Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 3, March 2021, pp. 1417~1423
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i3.pp1417-1423

A new framework based on KNN and DT for speech identification through emphatic letters in Moroccan dialect

Bezoui Mouaz[1], Cherif Walid[2], Beni-Hssane Abderrahim[3], Elmoutaouakkil Abdelmajid[4]
[1]Faculty of Science El jadida (FSJ), Université Chouaïb Doukkali (UCD), Morocco
[2,3,4]SI2M Labortory, National Institute of Statistics and Applied Economics Rabat, Morocco

Article Info

Article history:

Received Apr 17, 2020
Revised Jul 8, 2020
Accepted Aug 30, 2020

Keywords:

Decision tree
Hidden markov model
K-nearest neighbor
Machine learning
Speaker identification

ABSTRACT

Arabic dialects differ substantially from modern standard arabic and each other in terms of phonology, morphology, lexical choice and syntax. This makes the identification of dialects from speeches a very difficult task. In this paper, we introduce a speech recognition system that automatically identifies the gender of speaker, the emphatic letter pronounced and also the diacritic of these emphatic letters given a sample of author's speeches. Firstly we examined the performance of the single case classifier hidden markov models (HMM) applied to the samples of our data corpus. Then we evaluated our proposed approach KNN-DT which is a hybridization of two classifiers namely decision trees (DT) and K-nearest neighbors (KNN). Both models are singularly applied directly to the data corpus to recognize the emphatic letter of the sound and to the diacritic and the gender of the speaker. This hybridization proved quite interesting; it improved the speech recognition accuracy by more than 10% compared to state-of-the-art approaches.

Corresponding Author:

Bezoui Mouaz
Department of Computer Science
Chouaïb Doukkali University
24000, El jadida, Morocco
Email: mbezoui@gmail.com

1. INTRODUCTION

Automatic speech recognition is gaining big interest due to the high viability in speech signals. Indeed, authors may express their ideas in different accents, dialects, and pronunciations. They also differ from person to person, and between the two genders.

The existence of ambient noise, sound echoes, sound recording devices and stereo microphones results in additional variability. The hidden markov model (HMM) is used by conventional speech recognition systems to represent the sequential structure of speech signals. There are four emphatic consonants in Arabic which are of interest here: two of them are plosives: / d / and / t / and the other two are fricatives: / s / and / z / [1]. The letter / d / is an empathic plosive expressed with an alveo-dental articulation point, as this phoneme is rare in human languages [2].

Arabic is commonly referred to as "The Dhaad language", where Dhaad is the name of the spoken Arabic letter that carries the phoneme / d /. Additionally, this name was given to Arabic based on the classical Arabic version of / d / phoneme which is an emphatic lateral fricative, but not plosive as given by the MSA version. The letter / t / is an unvoiced emphatic plosive with an alveo-dental articulation point while the letter / s / is an unvoiced emphatic fricative with an alveo-dental articulation point. The famous letter / z / is an empathetic fricative expressed with an interdental point of articulation [3]. We explain in Table 1 the four Arabic emphatic sounds as well as their non-emphatic counterparts.

Table 1. Arabic emphatic consonants (ظ, ط, ص, ض)

Arabic letter	LDC Symbol	Non-Emphatic Conterparts
Dhaad ض	d	/d/ Daal
Saad ص	s	Voiced: /z/ (Zain); Unvoiced: /s/ (Seen)
Taà ط	t	Voiced: /d/ (Daal); Unvoiced: /t/ (Taà)
Thaà ظ	z	/th/ (Thaal)

In the literature, the majority of previous works for Arabic Speaker Identification has focused on MSA speech recognition. Speech data are usually news broadcasts where MSA is the formal language [4, 5]. Different classifiers have been investigated in this sense such as neural networks (NN) [6, 7], K-nearest neighbors (KNN) [8] and hidden markov models (HMM) [9, 10]. The common objective of these works was the improvement of models' accurccies [11]. Other researchers opted for ensemble methods [12]. They combined different classifiers during the training stage [13, 14] and the final decision was based on a comparison of obtained individual results. Another category of works proposed hybridizations of these classifiers by making optimizations to their algorithms. In this paper, we propose a new hybrid KNN-DT classifier. The main idea is to combine the robustness of KNN with the representativeness of DT classifier. The computational time was also investigated, and we propose in our hybrid model an optimal accuracy in a reasonable time for the recognition. The rest of this paper is organized as follows. In Section 2, we start by describing the datasets and its preprocessing steps before summarizing main used classifiers. The last part of section introduces the proposed framework which is based on a combination of KNN and DT. Section 3 discusses the experimental results. Finally, the last section concludes this work.

2. METHOD

2.1. Data description

We experiment three systems, hidden markov models, decision tree and KNN. In the case of the one based on a decision tree, the emphatic consonant /d/ was not always recognized properly. This was due to the specific characteristics of consonant's acoustics which make it difficult to be pronounced and therefore to be recognized [15]. Only few native speakers are able to pronounce it correctly. On all compounds, emphatic consonants yielded considerably lower accuracies compared to fricatives, nasals and plosives consonants [16].

While experimenting the proposed model for Speech recognition, we used a dataset consisting of 720 sounds. 12 people participated in this collection: 4 women and 8 men. 4 letters with 3 diacritics have been considered during the experimentation. Each participant recorded each sound 5 times. We divided this collection as follows:

The first four records of participants: woman1, woman2, woman3, man1, man2, man3, man5, man6, man7 are considered as a training set; while the fifth record of each of them, and all records of woman4, man4 and man8 are used for the test. To evaluate the proposed model, we used as metric the accuracy which is the ratio of sounds correctly predicted divided by the global number of sounds in the test set. In Table 2 the four Arabic consonants grapheme-phoneme correspondences.

Table 2. Arabic consonants table: grapheme-phoneme correspondences

Arabic Consonants	Occlusive	Emphatic	Fricative
Labial	ب		ف
Inter-dental		ظ	ث ذ
Dental	ت د	ط ض	
Pharyngeal			ح ع
Velar	ك		
Glotal	ء ‘		ه

2.2. Data preparation

The training corpus was collected in both MSA and Moroccan Arabic dialect. The records were carried out at the frequency: 16Khz. The collected recordings were segmented in order to clean the speech from external sounds such as noise and background music [17]. A high-quality microphone (Labtec AM-232) was used to record speechs from the authors over the period of October 2019 until December 2019.

2.3. The used classifiers overview

2.3.1. Hidden markov model

Among many other approaches, HMMs are proven to be the most efficient method in speech recognition. The reasons behind this method's popularity are the following ones: the availability of training algorithms for estimating the parameters of the models from finite training sets of speech data, its solid mathematical foundation as well as its ability to model time series with variable lengths. Despite its great success, it is well known that one of the most weaknesses of the conventional HMM classifier is the great number of tuning parameters (e.g., the states number, the number of Gaussian per state and the number of training iterations). These parameters have to be set experimentally and they are crucially dependent on the training and test data which affects the robustness of the classifier. Hidden markov models, introduced in the early 1970s, became the perfect solution to the problems of this subfield, namely the automatic speech recognition. The acoustic signal of speech is modeled by a small set of acoustic units, which can be considered as elementary sounds of the language. Traditionally, the chosen unit is the phoneme, thereby the word is formed by concatenating them [18]. More specific units can be used as syllables, disyllables, phonemes in context, and by such means making the model more discriminating, but this theoretical improvement is limited in practice by the complexity involved and estimation problems. The speech signal can be likened to a series of units. In the context of Markov ASR, the acoustic units are modeled by HMM which are typically left-right tristate as shown in Figure 1.

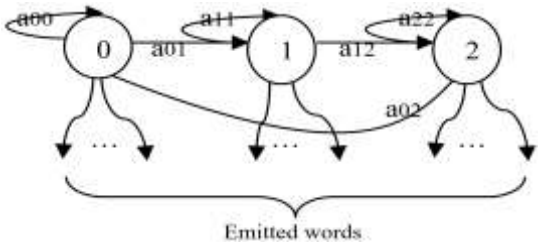

Figure 1. Hidden markov model (HMM) used topology

At each state of the Markov model, there is a probability distribution associated; modeling the generation of acoustic vectors via this state. An HMM is characterized by several parameters [19]:

N: the number of the states of the model.

The matrix of transition probabilities on the set of states of the model is calculated by the equation below:

$$A = \{a_{ij}\} = \{P(q_{t=1}|q_{t-1=i})\}$$

The matrix of emitting probabilities of the observations X_t for the state q_k is defined by:

$$B = \{b_k(X_t)\} = \{P(X_t|q_{t=k})\}$$

π is the initial distribution of states $(q_{i=0})$.

2.3.2. Decision trees

Our first experimentation concerns the prediction of sounds contents by using decision trees. The choice of this model is justified by the explicit rules it generates. The decision tree method is very easy to read and interpret. It illustrates that machine learning is not always synonymous with statistical models, but it can also target symbolic objects [20]. In the case of road accident, it is of great importance as the main purpose behind the use of machine learning techniques is to show visible rules to orient the actions.

Several algorithms have been proposed to build the optimal tree, the best-known ones are the ID3 algorithm [21] which was designed for the nominal attributes and its successors C4.5 and C5 [22] which also support the quantitative attributes. Technically, the information gain of the set *L* with respect to the feature x_j is therefore the variation of entropy [23] caused by the partition of *L* according to x_j:

$$Gain\ (L, x_j)\ =\ H(L)\ -\sum_{v\ \in\ valeurs\ (x_j)} \frac{card(l_{x_j=v})}{card(L)} H(l_{x_j=v})$$

$l_{x_j=v}$ refers to the set of accident for which the feature x_j has the value v.

Similarly, the gain ratio is computed by:

$$Gain\ ratio\ (L, x_j) = \frac{Gain\ (L, x_j)}{SplitInfo\ (L, x_j)}$$

Then, the feature having the highest information gain / gain ratio is the most significative.

2.3.3. K-nearest neighbors

Our second experimentation concerns another lazy machine learning model, namely k-nearest neighbors. This second technique is robust to noise. It was used in several pattern recognition applications [24-26]. However, it needs very high storage and computational time for large volumes of data. Its algorithm is based on similarities. It is considered as one of the simplest learning algorithms. When classifying a given sample, the algorithm votes its most similar samples in the sense of a predefined distance, and the class of the new sample is then determined by the majority among these k most similar samples (nearest neighbors) [27]. The performance of KNN depends largely on two factors: the value of k number and the measure used [28].

2.3.4. The proposed approach

The proposed model for Speaker Identification is a hybridization of decision trees (DT) and k-nearest neighbors (KNN). Both models are first applied directly on the data to recognize the letter of the sound, the diacritic and the gender of each speaker.

Instead of directly predicting the content of a sound, we judged better practice to detect first the gender of the author and the diacritic of the sound. Once these two parameters predicted, we add them as additional features to recognize the overall content. By doing so, we considerably refine the prediction by eliminating sounds that may contain different diacritics or letters pronounced by participants from the other gender.

Overall sound content prediction:
Accuracy of the decomposed model: 71.43%.
Improvement: 12.1%.

3. RESULTS AND DISCUSSION

First, we applied previously cited algorithms directly:

In Tables 3-5, we note that KNN is the most appropriate to recognize diacritics with an accuracy exceeding 71%, followed by HMM, while DT returns its highest accuracy for gender prediction. Our proposed approach will combine both predictors in an attempt to improve the overall performance.

The comparison of the three techniques: DT, KNN and the decomposed approach are shown in the following Figure 2. Figure 2 shows that the sound content prediction is improved by the sub-predictions: gender, letter and diacritic. The accuracy increased by 12.1%. Indeed, the delimitation of author's gender, as well as the designation of the letter reduces the risk of error caused by severe female accents or acute male accents; then the designation of the letter guides the prediction of diacritics towards a reduced list, thereby increasing the accuracy of the overall sound content prediction.

Table 3. HMM Accuracy for different predictions

Prediction	Gender prediction	Diacritic prediction	Sound content prediction
HMM Accuracy	63.42%	70.64%	67.03%

Table 4. DT Accuracy for different predictions

Prediction	gender prediction	diacritic prediction	sound content prediction
DT Accuracy	66.67%	63.56%	59.33%

Table 5. KNN Accuracy for different predictions

Prediction	gender prediction	diacritic prediction	sound content prediction
KNN Accuracy	52.78%	71.62%	59.33%

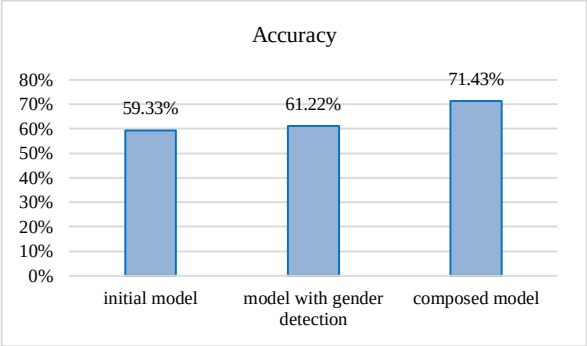

Figure 2. Accuracy of the three different experimented models

4. CONCLUSION AND PERSPECTIVES

In this paper, we presented a new model of speech identification for the Arabic language. Instead of applying traditional machine learning classifiers to recognize speechs, we investigated the feasibility of decomposing the identification task into gender, emphasis letters, and speech categorization. We proposed this hybridization to supersede the common mapping of the global sound to its corresponding dialect. By dividing this speech identification task, we improved the accuracy of our model by 12.1%. This novel approach can be used for various language processing applications such as sentiment analysis, voicebots… In our future works, we will continue our investigations for approaches that would directly map the raw acoustic waveform to the corresponding dialects. Actually, we are exploring long short-term memory (LSTM) and recurrent neural network (RNN) to further improve our hybrid model. Another line of research worth exploring is the effect of integrating social data during the training stage. This could help to detect new correlations between records and to highlight new important features for speech recognition.

ACKNOWLEDGEMENTS

We would like to acknowledge the main site where research was carried out; Department of Computer, Chouaib Doukkali University, Faculty of Science, EI Jadida, Morocco.

REFERENCES

[1] Ouni, S., Cohen, M., Massaro, W., "Training Baldi to be Multilingual: A Case Study for an Arabic Badr", *Speech Communication*, vol. 45, pp. 115-37, 2005.
[2] Al-Muhtaseb, H., Elshafei, M., & Alghamdi, M. "Techniques for High Quality Arabic Text-tospeech", *In The Third Workshop on Computer and Information Sciences* pp. 73-83, 2000.
[3] Selouani, S., Caelen, J., "Arabic Phonetic Features Recognition using Modular Connectionist Architectures", *Interactive Voice Technology for Communication, IVTTA'98, Proceedings 1998 IEEE 4th Workshop 29*, pp. 155-160, 1998.
[4] Maamouri, M., Bies, A., & Kulick, S. "Diacritization: A challenge to Arabic treebank annotation and parsing", *In Proceedings of the Conference of the Machine Translation SIG of the British Computer Society*.
[5] Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., ... & Haddad, B. "Building Annotated Written and Spoken Arabic LRs in NEMLAR Project", *In LREC*, pp. 533-538, 2006.
[6] Masmoudi, S., Frikha, M., Chtourou, M., & Hamida, A. B. "Efficient MLP constructive training algorithm using a neuron recruiting approach for isolated word recognition system", *International Journal of Speech Technology*, vol. 14, no. 1, pp. 1-10, 2011.
[7] Dhanashri, D., & Dhonde, S. B. ''Isolated word speech recognition system using deep neural networks", *In Proceedings of the international conference on data engineering and communication technology*, pp. 9-17, 2017.
[8] Xu, B., Wang, N., Chen, T., & Li, M. "Empirical evaluation of rectified activations in convolutional network", 2015. arXiv preprint arXiv:1505.00853.
[9] Khelifa, M. O., Elhadj, Y. M., Abdellah, Y., & Belkasmi, M. "Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system", *International Journal of Speech Technology*, vol. 20, no. 4, pp. 937-949, 2017.

[10] Rabiner, L. R., Wilpon, J. G., & Soong, F. K. "High performance connected digit recognition using hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 8, pp. 1214-1225, 1989.
[11] Zhang, X., Sun, J., & Luo, Z. ''One-against-all weighted dynamic time warping for language-independent and speakerdependent speech recognition in adverse conditions", PLoS ONE, vol. 9, no. 2, e85458, 2014. https ://doi.org/10.1371/journ al.pone.00854 58.
[12] Hazmoune, S., Bougamouza, F., Mazouzi, S., & Benmohammed, M. "A new hybrid framework based on Hidden Markov models and K-nearest neighbors for speech recognition", *International Journal of Speech Technology*, vol. 21, no. 3, pp. 689-704, 2018.
[13] Hazmoune, S., Bougamouza, F., Mazouzi, S., & Benmohammed, M. "A novel speech recognition approach based on multiple modeling by hidden Markov models", *In International Conference on Computer Applications Technology (ICCAT)*, pp. 1-6, 2013. Sousse: IEEE.
[14] Zhang, X., Povey, D., & Khudanpur, S. "A diversity-penalizing training method for deep learning", *In INTERSPEECH*, pp. 3590-3594, 2015.
[15] Hamza, M., Khodadadi, T., & Palaniappan, S. A "Novel automatic voice recognition system based on text-independent in a noisy environment", *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3643, 2020.
[16] Ouisaadane, A., Safi, S., & Frikel, M. ''Arabic digits speech recognition and speaker identification in noisy environment using a hybrid model of VQ and GMM", TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 18, no. 4, pp. 2193-2204, 2020.
[17] Mouaz, B., Abderrahim, B. H., & Abdelmajid, E., "Speech Recognition of Moroccan Dialect Using Hidden Markov Models", *International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 1, 2019, DOI: 10.11591/ijai.v8.i1.pp7-13.
[18] Rabiner L-R., Juang B-H., "Fundamentals of Speech Recognition", Prentice-Hall, 1993.
[19] Cing, D. L., & Soe, K. M. "Improving accuracy of part-of-speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language", *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 10, pp. 2088-8708, 2020.
[20] Quinlan, J.R, "Simplifying decision trees", *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221-234, 1987.
[21] Quinlan, J.R. "Induction of decision trees", Mach Learn vol. 1, pp. 81–106, 1986, https://doi.org/10.1007/BF00116251.
[22] Quinlan, J. R., & Cameron-Jones, R. M. "FOIL: A midterm report", In European conference on machine learning pp. 1-20, 1993.
[23] Graja S., and Boucher J., "Hidden Markov tree model applied to ECG delineation," *in IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 6, pp. 2163-2168, 2005.
[24] Alkhateeb, F., Baget, J-F, et Euzenat, J., "Extending SPARQL with regular expression patterns (for querying RDF)", *Journal of web semantics*, vol. 7, no 2, pp. 57-73, 2009.
[25] LI, J. et WANG, J., "System and method for automatic linguistic indexing of images by a statistical modeling approach". U.S. Patent no. 7, pp. 394,947, 2008.
[26] Wang, X. H., Liu, A., & Zhang, S. Q. "New facial expressionrecognition based on FSVM and KNN", *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 21, pp. 3132-3134, 2015.
[27] Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. *Procedia Computer Science*, vol. 127, pp. 293-299, 2018.
[28] Cherif, W., Madani, A., & Kissi, M. "A combination of low-level light stemming and support vector machines for the classification of Arabic opinions", *In 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pp. 1-5, 2016. IEEE.

BIOGRAPHIES OF AUTHORS

Bezoui Mouaz (born in 1985) received a master's degree of networks and telecommunications in 2011 from Chouaïb Doukkali University, Faculty of Science EI Jadida, Morocco. He is currently pursuing his Ph.D degree in Computer Science at the Chouaïb Doukkali University Faculty of Science, El Jadida, Morocco. Bezoui Mouaz is the author of over 3 technical publications. His research interests include Artificial Intellegence, Natural Language Processing, Machine Learning, Speech Recognition and Signal Processing.

Walid Cherif has completed his PhD in Data Science from Chouaib-Doukkali University (Morocco). He is an IT engineer from the National Institute of Statistics and Applied Economics (Rabat, Morocco) in which he is now member of the SI2M laboratory. He is author of many papers in reputed journals and international conferences. His research interests include: Data Science, Artificial Intelligence, Machine Learning, Deep Learning, Text and Data Mining.

Beni-hssane Abderrahim is currently an Full Professor in the Department of Computer Science at Chouaïb Doukkali University, Faculty of Science, EI Jadida, Morocco, in which he is now member of the LAROSERI laboratory. His research interests include Data Mining and Knowledge Discovery. He is author of many papers and technical publications in reputed journals and international conferences. His research interests include Artificial Intelligence, Natural Language Processing, Text and Data Mining, Machine Learning, Internet of Things and Big Data.

Elmoutaouakkil Abdelmajid is currently a Full professor in the Department of Computer Science at Chouaïb Doukkali University, Faculty of Science, EI Jadida, Morocco, in which he is now member of the LAROSERI laboratory. He is author of many papers and technical publications in reputed journals and international conferences. His research interests include Image Processing, Speech Recognition, Data Mining, Knowledge Discovery, Machine Learning and Big Data.