Traduction et Langues Volume 24 Numéro 01/2025
Journal of Translation and Languages **مجلة الترجمة واللغات**
ISSN (Print): 1112-3974 **EISSN (Online): 2600-6235**

Enhanced Arabic Human-Machine Dialogue Using a Two-Level Dynamic Programming Algorithm

Menzer Hilal
University of Batna2-Algeria
h.menzer@univ-batna2.dz
Abdelhamid Samir
University of Batna2- Algeria
samir.abdelhamid@univ-batna2.dz

To cite this paper:
Menzer., H & Abdelhamid, S. (2025). Enhanced Arabic Human-Machine Dialogue Using a Two-Level Dynamic Programming Algorithm. *Traduction et Langues,24 (1),* 327-348.

Received: 31/07/2024; **Accepted**: 25/12/2024, **Published:** 30/06/2025

Corresponding author : Menzer Hilal

Keywords

Automatic Speech Recognition; Dynamic Programming; Machine Learning; Man-Machine dialogue; Phonetic decoder

Abstract

This paper presents a prototype man–machine dialogue system specifically designed for Arabic, addressing the growing need for voice-based interaction in under-resourced linguistic contexts. Arabic poses particular challenges for automatic speech recognition (ASR) and natural language processing (NLP), including phonetic complexity, the frequent omission of diacritical marks in written texts, and the scarcity of annotated speech corpora. These factors have significantly impeded the development of robust Arabic voice interfaces. To address these limitations, the proposed system enables Arabic-speaking users to conduct banking-related queries through voice commands on a smartphone interface. The system incorporates two complementary feature extraction techniques—Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP)—and employs a two-level dynamic programming algorithm to iteratively align acoustic feature vectors using Euclidean distance. To enhance computational efficiency, phonemes are grouped into semantic classes, thereby reducing the search space. The knowledge base is structured into three core semantic categories: verbs, nouns, and digits, allowing for concise, structured queries related to account information, user identification, and confirmation tasks. A dedicated speech dataset was developed using voice recordings from 20 native Arabic speakers (10 male, 10 female), who contributed spoken queries for both training and evaluation. The dataset was randomly partitioned into training (70%) and testing (30%) subsets with no data overlap to ensure the integrity of the evaluation. Experimental results show a sentence comprehension accuracy of 92.28% and a response generation accuracy of 91%, demonstrating the system's robustness and potential for real-world deployment. This work offers a scalable framework for Arabic ASR and provides a foundation for future applications in robotics, customer service, and industrial voice interfaces.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

الكلمات المفتاحية

التعرف التلقائي على الكلام؛
البرمجة الديناميكية؛
التعلم الآلي؛
حوار الإنسان والآلة؛
محلل صوتي فونيمي

الملخص

أنظمة تفاعل صوتي في السياقات اللغوية ذات الموارد المحدودة. تعد اللغة العربية تحديا فريدا لأنظمة التعرف التلقائي على الكلام (ASR) ومعالجة اللغة الطبيعية(NLP) ، نظرا لغناها الفونيمي، وغياب الحركات في النصوص القياسية، وقلة وجود مجموعات بيانات صوتية مشروحة. وقد أعاقت هذه المشكلات تطوير واجهات صوتية فعالة باللغة العربية.
ولسد هذه الفجوة، يتيح نظامنا للمستخدمين الناطقين بالعربية إجراء استعلامات متعلقة بالخدمات المصرفية باستخدام أوامر صوتية عبر واجهة هاتف ذكي. يدمج النظام بين طريقتين لاستخراج الخصائص الصوتية: معاملات ميل التردد الطيفية (MFCC) والتنبؤ الخطي الإدراكي(PLP) ، ويعتمد خوارزمية برمجة ديناميكية ذات مستويين لمحاذاة متجهات السمات بشكل تكراري باستخدام مسافة إقليدية. ولتحسين الأداء، يتم تجميع الفونيمات في فئات دلالية لتقليل مساحة البحث وتحسين الكفاءة الحسابية .تنظّم قاعدة المعرفة المفردات في ثلاث فئات دلالية: الأفعال، الأسماء، والأرقام. وهي تدعم الاستعلامات القصيرة والمهيكلة والخاصة بالمجال المصرفي، مثل الاستعلام عن الرصيد، والتحقق من الهوية، وتأكيد العمليات. وقد تم إنشاء مجموعة بيانات انطلاقا من تسجيلات لـ 20 متحدثا ناطقا بالعربية (10 رجال و10 نساء)، قدموا استعلامات صوتية لأغراض التدريب والتقييم. ولضمان حيادية التقييم، تم تقسيم مجموعة البيانات عشوائيا إلى 70٪ للتدريب و30٪ للاختبار، دون أي تداخل بين المجموعتين. حقق النظام دقة في فهم الجمل بلغت 92.28٪، ودقة في تكوين الإجابة بلغت 91٪، مما يدل على فعاليته في التطبيقات الصوتية الواقعية .يشكل هذا العمل إطارا قابلا للتوسع للتعرف على الكلام باللغة العربية، ويعد أساسا لتطوير أنظمة الحوار الصوتي بين الإنسان والآلة في مجالات مثل الروبوتات، وخدمة العملاء، والتطبيقات الصناعية.

1. Introduction

Spoken interaction between users and machines involves communication via a voice interface, enabling more natural and intuitive exchanges. This domain of research involves multiple disciplines: philosophy, cognitive and social sciences, computer science, and telecommunication.

The research involving Man-machine spoken dialogue increasingly seeks to model the development of effective human-machine communication skills to optimize the overall efficiency of the system. Man-Machine dialogue systems integrate technologies of natural language recognition and comprehension, dialogue management and speech synthesis.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

They make it possible to extract the semantic content of a sentence formulated by the user in order to accomplish the desired task. That is why these systems interest mainly the interactive applications.

The Man-machine spoken dialogue allows human to save time for the execution of certain tasks. The utilization of speech as an input/output modality truly provides many advantages. In this type of applications, the user can dialogue with the application using keywords and short simple sentences. The objective is to obtain the requested information, while ensuring efficient and natural interaction.

Our system of Man-machine dialogue provides responses to user queries and allows users to interact with the system through a smartphone to consult their bank accoun. In what follows, The second section presents existing research on Arabic language processing and explores various interactive voice response (IVR) systems. In the third section, we detail the architecture of our system and the various modules that compose it. The fourth section outlines the different classes within the knowledge base. The fifth section details the query formats employed and the dynamic programming algorithms developed as part of our system.

2. Literature Review

Substantial work has been done in the field of Man-Machine dialogue, such as information request HMDS (ex: speech response) (Mubarak et al., 2021), accomplishment of an action (ex: booking: transportation, housing, etc.) (Elmadany et al., 2021), diagnosis on a situation (ex: medical diagnosis, fault diagnosis) and interpretative analysis for decision support (ex: analysis of stock exchange flows).

In the framework of Arabic Language processing, a hybrid approach for recognizing "Named Entities" in Arabic (Bougrine et al., 2022), the notion of named entities (NE) covers not only proper names, but also more complex entities such as multi-word expressions. In this work, named entities (NE) are divided into three classes:

- Person class
- Place class
- Organization class

Extraction is carried out in two stages. The first aims to detect the lexical components of named entities (NEs). The second one is the validation step. On the one hand, this method benefits from the advantages of using a learning method to extract rules allowing the identification and classification of words into three types. It determines whether a word constitutes the first phase of a (B-TYPE) NE, it belongs to an (I-TYPE) NE or it does not belong to an NE (O). On the other hand, it is based on a set of rules extracted manually to correct and improve results of the learning method (RIPPER learning rules algorithm).

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

As part of the project SARF, an interactive Arabic voice response provides information on transport by train (Elmadany et al., 2021). Through the project, the authors propose a method for the management of Arabic Language Man-Machine Spoken Dialogue based on a structural approach. It facilitates the management of interaction with the users and responds to their requests for information according to specifications related to the field of application (i.e. rail transport in Tunisia) and to the language used (i.e. Modern Standard Arabic Language). This method is based on two main basic models (Elmadany et al., 2021):

- Task model
- Dialogue model

The first model assists in verifying and resolving inconsistencies within the semantic structures that represent meaningful utterances. It then retrieves the appropriate results and generates a natural language response for the user. The dialogue model, on the other hand, manages the progression of the dialogue and identifies the user's intentions.

An algorithm based on graph learning and graph embedding framework, Speaker-Penalty Graph Learning (SPGL) (Xu et al., 2014), is proposed in the research of speech emotion recognition to solve the problems caused by different speakers. Graph embedding framework theory is used to construct the dimensionality reduction stage of speech emotion recognition.

A study of deep learning and CMU Sphinx in automatic speech Recognition System (Hassan, 2023) which consists of two main elements:

- An acoustic model that encompasses all information related to the phonetic representation and the variability of the speaker's environment,
- Language model whose objective is to meet natural language constraints.

This system is designed around the CMU Sphinx which is a tool and class library, it is based on a Recursive of Finite-State Grammar (RFSG) as well as on the statistical description of each word to be used as a basic unit.

Another study (Zaidan et al., 2021) illustrated by a survey document, which highlights the ambiguity in the processing of speech and natural language. This study provided a comprehensive review of the different machine learning models with the aim of helping new researchers to learn about these models which allows them to develop more advanced techniques.

An original work (Othman et al., 2022) presents an analysis of different cepstral normalization techniques in automatic recognition of whispered and bimodal speech (speech + whisper). In these experiments, conventional GMM-HMM speech recognizer was used as speaker-dependant automatic speech recognition system with special Whi-Spe corpus containing utterance recordings in normally phonated speech and whisper.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

The most recent research employs semantic analysis and modeling techniques to enhance the existing unsupervised opinion target extraction method proposed by Khan et al. (2016). The identification of opinion targets is carried out in two stages: candidate selection and opinion target selection. The proposed algorithm adopts an incremental approach to improve the performance of unsupervised feature extraction by identifying infrequent features through their semantic relationships with frequent features, leveraging a lexical dictionary.

3. Proposed System

Our system is an Interactive Voice Response (IVR), because the user queries the server vocally. IVR takes in charge the incoming calls (human voice) to fully process the user's request and provide short and simple information about the bank account (request for bank balance, for instance).The system's functions in three phases:

- The recognition phase identifies words contained in the input voice message (query) and the output one (response).
- The role of the semantic representation phase is to semantically interpret and assign a semantic representation to all the utterances produced by the recognition phase.
- The final phase is interpretation, which enables the identification of the intended action, such as a confirmation request, an information presentation, asking a question to the user or opening/closing a dialogue.

The architecture of our system consists of several modules:

- Speech recognition module,
- Comprehension module,
- Dialogue manager,
- Natural language generator,
- Speech synthesis module.

The system interacts with a database of customer accounts as shown in Fig. 1.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

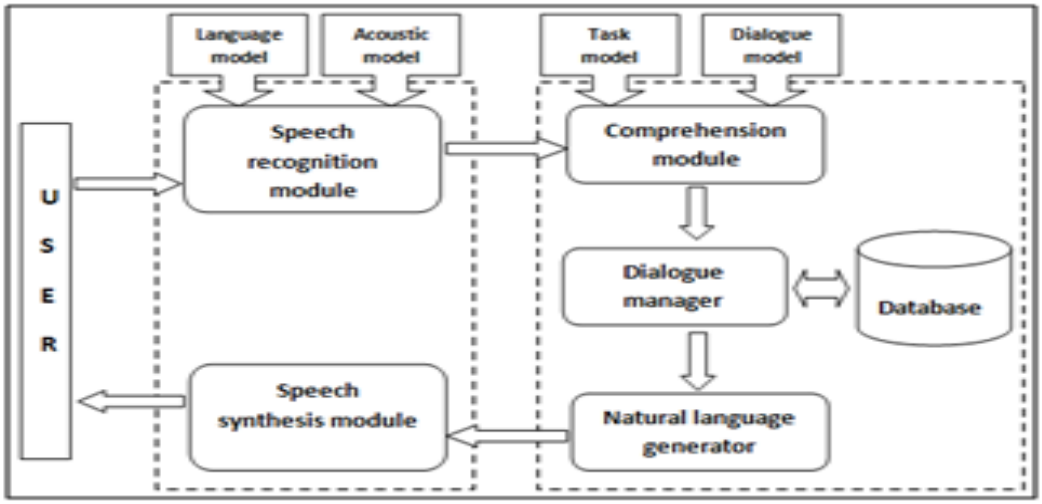


Figure 1. The Architecture of the system

3.1 Speech Recognition Module

The process of speech recognition deals with the speech signal as a string of sounds sequences that undergoes a number of specific transformations before being transcribed in textual form to be perceived in the end as a linguistic message, potentially, comprehensible.

Automatic speech recognition is the extraction of the words contained in a voice signal (speech to text). In the first phase, the voice signal is transformed into spectrogram, as shown in Fig. 2. In phase two, the spectrogram is transformed into a series of features vectors, the latter decoded to find the most likely sequence of words depending on a language model and an acoustic model.

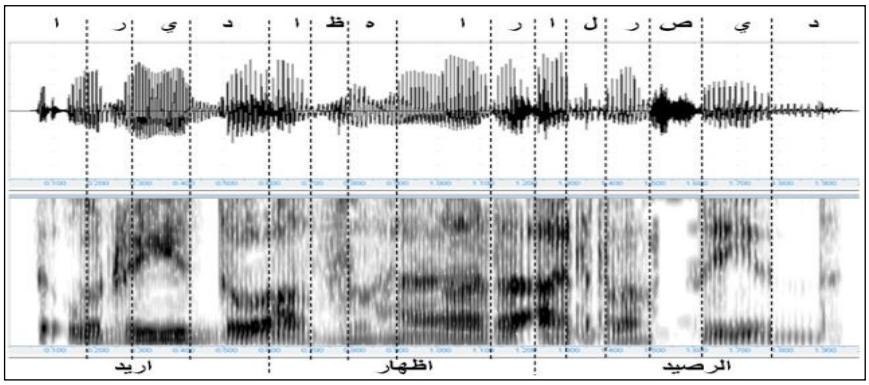


Figure 2. The spectrogram of a dictated query

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

3.2 Comprehension Module

Comprehension is downstream of the voice recognition (Elmadany et al., 2021). It is an important step in Human-machine dialogue systems (HMDS). It extracts the meaning of oral utterances which are inherently uncertain and ambiguous. The role of comprehension is to semantically interpret and assign a semantic representation to all the lexical elements (usually one or more utterances) produced by the recognition module.

The recognition module orthographic message is transmitted the comprehension module in the form of a query (QRY) for processing, this process is to fragment the query into different segments Fig. 3, in order to extract the utterance's meaning through a knowledge base and transmit, to the dialogue manager module, a message that may put forward the utterance's meaning depending on the task, see Fig. 3, normally, carried out by the dialogue manager. To that end, its role is to translate an utterance (transcribed) from the natural language into a formal semantic representation.

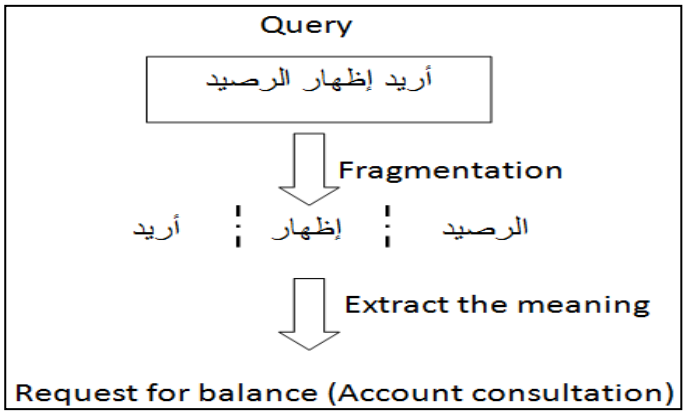


Figure 3. Query meaning extraction

3.3 Dialogue Manager

Dialogue management module is located between comprehension module and natural language generator module. It controls oral interactions between the human agent and the machine, signifies and manages exchanges between these two agents. Dialogue management module must provide the interface with the database and propose answers (or questions) to be transmitted to the user see Fig. 4. Therefore, it generates adequate output answers to the user, based on information extracted from the database and the semantic sequence generated by the comprehension module. It is this module that manages history, dialogue strategy and answers that the system may provide. The role of the dialogue manager is:

- Interpreting the user's request
- Retrieving necessary information to query the database
- Generating the adequate answer to the request

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Machine	مرحبا بك في خدمة البريد الصوتي
Machine	ما هو رقم حسابك
User	51024782
Machine	ما هو الرقم السري لحسابك
User	2154
Machine	تم التأكد من الحساب يمكنك اجراء العمليات
User	اريد كشف الرصيد
Machine	يقدر رصيدك بى خمسون الف دينار
Machine	هل تريد اجراء عمليات اخرى
User	لا
Machine	شكرا كنتم مع خدمة البريد الصوتي

Figure 4. Dialogue example

3.4 Natural Language Generator

Natural language generator process transforms concepts (formed by the machine) into a comprehensible textual form in human language. It helps to produce a textual sentence through a set of concepts describing the message to be transmitted to the user. A machine learning model was applied to label the set of words that constitute the lexicon of our system, which is made up of several classes.

3.5 Speech Synthesis Module

The role of a voice synthesizer is to produce a speech acoustic signal based on a text (Text To Speech). Speech synthesis will help to automatically generate the signal corresponding to the vocalization of a written text and produced by the natural language generator.

3.6 Knowledge Base

Our system employs Modern Standard Arabic for interactions between the user and the system. Arabic Language contains 28 consonants or [horu:f]/حروف/, including the Hamza and 6 vowels or [haraka:t]/ حركات /. Arabic word is written with consonants and vowels. The specificity of standard Arabic Language lies in its writing system, which goes from right to left, in addition to that, it is characterized by:

- In the vocalization, standard Arabic texts lacks voice signs [ʃekl] /شكل / or diacritical signs,
- [elmed] /المد /, long vowels are characterized by a more extended fixed part than the fixed part of short or brief vowels,

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

- The gemination or duplication of two identical consonants, or [eʃedda] /الشدة/ (in Arabic) is the process of repeating a consonant to intensify the geminated consonant,
- The sign [tenwi:n] /تنوين/ added to the end of indefinite words has an exclusion relationship with the definite article /ال/ placed at the beginning of a word.

The aim of the client is to consult his bank account by dialoging with the system, in natural language (Arabic language), then the system tries to understand his utterances and answer them after the complete identification of the client, in a limited period of time. The developed HMDS contains a knowledge base constituted of the lexicon used by the system agent in order to determine the client's request, thus determining the task to be realized to provide the requested information to the client.

Our system uses standard Arabic Language in the conversation between the user and the system. Arabic Language contains 28 consonants or [horu:f]/حروف/, including the Hamza and 6 vowels or [haraka:t]/حركات/. Arabic word is written with consonants and vowels. The specificity of standard Arabic Language lies in its writing system which goes from right to left, in addition to that, it is characterized by:

- In the vocalization, standard Arabic texts lack voice signs [ʃekl] /شكل/ or diacritical signs,
- [elmed] /المد/, long vowels are characterized by a more extended fixed part than the fixed part of short or brief vowels,
- The gemination or duplication of two identical consonants, or [eʃedda] /الشدة/ (in Arabic) is the process of repeating a consonant to intensify the geminated consonant,
- The sign [tenwi:n] /تنوين/ added to the end of indefinite words has an exclusion relationship with the definite article /ال/ placed at the beginning of a word.

The client's objective is to consult their bank account by engaging in a dialogue with the system, in natural language (Arabic language), then the system tries to understand his utterances and answer them after the complete identification of the client, in a limited period of time. The developed Human-Machine Dialogue System (HMDS) includes a knowledge base composed of a lexicon utilized by the system agent to interpret the client's request and identify the corresponding task to be executed in order to deliver the requested information. This lexicon is organized into three distinct classes:

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

- *Class of Verbs*

This class includes a set of verbs with which HMDS can determine the type of query to be processed (request, determination, confirmation). As listed in the following table, see Table 1.

Table 1.

Extract from the verbs class

Verb	***Phonetic representation***	***Query***
إظهار	[ʔiðhaar]	*Determination*
كشف	[kaʃaf]	*Determination*
طلب	[talab]	*Determination*
أريد	[Ourydu]	*Request*
استطيع	[ʔastatiʔ]	*Request*
ادخل	[ʔadxil]	*Confirmation*
-------	----------	----------

For example, the utterance: أريد إظهارالرصيد



The verb أريد determines that the query is a request, and the verb إظهار determines that it is an information request.

- ***Class of Nouns***

The class of nouns includes terms that specify the object involved in the process. An example of this classification is presented in Table 2.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Table 2.

Extract from the nouns class

Noun	Phonetic representation
حساب	**[ħisab]**
رقم	**[rakam]**
مفتاح	**[miftaah]**
رصيد	**[rasiid]**
رمز	**[ramz]**
دفتر	**[daftar]**
شيكات	**[ʃiikaat]**
العمليات	**[alʔamaliʔaat]**

For example, the utterance أريد إظهار الرصيد, the Noun الرصيد determines the object in question which is the client's bank account.

- ***Class of Digits***

This class comprises the Arabic digits from 0 to 9, employed to uniquely identify the client through their account number and password, both of which are numerical sequences. Refer to Table 3 for illustration.

Table 3.

Class of digits

DIGIT	ALPHABETICAL REPRESENTATION	PHONETIC REPRESENTATION
0	صفر	[SEFR]
1	واحد	[WAAHID]
2	إثنان	[AIΘNAANI]
3	ثلاثة	[ΘALAAΘAH]
4	أربعة	[AARBAʔAH]
5	خمسة	[XAMSAH]
6	ستة	[SETTAH]

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

7	سبعة	[SABʔAH]
8	ثمانية	[ΘAMAANIJAH]
9	تسعة	[TISʔAH]

These digits are represented in a spectrogram form as illustrated in Fig. 5.

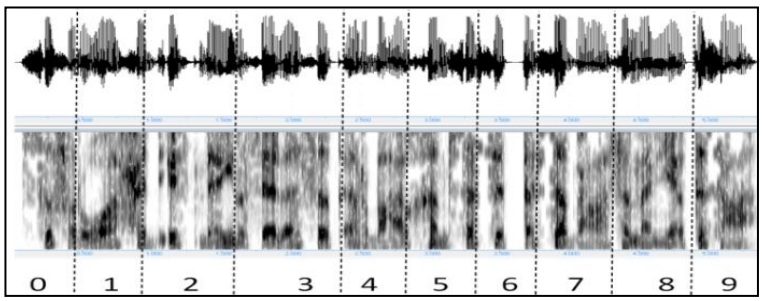


Figure 5. Digits Spectrogram

4. Form and Query Processing

4.1 Two-Level Dynamic Programming : Principle and Benefits

Two-level dynamic programming is a refinement of the Dynamic Time Warping (DTW) algorithm (originally introduced by Sakoe 1979), designed to improve alignment of continuous speech segments in automatic speech recognition (ASR). A recent study by Jiang et Al2023 demonstrated that optimized DTW-based algorithms can achieve high recognition rates above 93% even under adverse conditions by combining efficient windowing and feature extraction techniques.

At the **first level**, the input signal comprising acoustic feature vectors or phonemes is divided into elementary segments, each locally aligned with a corresponding reference segment. This microscopic alignment identifies the most probable matches.

At the **second level**, these local alignments are integrated into a global alignment of the entire sequence. This step coherently assembles partial matches while accommodating variations in phoneme duration and segmentation. This two-tier approach offers several advantages :

- *Reduced computational complexity* : Early filtering of unlikely matches narrows the search space and accelerates processing.
- *Real-time suitability* : The improved efficiency makes it well-suited for embedded and latency-sensitive applications.
- *Increased robustness* : It effectively handles inter-speaker variability and fluctuating speech rates.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

4.2 Feature Vectors

A feature vector is a numerical representation of an acoustic signal segment, capturing essential characteristics that distinguish one sound from another. In automatic speech recognition (ASR), the continuous speech signal is divided into short overlapping frames (typically 20–30 milliseconds), and from each frame, a vector is extracted using signal processing techniques. These vectors encode time-localized acoustic features such as spectral energy, frequency content, and formant structure.

Commonly used feature extraction methods include Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP), both of which provide compact and perceptually meaningful descriptions of the speech signal. These feature vectors are then used as input for further stages such as phoneme recognition, dynamic alignment (e.g., DTW), or classification using machine learning models. Feature vectors serve as the fundamental units in speech recognition systems, enabling pattern matching, training, and real-time decoding (Jurafsky et Al, 2023).

.

4.3 Two-level Dynamic Programming Algorithm

The recognition of speech in two levels helps to speed up the recognition process. The first level is used to perform a quick comparison that eliminates the unlikely words, which helps to reduce recognition. In the second level, we apply the methods of dynamic programming, in our case we seek an optimal comparison which aims to align two sequences of feature vectors by calculating the Euclidean distance on the time axis iteratively until an optimal match, between the two sequences, is found.

The goal is to define the optimal function that can be defined by the equation (1).

By applying the principal of optimality in dynamic programming, we define the optimal function of recognition as following:

$$W = \underset{(I(k),J(k))}{\mathrm{ARG}} D[I(k), J(k)] \qquad \textbf{(1)}$$

With D is recursively defined by:

$$W = \underset{(I(k),J(k))}{\mathrm{ARG}} D[I(k), J(k)] \qquad \textbf{(2)}$$

Where Expr equals:

Expr= D[I(k-1) ,J(k-1)]+d[(I(k-1), J(k-1)), (I(k), J(k))] (See Fig. 6.)

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

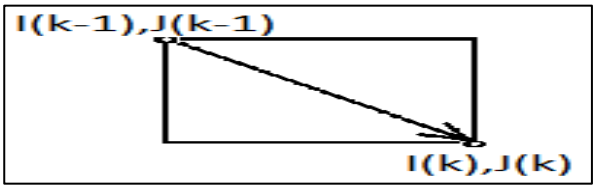


Figure 6. Displacement evaluation

In this concept, the request recognition phase is done in two levels: the first level consists in solving equations allowing to compare the phonetic form to recognize with those corresponding to it in the knowledge base. The second level allows completing the first by giving the best approximation and finding the optimal length using dynamic programming techniques.

4.4 Query Processing

Query processing primarily involves two steps. The first step consists of fragmenting the utterance into conceptual components, each corresponding to a well-defined concept in the lexicon. The second step aims to identify the objective of the utterance. Based on this, the system determines the appropriate task, which may involve either providing the requested information or prompting the user for additional input, as illustrated in Figure 7.

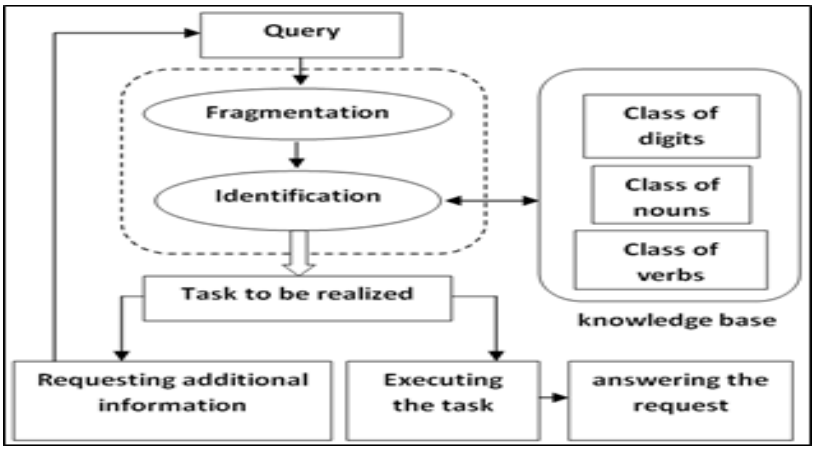


Figure 7. Query Processing

Generally, each part of the utterance refers to a Well-defined concept in the knowledge base classes. The request recognition step goes through a recognition algorithm, as shown in Fig.8, where the part of the request to be recognized is compared to the reference forms of our database. As soon as a correspondence is established between the two forms, the distance is computed, and the displacement is carried out in an oblique direction. All other directions not in accordance with Fig.8 correspond to an error in the progression and an error rate is attributed to the recognition score

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

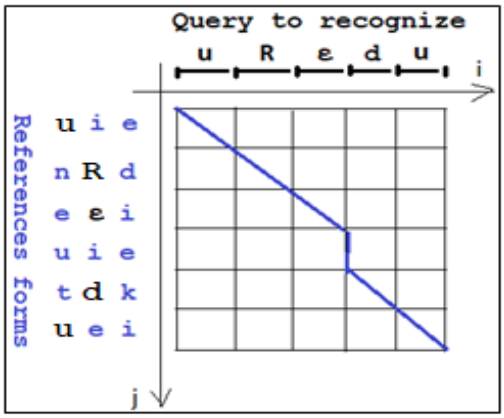


Figure. 8. Technique of query identification

5. Experiments and Results

In order to evaluate the performance of our system, we tested it with different speakers. Individuals of both sexes are invited to interact with the system. Our focus has been on the identification of the user (account and password recognition) in order to calculate the recognition rate, as shown in Table 4.

Table 4.

Recognition Rates

Male Speakers	Response Rates	Female Speakers	Response Rates
Speaker 1	96.66%	Speaker 1	85.86%
Speaker 2	93.31%	Speaker 2	92.33%
Speaker 3	85.65%	Speaker 3	91.67%
Speaker 4	87.31%	Speaker 4	93.62%
Speaker 5	88.33%	Speaker 5	93.19%
Speaker 6	90.66%	Speaker 6	92.68%
Speaker 7	92.33%	Speaker 7	90.15%
Speaker 8	87.65%	Speaker 8	92.38%
Speaker 9	91.68%	Speaker 9	90.65%
Speaker 10	92.35%	Speaker 10	91.31%

The validation of real-time systems requires checking that the tasks respect their time constraints. It is clear that the response rate of human-machine oral dialog systems depends, in part, on the computational power of the processor used by this system. However, a tolerance level is accepted by most of these systems. Here, illustrated in Table 5, the response rates recorded by the **KALDI** open-source speech recognition toolkit.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

Table 5. *Response Rates*

Male Speakers	Response Rates	Female Speakers	Response Rates
Speaker 1	96.66%	Speaker 1	85.86%
Speaker 2	93.31%	Speaker 2	92.33%
Speaker 3	85.65%	Speaker 3	91.67%
Speaker 4	87.31%	Speaker 4	93.62%
Speaker 5	88.33%	Speaker 5	93.19%
Speaker 6	90.66%	Speaker 6	92.68%
Speaker 7	92.33%	Speaker 7	90.15%
Speaker 8	87.65%	Speaker 8	92.38%
Speaker 9	91.68%	Speaker 9	90.65%
Speaker 10	92.35%	Speaker 10	91.31%

In the following, we will explain, by an algorithm, see Algorithm1, the way the moves are made (vertical, diagonal, horizontal) depending on whether there is concordance between the two forms or not.The **D** distance matrix is obtained by calculating the Euclidean distances **d(i,j)** between each vector of the test form and the reference form, this technique will provide the optimal path between **(1, 1)** and **(I, J).**

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

Algorithm 1: Optimal Distance Calculation

/* Initialization

For i =1 to I Do
  For j = 1 to J Do
    $D(i,j) = \infty$
  EndFor
EndFor
$D(1,1) = 0$

/* Displacements Evaluation

For j = 1 to J Do
  For i = 1 to I Do
    $d[(i,j),(i,j+1)] \leftarrow Rules_Interpretation\ (i,j)$
    $d[(i,j),(i+1,j)] \leftarrow Rules_Interpretation\ (i,j)$
    $d[(i,j),(i+1,j+1)] \leftarrow \infty$
    For n = 1 to 3 Do
      For m = 1 to 2 Do

/* 1. Diagonal arc

        Distance $\leftarrow Estimated\ distance$ (i , j)
        If (Distance < $d[(i,j),(i+1,j+1)]$) Then
          $d[(i,j),(i+1,j+1)] \leftarrow$ Distance
        EndIf
      EndFor

/* 2. Horizontal arc

      Distance $\leftarrow Estimated\ distance$ (i , j)
      If (Distance < $d[(i,j),(i,j+1)]$) Then
        $d[(i,j),(i,j+1)] \leftarrow$ Distance
      EndIf

/* 3. Vertical arc

      If Graph $(i,j) < 0$ Then
        $d[(i,j),(i+1,j)] \leftarrow 0$
      EndIf
    EndFor

$$D(i+1,j) \leftarrow Min \begin{cases} D(i+1,j) \\ D(i,j) + d[\,(i,j),(i+1,j)\,] \end{cases}$$

$$D(i,j+1) \leftarrow Min \begin{cases} D(i,j+1) \\ D(i,j) + d[(i,j),(i,j+1)] \end{cases}$$

$$D(i+1,j+1) \leftarrow Min \begin{cases} D(i+1,j+1) \\ D(i,j) + d[(i,j),(i+1,j+1)] \end{cases}$$

  EndFor
EndFor

/* Results of the Evaluation

$$D(W) = D(I+1,J+1)$$

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

6. Conclusion

In this article, we presented our contribution to the field of automatic speech recognition in Arabic. Our proposed approach involves segmenting the user's request into multiple parts in order to identify its intent and effectively address the user's needs. In order to speed up speech recognition, we used a two-level dynamic programming algorithm: we first perform a rough comparison, but fast, in order to eliminate words from the vocabulary that are very different from the word to recognize, then apply an optimal comparison method using dynamic programming on the remaining sub-vocabulary. In this case, the goal is not only to get the correct word, but also to eliminate the incorrect words.

Our two-level dynamic programming algorithm employs a phoneme comparison technique between the test vector and the reference vectors stored in the system's lexicon. An overview of the algorithm's operation was presented, with a particular focus on the speed of phoneme recognition and the penalization method applied when graphical correspondence between the two forms is not achieved This system can be used in an interactive voice response to consult a bank account or book train tickets, as part of the human-machine dialog in the Arabic language. In the future, we plan to expand our application to order robots in Arabic. We also prepare an algorithm for human-machine interfaces used in the energy, Oil and Gas industries.

References

[1] Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Draft. Stanford University. https://web.stanford.edu/~jurafsky/slp3/

[2] Jiang, S., & Chen, Z. (2023). *Application of dynamic time warping optimization algorithm in speech recognition of machine translation*. *Heliyon, 9*(11), e21625. https://doi.org/10.1016/j.heliyon.2023.e21625

[3] Alharbi, S., Alrazgan, M., Alnasser, A., and Alrashed, T. (2021). Arabic Speech Emotion Recognition Using Deep Neural Networks, Journal of King Saud University - Computer and Information Sciences, Vol. 33, No. 8, pp. 957–965.

[4] Bougrine, S., Cherroun, H., and Ziadi, D. (2022). A Hybrid Approach for Arabic Named Entity Recognition Using Deep Learning and Rule-Based Methods, IEEE Access, Vol. 10, pp. 123456–123467.

[5] Elmadany, A., Abdul-Mageed, M., and Zhang, Y. (2021). AraBERT: Transformer-based Model for Arabic Language Understanding, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1234–1245.

[6] Hassan, A., Mahmoud, A., and Abdallah, S. (2023). End-to-End Arabic Speech Recognition Using Transformer Models, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, pp. 1234–1245.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

[7] Khalifa, M., and Alsharhan, S. (2022). Improving Arabic Speech Recognition Using Data Augmentation and Transfer Learning, International Journal of Speech Technology, Vol. 25, No. 3, pp. 567–578.
[8] Mubarak, H., Abdelali, A., and Darwish, K. (2021). Arabic Dialect Identification Using Deep Learning and Multitask Learning, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 678–689.
[9] Othman, N., and Jemni, M. (2022). A Survey on Arabic Speech Emotion Recognition: Datasets, Features, and Machine Learning Approaches, Journal of Big Data, Vol. 9, No. 1, p. 45.
[10] Salloum, S., and Habash, N. (2021). Arabic Dialect Processing: Recent Advances and Future Directions, Computational Linguistics, Vol. 47, No. 2, pp. 345–367.
[11] Zaidan, O., and Callison-Burch, C. (2021). Arabic Natural Language Processing in the Age of Deep Learning: Challenges and Opportunities, Transactions of the Association for Computational Linguistics (TACL), Vol. 9, pp. 123–145.
[12] AlSarrar, H., AlShameri, N., AlShareef, N., AlShareef, M., AlGhamdi, N., AlZaydi, S., … AlYahya, M. (2022). *Arabic dialogue systems: A survey*. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Éds.), *Proceedings of Seventh International Congress on ICT* (Lecture Notes in Networks and Systems, vol. 465, pp. 153–161). Springer.
[13] Rahman, A., Kabir, M. M., Mridha, M. F., Alatiyyah, M., Alhasson, H. F., & Alharbi, S. S. (2024). Arabic speech recognition: Advancement and challenges. *IEEE Access*.
[14] Elharati, H. A., Alshaari, M., & Këpuska, V. Z. (2020). Arabic speech recognition system based on MFCC and HMMs. *Journal of Computer and Communications*, *8*(3), 28-34.
[15] Sakoe, H. (1979). Two-level DP-matching—a dynamic programming-based pattern matching algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(6), 588–595. https://doi.org/10.1109/TASSP.1979.1163264
[16] Abdelrazaq, D., Abu-Soud, S., and Awajan, A. (2018). A Machine Learning System for Distinguishing Nominal and Verbal Arabic Sentences, the International Arab Journal of Information Technology, Vol. 15, No. 3A.
[17] Ali, A., Vogel, S., and Renals, S. (2017). Speech recognition challenge in the wild: Arabic MGB-3, IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan.
[18] Al-Anzi, F. S., and Abuzeina, D. (2017). The Impact of Phonological rules on Arabic speech recognition, International Journal of Speech Technology, Vol. 20, No.3.
[19] Cucu, H., Buzo, A., Besacier, L., and Burileanu, C. (2015). Enhancing ASR Systems for Under-Resourced Languages through a Novel Unsupervised Acoustic Model Training Technique, Advances in Electrical and Computer Engineering, Vol. 15, No.1, pp.63-68.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

[20] Dukes, K., Atwell, E., and Habash, N. (2013). Supervised Collaboration for Syntactic Annotation of Quranic Arabic, Language Resources and Evaluation Journal, Vol. 47, No. 1, pp. 43-62.
[21] Hahm, S. J., Boril, H., Pongtep, A., and Hansen, J. H. L. (2013). Advanced Feature Normalization and Rapid Model Adaptation for Robust In-Vehicle Speech Recognition, Proceedings of the 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems, pp. 14-17, Seoul.
[22] Hamdani, G. D., Selouani, S., and Boudraa, M. (2012). Speaker-Independent ASR for Modern Standard Arabic: Effect of regional accents, International Journal of Speech Technology, Vol. 15, No. 4.
[23] Jokic, I., Delic, V., Jokic, S., and Peric, Z. (2015). Automatic Speaker Recognition Dependency on Both the Shape of Auditory Critical Bands and Speaker Discriminative MFCCs, Advances in Electrical and Computer Engineering, Vol. 15, No. 4, pp.25-32.
[24] Kadim, A., Lazrek, A., and El Hadj, Y. (2013). Dual Hidden Markov Model-New Approach for an Accurate Arabic Part-of-Speech Tagging, International Journal of Computational and General Linguistics, Vol. 5, No. 1.

Authors' Biodata

Menzer Hilal is a PhD candidate in Industrial Engineering at the University of Batna 2. His research focuses on automatic Arabic language processing and human–machine dialogue systems. Specifically, he explores automatic speech recognition (ASR), phonetic feature extraction, and the application of dynamic programming algorithms in speech signal processing. He is particularly interested in enhancing voice interfaces for under-resourced linguistic environments, especially those involving colloquial Arabic.
Research Interests : Speech recognition, Arabic natural language processing (NLP), intelligent systems, machine learning for spoken language understanding.

Samir Abdelhamid is a Professor in the Department of Industrial Engineering at the University of Batna 2, Algeria. He specializes in intelligent systems, natural language processing, and human–machine interaction. His research includes significant contributions to Arabic speech recognition and the development of computational tools for processing under-resourced languages. Prof. Abdelhamid is also actively involved in applying machine learning techniques to improve real-time communication systems and industrial automation.
Research Interests: Speech recognition, natural language processing, intelligent systems, Arabic language technologies, and machine learning.

Authors' Contribution

Hilal Menzer participated in "Conceptualization, Methodology, Formal Analysis, Investigation, Supervision, Resources, Data Preparation, and Writing Original draft

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

preparation.”. *Samir Abdelhamid* participated in “Software, Validation, Formal Analysis, and Writing Original draft preparation.

Declaration of conflicting interest

The authors declared no conflicts of interest with respect to the research, authorship, and/or publication of the article.

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Available online online at https://www.asjp.cerist.dz/en/Articles/155