MNK Publication

International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume 3, Issue 2: Page No.50-54 ,March-April, 2014
https://www.mnkpublication.com/journal/ijlrst/index.php

COMPARATIVE STUDIES FOR SPEECH ANALYSIS BASED ON MULTIRESOLUTION SPECTROGRAMS

Nefissa Annabi-Elkadri
[1] LIPAH, Faculté des Sciences de Tunis, Tunis, Tunisie

Abstract- This paper presents an evaluation study of the Multiresolution Spectral Analysis (MRS) method which provides 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 transition zone detection and an automatic silence/sonorant/non-sonorant classification. Our approach is compared to existing methods based on the MRS and classical spectral analysis by the means of our Visual Assistance of Speech Processing (VASP) System and two corpora. Our approach appears to yield better results in the two tasks in question than the other methods.

Keywords **-** Multiresolution Spectrogram; transition zones detection; Silence/Sonorant/Non-Sonorant detection ; Comparative studies

I. 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 to 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. [18] thus 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 o□ers 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.

Mallat [20, 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 calculate a multiresolution (MR) spectrum. In classical litterature, the MR analysis is based on discrete wavelet transforms [15,20–22]. It has since been applied to several domains: image analysis [20], time-frequency analysis [11], speech enhancement [14,23], automatic signal segmentation from the scalogram [19].

The MR spectrum, a compromise that provides both a higher frequency and a higher temporal resolution, is not a novelty. In phonetic analysis, Annabi-Elkadri and Hamouda [1] presents a study of two common vowels /a/ and /E/ in the Tunisian dialect and in French. Cheung and Lim [9] presents a method for combining a wideband and a narrowband spectrogram by evaluating the geometric mean of their corresponding pixel values. Chan and al. [7] describes the use of MR for clean connected speech and noisy phone conversation speech. For music signals, Cancela and al. [6] presents two algorithms, the efficient constant-Q transform and the MR Fast Fourier transform (FFT). These two are reviewed and compared to a new tool based on the Infinite Impulse Response filtering of the FFT. Additionally, MR FFT has been used as a part of an effective melody extraction algorithm. In this context, Dressler [13] advance a melody extraction algorithm based on an MR FFT whose aim is to extract the sinusoidal components of the audio signal. The MRS has also been used in speech enhancement [24] and speech synthesis [10].

The aim of this paper is to compare existing tools to a classic spectral analysis. We have implemented and applied a series of existing methods in the context of speech analysis with a focus on transition zones detection and silence/sonorant/non-sonorant classification.

II. MULTIRESOLUTION FFT

It is so difficult to choose the ideal window with the ideal characteristics. The size of the ideal window [4] 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 energy was concentrated at the center of the frame [16, p.41]. We estimated the center Cp of the frame and choosed an overlap equal to 75% [29]. The spectral $S_{i,k}(p_i)$ of each step *i* was :

Publication History
Manuscript Received : 22 April 2014
Manuscript Accepted : 25 April 2014
Revision Received : 28 April 2014
Manuscript Published : 30 April 2014

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

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%.

The multiresolution spectral MRS [29] was:

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

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

Diagrams displayed in Fig. 1 illustrates the difference between the standard FFT and the MRS. 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.

Fig. 1 Standard FFT (on the left) and MR FFT (on the right)

Fig. 2 shows the classical sonagram; Hamming window, 11 ms with an overlap equal to 1/3. The sentence pronounced is: "Le soir approchait, le soir du dernier jour de l'année".

Fig. 3 shows the multi-resolution sonagram of the same sentence. It offers several time integrations which are combinations of several FFT of different lengths depending on frequency bandwidth.

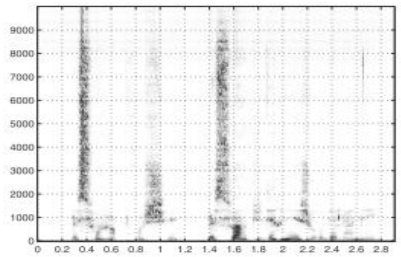

Fig. 2 Classical sonagram (Hamming, 11 ms, overlap 1/3) of this sentence: "Le soir approchait, le soir du dernier jour de l'année"

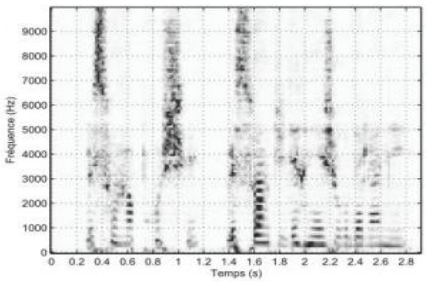

Fig. 3 MR sonagram; Hamming (23, 20, 15, 11 ms), overlap 75%, Band-limits in Hz were [0, 2000, 4000, 7000, 10000] of this sentence: "Le soir approchait, le soir du dernier jour de l'année"

III. PRINCIPLES OF THE EXISTING APPROACHES

Approach of Cheung

Cheung and Lim [9] presents a method for combining a wideband and a narrowband spectrogram by estimating the geometric mean of their pixels values. The combined spectrogram appears to preserve the visual features associated with high resolution in both the frequential and the temporal domain.

Approach of Chan

Chan and al. [7] describes a use of the MR analysis for spontaneous speech in a phone conversation. Their experiments show that MR cepstra result in a significantly lower error rates when compared to Mel-frequency cepstral coefficients.

Approach of Cancela

For music signals, Cancela and al. [6] presents two algorithms, efficient-constant-Q-transform and the multiresolution FFT, and compare them to a new proposal based on the Infinite Impulse Response filtering (IIF filtering) of the FFT. The depicted method appears to be a good compromise between design flexibility and reduced computational effort.

Approach of Dressler

In a melody extraction context, Dressler [13] is interested in describing the spectrum analysis for melody extraction based on multiresolution spectrograms. 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.

Approach of Shin

Shin and al. [28] combines a wideband and a narrowband spectrogram by calculating the average arithmetic and geometric mean of both spectra followed by a non-linear transformation and a spatial filter. This study is applied on medical signals.

IV. MATERIALS AND METHODS

Corpus

We used two corpora. The French corpus included CiVCiV with Ci being a stop consonant [p t k] and V a vowel [i e] [17]. The second corpus included read speech in Belgian French [26]. The sampling frequency was equal to 44.1 KHz, the wav format was adopted in mono-stereo.

VASP Software: Visual Assistance of Speech Processing Software

For our study, we created a prototype System for Visual Assistance of Speech Processing (VASP) [29]. VASP depicts sound on both time (by means of a waveform) and time-frequency (spectral representation; classical spectrogram - narrowband and wideband; spectrograms calculated with linear prediction and cepstral coefficients; Multiresolution spectrogram, etc.). Our system can automatically detect silence from speech on the basis of a waveform. From the spectrogram, the system can detect acoustic cues such as formants, and classify acoustic

landmarks automatically into classes: sonorant, silence, non-sonorant.

Analysis of Variance (ANOVA)

ANOVA provides a statistical tool for verifying whether the means of several groups are all equal, and therefore generalizes the test to more than two groups. We have obtained two types of results: the ANOVA table and the Tukey box-and-whisker plots [22,23,8,12].

Interquartile Range (IQR)

IQR is the distance between the 25th percentile and the 75th percentile [25]. The IQR is essentially the range of the mid 50% of the data which means that it not affected by outliers [25].

V. EXPERIMENTAL RESULTS

We realised two simulations where we tested for transition zone detection and silence/sonorant/non-sonorant classification.

Multiresolution Spectral Analysis (MRS)

A MRS was calculated for each speech signal. Hamming windows were chosen with the following properties: frame length = [23, 20, 15, 11] ms; frequency bandwith = [0-2000, 2000-4000, 4000-7000, 7000-10000] Hz; degree of overlapping = 75%.

Automatic detection of transition zones

We calculated the IQR for each frame. All components of a frame were real numbers between 0 and 255. Each diagram should allow us to clearly visualize the Tukey box plot and the areas of transitions between the different Tukey box-and-whisker plots and thus between the different classes. The length of each frame was 10 ms for classical spectrograms and 1.7 ms for MRS.

We calculated the Q1, Q2, Q3 and IQR for each frame and we plotted the Tukey box diagrams. We then presented our decision rules for transition zone detection and applied them to our corpora. We compared our results to experimental thresholds for Q3th, Q1th and IQRth [2,27].

Automatic Silence/Sonorant/Non-Sonorant detection

We defined as non-sonorants the fricatives and the stop consonants. All other sounds are defined as sonorants. When we found many successful sonorants or non-sonorants or silence, we considered them to belong to the same group. Fig. 4 shows an example of Multiresolution Spectrogram calculated with VASP.

We calculated MRS for each speech signal and applied our decision rules for classifying the signal into a silence/sonorant/non-sonorant class. For each MR FFT spectral frame number i, we performed an ANOVA for each group of N frames. All components of a frame were reals between 0 and 255.

Each diagram should allow us to clearly visualize the Tukey box plot and the areas of transitions between the different Tukey box plots and thus between the different classes. The length of each was fixed at 3 ms. We applied our decision rules for classifying the signal into a silence/sonorant/non-sonorant class. The hypothesis H0 to be rejected was that a frame was not classified as a non-sonorant. We calculated the probability p for each group and we plotted the Tukey box diagram [3].

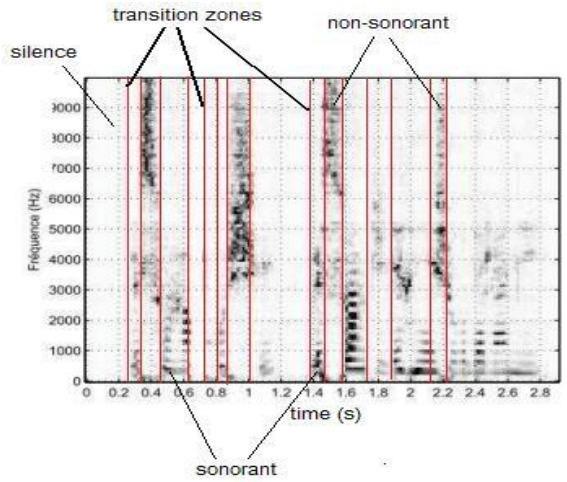

Fig. 4 Example of MRS calculated with VASP. Different classes are represented; silence/sonorant/non-sonorant and transition zones.

VI. DISCUSSIONS

In this study we presented and tested the performance of a method for automatically detecting transition zones by means of our proprietary VASP system, MRS analyses and two corpora. VASP is an inhouse system regrouping all needed tools and presents a visual improvement compared to standard spectrograms. It allows for better acoustic cue extraction and is an automated, open system. In comparison to Praat (freeware), VASP doesn't allow phonetic/phonemic transcription and includes offers less tools; it offers, nonetheless, more time-frequency representations and allows for an automatic detection of transition zones.

MR FFT were calculated for each signal and transition zones were detected by means of decision rules. We observed that the values for Q1, median, Q3 and IQR differed when the frame represented silence, a stop consonant or a vowel. Decision rules were defined on the basis of these variations. All values were compared to experimental thresholds for Q3th, Q1th and IQRth.

The first task consisted in detecting transition zones on the basis of IQR variation. This variation was presented as a graph with significant peaks which were detected by rules. For transition zone detection based on MRS FFT and IQR calculation, we obtained a score of 57.5%. The efficiency of classical spectral analysis and IQR calculation for the same task was of 23.75%.

In our second task, we classified each sound frame into silence/sonorant/non-sonorant according to decision rules. Our method for automatically detecting silence/sonorant/non-sonorant on the basis of MRS provided better results than classical spectral analyses. Detection was better and errors were fewer. For detection of silence based on ANOVA calculation, a score of 77% was obtained. The score of non-sonorant detection with MRS was 75% and that for sonorants was 80%.

When we calculated the mean and standard deviation for each frame, we observed an overlapping between the two groups; the silence class and the sonorant class and between

the two groups; the sonorant class and the non-sonorant class. The score of non-sonorant detection is less then the sonorant detection because it is related to the transition CV. When the energy on the spectrogram is not visible, the non-sonorant is omissed and considered as a sonorant. It's difficult to detect the locus of each non-sonorant. For ANOVA results, the scores were better to the mean and the STD. All results of detection for MRS analysis are better then the classical spectral analysis.

To evaluate the efficiency of our approach, we implemented and tested ten methods found in literature; Table 1 summarizes the obtained detection rates. Peak distinction representing the transition zones was rather low in the case of wideband spectrogram SBL1 and SBL2, the rates of which did not exceed 50 %. Similar results were obtained for Cheung's M1 and M2 approaches.

This was due to the limited resolution of the output spectrogram. Rates of detection average characterize the spectrogram to narrow band SBE and the spectrogram of the Cheung approach M3 the rate of which amounts to 60 %.

TABLE I RESULTS OF COMPARATIVE STUDY

Results	**% Zone Detection**	**% silence**	**% sonorant**	**% non-sonorant**
S BL 1	20	64	65	40
S BL 2	33.3	70	40	33
S BE	60	65	60	33
S Dressler	67	75	67	67
Cheung M1	30	20	30	35
Cheung M2	50	25	34	37
Cheung M3	60	27	37	38
Cancela	65	76	64	70
Shin	74	75	66	67
MRS	84	77	80	75

In the case of the approach of Cheung M1, M2 and M3, It was to be difficult to classify correctly frames. This is due to the quality of the resultant spectrogram. The rates of classification not overtaking the 38%. In the case of the wide-band classic spectrograms SBL1 and SBL2 and to narrowband SBE, the classification rates are considered rather average, even mediocre, for certain classes. They are between 33% for non-sonorants and 70% for silence.

We implemented and tested the approach of Cheung of three different manners. M1 corresponds to combining a 128 samples and a 512 samples spectrogram. M2 combines a 128 samples and a 1024 samples spectrogram. For M3, it is 512 and 1024 samples. We noticed that matrice results were relatively fuzzy, in the sense that the borderline edges of the energy zones are not clear. This lack of precision around the edges influenced the accuracy of transition zone detection as well as the classification rates for silence/sonorant/non-sonorant.

Indeed, these results were lower than 60 % for the detection of the zones of transition and varied between 20 % and 38 % for the classification.

The wideband and narrowband spectrograms were characterized by a variation of energy among lower and higher frequencies, which influenced frequency and time precisions in function with the size of the applied window.

Detection rates for Dressler's, Cancela's and Shin's methods were more similar to those for our MRS approach. The rates of detection of the transition zones was until 74 % for the approach of Shin.

Our approach MRS give better result with a rate of 84 % of detection. This rate strengthens the relevance of our approach as well as its precision during has it display of the energy zones characterizing phonemes in the spectrogram MRS.

The classification results of Dressler, Cancela and Shin approaches yielded similar results as our MRS approach. These rates were due to these approaches clear resolution.

The tested methods were initially implemented and applied for domains other than speech analysis.

We implemented them and adapted them to our working context; the opposite was not possible since we didn't have access to those corpora.

Our MRS approach, however, revealed a better classification rate for about 80% of the sonorants.

This experimental rate supported the theoretical contribution of the MR representation and its faculty to highlight the phonetic segments energy zones.

VII. CONCLUSIONS

In this paper, we showed that classification rates for our MRS analysis turned out better than those of the traditional spectral analysis as well as those of the other existing multiresolution methods that we studied. We also provided evidence that the MRS method yields better results for silence/sonorant/non-sonorant classification when coupled with ANOVA rather than when it is based on standard deviation and mean. For non-sonorants, other decision rules have to be implemented before a phoneme classification and recognition can be attempted.

REFERENCES

[1] N. Annabi-Elkadri and A. Hamouda. Spectral analysis of vowels /a/ and /E/ in tunisian context. International Conference on Audio, Language and Image Processing, Novembre 2010.

[2] N. Annabi-Elkadri, Automatic Detection of Transition Zones in Tunisian Dialect, International Journal of Advanced Science and Technology, Vol. 60, p. 67-82, November, 2013

[3] N. Annabi-Elkadri and A. Hamouda. Automatic Silence/Sonorant/Non-Sonorant Detection based on Multiresolution Spectral Analysis and ANOVA Method. International Workshop on Future Communication and Networking, Szczecin, Poland, 2011. IEEE.

[4] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. Traitement de la parole. Presses Polytechniques et Universitaires Romandes, 2000.

[5] Calliope. La parole et son traitement automatique. collection technique et scientifique des télécommunications, MASSON et CENT-ENST, Paris, 1989.

[6] P. Cancela, M. Rocamora, and E. Lopez. An e cient multi-resolution spectral transform for music analysis. 10th

International Society for Music Information Retrieval Conference, pages 309–314, 2009.

[7] 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. International Conference on Acoustics, Speech, and Signal Processing, volume 1, pages 405–408. IEEE, Mars 1999.

[8] A. Chauvin and R. Palluel-Germain. Les principes de l'Anova . Journées RJCP, 2011.

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

[10] T. Chi and C. Hsu. Multiband analysis and synthesis of spectro-temporal modulations of fourier spectrogram. Journal of the Acoustical Society of America, 129(5):EL190–EL196, May 2011.

[11] L. 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.

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

[13] K. Dressler. Sinusoidal extraction using an e cient implementation of a multi-resolution FFT. Proceeding of the 9th International Conference on Digital Audio E ects, pages 247–252, September 2006.

[14] Q. Fu and E. 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.

[15] 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.

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

[17] C. Karypidis. Asymétries en perception et traitement de bas niveau: traces auditives, mémoire à court terme et représentations mentales. PhD thesis, Université Paris3-Sorbonne Nouvelle, Paris, France, 2010.

[18] Ladefoged. Elements of Acoustic Phonetics. University of Chicago Press, 1996.

[19] H. Leman and C. Marque. Un algorithme rapide d'extraction d'arêtes dans le scalogramme et son utilisation dans la recherche de zones stationnaires. Traitement du Signal, 15(6):577–581, 1998.

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

[21] S. Mallat. Une Exploration des Signaux en Ondelettes. Editions de l'Ecole Polytechnique, Ellipses di usion, 2000. □

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

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

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

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

[26] Audiocite (2011). 'Belgian French Corpus'.

[27] N. Annabi-Elkadri and A. Hamouda, The Multiresolution Spectral Analysis for Automatic Detection of Transition Zones, International Journal of Advanced Science and Technology, Vol. 36, p. 95-110, November, 2011

[28] L. Shin, et al. (1997). 'Visual imagery and perception in posttraumatic stress disorder : A positron emission tomographic investigation'. Archives of General Psychiatry 54 :233–241.

[29] N. Annabi-Elkadri. Spectre à Multirésolution dans l'Analyse et le Traitement de la Parole. PhD thesis, Université Tunis-ElManar, Faculté des Sciences de Tunis, Tunis, Tunisie, 2014.