Towards an Automatic Speech-to-Text Transcription System: Amazigh Language

Ahmed Ouhnini[1]*, Brahim Aksasse[2], Mohammed Ouanan[3]
Dept. of Computer Science FST, Moulay Ismail University, Errachidia, Morocco[1]
Dept. of Computer Science FS, Moulay Ismail University, Meknes, Morocco[2, 3]

***Abstract*—Various studies inside the domain of research and the development of automatic speech recognition (ASR) technologies for several languages have not yet been published and thoroughly investigated. Nevertheless, the unique acoustic features of the Amazigh language, for example, Amazigh's consonant emphasis, pose many obstacles to the development of automatic speech recognition systems. In this study, we examine Amazigh language voice recognition. We treat the problem by focusing on transitions in vowel and consonant sounds and formant frequencies of phonemes. We present a hybrid strategy for phoneme separation based on energy differences. This includes analysis of consonant and vowel features, and identification methods based on formant analysis.**

Keywords—Speech recognition system; Amazigh language; analyzing formants and pitch; speech corpus; artificial intelligence

I. Introduction

Automatic Speech Recognition (ASR) is used to transcribe human speech captured via a microphone into text that computers can understand in order to enhance human-machine (HM) communication. ASR has long been the subject of intense research. Formant frequencies have been studied for decades. ASR has long been a topic of active investigation. For many years, formant frequencies are believed to be an important factor in recognizing speech phonetic content [1]. To arrive at the stage of recognizing phonemes, we examine a specific case of this problem and focus on the vowels and consonants transition in the Amazigh language and the formant frequencies of phonemes that are important for determining the phonetic content of speech [2]. We trust that this effort, will highlight the importance of consonant-vowel changes and vocal parameter analysis in speech recognition.

We provide a strategy that includes separation of phonemes by differences in energy between consonants and vowels, vocal characteristics processing of phonetics units, and a recognizer algorithm based on formants. Formant analysis methods focus on associating physical aspects with phonology. This method determines speech types by analyzing linguistically distinct features of the speech.

The rest of this article is organized as follows: Section II describes the characteristics of the human voice and the Amazigh language. The remainder of Section III describes some mathematical and engineering techniques for language modeling, followed by a discussion of proposed phoneme recognition methods. Section IV provides further insight into the results and Section V concludes the article.

II. Characteristics of Speech and Phonetics

A. Human Ear and Acoustic Sound

Sound constitutes a wave that propagates in a material environment like small variations of pressure. This is perceptible via human ears at frequencies ranging from 20 Hz to 20 kHz. Nonetheless, the phonetic information is judged to be less than 10 kHz [3].

Due to the way sound signals are sifted, we modify the sufficient range because ears are not sensitive to stage distortion. This allows us to focus exclusively on complementary application modules.

B. Human Voice

The human voice is a collaboration of breathing and multiple phonatory organs. In voiced phonemes, sound is first produced by the vibration of the vocal cords [4]. It is manipulated differently depending on the cavities it passes through, primarily the pharynx and mouth. These cavities act as resonators, increasing frequencies corresponding to the resonant frequencies of specific phonemes. These enhanced frequencies are known as "formants" and are the features that phonologists search in spectrograms to identify phonemes being pronounced [5].

The Fig. 1 shows the formants (F1, F2, and F3) superimposed on spectrogram of speech signal « he took holidays » showing the alternating voiced and unvoiced sounds. In the voiced situation, a formant structure is presented.

C. Phonology and Phonetic

Phonetic by definition is study of phonetic units, the smallest particular phonetic unit being regularly defined as a phoneme. The opposition between the terms bath and bread, well, suggests that [b] and [p] are phonemes. In general, we can classify then as follows: classes and subclasses, more significant of which is "vowel" and "consonant".

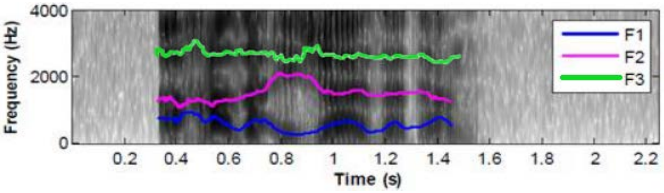

Fig. 1. Illustration of formants F1, F2, F3 of speech signal "he took holidays".

*Corresponding Author.

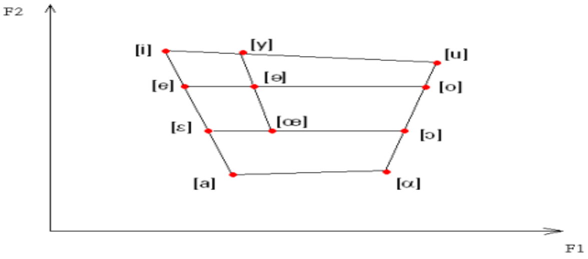

Fig. 2. Categorization of vowels using the vocal trapezoid.

Relatively vowels are lengthy-duration sounds with a flows and with considerable frequency characteristics consistency during times: in absence of highly pronounced prosodic features, formants appear horizontal on the spectrogram [6]. The vowel trapezius, shown in Fig. 2 illustrates their location in the planes specified by initial both formants, indicating articulation position of language.

However, consonants represent phonemes that encounter an obstacle when articulated (such as labial vowels, toothy teeth, palate closure in [k], etc.). In comparison, they are considerably shorter than vowels and significantly variable in length during time. Could be sonorous or loud. In the resonant scenario, only current formants are present, see the Fig. 3.

D. *Speech Signal Frequency Parameters*

The bandwidth of voice signal is much larger than the telephone bandwidth (4 kHz) and includes all information's necessary to know to decode human voice.

The fundamental frequency refers to the speed of opening and closing of the vocal chords during phonation. Its value is proportional to the individual's phonatory system size [7]. Voice frequency vary between 80 and 600 Hz based on age and gender.

The spectrogram is a representation in three dimensions, where the X-axis represents time, the Y-axis represents frequency, and the Z-axis represents frequency levels (symbolized by gray levels). Fast Fourier transform (FFT) with sliding window is applied to acquire the voice signal.

E. *Amazigh Language*

The Amazigh language, commonly called Tamazight or Berber, is one of humanity's earliest languages. Now it extends from the Red Sea to the Canary Islands and from Niger in the Sahara to the Mediterranean Sea, including the northern section of Africa.

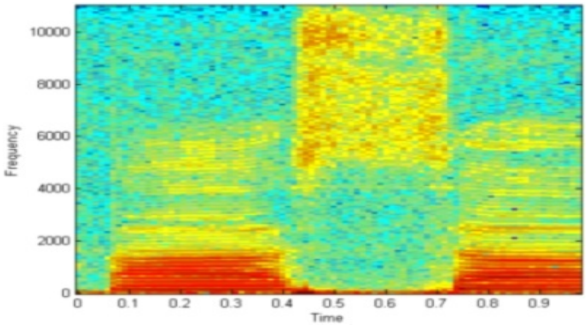

Fig. 3. Spectrogram of the syllable (asa).

In Morocco, Amazigh language is classified into three main regional's varieties, based on historical, geographical, and sociolinguistic factors: Tarifite in the north, Tamazight in Central Morocco and the south-east, and Tachelhite in the south-west and the High Atlas. Even though half of the Moroccan population are Amazigh speakers, the Amazigh language has been reserved exclusively for informal and familial domains: Boukous (1995) [8]. In the last decade, with royal generosity, the language was institutionalized and included in Moroccan school system.

Since February 2003, Morocco's official Amazigh alphabet system, known as Tifinaghe-IRCAM, has been used in Moroccan school programs and Amazigh historical studies. The system uses the alphabet describe in Table I [9]:

TABLE I. AMAZIGH LANGUAGE ALPHABET

27 Consonants	***labial***	ⴼ, ⴱ, ⵎ
	dental	ⵜ, ⴷ, ⵟ, ⴹ, ⵏ, ⵔ, ⵕ, ⵍ
	alveolar	ⵙ, ⵣ, ⵚ, ⵥ
	palatal	ⵛ, ⵊ
	velar	ⴽ, ⴳ
	labiovelar	ⴽⵯ, ⴳⵯ
	uvular	ⵇ, ⵅ, ⵖ
	pharyngeal	ⵃ, ⵄ
	laryngeal	ⵀ
2 Semi-consonant	ⵢ and ⵡ	
4 Vowels	***full vowel***	ⴰ, ⵉ, ⵓ
	Neuter vowel (schwa)	ⴻ it has a specific status in phonology Amazigh.

The correct writing of words in the Latin letters closely resembles phonetic transcription and correctly conveys their pronunciation, includes twinned and vowel sounds.

Examples: /illa/—> « illa » —> « il existe », « il est ». [10]

The pronunciation of Amazigh language varies from region to region [11].

III. TECHNOLOGIES AND METHODS

A. *Fourier Analysis*

Transform of Fourier permits a time-frequency processing at a resolution suitable for speech signals that are quasi-stationary on intervals of 10-100 ms.

1) Fourier transform: We are dealing with the pre-Hilbert sets of square integrables function $L^2(\mathbb{R})$, and the orthogonal family of sine functions $e_f: t \rightarrow e^{j2\pi t} / f \in R$ (we restrict ourselves physically in concret pulses)). Preparing the projection portion of each sinusoid provided by the scalar product, and not ignoring the complex conjugate with respect to the second number, the Fourier transform H(f) (or FT(g(t)) of a functions g(.) allows projection over the vector space through which the sinusoids:

$$\mathrm{FT(f)} = < \mathrm{g(t)}|\mathrm{e_f} > = \int_{-\infty}^{+\infty} \mathrm{g(t)e^{-j2\pi ft}\,dt} \qquad (1)$$

Where the variables t and f refer to time and frequency, respectively. FT(f)(t) is the transform of g(.).

These transform would be employed in our study to investigate the contributions of each frequency range in the speech signal more qualitatively by evaluating the spectrogram, (Ohm's rule states that human ear is insensitive to the acoustic signal's phase) [12].

2) Discrete fourier transformation: The signal has now been expressed by sampling taken evenly throughout time by sampling continuous time $\{x(n)/n \in [0, N-1]$

To avoid edge effects, we convert them into an N-period signal. Following that; the discrete transform expressed as:

$$\mathrm{F(k)} = \sum_{n=0}^{N-1} x(n)e^{\frac{-2i\pi kn}{N}} \tag{2}$$

The frequency observations is related to the $\frac{k}{N}$ factor for $k\epsilon \left[0, \frac{N}{2}\right]$ If f_e represents the sample frequency, so $= f_e \times \frac{k}{N}$. It should be emphasized that because the received information is restricted to half of the period $[0, \mathrm{N}-1]$, this transformation is redundant. Indeed:

$$\forall k \in Z \ \ \mathrm{F(N-k)} = \sum_{n=0}^{N-1} \mathrm{x(n)e}^{\frac{2i\pi kn}{N}} \mathrm{e}^{-2i\pi n}$$
$$= \sum_{n=0}^{N-1} \mathrm{x(n)e}^{\frac{2i\pi kn}{N}} = \overline{\mathrm{F(k)}} \tag{3}$$

As a result, modules are similar, with the term simply having the reverse indication. If we wish to evaluate the signal's about frequency scale of 10 kHz, we need to use a sampling frequency with 20 kHz [13]. Within reality, we employ the FFT (Fast Fourier Transform), which has a computational cost of O(Nlog2(N)) rather than O(N^2) for straight computing, also use this redundancy to improve the computation.

B. Convolution and Transformation

The convolution product in continuous-time of two functions is given as:

$$\forall t \in R, (f * h)(t) = \int_{-\infty}^{+\infty} f(\tau)h(t-\tau)d\tau \tag{4}$$

Convolution operators are commutative. Also, since there are associative equations, the following two are valid:

$$\forall t \in R, (f * h)(t) = (h * f) \tag{5}$$

As a consequence of commutativity:

$$\big((h * g) * f\big)(t) = ((h * f) * g)(t) \tag{6}$$

The Fourier transform of the two functions ordinary product is the convolution product of Fourier transforms. In addition, the Fourier transform of the two functions convolution product is the usual product of Fourier transforms:

$$\begin{cases} FT(g * h) = FT(g) \times FT(h) \\ FT(g \times h) = FT(g) * FT(h) \end{cases} \tag{7}$$

This finding is also valid for discrete cyclic representations. It will be applied in formants analysis as a result of the speech signal modeling adopted.

C. Windowing Issue

Practically, in addition to the **x**-signal discrete, the observation time of 2τ has over. Consequently, we see the signal convolution using a window function [14]:

$$\prod_\tau(t) = \begin{cases} 1 \ if \ t \in [-\tau, \tau] \\ \ \ 0 \ if \ not \end{cases} \tag{8}$$

As according (7), Fourier transformation of $s(t) \times \prod(t)$ being a convolution product between window and signal, following equation give gate function:

$$\int_{-\infty}^{+\infty} \prod(t) e^{-i\omega t} dt = \int_{-\tau}^{\tau} e^{-i\omega t} dt = 2\tau \operatorname{sinc}(\omega\tau) \tag{9}$$

This cardinal sinus has a central lobe with a width of 2/τ. As a consequence, when the observation period approaches zero, the spectrum expands. To resolve this issue, we use a zero-energy concentration window that restricts this phenomena. A Hamming window was used in our study [15], [16].

$$\mathrm{H(n)} = 0.54 - 0.46 \times \cos(2\pi \frac{\mathrm{n}}{\mathrm{N-1}}) \tag{10}$$

D. Formants and Pitch Examination

1) Decoding problem from acoustic to phonetic: Due to the continuous nature of speech signals, it is difficult to identify different linguistic units such as words, syllables and phonemes in the recorded signal. This problem is known as phonetic acoustic decoding. [17], [28]. We used a process allowing us to identify the transitions between consonants and vowels. The syllables corresponding to this case are available to qualify [18].

2) Vowel-consonant (VC) and Consonant-vowel (CV) Transitions Detection: Digital filtering is employed first to remove as much background noise as feasible [19].

The flowchart in Fig. 4 illustrates the identification of vowel-consonant and consonant-vowel transitions.

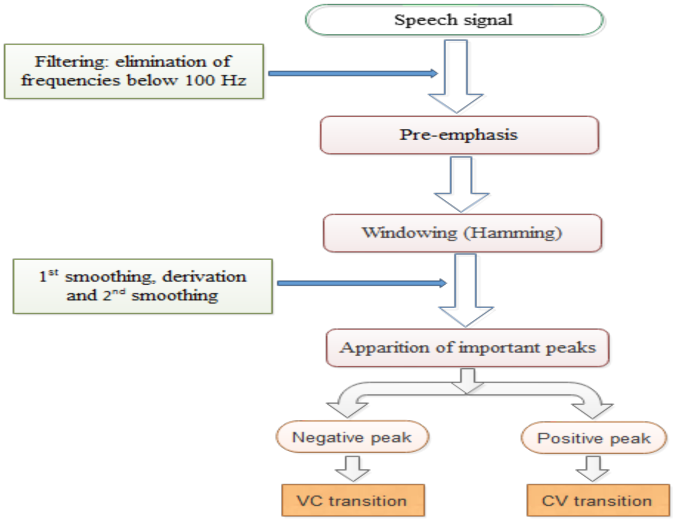
Speech signal
Filtering: elimination of frequencies below 100 Hz
Pre-emphasis
Windowing (Hamming)
1st smoothing, derivation and 2nd smoothing
Apparition of important peaks
Negative peak
Positive peak
VC transition
CV transition

Fig. 4. CV and VC flowchart for detecting transitions [20].

3) Formants: Minimums and maximums presents in the vocal signal spectrum correlate to resonance and anti-resonance tract vocal, also named formants and anti-formants.

In general, the frequencies of analysed formants are: F1 is 200-900 Hz, F2 is 500-2500 Hz, F3 is 1500-3500 Hz, and F4 is 2500-4600 Hz.

The formants (F1, F2, F3, and F4) used in this study, for vowels and consonants found in word that recovered in order to differentiate the phoneme formants [21].

Formant analysis is used to identify consonants and vowels. Before formants can be recognized, they must be processed to make them clearer [22].

The chart in Fig. 5 represents this preprocessing.

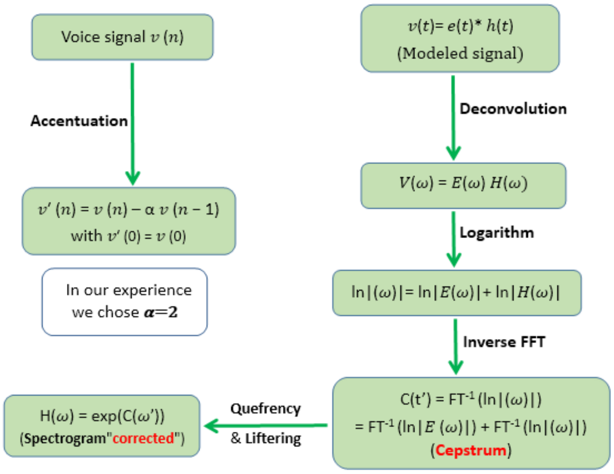
Voice signal $v(n)$
Accentuation
$v'(n) = v(n) - \alpha\, v(n-1)$
with $v'(0) = v(0)$
In our experience
we chose $\alpha=2$
$v(t)= e(t)^{*} h(t)$
(Modeled signal)
Deconvolution
$V(\omega) = E(\omega)\, H(\omega)$
Logarithm
$\ln|(\omega)| = \ln|E(\omega)| + \ln|H(\omega)|$
Inverse FFT
$C(t') = FT^{-1}(\ln|(\omega)|)$
$= FT^{-1}(\ln|E(\omega)|) + FT^{-1}(\ln|(\omega)|)$
(Cepstrum)
Quefrency
& Liftering
$H(\omega) = \exp(C(\omega'))$
(Spectrogram"corrected")

Fig. 5. Diagram for parameter preprocessing.

Once spectrum corrected was acquired, we search for every formant's in a carefully selected frequency band to increase the probability of finding the formant with the highest average amplitude.

Tables II and III show the predicted formant values (in Hertz) for a male voice. The extreme values of each formant have been specified.

TABLE II. A MAN'S VOICE'S ESTIMATED VOWEL FORMANT VALUE

Voyelle	Latin correspondence	F1	F2	F3	F4
[ⵉ]	[i]	250	2250	2980	3280
[ⵓ]	[u]	420	2050	2630	3340
[ⴰ]	[a]	760	1450	2590	3280

TABLE III. A MAN'S VOICE'S ESTIMATED CONSONANT FORMANT VALUE

Consonne	Latin correspondence	F1	F2	F3	F4
[ⵎ]	[m]	300	1300	2300	2770
[ⵏ]	[n]	350	1050	2300	3470
[ⵍ]	[l]	360	1700	2500	3300
[ⵔ]	[r]	550	1300	2300	2700

After the formant values have been obtained, we calculate the distance with every formant values in memory. Formant distance is strictly Euclidean; between both phonemes A and B, it is calculated as follows:

$$\mathrm{dist}(A,B) = \sqrt{\sum_{i=1}^{4}\left(F_A^i - F_B^i\right)^2} \qquad (10)$$

4) Pitch: Pitch is s a crucial component of human voice and widely recognized as perceptual fundamental of sound that is strongly attached to frequency and can be related to the vocal cord's vibration fundamental frequency, permitting audio frequency recognition. It is among the most essential auditory features of sounds, as well as quality and loudness [23], [30].

We used the "Get Pitch" command to extract the pitch, with set the pitch floor to 75 Hz, and set the pitch ceiling to 500 Hz.

E. Measurement Tools and Corpus

1) Tools: Phoneticians and academics utilize the open source program PRAAT [24] to identify various phonetic properties of speech. It is a very efficient software for analysing and recreating acoustic speech signal [25].

For collecting all of the characteristics presented in this study, wav files were registered and analyzed by PRAAT (see Fig. 6).

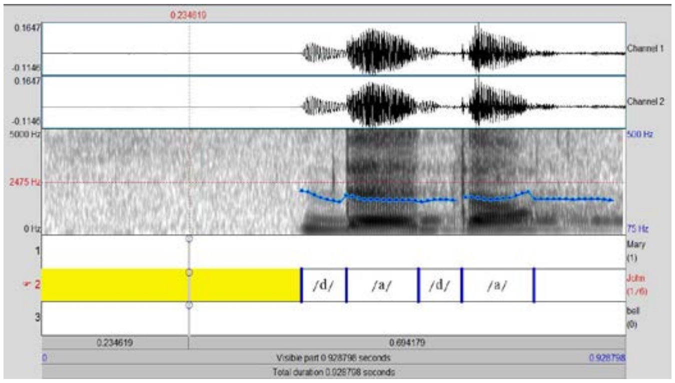

Fig. 6. Manual segmentation of the /dada/ syllable for the CVCV context.

2) Measurements: The foundations for measuring voice signals acoustically are pitch and the four formants, which are widely utilized as indicators of perceived speech quality [26]. The Table IV shows the Amazigh vowels used in this work.

TABLE IV. THE VOWELS AMAZIGH USED IN THIS STUDY

Tifinagh	English transcription	Latin correspondence	Arabic transcription	IPA
ⴰ	YA	A	يا	Æ
ⵉ	YI	I	يي	I
ⵓ	YO	U	يو	ʊ

3) Preparation of the corpus: Ten persons (five women and five men) are chosen among a vocal database of Amazigh people in Morocco comes from various regions with no distinctive geographical distribution. Age was used to coordinate subjects. The average age of the women was 35, ranging from 23 to 50 years. The men in this group range in age from 25 to 50 years, with an average of 36 years. Each speaker repeated the process 10 times. The total amount of evaluated words (10 speakers x 8 words x 10 repetitions), giving us 800 files to examine.

Our objective is to analyse the consonants, semi consonants, and vowels that are pronounced by studying the important voice parameters [27]. We manually recovered the vowels A, I, and U from Krad, Tanmirt, and Ayur words spectrograms. Based on the spectrogram of words Aghrum, Attas, and Tazalit, R, T, and Z are obtained consonants. More information about the database is shown in Table V.

TABLE V. RECORDING PARAMETERS USED IN THE CORPUS AMAZIGH PHONEMES PREPARATION

Parameter	Value
Sampling rate	22.05 kHz
Quantization	16 bits
Duration	2 second / syllabe
Wave format	Mono, wav
Corpus	10 Amazigh words
Speaker	10 (5 females + 5 males)
Accent	Moroccan Tamazight

4) Materials: In this study, we use a microphones and a computer having 8 GB of RAM and an Intel Core i7 processor running at 2.5 GHz. Our experience indicates that Windows 10 LTSB is the prevalent operating system. In a silent room, the microphone were placed between 4 and 10 centimeters from the individual's lips. We recorded the wav file with the parameters shown in Table V.

IV. RESULTS AND DISCUSSION

Fig. 7 represents our approach to determining the acoustical power of syllable [ara]. The Fig. 8 give the temporal derivatives of acoustical power shown in Fig. 7.

A. Authors and Affiliations

After a series of studies [29], it has been experimentally estimated that the transition occurs near the peak where the signal has lost or gained 66 percent in extreme difference of intensity while comparing both phonemes. Crosses appear on the chart to indicate the transitions.

The algorithm for phoneme separation is very effective as follows:

- The initial and final moments of silence were deleted;
- The speakers didn't blow into the microphone during recording, causing audio signal saturation and resulting in extremely big peak which the program interprets like a transition;

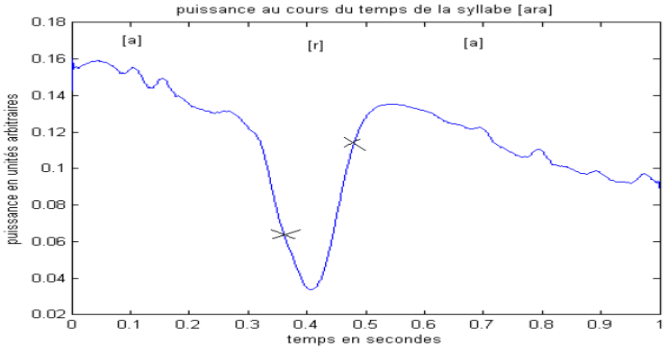

Fig. 7. Acoustical power of the word [ara].

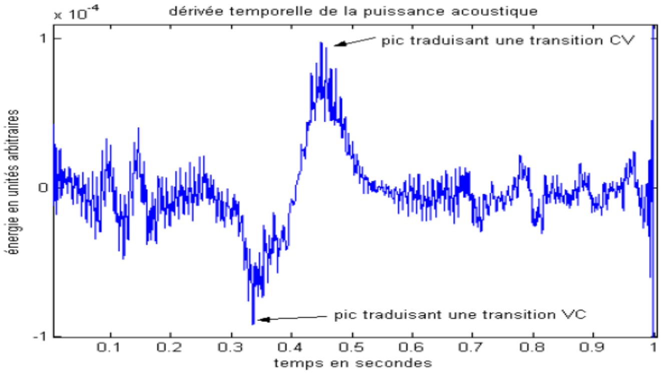

Fig. 8. Derivative of word's acoustical energy [ara].

This method achieves a 75% success rate, which is dependent on parameters such as geographic place, recorded material, as well as verbal difficulties of speakers.

V. CONCLUSION

This study examined the difficulty of automatic speech recognition for the Amazigh language. In specifically, we attempted to extract the vowels and consonants from the Amazigh voice signal. It is now possible to develop an Amazigh corpus to complement those that already exist. This approach opens the door to the socioeconomic growth of the Amazigh community in Morocco.

This methodology to voice recognition enabled us to detect and to exploit by solutions implemented, a number of phonetic and spectral characteristics of the Amazigh voice signal. Better identification of the issue parameters (accentuation coefficients, quefrence cut cepstrum, etc.) in addition to a more precise analysis of formant transitions and trajectories and a main aspects of prosody in the speech are mostly feasible strategies to produce improved outcomes. In addition, speakers of a language like Amazigh should be proficient in both the language and the use of Information Technology resources.

REFERENCES

[1] B. H. Juang et L. R. Rabiner, « Automatic Speech Recognition – A Brief History of the Technology Development », Ga. Inst. Technol. Atlanta Rutgers Univ. Univ. Calif. St. Barbara, p. 24, 2004.

[2] D. Gerhard, « Pitch Extraction and Fundamental Frequency: History and Current Techniques », Tech. Rep. Regina Dep. Comput. Sci. Univ. Regina, p. 23, 2003.

[3] T. W. Parsons, « Separation of speech from interfering speech by means of harmonic selection », J. Acoust. Soc. Am., vol. 60, no 4, p. 911-918, oct. 1976, doi: 10.1121/1.381172.

[4] K. Samudravijaya, « Modeling Natural Language for Automatic Speech Recognition », Tata Institue Fundam. Res. Homi Bhabha Road Mumbai India, p. 8.

[5] J. Li, L. Deng, R. Haeb-Umbach, et Y. Gong, « Fundamentals of speech recognition », in Robust Automatic Speech Recognition, Elsevier, 2016, p. 9-40. doi: 10.1016/B978-0-12-802398-3.00002-7.

[6] T. Koizumi, M. Mori, S. Taniguchi, et M. Maruya, « Recurrent neural networks for phoneme recognition », in Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Philadelphia, PA, USA, 1996, vol. 1, p. 326-329. doi: 10.1109/ICSLP.1996.607119.

[7] N. Dave, « Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition », Int. J. Adv. Res. Eng. Technol., vol. 1, no 6, p. 5, 2013.

[8] A. Boukous, « Phonologie amazighe », Inst. R. Cult. Amaz. Rabat Maroc, p. 445, 2009.

[9] F. Ataa Allah et S. Boulaknadel, « Amazigh Search Engine: Tifinaghe Character-Based Approach », Int'l Conf Inf. Knowl. Eng., p. 5, 2010.

[10] M. Makhlouf, D. Legros, et B. Marin, « Influence de la langue maternelle kabyle et arabe sur l'apprentissage de l'orthographe française », Univ. Mouloud Mammeri Tizi Ouzou IUFM Créteil, p. 7, 2006.

[11] H. Satori et F. El Haoussi, « Investigation Amazigh speech recognition using CMU tools », Int. J. Speech Technol., vol. 17, no 3, p. 235-243, sept. 2014, doi: 10.1007/s10772-014-9223-y.

[12] R. Dufour, « Transcription Automatique da la Parole Spontanée », Inform. Cs Univ. Maine Fr. Tel-00595465, p. 190, 2010.

[13] S. K. Saksamudre, P. P. Shrishrimal, et R. R. Deshmukh, « A Review on Different Approaches for Speech Recognition System », Int. J. Comput. Appl., vol. 115, no 22, p. 23-28, avr. 2015, doi: 10.5120/20284-2839.

[14] H. Hosni, Z. Sakka, A. Kachouri, et M. Samet, « Étude de la Paramétrisation RASTA PLP en vue de la Reconnaissance Automatique de la Parole Arabe », 5th IEEE Int. Conf. Sci. Electron. Technol. Inf. Telecommun. Tunis., p. 7, 2009.

[15] M. Agrawal et T. Raikwar, « Speech Recognition Using Signal Processing Techniques », Int. J. Eng. Innov. Technol., vol. 5, no 8, p. 4, 2016, doi: 10.17605/osf.io/zab7g.

[16] T. W. Parsons, Voice and speech processing / Thomas W. Parsons. New York: McGraw-Hill, 1987.

[17] A. Abenaou, F. Ataa Allah, et B. Nsiri, « Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables », Asinag, no 9, p. 133-145, 2014.

[18] B. Lecouteux, « Reconnaissance Automatique de la Parole guidée par des transcriptions a priori », Inform. Lang. CsCL Univ. D'Avignon Pays Vaucluse Fr. Tel-01381704, p. 170, 2008.

[19] S. El Ouahabi, M. Atounti, et M. Bellouki, « Toward an automatic speech recognition system for amazigh-tarifit language », Int. J. Speech Technol., vol. 22, no 2, p. 421-432, juin 2019, doi: 10.1007/s10772-019-09617-6.

[20] A. Ouhnini, B. Aksasse, et M. Ouanan, « Phonemes Recognition Using Formant Analysis in the Case of Consonant Vowel Transition Case "Amazigh Language" », in : Int'l Conf. Advanced Intelligent Systems for Sustainable Development (AI2SD'2020), vol. 1417, Springer International Publishing, 2022, p. 348-358. doi: 10.1007/978-3-030-90633-7_30.

[21] C. Gendrot, « Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande », MIDL Paris, p. 6, 2004.

[22] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, et A. Stolcke, « The Microsoft 2017 Conversational Speech Recognition System », in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), avr. 2018, p. 5934-5938. doi: 10.1109/ICASSP.2018.8461870.

[23] E. V. Bonzi, G. B. Grad, A. M. Maggi, et M. R. Muñóz, « Study of the characteristic parameters of the normal voices of Argentinian speakers », Pap. Phys., vol. 6, p. 060002, juill. 2014, doi: 10.4279/PIP.060002.

[24] M. Labied et A. Belangour, « Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison », Int. J. Adv. Comput. Sci. Appl., vol. 12, no 8, 2021, doi: 10.14569/IJACSA.2021.0120821.

[25] J. Kreiman et B. R. Gerratt, « Perception of aperiodicity in pathological voice », J. Acoust. Soc. Am., vol. 117, no 4, p. 2201-2211, avr. 2005, doi: 10.1121/1.1858351.

[26] R. Rehman, K. Bordoloi, K. Dutta, N. Borah, et P. Mahanta, « Feature Selection and Classification of Speech Dataset for Gender Identification: A Machine Learning Approach », Journal of Theoretical and Applied Information Technology, Vol. 98, no 22, p. 11, nov. 2020.

[27] L. Besacier, V.-B. Le, E. Castelli, S. Sethserey, et L. Protin, « Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer », TALN Dourdan, p. 12, 2005.

[28] Z. Yin, « Training & Evaluation System of Intelligent Oral Phonics Based on Speech Recognition Technology », Int. J. Emerg. Technol. Learn. IJET, vol. 13, no 04, p. 45, mars 2018, doi: 10.3991/ijet.v13i04.8469.

[29] F. D. Wang, X. Wang, et S. Lv, « An Overview of End-to-End Automatic Speech Recognition », Symmetry, vol. 11, no 8, p. 1018, août 2019, doi: 10.3390/sym11081018.

[30] F. Jiao, J. Song, X. Zhao, P. Zhao, et R. Wang, « A Spoken English Teaching System Based on Speech Recognition and Machine Learning », Int. J. Emerg. Technol. Learn. IJET, vol. 16, no 14, p. 68, juill. 2021, doi: 10.3991/ijet.v16i14.24049.