

New parameters for resolving acoustic confusability between Arabic phonemes in a phonetic HMM recognition system

I. A. Maaly¹, A.R. Elobeid² & K.M. Ali Ahmed²

¹*Department of Electrical Engineering, University of Khartoum, Sudan*

²*Physics Department, Faculty of Science, Qatar University, Qatar*

Abstract

A Hidden Markov Model (HMM) recognition system is implemented for Arabic Phonemes as units of recognition. An important result of this system is the confusion encountered between some phonemes of Arabic (e.g., /h/ (Ha), /ʔ/ (Hamza)), i.e., the recognizer could not distinguish between them. New parameters, which are based on a new classification of Arabic phonemes, are added at a higher level of the system for resolving this acoustic confusability and improving the recognition accuracy. These new parameters, which are based on the shape of the tongue and the place of articulation along the vocal tract, are called the Emphatic/nonemphatic, Root and Hamz parameters. The “Emphatic/nonemphatic” parameter gives a performance of 44%, the “Root” parameter gives a performance of 40%, while the “Hamz” parameter could resolve confusability encountered between the phonemes /h/ and /ʔ/ with a performance of 90%. These new parameters may need accurate estimates of the distances along the vocal tract to improve their performance.

1 Introduction

The objective of this research is to achieve a robust recognition system for Arabic language. Arabic speech has its distinctive features, e.g., it has a large number of pharyngeal sounds (e.g., /x/ (Kha), /ʔ/ (Hamza) and a large number of

emphatic ones (e.g., /S/ ($\dot{S}a$). It also has the two palatal sounds /T/ ($\dot{T}a$) and /D/ ($\dot{D}a$) that are rarely used in other languages.

Parametric representation of Arabic speech is used to help realize an Arabic speech recognition system. Each Arabic phoneme is modeled as a Hidden Markov Model (HMM). In the HMM the speech phonemes could be recognized through the observed parameters that are derived from the speech signal.

In section (2) the signal processing methodology for deriving the speech signal parameters is explained. Also, the vector quantization method, which is used to reduce the computational complexity, is outlined. In section (3) the implementation of the phoneme HMM is discussed. The new parameters derived for a second level of the recognizer are given in section (4).

2 Signal processing

2.1 Spectral analysis

The database used in this work consists of five repetitions of a set of 31 monosyllabic words uttered by three Sudanese Arabic speakers. These words are meaningless, but are of the same structure /cvc/ or /cv/. Each word in the database starts with the plosive sound /ʔ/ (Hamza), followed by the short vowel /a/ (Fatha), and ends with one of the 28 Arabic consonants, or one of the three vowels (the structure becomes (/cv:/). These words are chosen according to Sibawaihi advice (the famous Arabic phonetician in the 8th century [2]). He mentioned that a good perception of Arabic phonemes can be achieved by articulating any of them preceded by (Alif Hamz), (e.g., /ʔar/, /ʔam/. The semivowels /j/ (Ya), and /w/ (Wa) are articulated in a /cv/ structure (e.g., /ja/ (Ya), /wa/ (Wa)). The three short vowels of Arabic are not included because we are interested in the features of Arabic phonemes and not in their duration.

After a comprehensive study of Arabic phonetics [1][2], each of the 31-monosyllabic words has been hand-labeled to the level of individual phonemes. These Arabic phonemes are ideal since they are not affected by context. A total of 31 HMMs are trained and decoded using these phonemes.

Analysis is accomplished as follows [3], [4], [5]:

Firstly, speech is sampled at 10 kHz. Then, the pre-emphasis filtering is accomplished on speech signal samples using a first order differentiator [5].

Sections of N_s consecutive speech samples are used as frames. In this research, we used $N_s = 200$ samples corresponding to 20 ms of signal. The total number of frames is denoted N_f . The combination of 20 ms frame duration and (5 ms) overlap is used in this system.

Parameters used for representing the speech signal contain a set of LP-derived cepstral coefficients plus power, delta power, and delta-delta power. The covariance lattice method of the Linear Predictive Coding (LPC) analysis technique is implemented [3], [4], [6], [7], [8]. For each frame (l) a set of 13 LPC coefficients denoted (In next sections, the subscript will denote frame

number (l) and index of prediction parameter will be included between brackets.) $\{a_l(i)\}$ plus a set of 13 reflection coefficients denoted $\{k_l(i)\}$ are computed (where $1 \leq i \leq P$ and $1 \leq l \leq N_f$). $P=13$ is the number of prediction coefficients.

Then a set of 13 cepstral coefficients is obtained from the predictor coefficients for each time frame [4][5].

Power is also computed on a frame-by-frame basis. A single value of power is obtained for each frame of the signal [5]. For obtaining the delta power parameters, a first order differentiator is used.

We obtain the second order derivative of power by reapplying the differentiator.

The final set of parameters used for the recognition is the cepstral coefficients with power, delta power, and delta-delta power, which contains 16 parameters.

2.2 Area function

The area function is the estimation of the vocal tract shape. It is the dependence of cross sectional area upon the length along the tract, which can represent the place of articulation. Since the perception of Arabic speech depends mainly on a correct place of articulation, the place of articulation of each phoneme is supposed to be one of the important factors for Arabic speech recognition systems. For this reason, using the area function as a combination of acoustic and articulatory parameter is expected to improve the efficiency of the Arabic speech recognition system. These coefficients are used for deriving new parameters that can be used in a second level of the recognition system

The reflection coefficients that are obtained directly from the LP analysis technique, are used in a direct and simple formula to obtain the area function as follows [9],[10]

$$A_l(P+1) = 1 \quad (1)$$

$$A_l(m) = \frac{1 + k_l(m)}{1 - k_l(m)} A_l(m+1) \quad (2)$$

Where $k_l(m)$ is the m^{th} reflection coefficient of frame l , and $1 \leq m \leq P$.

2.3 Statistical modeling and vector quantization

In this section we consider joint quantization of a block of signal parameters. This type of quantization is called block or Vector Quantization (VQ) [4], [5], [11] that is implemented as a compression technique to reduce the computational complexity. The output of this stage is a set of observation vectors representing the signal. These observation vectors are used in the HMM recognition system. The first step in this process is the initial guess of the codebook. An initial guess

by splitting is implemented on this system [12]. The vector quantization codebook is generated by employing the Linde-Buzo-Gray (LBG) algorithm [4], [5],[12],

3 The HMM recognition system

As vocabularies become larger and recognition tasks more complicated, training and storing models for each word is generally impossible and therefore models for subword units (e.g., syllables, phonemes) are employed [4]. For this reason phoneme models are used in this Arabic speech recognition system as discussed below.

Extraction of phonemes from context is accomplished through hand labeling methods. For this Arabic speech recognition system a state emitter, left to right model is used for each phoneme. For each phoneme HMM a 3-states model is constructed without skipping. The discrete observation HMM is restricted to the production of a finite set of discrete observations. In this case the naturally occurring observation vectors are quantized into one of the permissible sets using the vector quantization (VQ) method described above. Prior to training any of the HMMs for the individual utterance, a set of continuous observation vectors from a large corpus of speech was used to derive the codebook. Subsequently, any observation vector used for either training or recognition was quantized using this codebook.

In the training phase, the Viterbi reestimation algorithm is applied [4]. First, the test utterance is analyzed and vector quantized. Its observation sequence is now the only known parameter. Second, we represent this utterance by a HMM. Its parameters are those obtained for the existing trained models \square_i , $1 \leq i \leq W$, and W is the number of models trained. We use the parameters of each reference model in conjunction with the observation sequence O of the test utterance. This procedure is implemented using the Viterbi decoding algorithm.

In the recognition phase, given a test utterance we seek to model it and to find its P^* , where $P^* = P(O | \square_i)$ is the probability that the HMM could generate the observation sequence O given the model \square_i using the best possible sequence state.

The resulting measures computed for each model, \square_i , will be $P(O | \square_i)$. These measures provide meaningful likelihood measurement with which appropriate relative comparisons across models are made. Finally, the reference model, which provides the maximum probability, is chosen as the recognized one.

4 A second level of the recognizer

In the light of the results obtained, acoustic confusability is encountered between some Arabic phonemes. Confusability refers to the extent to which phonemes can be easily confused because of some acoustic similarities [4]. The extent to which similarities occur in the phonemes will have an impact on the performance

of the recognizer. A higher level of the recognizer is called upon to make the correct recognition. In this second level of the recognition system, additional parameters are presented for resolving the acoustical confusability between some Arabic phonemes. The new parameters are based on a new classification for Arabic phonemes as described in the next section.

4.1 Classification of Arabic phonemes

Based on the shapes of the front, the back and the root of the tongue, Arabic speech can be classified into three classes as follows:

1. Emphatic phonemes.
2. Non-emphatic phonemes.
3. Root phonemes.

In the Arabic system of emphatic vs. non-emphatic phonemes, the emphatics are more backed than their non-emphatic counterpart [2]. An emphatic articulation is one in which the back of the tongue is elevated towards the palate while the front of the tongue is lowered. Therefore, the tongue forms a narrower area at the position of the back of the tongue as compared to that at the position of the front of the tongue. In contrast, for the non-emphatic phonemes such a tongue shape does not occur.

The 7 emphatic phonemes of Arabic speech are /q/ (*Ķa*), /ɣ/ (Gha), /x/ (Kha), /D/ (*Ďa*), /T/ (*Ṭa*), /S/ (*Ṣa*), and /ð/ (Dha) [2]. Non-emphatic phonemes are all other phonemes except the pharyngeal /H/ (*Ĥa*) and /□/ (*Āa*), and the glottal /a/ (Aa), /h/ (Ha), and the /ʔ/ (Hamza) which will be defined here as root phonemes.

Root phonemes are considered to be unusual speech sounds. Not all languages use them and those that do have very few of them. In Arabic speech the 5 root phonemes are the consonants /□/ (*Āa*), /H/ (*Ĥa*), /ʔ/ (Hamza), /h/ (Ha), and the vowel /a/ (Aa). A root articulation is one in which the root of the tongue assumes the shape of a bulge and is drawn back toward the vertical back wall of the pharynx to form a stricture [1], [2]. This radical bulge generally divides the vocal tract into 2 cavities, one below extending from the stricture to the glottis, the other above extending from the stricture to the lips.

4.2 The new parameters

Based on the shape of the tongue, some parameters are derived. and can be used in a higher level of the recognizer to resolve acoustic confusability between the phonemes. The new parameters are as follows:

1. Emphatic/non emphatic parameter.
2. Root parameter.
3. Hamz parameter.

These parameters are presented in the following subsections.

4.2.1 "Emphatic/nonemphatic" parameter

According to the X-ray tracing for the area function by Fant [10], an approximate length of the back of the tongue for a constant vocal tract of length (17-cm) is calculated. The back of the tongue starts approximately at a distance 6 cm and ends at 9 cm from the glottis. The front of the tongue starts at 9 cm and ends at 12 cm from the glottis. Hence, for the lossless tube model the area function of $P=13$ sections, the back of the tongue starts (approximately) at section 5 and ends at section 7, and the front of the tongue starts at section 8 and ends at section 10. The areas between section 5 and 7 are the areas that have smaller values in the case of emphatic articulation [2].

For deriving the Emphatic/nonemphatic parameter, the summation of areas from section 5 to 7 (X) is compared to the summation of areas from section 8 to 10 (Y). It is observed that (X) is smaller than (Y) in case of emphatic articulation. Accordingly, a ratio factor is derived as follows,

$$R_{emp} = X / Y \quad (3)$$

It is found that this ratio is less than one for the emphatic sounds.

4.2.2 "Root" parameter

A new parameter is derived for the root phonemes /□/ ($\dot{A}a$), /H/ ($\dot{H}a$), /?/ (Hamza), /h/ (Ha), and the vowel /a/ (Aa) which are not emphatic. According to X-ray results by Fant [10], the root of the tongue starts at the glottis and ends at 6 cm from the glottis. Hence, for the lossless tube model the area function of $P=13$ sections, the root of the tongue starts (approximately) at section 1 and ends at section 5. This constricted area at the root of the tongue is the feature that can be used to distinguish between root phonemes and other phonemes of Arabic.

For deriving the Root parameter, the summation of areas from section 1 to 5 (C) is compared to the summation of areas from section 6 to 10 (D). It is observed that (C) is smaller than (D) in case of root articulation. Accordingly, a ratio factor is derived as follows,

$$R_{root} = C / D \quad (4)$$

It is found that this ratio is less than one for the root phonemes.

4.2.3 "Hamz" parameter

On the other hand, the root phonemes have some acoustic confusability between them. This confusability could be resolved for the phonemes /h/ and /?/. The summation of areas at sections 1 and 2 are smaller than that at sections 3 and 4

for the pharyngeal /ʔ/. While the reverse is true for the pharyngeal /h/. So, another parameter is derived to distinguish between the phonemes /h/ and /ʔ/. We define the following,

$$R_h = \frac{A(1) + A(2)}{A(3) + A(4)} \quad (5)$$

From this equation we obtain the following two cases:

The case of the phoneme /ʔ/, $R_h < 1$.

The case of the phoneme /h/, $R_h > 1$.

4.3 Algorithm of the second level of the recognizer

The second level of the recognizer runs after the decision step taken in the HMM recognizer (the first level of the recognizer), and its algorithm is implemented as follows,

1. We have the area function of the test utterance at the l^{th} time frame $\{A_l(i)\}$, where $1 \leq i \leq 13$, and $1 \leq i \leq 13$
2. Calculate the average area function of all frames $A_{ave}(i)$
3. Calculate the Emphatic parameter

$$R_{emp} = \frac{\sum_{i=5}^7 A_{ave}(i)}{\sum_{i=8}^{10} A_{ave}(i)} \quad (6)$$

If $R_{emp} < 1 \Rightarrow$ This is an emphatic phoneme,

Else \Rightarrow This is either a non-emphatic or root phoneme.

4. Calculate the Root parameter

$$R_{root} = \frac{\sum_{i=1}^5 A_{ave}(i)}{\sum_{i=6}^{10} A_{ave}(i)} \quad (7)$$

If $R_{root} < 1 \Rightarrow$ This is a root phoneme,

Else \Rightarrow This is a non-emphatic.

5. Calculate the Hamz parameter

If the phoneme is recognized (in the first level) as either /h/ or /?/ then calculate,

$$R_h = \frac{\sum_{i=1}^2 A_{ave}(i)}{\sum_{i=3}^4 A_{ave}(i)} \quad (8)$$

If $R_h < 1 \Rightarrow$ This is the root phoneme /?/,

Else \Rightarrow This is the root phoneme /h/.

6. Compare the previous decisions of phoneme classification with the dictionary of classified phonemes and chose the nearest one to the first decision in the first level recognizer.

Finally, these new parameters are based on the length of the vocal tract, it should be pointed out that accurate estimate of the vocal tract of each speaker is necessary, and adjustment of the mentioned distances of the vocal tract is necessary.

5 System evaluation

The new parameters, which are based on a new classification of Arabic phonemes, are added at a higher level of the recognition system for resolving the acoustic confusability and improving the recognition accuracy. The “Emphatic/nonemphatic” parameter gives a performance of 44%, the “Root” parameter gives a performance of 40%, while the “Hamz” parameter could resolve confusability encountered between the phonemes /h/ and /?/ with a performance of 90%. Accurate data of the vocal tract distances are needed for deriving the new parameters.

Acknowledgment

The authors would like to thank Dr. Ga’afar Mirghani, Department of Art, University of Khartoum, for his valuable comments in the field of Arabic phonetics.

6 References

- [1] A.R. Ayoub, *Arabic Speech Production and Analysis*, Kuwait University, Kuwait, 1984. (Published In Arabic).
- [2] G. Mirghani, *Arabic Speech & Arabic Phonetics*, ALESCO, Khartoum, 1985. (Published in Arabic).

- [3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, N. J.: Prentice Hall, 1978.
- [4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, Inc., 1993.
- [5] J. W. Picone, Signal Modeling Techniques in Speech Recognition, *Proceedings of the IEEE*, Vol. 81, No. 9, Sep. 1993.
- [6] Iman A. Maaly, Abdel R. Elobeid, and Asim A. Sagayroon, On Obtaining Parameters for a Model of Arabic Speech Production, *Proceedings of the IEEE International Conference on Signal Processing Applications and Technology, Dallas, Texas USA*, pp. 1628-1633, Oct. 1994.
- [7] A.R. Elobeid Ahmed, I.A. Maaly, and M.A.H. Abbas, Performance Tests On Several Parametric Representations for an Arabic Phoneme Recognition System Using HMMs, *Proceedings of the Thirteenth International Conference on Applications of Artificial Intelligence in Engineering, AIENG XIII, Wessex Institute of Technology, Galway Ireland*, pp. 45 – 48, July 1998.
- [8] J. Makhoul, Stable and Efficient Lattice Methods for Linear Prediction, *IEEE Transaction on Acoustics, Speech and Signal Processing*, Vol. ASSP 22, No.2, April 1974.
- [9] H. Wakita, Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveform, *IEEE Transaction on Audio and Electroacoustics*, Vol. AU-21, No. 5, OCT. 1973.
- [10] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, Paris, 1970.
- [11] Y. Linde, A. Buzo, and R. M. Gray, An Algorithm for Vector Quantizer Design, *IEEE Transactions on communications*, Vol. 28, pp. 84-95, Jan. 1980.
- [12] G. D. Forney, The Viterbi algorithm, *Proceedings of the IEEE*, Vol. 61, pp. 268-278, Mar. 1973.