View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Lincoln Institutional Repository

# A Speech Recognition Model Based on Tri-Phones for the Arabic Language

B. Al-Diri, A. Sharieh, M. Qutiashat

Computer Science Department, The University of Jordan
Amman, Jordan
sharieh@ju.edu.jo

**Abstract**

One way to keep up a decent recognition of results- with increasing vocabulary- is the use of base units rather than words. This paper presents a Continuous Speech Large Vocabulary Recognition System-for Arabic, which is based on tri-phones. In order to train and test the system, a dictionary and a 39-dimensional Mel Frequency Cepstrum Coefficient (MFCC) feature vector was computed. The computations involve: Hamming Window, Fourier Transformation, Average Spectral Value (ASV), Logarithm of ASV, Normalized Energy, as well as, the first and second order time derivatives of 13-coefficients. A combination of a Hidden Markov Model and a Neural Network Approach was used in order to model the basic temporal nature of the speech signal. The results obtained by testing the recognizer system with 7841 tri-phones. 13-coefficients indicate accuracy level of 58%. 39-coeefficents indicates 62%. With Cepstrum Mean Normalization, there is an indication of 71%. With these small available data-only 620 sentences-these results are very encouraging.

## 1. Introduction

Automatic Speech Recognition (ASR) is a process upon which speech signals are automatically converted into the corresponding sequence of words in a text. In real life applications, however, speech recognizers are used in adverse environments. The recognition performance is typically degraded-if the training and the testing environments are not the same.

Speech recognition is useful as a form of input: it is useful when someone's hands or eyes are busy; it allows people who are working in an active environment-such as in hospitals-to use computers; it also allows people with handicaps such as: blindness or palsy to use computers. Although everyone knows how to talk, not every one knows how to type. With speech recognition, typing would no longer be a necessary skill for using a computer. If we are successful enough in combining it with the natural understanding of language, it would make computers accessible to people who don't want to learn the technical details of using computers (Deroo, 1998).

In the last years, there were significant developments of large vocabulary speaker independent continuous speech recognition systems. These developments have been mainly done for the English language. However, this success in terms of both research and commercial systems has been pushing the development of this kind of systems for other languages.

With recent advances, speech recognizers based upon Hidden Markov Models (HMMs)-have achieved a high level of performance in controlled environments (Bahl et al. 1995). The HMMs are used in most of the state-of-the-art continuous-speech recognition systems. This approach is limited by the need for strong statistical assumptions that are unlikely to be valid for speech. Training of the phonetic models is based on the Maximum-Likelihood Estimation using the Forward-Backward Algorithm (Levinson et al. 1983). Recognition uses the Viterbi Algorithm (Levinson et al. 1983) in order to find the HMM state sequence (corresponding to a sentence) that has the highest probability of generating the observed acoustic sequence.

The hybrid Multi-Layer Perception (MLP) / HMM system substitutes probability estimates which are computed with MLPs for the tied-mixture HMM state-dependent observation probability densities. The initial hybrid system used an MLP to compute context-independent phonetic probabilities for phone classes in the DECIPHER (TM) system (Murveit, et al. 1989).

The large vocabulary, speaker-independent continuous speech recognition hybrid system for the European Portuguese language combines the advantages of the HMM model and the Neural Network approaches by using MLPs to estimate the state-dependent observation probabilities of The HMM (Neto et al. 1998).

Several authors (Richard & Lippman, 1991; Bourlard & Morgan, 1994) have shown that the outputs of artificial neural networks (ANNs) used in classification mode can be interpreted as estimates of posterior probabilities of output classes-which are conditioned upon the input. It has been proposed to combine ANNs and HMMs into what is now referred to as hybrid HMM/ANN Speech Recognition Systems.

There is a lack of ASR for Arabic, so this research which is based upon the work by Al-Diri (2002), aims to build a large vocabulary continuous speech recognition for the Arabic language. Section 2 describes the procedure in which to build a baseline clean speech recognizer. A dictionary for Arabic phones is described and implemented, and the hybrid HMM/NN will be described. Section 3 presents results of testing the proposed model. Section 4 concludes the paper.

## 2. A Speech Recognizer for Arabic

A speech recognizer was built by using the clean speech corpus ARABIC_DB (Al-Diri & Sharieh, 2000). In this section, the training and testing procedures of forward large vocabulary continuous speech recognizers are described.

### 2.1 The Dictionary Model

The design and implementation of a dictionary for Arabic is one of the challenging problems that were faced. The design of the dictionary must adhere to certain requirements:

1. Most of the phones that could be spooked should be included.
2. A dictionary must allow for fast and accurate search algorithms.
3. The morphological structure of the Arabic Language (AL) must be taken into consideration in the design of a dictionary.
4. The memory requirements must be as small as possible.

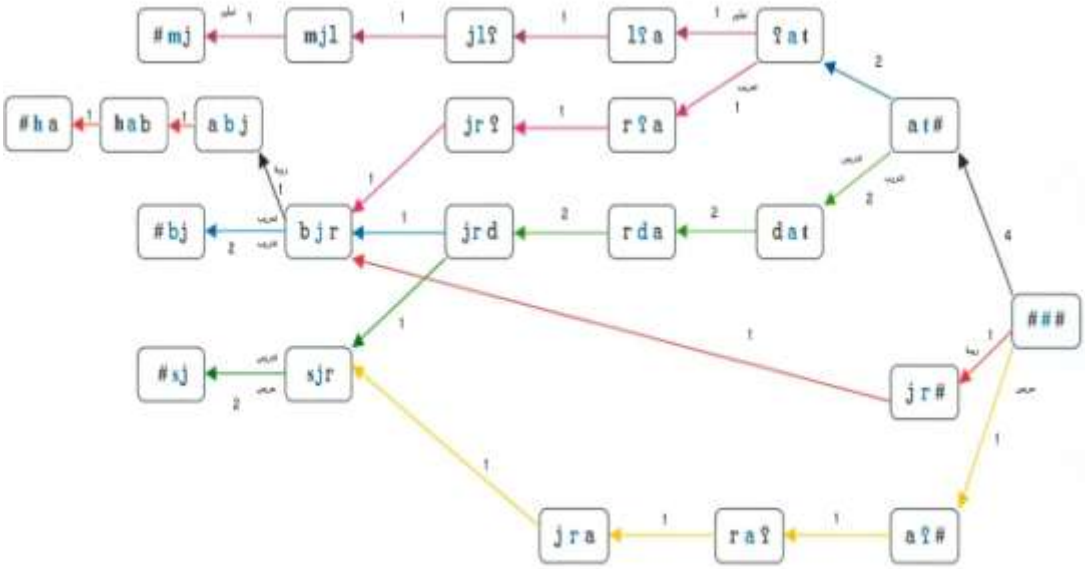


Figure 1: Examples from the dictionary, which has the Arabic words: "تعليم، تدريب، تدريس، عريس، ريبة" and "تعريب" and their spelling correspondents in English are: "Tareeb, Taleem, Tadreeb, Tadreeb, Tadrees, Arees, and Reebah", respectively.

In order to meet these requirements, the dictionary model employed in this research is based upon a phone structure-as illustrated in Figure 1. Each node in the figure represents a unique triphone- each link $L_{ij}$ between two nodes has a number of occurrences of node j after node i. For example, the link from triphone "###" for a triphone "at#" is 4, because the triphone "at#" occurred 4 times after the occurred of "###" triphone. This dictionary structure saves space because: of its characteristics and the morphological structure of the AL. Each Arabic letter has an International Phone Alphabet (IPA) as shown in Table 1. All the triphones in the structure-as shown part of them

in Table 2-must be trained in the training stage. It is not possible to recognize a word that contains a triphone, unless it is trained.

Table 1: A list of Arabic Monophones and their International Phonetic Alphabet.

| IPA | M. | IPA | M. | IPA | M. | IPA | M |
|---|---|---|---|---|---|---|---|
| uː | وْ: | q | ق | z | ز | ʔ | ء |
| u | ـُ | k | ك | s | س | b | ب |
| aː | ا | l | ل | ʃ | ش | t | ت |
| a | ـَ | m | م | sˤ | ص | θ | ث |
| ɑː | اْ: | n | ن | dˤ | ض | dʒ | ج |
| ɑ | ـَ ـْ: | h | ه | tˤ | ط | ħ | ح |
| \| | + | w | و | ðˤ | ظ | x | خ |
| ǂ | * | y | ي | ʕ | ع | d | د |
| | | iː | ي : | ʁ | غ | ð | ذ |
| | | i | ـِ | f | ف | r | ر |

Table 2: Some entries of a list of triphones that were extracted from the ARABIC_DB.

| TriPhone_ID | Right Monophone | Middle Monophone | Left Monophone |
|---|---|---|---|
| 1 | * | * | ء |
| 2 | * | * | ب |
| … | … | … | … |
| 1000 | خ | ـَ- | ش |
| 1001 | خ | ـَ- | ف |
| … | … | … | … |
| 5033 | ـَ-: | ي | ف |

| | | | |
|---|---|---|---|
| 5034 | ◌َ-: | ي | ي |

**2.2 Features Extraction**

The continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M samples (where M < N). The first frame consists of the first N samples. The second frame begins with M samples after the first frame, and overlaps it by N - M samples. Similarly, the third frame begins with 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples. The process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 400 (which is equivalent to ~ 25 msec) and M = 100.

A 39-dimensional Mel Frequency Cepstrum Coefficients (MFCCs) feature vector is computed from 25 ms of a window-with 15 ms overlapping-using the following steps:

1. Re-emphasize and weight the speech signal by a Hamming Window (HW). The HW, w (n), is defined in equation (1). The N, here, is the total number of samples in a given interval.

$$w(n) = 0.45 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \le n \le N-1 \qquad (1)$$

2. Find the Fourier transform of the weighted signals.
3. Compute the average of the spectral magnitude values using a triangular window at uniform spaces on the Mel scale in order to take auditory characteristics into consideration. The Mel scale is defined as in equation (2), where f is the frequency.

$$mel(f) = 2595 * \log_{10}(1 + f/700) \qquad (2)$$

4. Apply the logarithm for the averaged spectral values. The convolution between sound source (pitch) and articulation (vocal tract impulse response) becomes addition due to the logarithm operation.
5. Convert the log (Mel) spectrum back to time, the result is called MFCC. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, you can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore, if you denote those Mel power spectrums coefficients-

resulted from the last step: $S_k$, k=1,2, ..., K, then you can calculate the MFCC's ($C_n$), using equation (3).

$$\tilde{C}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[ n\left( k - \frac{1}{2} \right) \frac{\pi}{K} \right] \qquad n = 1,2,...,K \qquad (3)$$

The first component $C_0$ was excluded from the DCT-since it represents the mean value of the input signal, which carried little speaker specific information.

6. Append the normalized frame energy (Lin & Che 1995), producing a 13-dimensional feature vector.

7. Compute the first and the second order time derivatives of the 13 coefficients using Equation (4). The first and the second derivatives are good features for representing dynamic information. The first leads to fewer coefficients to estimate and more robust estimates. The second is good for initialization of connection weights, leading to better –behaved optimization. The $d_t$ is activity of the delta coefficient unit and $c_{t-x}$ are activities of the original input unit delayed x frames. Equation (4) is equivalent to the delta coefficients of Hidden Markov Model Toolkit (HMT) (young, 1995).

$$d_t = \frac{1}{10}(2c_{t+2} + c_{t+1} - c_{t-1} - 2c_{t-2}) \qquad (4)$$

The MFCC feature extraction is applied to the speech signal. From this preprocessing phase, a frame results with the log energy, 12 MFCC and their first and second derivatives. Therefore, the feature vector has a total of 39 coefficients.

**2.3 Training Speech Recognizer**

Thirty-six phone models and two silence models were trained using the 620 sentences in the ARABIC_DB. Each phone model is a 3-state left to right HMM as in Figure 2. There are two silence models, one represents a long silence or noise if it exists-usually at the beginning and at the end of utterances. A 3-state left to right HMM is used for the long silence. The other silence "-", which represents a short pause between words-uses, is a 3state left to right HMM with a possible skips transition depending on the Arabic specialists.

The training method is as follows:

1 Build a network of alternative pronunciations for each utterance.

2 Construct the HMM model topology of each triphone.

3 Guess the initial set of model parameters for the HMM.

4 Improve the HMM.

5 Save individual HM models for each triphone of the phones separately.

In this way, the words that are not trained can be recognized. For instance, once the word “طائرة” -which is spelled as: "taerah"- is trained, there will be six HMM models for the triphones: [□ τ0α⌉], [τ0α⌉?], [α⌉?ι], [?ιρ], [ιρη]and [ρη□ ], correspondents to t, a, e, r, a, and h, respectively. In the recognition stage, you can use the models for the triphones [□ τ0α⌉], [τ0α⌉?], [α⌉?ι] ανδ [?ιρ]in order to construct the HMM of the word “طائر” (the triphones for “طائر” are [□ τ0α⌉], [τ0α⌉?], [α⌉?ι], [?ιρ]and [ιρ□ ] assuming that the triphone model for [ιρ□ ]is also trained. Therefore, the word “طائر”, with English spelling as "taer", does not need to be trained.

A training algorithm must start with an initial guess. The model parameters are A and π, where A is the state transition probability distribution and π is the initial state distribution. An initial guess of the state transition probability distribution is given in Figure 3.

The triphones that have [□] at the rightmost phoneme position have only one state transition for the rightmost state. This transition goes from the rightmost state to itself-because there isn’t any other state on the right. Since the sum of the outgoing transition probabilities for a state must be 1, the probability of taking this transition is 1.

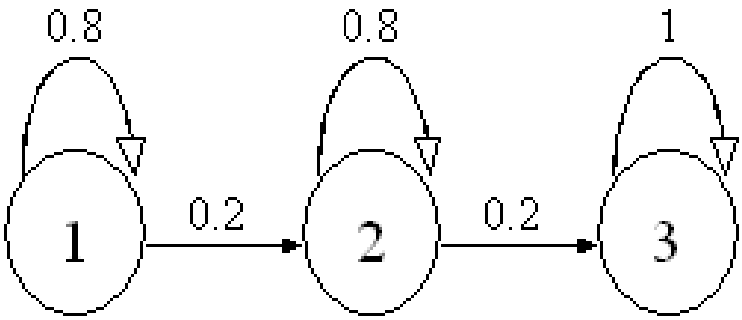


Figure 3: Initial estimate of the state transition probability distribution, where $\pi_1 = 1$ and $\pi_2 = \pi_3 = 0$.

In general, the length of the same phoneme and a different one depend on their position. But, it is very difficult to cover this problem because there isn’t any definite mathematical relation between the phones and their position.

Also, the length of vowels and semivowels are longer than other phones. Yilmaz (1999) used isolated word recognition, and also assigned double feature vectors to the vowels and

semivowels in the middle of word than other phones. In this research, the vowels and semivowels are assigned double feature vectors than any other phones in any position.

**2.4 Evaluation of the Adapting Neural Networks**

The Neural Network (NN) which is based upon the feature transformation approach-that uses Mean Square Error (MSE)-was evaluated. The convolution and additive noise becomes non-linear in the cepstral domain. Some procedures such as: Time Derivative Computation and cepstral mean normalization (CMN) operation make it difficult to handle the noise in a linear frequency domain. Neural networks are popular mathematical tools that can model arbitrary non-linear functions without any expert knowledge (Yuk, 1999). The feature transformation neural network converts distorted speech feature vectors, as well as speaker independent, to those that correspond to clean speech and speaker dependent. The optimal architecture of the NN is decided empirically after a series of experiments.

The input and the output of the NN should be fixed for all kinds of phones. So, the input and output will be one vector of 39 coefficients-unless we define otherwise.

Additional noise causes non-linear distortion in the cepstral domain. A feature transformation neural network that uses stereo data and MSE as its objective function has been used to handle the non-linear distortion. The two-hidden-layer neural networks can represent what an one-hidden-layer network can represent. When hidden nodes are used, the derivatives information plays a positive role. This is because the derivatives affect the weights of hidden layers. The two-hidden-layer networks, which incorporate time derivatives, are used for the current experiments-unless stated otherwise.

**2.5 Hybrid HMM/NNS**

In this work, a hybrid system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with Artificial Neural Network (ANN) as feature transformation model is employed. In the HMM/NN system, a HMM process is used to model the basic temporal nature of the speech signal. The NN is used as the feature transformation model as a robust adaptation technique within the HMM framework.

**2.6 Implementation issues**

A lot of functions were written to deal with: Database's Files, utterance's statements and preparation of data for training and testing ever for HMM or NN. A group of functions to create and deal with a lexical of an AL model were implemented in Matlab.

Some of HMM's functions were found through the Internet (Murphy, 1998; Cappe, 2001). Some were re-implemented to fit the need of this research-some were fully implemented. The set of functions were implemented in Mat lab tool and used for training and testing the NN.

**3. Experiments and Results**

The development of a robust large vocabulary continuous speech recognition system needs a large vocabulary speech corpus for the Arabic language. There is not a known, well-established database for the AL. However, the database in (Al-Diri & Sharieh, 2002) was used. This database has 5034 unique tri-phones, which are driven from 750 sentences that are spoken uttered by three males.

**3.1 Baseline Performance**

The speech recognition system described in Section 2 is evaluated. Table 3 shows the phone recognition accuracies of the recognizer for the six speakers: A, B, C, D, E and F. The recognizer was trained with 620 statements for the speaker A. The number of coefficients without using Cepstral Mean Normalization (CMN) was thirteen coefficients. The recognizer was tested with 171 statements (7841 triphones) for each speaker.

Table 3: The accuracy of the base line for the six speakers, using 13 coefficients without CMN.

| Speaker (S) | Gender M/F | Accuracy % Level 1 | Accuracy % Level 2 | Accuracy % Level 3 | Accuracy % |
|---|---|---|---|---|---|
| A | F | 34.69 | 13.68 | 9.37 | 57.75 |
| B | M | 33.16 | 13.23 | 8.63 | 55.02 |
| C | F | 34.50 | 13.76 | 9.96 | 58.22 |
| D | M | 28.85 | 9.69 | 6.54 | 45.08 |
| E | M | 29.37 | 10.51 | 6.84 | 46.72 |
| F | F | 33.52 | 14.07 | 9.37 | 56.96 |

Level 1 is the maximum percentage of the correct tri-phones. Level 2 and Level 3 are the preceding percentage of the correct triphones of Level 1. The accuracy is computed by summing up the accuracies of the three levels. The accuracy of the recognizer is better for the female gender than the male because the female person trained the recognizer. The maximum accuracy here is 58.22%.

Table 4 shows the phone recognition accuracies of the recognizer for the speakers-using 39 coefficients with both CMN and without CMN. The accuracy of using 39 coefficients is better than using 13 coefficients. The maximum accuracy without CMN is 62.13%; with CMN it is 70.88%.

Table 5 shows the accuracy of using 50 statements with 1977 triphones. The accuracy of recognition in Table 5 is less than those in Table 4. This is because the amount of the testing triphones were less and different. The accuracy is best for the speaker who trained the recognizer.

Table 4: The accuracy of the base line for the six speakers, with 39 coefficients without and with using CMN.

| | Accuracy % Level 1 | | Accuracy % Level 2 | | Accuracy % Level 3 | | Accuracy % | |
|---|---|---|---|---|---|---|---|---|
| Speaker | Without | With | Without | With | Without | With | Without | With |
| A | 35.20 | 41.18 | 15.89 | 17.80 | 10.74 | 11.90 | 61.83 | 70.88 |
| B | 35.44 | 35.66 | 15.23 | 15.06 | 10.48 | 10.51 | 61.15 | 61.23 |
| C | 35.59 | 35.65 | 15.65 | 15.75 | 10.89 | 10.74 | 62.13 | 62.13 |
| D | 32.23 | 32.56 | 11.02 | 11.73 | 7.96 | 8.63 | 51.21 | 52.93 |
| E | 34.01 | 34.51 | 14.12 | 14.02 | 9.27 | 9.27 | 57.40 | 58.09 |
| F | 35.30 | 35.48 | 15.16 | 15.34 | 10.98 | 10.83 | 61.45 | 61.65 |

Table 5: The accuracy of the baseline for the six speakers, 39 coefficients with using CMN and 50 statements.

| Speaker | Gender | Accuracy % Level 1 | Accuracy % Level 2 | Accuracy % Level 3 | Accuracy % |
|---|---|---|---|---|---|
| A | F | 39.91 | 17.53 | 11.84 | 69.28 |
| B | M | 34.28 | 13.58 | 9.77 | 57.63 |
| C | F | 34.47 | 14.87 | 10.10 | 59.44 |

| | | | | | |
|---|---|---|---|---|---|
| D | M | 31.51 | 12.27 | 7.26 | 51.04 |
| E | M | 32.88 | 13.45 | 7.91 | 54.24 |
| F | F | 33.18 | 14.63 | 10.37 | 58.18 |

The same recognizer was tested; using the same statements after the NN transformation was evaluated. The gain ($G_I$) of the NN can be calculated by formula (5)-where H is the accuracy of a speaker after the NN features were transformed. S is the accuracy of that speaker. B is the accuracy of the speaker who trained the recognizer. Thus, the NN based transformation methods improve the performance by doing feature transformation.

$$G_I = \frac{H - S}{B - S} \qquad (5)$$

Table 6: Gains that were obtained after features transformed using NN for independent speakers.

| Speaker | Gender | Accuracy % Level1 | Accuracy % Level2 | Accuracy % Level3 | Accuracy % | Gain % GI |
|---|---|---|---|---|---|---|
| A | F | 39.52 | 16.96 | 11.58 | 68.06 | - |
| B | M | 37.58 | 15.88 | 11.19 | 64.65 | 67.31 |
| C | F | 38.03 | 16.45 | 10.65 | 65.13 | 66.01 |
| D | M | 36.68 | 16.32 | 9.23 | 62.23 | 65.75 |
| E | M | 35.59 | 15.59 | 10.41 | 62.59 | 60.42 |
| F | F | 37.42 | 16.21 | 11.28 | 64.91 | 68.12 |

Table 6 shows the gains of the feature transformation for different speakers. The minimum gain was 60.42%, the maximum gain was 68.12% and the average gain was 65.52%.

**3.2 Feature Transformation for Adverse Environment**

In this section, the NN based transformation methods have been evaluated on a large vocabulary continuous speech recognition task in an independent and noisy reverberant enclosure. The combined NN further improves the performance by doing feature transformation.

The adverse environments are simulated through two sets of multiple distortion databases. The recognizer was tested using 40 statements (1977 triphones) against the two sets. Table 6 shows the accuracy of sets 1 and 2 before using NN.

The gain ($G_n$) of the NN can be calculated by formula (6), where Y is the accuracy of that set after the NN feature was transformed. T is the accuracy of that set.

$$G_n = \frac{Y - T}{B - T} \qquad (6)$$

Table 7 shows the gain of the feature transformation for the two sets. The minimum gain was 55.14%, the maximum gain was 57.5%, and the average gain was 56.32%.

The recognition of a phone is very sensitive. As a result, you will catch it or you will fail. The recognition of a word can be achieved by the maximum probability of its phones. So recognizing the correct word can be achieved without recognizing all phones of that word. For example, if we have a word (بي), which consists of two phones (ب and ي) and we recognize the phone (ب)- then what is the probability of recognizing the word (بي)? If we try to expect other phone by experiments, then no other than (ي) can produce (ب) a word. Thus if you recognize the phones by 50%, then some of the words can be recognized by 100%

Table 7: The accuracy of the baseline for the two sets of the distortion database with six speakers, 39 coefficients with CMN tested using 50 statements.

| Set | Accuracy % Level 1 | Accuracy % Level 2 | Accuracy % Level 3 | Accuracy |
|---|---|---|---|---|
| A | 39.62 | 16.92 | 11.54 | 68.08 |
| Set 1 | 10.32 | 4.87 | 3.68 | 18.87 |
| Set 2 | 9.89 | 4.12 | 3.29 | 17.30 |

Table 8: Gains that were obtained after the features were transformed using NN for the two sets of multiple distortion databases.

| Set | Accuracy % Level 1 | Accuracy % Level 2 | Accuracy % Level 3 | Accuracy % | Gain $G_n$ |
|---|---|---|---|---|---|
| A | 39.23 | 16.26 | 11.19 | 66.68 | - |
| Set 1 | 26.45 | 12.39 | 7.52 | 46.36 | 57.50 |
| Set 2 | 25.39 | 12.23 | 6.91 | 44.53 | 55.14 |

Let Reco_Phone denotes the number of phones in a word that can recognize it. Let Ratio_Phone denotes the ratio of Reco_Phone over the number of the word's phones. If the

accuracy of recognizing a phone is Phone_Acc, then the accuracy of recognizing a word (word_Acc) is computed by Equation (7).

$$Word\_Acc = \frac{Phone\_Acc}{Ratio\_Phone} \qquad (7)$$

For example, if the Phone_Acc 40% and Ratio_Phone =0.5, then

$$Word\_Acc = \frac{40\%}{0.5} = 80\%$$

The worst case is when we need each of the word's phones to recognize it. Then, word Accuracy can be computed according to Equation (7).

$$Word\_Acc = \frac{40\%}{1} = 40\%$$

The constant phones in a word are sufficient most of the time in order to recognize it. From the ARABIC_DB the ratio of constants in statements equals 56% of phones. So, in general recognizing a word based on the triphones will be faithful and more preferable than recognizing a triphone.

## 4. Conclusion and Future Work

In this paper, automatic speech recognition methods that are robust to the environmental mismatches are explored and applied on the Arabic language. The proposed model has been applied to large speech recognition and evaluated under various adverse acoustical environments. It has also been applied to unsupervised speaker adaptation. Because it requires only a small amount of training data, the proposed approach is cost-effective while, at the same time it is expensive to collect data in a new environment. Therefore, it permits the recognizer-which has been trained once on clean close talking speech-to be used in a wide variety of less favorable environments.

A neural network based transformation approach for robust speech recognition has been proposed, developed, and evaluated. The major advantages of this are: it does not require retraining of the speech recognizer, it automatically learns the mapping function between the training and testing environments, it is able to handle nonlinear distortions, and it is able to handle different speakers with different genders.

It was found that the amount of the text and the database were insufficient for a robust estimation of language models for Arabic. Thus, the database must be enlarged, and the dictionary of a speech recognition system which contains the phones that can be recognized by the system-need to be worked on.

## 5. References


1. Al-Diri, B. and A. Sharieh (2002). A Database for Arabic Speech Recognition ARABIC_DB, Technical Report, The University of Jordan, Amman, Jordan.
2. Al-Diri, B. (2002). A large Vocabulary Speech Recognition for ARABIC. Master Thesis, Computer Science Dept., The University of Jordan, Amman, Jordan.
3. Bahl, L. Balakrishnan, S. Bellegarda, J. Franz, M. Gopalakrishnan, P. Nahamoo, D. Novak, M. Padmanabhan, M. Picheny, M. & Roukos. S. (1995). Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task. In IEEE International Conference on Acoustics, Speech, and Signal Processing, 1:41-44.
4. Bourlard, H. & Morgan, N. (1994). Connectionist Speech Recognition - A Hybrid Approach. Kluwer Academic Publishers.
5. Cappé, O. (2001). A set of MATLAB/OCTAVE Functions for the EM Estimation of Mixtures and Hidden Markov Models. http://tsi.enst.fr/~cappe/h2m/.
6. Deroo, O. (1998). Modèle Dépendant du Contexte et Fusion de Données Appliqués à la Reconnaissance de la Parole par Modèle Hybride HMM/MLP. Ph.D. thesis, Facult'e Polytechnique de Mons Laboratoire TCTS, Mons, Belgium.
7. Levinson, S. Rabiner, L. & Sondhi, M. (1983). An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. Bell Syst. Tech. Journal 62, 1035-1074.
8. Lin, Q. & Che, C. (1995). Normalizing the Vocal Tract Length for Speaker Independent Speech Recognition. IEEE Signal Processing Letters, 2(11), 201-203.
9. Murveit, H. Cohen, M. Price, P. Baldwin, G Weintraub, M. & Bernstein, J. (1989). SRI's DECIPHER System. In DARPA Speech and Natural Language Workshop, 238-242, Philadelphia, USA.
10. Murphy, K. (1998). Hidden Markov Model (HMM) Toolbox. www.cs.berkeley.edu.
11. Neto, J. Martins, C. & Almeida, L. (1998). A Large Vocabulary Continuous Speech Recognition Hybrid System for the Portuguese Language. In Proceedings ICSLP 98, Sydney, Australia.
12. Richard, D & Lippman, R. (1991). Neural Network Classifiers Estimate Bayesian Posteriori Probabilities. Neural Computation, No 3, pp 461-483.
13. Yilmaz, C. (1999). A Large Vocabulary Speech Recognition System For Turkish. Master, thesis, Department of Computer Engineering and Information Science, Bilkent University, Turkey.

14. Yuk, D. (1999). Robust Speech Recognition using Neural Networks and Hidden Markov Models–Adaptations using Non-linear Transformations. Ph.D. thesis, Department of Computer Science, The University of New Brunswick, USA.