

Performance tests on several parametric representations for an Arabic phoneme recognition system using HMMs

A.R.Ellobeid Ahmed,^a L.A. Maaly,^b M.A.H. Abbas^b

*U*Department of Physics, Faculty of Science, University of Qatar,
PO. Box 2713, QATAR

E-mail: aelobeid@qu.edu.qa

*h*Department of Electrical Engineering, University of Khartoum,
PO.Box 321, SUDAN

E-mail: imaaly@hotmail.com

Abstract

The study of speech recognition is part of a quest for "artificially intelligent" machines that can "hear", "understand", and "act upon" spoken information, and "speak" in completing the information exchange. As the objective of a robust, intelligent, fluent, Arabic machine remains a distant goal, we introduce an Arabic speech recognition system using Hidden Markov Models (HMM). It is a phoneme recognition system which is one of the important steps towards a continuous speech recognition system with a large vocabulary size for Arabic.

Signal Modeling of this recognition system is accomplished in four main steps:

- **First**, spectral shaping is accomplished by recording the speech signal at a sample frequency of 10 kHz, and then emphasizing important frequency components in the signal by the pre-emphasis filtering using a first order differentiator.
- **Second**, spectral analysis techniques are implemented after segmenting the signal into frames of duration 20 ms each, and with overlapping of 5 ms. Different sets of parameters are extracted. Most of these parameters are based on the Linear Prediction Coding (LPC) technique. Another parameter (power) is extracted directly from the signal.
- **Third**, parameters transformation is accomplished by the differentiation process to better characterize temporal variations in the signal, and by the filtering process to enhance those portions of the cepstrum representing vocal tract information. The final sets of parameters obtained are the following:

Artificial Intelligence in Engineering

1. Prediction coefficients.
2. Reflection coefficients.
3. LP-derived cepstral coefficients.
4. Liftered cepstral coefficients.
5. Delta cepstral coefficients.
6. Delta-delta cepstral coefficients.
7. Area function parameters.
8. Log area ratio parameters.
9. Power.
10. Delta power.
11. Delta-delta power.

• Fourth, the prewhitening transformation is implemented on some combinations of the above sets of parameters to remove correlation between their elements. Then, Vector Quantization (VQ) 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 speech recognition system adopted, is based on HMMs. A left-to-right HMM with 3 states is constructed for each of the 31 Arabic phonemes. The Viterbi reestimation method is used to estimate model parameters in the training phase of the recognition system. The Viterbi decoding algorithm is implemented for solving the problem of choosing an optimal states sequence corresponding to the observation sequence and the model. In the recognition phase, the probability of the observation sequence, given the model, is computed using the Viterbi decoding algorithm. At the final recognition stage, performance tests are accomplished on different sets of parameters.

This system is constructed using the technique of Object Oriented Programming which makes management of the computer memory easier, and makes code reuse practical.

The results of the performance tests confirm the superiority of the cepstral coefficients representation with recognition performance 74% over the other representations. Combinations of cepstral coefficients and each of the other parameter sets are tested. Results showed that supplementing the cepstral coefficients with delta power and delta-delta power improves the performance to 81 %. These results are evaluated and compared to similar results on the recognition of English vowels.

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

Artificial Intelligence in Engineering

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 raise the efficiency of the Arabic recognition system. But, results showed that the performance of using the area function parameters is 48.4 %. Such a low performance may be due to the fact that the area function derived from the reflection coefficients can not perfectly characterize the nasal sounds and unvoiced fricative sounds.

The recognition performance using the log area ratio parameters is 64.5% which is better than that of the area function. This improvement may be due to the fact that these coefficients have lower inter-parameter correlation.

In this Arabic recognition system, the LP derived formants which are more accurate than those derived from other techniques, e.g., Fast Fourier Transform (FFT), are tested as a parameter set. Formants give performance of 35.5% . This low performance using the formants may be due to the recognized theory that vowels are perceived by static conditions steady-state frequencies; whereas consonants are perceived by dynamic conditions, i.e., rapid shifts of their formant frequencies, known as formant transitions. Since phoneme units used in this recognition system do not include transitional units, these formants do not perfectly characterize consonants.

In this system performance tests are implemented on power and its derivatives as basic parameters. Results show that the use of power, delta power and delta-delta power in one parameter set gives a performance of (22.5%). In this case we mixed power and the time derivatives of the power that have completely different numerical scales, i.e., the range and variance of the power term will be much larger than the range and variance of the time derivatives of power. The result will likely be dominated by the terms with large amplitudes and variances, even though the true information may lie in the smaller amplitude parameters. Consequently, the use of delta power and delta-delta power as a parameter set gives a little improvement of performance (38.7%).

Results showed that supplementing the cepstral coefficients with delta power and delta-delta power gives the best performance for this Arabic speech recognition system (81 %). Small improvement is also obtained by supplementing the reflection coefficients with delta power

Artificial Intelligence in Engineering

and delta-delta power. However, supplementing the log area ratios with delta power and delta-delta power degrades the performance.

Implementation of the liftering process on the cepstral coefficients is suggested to enhance those portions of the cepstrum representing vocal tract information. However, the use of the liftered cepstral coefficients in this recognition system gives a performance of (64.5%) which is less than that of the cepstral coefficients (74.2%).

On the other hand, to better characterize temporal variations in the signal, higher order time derivatives of signal measurements were added to the signal model, i.e., the use of delta cepstral and delta-delta cepstral coefficients are standard in the up-to-date speech recognition systems. However, each of the delta cepstral coefficients and delta-delta cepstral coefficients in this recognition system gives a performance of (61.3%), which are also less than that of the cepstral coefficients.

In the light of the results obtained, some characteristics of Arabic phonemes are observed. A new classification of Arabic phonemes, which is based on the positions of the front, back, and root of the tongue during the articulation of Arabic phonemes, is presented. Three new parameters are derived for resolving the confusability encountered between Arabic phonemes, and for improving the performance of the system. They are used in a second level of the recognizer. Performance results of these parameters are 44% for the "Emphatic/non-emphatic" parameter, 40% for the "root" parameter, and 90% for the "Hamza" parameter. An interesting fact is that confusion between vowels and consonants is very small in all the results. The second important result is the confusion between the glottal and pharyngeal phonemes of Arabic, i.e., sometimes the recognizer can not distinguish between them.