Impact of Acoustic Parameters on Speech Signal Processing: Optimization of the Recognition for Arabic Numerals

Lamkadam Abdelmajid[1] Karim Mohamed[2]

[1,2]LISTA, Faculty of Sciences, University of Sidi Mohamed Ben Abdellah, P.O Box 1796 Atlas, Dhar El Mehraz, Fez, Morocco

Abstract— In this work, the general characteristics [2] of a speech signal have been presented and described, and they have been organized judiciously according to the different aspects [5] that can be contained and characterized by such a signal, and Cited various tools [10] that are currently used in a signal, to arrive at a clear and fairly complete description of the latter. Indeed, there are a large number of characteristics that can be extracted from the speech signal; we cite the acoustic, phonetic, morphological characteristics ... Also those which make it possible to discriminate particular classes [8] sound. However, given our goal of a complete description of the speech signal, we will focus more on general characteristics not related to particular classes. Finally, the speech signal can be processed by a set of mathematical and physical tools [8], also via various acquisitions and editing equipment. Once the processing steps have been completed, acoustic vectors [10] will be present which will facilitate their use for recognition [3].

Key words: Speech Signal, Channel Analysis, Speech Processing, Acoustic Characteristics, VAD, Segmentation, Windowing, Extraction, DTW, ASR

I. INTRODUCTION

Voice technologies sweep our daily environment. Interactive voice systems [1], onboard voice control, and a large number of products are currently available on the market (Games, mobile phone..) to the professional (TV, security, avionics..).

In order to fully exploit this speech signal, the multiple signal processing techniques [3] are used, in order to benefit from perfect analysis, optimal processing, correct coding, and tricks of calculations sufficient to extract the information that completely represents the true acquisition signal [5] to subsequent uses.

In addition, voice interaction allows interacting during treatment with interfaces in the user environment, allowing fighting against the success of the task. For example, a simple orientation of the microphone may cause annoying disruptions good recognition [10].

Before proceeding further, it is advisable to identify and specify the maximum of the characteristics in order to surround and to capture more this speech signal in its totality.

Therefore, the aim of this work is to optimize the processing chain [10]; and selecting the relevant parameters [13] from the occurrence of the existing values; in order to improve and model the analysis and treatment procedure.

II. PROBLEMS AND OBJECTIVES

A. Description of the Acoustic Signal

1) General Signal

The signal is generally a complex wave which can be considered as a linear combination of sinusoidal single waves of different frequencies.

$$\psi(t) = \psi_0 e^{i(n\omega t+\varphi)} + C_0 \quad (1)$$

For simple signals, and taking into account the non-complexity of the calculations, we can write:

$$\psi(t) = \sum s(t) g_n(t) \qquad n\epsilon[1, +\infty] \quad (2)$$

s(t) Is the signal to be analyzed and $g_n(t)$ are the functions of representation as sinus and cosinus (ideal functions for this representation).

2) Sound Signal

More generally, in the indefinite domain of frequencies, taking into account parameters and characteristics [2] of the signal, and without neglecting certain initial conditions, we arrive at the simple sinusoidal form:

$$s(t) = A_0 \sum \sin(n\omega t + \varphi_0) + B_0 \qquad n\epsilon[1, +\infty[\quad (3)$$

The simple sound signal that exists is the monochromatic sound, which is a constant amplitude vibration, on a single frequency, during indefinitely, like a continuous whistle. To make the signal carries information, it must be modulated; and applied for duration T_0, with amplitude A_0.

$$s(t) = A_0 \sin[\omega t + \varphi_0] + B_0 \quad (4)$$

3) Speech Signal

Speech is thus a series of sounds produced either by vibrations of the vocal cords (quasi-periodic source of voicing) or by turbulence created by the air flowing in the vocal tract or when an occlusion of that voice is released (Noise sources and unvoiced sounds).

In reality, we can't describe the signal of speech in a comprehensive manner, it always presents ambiguities.

B. General characteristics of the speech signal

To properly prepare this vibration of speech for good recognition, the elementary components [13] must be identified and described with a judicious and detailed manner.

Here we have cited some important notions and characteristics [6] of the speech signal, in order to highlight the problems posed during its treatment. Then the signal of the speech can be represented taking into account several notions [5] such as:

- Prosody,
- Acoustic,
- Phonetic,
- Spectral
- Morphology,
- Semantics,
- Pragmatic,
- ...

Taking into account of all these aspects, i.e., make a maximum analysis and an effective and selective treatment for the treatment parameters, allows us to reduce the error rate and increase the recognition rate. Really is not the case.

C. Classification and organization of acoustic parameters

These characteristics can be classified according to whether they are input characteristics (test parameter sets to be

modified), or output characteristics, as parameters resulting from the application of processing techniques on the realized corpus.

1) *Input Parameters*

The values of these parameters are gradually varied (Illustrated in the Table 1), in order to study and analyze the effects of these variations on speech recognition.

Parameters	Aspect				
	Temporal	Frequency	Spectral	Statistical	Determinist
Recording time	x				
Sampling		x			
Coding				x	
Formatting	x				
Segmentation	x				
Windowing	x				
Splitting				x	
Amplification	x				
Filtering		x			
Emphasizing		x			
Smoothing	x				

Table 1: Input parameters according to the analysis aspects.

2) *Output parameters*

These parameters (Illustrated in the Table 2) are taken into account to evaluate the effects of changes in input parameters.

Parameters	Aspect				
	Temporal	Frequency	Spectral	Statistical	Determinist
Energy, Intensity	x				
Amplitude, Height	x				
Average, Max, Min	x			x	
Covariance, Deviation				x	
Pitch (F_0), Formants		x			
Spectre			x		
Cepstre	x				
Stamp			x		
Zero crossing rate	x				
Partition function					x
Spectral flow			x		
Spectral centroid			x		
Spectral rolloff point			x		

Table 2: Output parameters according to the analysis aspects.

D. Analysis and Signal Processing

As shown in our study [10], the processing method is schematized in the following figure; this is a digital processing to make the signal to handle, effective and robust. In this work we interested only to the part of acquisition and analysis (Segmentation and Windowing).

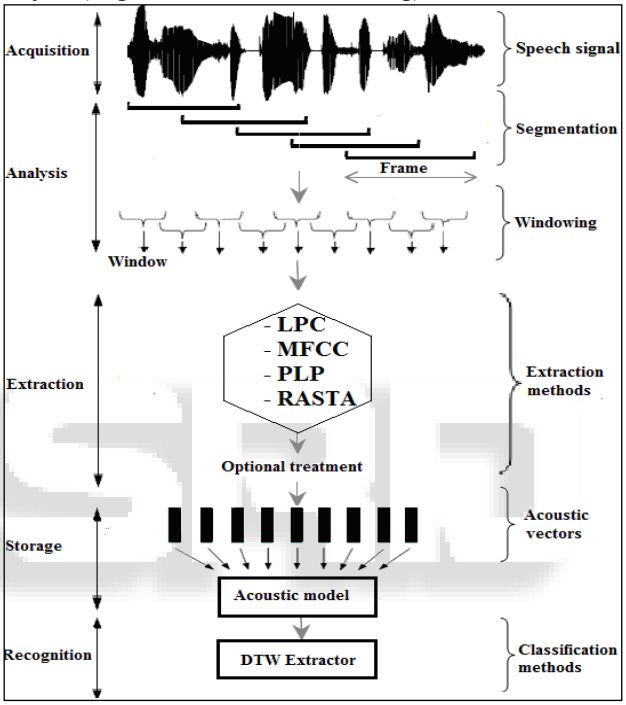

Fig. 1: Channel analysis and signal processing.

III. MATERIALS AND METHODS

A. Context of work

This work is carried out in an empirical manner, and without any clear methodology to guide us in the choice of characteristics to be included in the acoustic vector. Our objective is to propose and use several analysis parameters encountered in the field of ASR [9] (Automatic Speech Recognition).

The extraction is done by combining the 4 methods (LPC-MFCC-PLP-RASTA) of extraction; already realized in our work [10]. To facilitate the phase of the recognition it was limited to the DTW extractor.

B. Preparation of the reference corpus

The choice of equipment and acquisition [2] conditions have a direct reflection on the quality of the speech signal, consequently on the quality of the recognition. As our comparative study [10] of acoustic vectors, The Table 3 collecting the acoustic characteristics, to realize the basic references containing Arabic numbers (0 to 9). The corpus record is done for one Moroccan speaker.

Parameter	Value
Format	Mono, (.wav)
Sampling	8 Khz
Codage	16 Bits
Frames number	50
Recording time	5 Secondes/digit
Windowing	Hamming
Corpus	10 Arabic numerals

Table 3: Acoustic characteristics of the speech corpus.

Each number {0 to 9}, constituting the basic test given by a single speaker; tested with other numbers {0 to 9} (forming the base for references), these files are distributed as follows:

- Directory "dictionaries" contains 100 files form the references: 10 digits, each digit pronounced 10 times (10 trials).
- Directory "tests" contains 10 files form the tests: 1 digit, each digit pronounced once (1 trial).

C. Registration and formatting the test corpus

The process of learning [11] and manipulation consists in changing the acoustic parameters of acquisition and analysis each time in both parts: reference and test. For the other parameters are fixed to the default values (most used values).

D. Parameters used and changed during the tests

- Data formats.
- Sampling frequencies.
- Number of coding bits.
- Number of frames.
- Duration of frame.
- Windowing.

E. Cleaning and strengthening of the signal

This task is performed directly on the recorded "sounds" files, using the mathematical tools [3] known in signal processing. The different pre-processing sound effects [6] that a speech signal can undergo are summarized as follows:

- Delete Silence [9] (Silence-Speech Separation): Reduce the stored files sound on the memory.
- Voice Activity Detection (VAD)[7].
- Filtering (Speech-Noise Separation): Multiply the analyzed signal in the frequency domain by a weighting function.
- Emphasis: Increase the amplitude of the treble sounds, or reduce the amplitude of the bass sounds.
- Smoothing [6]: Remove the insignificant segments from the assembly.
- Amplification: Multiply the signal by the windowing signal (Hamming), keep and amplify that the main part of the signal.

F. Recognition Rating Factor

To evaluate the empirical results obtained, a Recognition Performance Factor [12] was used RPF [10] defined as follows:

RPF=(Number of recognized trials)/Number of total trials)*100 (5)

G. Signal Comparison Domains

These are the two domains [7] of presentation and comparison of the signal, the temporal and spectral domain, the passage is ensured by application of the FFT and FFT^{-1}.

1) *Temporal Comparison*

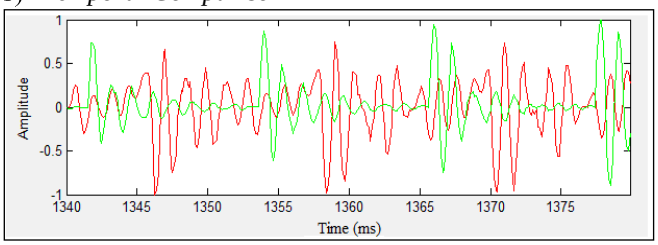

Fig. 2: Comparison of two signals in the temporal domain.

2) *Spectral comparison*

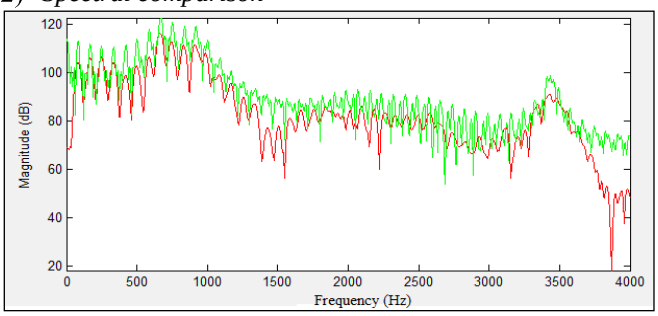

Fig. 3: Comparison of two signals in the frequency domain.

It clearly appears that the comparison of the signal in the frequency domain and more reliable and clear than that in the time domain.

IV. Results and Discussion

A. Quantification Effect: (Amplitude Cutting)

This number reflects how many bits of coding treated simult aneously; it follows a law of 2^n. The Table 4 shows the result of this effect.

1) *Results*

Coding (Number of bits)	Recognition RPF (%) for {0-9} digits	Processing Time (s)
8	90	4 (Most used)
16	90	4
24	90	4
32	90	4

Table 4: Result of the coding effect on the recognition.

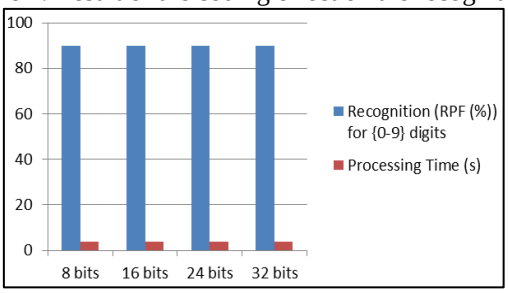

Fig. 4: Effect of coding on recognition and processing time.

2) *Observation and Discussion*

- The increase in the number of coding bits necessitates the increase of the recording time.
- Increase the number of coding bits has no effect on the recognition or on the processing time.

B. Sampling effect: (Sampling frequency: Fe)

To correctly represent the signal, the compatible sampling frequency must be used. The result is shown in Table 5.

1) *Results*

Sampling frequency(Khz)	Recognition RPF (%) for {0-9} digits	Processing Time (s)
8 (Telephony)	60	3 (Most used)
11,025	40	4
16 (PC, Audio conferencing)	40	7
22,05 (PC, Audio ADPCM)	60	18
32(Radio diffusion: DAB, NICAM)	70	12
44,1 (CD Audio, Digital Studio, DAT)	70	52
96 (Modem, Telecom)	50	636

Table 5: Result of the sampling effect on the recognition.

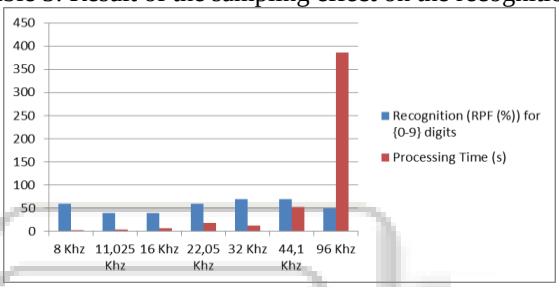

Fig. 6: Effect of sampling on recognition and processing time.

2) *Observation and discussion*

- The most used sampling frequency (8 kz) is fast but has an average recognition rate.
- When the sampling frequency increases the recognition and the processing time also.
- For the transmission frequency (modem, telecom), an average recognition rate is cited, but very slow.

C. Effect of the data format

This is the format for recording [10] of the sound track in the file. The test result is stored in the Table 6.

1) *Results*

Formatting (Data format)	Recognition *RPF (%) for {0-9} digits*	Processing Time (s)
uint8	0	1
int16	30	1
double	80	1 (Most used)
single	100	1

Table 6: Result of the data format effect on the recognition.

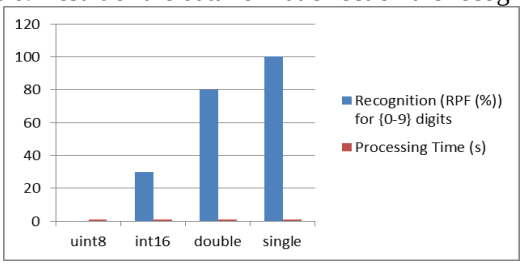

Fig. 5: Effect of formatting on recognition and processing time.

2) *Observation and discussion*

- The data format does not affect the processing time.
- The best data format is "single", is different from the most used one.

D. Effects of the number of frames: Segmentation

This number (form 2^n) avoids artifacts related to the side eff ects during the transformation of the temporal [7] domain in the frequency domain. Table 7 shows the result of segmentation [4] of the signal over several frames.

1) *Results*

Fragmentation (Number of frames)	Recognition RPF (%) for {0-9} digits	Processing Time (s)
8	70	0.5
16	50	0.5
32	70	0.7
64	90	1.5 (Most used)
128	70	3.8
256	90	11.4

Table 7: Result of the splitting effect on the recognition.

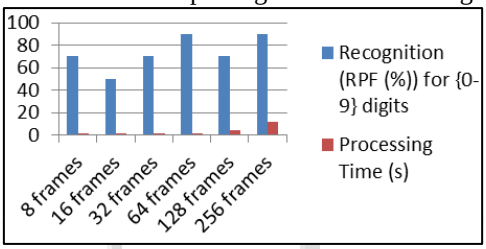

Fig. 7: Effect of fragmentation on recognition and processing time.

2) *Observation and discussion*

- As the number of frames increases, the processing time also increases.
- As the number of frames increases, the processing time and the recognition rate also increase.

E. Effect of the duration of the frame

The sampling period is the time separating two successive samples [7], is the inverse of the period: Te=1/Fe. This parameter is related to the number of frames, if the total number of samples is N, the signal will have a total duration of NTe. The time scale will therefore extend from 0 (s) to NTe = N / Fe (s).

The report of this effect is inversely proportional to the results obtained for the number of frames.

F. Windowing effect

In order not to lose important information at the beginning or at the end of windows, and to attenuate the discontinuities, for each slice there is an offset of 10 ms, called windowing technique [11], for have a continuous sound signal of reasonable size compared to the computing capabilities of the recognition systems. The speech signal is thus transformed into a series of vectors computed for each frame [7]. The results of the application of windowing types on recognition are shown in Table 8 below.

1) *Results*

Windowing type	Recognition RPF (%) for {0-9} digits	Processing Time (s)
Hanning	70	1.4

Hamming	80	1.5 (Most used)
Rectangular	50	1.6
Triangular	60	1.6

Table 8: Result of windowing effect on the recognition.

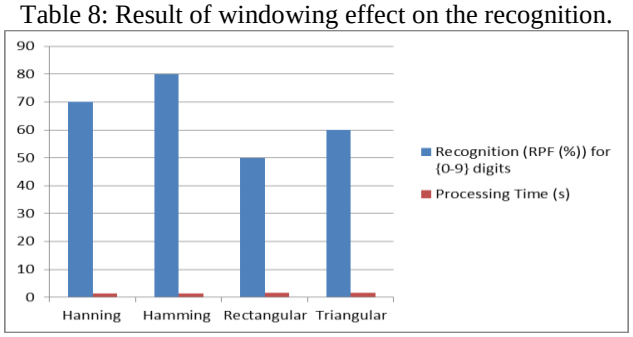

Fig. 8: Effect of windowing on recognition and processing time.

2) *Observation and discussion*

– The processing time varies slightly between these windowing methods.
– The best method of windowing is that of hamming (more used).

G. Best result collection

1) *Results*

In the table (Table 9) the best values for the parameters used in the higher tests were identified, in order to arrive at the selection of the relevant parameters [13].

Parameters	Best Values	Recognition RPF (%) for {0-9} digits
Coding	All Values	90
Sampling	32 Khz	70
Formatting	Single	100
Fragmentation	64	90
Windowing	Hamming	80

Table 9: Collection of best values of parameters applied to the speech signal.

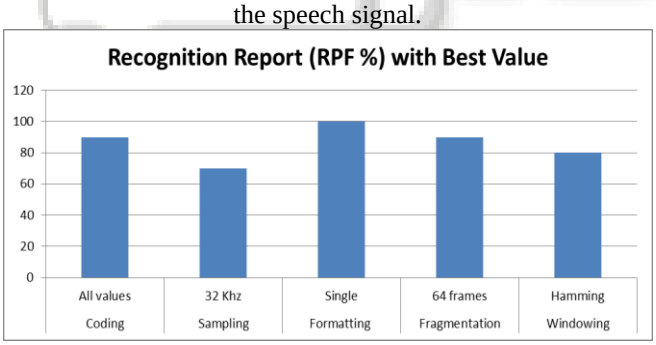

Fig. 9: Result of recognition with best value.

2) *Observation and discussion*

Using these parameters selected, and their best values, a rather high recognition has been obtained; with optimal processing time.

V. Conclusion

Recognition is more efficient by examining some primordial values, only as the data format, frame duration, and fragmentation report. By adding or combining other characteristics, such as scanning parameters, and windowing ... We got a considerable rate of recognition, with an acceptable processing time.

In order to improve the performance of the system, the choice of these parameters must be judicious and careful to better separate the speech signal to that of any disturbances and parasites (silence, noise, degradation, assembly, crackling, melting, clicks and cracks ...).

It should be noted that its acoustic vectors do not present the entire signal in question. The human perception is modeled by the frequency and spectral characteristics, since they are the same characteristics as the human brain.

References

[1] H. Cerf-Danon, M. El-Bèze, B. Merialdo "Reconnaissance automatic de la parole", Informatique et Santé, Volume 4 - 1991, Paris, Springer-Verlag France.
[2] Lotfi Amiar, Mokthar Sellami " Un système basé sur une modélisation Markovienne pour la reconnaissance de la parole Arabe", SETIT, March 27-31, 2005 - Tunisia.
[3] Stéphane Dupont "Etude et development d'architectures multiband et multi-models pour la reconnaissance robuste de la parole", Laboratories TCTS, Juin 2000, Mont France.
[4] Ali Sadiqui, Noureddine Chenfour "Modélisation statistiques basée sur la morphologie pour la langue arabe", Annals. Computer Science Series. 8th Tome 2nd Fasc. 2010.
[5] Jorge Arturo GUTIÉRREZ CELAYA "Fusion d'informations en identification automatique des langues", 25 juillet 2005, Toulouse III, France.
[6] Lise REGNIER "Détection de la voix chanté dans un morceau de la musique", IRCAM, 2008, 75004 Paris, France.
[7] Marco Martalò* et al. "Low-Complexity Hybrid Time-Frequency Audio Signal Pattern Detection", IEEE Sensors Journal, vol. 13, no. 2, February 2013.
[8] Asma Rabaoui* et al. "Sélection de descripteurs audio pour la classification des sons environnementaux avec des SVMs mono-classe", Colloque GRETSI, 11-14 septembre 2007, Troyes. France.
[9] Johan Olsson "Text Dependent Speaker Verification with a Hybrid HMM/ANN System", CTT, November 2002, Stockholm.
[10] A.LAMKADAM et al. "Comparative study and improvement of acoustic vectors extractors: Multiple streams applied to the recognition of Arabic numerals", ISCV 2015, 25-26 MARCH, FEZ, Morocco.
[11] Anne Spalanzani "Algorithmes évolutionnaires pour l'étude de la robustesse des systèmes de reconnaissance automatique de la parole", Soutenue 28 Octobre 1999, Université Joseph Fourier - Grenoble I, France.
[12] Hachem KADRI* et al. "Speaker Change Detection Method Evaluated on Arabic Speech Corpus", ISCCSP, 13-15 March 2006n Marrakech, Morocco.
[13] Abdenour Hacine-Gharbi "Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole", Soutenue le 09 décembre 2012, Université d'Orléans. France.

