Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation

Dr Hebah H. O. Nasereddin
Faculty of Information Technology
Middle East University (MEU)
Amman, Jordan
hnasereddin@meu.edu.jo

Ayoub Abdel Rahman Omari
Web Developer, iHorizons
Middle East University (MEU)
Amman, Jordan
Omari.ayoub90@gmail.com

***Abstract*—Speech processing is considered to be one of the most important application area of digital signal processing. Speech recognition and translation systems have consisted into two main systems, the first system represents an ASR system that contains two levels which are level one the feature extraction level As well as, level two the classification technique level using Data Time Wrapping (DTW), Hidden Markov Model (HMM), and Dynamic Bayesian Network (DBN). The second system is the Machine Translation (MT) system that mainly can be achieved by using three approaches which are (A) the statistical-based approach, (B) rule -approach, and (C) hybrid-based approach. In this study, we made a comparative study between classification techniques from ASR point of view, as well as, the translation approaches from MT point of view. The recognition rate was used in the ASR level and the error rate was used to evaluate the accuracy of the translated sentences. Furthermore, we classified the sample text audio files into four categories which were news, conversational, scientific phrases, and control categories.**

Keywords—AF (Acoustic Feature); AM (Acoustic Model); ANN (Artificial Neural Network); ASR (Automatic Speech Recognition); BN (Bayesian Network); CNN (Convolutional Neural Networks); CVC (Consonant-Vowel-Consonant); DAG (Direct Acyclic Graph); DBN (Dynamic Bayesian Networks); DCT (Discrete Cosine Transformation); DNN (Deep Neural Network); DTW (Dynamic Time wrapping); HMM (Hidden Markov Model); JPD (Joint Probability Distribution); LM (Language Model); LPCC (Linear Predictive Cepstral Coefficients); MFCC (Mel Frequency Cepstral Coefficients); MT (Machine Translation); PD (Punctuation Dictionary); PLP (Perceptual Linear Prediction Coefficient); QV (Quantization Vector); RV (Random Variables); SLT (Spoken Language Translation); SVM (Support Vector Machine); WER (Word Error Rate); WRR (Word Recognition Rate)

I. INTRODUCTION

The most important tool for the interaction between the human being is the speech. Thus, using speech human beings can easily communicate, explain, and investigate their ideas in different fields of life. Hence, human beings would like to interact with computers via speech, rather than using primitive interfaces such as keyboards and pointing devices (Vimalaand Radha V., 2012). Therefore, achieving the interaction between the human beings and the computer could be established using Automatic Speech Recognition (ASR) systems. Several computer applications employing speech recognition functions such as electronic dictionaries, customer call centers in communication organizations, the modern generation of automobiles, or the smart house security detecting and authorization processes or more.

The main aim of ASR systems is the transcription of human speech into spoken words. It is a very challenging task because human speech signals are highly variable due to various speaker attributes, different speaking styles, uncertain environmental noises, and so on (Abdel-Hamid, O. Abdel-Hamid, O., Mohamed, A., Jiang, H., Peng, L., Penn, G., and Yu, D., 2014). Thus, the key components of ASR systems namely the Acoustic Feature (AF), the Language Model (LM), the Pronunciation Dictionary (PD), the Acoustic Model (AM), and the decoder.

Any Automatic Speech Recognition ASR contains the following main components:

a) The Acoustic Feature (AF)

The raw data in this level is a raw audio signal that is transmitted from the microphone which needs to be converted into a manageable form in order to make it capable to deal with speech recognition tasks. The input audio signal is converted into a series of sequential frames that are divided based on a specific time interval. Thus, the redundant data is eliminated in this level in order to obtain the representative vector for each frame in the signal.

b) The Language Model (LM)

This component is responsible on describing the combination of words in the target audio signal.

c) Pronunciation Dictionary (PD)

This component is represented as a container of the words and its pronunciation of the source language, as well as, a set phonemes are used for the acoustic models. Furthermore, multiple entries can appear for a word depending on the pronunciation that called homonyms.

d) The Acoustic Model (AM)

This component contains a data describing the acoustic nature of all phonemes that was understood in the system.

978-1-5090-5443-5/17/$31.00 ©2017 IEEE

Thus, the AM usually specific for one language and could be adjusted for a particular language accent. Furthermore, each context dependent phoneme called triphone. One challenge in this level is that the phonemes are context dependent so it important to tend to sound different based on the next and previous.

e) The Decoder

The most important component on the ASR systems and it was represented as the reason behind the ASR system. For each audio frame there is a process of pattern matching. Hence, the decoder evaluates the received feature against all other patterns. The best match can be achieved when more frames are processed or when the language model is considered.

The classification methods for ASR system's implementation

Many modern ASR systems recognizing and classifying methods should be built using one of the following methods which are (Cutajar, M., et al., 2013):

- Hidden Markov Model (HMM).
- Dynamic Time wrapping (DTW).
- Dynamic Bayesian Networks (DBN).
- Support Vector Machine (SVM).

In order to cover some ASR sub tasks such as acoustic modeling / language modeling two basic techniques can be used which are (Cutajar, M., et al., 2013):

- Artificial Neural Network (ANN).
- Deep Neural Network (DNN).

From HMM point of view, the reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. Using HMMs representing a complete words can be easily constructed (using the pronunciation Dictionary) from HMMs and word sequence probabilities added and complete network searched for best path corresponding to the optimal word sequence. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model and modeling the short-term spectra associated with each state with the mixtures of multivariate Gaussian distributions (the state output transition probabilities and the means, variances and mixture weights that characterize the state output distributions). For each word or phoneme, will have a different output distribution; a HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes (Gemmeke, J. et al., 2013).

From DTW point of view, it is an algorithm for measuring the similarity between two sequences which may vary in time or speed. Generally, it is a method that allow computer to find an optimal match between two given sequences (i.e. the sequence are wrapped none directly to match each other) (Vamila C., and Radha, V., 2012).

From DBN point of view, A Bayesian Network (BN) is a way of representing the conditional independence properties of set of Random Variables (RV). Thus, the independences are encoded via missing edges in the graph. in order to clarify the idea DBN and BN is consisting of an indefinite number of frames contains two variables which are from state and the observation, and two sedges (state to the observation, from state in the previous frame to the current state) (Livescu, K, Bilmes, J., and Glass, J., 2003).

Problem Statement

Automatic Speech Recognition (ASR) systems provide an efficient way to extract the spoken text from speech signals by implementing several feature extraction approaches, as well as, employing different types of classification methods. Finding the best feature extraction approach and classification method – with regard to the translated speech- that suits Machine Translation (MT) systems is a challenge. Different approaches were referenced to implement robust MT systems that were varied from statistical approaches to rule based approaches; this represent a challenge in selecting the needed MT system's approach. The main focus in this study was to find the ASR classification technique that suits MT approach that achieves the most accurate translation of a specific type of speech. This study concentrated on finding the effects of classification methods and MT system's approaches on the translated speech.

II. THEORITICAL BACKGROUND AND RELATED WORK

In this section, the author provides a theoretical background about using the ASR classification techniques. Therefore, DTW, HMM, and DBN techniques were discussed in details.

A. Dynamic Time Wrapping (DTW)

Any two time series can be varied in time and speed which is called wrapping points. Thus, data time wrapping technique is one of the most used feature matching (i.e. classification) techniques. Consequently, this technique is used to find the optimal alignment between two time series, as well as, measuring the similarity between those time series (Zhang et al., 2013).

The DTW is employing linear time wrapping by comparing signals of two time series based on linear mapping of the two temporal dimensions (Chapaneri, 2012). Thus, DTW allow non-linear alignment of one signal to another by minimizing the distance between two signals. Therefore, this wrapping can be used to extract figure recognition based on the similarity and dissimilarity between those signals. From speech signals point of view, the duration of each spoken word or digit can vary but the overall speech waveform are similar for same word or digit. Therefore, by applying the DTW technique the corresponding regions between the two time series can be extracted easily to be used in matching processes (Muda et al., 2010).

In more details, Chapaneri, S. measured the optimal wrap path for a given two time series A and B where the length of A is |A| and the length of B is |B|. Thus, A = A1, A2, A3… A|A|,

978-1-5090-5443-5/17/$31.00 ©2017 IEEE

and B = A = B1, B2, B3… B|B|, a wrap can be constructed W = W1, W2, W3 … W k. Where k is the length of wrap path for a kth element that is max (|A|, |B|) ≤ k ≤ |A| + |B|, and W k = (i, j) where i is the length of time series A, and j is the length of time series B. Consequently, the optimal wrap path that represents the minimum distance between two time series can be calculated using (equation 1) (Chapaneri, 2012).

$$Dist\left(w\right)=\sum_{k=1}^{|k|}\frac{Dist\left(w_{ki},\mathrm{w}_{kj}\right)}{(|A|+|\mathrm{B}|)} \qquad (1)$$

In our research, the two time series corresponds to the two number of coefficient features from MFCC phase. Thus, each one of these time series was represented as a vector from different speech signal with two dimensional cost matrix for each feature vector Ai and Bj. The spoken feature vector was compared to template feature vector using DTW (i.e. in case of using DTW classification technique) and the one of the minimum distance is chosen as a recognition output.

B. Hidden Markov Model (HMM) technique

The core idea in using HMM for speech recognition applications is to create a stochastic models from known utterances and compares it with the unknown utterances was generated by speaker. An HMM M is defined by a set of states N that have K observation symbols as well as, three possibility metrics for each state which are in (equation 2) (Ghahramani, 2001)..

$$M=\left\{\Pi,A,B\right\} \qquad (2)$$

Where:

- ∏: initial state probability.
- A: at,j state transition probability.
- B: bt, j, k symbol emission probabilities.

For each HMM system, it could be use three different types of topologies to employ Markov chain which are ergodic model, general left to right model, and linear model. Figure 1 illustrates HMM topologies for a system with four states. Consequently, each state has its own probability which leads to compute the probability for an occurrence of state in a given situation of another state using Bayesian rule.

In this context, for any system employs HMM technique three basic algorithms which are classification, training, and evaluation algorithms. In classification algorithm, the recognition process is enabled for any unknown utterance by identifying the unknown observations sequence via choosing the most likely class to have produced the observation sequence. In training algorithm, the model is responsible to store data collected for a specific language (i.e. in our research the language was the English language). In the evaluation algorithm, the probability of an observation sequence is computed for matching processes.

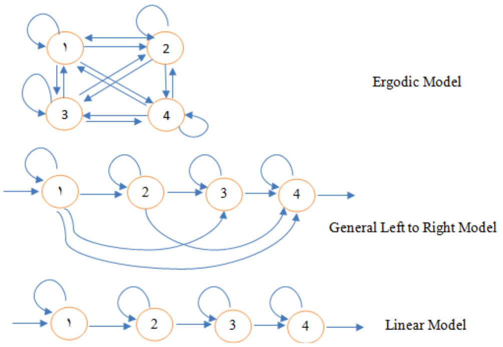

Fig. 1. HMM three states topologies for a system with four states 1, 2, 3, and 4 (Paul, 1990)

The classification algorithm was employed for a given observations O = O1, O2, O3 … OT. A chosen class was computed using (equation 3) (Paul, 1990).

$$\mathrm{Chosen\_Class}=\arg MAX\left[P\left(M_{class}\,|O\right)\right] \qquad (3)$$

Therefore, by applying Bayesian rule to find the probability was computed using (equation 4) (Paul, 1990).

$$P\left(M_{class}\left|O\right.\right)=\frac{P(\mathrm{O}|M_{class})\,\mathrm{P}(\mathrm{M}_{class})}{P(O)} \qquad (4)$$

In this research, we trained the proposed system by several records. Thus, each record had a several tokens for each word in the vocabulary in order to process them with a front end to create the observation sequences based on each token. Furthermore, each training data was identified by word label (i.e. word spelling). In the recognition process, a probability of each word was calculated in order to match the occurrence of specific word with another one in the vocabulary table. In contrast, for an unknown observation sequence the same processes were implemented in order to pass them to HMM.

C. Dynamic Bayesian Networks (DBN) technique

Several research works described DBN technique as the general and flexible model because of its capability in representing complex temporal stochastic processes (Franklen et al., 2007). Thus, this technique also called dynamic probabilistic networks. Furthermore, DBN technique include directed edges pointing in the direction of time that provide a computation of Joint Probability Distribution (JPD) among random variables. In contrast to HMM, a DBN technique allow each speech frame to be associated with an arbitrary set of random variable (Garg and Rehg, 2011).

The Bayesian network is defined by a graphical model structure M and a family of conditional distribution F and their parameters O. The model structure M consists of a set of nodes N and a set of direct edges E connecting the nodes which results a Direct Acyclic Graph (DAG). Consequently, the nodes represents the random variable in the network as well as, the edges encodes a set of conditional dependencies.

Therefore, in Bayesian network the direction of arrow is important between nodes. For instance if the arrow direction from node 1 to node 2 that means that node 1 influence node 2.

In this context, Stephenson, T. et al. concluded the Bayesian network for a given set of random variables that were denoted by X={X1, X2 …, Xn} that correspond to the nodes V in the network. And the values of corresponding variables were denoted by x = {x1, x2 …, xn}. Hence, the joint distribution over a random variable x as in (equation 5) (Stephenson et al., 2000).

$$P\left(x_1, x_2 ..., x_n\right) = \prod_{i=1}^{n} P\left(x_i Pa\left(X_i\right)\right) \qquad (5)$$

A DBN technique is a temporal extension of Bayesian network technique. Figure 3.10 illustrates the DBN model for recognition an isolated word (Arab) that was pronounced as (/'æ/ - /r/ - /ə/ - /b/). Therefore, we employed the following calculations to find the joint distribution of a finite length time series for speech signals. Thus, let X[t] = {X1[t], X2[t] …, Xn [t]} to denote to the random variable in X at a time t {1, 2, 3…}. For all t > 1 and for all values of X [1], X [2] …, X [t] the joint distribution was computed based on (equation 6) (Stephensonet al., 2000).

$$P\left(\mathrm{X}[1], \mathrm{X}[2] ..., \mathrm{X}[\mathrm{t}]\right) = P\left(X\,[1]\right) \prod_{t=2}^{T} P\left(x\,[t]\, x\,[t-1]\right) \qquad (6)$$

In figure 2 the gray vertices (articulators) are observed in training (i.e. when available but not in normal recognition. Hence, the acoustic and final position as well as the transition variables were always observed.

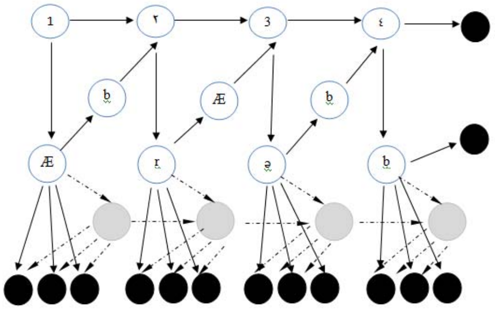

Fig. 2. A dynamic Bayesian network for isolated word (Arab) recognition covering five time steps

D. Feature Extraction phase using MFCC

By using the technique in feature extraction process a fingerprint is created of sound files (Kurzekaret. al, 2014). Hence, MFCC technique is capable to capture the important properties of audio signals in term of time and frequency (Hasan et. al, 2004). Therefore, to employ MFCC feature extraction technique several steps should be implemented. Hence, MFCC can be implemented using six primary steps which are preprocessing, framing, hamming windowing, Fast Fourier Transform (FFT), Mel bank filtering, and Discrete Cosine Transformation (DCT) steps.

In preprocessing step, the speech input was recorded at sampling rate of 22050 Hz in order to minimize the effects of aliasing because of the conversion from analogue to digital form. Furthermore, in this step the energy of speech signal was increased in order to emphasize a higher frequency. Hence, the output signal from this step (Kurzekar et. al, 2014).

In framing step, the speech signal was segmented into small frames (n) of the length which was varied between 20 to 40 msec in order to pass these frames to the hamming windowing step. Consequently, hamming windowing was responsible to create a window shape by considering the next block of feature extraction processing chain as well as integrating all the closest frequency lines. Thus, hamming windows was computed based on equation (7) and (equation 8) (Kurzekaret. al, 2014).

$$Y\,[n] = X\,[n] * W\,[n] \qquad (7)$$

$$W\,[n] = 0.54 - 0.46 * cos\left(\frac{2\pi n}{N-1}\right) \qquad (8)$$

In FFT step the frame was converted of N samples from time domain into frequency domain to preserve the convolution of glottal pulse and vocal tract impulse response in the time domain. Therefore, the computation in this step was conducted based on equation (9) (Kurzekar et. al, 2014).

$$Y\,[w] = \mathrm{FFT}(h\,[t] * x\,[t]) \qquad (9)$$

Based on the results of FFT step the spectrum frequencies were very wide as well as, the voice signal does not follow the linear scale. Therefore, the Mel filter bank was used to ease the conversion to get a Mel frequency signal that is appropriate for human hearing and perception. The Mel frequency was computed in this step based on (equation 10) (Kurzekaret. al, 2014).

$$F\left(Mel\right) = \left[2595 * \log_{10}\left[\frac{1+f}{700}\right]\right] \qquad (10)$$

III. THR PROPOSED METHOD ARCHITECTURE

The main theme of this research is to find a suitable ASR system that suits MT system for real time speech translation from English to Arabic. Thus, to meet this aim we ran several experiments to cover several environments and techniques to calculate the accuracy of the recognized and translated sentences in each combination. Therefore, we examined nine environmental combinations to cover feature extraction using MFCC with classification approaches (i.e. HMM, DTW, and DBN), and MT translation approaches (knowledge based approach, rule based approach, and hybrid based approach).

Furthermore, for each combination we applied hundred speech audio files in two clusters which were (i) training

978-1-5090-5443-5/17/$31.00 ©2017 IEEE

cluster with fifty sentences and 297 words for each sentence and (ii) testing cluster with fifty sentences -that covered four speech type categories (i.e. 14 sentences in news category, 15 sentences in conversational category, 10 sentences in scientific phrases category, and 11 sentences in control category). Hence, all speech audio files were recorded on the same microphone type, storage machine, and in the same place to guarantee an identical situation for all speech audio files recording operation. In this context, the noise factor was eliminated in speech audio files by adding artificial silence between words in order to identify the boundaries of speech. The number of experiments were examined in this study was for training cluster (747) experiments (i.e. 1 type of feature extraction, 3 types of classification techniques, (3) types of MT approaches, (50) sentences were applied, and (297) words) as well as the number of experiments in testing phase was 450 experiments. Therefore, totally the number of experiments in this study was 1197 experiments. Figure 3 illustrates the main environmental combination processes examined in this research.

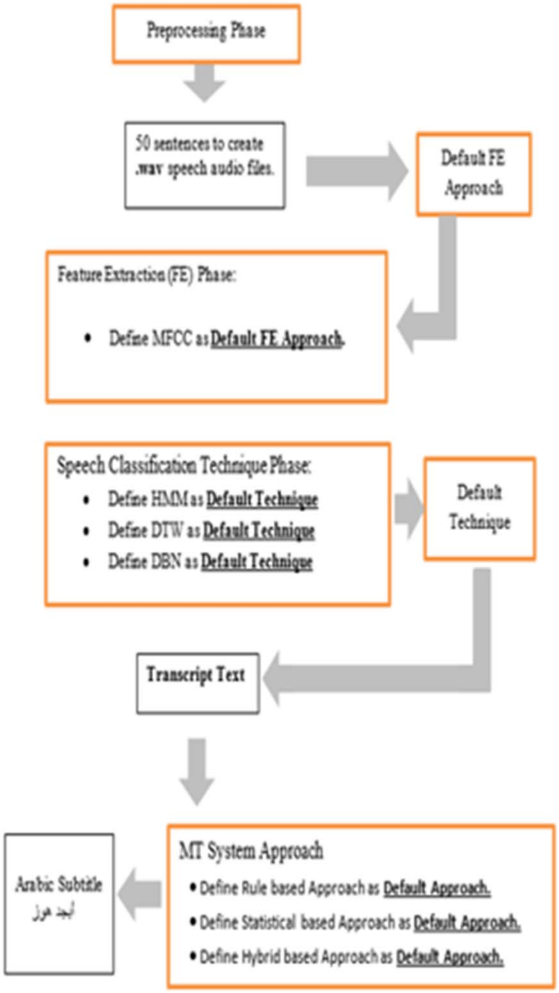
Preprocessing Phase
50 sentences to create .wav speech audio files.
Default FE Approach
Feature Extraction (FE) Phase:
• Define MFCC as Default FE Approach.
Speech Classification Technique Phase:
• Define HMM as Default Technique
• Define DTW as Default Technique
• Define DBN as Default Technique
Default Technique
Transcript Text
MT System Approach
• Define Rule based Approach as Default Approach.
• Define Statistical based Approach as Default Approach.
• Define Hybrid based Approach as Default Approach.
Arabic Subtitle
أبجد هوز

Fig. 3. The main phases were examined in this research to identify the accuracy and performance of real time speech recognition and translation from Arabic to English

IV. THE EXPERIMENTAL RESULTS OF ASR LEVEL

In this section, we provided the experimental results of comparing the speech recognition results after employing three classification techniques which are the DTW, HMM, and DBN techniques for different types of speech. Therefore, the experiments in this research took into consideration four types of speech which are news, scientific phrases, conversational, and control phrases. Thus, we examined each speech category with the three matching models by computing WRR to identify the accuracy of speech recognition.

Table 1 shows the accuracy results in term of WER measure after applying news category speech files in an ASR system that employs the DTW matching approaches.

TABLE. I. THE WER PARAMETER RESULTS AFTER ANALYZING THE NEWS SPEECH CATEGORY USING AN ASR SYSTEM USES DTW

Sentence No.	**WER**
1.	42.857143
2.	42.857143
3.	71.428571
4.	42.857143
5.	50
6.	33.333333
7.	50
8.	33.333333
9.	0
10.	71.428571
11.	37.5
12.	0
13.	22.222222
14.	0
Average	**35.55839**

Consequently, the DTW matching approach achieved 35.5% as an average WER via applying news speech files. Hence, the second stage in this level is to find the WER for ASR system that uses HMM matching approach. Table 2 shows the WER results after analyzing the news speech category using an ASR system that employs HMM for matching criteria.

TABLE. II. THE WER PARAMETER RESULTS AFTER ANALYZING THE NEWS SPEECH CATEGORY USING AN ASR SYSTEM USES HMM

Sentence No.	**WER**
1.	0
2.	71.428571
3.	33.333333
4.	33.333333
5.	0
6.	33.333333
7.	0
8.	87.5
9.	0
10.	88.888889
11.	87.5
12.	0
13.	12.5
14.	0
Average	**31.9569**

978-1-5090-5443-5/17/$31.00 ©2017 IEEE

In the third stage in this level, we examined the results in term of WER on an ASR system that employs DBN as a matching approach.

The average values of WER results drew several gaps between the classification techniques and MT systems. For instance, In Statistical-based system using DTW classification technique the lowest WER achieved in case of scientific phrases, In contrast, this combination showed highest WER in conversational category. Thus, by choosing the HMM as a classification technique; the results showed that the lowest WER achieved in conversational category. Furthermore, by using the DBN as a classification technique a WER average results showed the lowest results can be achieved after applying the control category.

In rule-based system using DTW classification technique the lowest WER achieved in case of news category, In contrast, this combination showed highest WER in control category. Thus, by choosing the HMM as a classification technique; the results showed that the lowest WER achieved in news category. Furthermore, by using the DBN as a classification technique a WER average results showed the lowest results can be achieved after applying the news category. Consequently, these results showed by using the rule-based system in MT it had a positive effect in speech recognition and translation.

The empirical findings of this study were summarized as graph in figure 4 for both levels; the classification techniques results and MT results.

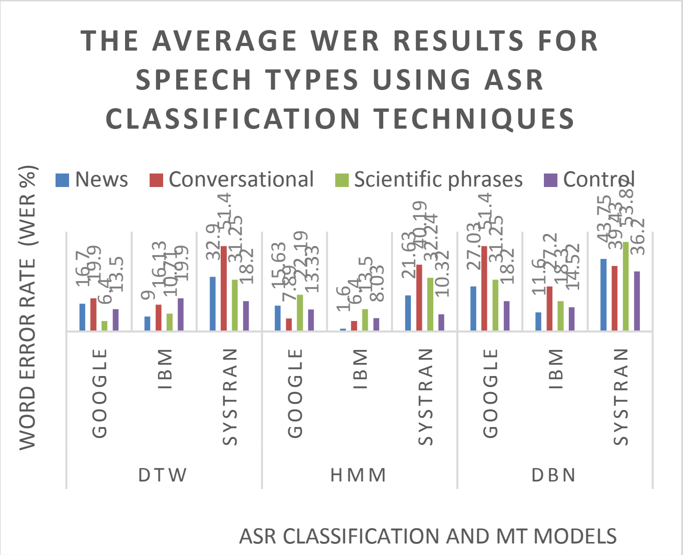

Fig. 4. The average WER results for speech types using ASR classification techniques

V. Conclusion

The main theme of this research was to measure the accuracy of ASR system classification technique with MT base system for real time speech translation from English to Arabic. In this study we ran several experiments to cover several environments and techniques to calculate the accuracy of the recognized and translated sentences in each combination. Nine environmental combinations were covered in the conducted experiments of feature extraction using MFCC with classification approaches (i.e. HMM, DTW, and DBN), and MT translation approaches (knowledge based approach, rule based approach, and hybrid based approach). For each combination we applied fifty speech audio files that covered four speech type categories (i.e. 14 sentences in news category, 15 sentences in conversational category, 10 sentences in scientific phrases category, and 11 sentences in control category). The empirical findings showed that the DBN as a classification technique achieved the best recognition rate with 79.2% compared with HMM and DTW for news category. However, the HMM classification technique achieved the best recognition rate for conversational with 80.1%, scientific phrases with 86%, and control with 63.8 % recognition rates. In contrast, using DTW as a classification technique in ASR had a negative behavior on the recognition rate for all speech categories.

On the other hand, the empirical findings from MT point of view the rule based model which was represented by IBM Watson cloud achieved the best results for in the majority of speech categories with 13.93% in conversational, 7.38% in scientific phrases, and 17.91% in control categories. The statistical based model – that was represented by Google Translate - in translation the empirical findings showed that for conversational and scientific phrases the error rate was close to rule based with an intangible difference. The hybrid translation based model influenced the error rate in the three ASR classification techniques and for all speech categories which was assigned as a negative effect.

VI. Future Work

The empirical findings of this study was for fifty sentences that were segmented in word by word speech database. Thus, for future works increasing the number of training speech audio files is recommended to increase the accuracy of research findings as well as defining new speech categories is recommended too in order to specialize the English language for Arabic automatic translation systems. In the future, we recommend to work on combining two classification techniques such as DTW and HMM over a several speech categories. As well as, we recommend to increase the number of words in each sentences to cover continues speech for more than twenty seconds. Furthermore, we recommend to work on optimizing the results of MT by using the hybrid-based rules depend on the type of speech. The Arabic language contains a lot of grammar rules. Therefore, it is recommended to enhance the data MT systems with a grammar dictionary in order to be used in a WEB-API's.

Acknowledgment

The authors are grateful to Middle East University, Amman, Jordan for the financial support granted to cover the participation in the conference and the publication fee of this research article.

978-1-5090-5443-5/17/$31.00 ©2017 IEEE

References

[1] Abdel-Hamid, O., Mohamed, A., Jiang, H., Peng, L., Penn, G., and Yu, D. (2014). Conventional Neural Network for Speech Recognition. ACM Transaction on Audio Speech, and Language Processing, 22 (10), PP. 2329-2339.

[2] Alotaibi, Y., and Hussein, A. (2010). Comparative Analysis of Arabic Vowels Using Format and an Automatic Speech Recognition System. International Journal of signal processing, image processing and pattern recognition, 3(2), PP. 11-22.

[3] Alsuliaman, M., Muhammad, G., Bencherief, M., Mahmooud, A., Ali, Z, and Al-Jabri M. (2011).Building Rich Arabic Speech Database. IEEE Fifth Asia Modeling Symposium, 3 (1), PP. 100-105.

[4] Antony, J. (2013). Machine Translation Approaches and Survey for Indian Languages. Computational Linguistics and Chinese Language Processing, 18 (1), PP. 47-78.

[5] Baker, J., Deng L., Glass, J., Khudanpur, S., Chin-hui L., Morgan, N., and O'Shaughnessy, D. (2009). Developments and Directions in Speech Recognition and Understanding, Signal Processing Magazine, IEEE, 26 (3), PP.75-80, May 2009

[6] Benzeguiba, M., Mori, R.D., Deroo, O., Dupon, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., Automatic Speech Recognition and Speech Variability: a Review, Speech Communication (2007).

[7] Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic Speech Recognition for Under Resourced Languages: A Survey. Speech Com**munication, 56 (1), PP. 85-100.**

[8] Chapaneri, S. (2012). Spoken Digits Recognition Using Weighted MFCC and Improved Feature for Dynamic Time Wrapping. International Journal of Computer Applications. 4 (3), PP. 6-12.

[9] Cutajar, M., Gatt, E., Grech, I., Casha, O., and Micallef, J. (2013). Comparative Study of Automatic Speech Recognition Techniques. The Institution of Engineering and Technology, 7 (1), PP. 25-46.

[10] Essa, E., Tolba, A., and Elmougy, S. (2008). Combined Classifier Based Arabic Speech Recognition.International Journal in Speech Recognition and Computer-Human Interaction. 4 (2), PP. 11-15.

[11] Franklen, J., West, M., and King, S. (2007). Articular Feature Recognition Dynamic Bayesian Network (DBN).Computer Speech and Language Conference, PP. 35-70.

[12] Garg, A. and Rehg, V. (2011). Audio-Visual Speaker Detection Using Dynamic Bayesian Network.The Institution of Engineering a**nd** Technology 1 (1), PP. 19-27.

[13] Gemmeke, J., Virtanen, T., and Demuynck, K. (2013). Exemplar-Based Joint Channel and Noise Compensation. IEEE [14]International Conference on Acoustic**,** Speech and Signal Processing.

[14] Ghahramani, Z. (2001). An Introduction to Hidden Markov Models and Bayesian Networks. International Journal of Pattern Recognition and Artificial Intelligence, 15 (1), PP. 9-42.

[15] Giannoluious, P., and Patamins, G. (2012). A Hierarchical Approach with Feature Selection for Emotion Recognition From Speech.Proceeding of the Eight International Conference on Language Resources and Evaluation LREC – 2012, Istanbul, Turkey, 1203-1206. ISBN: 978-951-17408-7-7.

[16] Hammo, B., Sleit, A., El-Haj, M, Baarah, A., and Abu-Salem, H., (2012). A Computational Approach for Identifying Quranic Theme. International Journal of Computing Proceeding Oriental Language. 22 (4), 189-196.

[17] Jouvet, D., and Vinusea, N. (2012). Classification Margin for Improved Class Based Speech Recognition Performance.IEEE International Conference on Acoustic, Speech and Signal Processing, Kyoto, Japan, PP. 4285-4288.

[18] Kazuma Nishimura, HiromichiKawanami, Hiroshi Saruwatari and KiyohiroShikan (2011). Investigation of Statistical Machine Translation Applied to Answer Generation for a Speech-Oriented Guidance System. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, PP. (11-18).

[19] Kurzekar, P., Deshmukh, R., Waghmare, V., and Shrishrimal, P. (2014). A Comparative Study of Feature Extraction Techniques for Speech Recognition System. International Journal of Innovative Research in Science, Engineering and Techno**logy**, 3 (12), PP. 18006-180016.

[20] Lahdesmaki, H., and Shumleuch, A. (2008). Learning the Structure of Dynamic Bayesian Networks from Time Series and Steady state Measurements. Machine Learning (ML), 71 (2), PP. 185-217.

[21] Lei, X., Senior, A., Gruenstein, A., and Sorensen, J. (2013). Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices. 14th Annual Conference of [23] International Speech Communication Association, PP. 662-665.

[22] Livescu, K, Bilmes, J., and Glass, J.(2003). Hidden Feature Model for Speech Recognition Using Dynamic Bayesian Networks. 8th European Conference on Speech Communication and Technology.

[23] Livescu, K., Glass, J., and Bilmes, J. (2013) Hidden Feature Model for Speech Recognition Using Dynamic Bayesian Networks. 8th European Conference on Speech Communication and Technology

[24] Mamta, A. and Wala, T. (2015). A Review of Various Approaches for Machine Translation. International Journal of Advance Research in Computer Science and Management Studies, 3 (2), PP. 108-113. ISSN: 2321-7782.

[25] Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice Recognition Algorithm Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Wrapping (DTW) techniques. Journal of computing, 2 (3), PP. (138-143). ISSN: 2151-9917.

[26] Ng., R., Shah, K., Aziz, W., Specia, L., and Hain, T. (2015). Quality Estimation for ASR K-Best List Rescoring in Spoken Language Translation. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, 5226-5230, ISBN: 978-4677-6997-8.

[27] Och, F. and Ney, H. (2002). Discriminative and Maximum Entropy Models for Statistical Machine Translation. Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), PP. 295-302

[28] Park, Y., Patwardhan, S., Visweswariah, k., and Gates, S. (2008). An Empirical Analysis of Word Error Rate and Keyword Error Rate. Proceeding of the International Conference on Spoken Language Processing, PP. 2070-2073.

[29] Paul, D. (1990). Speech Recognition Using Hidden Markov Model.The Lincoln Laboratory Journal- Journal of Computer Science. 3 (1), PP. 41-62.

[30] Rabiner, L, Schafer, R. (2014). MATLAB Exercises in Support of Teaching Digital Speech Processing.IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, PP. (2480-2483).

[31] Rajhona, J. and Pollak, P., (2011). ASR System in Noisy Environment Analysis and Solution for Increasing Noise Robustness. Radio Engineering, 20 (2), PP. 74-84.

[32] Rhaman, K. and Tarannum, N. (2012). A Rule Based Approach for Implementation of Bangla to English Translation. 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT).

[33] Sharwanker, V and Thakare, V. (2013). Techniques for Feature Extraction in Speech Recognition System: A Comparative Study. Computer and Information Sciences, 6 (1), 58 – 69. ISSN: 1913-8989.

[34] Singh, N., Khan, R. A., & Shree, R. (2012). MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. International Journal of Computer Applications IJCA, 54 (1), PP. 9-13.

[35] Stephenson, T., Bourland, H., Bengio, S., and Morri, A. (2000). Automatic Speech Recognition Using Dynamic Bayesian Networks with both Acoustic and Articular Variables.6th International Conference on Spoken Language Processing (ICSIP'00) China, PP. 951-954.

[36] Syahrina, A. and Lind, B. (2011). Online Machine Translator System and Result Comparison. University of Boras, School of Computing, 1 (3),PP. (18-26).

[37] Vimala, C. and Radha, V. (2012). A Review on Speech Recognition Challenges and Approaches. World of Computer Science and Information Technology Journal (WCSIT), 2 (1), PP. (1-7). 2221-0741

[38] Watanbe, S., and LeRoux, J. (2014). Black Box Optimization for Automatic Speech Recognition.IEEE International Conference on

Acoustic Speech and Signal Processing ICASSP-2014, Fortena, Italy. PP. 3256-3260.

[39] Yadav, K. and Mukhedkar, M. (2013). Review on Speech Recognition. International Journal of Science and Engineering, 1 (2), PP. 61-70. ISSN: 2347.

[40] Zhang, Y., Adl, K., and Glass, J. (2014). Fast Spoken Query Detection Using Lower Bownd Dynamic Time Wrapping on Graphical Processing. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, PP. 5173-5176.

[41] Zhu, S. and Wang, Y. (2015). Hidden Markov Induced Dynamic Bayesian Network for Recovering Time Evolving Gene Regulatory Networks. Scientific Reports, 1 (4), PP. (1-17).

[42] Mon, S. and Tun, H. (2015). Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM). International Journal of Scientific and Technology Research (IJSTR), 4 (6), PP. 349-352.

[43] Mishra, K., Bhagat, P., and Kazi, A. (2016). Automatic Subtitle Generation for Sound in Videos. International Journal of Engineering and Technology (IRJET) 3 (2), PP. 915-918.

978-1-5090-5443-5/17/$31.00 ©2017 IEEE