

# Enhancing Amazigh Speech Recognition System with MFDWC-SVM

Fadwa Abakarim(✉) and Abdenbi Abenaou

Research Team of Applied Mathematics and Intelligent Systems Engineering, National School of Applied Sciences, Ibn Zohr University, 80000 Agadir, Morocco
fadwa.abakarim@gmail.com

**Abstract.** In this research, a new Automatic Speech Recognition (ASR) system is introduced for the Amazigh language, more precisely Tachelhit, spoken in the High Atlas and the South-West of Morocco. The implementation of this system is a challenging task because the availability of Amazigh speech databases is limited, so the first step is to create a database and then to develop the system. Our database contains 33 letters and 10 digits (0 to 9). 120 speakers participated in the recording of the speech corpus: 80 adults (40 males and 40 females) and 40 children (20 boys and 20 girls). Among them, 60 are native speakers and the other 60 are speaking the language for the first time. After creating the database, we used Short-Time Energy (STE) to remove silence and compress the length of the input speech signals. Next, the features are extracted by a combination of two well-known approaches: the Mel-Frequency Cepstral Coefficients (MFCC) and the Discrete Wavelet Transform (DWT). The result is the Mel-Frequency Discrete Wavelet Coefficients (MFDWC) method. Finally, the classification step is conducted using Support Vector Machine (SVM) model. The experimental results prove that the system performs well by achieving an accuracy of 94.42% for the recognition of 10 digits, 90.82% for the recognition of 33 letters and 92.22% for the recognition of 33 letters and 10 digits.

**Keywords:** Automatic Speech Recognition · Amazigh Language · Short-Time Energy · Support Vector Machine · Mel-Frequency Cepstral Coefficients · Discrete Wavelet Transform · Mel-Frequency Discrete Wavelet Coefficients

## 1 Introduction

Speech is a type of communication used by humans to interact with each other. In the same context, there is also human-machine interaction, where a machine detects and recognizes the spoken words. This is called Automatic Speech Recognition (ASR).

Several languages have been used for the achievement of a speech recognition system such as Hindi [1], Arabic [2], French [3] and English [4]. However, it is rare to find an Amazigh-Tachelhit speech recognition system, and our goal is to develop one. This sort of system can be used in various applications, including learning new languages, controlling devices and translation.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13956, pp. 471–488, 2023.
https://doi.org/10.1007/978-3-031-36805-9_31

The Amazigh language is a vital component of Moroccan culture, with around half of the population speaking it as their first language. Although the Royal Institute of Amazigh Culture (IRCAM) has facilitated its integration into various areas, the language still suffers in Information and Communication Technologies (ICT) because of the lack of Natural Language Processing (NLP) compared to other languages [5]. Therefore, several researchers are focused on the development of Amazigh speech recognition systems.

Most ASR systems are developed using free software such as HTK [6], CMU Sphinx [7], ISIP [8] and AVCSR [9]. Of these, HTK and CMU Sphinx are commonly used by researchers to build Amazigh speech recognition systems.

El Ouahabi et al. [10] presented an Amazigh ASR system able to identify 43 isolated words, including 33 letters and 10 digits. The MFCC was used for feature extraction and the Hidden Markov Model Toolkit (HTK) for system development. The database contained 19500 audio files: 16500 for letters and 3000 for digits. The number of speakers was 80, including 40 males and 40 females. The system achieved a high recognition rate of 91.31% using Hidden Markov Models (HMMs) with 3 states and 32 Gaussian Mixture Models (GMMs).

Telmem and Ghanou [11] realized an Amazigh ASR system using CMU Sphinx-4. The database contained 2970 audio files of 33 letters recorded by 9 speakers. The system achieved a high accuracy of 88.00% using 3 HMMs and 2 GMMs.

Satori and Elhaoussi [12] developed an Amazigh ASR system using CMU Sphinx tools. The features were extracted using an MFCC approach. The database contained 25800 audio files of 33 letters and 10 digits. The number of speakers was 60, including 30 males and 30 females. The system achieved an improved performance of 89.07% using 5 HMMs and 16 GMMs.

El Ouahabi et al. [13] created an Amazigh ASR system using CMU Sphinx-4. The features were extracted by the MFCC approach. The database used in this study contained 46750 audio files of 187 isolated words. The number of speakers was 50, including 25 males and 25 females. The results show that the combination of 6 and 8 GMMs with 2000, 3000 and 4000 tied states (senones) gives the best recognition rate of 91.80%.

Telmem and Ghanou [14] proposed an Amazigh ASR system using CMU Sphinx-4. The database contained 11220 audio files of 33 letters recorded by 34 speakers. A high recognition rate of 90.00% was obtained using 3 HMMs and 128 GMMs.

The Moroccan Amazigh language has three regional variants: Tachelhit in the High Atlas and the South-West, Tamazight in the Central and the South-East and Tarifit in North regions. Most of the Amazigh ASR research is conducted on Tarifit [12, 13, 15]. Therefore, in this research, we focus on the Amazigh language spoken in the city of Agadir and its surroundings, which is Tachelhit. Moreover, the MFCC approach and the open-source toolkits based on HMMs: CMU Sphinx 4 and HTK, are commonly used for the development of Amazigh ASR systems.

In this study, a new Amazigh ASR system is created, based on a combination of two well-known feature extraction methods, MFCC and DWT. This combination produces the MFDWC method. The classification process is realized using SVM. A self-created database is used to train and test the system performance. It contains 43 isolated words (33 letters and 10 digits) recorded by 120 speakers.

The structure of the article is as follows: A concise summary of the Amazigh language is given in Sect. 2. Section 3 introduces the self-created database. Section 4 describes the applied methods used to create the proposed system. The experimental results are presented in Sect. 5. Finally, Sect. 6 concludes the paper.

## 2 Amazigh Language

The Amazigh language, also known as Tamazight or Berber, is a subcategory of the Afro-Asiatic language family and consists of two parts: the Berber languages of the North and the South. They are spoken in several countries, from Morocco, which has a population of around 50% Amazigh speakers, to Egypt, through Algeria with 25%, Libya, Mauritania, Mali, Tunisia, and Niger [5].

In Morocco, the Amazigh language is classified into three main regional dialects: Tachelhit in the High Atlas and the South-West, Tarifit in the North and Tamazight in Central and South-East Morocco [16].

In 2001, the Royal Institute of Amazigh Culture (IRCAM) was founded to standardize the Amazigh language and integrate it into various fields. As such, it has become a national language and since 2003 it has been introduced in administration and the media. It is taught in the educational system as an essential subject in elementary school, and as a course in the faculty of letters and humanities [17].

The Tifinaghe-IRCAM alphabet system, developed by IRCAM, is written from left to right and is built on the basis of a graphic system that shows phonological tendencies. However, it does not preserve all the phonetic realizations produced, but only the functional ones. This system contains [5, 18, 19]:

27 consonants including: the alveolars ( ⵔ, ⵙ, ⵥ, ⵣ), the dentals ( ⵔ, ⵕ, ⵍ, ⵟ, ⵜ, ⵏ, ⵟ, ⵏ), the labials ( ⴱ, ⴼ, ⵎ), the labiovelars ( ⴽⵯ, ⴳⵯ), the laryngeal ( ⵀ), the palatals ( ⵊ, ⵛ), the pharyngeals ( ⵃ, ⵄ), the uvulars ( ⵅ, ⵇ, ⵖ) and the velars ( ⴽ, ⴳ).

2 semi-consonants: ⵡ and ⵢ.

4 vowels: neutral vowel (or schwa) ⵓ which has a distinctive position in Amazigh phonology, and three full vowels ⵉ, ⴰ, ⵓ.

Tifinaghe does not have a specific set of punctuation marks. Instead, IRCAM suggests using international symbols such as “ ” (space), “.”, “,”, “;”, “ ?”, “ !” and “…” as punctuation marks. For numerals, in Morocco, the standard ones are used (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) which are recommended for Tifinaghe writing [18, 19].

Tables 1 and 2 show the 33 characters of the Amazigh alphabet and the 10 Amazigh digits along with their transcriptions in Arabic and English and their syllables.

**Table 1.** The 33 Amazigh characters with their transcriptions in Arabic and English and their syllables [12, 15].

| Tifinaghe-IRCAM | Arabic Transcription | Arabic correspondence | English Transcription | Syllables |
|---|---|---|---|---|
| ⴰ | يا | ا | YA | CV |
| ⴱ | ياب | ب | YAB | CVC |
| ⴳ | ياك | - | YAG | CVC |
| ⴳⵯ | ياك | - | YAGG | CVC |
| ⴷ | ياد | د | YAD | CVC |
| ⴹ | ياض | ض | YADD | CVCC |
| ⴻ | ياي | ي | YEY | CVC |
| ⴼ | ياف | ف | YAF | CVC |
| ⴽ | ياك | ك | YAK | CVC |
| ⴽⵯ | ياكك | - | YAKK | CVCC |
| ⵀ | ياه | ه | YAH | CVC |
| ⵃ | ياح | ح | YAHH | CVC |
| ⵄ | ياع | ع | YAAA | CVC |
| ⵅ | ياخ | خ | YAKH | CVC |
| ⵇ | ياق | ق | YAQQ | CVC |
| ⵉ | يي | ي | YI | CV |
| ⵊ | ياج | ج | YAJ | CVC |
| ⵍ | يال | ل | YAL | CVC |
| ⵎ | يام | م | YAM | CVC |
| ⵏ | يان | ن | YAN | CVC |
| ⵓ | يو | و | YO | CV |
| ⵔ | يار | ر | YAR | CVC |
| ⵕ | يار | - | YARR | CVC |
| ⵖ | ياغ | غ | YAGH | CVC |
| ⵙ | ياس | س | YAS | CVC |
| ⵚ | ياص | ص | YASS | CVC |
| ⵛ | ياش | ش | YASH | CVC |
| ⵜ | يات | ت | YAT | CVC |
| ⵟ | ياط | ط | YATT | CVC |
| ⵡ | ياو | و | YAW | CVC |
| ⵢ | ياي | ى | YAY | CVC |
| ⵣ | ياز | ز | YAZ | CVC |
| ⵥ | ياز | - | YAZZ | CVCC |

**Table 2.** The 10 Amazigh digits with their transcriptions in Arabic and English and their syllables [12, 15].

| Tifinaghe-IRCAM | Arabic Transcription | Arabic correspondence | English Transcription | Syllables |
|---|---|---|---|---|
| ⵎⵎⵢⴰ | أميا | صفر | AMYA | VC-CV |
| ⵢⴰⵏ | يان | واحد | YEN | CVC |
| ⵙⵉⵏ | سين | اثنان | SIN | CVC |
| ⴽⵔⴰⴹ | كراض | ثلاثة | KRAD | VC-CVC |
| ⴽⵓⵣ | كوز | أربعة | KOZ | CVC |
| ⵙⵎⵎⵓⵙ | سموس | خمسة | SMMUS | CC-CVC |
| ⵙⴹⵉⵙ | سضيص | ستة | SDES | CCVC |
| ⵙⴰ | سا | سبعة | SA | CV |
| ⵜⴰⵎ | تام | ثمانية | TAM | CVC |
| ⵜⵣⴰ | تزا | تسعة | TZA | CC-CV |

## 3 Amazigh Speech Database

For intelligent systems in the field of signal processing, a database is always needed for training and testing. However, a common problem that researchers may encounter during implementation of their work is the lack of input data that can be processed to train and test the system. This work will focus on the realization of a speech recognition system.

Speech databases are available in several different languages, such as Mandarin Chinese, English, Javanese, Nepali, Korean and French, to serve researchers [20]. However, the availability of Amazigh speech databases is limited. Therefore, this gives us the opportunity and motivation to create our own database. Furthermore, it will be available to all researchers and could be used to perform various tasks, including speaker recognition, gender recognition, and keyword recognition.

The database consists of 33 letters (see Table 1) and 10 digits (see Table 2) of the Amazigh language, collected from 120 speakers. Among them, 60 are native Amazigh-Tachelhit and the other 60 are speaking the language for the first time.

The advantage of our database is that the speakers are of various ages. There are 80 adults from 19 to 55 years old, and in addition, there are 40 children from the ages of 7 to 13. Tables 3 and 4 show more details.

The audio files were recorded using Audacity software [21] in a normal environment. The recording has a resolution of 16 bits and a sampling rate of 16 kHz. Each speaker was asked to record each word with 10 repetitions, for a total of 51600 audio recordings in ".wav" format. More details are provided in Table 5. In addition, to simplify the recording process, all speakers were asked to record 10 repetitions of each word in a single voice file (see Fig. 1). Then, it was necessary to manually segment the result files and save each word as an audio file (see Fig. 2). In the end, each word had 10 audio files for each speaker.

**Table 3.** Gender and age of adult speakers.

| Age Category | Gender | | |
|---|---|---|---|
| | Male | Female | Total |
| 19–29 | 20 | 20 | 40 |
| 30–40 | 10 | 10 | 20 |
| 41–55 | 10 | 10 | 20 |
| Total | 40 | 40 | 80 |

**Table 4.** Gender and age of child speakers.

| Age Category | Gender | | |
|---|---|---|---|
| | Male | Female | Total |
| 7–9 | 10 | 10 | 20 |
| 10–13 | 10 | 10 | 20 |
| Total | 20 | 20 | 40 |

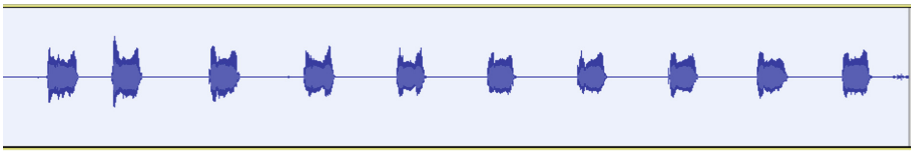

**Fig. 1.** Voice recording of the digit "Amya" with 10 repetitions from a female speaker before cutting.



**Fig. 2.** Single voice recording of the digit "Amya" from a female speaker after cutting.

**Table 5.** Database technical parameters.

| Parameter | Value |
|---|---|
| Corpus | Amazigh 33 letters and 10 digits |
| Accent | Moroccan Amazigh-Tachelhit |
| Number of speakers | 120 |
| Origin of speakers | Morocco, the city of Agadir and its surroundings |
| Age of speakers | 7–55 years old |
| Gender of speakers | 60 males and 60 females |
| Native speakers' number | 60 |
| Non-native speakers' number | 60 |
| Words' number | 43 (33 letters and 10 digits) |
| Number of repetitions per word | 10 |
| Number of tokens per speaker | 430 (43 words × 10 repetitions) |
| Total number of audio files | 51600 (43 words × 10 repetitions × 120 speakers) |
| Each audio file duration | Between 1 and 3 s |
| Number of bits | 16 bits |
| Sampling rate | 16 kHz |
| Audio data file format | .wav |
| Channels | 1 (Mono) |
| The recording software | Audacity |

## 4 Applied Methods

In our study, the achievement of an Amazigh speech recognition system consists of three steps:

First, pre-processing. After collecting the database, it was observed that the input speech signals contained silence at the beginning and end of each spoken word. Therefore, we used Short-Time Energy (STE) to remove the silence and compress the length of the speech signals.

Second, feature extraction, in which we combined two well-known feature extraction methods, MFCC and DWT. The result is MFDWC, which provides good time-frequency localization and is robust in noisy environments.

Finally, the classification, which affects the system performance and the final decision. Therefore, we chose one of the most used models for classification problems and pattern recognition, the SVM model.

### 4.1 Pre-processing

To create a fast and efficient ASR system, the input speech signals should be pre-processed to compress them and preserve their quality by removing irrelevant information. In this study, we used the Short-Time Energy (STE) approach to detect and remove the unvoiced parts of the input signal, producing a fully voiced signal.

The speech signal amplitude varies over time, and typically, the voiced segments have a higher amplitude than unvoiced speech segments. STE can be used to represent these amplitude variations. It is a measure of the energy in a voice signal over a short period of time [22].

The short-time energy can be defined as follows [22]:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{1}$$

The choice of window function can significantly affect the characteristics of the short-time energy representation. For the proposed system, we opted for the Hamming window because it provides higher out-of-band attenuation compared to the rectangular window. This means that it reduces the amplitude of frequencies that are not within the window.

$$w(n) = (1-\alpha) - \alpha\, cos\left[\frac{2\pi n}{N-1}\right] \tag{2}$$

The value of $\alpha$ is between 0 and 1. We used $\alpha = 0.46$, which gives:

$$w(n) = \begin{cases} 0.54 - 0.46\, cos\left[\frac{2\pi n}{N-1}\right], & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \tag{3}$$

where

$n$ represents the current sample, while $N$ represents the total number of samples.

Figure 3 presents the process of STE.

Figure 4 shows the speech signal of the Amazigh digit "Amya" before and after applying STE.

### 4.2 Feature Extraction

Feature extraction is a major phase in signal processing, as it transforms the input data into more valuable information that can be used for classification.

The quality of the extracted features has a direct impact on the classification performance. Indeed, they must be sufficiently discriminating to ensure a good separation between the different classes of signals. Moreover, they should be robust to noise and other disturbances to provide a good generalization of the classification.

In speech recognition, The Mel-Frequency Cepstral Coefficients (MFCC) are the most commonly used features, and their extraction involves an important step known as the Discrete Cosine Transform (DCT). However, the DCT basis vectors have similar time and frequency resolution, and they cover all frequency bands. This means that noise in one frequency band can affect all of the coefficients, which can be problematic in noisy

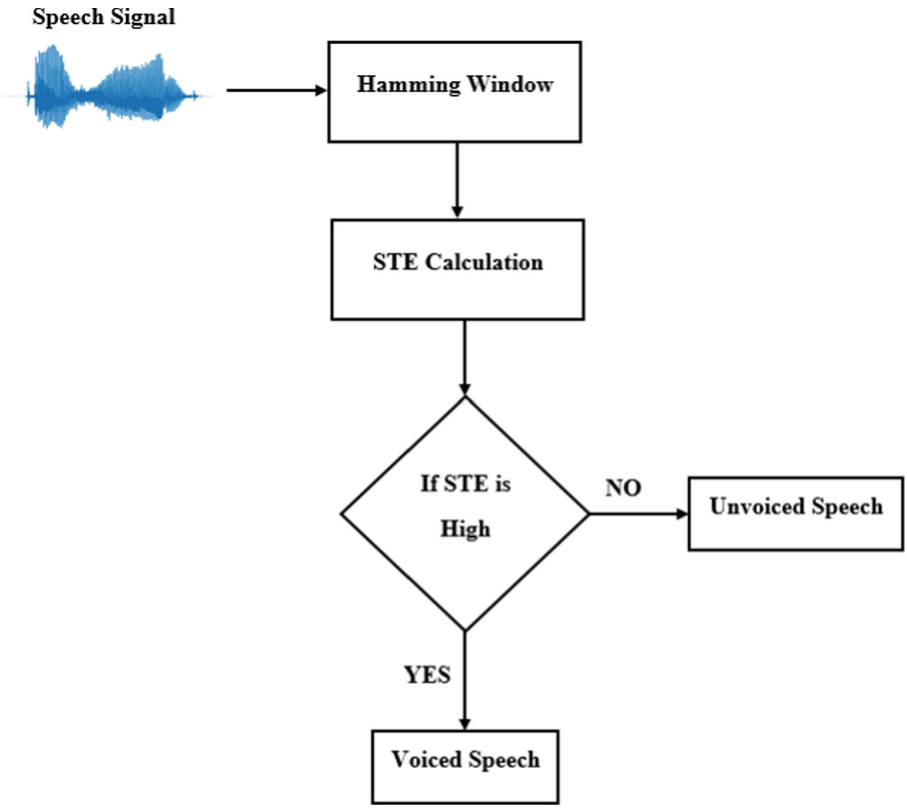


**Fig. 3.** The process of STE.

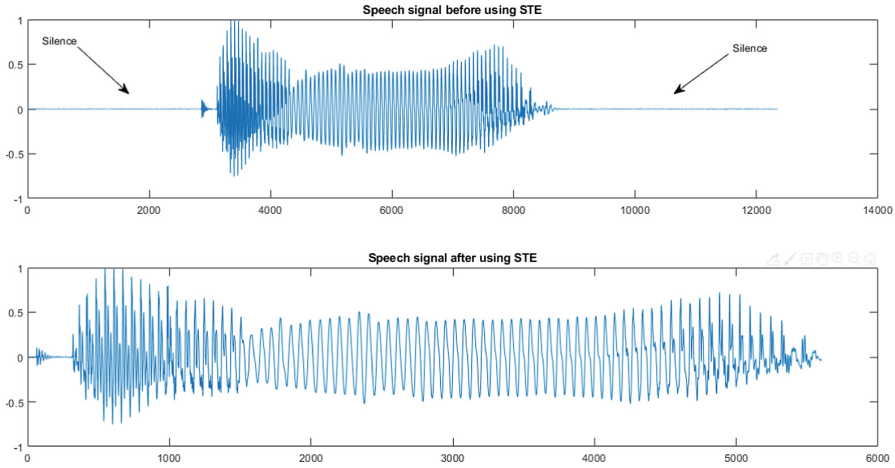


**Fig. 4.** Removing silence from the speech signal of the Amazigh digit "Amya" using STE.

environments [23]. These limitations can be overcome by using the Discrete Wavelet Transform (DWT) method because of its better temporal and frequency localization capability [23, 24].

The DWT method is applied to the energies of the Mel scaled log filter bank. The result of this operation gives the Mel-Frequency Discrete Wavelet Coefficients (MFDWC). The process of extracting MFDWC is the same as for MFCC, except that the DWT approach is applied instead of DCT (see Fig. 5).

To extract the MFDWC features, we first improve the quality of the input speech signal by reducing the low frequencies. Therefore, a high-pass pre-emphasis filter is used. Equation (4) shows the formula for the high-pass filter [25, 26]:

$$y(s) = 1 - 0.97s^{-1} \tag{4}$$

where

$s$ is the input voice signal.

Then, the signal is segmented into several frames, neither long nor short in size. In this study, we segmented the signal into 25 ms frames with a 10 ms overlap. To reduce the signal discontinuities before and after each frame, the Hamming window is implemented in each frame. Equations (2) and (3) mentioned in the pre-processing part present the Hamming window calculation.

The next step is to convert each frame from the time domain to the frequency domain using the Fast Fourier Transform (FFT) algorithm.

$$s_i(k) = \sum_{n=1}^{N} s_i(n)w(n)e^{-2j\pi kn/N}, 1 \leq k \leq N/2 \tag{5}$$

where

$s_i(n)$ is the speech signal after framing and $i$ is the frame index.

After applying FFT to each frame of the voice signal, the power spectrum of each frame is found using the following equation:

$$p_i(k) = \frac{1}{N}|s_i(k)|^2 \tag{6}$$

Psychophysical studies show that human ears do not detect frequencies on a linear scale [27]. Thus, the filter bank is processed based on the Mel scale.

The formula to convert the linear scale to the Mel scale is the following:

$$Mel(f) = 1125 * \ln\left(1 + \frac{f}{700}\right) \tag{7}$$

where

$Mel(f)$ represents the perceived Mel frequency and $f$ represents the given frequency in Hz.

A filter bank of 25 triangular filters is used, obtained by the formula below:

$$b_m(k) = \begin{cases} \frac{k - f(mel-1)}{f(mel) - f(mel-1)}, & f(mel-1) \leq k \leq f(mel) \\ \frac{f(mel+1) - k}{f(mel+1) - f(mel)}, & f(mel) \leq k \leq f(mel+1) \\ 0, & otherwise \end{cases} \tag{8}$$

where

$f(mel)$ is the frequency in the Mel scale.

To calculate the energy of Mel filter bank, we multiply each filter bank defined in Eq. (8) by the power spectrum defined in Eq. (6), as shown in Eq. (9):

$$\tilde{e}_m = \sum_{k=0}^{K/2} p_i(k) b_m(k) \tag{9}$$

Finally, we apply the DWT approach to the logarithm of the Mel filter bank energies. This gives as a result the Mel-Frequency Discrete Wavelet Coefficients (MFDWC).

$$\tilde{c}_n = DWT(\log(\tilde{e}_m)) \tag{10}$$

The wavelet transform creates a set of basic functions called "child wavelets" using a prototype function known as "mother wavelets". The choice of the mother wavelet is crucial in achieving good temporal and frequency resolution. After trying several families of wavelets to get a good performance, we chose the "Daubechies4", which produces a result of 16 MFDWC features.

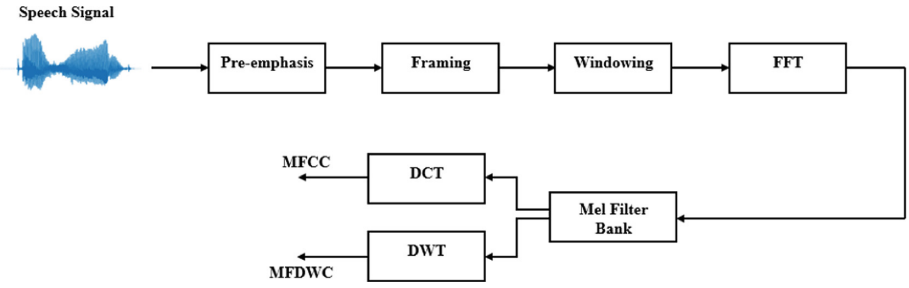


**Fig. 5.** The extraction process of MFCC and MFDWC features.

### 4.3 Classification

After feature extraction comes the classification step. The choice of a classification model is essential in achieving good performance. In this study, we used a powerful and commonly used classifier in the speech recognition field, the SVM model [24, 28–30].

SVM performs complex transformations of the data according to the selected kernel function. This transformation aims at maximizing the separation boundaries between data points based on defined labels or classes.

In our study, we do multiclass classification and SVM does not support this. SVM supports binary classification and separation of data points into two classes. Therefore, we decompose the multiclass classification problem into several binary classification problems, and this is what the one-versus-one and the one-versus-all approaches use. In the one-versus-one approach, the decomposition is set to a binary classifier for each pair of classes ($N$ classes require $N(N-1)/2$ models). In the one-versus-all approach, the decomposition is set to a binary classifier for each class ($N$ classes require $N$ models).

Although the one-versus-one approach uses a high number of binary SVMs ($N(N-1)/2$), it requires less training data and is less sensitive to unbalanced datasets. On the other hand, the one-versus-all method generates a relatively lower number of binary SVMs (N), but it needs a large dataset to train each binary SVM. Therefore, in this research, we chose the one-versus-one approach and used LIBSVM because it is more practical. In addition, as a kernel function, we used the Radial Basis Function (RBF).

Figure 6 presents the three steps involved in implementing the proposed system.

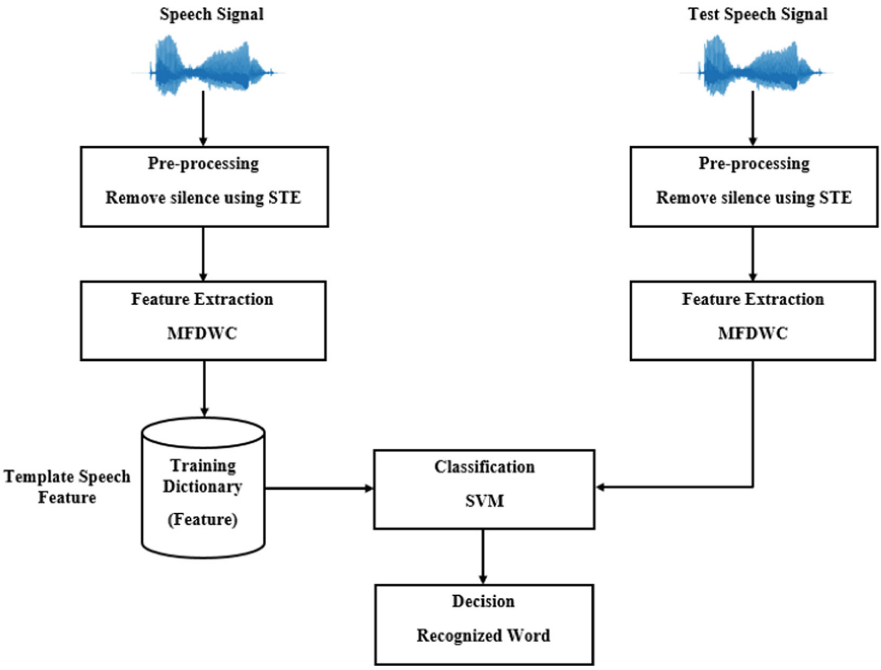


**Fig. 6.** The proposed system steps.

## 5 Experimental Results and Discussion

One of the most rarely used languages in technology is the Amazigh language. It is rare to find recognition systems based on the Amazigh language including speaker recognition, speech recognition and voice pathology recognition. The goal is to create an Amazigh-Tachelhit ASR system, using MFDWC for feature extraction and SVM for classification. Our database consists of 51600 audio files of 33 letters and 10 digits, 70% is used for training and 30% for testing (see Table 6). Moreover, to evaluate the system performance, different speakers are used in the testing process than the ones used in the training process.

**Table 6.** The training and test datasets used in this study.

| Data | Training data | Test data | Total |
|---|---|---|---|
| Letters | 27720 | 11880 | 39600 |
| Digits | 8400 | 3600 | 12000 |
| Total | 36120 | 15480 | 51600 |

To test the effectiveness of the system, three experiments were conducted. The first involved training and testing the system using only Amazigh digits, while the second used exclusively Amazigh letters and the third a combination of letters and digits (see Table 7).

**Table 7.** The various experiments conducted to test the proposed system effectiveness.

| Experiments | Training dataset | Test dataset |
|---|---|---|
| Experiment 1: 10 digits (0 to 9) | 8400 | 3600 |
| Experiment 2: 33 letters | 27720 | 11880 |
| Experiment 3: 10 digits + 33 letters | 36120 | 15480 |

The experiments were performed using MATLAB 2019 on a 64-bit operating system equipped with 8 GB of RAM and an Intel Core i7 processor.

Table 8 shows the recognition rates obtained from the first experiment. The test dataset of this experiment consisted of 3600 audio files. A total recognition rate of 94.42% was obtained. We can see that the digits that are well recognized by the system are: YEN ( ⵢ ⴰ ⵏ) and KRAD ( ⴽ ⵕ ⴷ), with accuracies of 97.50% and 98.06% respectively. On the other hand, the digits that are poorly recognized by the system are: KOZ ( ⴽ ⴽ ⵓ ⵣ) and SA ( ⵙ ⴰ), with recognition rates of 92.22% and 90.56% respectively.

Table 9 shows the recognition rates from the second experiment. The test dataset of this experiment consisted of 11880 audio files. A total recognition rate of 90.82% was obtained. We can see that the letters that are well recognized by the system are: YAD ( ⴷ) and YAT ( ⵟ), with recognition rates of 96.39% and 96.94% respectively. On the other hand, the letters that are poorly recognized by the system are: YA ( ⵄ) and YI ( ⵉ), with recognition rates of 84.72% and 85.00% respectively.

**Table 8.** The recognition rates of 10 Amazigh digits using the MFDWC method and the SVM classifier.

| Digits | Recognition rate (%) |
|---|---|
| AMYA | 97.22 |
| YEN | 97.50 |
| SIN | 92.50 |
| KRAD | 98.06 |
| KOZ | 92.22 |
| SMMUS | 92.78 |
| SDES | 94.17 |
| SA | 90.56 |
| TAM | 96.67 |
| TZA | 92.50 |
| Total recognition rate (%) | **94.42** |

**Table 9.** The recognition rates of 33 Amazigh letters using the MFDWC method and the SVM classifier.

| Letters | Recognition rate (%) |
|---|---|
| YA | 84.72 |
| YAB | 86.94 |
| YAG | 93.06 |
| YAGG | 91.39 |
| YAD | 96.39 |
| YADD | 95.83 |
| YEY | 90.00 |
| YAF | 91.94 |
| YAK | 93.33 |
| YAKK | 93.06 |
| YAH | 92.50 |
| YAHH | 92.22 |
| YAAA | 86.39 |
| YAKH | 88.61 |
| YAQQ | 93.06 |
| YI | 85.00 |
| YAJ | 94.17 |
| YAL | 93.61 |

(*continued*)

**Table 9.** (*continued*)

| Letters | Recognition rate (%) |
|---|---|
| YAM | 93.06 |
| YAN | 92.22 |
| YO | 86.11 |
| YAR | 91.39 |
| YARR | 89.17 |
| YAGH | 91.67 |
| YAS | 87.22 |
| YASS | 87.78 |
| YASH | 87.50 |
| YAT | 96.94 |
| YATT | 95.83 |
| YAW | 93.61 |
| YAY | 91.11 |
| YAZ | 85.83 |
| YAZZ | 85.28 |
| Total recognition rate (%) | **90.82** |

Tables 8 and 9 show that the total recognition rate of the 33 letters is the least accurate compared to the total recognition rate of the 10 digits. In addition, we can observe that words containing the letters "Z" or "S" are the most misrecognized.

Table 10 shows the overall recognition rate from the third experiment, which is 92.22%. The test dataset of this experiment consisted of 15480 audio files including 33 letters and 10 digits.

**Table 10.** Overall recognition rate of the global data (Amazigh 10 digits and 33 letters) using the MFDWC method and the SVM classifier.

| Global data | Recognition rate (%) |
|---|---|
| 10 digits + 33 letters | 92.22 |

Although we used different speakers in the testing process than those used in the training process, the system performed well, achieving more than 92.00%. Many conditions affect system performance and cause recognition results to differ from system to system, depending on the number of speech files used in the training and test datasets. These conditions include: the quality of the audio files, the environment in which they

are recorded, whether it is a noisy or normal environment, the number, age, gender and accent of the speakers, as well as the methods used to train and test the system.

Table 11 shows the results obtained by other systems and our system.

**Table 11.** The recognition rates obtained by other systems and our system.

| State of the art | Proposed System | Amazigh Corpus | Number of speakers | Number of training voice recordings | Number of test voice recordings | Recognition rate (%) |
|---|---|---|---|---|---|---|
| [11] | System based on CMU Sphinx tools | 33 letters | 9 | 2640 | 330 | 88.00 |
| [12] | System based on CMU Sphinx tools | 33 letters + 10 digits | 60 | 18060 | 7740 | 89.07 |
| [14] | System based on CMU Sphinx tools | 33 letters | 34 | 9900 | 1320 | 90.00 |
| Current study | System based on MFDWC and SVM | 33 letters + 10 digits | 120 | 36120 | 15480 | 92.22 |

## 6 Conclusion

In this work, our goal was to focus on a new language, one less documented and less explored by researchers, which is the Amazigh language, Tachelhit. Therefore, a new speech recognition system has been proposed for the Amazigh-Tachelhit language. The database used in this system is a self-created database. It is collected from 120 speakers: 60 are native speakers and the other 60 are not. 30% of the database is used for testing and 70% for training. The system is based on a combination of two widely used feature extraction methods, MFCC and DWT, to form a new method called MFDWC. The SVM model is used for classification, and the one-versus-one approach is used to handle the multiclass classification problem. To test the effectiveness of the system, we used different speakers in the testing process to those used in the training process. The system performed well, with an overall rate of 92.22%. According to the experimental results,

the total recognition rate of the 33 Amazigh letters is less accurate than that of the 10 Amazigh digits. In addition, words containing the letters "Z" or "S" are the least recognized. The proposed system can be used in various applications, including language learning, translation and dictation.

## References


1. Patil, U.G., Shirbahadurkar, S.D., Paithane, A.N.: Automatic speech recognition of isolated words in Hindi language using MFCC. In: Proceedings of the 2016 International Conference on Computing, Analytics and Security Trends (CAST), pp. 433–438. IEEE (2016). https://doi.org/10.1109/CAST.2016.7915008
2. Satori, H., Harti, M., Chenfour, N.: Introduction to Arabic speech recognition using CMUSphinx system. arXiv Preprint (2007). arXiv:0704.2083
3. Youcef, B.C., Elemine, Y.M., Islam, B., Farid, B.: Speech recognition system based on OLLO French Corpus by using MFCCs. In: Chadli, M., Bououden, S., Zelinka, I. (eds.) Recent Advances in Electrical Engineering and Control Applications. LNEE, vol. 411, pp. 326–331. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48929-2_25
4. Naithani, K., Thakkar, V.M., Semwal, A.: English language speech recognition using MFCC and HMM. In: Proceedings of the 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), pp. 1–7. IEEE (2018). https://doi.org/10.1109/RICE.2018.8509046
5. Nejme, F.Z., Boulaknadel, S., Aboutajdine, D.: Analyse automatique de la morphologie nominale Amazighe. In: Proceedings of the 2013 Actes de la conférence du Traitement Automatique du Langage Naturel (TALN), pp. 5–18 (2013)
6. Young, S.J., Young, S.: The HTK Hidden Markov Model Toolkit : Design and Philosophy. Cambridge University Engineering Department (1993)
7. Lamere, P., et al.: The CMU Sphinx-4 speech recognition system. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2–5. IEEE (2003)
8. Ordowski, M., Deshmukh, N., Ganapathiraju, A., Hamaker, J., Picone, J.: A public domain speech-to-text system. In: Proceedings of the 1999 European Conference on Speech Communication and Technology. (1999)
9. Liu, X., Zhao, Y., Pi, X., Liang, L., Nefian, A.V.: Audio-visual continuous speech recognition using a coupled hidden Markov model. In: Proceedings of the 2002 International Conference on Spoken Language Processing (ICSLP) (2002)
10. El Ouahabi, S., Atounti, M., Bellouki, M.: Optimal parameters selected for automatic recognition of spoken Amazigh digits and letters using Hidden Markov Model Toolkit. Int. J. Speech Technol. **23**(4), 861–871 (2020). https://doi.org/10.1007/s10772-020-09762-3
11. Telmem, M., Ghanou, Y.: Amazigh speech recognition system based on CMUSphinx. In: Ben Ahmed, M., Boudhir, A.A. (eds.) SCAMS 2017. LNNS, vol. 37, pp. 397–410. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74500-8_37
12. Satori, H., ElHaoussi, F.: Investigation Amazigh speech recognition using CMU tools. Int. J. Speech Technol. **17**(3), 235–243 (2014). https://doi.org/10.1007/s10772-014-9223-y
13. El Ouahabi, S., Atounti, M., Bellouki, M.: Toward an automatic speech recognition system for amazigh-tarifit language. Int. J. Speech Technol. **22**(2), 421–432 (2019). https://doi.org/10.1007/s10772-019-09617-6
14. Telmem, M., Ghanou, Y.: Estimation of the optimal HMM parameters for amazigh speech recognition system using CMU-Sphinx. Procedia Comput. Sci. **127**, 92–101 (2018). https://doi.org/10.1016/j.procs.2018.01.102

15. El Ouahabi, S., Atounti, M., Bellouki, M.: Amazigh isolated-word speech recognition system using hidden Markov model toolkit (HTK). In: Proceedings of the 2016 International Conference on Information Technology for Organizations Development (IT4OD), pp. 1–7. IEEE (2016). https://doi.org/10.1109/IT4OD.2016.7479305
16. Ouakrim, O.: Fonética y fonología del Bereber. Survey: University of Autònoma de Barcelona (1995)
17. Boulaknadel, S., Talha, M.: Analyse syntactico-sémantique de la langue amazighe (2013)
18. Ataa Allah, F., Boulaknadel, S.: Natural language processing for Amazigh language: challenges and future directions. In: Proceedings of the 2012 workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012) (2012)
19. Ataa Allah, F., Boulaknadel, S.: Convertisseur pour la langue amazighe: script arabe-latin–tifinaghe. In: Proceedings of the 2011 Symposium International sur le Traitement Automatique de la Culture Amazighe, pp. 3–10 (2011)
20. Open Speech and Language Resources. https://www.openslr.org/resources.php. Accessed 20 Jan 2023
21. Audacity software. https://www.audacityteam.org/. Accessed 17 Dec 2022
22. Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B.D.: Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In: Proceedings of the 2008 American Society for Engineering Education (ASEE), pp. 1–7 (2008)
23. Tufekci, Z., Gowdy, J.N.: Feature extraction using discrete wavelet transform for speech recognition. In: Proceedings of the 2000 IEEE SoutheastCon. 'Preparing for The New Millennium' (Cat. No. 00CH37105), pp. 116–123. IEEE (2000). https://doi.org/10.1109/secon.2000.845444
24. Cutajar, M., Gatt, E., Grech, I., Casha, O., Micallef, J.: Comparative study of automatic speech recognition techniques. In: Proceedings of the 2013 IET Signal Processing, vol. 7, no. 1, pp. 25–46 (2013). https://doi.org/10.1049/iet-spr.2012.0151
25. Hammami, N., Lawal, I.A., Bedda, M., Farah, N.: Recognition of Arabic speech sound error in children. Int. J. Speech Technol. **23**(3), 705–711 (2020). https://doi.org/10.1007/s10772-020-09746-3
26. Pandit, P., Makwana, P., Bhatt, S.: Automatic speech recognition of continuous speech signal of Gujarati language using machine learning. In: Sahni, M., Merigó, J.M., Jha, B.K., Verma, R. (eds.) Mathematical Modeling, Computational Intelligence Techniques and Renewable Energy. AISC, vol. 1287, pp. 147–159. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9953-8_13
27. Senthil Raja, G., Dandapat, S.: Speaker recognition under stressed condition. Int. J. Speech Technol. **13**, 141–161 (2010). https://doi.org/10.1007/s10772-010-9075-z
28. Padmanabhan, J., Johnson Premkumar, M.J.: Machine learning in automatic speech recognition: a survey. IETE Tech. Rev. **32**(4), 240–251 (2015). https://doi.org/10.1080/02564602.2015.1010611
29. Ghai, W., Singh, N.: Literature review on automatic speech recognition. Int. J. Comput. Appl. **41**(8), 42–50 (2012). https://doi.org/10.5120/5565-7646
30. Ali, H., Jianwei, A., Iqbal, K.: Automatic speech recognition of Urdu digits with optimal classification approach. Int. J. Comput. Appl. **118**(9), 1–5 (2015). https://doi.org/10.5120/20770-3275