Hotspots for Enhancing Quranic Speech Recognition

Mubarak Al-Marri[1,a], Hazem Raafat[1,b]

[1]Computer Science Department, Kuwait University , Kuwait
[a]aljunobi@hotmail.com,[b]hazem@cs.ku.edu.kw

Abstract

This paper discusses hotspots where minor changes in these spots can lead to significant changes in results. Speech recognition is a process with multiple stages and in each stage; there are spots to be enhanced. Recording principles and avoiding issues are essential steps. Collecting the relevant data or designing the data in such form or structure can enhance the training model. In addition, extracting phonemes using a smart acoustic model or using a robust language model tool are another important spots for improvement. Finally, using Deep Neural Network (DNN) as a speech recognition model has been proved to achieve very good results.

Keywords: Computer Aided Language Pronunciation, Hidden Markov Model, Automatic Speech Recognition, Deep Neural Network, Arabic Speech Recognition, Quran Speech Recognition

1. Introduction

Automatic Speech Recognition (ASR) is a technique that enables a computer to receive and analyze human speech, and to execute relevant procedures. Speech recognition is widely used in many applications around the world. Voice dialing (Fissore, Laface, & Ruscitti, 1992), speech-to-text processing (Lamel, Gauvain, Le, Oparin, & Meng, 2011), and language learning (Pieraccini, 2012) are some examples of these applications. ASR can also play an important role in Qur'an recitation.

1.1. ASR & Qur'anic applications

The domain of the Holy Qur'an is huge. It is read at least 3.78 billion times a day, and more than 5 million people have memorized it (What is the most read book in the world?, 2014) , (Ahmed, 2013).

In fact, *"participation in recitation, as reciter or listener, is itself an act of worship"* (Nelson, 2001). Millions of Muslims around the world are keen on reciting the Holy Qur'an every day. However, there are rules of pronunciation, intonation, and caesuras that must be respected by the reciter to perform the recitation correctly. These rules, known as Tajweed rules, are themselves a study (Czerepinski, 2003). Although there are numerous teachers who know these rules and can teach them, many people cannot recite the Qur'an correctly because of certain limitations and restrictions. Moreover, people who have already memorized the Holy Qur'an must recite it from memory and confirm that they do not miss any of the verses. Therefore, supporting this holy book and contributing to this field involve helping millions of people around the world.

It is essential for any researcher, especially the beginner, to know the hotspots of Qur'anic Speech Recognition (QSR) to enhance and improve the system. Such that any change occuring in one of them, can lead to a remarkable enhancement in results. This paper will discuss those hotspots. Section two is about related work. Section three shows data collection and definition. Section four discusses recording methods and issues. Section five describes how to build a Language Model (LM). Section six includes conclusions and future work.

2. Related Work

An important topic with ASR is to know the kind of problems that researchers may face. In particular, with the Arabic language, there are some unique problems that may not exist in other languages. There are two main issues for the Arabic language with ASR that have been mentioned in (Kirchhoff, et al., 2003). First, only long vowels and consonants are represented by letters in the Arabic alphabet. Arabic diacritics represent other vowels. Since Arabic diacritics have an impact on the word's meaning, then they should exist in most words. The problem comes if diacritics don't exist when text recognition is used. However, the previous situation has no impact on voice recognition.

Second, Arabic is a rich language that has a complex morphological system. Thus, wordforms have a very high potential that affects the model stability and increases phoneme probabilities. Also, this problem should not be an issue in our model since the domain is the Holy Qur'an only and not the whole Arabic language. The Holy Qur'an has well-known content and limited wordforms. Additionally, the Holy Qur'an is one of Arabic language references and resources, so all sentences, wordforms, and grammars in Qur'an are definitely correct. The program can always compare the speech with the original Qur'anic text.

A system without experience or historical database is like a human child. People enhance their understanding and knowledge by learning. Similarly, an ASR system must be trained very well before the recognition stage, and the relationship between training and recognition is somewhat proportional: the more the system is trained, the better the recognition result obtained. The training stage is essential in any ASR system, and enhancing training results enhances recognition.

There are two important stages that should be considered to enhance the ASR model. The first stage is the extraction, where speech features are extracted from the speech signal. Enhancing this stage will have a significant impact on the next stage since all the later stages depend on the result of this stage. The second stage is the training, where the data is manipulated to build system knowledge. An attempt to enhance the extracting tool is made by Iqbal, Awais, Masud, & Shamail (2008). Usually, the extracting tool is not created for a specific language. However, their paper is discussing a tool that is specially prepared to identify Arabic vowels based on formant frequencies. The system shows about 90% as an overall accuracy. Our research is an introduction to the topic and could be enhanced in future. ASR researches and studies are going deeper with more techniques and concepts gradually.

The main four components of ASR; Feature Extraction (FE), Acoustic Model (AM), Pronunciation Model (PM), and Language Model (LM), have been discussed in (Jelinek, 1976). The first component will convert the audio waves to acoustic features using one of the filters that will be discussed later. AM depends on acoustic features and will create probability representation of each phone. The job of PM is to find the right sequence of phones that can form a valid word. Finally, LM should select the correct sequence of words that represent a correct sentence based on the defined grammars.

Zhao (1993) introduced Gaussian Mixture Models (GMM) with Hidden Markov Model (HMM) or GMM-HMM model with speaker-independent speech recognition system. The system was evaluated on TIMIT database with Viterbi algorithm (Forney, 2005) for the decoding stage. A comparison has been made in the training stage between merging algorithm and segmental K-means. System trained by merging algorithm achieved 4.1% higher accuracy in decoding than segmental K-means. Another comparison has been made between standard dictionary and compressed dictionary. In the compressed dictionary, audio signal's dynamic range has been compressed by using a certain compression threshold. The system with the compressed dictionary achieved 3.0% higher decoding accuracy than the standard dictionary. In this research, Baum-Welch algorithm (Shu, Hetherington, & Glass, 2003) is used in the training stage.

There is no clear methodolgy for collecting data in most reseaches. Some researches like (Tabbal, El Falou, & Monla, 2006) used only one Sura with four verses in training and recognition stages. Others used a ready speech corpus like (Abushariah, Ainon, Zainuddin, Elshafei, & Khalifa, 2010) and (Vergyri, Kirchhoff, Duh, & Stolcke, 2004). However, the Arabic language that has been used for both resources is not the original Arabic which has been used in the Quran. It is a local language that is related more to the country where it is spoken. Also, there is no clear structure for this data. Structured data has a positive impact on the training stage and consequently to the recognition stage. A recommended structured data will be discussed in section 3.

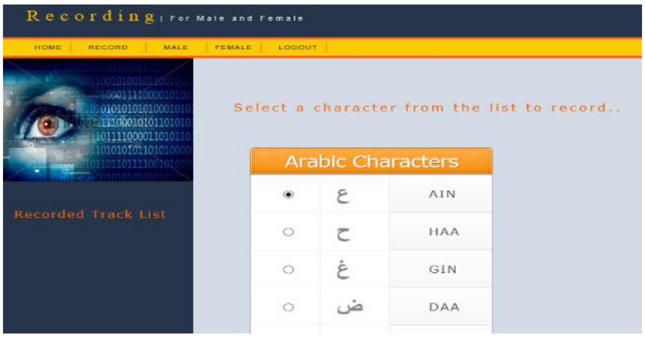
3. Data Collection & Definition

Most researchers are not paying attention to the collected data either data contents or structure. However, data quality, structure, and contents affect the result significantly. Especially with the Qur’anic speech recognition application, data should be selected with more care. The researcher should make sure that the collected data cover all concerned phonemes before the learning session. The best practice is to use Al-Qaida Noorania, Haqqani (ar-Ra'ee, 2009) since it covers all phonemes including diacritics with minimum size of data. Al-Qaida Noorania, Haqqani has been verified, used in (Abdallah, et al., 2015) , and achieved good results. Also, each phoneme has its own module which contains all related data.

3.1. Data Collection Methods

There are two primary methods for collecting the required data. The first one is the traditional method, in which we have a face to face meeting with the speaker and have a real recording and teaching session with him/her. Speaking and familiarity can be learned easily owing to direct communication with the instructor and from the context. Therefore, this method is co-operative and simple. However, the instructor must be dedicated and must give hi/her full attention to the speaker during the session. Also, the speaker is usually under stress, since all his/her mistakes will immediately be discovered and corrected by the instructor. Therefore, this method is suitable for short distances and time availability. The quality of recordings is assured because the speaker will use our microphone and recording device.

The second data collection method breaks free from most limitations in which an internet site is used to collect data (see Figure 1). The speaker selects when to record since the site is available all the time. It does not matter whether the speaker is close or far away and can record from any part of the world. Also, the speaker is free of stress, since (s)he can repeat and listen to the recorded tracks and then submit them when satisfied. However, the quality of the waves is not guaranteed, since the speaker will use his/her own devices, which might not meet our standards. We have used the first method to obtain the best quality and performance.



Select a character from the list to record..

Arabic Characters		
●	ع	AIN
○	ح	HAA
○	غ	GIN
○	ض	DAA

Figure 1: Recording using an Internet Site

3.2. Cleansing Data Process

This process started immediately after the recording stage to prepare the acoustic waves for the training process. Data purity affects the training process positively, and good training process will lead to a good result in the recognition process. Steps for this process are mentioned in the next sections.

4. Recording

4.1. Recording Session

The recording session duration is based on the script to be read and the speaker capabilities. There are two types of recording based on speaker capabilities. First, speakers who can read Arabic very well just need to read and record their voice. Second, some of them can repeat the Arabic word only rather than reading it, so for this type of speakers, a recorded tap for the whole script can be used. A speaker should listen to the phoneme, and record it with his/her voice at the same time. Usually, this may take more effort and longer time.

4.2. Recording Criteria

Adobe Audition CC 2015[1] was used in all figures. There are some criteria that should be fulfilled to assure recording quality and correctness. Missing any of these criteria is not acceptable, and it may cause to discard the acoustic wave later.

4.3. Signals should not exceed the recording boundaries.

In the extracting stage, only signals between the boundaries will be extracted. The top and the red bottom lines in Figure 2 show the boundaries. The normal status is when signals lie between boundaries and do not exceed them.

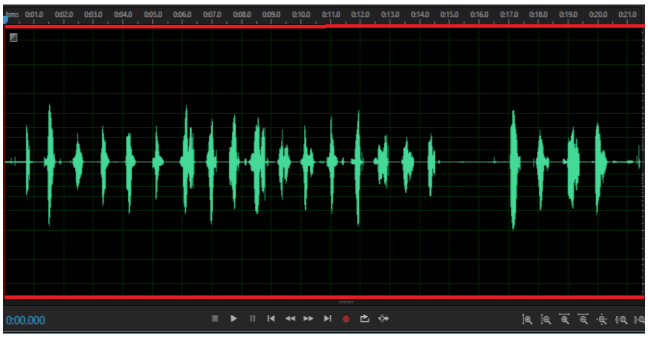


Figure 2: Bounded Signals

[1] Adobe Creative Suite, Adobe Audition CC 2015, https://www.adobe.com/mena_en/products/audition.html

However, from Figure 3 it is very clear that signals exceed the boundaries. In this situation, the one who records should tune the microphone volume until the signals lie within the boundaries. Then the speaker should repeat the recording.

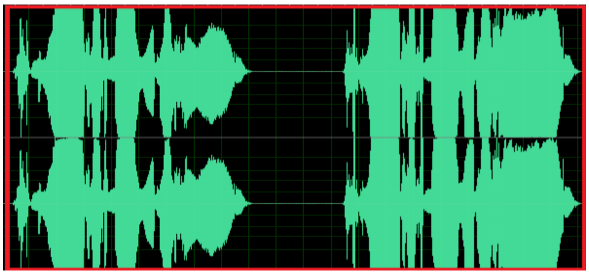


Figure 3: Signals Exceed the Boundaries.

4.4. Avoid Merged Noise.

Noise is any sound other than speaker's voice. Phonemes can be divided into three types. First, the actual signal that contains speaker's voice and has a meaning in the system's dictionary. Second, phonemes that have no meaning like any sounds other than speaker's voice. Finally, the silence that has low-frequency noise. The most important phonemes to us are first and third ones. There are two types of noise in this regard. First, noise that is located either before or after the actual signal, which is caused by clicks for example. In Figure 4, signals circled in orange are normal noise. This type of noise can be removed using the editing program.

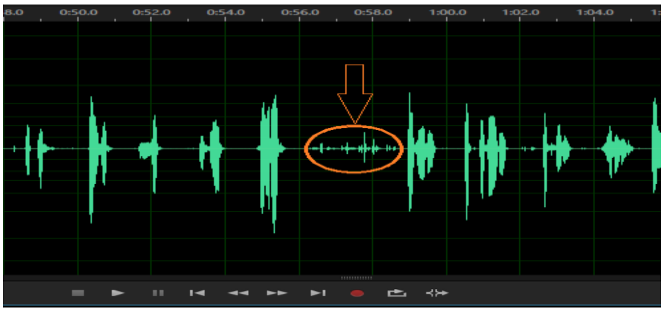


Figure 4: Normal Noise

Second, noise that is merged with the actual signal. For example, when some sounds produced by cars, animals, or any other sounds interfere or overlap with the speaker's voice. This type of noise is the most complex; splitting between noise and the speaker's voice is not an easy job. Also, the signal's quality will be affected even if splitting is successful. Figure 5 shows an example of Merged Noisy Signal where noisy segments are circled in orange. Notice that the selected phoneme in Figure 6 is the same phoneme that has been selected in Figure 5. Since it is hard to recognize the phoneme in Figure 5 visually because of the merged noise, it is hard also to be extracted. This issue is a complex topic that still needs more analysis and investigation. In this case, the speaker needs to repeat the recording.

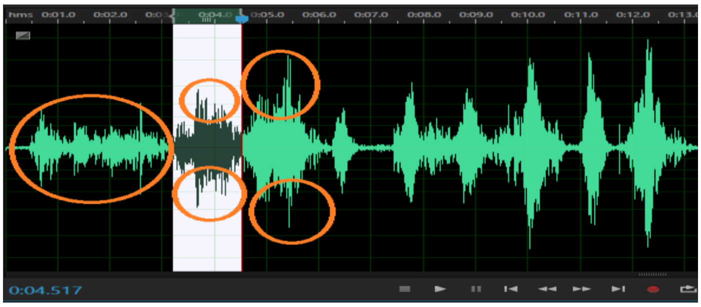


Figure 5: Merged Noisy Signal

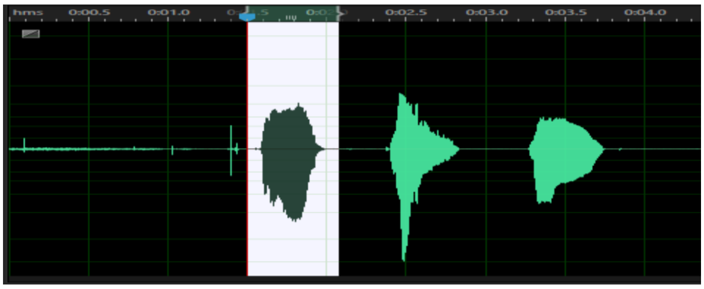


Figure 6: Signal Without Noise

4.5. Avoid Laptop-Charging while Recording

This issue was only faced in Hyderabad City, India. Usually, the recording device is a laptop. We have used about nine laptops, six of them for male recording and the rest of them for female recording. We have noticed that if the recording is done while a laptop is charging, then a thick line will appear instead of the silence stage in all laptops! It is important (Vergyri, Kirchhoff, Duh, & Stolcke, 2004) to detect the silence between all phonemes; otherwise, an extracting tool will consider the whole record as one phoneme! Also, this type of problems cannot be solved later since it affects the whole recording. Therefore, it should be avoided initially. Figure 7 shows that silence is obvious between all phonemes and is easy to detect. However, in Figure 8, the thick line noticeable between the two red lines is considered as a continuous phoneme. The silence is not clear at all, and it does not exist!

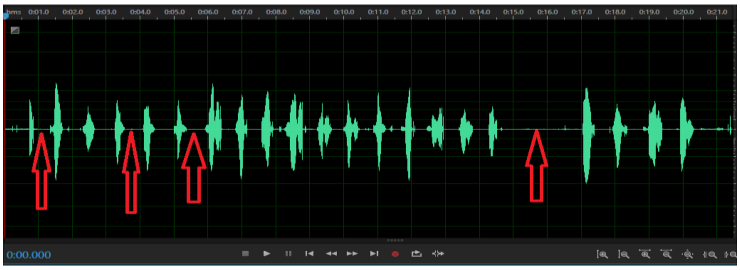


Figure 7: Silent Phoneme is Obvious

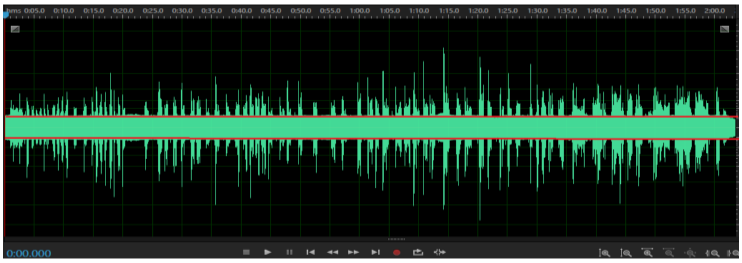


Figure 8: No Silent Phoneme

4.6. Fix Normal Recording Issues.

An example of normal recording issues is vibration noise that has been shown in Figure 4. Using the editing program tools, we should remove all clicking noise from all acoustic files. Another example is to add silence between some signals. Sometimes, the speaker is so fast during recording that (s)he merges two phonemes together without any silence. Therefore, we should check this type of mistakes and add the silence manually in the right place.

5. Building a Language Model (LM)

An example to understand the definition of LM mechanism is to imagine someone hearing only 70%–80% of voices and depending on prediction to understand utterances. One of the important questions that must be answered to solve that problem is knowing what sequence of words could form a voice. In fact, there is more than one answer, but usually, one of those answers receives a higher score than the others. This is exactly what LM is: assigning a value that represents the probability of each sequence of words per a logic.

There are many LMs based on many theories, but the most popular one is the n-gram language model (Roark, Saraclar, & Collins, 2007), and this is what we have used in our application. An n-gram is an algorithm to predict the next word after observing N-1 words using the probabilistic method. Usually, LM is used with words to find the probability of any word following another word. For example, in Arabic, the word “الخير” has the high probability to follow the word “صباح” which means “good morning” in English. However, in our case, the prediction goes deeper to reach more than one level.

Markov assumption adopts that some words are likely to follow other words depending on the context. Bigram or 2-gram assumes that probability of upcoming word can be predicted by discovering the last encountered word. The model can be extended to trigram or 3-gram by looking to the last two words in the series. Also, the number can be increased by N-gram, which is considering N-1 words in the history. Equation. 1 is used to calculate the probability of the next word (Jurafsky & Martin., 2006):

P(W_n| W_1^{n-1}) ≈ P(W_n |W_{n-N+1}^{n-1}) , where W_1^{n-1} represents the word sequence (w1 , w2 , …., wn-1), N is the gram degree, and n is number of words.

$$P_{\text{katz}}(z|x,y) = \begin{cases} P^*(z|x,y), & \text{if } C(x,y,z) > 0 \\ \alpha(x,y)P_{\text{katz}}(z|y), & \text{else if } C(x,y) > 0 \\ P^*(z), & \text{otherwise.} \end{cases} \tag{1}$$

One of the N-grams problems is when there is no example matching certain word sequence! In this case, the model will return zero. Usually, this problem will appear from bigram model and above. Unigram should always have the basic probability depending on the context. To avoid such case, the model will be decreased by one until either we reach unigram that should have a value or non-zero value before that. The previous one is called Katz back-off algorithm. Equation. 2 represents the algorithm for trigram (Jurafsky & Martin., 2006):

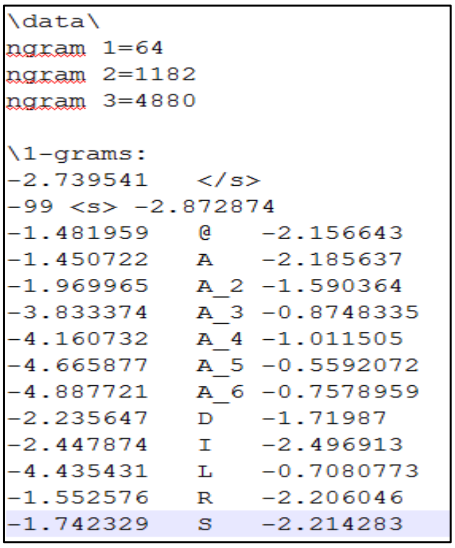
Where (P* = probability, and (x,y,z) are words in sequence, and α is the normalization factor). Since all probabilities of the N-grams are in the interval [0,1], this may cause an arithmetic underflow during multiplication expressions. Therefore, storing the probabilities as log will prevent the arithmetic underflow for two reasons. First, small numbers are not small in log space. Second, multiplying is converted to addition in log space as the following:

$$P1 * P2 = \exp(\log P1 + \log P2) \tag{2}$$

The backoff models are usually stored in ARPA[2] format like in Equation. 3 (Jurafsky & Martin., 2006):

$$\begin{array}{llll} \text{unigram:} & \log p^*(w_i) & w_i & \log \alpha(w_i) \\ \text{bigram:} & \log p^*(w_i|w_{i-1}) & w_{i-1}w_i & \log \alpha(w_{i-1}w_i) \\ \text{trigram:} & \log p^*(w_i|w_{i-2},w_{i-1}) & w_{i-2}w_{i-1}w_i & \end{array} \tag{3}$$

A part of a real LM module is shown in Figure 9.



```
\data\
ngram 1=64
ngram 2=1182
ngram 3=4880

\1-grams:
-2.739541   </s>
-99 <s> -2.872874
-1.481959   @   -2.156643
-1.450722   A   -2.185637
-1.969965   A_2 -1.590364
-3.833374   A_3 -0.8748335
-4.160732   A_4 -1.011505
-4.665877   A_5 -0.5592072
-4.887721   A_6 -0.7578959
-2.235647   D   -1.71987
-2.447874   I   -2.496913
-4.435431   L   -0.7080773
-1.552576   R   -2.206046
-1.742329   S   -2.214283
```

Figure 9: Language Model

N-gram language model has been used in many researches recently. According to Kaur (2014), *“N-gram models can be imagined as placing a small window over a text in which only n words are visible at the same time”* (p. 853). The model mechanism is to find the probability of each possible transition among n phonemes. The results proved that the increase of n ensures more accuracy. However, since the number of comparisons increases with n, the time also increases, and this is the only disadvantage of this technique.

Julius toolkit (Lee, Kawahara, & Shikano, 2001), is a recommneded open source speech recognition engine. It was selected because of its high performance and ability to deal with large vocabulary, i.e., 9-gram, in speech recognition. One of the techniques that has been used to enhance the overall accuracy was the HTK word lattice option (Young, et al., 2006). In fact, using a lattice of each word that contains all possible recitation and pronunciation errors might enhance the performance slightly and this was proved in (Abdallah, et al., 2015).

El-Desoky Mousa, Kuo, Mangu, & Soltau (El-Desoky Mousa, Kuo, Mangu, & Soltau, 2013) reported that replacing an n-gram Language Model (LM) with a Deep Neural Network(DNN) or DNN–LM enhanced the performance and reduced Word Error Rate (WER) significantly. The previous resource is a good example where DNN is used to enhance LM. Also, it was

[2] “The ARPA backoff format was developed by Doug Paul at MIT Lincoln Labs for research sponsored by the U.S. Department of Defense Advanced Research Project Agency (ARPA).”

offered a solution to Large-Vocabulary Continuous Speech Recognition (LVCSR) problem. However, in our research, DNN is used to enhance the acoustic model. Additionally, phonetics are the most suitable approach when there are limited grammars and words. Another example of discussing LVCSR problem is in Xue, Abdel-Hamid, Jiang, Dai, & Liu (Xue, Abdel-Hamid, Jiang, Dai, & Liu, 2014) as hybrid DNN and HMM models that have been revived in acoustic modeling for LVCSR. BP was also used in their training procedure.

6. Speech Recognition Models Evaluaiton

Even though there were many speech recognition models used in the past, recent researches are focusing mainly on either GMM-HMM or DNN-HMM. Also, most results indicate progress in DNN's performance over GMM in general (Abdallah, et al., 2015). There are many justifications for this results but the most important reason is related to the acoustic model. Speech recognition data is Non-Linear object, manifold object, and manipulating this type needs intensive analysis to extract the data.

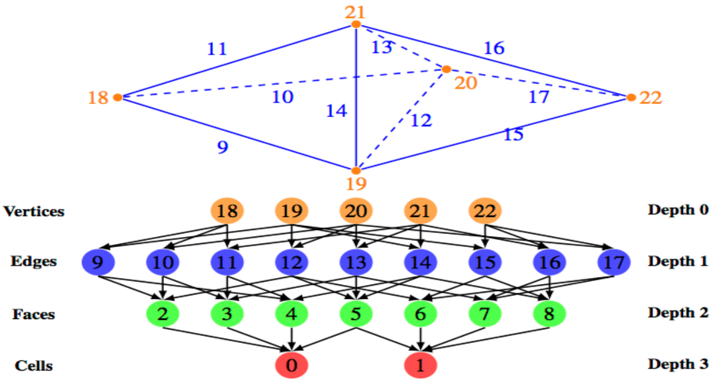


Figure 10: Manipulating Manifold in DNN (Matt Knepley's HomePage)

Figure 10 shows how DNN manipulate the manifold object. This example indicates the depth of this type of problems. For example, to determine Face 2 located in layer 2, we need to determine vertices and edges first. Therefore, we need two levels of computation to determine one or more faces. Also, another level is needed to determine the final answer. A simple example has been shown just to explain and understand the methodology. Acoustic waves are more complex than this example. This example shows how DNN can deal with this type of objects smoothly and efficiently.

7. Conclusions and Future Work

Speech recognition is an integrated process that starts with recording and ends with recognition. There are some hotspots in each part of this process where minor changes in these spots can lead to significant changes in the results. This paper has shown that Recording, Data structure, Acoustic Model, Language Model, and Speech recognition models are very important spots for enhancements. Focusing on these spots in the future will definitely enhance and improve the speech recognition applications.

8. Acknowledgements

The authors would like to acknowledge the support of their institute.

References

Abushariah, M. A., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010). Natural Speaker-Independent Arabic Speech Recognition System Based on Hidden. *ICCCE 2010*. Kuala Lumpur, Malaysia.

Abdallah, M., Al-Marri, M., Abdou, S., Raafat, H., Rashwan, M., & El-Gamal, M. A. (December 2015). Improving Holy Qur'an recitation system using Hybrid Deep Neural Network-Hidden Markov Model approach. *International Journal on Islamic Applications in Computer Science And Technology, 4*(3), 1-8.

Ahmed, M. (2013, November 19). *The Holy Quran – A Linguistic Miracle*. Retrieved January 20, 2014, from Center for Islamic Studies: http://cisweb.lk/the-miracle-of-the-quran-by-khalid-baig/

ar-Ra'ee, S. M. (2009). *Noorani Qa'idah.* India: Darul Salaam; 2nd edition (2009-01-01) (1656).

Czerepinski, K. C. (2003). *Tajweed rules of the Qur'an - Part One.* Syria - Damascus: Dar Al-Khair Islamic Books Publisher.

El-Desoky Mousa, A., Kuo, H.-K., Mangu, L., & Soltau, H. (2013). Morpheme-based feature-rich language models using Deep Neural Networks for LVCSR of Egyptian Arabic. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, (S. 8435 - 8439). Vancouver, BC.

Fissore, L., Laface, P., & Ruscitti, P. (1992). HMM modeling for speaker independent voice dialing in car environment. *1992 IEEE International Conference*, 249 - 252 vol.1.

Forney, G. J. (2005). The viterbi algorithm. *Proceedings of the IEEE, 61*, S. 268 - 278.

Iqbal, H., Awais, M., Masud, S., & Shamail, S. (2008). On Vowels Segmentation and Identification Using Formant Transitions in Continuous Recitation of Quranic. In R. Katarzyniak, *New Challenges in Applied Intelligence Technologies* (S. 155-162). Springer-Verlag Berlin Heidelberg.

Jelinek, F. (April 1976). Continuous Speech Recognition by Statistical Methods. *PROCEEDINGS OF THE IEEE*, S. 532-556.

Jurafsky, D., & Martin., J. H. (2006). *An introduction to natural language processing,computational linguistics, and speech recognition.*

Kaur, B. (2014). Review On Error Detection and Error Correction Techniques in NLP. *International Journal of Advanced Research in Computer Science and Software Engineering*, 851-853.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., . . . Vergyri, D. (2003). Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop. *The 2002 Johns-Hopkins Summer Workshop*, (S. I-344 - I-347).

Lamel, L., Gauvain, J., Le, V., Oparin, I., & Meng, S. (2011). Improved models for Mandarin speech-to-text transcription. *IEEE International Conference*, 4660 - 4663.

Lee, A., Kawahara, T., & Shikano, K. (2001). Julius --- An Open Source Real-Time Large Vocabulary Recognition Engine. *EUROSPEECH2001* (S. 1691-1694). Aalborg, Denmark: ISCA.

Matt Knepley's HomePage. (kein Datum). *http://www.caam.rice.edu/~mk51/fem.html*. Abgerufen am 24. 10 2015 von http://www.caam.rice.edu/~mk51/pictures/PLEX_0.png

Nelson, K. (01. January 2001). *The Art of Reciting the Qur'an*. Abgerufen am 20. January 2014 von books.google.com.kw: http://books.google.com.kw/books/about/The_Art_of_Reciting_the_Qur_an.html?id=faS0GZ6wPCMC&redir_esc=y

Pieraccini, R. (2012). *The Voice in the Machine: Building Computers That Understand Speech.* London, England: MIT Press.

Roark, B., Saraclar, M., & Collins, M. (April 2007). Discriminative n-gram language modeling. *Computer Speech and Language, 21*(2), 373–392.

Shu, H., Hetherington, L., & Glass, J. (2. October 2003). Baum-Welch training for segment-based speech recognition. *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on* (S. 43-48). IEEE.

Tabbal, H., El Falou, W., & Monla, B. (2006). Analysis and implementation of a" Quranic" verses delimitation system in audio files using speech recognition techniques. *Information and Communication Technologies.* Damascus, Syria.

Vergyri, D., Kirchhoff, K., Duh, K., & Stolcke, A. (2004). Morphology-Based Language Modeling for Arabic Speech Recognition. *INTERSPEECH 2004 - ICSLP.* Korea.

What is the most read book in the world? (2014, January 20). Retrieved from Answers: http://wiki.answers.com/Q/What_is_the_most_read_book_in_the_world#slide=10&article=What_is_the_most_read_book_in_the_world

Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L., & Liu, Q. (08. August 2014). Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions, 22*(12), S. 1713 - 1725.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Woodland, P. (2006). *The HTK book* (3.4 Ausg.). Cambridge: Entropic Cambridge Research Laboratory.

Zhao, Y. (1993). A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units. *IEEE Transactions on Speech and Audio Processing* (S. 345-361). IEEE.

Biodata

	Mubarak H. Al-Marri is a researcher in computer science field. He graduated with a B.Sc. (Hons.) degree in Computer Science from American University of Kuwait in 2009. He received his M.Sc. in Computer Science from University of Kuwait in 2016. He worked as a meritorious analyst with over 20 years of experience in key areas, including software analysis and development, programming and IT. Excellent planning, organizational, prioritizing, and communication skills. Cooperative team player capable of performing challenging assignments on time. His research interests include Artificial Intelligence, Neural Networks, Speech Recognition, and Cryptography. Have extensive experience in web-based applications, including designing, coding, and developing. He is a memebr of IEEE. He speaks Arabic and English languages.

	Hazem M. Raafat received the B.Sc. (Hons.) degree in Electronics & Communications Engineering from Cairo University, Egypt, in 1976, and a Ph.D. degree in Systems Design Engineering from the University of Waterloo, Canada in 1985. He worked as an Associate Professor with the Department of Computer Science at the University of Regina, Canada where he also held a joint appointment with the Electronics Information Systems Engineering Department. He is currently with the Computer Science Department at Kuwait University. His research interests include data mining, computer vision, pattern recognition, multiple-classifier systems, texture analysis, and natural language processing. He is a member of IEEE and ACM.