IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 8, No. 1, March 2019, pp. 7~13
ISSN: 2252-8938, DOI: 10.11591/ijai.v8.i1.pp7-13

Speech recognition of moroccan dialect using hidden markov models

Bezoui Mouaz, Beni-hssane Abderrahim, Elmoutaouakkil Abdelmajid
Department of Computer Sciences, Faculty of Sciences, Chouaib Doukkali University, Eljadida, Mororcco

Article Info

Article history:

Received Dec 27, 2018
Revised Feb 18, 2019
Accepted Feb 26, 2019

Keywords:

ASR
DA
HMM
MFCC
MSA

ABSTRACT

This paper addresses the development of an Automatic Speech Recognition (ASR) system for the Moroccan Dialect. Dialectal Arabic (DA) refers to the day-to-day vernaculars spoken in the Arab world. In fact, Moroccan Dialect is very different from the Modern Standard Arabic (MSA) because it is highly influenced by the French Language. It is observed throughout all Arab countries that standard Arabic widely written and used for official speech, news papers, public administration and school but not used in everyday conversation and dialect is widely spoken in everyday life but almost never written. we propose to use the Mel Frequency Cepstral Coefficient (MFCC) features to specify the best speaker identification system. The extracted speech features are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later. The Euclidean distance between the MFCC's of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification satisfactorily.

Corresponding Author:

Bezoui Mouaz,
Department of Computer Sciences,
Chouaib Doukkali University,
73, Lot Firdaouss, Appt 1, Av Jabrane Khalil Jabrane, El jadida 24100, Morocco.
Email: mbezoui@gmail.com

1. INTRODUCTION

The majority of previous work in Arabic ASR has focused on the formal standard Arabic language that is known as Modern Standard Arabic (MSA). MSA is not the language of ordinary discussions and several communications in all Arabic countries. The population use other Arabic varieties in everyday life that is known as Dialectal Arabic (DA). A significant problem in Arabic ASR is the existence of quite many different dialects e.g. Moroccan, Tunisian, Egyptian, Saudi, Iraqi etc. Every country has its own dialect, and sometimes there exist different dialects within the same country [1]. Moreover, the different Arabic dialects are only spoken and not formally written and significant, syntactic, lexical, morphological, phonological and differences exist between the dialects and the standard form. In this work we propose to use the Mel Frequency Cepstral Coefficient (MFCC) features for designing a sound-dependent and specify the best speaker identification system. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker [2]. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later. The Euclidean distance between the MFCC's of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification satisfactorily.

2. HIDDEN MARKOV MODELS

Hidden Markov Models, introduced in the early 1970s [3], became the perfect solution to the problems of 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. More specific units can be used as syllables, disyllables, phonemes in context, thereby 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 as shown in Figure 1 which are typically left-right tristate.

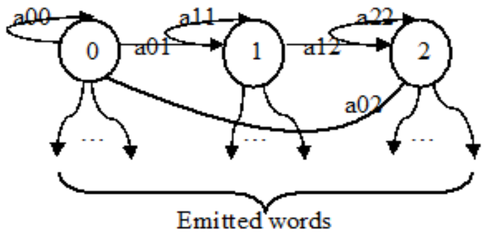


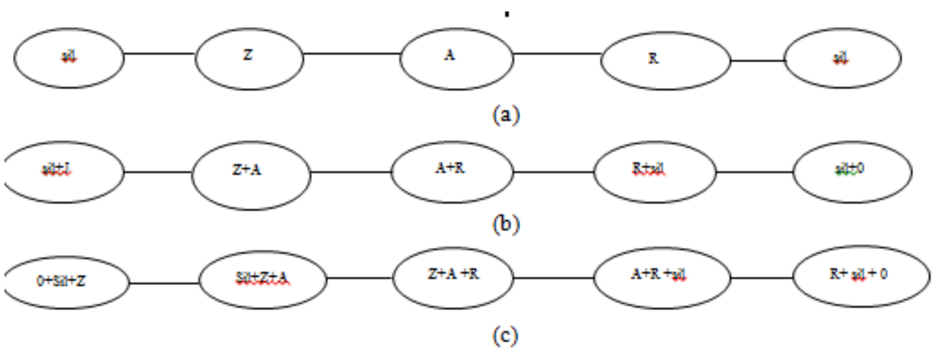
Figure 1. HMM used topology.

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:

- N: the number of the states of the model.
- $A = \{a_{ij}\} = \{P(q_{t=1}|q_{t-1=i})\}$ is the matrix of transition probabilities on the set of states of the model.
- $B = \{b_k(X_t)\} = \{P(X_t|q_{t=k})\}$ is the matrix of emitting probabilities of the observations X_t for the state q_k.
- π is the initial distribution of states $(q_{i=0})$.

3. ACOUSTIC MODELS AND PARAMETERS

The speech signal contains many other elements more than the linguistic message: information related to the speaker, the recording conditions, etc... In addition, the variability and redundancy of the speech signal makes it difficult to use as such. It is therefore necessary to extract the parameters that are dependent on the linguistic message. These parameters are estimated via sliding windows on the signal. This analysis window used to estimate the signal on a stationary portion of a considered signal: typically 10 to 30 ms limiting side effects and discontinuities of the signal via a Hamming window. In our experiments, we use 25ms as a window size. The majority of parameters represent the frequency spectrum and its evolution over a window size. Parameterization techniques that are the most commonly used are: PLP Perceptual Linear Prediction: spectral domain, LPCC Linear Prediction Cepstral Coefficients: time domain, MFCC Mel Frequency Cepstral coefficients: cepstral domain [4]. For our work, we have used MFCC parametrization for the feature extraction. Our first intervention in the recognition system is in the phase of labeling sound files. In large vocabulary ASR systems, DBNs are used to represent sub units of words (such as phones). For the Arabic language, it is typical to have around 38 models (phones). The exact phone set depends on the dictionary that is used. Word models can be constructed as a combination of the sub word models. In practice, the realization of one and the same phone differs a lot depending on its neighboring phones called 'phone context'[5]. Speech recognition use context depends on phonetic alphabets, in which there are one or more units for each phoneme in the context of surrounding phonemes. Several of the more common schemes are monophones, biphones and triphones. Figure 2 shows the arabic word 'زار' : [Z A R] in a monophone, biphone and triphone representation.



'sil' refers to the silence at the start and the end of the utterance, which is modeled as a 'phone' too.

Figure 2. (a) Monophone, (b) Biphone and (c) Triphone with one hidden variable (HMM) for the Arabic word 'ZAR'.

4. SPEECH RECOGNITION

Like any other pattern recognition systems, the process of performing speaker recognition consists on two phases namely: training and testing. Training is the process of familiarizing the system with the voice characteristics of the speakers registering by extract features from each speaker [6]. The block diagram of training phase is shown in Figure 3. Feature vectors representing the voice characteristics of the speaker are extracted from the training utterances and are used for building the reference models. During testing, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using several matching algorithms. feature matching process is performed to decide whether these features belong to a previously known speaker pattern or not. A schematic diagram of the testing phase as shown in Figure 4.

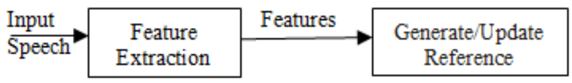


Figure 3. The block diagram of the training mode.

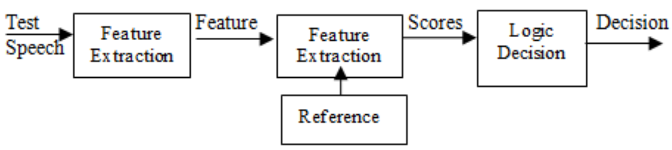


Figure 4. The block diagram of the recognition mode.

The process of performing speaker identification consists of two modes: a training mode and a recognition mode. In the training phase, a database of speaker's pattern is used to extract features from each speaker. These features are used to train a neural network. In the testing phase, features are extracted from every incoming speaker and a feature matching process is performed to decide whether these features belong to a previously known speaker pattern or not. A schematic diagram of the steps of the proposed detection system is shown in Figure 4.

The steps of the feature extraction process from a flow chart can be summarized as follows:

- The speech signal can be used in time domain or in another discrete transform domain. The DCT, DST and DWT can be used for this purpose.
- MFCCs and polynomial shape coefficients are extracted from either the speech signal, the discrete transform of the signal or both of them.

Both the training and the recognition modes include feature extraction, sometimes called the front-end of the system. The feature extractor converts the digital speech signal into a sequence of numerical descriptors, called feature vectors. The features provide a more stable, robust, and compact representation than the raw input signal. Feature extraction can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate [7]. During the training mode, each speaker in the set is modeled using a set of training data. Features are extracted from the training data essentially striping away all unnecessary information in the training speech samples leaving only the speaker characteristic information, with which speaker models can be constructed. In the recognition mode, features are extracted from the unknown speaker's voice sample. Pattern matching refers to an algorithm, or several algorithms, that compute a match score between the unknown speaker's feature vectors and the models stored in the database. The output of the pattern matching module is a similarity score. The last phase in the recognition chain is decision making. The decision module takes the match scores as its input, and makes the final decision of the speaker identity. It is clear that the feature extraction process (obtaining speaker discriminatory information) and the classification process (using the features to determine the correct speaker) algorithms are of critical importance to any speaker identification system.

4.1. Feature extraction

A single human speech signal contains a large amount of speaker dependent information. While the human brain is able to distinguish between speakers based on 'high-level' properties such as dialect, speaking style, context of the speech and the emotional state of the speaker, designing identification algorithms based on these properties is infeasible due to the required high complexity. It is possible however to build efficient identification algorithms based on the low-level properties of the signal such as pitch, intensity, formant frequencies and their characteristics. The concept of feature extraction contributes to the goal of identifying speakers based on the low-level properties in two ways. Firstly, the extraction produces sufficient information for good speaker discrimination and captures this information in a form and size that allow efficient modeling. Secondly, feature extraction can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. The feature extractor converts the digital speech signal into a sequence of numerical descriptors called feature vectors. Several feature extraction techniques are used in speaker recognition systems. The concept of feature extraction using the MFCCs is widely known in speaker identification, It contributes to the goal of identifying speakers based on the low-level properties [8]. It is clear that the speech signal has oscillatory patterns, which supports the application of the cepstral method for feature extraction from our speech signals. In speaker identification, the extraction produces sufficient information for good speaker discrimination. In the following subsection, an explanation for the extraction of the MFCCs and the polynomial coefficients is presented.

4.2. Extraction of MFCCs

The MFCCs are commonly extracted from speech signals through cepstral analysis. The input signal is first framed and windowed, the Fourier transform is then taken and the magnitude of the resulting spectrum is warped by the Mel-scale. The log of this spectrum is then taken and the DCT is applied shown in Figure 5. The 1-D signal must first be broken up into small sections; each of N samples. These sections are called frames and the motivation for this framing process is the quasistationary nature of the 1-D signals. However, if we examine the signal over discrete sections, which are sufficiently short in duration, then these sections can be considered as stationary and exhibit stable characteristics [9-11]. To avoid loss of information, frame overlap is used. Each frame begins at some offset of L samples with respect to the previous frame where LBN. For each frame, a windowing function is usually applied to increase the continuity between adjacent frames. Common windowing functions include the rectangular window, the Hamming window, the Blackman window and flattop window. Windowing in time domain is a pointwise multiplication of the frame and the window function. The magnitude spectrum $|\mathrm{X}(\mathrm{k})|$ is now scaled in both frequency and magnitude. First, the frequency is scaled logarithmically using the so-called Mel filter bank $\mathrm{H}(\mathrm{k}, \mathrm{m})$, and then the logarithm is taken, giving:

$$X'(m) = \ln\left(\sum_{k=0}^{N-1} |X(\mathrm{k})|.\,\mathrm{H}(k, m)\right)$$

for m = 1, 2,…, M, where M is the number of filter banks and M<<N.

The Mel filter bank is a collection of triangular filters defined by center frequencies calculated on the Mel scale (Srinivasan et al. 2004; Lungyun et al. 2006). The triangular filters are spread over the entire frequency range from zero to the Nyquist frequency. The number of filters is one of the parameters which affect the recognition accuracy of the system. Finally, the MFCCs are obtained by computing the DCT of $X'(m)$ using:

$$C_l = \sum_{m=1}^{M} X'(m) \cos\left(l \frac{\pi}{M}\left(m - \frac{1}{2}\right)\right)$$

for l = 1,2,…, M, where C_l is the lth MFCC.

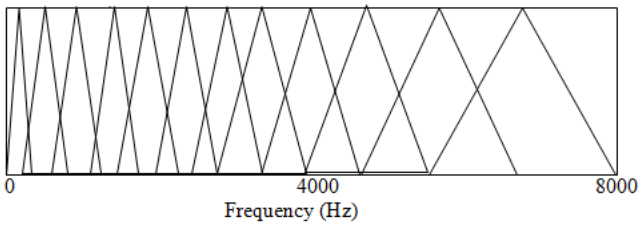


Figure 5. The Mel filter bank

The number of the resulting MFCCs is chosen between 12 and 20, since most of the signal information is represented by the first few coefficients. The 0th coefficient represents the average log energy of the frame. An experience with speech recognition[12-13], showed that it is beneficial to use also delta and delta-delta coefficients which decrease the word error rate. Even though the original set of features of the MFCC is more or less correlated then after addition of delta and delta-delta features the information redundancy of elements in feature vectors increases. Since in this system we are concerned with the spectral features such as MFCC features, we add different features related to the MFCC such as time derivatives[14]. The first order regression coefficients of the MFCC feature vector called Delta is included. Also, the second order regression coefficients, called Delta-Delta, is included.

5. RESULTS AND DISCUSSION

The database used in training and testing the system for each dialect is a combination of twenty speakers, eleven males and nine females. We chose 4 people, 3 women and a 2 mens to pronounce 4 words of the Moroccan dialect which are: (سلام) = (Hi), (كيداير) = (How are you), (لاباس) = (There is nothing wrong), (بخير) = (Fine) and we recorded the voices of the speakers in files in (.wav) format. later we began our work of the recognition and the lyrics of the Moroccan dialect by the part Learning [11], through "add a new sound from file" which invites the user to choose a file (.wav) and classify it by Identity , from ID:1 to ID:5[15]. we continued the training phase to build a database of files (.wav) with 4 classes, each class represents a well-defined speaker.The speech of eleven male speakers and nine female speakers are used for training, and the speech of one male speaker and three female speakers are used for testing. The speech for training from each speaker is one minute long. The speech for testing from each speaker is 10 second long [16]. Our HMM Speech Recognition System was programmed by using MATLAB to create a user interface and to enable user to add a new sound from audio files as shown in Figure 6.

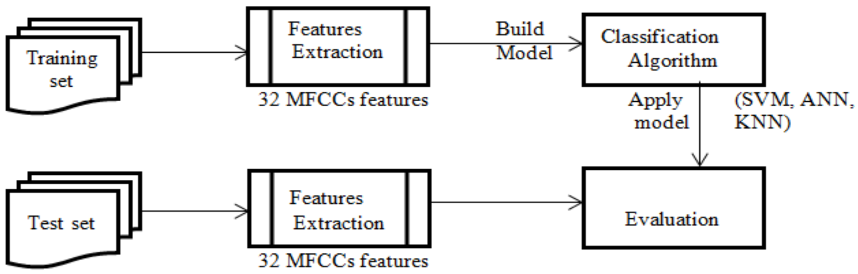


Figure 6. Flowchart of the deep learning model.

We present in this section the results of the identification process including the testing results. It should be mentioned that no detailed or word-level labelling was done for the database related to the testing step[17]; speech is labelled [18] according to the corresponding dialect data. For instance, a Moroccan speech file is labelled with the letter "M" while a MSA speech file is labelled with the letter "A". Training with both MSA and Moroccan Dialect as shown in Table 1.
We followed three ways of treatment for training step:
- Training with MSA files (speakers): The system could identify all MSA testing files and nothing of MOROCCAN files.
- Training with Moroccan files: The system could identify all MOROCCAN testing files and nothing from MSA files.
- Training with both MSA and Moroccan files: The system could identify all 10 testing files of MSA and 8 files from 10 testing files of Moroccan Dialect.

Table 1. Training with both MSA and Moroccan Dialect

Speakers	Number of attempts Salam سلام	 Kidayr كيداير	 Labas لاباس	 Bikhir بخير	Recognized word
ID 1			4		3
ID 2			4		4
ID 3			4		4
ID 4			4		3
ID 5			4		4
Total			20		18

6. CONCLUSION

The purpose of this work is to verify the ability of our HMM Speech Recognition System to distinguish the vocal print of speakers, and identify them by giving each of them a specific class. This is done through create a speech recognition system, and apply it to a Moroccan Dialect speech [19-20]. By investigating the extracted features of the unknown speech and then compare them to the stored extracted feature vectors for each different speaker, in order, to identify the unknown speaker. The model used in this work was the Hidden Markov Model [21]. The MFCC + Delta + Delta-Delta features performed best reaching an identification score. The accuracy of our HMMSRS (HMM Speech Recognition System) is about 90%.

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] Mohamed Amine Menacer, Odile Mella, Dominique Fohr, Denis Jouvet, David Langlois, Kamel Smaili. "Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect," 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates.

[2] S. Connel, A Comparison of Hidden Markov Model Features for the Recognition of Cursive Handwriting, Computer Science Department, Michigan State University, MS Thesis (1996).

[3] Rabiner L-R., Juang B-H., Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[4] Graves A,Mohamed A, Hinton G (2013) "Speech recognition with deep recurrent neural networks", In ICASSP 2013, pp. 1-5.

[5] Srinivasan et Thorpe and Shelton 1993 et al. 2004; Lungyun et al. 2006; Chiyi and Kubichek 1996; Lam et al. 1996.

[6] Afify, M., Sarikaya, R., kwang Jeff Kuo, H., Besacier, L., Gao, Y., 2006. "On the use of morphological analysis for dialectal arabic speech recognition", in: Proceedings of ICSLP06, pp. 277-280.

[7] Adnan Qayyum, Siddique Latif, Junaid Qadir, "Quran Reciter Identification: A Deep Learning Approach", Computer and Communication Engineering (ICCCE) 2018 7th International Conference on, pp. 492-497, 2018.

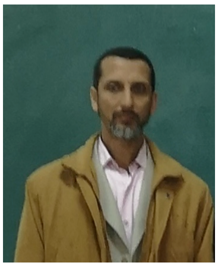
[8] Salima Harrat , Karima Meftouh, Mourad Abbas, Khaled-Walid Hidouci and Kamel Smaili, "An Algerian dialect: Study and Resources", Published in: IJACSA (International Journal of Advanced Computer Science and Applications), Vol. 7, No. 3, 2016.

[9] Tachicart Ridouane and Bouzoubaa Karim. 2014. "A hybrid approach to translate moroccan arabic dialect". In SITA'14, 9th International Conference on Intelligent Systems.
[10] Mouaz Bezoui, Abdelmajid Elmoutaouakkil, and Abderrahim Benihssane. Feature extraction of some Quranic recitation using melfrequency cepstral coeficients (MFCC). In Multimedia Computing and Systems (ICMCS), 2016 5th International Conference on, pages 127– 131. IEEE, 2016.
[11] Jyoti Guglani, A. N. Mishra, "Continuous Punjabi speech recognition model based on Kaldi ASR toolkit", International Journal of Speech Technology, 2018.
[12] Tachicart Ridouane, Karim Mohamed Bouzoubaa, Si Lhoussain Aouragh, Hamid Jaafar. "Automatic Identification of Moroccan Colloquial Arabic". January 2018 in book Arabic Language Processing: From Theory to Practice.
[13] Afify, M., Sarikaya, R., kwang Jeff Kuo, H., Besacier, L., Gao, Y., 2006. "On the use of morphological analysis for dialectal arabic speech recognition", in: Proceedings of ICSLP06, pp. 277–280.
[14] Ali A., Dehak N., Cardinal P., Khurana S., Yella S., Glass J., Bell P., Renals S., 2016. "Automatic dialect detection in arabic broadcast speech". CoRR 08-12-September-2016, 2934–2938. doi:10.21437/Interspeech.2016-1297.
[15] Ali, A., Mubarak, H., Vogel, S., 2014a. "Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian ASR", in: International Workshop on Spoken Language Translation (IWSLT 2014).
[16] Elmahdy, M., Hasegawa-Johnson, M., Mustafawi, E., 2012. "A baseline speech recognition system for levantine colloquial arabic". Proceedings of ESOLEC .
[17] Elmahdy, M., Hasegawa-Johnson, M., Mustafawi, E., 2014. "Development of a tv broadcasts speech recognition system for qatari arabic", in: LREC, pp. 3057–3061.
[18] Soltau, H., Mangu, L., Biadsy, F., 2011. "From modern standard arabic to levantine ASR: Leveraging gale for dialects", in: 2011 IEEEWorkshop on Automatic Speech Recognition Understanding, pp. 266–271. doi:10.1109/ASRU.2011.6163942.
[19] Vergyri, D., Kirchhoff, K., Gadde, V.R.R., Stolcke, A., Zheng, J., 2005. "Development of a conversational telephone speech recognizer for levantine arabic", in: Proceedings of Interspeech, Citeseer. pp. 1613–1616.
[20] Mouaz Bezoui, Abderrahim Beni-hssane and Abdelmajid Elmoutaouakkil, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models" International Symposium on Machine Learning and Big Data Analytics for Cybersecurity and Privacy (MLBDACP) April 29 – May 2, 2019, Leuven, Belgium.
[21] Wray, S., Ali, A., 2015. "Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic". International Speech and Communi-cation Association. volume 2015-January. pp. 2824–2828.

BIOGRAPHIES OF AUTHORS

Bezoui Mouaz (born in 1985) received a master's degree of networks and telecommunications from Chouaib Doukkali University, Faculty of Science EI Jadida, Morocco in 2011. He is currently pursuing his Ph.D. degree (Computer Science) at the Chouaib Doukkali University Faculty of Science, El Jadida, Morocco. His research interests include Speech Recognition and Signal Processing.

Beni-hssane Abderrahim is currently a professor in Chouaib Doukkali University Faculty of Science, Department of Computer science, LAROSERI Laboratory, EI Jadida, Morocco. His research interests include Data Mining and Knowledge Discovery.



Elmoutaouakkil Abdelmajid is currently a professor in Chouaib Doukkali University Faculty of Science, Department of Computer science, LAROSERI Laboratory, EI Jadida, Morocco. His research interests include Medical Image Processing and Speech Recognition.