The Multiresolution Spectral Analysis for Automatic Detection of Transition Zones

Nefissa Annabi-Elkadri[(1)], *Atef Hamouda*
URPAH, Computer Science Department, Faculty of Sciences of Tunis,
Tunis El Manar University, Tunis, Tunisia.
[(1)]*nefissa.annabi@gmail.com*
https://sites.google.com/site/nefissaannabielkadri/

Abstract

This paper presents an automatic method for detecting transition zones based on multiresolution spectral analysis (MRS). 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 an automatic detection of zones of transition by calculating the Interquartile Range (IQR) of each frame of the MRS FFT. We applied our Visual Assistance of Speech Processing (VASP) System to a corpus. This corpus was in French pronounced by french speakers and has the format C_iVC_iV with C_i was a stop consonant [p t k] and V was a vowel [i e]. The results showed 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

Choosing an appropriate window length for spectral analysis is not a straightforward 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, with masking or distorting rapid acoustic landmarks occurring in speech. [26] 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.

A classic speech spectrogram is a visual representation of log-magnitude amplitude (dB) versus time and frequency. It offers a single integration time which is the length of the window and implements a uniform bandpass filter, with spectral samples being regularly

spaced and corresponding to equal bandwidths.

[31, p.674] remarks *"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 [22,31–33]. It has since been applied to several domains: image analysis [31], time-frequency analysis [15], speech enhancement [20, 34], automatic signal segmentation by search of stationary areas from the scalogram [28].

The MR spectrum, a compromise that provides both a higher frequency and a higher temporal resolution, is not a new method. In phonetic analysis, [2, 3] presents a study of two common vowels /a/ and /E/ 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 /E/ are, in some contexts, similarly pronounced. [4] applies the MRS for an automatic method for Silence/Sonorant/Non-Sonorant detection used the ANOVA method. Results are compared the classical methods for classifications such as Standard Deviation and Mean with ANOVA who were better. The method for automatic Silence/Sonorant/Non-Sonorant detection based on MRS provides better results compared to classical spectral analysis. [13] present 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. [11] 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, [10] 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, [17] 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 [35] and speech synthesis [14].

2 Characteristics of stop consonants

In articulatory phonetics, the point of articulation of a consonant is the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator and a passive location. Along with the manner of articulation and the phonation, this gives the consonant its distinctive sound [26, 27].

[5, p.272] defines the stops consonants that *"are unique among the sounds of speech in that they include a variable period of total blockage of airflow during which sound output may cease. During this interval air pressure rises behind the point of closure to be released as a burst of acoustic energy"*.

Three major parameters discriminated the stop consonants [5]:

- the characteristics of the burst,
- the nature of formant transitions before and after the closure period,
- the time required for the reestablishment of the Voice Onset Time following release of the closure.

When a voiced stop is followed by a vowel sound, its evolution is as follows [40]:

- a momentary silence of the occlusion;
- a burst;
- friction noise produced at the constriction, the spectrum is that of a sharp band noise;
- glottal flow noise and the spectrum is that of a noise band sharper than the previous one, but can temporarily coexist with him;
- vibration of vocal cords - due to the vowel following the consonant articulation that produces a harmonic spectrum, the intensity decreases regularly from low to high and this signal appears with a some VOT delay with respect to the burst that oscillates between 10 and 30 ms for French.

The production of the voiceless stop is the same as that of its voiced counterpart. The glottal flow noise disappears and the sounds of bursts and friction are reduced.

[19, p.B-21] studied the spectra evolution of stop consonants: *"The [k] has a single, centrally located peak, which accounts for the ompactness. There is a high frequency emphasis of the [t], qualifying for acuteness and a low frequency emphasis for [p] which is grave"* [18]. [38, 39] presented a method based on locus equations for stop consonants classification [38] and phonetically described stop place categories as a function of syllable -initial, -medial, and -final position [39].

When analysing a stop consonant acoustically, we generally take into consideration the duration of the closure phase and that of the burst, the duration of the friction, the Voice Onset Time (VOT) [29] and the number of bursts [1, 6, 29]. [36] has presented an overview of studies on the features of the laryngeal movements for plosives in several languages using endoscopy and photo-electroglottography. [42] proposed a method to find the burst. A breakdown of the plosive in 3 segments burst, suction and friction noise, has been made. Formant transitions are often visible in the noise and friction in aspiration; noise begins to relaxation of the articulation (at the end of the keeping) and CV sequences end with the first vowel period. The noise spectrum is changing rapidly as a function of time. In most cases, it is more discriminating in its initial part. More spectrum is calculated by the following vowel, the more it is dominated by the formants of the vowel. [42] calculate spectra corresponding to [25] voiced plosives at 2 different times to evaluate the cues for the place articulation of plosives just after the relaxation of the eye consonantal articula-

Table 1. The duration of burst in stop consonants (in ms)

Unvoiced	ka	ta	pa
Duration (ms)	32	23	12
Voiced	ga	da	ba
Duration (ms)	19	12	7

tion. Segmentation noise is very difficult to realize, only a segment of fixed length at the beginning of the noise is generally taken into account; [7] for 26 ms, or 10 to 15 ms for [43].

[41] propose some nonspectral methods for analysis of stop consonants based on the excitation source information in speech signal. This method attempts to make a case for nonspectral methods for analysis of stop consonants. Table 1 shows the duration of burst in stop consonants (in ms) in Indian Languages [41].

[21] characterize an incomplete stop consonant by an indistinguishable closure or a missing burst. He shows that closure duration can be used as a feature to classify the incomplete stops due to Stop-Stop Interaction (SSI) and the complete stops in read speech of the TIMIT with 69.66% (79.14%) accuracy for automatically estimated [21] durations.

[30] presented a method based on the wavelet transform for detection of stop consonants in the French language. Correlation functions of the Gaussian wavelet was calculated. [30] makes the remark that *"Correlation functions have a minimum before each maximum, where that maximum is synchronous with the burst of the stop consonant"*. The rate of detection is 94% for the unvoiced stops and 75% for voiced stops.

3 Multiresolution FFT

It's so difficult to choose the ideal window with the ideal characteristics. The size of the ideal window [8] was equal to twice the length of the pitch of the signal. A wider window show 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 [23, p.41]. We noted the center C_p. So our problem now, is the estimated of C_p.

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

We would like to calculate the spectral 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 [9, 26] to meet the stationnarity condition. We choosed the Hamming window and we fixed the size to 512 points and the overlap to 50%. Figure 1 shows the principal of the center estimation.

For each frame p, the center C_p was estimated:

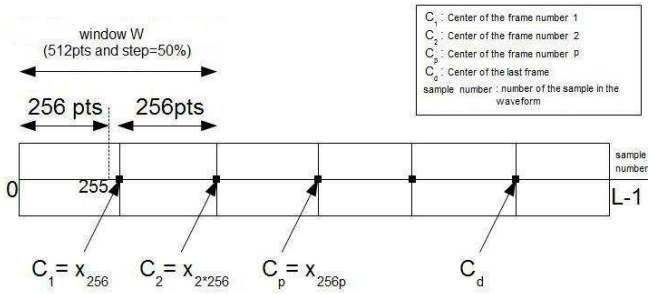

Figure 1. Signal sampling and windowing for center estimation C_p. The window length $N = 512$ points and overlap = 50%.

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

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 componant number l of s:

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

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

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

In general case:

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

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 componant of s number l of the frame p
$S_k(p)$: the componant 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 spectral, 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 NB. The number of steps NB was equal to the number of band frequency fixed a priori. For each step number i (i), 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 spectral $S_{i,k}(p_i)$ for each step i was:

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

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 give us a bad estimation of the energy dispersal. So our problem consisted on the overlap choosing. It was necessary that the frames overlap with a percentage higher to 50% of the frame length. We choosed an overlap equal to 75% (fig.2).

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

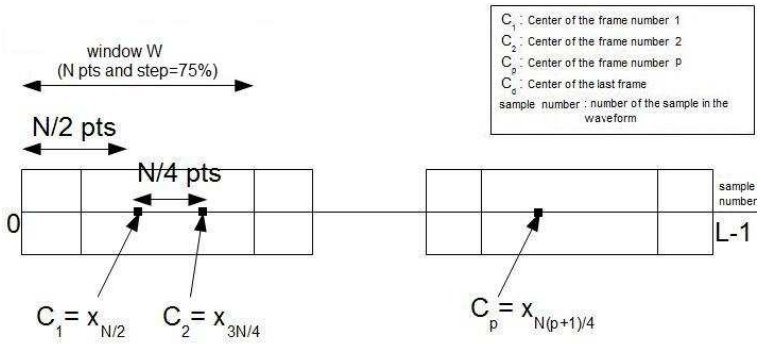

Figure 2. Signal sampling and windowing for center estimation C_{i,p_i} (overlap = 75%).

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

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:

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

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

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

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

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 spectral MRS was:

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

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

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

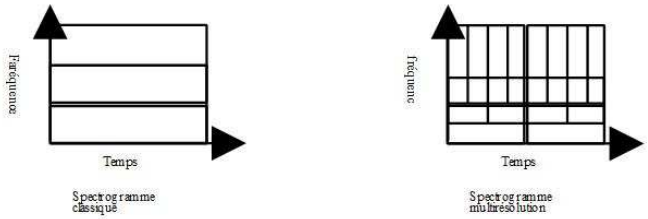

Figure 3. Difference between classical FFT and the MR FFT.

4 Method

4.1 Corpus

Our corpus was produced by 19 native French speakers [24]. This corpus was in French pronounced by French speakers and has the format C_iVC_iV with C_i was a stop consonant [p t k] and V was a vowel [i e].

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 developed our system with GUI Matlab. In the following subsection, we will present some of the functionalities offered by our system.

VASP reads sound files in wav format. It represent 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

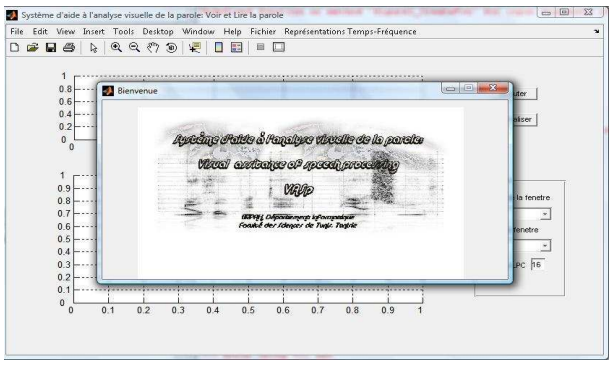

Figure 4. A screenshot of VASP.

(MR FFT) and Multiresolution spectrogram.

From 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, coordinates of any pixel can be read. For example, we can select a point from a spectrogram and read its coordinates directly (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 signal. Edge detection is calculated with gradiant method or median filter method. The second method is based on detecting energy from a time-frequency representation.

4.3 The interquartile Range (IQR)

The Tukey Box-and-Whisker plot is an exploratory graphic [12]. It is a powerful means 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 [12].

The Tukey box-and-whisker diagram display differences between populations without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box help indicate the degree of dispersion and skewness in the

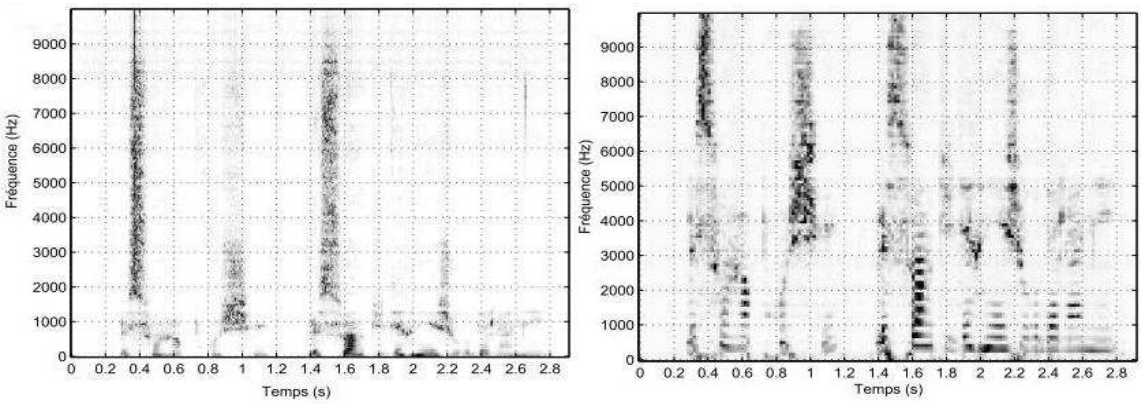

Figure 5. Classical sonagram (Hamming, 11ms, overlap 1/3) (on the left) and MR sonagram (on the right); Hamming (23, 20, 15, 11 ms), overlap 75%, Band-Limits in Hz were [0, 2000, 4000, 7000, 10000]Hz of this sentence: "Le soir approchait, le soir du dernier jour de l'année"

data, and identify outliers [16].

The Interquartile Range (IQR) is the distance between the 75th percentile and the 25th percentile [37]. 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 [37].

5 Experimental Results

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

5.1 Multiresolution Spectral Analysis

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

Figure 5 shows the classical sonagram (on the left); Hamming window, 11ms with an overlap equal to 1/3 and the MRS (on the right); Hamming (23, 20, 15, 11)ms, overlap 75%, Band-Limits in Hz were [0, 2000, 4000, 7000, 10000] 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 frequency bandwidth.

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 a Hamming window in lengths (23ms,

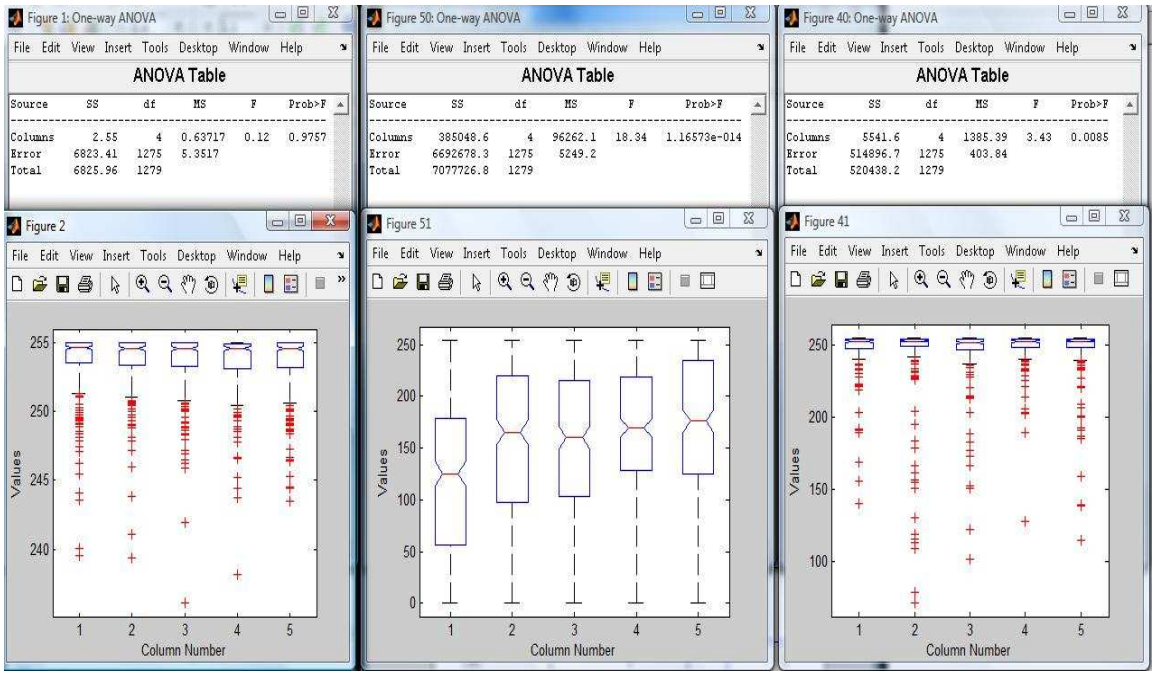

Figure 6. Examples of the Tukey Box-and-Whiskers diagram of Silence on the left, a Stop Consonant on the middle and a Vowel on the right in the case of MRS FFT.

20ms, 15ms and 11ms), a 75% overlap and Band-Limits were [0, 2000, 4000, 7000, 10000]Hz.

Our corpus was composed by words at the format C_iVC_iV with C_i was a stop consonant [p t k] and V was a vowel [i e]. We calculated the IQR for each frame. All the composants of each frame were reals between 0 and 255. Each diagram should allow us to clearly visualize the Tukey box-and-whisker plot and the areas of transitions between the different Tukey box-and-whisker plot and thus between the different classes. The length of each frame was $10ms$ for classical spectrogram and $1.7ms$ 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 ploted 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 6 shows examples of the Tukey Box-and-Whiskers diagram.

In this study, we compared our method with a classic spectral analysis. Figure 7 and figure 8 show results of automatic detection of transition zones based, respectively, on classical spectral analysis and on multiresolution spectral analysis.

6 Discussion

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 on our corpus. VASP is a self developed system

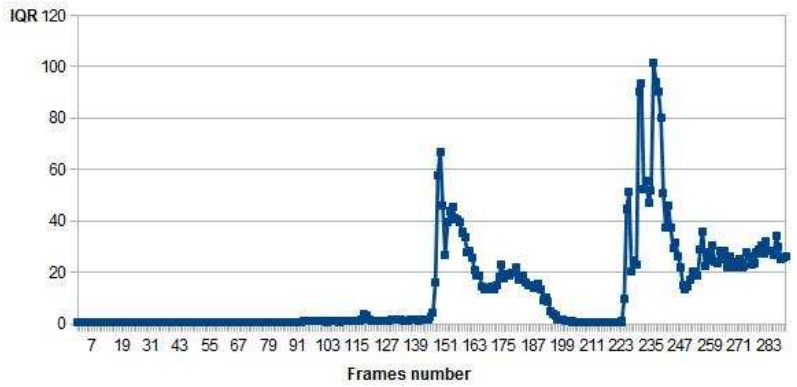

Figure 7. Transition zones detection based on classical spectral analysis and IQR calculation of a sentence at the format "[k]V[k]V" where [k] was a stop consonant and V was a vowel [i e] pronounced by one speaker.

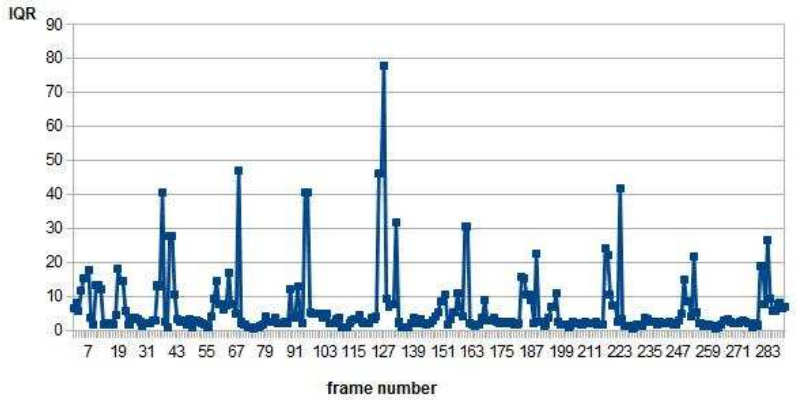

Figure 8. Transition zones detection based on MRS FFT and IQR calculation of a sentence at the format "[k]V[k]V" where [k] was a stop consonant and V was a vowel [i e] pronounced by one speaker.

regrouping all our needed tools. VASP presents a visual improvement compared to standard spectrogram. It enables a better extraction of acoustic cues of the signal. It is an automated open system. In comparison to Praat system (freeware), VASP doesn't allow phonemes transcription and has less tools. But VASP offers us more time-frequency representations and allows for an automatic detection of the transition zones.

We calculated the MR FFT for each signal. We then detected the transition zones of the sound based on decision rules. Figure 6 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 a decision rules based on these variations. We compared all values to experimental thresholds; $Q3_{th}$, $Q1_{th}$ and IQR_{th}.

Figure 8 represents the variation of the IQR in the case of MRS FFT and figure 7

represented the variation of the IQR in the case of classic spectral analysis. Each peak represented a transition between silence-stop consonant or stop consonant-vowel or vowel-silence.

For transition zone detection based on MRS FFT and IQR calculation, we obtained a score of 57.5%. The score of transition zone detection based on classical spectral analysis and IQR calculation was 23.75%. Our method based on MRS FFT provides better results compared to classical spectral analysis. Detection was better and errors were fewer.

7 Conclusion

The automatic detection of the transition zones based on MRS FFT seems to give better results than the classical spectral analysis. We are currently extending this study to another corpus composed by read text in Arabic language before developing a method for automatic segmentation and classification.

ACKNOWLEDGMENTS

We wish to thank Charalampos Karypidis for his advice in accomplishing this work.

References

[1] Arthur S. Abramson and Leigh Lisker. Voice Onset Time in Stop Consonants: Acoustic analysis and Synthesis. In *Rapports du 5ème Congrès Intl. d'Acoustique*, volume Ia, 1965.

[2] 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).

[3] 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.

[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*, Szczecin, Poland, 2011 (in press). IEEE.

[5] R. J. Baken and Robert F. Orlikoff. *Clinical Measurement of Speech and Voice*. Singular Publishing Group, 2000.

[6] J. Soto Barba. Los fonemas /b/ y /p/ se diferencian por la sonoridad. In *Estudios Filológicos*, volume 29, pages 33–37, 1994.

[7] Sheila E. Blumstein and Kenneth N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral charcteristics of stop consonants. *Journal Acoustica Society of America*, 66:1001–1017, 1979.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Alan Chauvin and Richard Palluel-Germain. Les principes de l Anova . In *Journées RJCP*, 2011.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] G. Fant. *Acoustic analysis and synthesis of speech with applications to Swedish*. Ericsson Technics, 1959.

[19] Gunnar Fant. Phonetics and phonology in the last 50 years. In *Sound to Sense at MIT*, volume B, pages 20–41, June 2004.

[20] 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*, 2003.

[21] Prasanta Kumar Ghosh and Shrikanth Narayanan. Closure duration analysis of incomplete stop consonants due to stop-stop interaction. *Journal Acoustica Society of America Express Letters*, 126(1):EL1–EL7, July 2009.

[22] 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.

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

[24] 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.

[25] D. Krull. Relating acoustic properties to perceptual responses: a study of swedish voiced stops. *Journal of the Acoustical Society of America*, 88(6):2557–2570, 1990.

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

[27] Peter Ladefoged and Keith Johnson. *A Course in Phonetics*. Wadsworth, Cengage Learning, 2010.

[28] 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.

[29] L. Lisker and A. S. Abramson. Stop categorization and voice onset time. In *Proceedings of the 5th International Congress of Phonetic Sciences*, pages 389–391, 1964.

[30] F. Malbos, M. Baudry, and S. Montrésor. Détection des occlusives à l'aide de la transformée en ondelettes. *Journal de Pysique*, 4:493–496, Mai 1994.

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

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

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

[34] S. Manikandan. Speech enhancement based on wavelet denoising. *Academic Open Internet Journal*, 17, 2006.

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

[36] R. Ridouane. Voisement, aspiration et pré-apiration : le comportement glottal.

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

[38] H. M. Sussman, Helen A. McCaffrey, and Sandra A. Matthews. An investigation of locus equations as a source of relational invariance for stop place categorization. *Acoustical Society of America*, 90(3):1309–1325, 1990.

[39] Harvey M. Sussman, Nicola Bessell, Eileen Dalston, and Tivoli Majors. An investigation of stop place of articulation as a function of syllable position: A locus equation perspective. *Acoustical Society of America*, 101(5):2826–2838, May 1997.

[40] Équipe de linguistique générale UNIL. Caractéristiques acoustiques des différentes réalisations. Université de Lausanne, 2011.

[41] B. Yegnanarayana, Sri Rama Murty K., and S. Rajendran. Analysis of stop consonants in indian languages using excitation source information in speech signal. In *Proceedings ISCA ITRW Speech Analysis and Processing for Knowledge Discovery, Aalborg, Denmark*, June 2008.

[42] Anne Bonneau Yves Laprie. Segmentation du bruit d'explosion des occlusives. *XXIVèmes journées d'étude sur la parole, Nancy*, 26:24–27, 2002.

[43] Victor Waito Zue. *Acoustic Characteristics of Stop Consonants: a controlled study.* PhD thesis, Massachusetts Institute of Technology, MIT, Lincoln Laboratory, May 1976.