International Journal of Advanced Science and Technology
Vol.60, (2013), pp.67-82
http://dx.doi.org/10.14257/ijast.2013.60.07

Automatic Detection of Transition Zones in Tunisian Dialect

Nefissa Annabi-Elkadri
LIPAH, Computer Science Department, Faculty of Sciences of Tunis,
Tunis El Manar University, Tunis, Tunisia.
nefissa_annabi@yahoo.fr
https://sites.google.com/site/nefissaannabielkadri/

Abstract

This study is an extension of our last researches about the detection of transition zones based on multiresolution spectral analysis (MRS). In this paper we present the fourth step for the realization of an automatic system for Tunisian Dialect segmentation and analysis. The MRS is calculated over several Fast Fourier Transforms (FFT) of different length. It can provide a higher temporal accuracy in the upper spectral region and a better frequency resolution in the lower spectral range. We showcase the importance of this tool by attempting to an automatic transition zones detection by calculating the Interquartile Range (IQR) of all frames of the MRS FFT. We applied our Visual Assistance of Speech Processing (VASP) System to a corpus. This corpus was in Tunisian Dialect pronounced by tunisian speakers. The analysis of the obtained results shows that the automatic detection of transition zones based on MRS provides better results compared to classical spectral analysis of the corpora used.

Keywords: *Multiresolution Spectral Analysis, IQR, Automatic Detection of Transition Zones.*

1 Introduction

The classic speech sonagram offers a single integration time which is the length of the window. It implements a uniform bandpass filter, the spectral samples are regularly spaced and correspond to equal bandwidths. The choice of the window length determines the time-frequency resolution for all frequencies of the sonagram. Choosing an appropriate window length for spectral analysis is not a straight forward process. A narrow window provides a low frequency resolution, approximating only roughly the spectral envelope, whereas a wider window provides a high frequency resolution and can even show the harmonics in the spectrum. The drawback of analysing a greater part of the signal can lead, however, to a lower temporal resolution, thus masking or distorting rapid acoustic landmarks occurring in speech. Ladefoged [21] suggests using a wide window for long steady-state vowels and a narrow window when investigating stop bursts in which the higher frequencies are more important.

ISSN: 2005-4238 IJAST

Mallat [23, p.674] makes the remark that *"it is difficult to analyze the information content of an image directly from the gray-level intensity of the image pixels... Generally, the structures we want to recognize have very different sizes. Hence, it is not possible to define a priori an optimal resolution for analyzing images."*. To improve the standard spectral output, we can calculate a multiresolution (MR) spectrum. In the original papers, the MR analysis is based on discrete wavelet transforms [18,23–25]. It has since been applied to several domains: image analysis [23], time-frequency analysis [13], speech enhancement [16,26], automatic signal segmentation by search of stationary areas from the scalogram [22].

The MR spectrum which is a compromise that provides both a higher frequency and a higher temporal resolution, is not a new method. In phonetic analysis, [1, 2] presents a study of two common vowels [a] and [ɛ] in Tunisian dialect and French language. Vowels are pronounced in Tunisian context. The analysis of the obtained results shows that due to the influence of French language on the Tunisian dialect, the vowels [a] and [ɛ] are, in some contexts, similarly pronounced. An extension of this study [5] presented and tested the concept of multi-resolution for the spectral analysis (MRS) of six common vowels [a], [ɛ], [i], [e], [u] and [o] in Tunisian words and in French words under the Tunisian context. [5] applied the MRS method to the signal, at the first step and applied the multi-resolution LPC [1] for the formant detection at the second step. The analysis of the obtained results shows that the vowels [a], [ɛ], [i], [e], [u] and [o] are, in some contexts, similarly pronounced in French language and in Tunisian dialect.

Annabi-Elkadri [4] applied the MRS for an automatic method for Silence/Sonorant/Non-Sonorant detection used the ANOVA method. Results compared the classical methods for classifications such as Standard Deviation and Mean with ANOVA which was better. The method for automatic Silence/Sonorant/Non-Sonorant detection based on MRS provides better results compared to classical spectral analysis. Annabi-Elkadri [3] presented an automatic method for the detection of transition zones based on multiresolution spectral analysis (MRS) by calculating the Interquartile Range (IQR) of each frame of the MRS FFT. The analysis of the obtained results showed that the automatic detection of transition zones based on MRS provides better results compared to classical spectral analysis of the corpora used.

Cheung [11] presents a method for combining a wideband and a narrowband spectrogram by evaluating the geometric mean of their corresponding pixel values. The combined spectrogram appears to preserve the visual features associated with high resolution in both frequency and time. Chan and al. [9] describe an approach of using MR for clean connected speech and noisy phone conversation speech. Their experiments showed that MR cepstra result in a significantly lower number of errors when compared to Mel-frequency cepstral coefficients. For music signals, Cancela and al. [8] present two algorithms, the efficient constant-Q transform and the MR Fast Fourier transform (FFT). These are reviewed and compared to a new proposal based on the Infinite Impulse Response filtering of the FFT. The proposed method appears to be a good compromise between design flexibility and reduced computational effort. Additionally, MR FFT has been used as a part of an effective melody extraction algorithm. In this context, Dressler [15] advances a melody extraction algorithm based on an MR FFT whose aim is to extract the sinusoidal components of the audio signal. The calculation of spectra of different frequency resolutions is executed so that sinusoids that are stable over different frames of the FFT can be detected. The results showed that the MR analysis improves the extraction of the sinusoidal. The MRS has also been used in speech enhancement [28] and speech synthesis [12].

This study aims to extend the researche of [1–3, 5]. We want to realize an automatic system for segmentation and analysis of the Tunisian Dialect. At the first step, we presented the concept of multi-resolution for the spectral analysis (MRS) [1–3]. At the second step, we tested our method to analyse vowels in Tunisian words and in French words under the Tunisian context [1, 2, 5]. At the third step, we proved that the MRS can be used for an automatic method for the detection of transition zones. Our corpus was produced by 19 native French speakers [20]. This corpus was in French pronounced by French speakers and has the format C_iVC_iV where C_i was a stop consonant [p t k] and V was a vowel [i e]. We wish to extend this research to the Tunisian dialect corpus pronounced by Tunisian speakers.

2 History of Tunisian Dialect and its relationship with Arabic and French

The official language in Tunisia is Arabic. But, the popular language is the Tunisian Dialect (TD). It is a mix of Arabic with a lot of other languages: French, Italian, English, Turkich, German, Berber and Spanish. This mixture is related to the history of Tunisia, since it was invaded and colonized by many civilizations like the Romans, Vandals, Byzantains, the Arab Moslems and French.

After French colonization, the French government wanted to spread the French language in the country. The French instituted a bilingual education system with the Franco-Arab schools. Programs of bilingual schools were modeled primarily on the model of French primary education for children of European origin (French, Italian and Maltese), which were added courses in colloquial Arabic. As for the Tunisian children, they received their education in classical Arabic in order to study the Quoran. Only a small Tunisian elite received a truly bilingual education, in order to co-administer the country. Tunisian Muslim mass continued to speak only Arabic or one of its many varieties. The report of the Tunisian Minister of Affairs, Jean-Jules Jusserand, pursuing the logic of Jules Ferry. In a "Note on Education in Tunisia", dated February 1882, Jusserand exposing his ideas: "*We have not at this time we better way to assimilate the Arabs of Tunisia, to the extent that is possible, that they learn our language, it is the opinion of all who know them best: we can not rely on religion to make this comparison, they do not convert to Christianity, but as they learn our language, a host of European ideas will prove to be bound to them, as experience has sufficiently demonstrated. In the reorganization of Tunisia, a large part must be made to education*".

After independence, education of the French language such as Arabic was required for all Tunisian children in primary school. This explains why French has become the second language in Tunisia. It is spoken by the majority of the population.

There are different varieties of TD depending on the region, such as dialect of Tunis, Sahel, Sfax, etc. Its morphology, syntax, pronunciation and vocabulary are quite different from the Arabic [27]. There are several differences in pronunciation between Standard Arabic and TD. Short vowels are frequently omitted, especially where they would occur as the final element of an open syllable. While Standard Arabic can have only one consonant at the beginning of a syllable, after which a vowel must follow, TD commonly has two consonants in the onset. For example Standard Arabic "book" is $/kita?b/$, while in TD, it is $/kta?b/$. The nucleus in TD may contain a short or long vowel, and at the end of the syllable, in the coda, it may have up to three consonants, but in standard Arabic, we cannot have more than two consonants at the end of the syllable. Word-internal syllables are generally heavy in that they either have a long vowel in the nucleus or consonant in the coda. Non-final syllables composed of just a consonant and a short vowel (i.e. light syllables) are very rare in TD, and are generally loaned from standard Arabic: short vowels in this position have generally been lost, resulting in the many initial CC clusters. For example $/?awa?b/$ "reply" is a loan from Standard Arabic, but the same word has the natural development $/?wa?b/$, which is the usual word for "letter" [17].

In TD's non-pharyngealised context, there is a strong fronting and closing of $/a?/$, which, especially among younger speakers in Tunis can reach as far as $/e?/$, and to a lesser extent of $/a/$.

This is an example of Tunisian Arabic sentence (SAMPA and X-SAMPA symbols): '$/ddZ \quad bA:k \quad Ukil\, ?\backslash adailkas\, mta?\backslash u \quad milEna\, bil \quad lE \quad sykse/$'. This sentence is a mixture of three languages; '$/ddZ \quad bA:k/$' in English, '$/Ukil \quad ?\backslash ada \quad ilkas\; mta?\backslash u\; milEna \quad bil/$' in Tunisian Arabic, which means '*as usual the show is interesting*' and finally '$/lE \quad sykse/$' in French, which means '*success*'.

3 Multiresolution FFT

It's so difficult to choose the ideal window with the ideal characteristics. The size of the ideal window [6] was equal to twice the length of the pitch of the signal. A wider window shows the harmonics in the spectrum, a shorter window approximated very roughly the spectral envelope. This amounts to estimate the energy dispersion with the least error. When we calculated the windowed FFT, we supposed that the eneregy was concentrated at the center of the frame [19, p.41]. We noted the center C_p. So our problem now, is the estimation of C_p.

3.1 The center estimation in the case of the Discrete Fourier Transform (DFT)

We would like to calculate the spectrum of the speech signal s. We note L the length of s. The first step is to sample s into frames. The size of each frame was between 10 ms and 20 ms [7, 21] to meet the stationnarity condition. We choose the Hamming window and we fixed the size to 512 points and the overlap to 50%. Figure 1 shows the principle of the center estimation.

For each frame p, the center C_p was estimated:

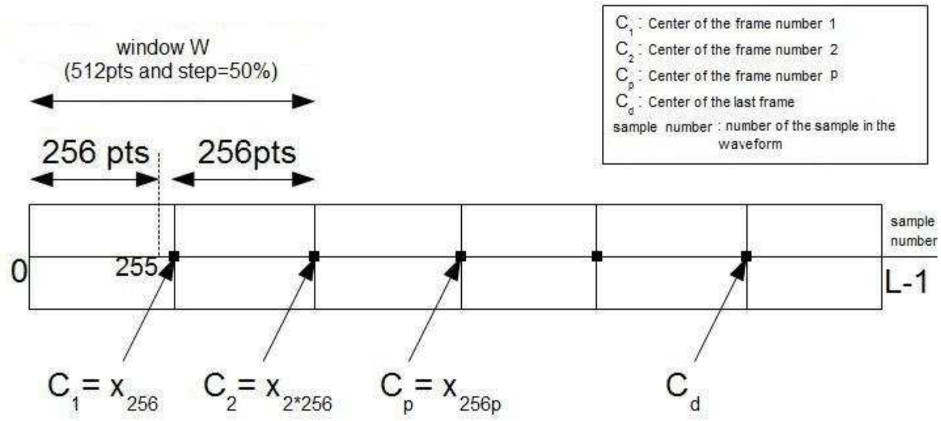

window W
(512pts and step=50%)
256 pts
256pts
0
255
L-1
sample number
$C_1 = x_{256}$
$C_2 = x_{2*256}$
$C_p = x_{256p}$
C_d
C_1 : Center of the frame number 1
C_2 : Center of the frame number 2
C_p : Center of the frame number p
C_d : Center of the last frame
sample number : number of the sample in the waveform

Figure 1. Signal sampling and windowing for center estimation C_p. The window length $N = 512$ points and overlap = 50%.

$$\begin{cases} C_1 & = & x_{256} & for \quad p = 1 \\ C_2 & = & x_{2*256} & for \quad p = 2 \\ & \vdots & & \\ C_p & = & x_{256p} & in \quad general \quad case \end{cases}$$

The center $C_p = x_{256p}$ with $p = 1...[\frac{L-1}{256} - 1]$ and [ ] the integer part.

Each signal s was sampled into frames. Each frame number p was composed by $N = 512$ points:

$$\begin{cases} s_0(p) & = & x_{256(p-1)} \\ s_1(p) & = & x_{256(p-1)+1} \\ & \vdots & \\ s_{511}(p) & = & x_{256(p-1)+511} \end{cases}$$

In general case, for the component number l of s:

$$s_l(p) = x_{256(p-1)+l}$$

The FFT windowing for the frame number p was calculated as:

$$S_k(p) = \sum_{l=0}^{511} s_l(p) e^{-\frac{2j\pi kl}{512}} w(s_l(p) - s_{256}(p)) \tag{0}$$

In general case:

$$S_k(p) = \sum_{l=0}^{N-1} s_l(p) e^{-\frac{2j\pi kl}{N}} w(s_l(p) - s_{\frac{N}{2}}(p)) \tag{1}$$

We noted $C_p = s_{\frac{N}{2}}(p)$ the center of the frame number p with $p = 1...[\frac{2(L-1)}{N} - 1]$, [ ] the integer part and :

$s_l(p)$: the component of s number l of the frame p
$S_k(p)$: the component of S number k of the frame p
L: the length of the signal s
N: the length of the window w

3.2 The center estimation in the case of the MRS FFT

To improve the standard spectrum, we calculated the MRS FFT by combining several FFT of different lengths. The temporal accuracy is higher in the high frequency region and the resolution of high frequency in the low frequencies.

We calculated the FFT windowing of the signal several times. The number of steps NB was equal to the number of band frequencies fixed a priori. For each step number i ($i \leq NB$), the signal s was sampled into frames $s_i(p_i)$ and windowed with the window w. We noted N_i the length of frames and of w for each step i. C_{i,p_i} was the center of w.

The spectrum $S_{i,k}(p_i)$ for each step i was:

$$S_{i,k}(p_i) = \sum_{l=0}^{N_i-1} s_{i,l}(p_i) e^{-\frac{2j\pi kl}{N}} w(s_{i,l}(p_i) - C_{i,p_i}) \tag{2}$$

with: $C_{i,p_i} = s_{i,\frac{N_i}{2}}(p_i)$ the center of the frame p_i when the overlap=$\frac{N_i}{2}$.

In MRS, the overlap $\frac{N_i}{2}$ can not satisfy the principle of continuity of the MRS in different band frequencies. A low overlap causes a discontinuity in the spectrum MRS and thus gives us a bad estimation of the energy disperse. So our problem consisted of choosing the overlap. It was necessary that the frames overlap with a percentage higher then 50% of the frame length. We chose an overlap equal to 75% (fig.2).

For the frame $p_i = 1$ of the step number i, we have N_i components:

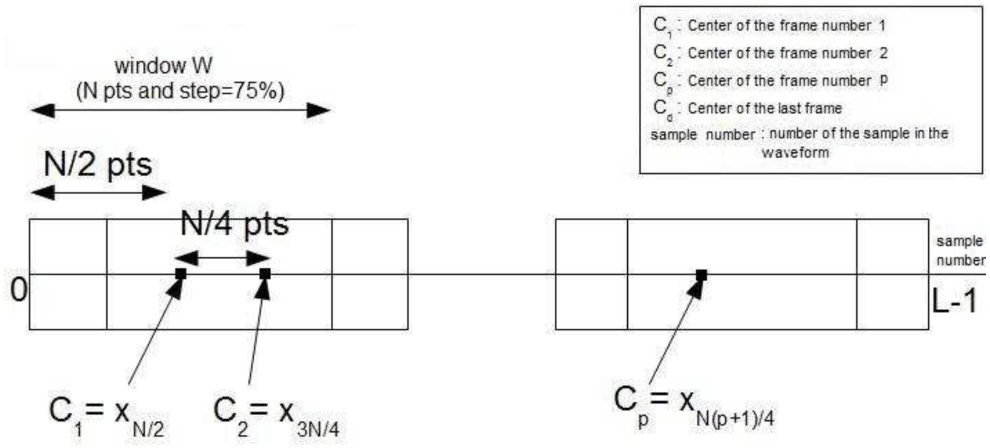

window W
(N pts and step=75%)
N/2 pts
N/4 pts
0
L-1
sample number
$C_1 = x_{N/2}$
$C_2 = x_{3N/4}$
$C_p = x_{N(p+1)/4}$
C_1 : Center of the frame number 1
C_2 : Center of the frame number 2
C_p : Center of the frame number p
C_q : Center of the last frame
sample number : number of the sample in the waveform

Figure 2. Signal sampling and windowing for center estimation C_{i,p_i} (overlap = 75%).

$$\begin{cases} s_0(1) & = & x_0 \\ s_1(1) & = & x_1 \\ & \vdots & \\ s_l(1) & = & x_l \\ \vdots & & \\ s_{N_i-1}(1) & = & x_{N-1} \end{cases}$$

For the frame $p_i = 2$ of the step number i, we have N_i components:

$$\begin{cases} s_0(2) & = & x_{N_i} \\ s_1(2) & = & x_{N_i} + 1 \\ & \vdots & \\ s_l(2) & = & x_{N_i} + l \\ & \vdots & \\ s_{N_i-1}(2) & = & x_{N_i} + N_i - 1 \end{cases}$$

In general case, for the frame p_i of the step number i, we have N_i components:

$$\begin{cases} s_0(p_i) & = & x_{(p_i-1)N_i} \\ s_1(p_i) & = & x_{(p_i-1)N_i} + 1 \\ & \vdots & \\ s_l(p_i) & = & x_{(p_i-1)N_i} + l \\ & \vdots & \\ s_{N_i-1}(p_i) & = & x_{(p_i-1)N_i} + N_i - 1 \\ s_{N_i-1}(p_i) & = & x_{p_iN_i} - 1 \end{cases}$$

The center C_{i,p_i} of $p_i = 1$ was:

$$C_{i,1} = \frac{N_i}{2}$$

The center C_{i,p_i} of $p_i = 2$ was:

$$\begin{cases} C_{i,2} & = & \frac{1}{2}(\frac{1}{4} + \frac{5}{4})N_i \\ & = & \frac{3}{4}N_i \end{cases}$$

In general case, the center C_{i,p_i} of p_i was:

$$\begin{cases} C_{i,p_i} & = & C_{i,p_i-1} + \frac{N_i}{4} \\ & = & C_{i,1} + (p_i - 1)\frac{N_i}{4} \\ & = & x_b \quad with \quad b = \frac{N_i(p_i+1)}{4} \end{cases}$$

with : $\frac{N_i(p_i+1)}{4} \leq L$ and $p_i \leq \frac{4L}{N_i} - 1$

The spectrum $S_{i,k}(p_i)$ of each step i was :

$$S_{i,k}(p_i) = \sum_{l=0}^{N_i-1} s_{i,l}(p_i)e^{-\frac{2j\pi kl}{N}} w(s_{i,l}(p_i) - C_{i,p_i}) \tag{2}$$

with: $C_{i,p_i} = x_{\frac{N_i(p_i+1)}{4}}$ the center of the frame p_i and the overlap equal to 75%.

So, the multiresolution spectrum MRS was:

$$S_k(p) = S_{i,k}(p_i) \quad with \quad k_i \leq k \leq k_{i+1} \tag{3}$$

with: $0 \leq k \leq N_0 + N_1 + \ldots + N_P$ and $\quad 1 \leq p \leq P$.

Figure 3 shows the difference between classical FFT and the MR FFT. For standard FFT, the size of the window is equal for each frequency band unlike the MRS windows size. It dependent of the frequency band.

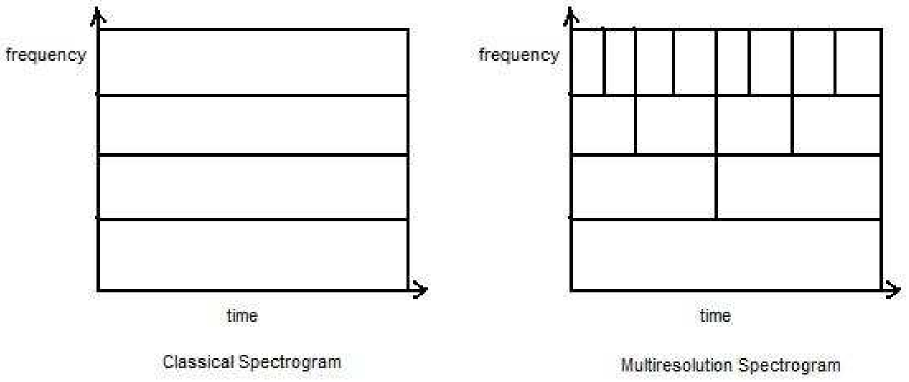


Figure 3. Difference between classical FFT and the MR FFT.

4 Materials and Methods

4.1 Corpus

Our corpus is composed of TD prounounced by Tunisian speakers. The sampling frequency is equal to 44.1 KHz, the wav format was adopted in mono-stereo. We avoided all types of noise filter that would degrade the quality of the signal and thus, cause information losts. We have recorded the real time spontaneous discussions of 4 speakers. We have removed noise and some sounds like laughing, music, etc. It was difficult to realize a spontaneous corpus because, in real time, it is impossible to have all phonemes and syllables. Another difficulty was the variability of discussion themes and pronounced sounds. For these reasons, we decided to complete our corpus with another one. We prepared a text in Tunisian dialect with all sounds to study. Every phoneme and syllable appeared 15 times. We asked four speakers: two men and two women, to read the text in a high voice in the same conditions of the first corpus records. All speakers are between 25 and 32 years old. Speakers did not know the text. Our corpus was transcribed in sentences, words and phonemes.

4.2 VASP Software: Visual Assistance of Speech Processing Software

For our study, we have created our first prototype System for Visual Assistance of Speech Processing VASP (figure 4). It offers many functions for speech visualization and analysis. We realized our system with GUI Matlab. In the following subsection, we will present some of the functionalities offered by our system.

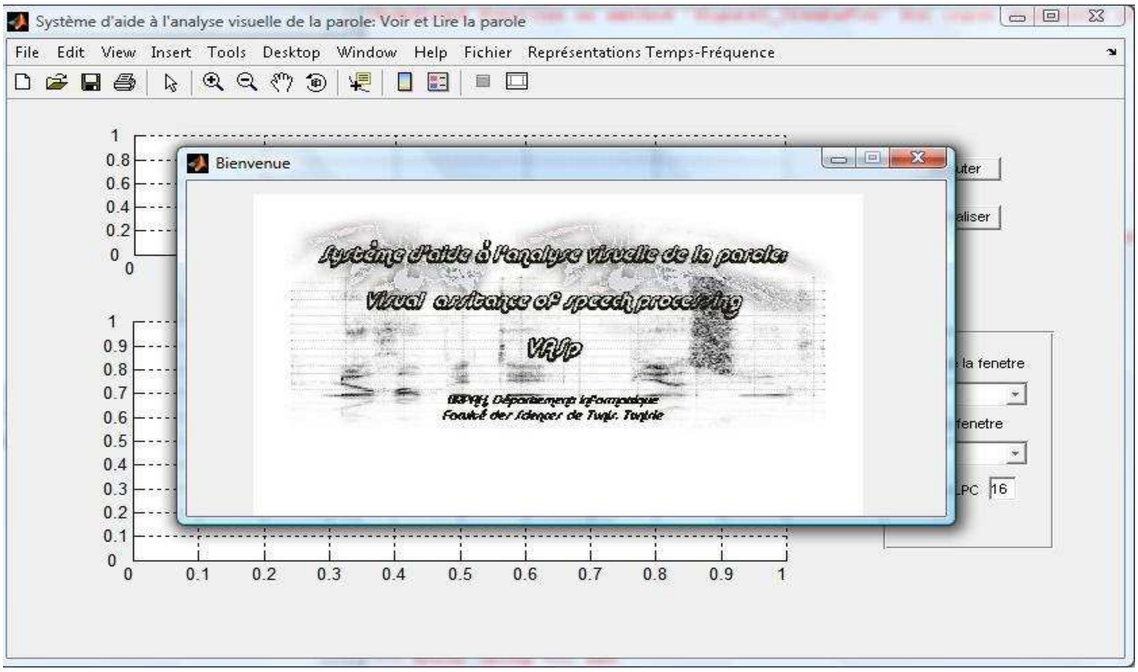

Système d'aide à l'analyse visuelle de la parole: Voir et Lire la parole
File
Edit
View
Insert
Tools
Desktop
Window
Help
Fichier
Représentations Temps-Fréquence
Bienvenue
Système d'aide à l'analyse visuelle de la parole
Visual assistance of speech processing
VASP
URPAH Département Informatique
Faculté des Sciences de Tunis. Tunisie
uter
aliser
la fenetre
fenetre
LPC
16
1
0.8
0.6
0.4
0.2
0
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4. A screenshot of VASP.

VASP reads sound files in wav format. It represents a wav file in time domain by waveform and in time-frequency domain by spectral representation, classical spectrogram in narrow band and wide band, spectrograms calculated with linear prediction and cepstral coefficients, gammatone, discrete cosine transform (DCT), Wigner-Ville transformation, Multiresolution LPC representation (MR LPC), Multiresolution spectral representation (MR FFT) and Multiresolution spectrogram.

From the waveform, we can choose, in real time, the frame for which we want to represent a spectrum. Parameters are manipulated from a menu; we can select the type of windows (Hamming, Hanning, triangular, rectangular, Kaiser, Barlett, gaussian and Blackmann-Harris), window length (64, 128, 256, 512, 1024 and 2048 samples) and LPC factor.

From all visual representations, the coordinates of any pixel can be read. For example, we can select a point from a spectrogram and read directly its coordinates (time, frequency and intensity).

VASP offers the possibility to choose a part of a signal to calculate and visualize it in any time-frequency representations.

Our system can automatically detect Silence/Speech from a waveform. From the spectrogram, the system can detect acoustic cues like formants, and classify it automatically to two classes: sonorant or non-sonorant.

Our system can analyse visual representations with two methods; image analysis with edge detection, and, sound analysis of signal. Edge detection is calculated with gradiant method or median filter method.

4.3 Interquartile Range (IQR)

The Tukey Box-and-Whisker plot is an exploratory graphic [10]. It is a powerful mean of observation more interesting than the histograms. It is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest observation (sample minimum), lower quartile ($Q1$), median ($Q2$), upper quartile ($Q3$), and largest observation (sample maximum). A boxplot may also indicate which observations, if any, might be considered outliers [10].

The Tukey box-and-whisker diagram displays differences between populations without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box help in indicating the degree of dispersion and skewness in the data, and identify outliers [14].

The Interquartile Range (IQR) is the distance between the 75th percentile and the 25th percentile [29]. The IQR is essentially the range of the middle 50% of the data. Because it uses the middle 50%, the IQR is not affected by outliers or extreme values. The IQR is also equal to the length of the box in a box plot [29].

5 Experimental Results

Our experimental part consisted of approving the MRS theory applied on our corpus. We presented a method for automatic detection of transition zones based on MRS and compared it to classical spectral analysis.

5.1 Multiresolution Spectral Analysis

We calculated a multiresolution spectrogram of each speech signal. We chose Hamming window with lengths [23,20,15,11] ms for frequency bandwiths [0-2000, 2000-4000, 4000-7000, 7000-10000] Hz and 75% overlap.

Figure 5 shows the classical sonagram; Hamming window, 11 ms with an overlap equals to 1/3 and figure 6 shows the MR sonagram; Hamming (23, 20, 15, 11) ms, overlap of 75%, Band-Limits in Hz were [0, 2000, 4000, 8000, 12000] of the sentence: "Le soir approchait, le soir du dernier jour de l'année". MRS offers several time integrations which are combinations of several FFT of different lengths depending on the frequency bandwidth.

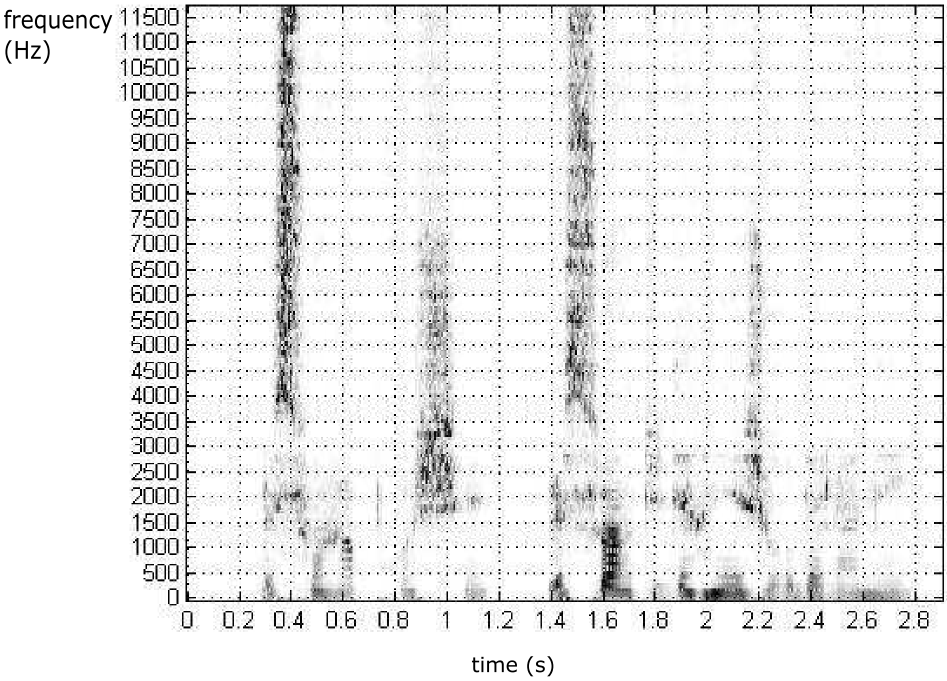


Figure 5. Classical sonagram (Hamming, 11 ms, overlap of 1/3) of this sentence: ”Le soir approchait, le soir du dernier jour de l’année”

5.2 Automatic detection of transition zones

Before starting the experimental phase, we must set the input parameters. The parameters of calculations MRS were fixed. We chose Hamming window in lengths (23 ms, 20 ms, 15 ms and 11 ms), an overlap of 75% and Band-Limits were [0, 2000, 4000, 8000, 12000] Hz.

We calculated the IQR for each frame to our corpora. All the composants of each frame were real numbers between 0 and 255. Each diagram should allow us to clearly visualize the Tukey box-and-whisker plot and the areas of transition between the different Tukey box-and-whisker graphs and thus between the different classes. The length of each frame was 10 ms for classical spectrogram and 1.7 ms for MRS.

We calculated the lower quartile $Q1$, the second quartile $Median$, the upper quartile $Q3$ and the interquartile range IQR of each frame and we plotted the Tukey box-and-whisker diagram. We presented our decision rules for detection of the transition zones and we applied it to our corpora. We compared our results to experimental thresholds; $Q3_{th}$, $Q1_{th}$ and IQR_{th}. Figure 7 shows examples of the Tukey Box-and-Whiskers diagram.

In this study, we compared our method with a classic spectral analysis. Figure 8 and figure 9 show results of automatic detection of transition zones based, respectively, on classical spectral analysis and on multiresolution spectral analysis.

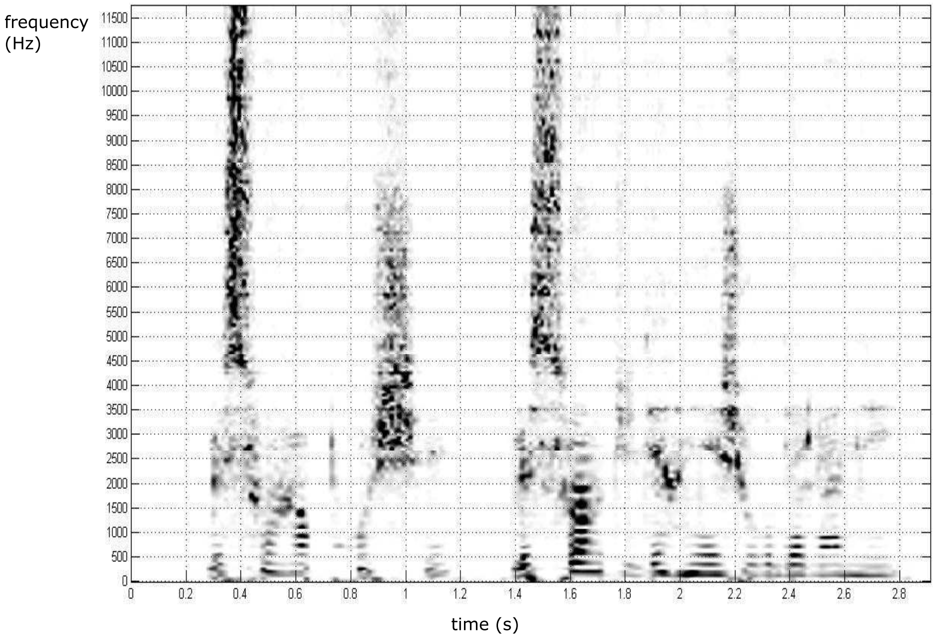


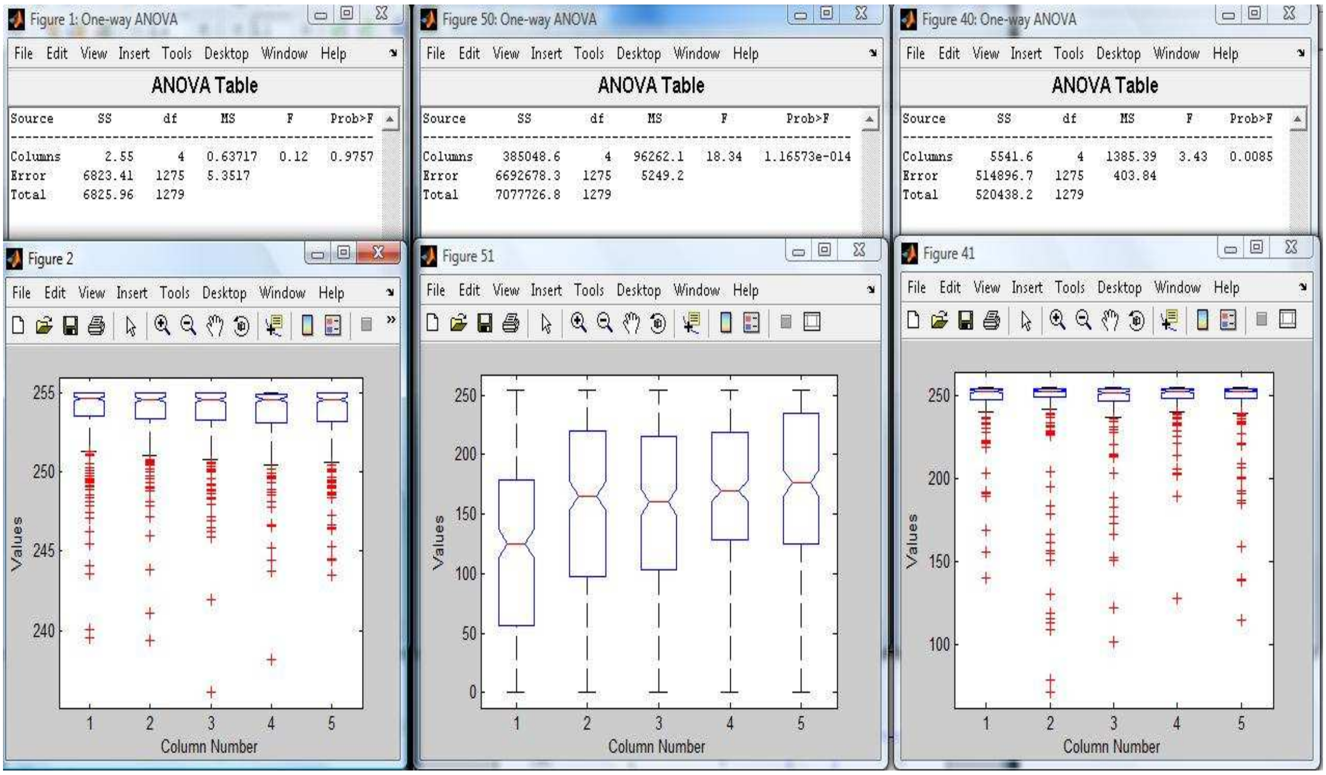
Figure 6. Multiresolution sonagram; Hamming (23 ms for the band-limit [0-2000] Hz, 20 ms for the band-limit [2000-4000] Hz, 15 ms for the band-limit [4000-8000] Hz and 11 ms for the band-limit [8000-12000] Hz), overlap of 75%, of this sentence: "Le soir approchait, le soir du dernier jour de l'année"

6 Discussions

In this study we presented and tested the performance of an automatic detection of the transition zones based on multiresolution spectral analysis. We applied our self developed system (VASP) based on multiresolution to our corpus. VASP is a self developed system regrouping all our needed tools. VASP presents a visual improvement compared to standard spectrogram. It enables a better extraction of the acoustic cues of the signal. It is an automated open system. In comparison to Praat system (freeware), VASP does not allow phonemes transcription and has less tools. But VASP offers us more time-frequency representations and allows the automatic detection of the transition zones.

We calculated the MR FFT for each signal. Then, we detected the transition zones of the sound based on decision rules. Figure 7 shows examples of the Tukey Box-and-Whiskers diagram in the case of MRS FFT. We remarked that the values of $Q1$, $Median$, $Q3$ and IQR was varied when the frame represented a silence or a stop consonant or a vowel. We defined decision rules based on these variations. We compared all values with experimental thresholds; $Q3_{th}$, $Q1_{th}$ and IQR_{th}.

Figure 9 represents the variation of the IQR in the case of MRS FFT and figure 8 represents the variation of the IQR in the case of classic spectral analysis. Each peak represents a transition between silence-stop consonant or stop consonant-vowel or vowel-silence.



```
Source       SS       df      MS        F     Prob>F
-----------------------------------------------------
Columns      2.55       4   0.63717   0.12   0.9757
Error     6823.41    1275   5.3517
Total     6825.96    1279
```

```
Source        SS        df      MS        F       Prob>F
-----------------------------------------------------------
Columns    385048.6      4   96262.1   18.34   1.16573e-014
Error     6692678.3   1275    5249.2
Total     7077726.8   1279
```

```
Source       SS       df      MS       F     Prob>F
----------------------------------------------------
Columns    5541.6      4   1385.39   3.43   0.0085
Error    514896.7   1275    403.84
Total    520438.2   1279
```

Figure 7. Examples of the Tukey Box-and-Whiskers diagram of Silence on the left, a Stop Consonant on the middle and a Vowel on the right in the case of MRS FFT.

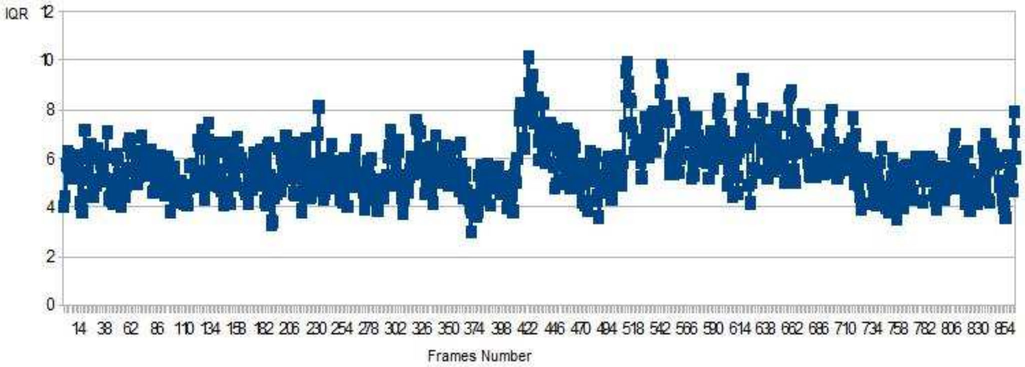


Figure 8. Transition zones detection based on classical spectral analysis and IQR calculation of this sentence : "$/hal \quad anna \quad essa?a : da/$" pronounced by a tunisian speaker in Tunisian Dialect which means '$Is\ that\ happiness$'.

For transition zone detection based on MRS FFT and IQR calculation, we obtained a score of **52**%. The score of transition zone detection based on classical spectral analysis and IQR calculation was **20.75**%. Our method based on MRS FFT provides better results compared to classical spectral analysis. Detection was better and errors were fewer.

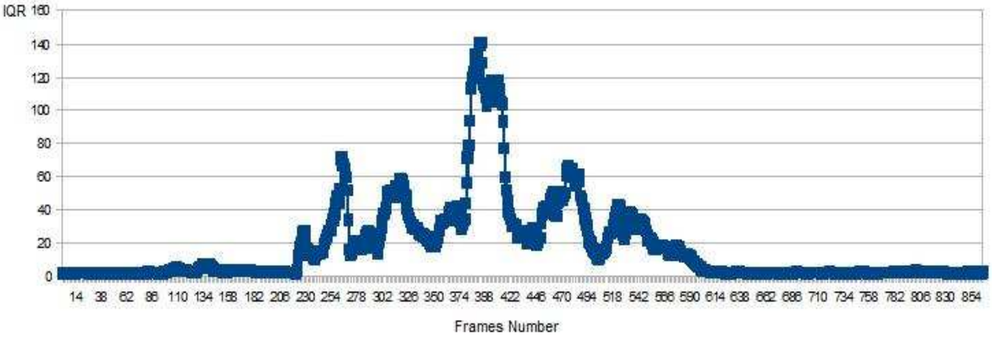


Figure 9. Transition zones detection based on MRS FFT and IQR calculation of this sentence : "$/hal \quad anna \quad essa?a : da/$" pronounced by a tunisian speaker in Tunisian Dialect which means '$Is\ that\ happiness$'.

7 Conclusion

We presented our approach to the analysis of the speech signal based on a pre-classification of frames extracted from spectrograms (classic and multiresolution). We proposed a classifier that aims to identify transition zones carrier information fields. In our experiment, we have shown that the classification results based on the analysis of MRS, gives better results than the classical spectral analysis. This is due to the wealth of information from the MRS. Our classifier is based on detection methods of dispersion energy. The MRS data from a sample of a size of the window is correlated with the frequency bands of which the cover is not selected so that it retains the continuity between the strips. These choices have facilitated the detection of energy zones and explain the effectiveness and contribution of MRS to detecting transition zones. We realised our experiments on the Tunisian dialect. We intend to extend this study to another corpus composed by a text reader in Arabic language before developing a method for automatic segmentation and classification.

References

[1] Nefissa Annabi-Elkadri and Atef Hamouda. Spectral analysis of vowels /a/ and /e/ in tunisian context. In *2010 International Conference on Audio, Language and Image Processing*, number CFP1050D-ART in 978-1-4244-5858-5. IEEE/IET indexed in both EI and ISTP, Novembre 2010. (in Press).

[2] Nefissa Annabi-Elkadri and Atef Hamouda. Analyse spectrale des voyelles /a/ et /e/ dans le contexte tunisien. In *Actes des IXe Rencontres des Jeunes Chercheurs en Parole RJCP*, pages 1–4. Université Stendhal, Grenoble, Mai 2011.

[3] Nefissa Annabi-Elkadri and Atef Hamouda. The multiresolution spectral analysis for automatic detection of transition zones. *International Journal of Advanced Science and Technology*, 36:95–110, November 2011.

[4] Nefissa Annabi-Elkadri and Atef Hamouda. Automatic Silence/Sonorant/Non-Sonorant Detection based on Multi-resolution Spectral Analysis and ANOVA Method. In *International Workshop on Future Communication and Networking*, Hong Kong, 2011 (in press). IEEE.

[5] Nefissa Annabi-Elkadri, Atef Hamouda, and Khaled Bsaies. *Speech Processing*, chapter Multiresolution Spectral Analysis of Vowels in Tunisian Context, page . 979–953–307–620–0. INTech, 2011 (Accepted).

[6] René Boite, Hervé Bourlard, Thierry Dutoit, Joel Hancq, and Henri Leich. *Traitement de la parole.* ISBN 2-88074-388-5. Presses Polytechniques et Universitaires Romandes, 2000.

[7] Calliope. *La parole et son traitement automatique.* collection technique et scientifique des télécommunications, MASSON et CENT-ENST, Paris, ISBN :2-225-81516-X, ISSN : 0221-2579, 1989.

[8] P. Cancela, M. Rocamora, and E. Lopez. An efficient multi-resolution spectral transform for music analysis. In *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 309–314, 2009.

[9] C.P. Chan, Y.W. Wong, Tan. Lee, and P.C. Ching. Two-dimensional multi-resolution analysis of speech signals and its application to speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP99*, volume 1, pages 405–408. IEEE, Mars 1999.

[10] Alan Chauvin and Richard Palluel-Germain. Analyse de la variance. In *Tutorial in "Journées Rencontre Jeunes Chercheurs en Parole (RJCP)"*, 2011.

[11] S. Cheung and J.S. Lim. Combined multi-resolution (wideband/narrowband) spectrogram. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, pages 457–460. IEEE, 1991.

[12] Tai-Shih Chi and Chung-Chien Hsu. Multiband analysis and synthesis of spectro-temporal modulations of fourier spectrogram. *Journal of Acoustical Society of America JASA Express Letters*, 129(5):EL190–EL196, May 2011.

[13] Laurence Cnockaert. *Analysis of vocal tremor and application to parkinsonian speakers / Analyse du tremblement vocal et application à des locuteurs parkinsoniens.* PhD thesis, F512 - Faculté des sciences appliquées - Electronique, 2008.

[14] Flowing Data. How to read (and use) a box-and-whisker plot, 2008.

[15] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proceeding of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 247–252, September 2006.

[16] Qiang Fu and Eric A. Wan. A novel speech enhancement system based on wavelet denoising. *Center of Spoken Language Understanding, OGI School of Science and Engineering at OHSU*, .:., February 2003.

[17] Michael Gibson. *Dialect Contact in Tunisian Arabic: sociolinguistic and structural aspects.* PhD thesis, University of Reading, 1998.

[18] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of consonant shape. *SIAM Journal on Mathematical Analysis*, 15(4):723–736, 1984.

[19] J.P. Haton and al. *Reconnaissance automatique de la parole.* DUNOD, 2006.

[20] Charalampos Karypidis. *Asymétries en perception et traitement de bas niveau: traces auditives, mémoire à court terme et représentations mentales (Asymmetries in perception and low-level processing: auditory traces, short-term memory and mental representations).* PhD thesis, Université Paris 3 – Sorbonne Nouvelle, Paris, France, 2010.

[21] Peter Ladefoged. *Elements of Acoustic Phonetics.* The University of Chicago Press, 1996.

[22] Hélène Leman and Catherine Marque. Un algorithme rapide d'extraction d'arêtes dans le scalogramme et son utilisation dans la recherche de zones stationnaires / a fast ridge extraction algorithm from the scalogram, applied to search of stationary areas. *Traitement du Signal*, 15(6):577–581, 1998.

[23] S. Mallat. A theory for multiresolution signal decomposition : the wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[24] S. Mallat. *Une Exploration des Signaux en Ondelettes.* Editions de l'Ecole Polytechnique, Ellipses diffusion, 2000.

[25] S. Mallat. *A wavelet Tour of Signal Processing.* Academic Press, 3rd edition edition, 2008.

[26] S. Manikandan. Speech enhancement based on wavelet denoising. *Academic Open Internet Journal*, 17(1311–4360):., 2006.

[27] William Marçais. *Les parlers arabes, Initiation à la Tunisie.* d. Adrien Maisonneuve, Paris, 1950.

[28] Rohini R. Mergu and Shantanu K. Dixit. Multi-resolution speech spectrogram. *International Journal of Computer Applications*, 15(4):28–32, February 2011.

[29] Steve Simon. What is the interquartile range?, 2008.