indicating a lack of periodicity in the waveform as can be seen in figure 7 where the autocorrelation of the unvoiced segment of speech depicted in figure 2 is shown.

The peaks of the short time autocorrelation function are reduced in amplitud as k increases because fewer data is involved in its computation as can be seen in figures 6 and 7, this undesirable effect was solved by the modified short time autocorrelation function (MSTAF) $\hat{R}_n(k)$ [8] defined in (7). The MSTAF is really a cross-correlation between two segments of speech since w_2 is a window that takes data from the next segment. Having to use the MSTAF is the first disadvantage of this method since affects modularity, it's always better to analyze a segment of speech without the need of the next segment specially in continuos speech recognition.

For the speech signal, the autocorrelation has many peaks, most of them are the result of oscillations of the vocal tract response which shapes the speech waveform. The autocorrelation retains too much of the information of the speech signal, in fact, the first 10 values are enough to estimate an LPC based model of the vocal tract. When trying only to decide whether the speech is voiced or fricative, the distracting features of the signal should be eliminated, this is what a "spectrum flattener" does, the most common technique for this issue is called "center clipping", it consists on zeroing the values of the speech signal whose magnitudes are below some threshold so that only prominent peaks are left. Now the autocorrelation can be computed and the periodicity peaks wont be confused with the peaks due to vocal tract oscillations. The full algorithm for determining the nature of the signal (voiced/unvoiced) and its pitch is according to [8], [4]:

1. The speech signal is sampled at 10 KHz and windowed into segments of 30 msec overlapped by 20 msec.

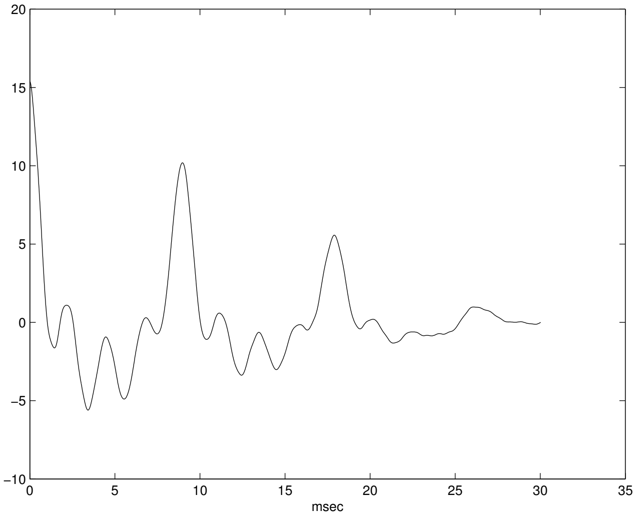

Figure 6. Autocorrelation of the voiced segment of speech shown in figure 1

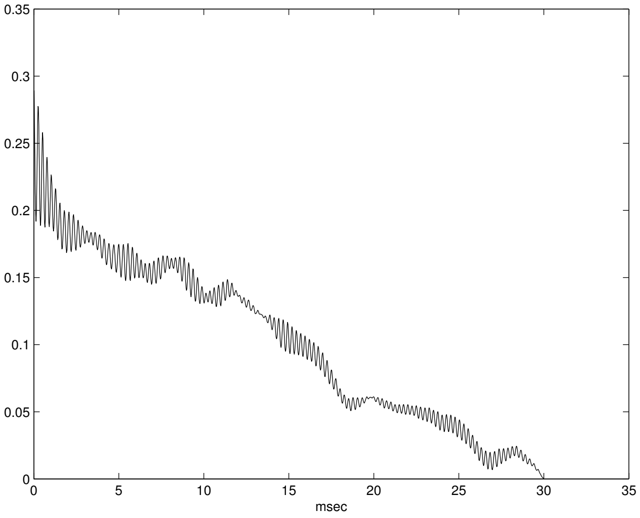

Figure 7. Autocorrelation of the unvoiced segment of speech shown in figure 2

2. Find the largest peak of the first and the last 10 msec of the segment.

3. set the clipping level as two thirds of the lowest of the two peaks determined in the preceding step.

4. Apply center clipping to the speech signal.

5. Compute the MSTAF.

6. Locate the largest peak of the autocorrelation function and compare it with $R_n(0)/3$, if the peak falls below this threshold classify the segment as unvoiced and voiced otherwise (the pitch period is the location of this peak).

$$\hat{R}_n(k) = \sum_{m=-\infty}^{\infty} s(m)w_1(n-m)s(m+k)w_2(n-k-m) \tag{7}$$

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

When using the DFT one is either assuming that the signal is periodic or accepting the fact that the transformation will take the whole signal as a single period of a waveform that is periodic, however, the speech signal is not periodic, specially when is fricative. The Continuos Fourier Transform (CFT) does not have such restriction, there is no need for applying the Hanning window or an equivalent. A discretization of the CFT was defined in [3], with this tool we found that discriminating voiced from unvoiced speech was rather trivial. The algorithm we implemented to find the discretization of the CFT is:

1. Convert the sequence $0, 1, 2, ..., N-1$ to N (the frame size) equidistant values from $-\pi$ to π

2. Find the coefficients of the trigonometric polinomial of degree $M < N/2$ given by (8) that best adjusts to the speech signal waveform using the formulas (9) and (10) to find a_j and b_j respectively [5].

$$\frac{a_0}{2} + \sum_{j=1}^{M} [a_j cos(jx) + b_j sin(jx)] \tag{8}$$

$$a_j = \frac{2}{N} \sum_{k=1}^{N} [f(x_k) cos(jx_k)] \quad \forall \quad j = 0, 1, \dots, M \tag{9}$$

$$b_j = \frac{2}{N}\sum_{k=1}^{N}[f(x_k)sin(jx_k)] \quad \forall \quad j = 1, 2, \ldots, M \tag{10}$$

3. Form vector x with the roots of the Hermite polynomial of degree P

4. Construct the Fourier's Kernel matrix F using equation (11) according to [2]. This Matrix is Hermitian so $F^{-1} = F^t$. Multiplying a signal vector by F is equivalent to finding a discretization of its CFT, multiplying by F^t would be a way of computing the inverse CFT.

$$F_{i,j} = \frac{\pi}{\sqrt{2n}}\sqrt[4]{\frac{4n+3-x_j^2}{4n+3-x_i^2}}[cos(x_i x_j) + jsin(x_i x_j)] \tag{11}$$

5. Evaluate the trigonometric polynomial found in step (2) in the hermite's zeros vector of step (3), call this vector f.

6. Compute $g = Ff$ so g will be the discretization of the CFT of f

Hermite's polynomial roots are not spaced at equal distances, however, for high degrees this roots have an interval where they are equidistant, then we chose to use only the zeros inside this interval which are the 170 roots of least magnitud of an Hermite's polynomial of degree 480.

In figure 11 the real and imaginary parts of the CFT of the unvoiced sound of the "j" is shown, observe how the Real part predominates over the imaginary part, the same can be concluded by looking at figure 8 where the real and imaginary parts of the CFT of the unvoiced sound of the "s" is shown. On the other hand, you may observe in figures 9 and in 10 how neither the real nor the imaginary part is predominant for the voiced sounds of "a" and "m" respectively. Based on this observations the simple proposed way of classifying a short speech segment as voiced or unvoiced is to check wether the average power of the imaginary part of the CFT falls below a threshold or not, if it does, then the segment is declared unvoiced and voiced otherwise.

3 Experiments

For our experiments we used a trigonometric polynomial of degree 100 ($M = 100$) since using a lower degree modified the waveforms of unvoiced speech signals, as for the other parameters we used a Hermite's polynomial of degree 480 (P=480) but used only 170 of its roots as indicated in the preceding section. The length of the speech frames

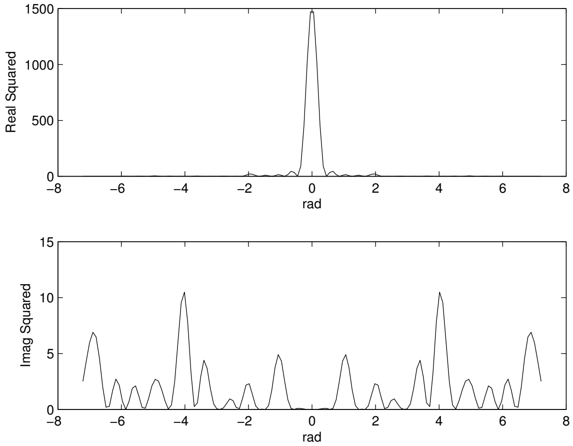

Figure 8. Top: Square of the real part of the CFT of a segment of the unvoiced speech sound of the "s". Bottom: Square of the imaginary part of the CFT of the same signal.

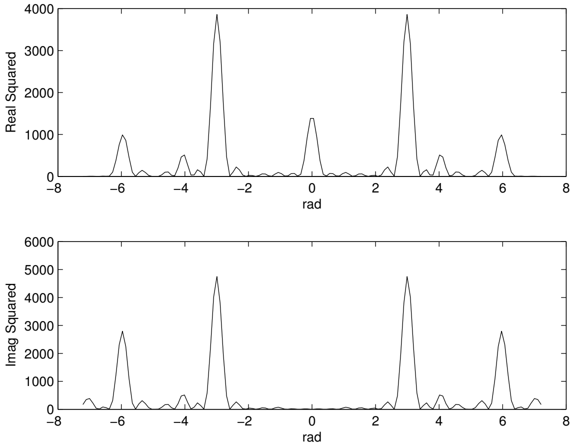

Figure 9. Top: Square of the real part of the CFT of a segment of the voiced speech sound of the spanish "a". Bottom: square of the imaginary part of the CFT of the same speech signal.

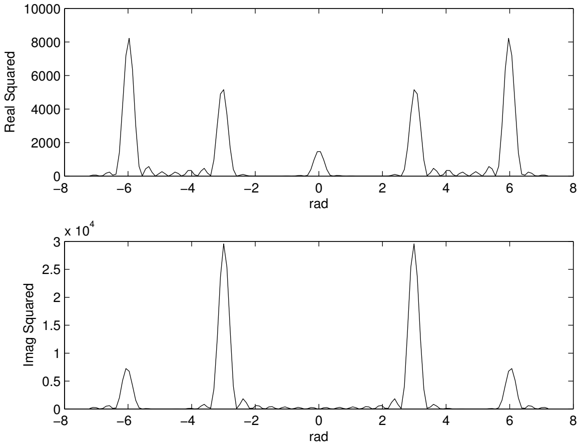

Figure 10. Top: Square of the real part of the CFT of a segment of the voiced speech sound of the "m". Bottom: square of the imaginary part of the CFT of the same speech signal.

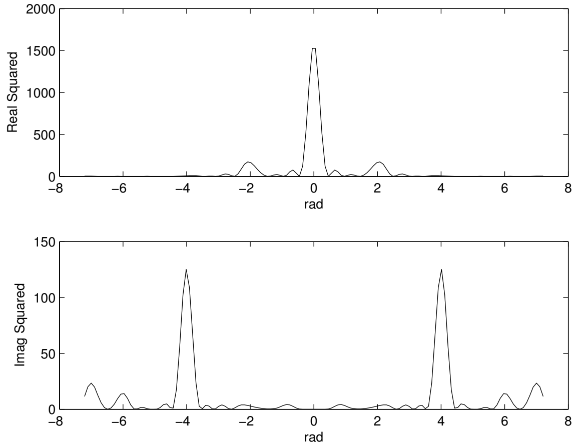

Figure 11. Top: square of the real part of the CFT of a segment of the unvoiced speech sound of the spanish "j". Bottom: squared of the imaginary part of the CFT of the same signal.

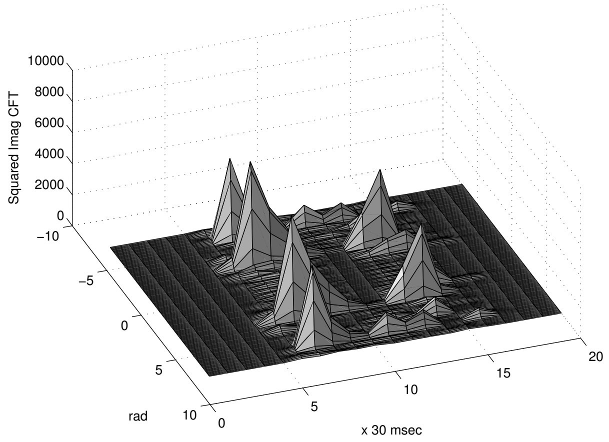

Figure 12. Twenty segments of the spanish word "seis", for each segment, the square of the imaginary part of the CFT is plotted.

were 256 samples ($N = 256$) with overlap of fifty percent, No window was applied to the frames (it is one of the advantages of the algorithm) and the sampling was made for telephonic quality (8 samples per second, 8 bits per sample, mono-aural).

In figure 12, the first five 30 msec segments would be declared unvoiced, then the next 10 segments are voiced and the last five are unvoiced again, the pronounced word is "seis". Figure 13 is the square of the imaginary part of the CFT of the word "feo" there the first four segments are fricative and the rest are voiced. In figure 14 the same was done for the word "ceja" with mexican pronunciation and we can clearly see that the first six segments are unvoiced, then then next six are voiced, then unvoiced again for five more segments and the last seven segments are voiced.

To make a false positive and negative analysis and determine the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), the ten digits in spanish and mexican pronunciation were recorded in a quiet room ("cero", "uno",...,"nueve"), then each 10ms frame was carefully classified by hand as voiced, unvoiced or silence. The threshold used in all v/u classification methods tested was modified until it reached a balanced optimum between the FAR and the FRR, for example a threshold that correctly recognizes all the voiced frames making FRR equal to zero might take some unvoiced segment as voiced increasing the FAR, this optimization of the threshold for each considered v/u classification method was also hand made. Silent frames were detected measuring the short time energy (remember that it was a quiet room) and so the v/u classification methods only had to decide upon speech content signal frames.

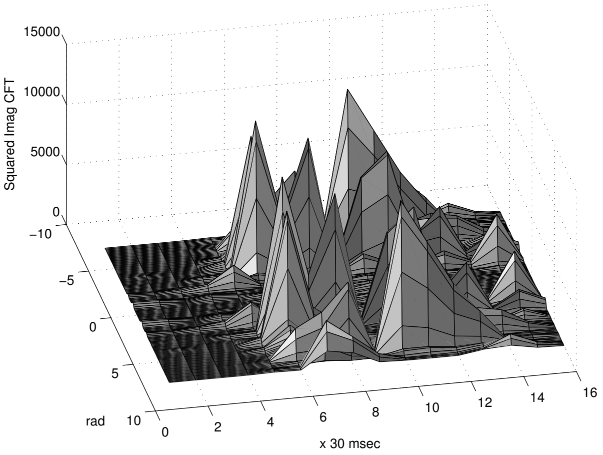

Figure 13. Sixteen segments of the spanish word "feo", for each segment, the square of the imaginary part of the CFT is plotted.

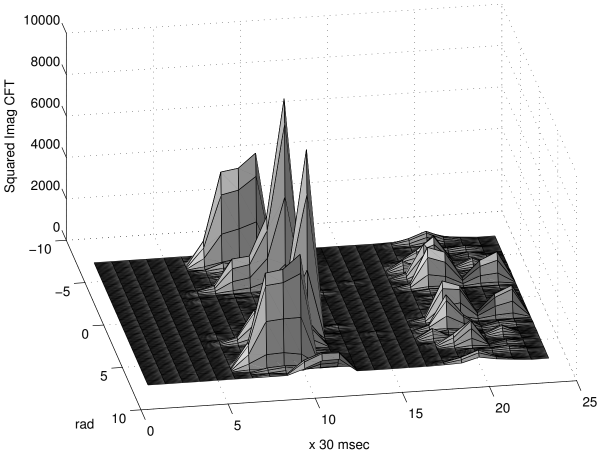

Figure 14. Twenty five segments of the spanish word "ceja" (mexican pronuntiation). For each segment, the square of the imaginary part of the CFT is plotted.

Table 1. False Positive and Negative Analysis

	ZCR	E_n	Ceps	R_n	MSTAF	CFT
FAR	0.07	0.18	0.106	0.016	0	0
FRR	0.06	0.16	0.136	0.02	0	0

The results are shown in table 1.

4 Conclusions and Future Work

A Method for classifying a speech signal segment as voiced or unvoiced was proposed which is not restricted since it makes no assumptions on the nature of the signal. The methods that use ZCR, Cepstrum and Prediction error are not guaranteed to work for all cases. The use of the MSTAF method is not restricted either, however it requieres the use of the next segment of speech in order to work well, this requirement complicates things for some applications, specially in continuos speech recognition. The False positive and negative analysis confirmed the expectations. In the future, we will try the use of the Discretization of the CFT in the issue of identifying individuals by its voice.

References

[1] Ahmadi, Sassan & Spanias, Andreas S. *Cepstrum-Based Pitch detection using a new statistical v/UV Classification Algorithm*. IEEE Transactions on speech and audio processing, vol 7, No 3, pp 333-338. May 1999

[2] Campos, R. G.*A Quadrature Formula for the Hankel Transform*. Numerical Algorithms, Volume 9, No. 3-4, pp 343-354. 1995

[3] Campos, R. G. & Juarez, L. Z. *A Discretization of the Continuos Fourier Transform*. Nuovo Cimento 107 B, pp 703-711. 1992

[4] Kunieda, Nobuyuky, Shimamura, Tetsuya & Suzuky Jouji *Characteristics of pitch extraction by ACLOS (Autocorrelation of LOg Spectrum)* The Journal of the Acoustical Society of America. Vol 100, Issue 4, pp 2602-2603. October 1996

[5] Mathews, Jhon H & Fink, Kurtis D. *Numerical Methods with Matlab 3rd Edition*. Prentice Hall, ISBN 84-8322-181-0. 2000

[6] Mitev, Petar & Hadjitodorov *Fundamental frequency estimation of voice of patients with Laryngeal disorders* Information Sciences-Informatics and Computer

Science: An International Journal. Volume 156, issue 1-2, pp 3-19 Special issue: Spoken language analysis, modeling and recognition. statistical and adaptive coneccionist approaches. November 2003

[7] Pineda, Luis; Villaseñor, Luis; Cuétara, Javier; Castellanos, Hayde & Lpez Ivonne. *DIMEx100: A new phonetic and speech corpus for Mexican Spanish*. Instituto de Investigaciones en Matemticas Aplicadas y en sistemas (IIMAS, UNAM). 2005

[8] Rabiner, Lawrence R. & Schafer, Ronald W. *Digital Processing of speech signals* Laboratorios Bell. Prentice Hall 1978

[9] Wildermoth, Brett R. & Paliwal, Kuldip K. *Use of Voicing and pitch information for speaker recognition*. Proc 8th Australian International Conf. Speech science