Jurnal Taman Vokasi
Vol. 6, Issue (2), 2018 (pp. 147-152)
p-ISSN: 2338-1825; e-ISSN: 2579-4159
http://jurnal.ustjogja.ac.id/index.php/tamanvokasi

Review of utilizing the speech identification system for spoken queries in the Qur'an

Taufik Ridwan[1], Nuur Wachid Abdul Majid[2]

[1,2] Universitas Pendidikan Indonesia, Bandung, Indonesia
E-mail: [1]taufikridwan@upi.edu, [2]nuurwachid@upi.edu

Abstracts. The speech recognition system will produce a transcription text from the sound being tested — some speech recognition systems at the Al-Qur'an show quite good accuracy. There is a big possibility to use the speech recognition system to be an input to other systems. The use of sound as input to the system to do searches in data in a text representation is called spoken query. This paper tries to conduct a discussion on the use of speech recognition systems to become spoken queries in the Qur'an data. As a result, the speech recognition system can be used as input to the Qur'anic verse retrieval system and the tajwid checking system.

Keywords: speech recognition, spoken query, Al-Qur’an

Introduction

Al-Qur'an is a holy book that becomes a life guide and the main legal source for Muslims. The entire contents of the Qur'an are written in Arabic. A Muslim is required to be able to read, understand, and carry out the contents of the Qur'an. Statistically, the Qur'an consists of 114 letters, 6,236 verses and 77,845 words (Hammo, et’al, 2007).

Some informatics research in the Qur'an has been developed. These studies are quite varied, from those in text representation to sound. In the text representation, most of the research that is built is based on search systems, both in the form of syntax and semantics, while for sound representations most systems built are speech recognition systems.

Hammo et al. Used Relational Database on Modern Standard (Hammo, et’al, 2007). Then Anwar et al. Performed Inexact String Matching through the process of stemming for verse searches of the Qur'an (Anwar, et’al, 2010). Istiadi developed a verse search system for Al-Qur'an based on phonetic similarities with Latin scripts in shipping using the rules of transliteration commonly used in Indonesia (Istiadi, 2013). Then, Rafe and Nozari try to search with Exact String Matching on the text piece to find out whether or not there is a verse from the Qur'an from the text fragment (Rafe, Vahid, & Nozari, 2014). Tarawneh et al. Conducted a search by combining syntactic and conceptual using a list of synonyms in the Al-Qur'an based on Regular Expressions (Tarawneh, Montherm, & Al-Shawakfa, 2014). Furthermore, Ta'a tried to do a verse search for the Qur'an with the ontological approach he had composed in the Al-Qur'an text (Ta’a, et’al, 2016).

Some research on the speech recognition system in the Qur'an has already been done. Noor et al. Checked automated automation on Al-Fatihah (Ibrahim, et’al 2010). Furthermore Aslam et al. Tried to develop E-Hafidz for speech recognition systems in the Qur'an with training data from the first five letters in the Qur'an (Muhammad, et’al, 2012). Then, Yuwan and Lestari take another approach to modeling the speech recognition system in the Qur'an by choosing 180 verses which have a degree of frequency similarity and high phoneme distribution plus verses ghoribah (verses in the first few letters that are read letters per message, for example QS Al-Baqarah verse 1 (Yuwan, Rahmi, Lestari, 2015). Furthermore, Yusuf tried to correct the shortcomings in his previous research by adding more speaker training data and using adaptation (Yusuf, 2016).

Actually, from search research and speech recognition systems can be combined to become a new system. The speech recognition system will produce transcription text; from this transcription text can then be an input query on a text-based search system.

Research Methods

Recognition And Query Spoken System

Speech recognition system (ASR) is a study of how the system can produce transcription text from sound signals obtained from speakers (Dahl, 2015). For humans, the process of recognizing speech sounds from people who speak is not tricky, but for a computer, it takes several methods that must be passed to be able to replicate the process.

Speech recognition systems are closely related to the pattern recognition process which is part of artificial intelligence. The introduction of patterns primarily aims to determine groups or categories of models based on the characteristics they have. Similar to other pattern recognition processes, speech recognition systems are divided into two major stages, namely the training stage and the testing phase. The training phase aims to form a model that will be used as a template used in the introduction process, while the testing phase aims to produce transcription text in the speech recognition process which is used as test data.

The idea of pattern recognition in sound is by digitizing words and converting sound waves into vectors and then matching the digital signals with a pattern (template) that has been made in the training phase. The patterns or unique characteristics possessed by the voice include amplitude, frequency, tone, and intonation.

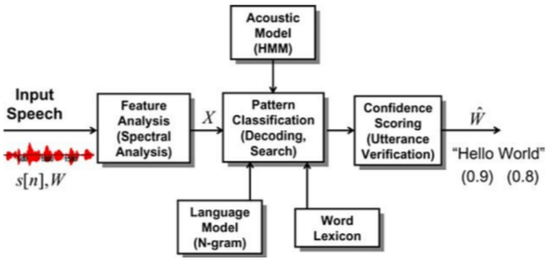

Acoustic Model (HMM)
Input Speech
Feature Analysis (Spectral Analysis)
X
Pattern Classification (Decoding, Search)
Confidence Scoring (Utterance Verification)
Ŵ
"Hello World"
(0.9) (0.8)
s[n],W
Language Model (N-gram)
Word Lexicon

Figure 1. Speech recognition system architecture (Rabiner, Lawrence, & Schafer, 2007)

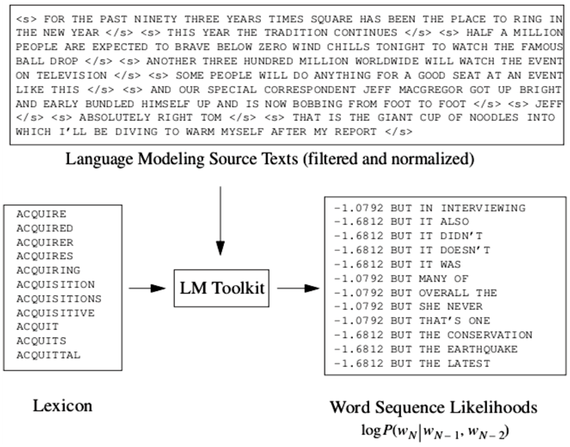
Figure 1 shows the speech recognition system architecture. In this architecture, voice input is carried out an analysis feature in the form of feature extraction. Having found the elements then go into the pattern classification stage which uses several models that have been built to compare with the data tested to become a transcription text eventually.

From Figure 1 you can see several models used by speech recognition systems. Some of these models are acoustic models, Lexicon models, and language models.

The acoustic model is a set of sub-phonetic units and statistical models used to calculate the probability of feature vectors for phoneme sequences. The acoustic model provides a match value between the possibility of transcription and sound input. Some training algorithms are used to find the highest estimated probability of likelihood (Siegler, 1999). The acoustic model must contain all sounds for each word used in the language. The speech recognition system will listen to the sound sequence that forms a word. When finding a particular series, the acoustic model will return the textual representation of the word. Thus, when the system identifies speech listening to words, it is the process of looking to a sound sequence that forms one of the words defined in the language.

The Lexicon model is a list that contains the possible pronunciation of pronunciation words and phonemes (Siegler, 1999). Lexicon will prevent some word sequences that are not permitted to avoid too much exploration to be done at the matching stage.

Language models function to estimate the probability of a sequence of words (Siegler, 1999). This probability value along with the probability value of the acoustic model will limit the search space of the word sequence to the maximum likelihood. The speech recognition system will return the highest probability of the word series hypothesis during the testing phase.



```
<s> FOR THE PAST NINETY THREE YEARS TIMES SQUARE HAS BEEN THE PLACE TO RING IN
THE NEW YEAR </s> <s> THIS YEAR THE TRADITION CONTINUES </s> <s> HALF A MILLION
PEOPLE ARE EXPECTED TO BRAVE BELOW ZERO WIND CHILLS TONIGHT TO WATCH THE FAMOUS
BALL DROP </s> <s> ANOTHER THREE HUNDRED MILLION WORLDWIDE WILL WATCH THE EVENT
ON TELEVISION </s> <s> SOME PEOPLE WILL DO ANYTHING FOR A GOOD SEAT AT AN EVENT
LIKE THIS </s> <s> AND OUR SPECIAL CORRESPONDENT JEFF MACGREGOR GOT UP BRIGHT
AND EARLY BUNDLED HIMSELF UP AND IS NOW BOBBING FROM FOOT TO FOOT </s> <s> JEFF
</s> <s> ABSOLUTELY RIGHT TOM </s> <s> THAT IS THE GIANT CUP OF NOODLES INTO
WHICH I'LL BE DIVING TO WARM MYSELF AFTER MY REPORT </s>
```

```
ACQUIRE
ACQUIRED
ACQUIRER
ACQUIRES
ACQUIRING
ACQUISITION
ACQUISITIONS
ACQUISITIVE
ACQUIT
ACQUITS
ACQUITTAL
```

```
-1.0792 BUT IN INTERVIEWING
-1.6812 BUT IT ALSO
-1.6812 BUT IT DIDN'T
-1.6812 BUT IT DOESN'T
-1.6812 BUT IT WAS
-1.0792 BUT MANY OF
-1.0792 BUT OVERALL THE
-1.0792 BUT SHE NEVER
-1.0792 BUT THAT'S ONE
-1.6812 BUT THE CONSERVATION
-1.6812 BUT THE EARTHQUAKE
-1.6812 BUT THE LATEST
```

Language Modeling Source Texts (filtered and normalized)
LM Toolkit
Lexicon
Word Sequence Likelihoods
$\log P(w_N | w_{N-1}, w_{N-2})$

Figure 2. The process of building a language model

The search system generally uses data in the form of text but does not rule out the possibility of using data in other ways such as sound. The idea used in a sound-based search system is to use the ability to recognize speech systems or Automatic Speech Recognition (ASR) to produce transcriptions of text from queries that are then used on other systems that use text representation. In theory, speech recognition systems can be used to support systems such as voice interfaces, voice search, dialogue systems, audio information retrieval, and other systems that utilize the use of voice data (Dahl, 2015).

The search system that uses spoken queries is almost the same as a search system that uses input in the form of text. The only difference is the use of the query in the kind of sound. Users have to say the word they want to search without having to care about how the word should be written.

One way to measure the performance of speech recognition systems is the value of Word Error Rate (WER) or the average number of errors. WER is obtained by finding the difference between the reference transcription and the results of the introduction using the Levenshtein Distance algorithm.

$$\text{WER}=\frac{\text{Insertions+Substitutions+Deletions}}{\text{Total word in Correct Transcript}}*100\% \ \ldots\ldots\ldots\ldots\ldots(1)$$

$$\text{Akurasi} = 100\ \% - \text{WER}\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Insertions are the number of word insertions found in the transcription hypothesis, while the Deletions are the opposite, namely the number of words deleted in the transcription hypothesis. Substitutions are the number of words replaced by the assumption of ASR.

Speech recognition system research is not easy and has several challenges because a person's voice signals are very varied, differences in pronunciation style and the presence of noise sounds from the surrounding environment (Abdel-Hamid, et'al, 2014).

Spoken Query Architecture In Al-Qur'an

The use of spoken queries is fundamentally changing the input queries that come from text representations to sound representations. Even so in the end, the sound-shaped question will eventually be converted back into text. Here are some systems that use spoken questions.

- Spoken Query for the Al-Qur'an Gathering System

The back-and-forth system that uses spoken queries is almost the same as the back-and-forth system that uses text queries. The difference that distinguishes it is only query representation. At the spoken query the user says the word he wants to search without needing to care about how the word should be written. One of the challenges to spoken use queries for the reverse system is the accuracy of the ASR transcription results. When the Word Error Rate (WER) value generated by ASR is high, it will also have an impact on reducing the accuracy of the system being built.

Ridwan and Lestari describe the feedback system in the Qur'an by using spoken queries (Abdel-Hamid, et'al, 2014).

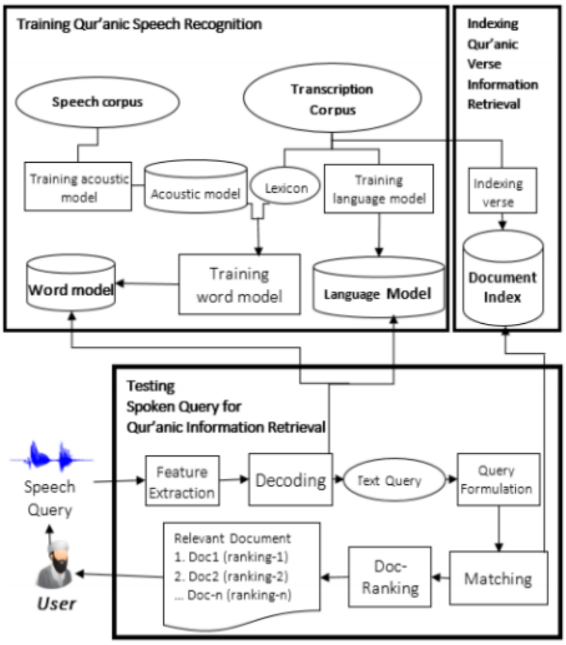


Figure 3. Spoken Query Architecture in the back-meet system in the Qur'an

From the system architecture, it can be seen to be able to speak queries on the back-to-back system to do the speech recognition process by ASR, then the system can search and return documents in a certain order.

- Spoken Query for Tajweed Checks

The process of checking tajwid law in the form of text is basically an attempt to find certain patterns that include certain tajwid laws on a data representation in the form of text. The method used is one of which can use regular expressions.

Ridwan and Majid describe the architecture of the Tajwid checking system. In general, the system consists of the process of building a speech recognition system, building a text-based

Tajweed checking system, and testing the Tajweed checking system using greeting input (Ridwan, Taufik, & Majid, 2018). As for more details can be seen in Figure 4.

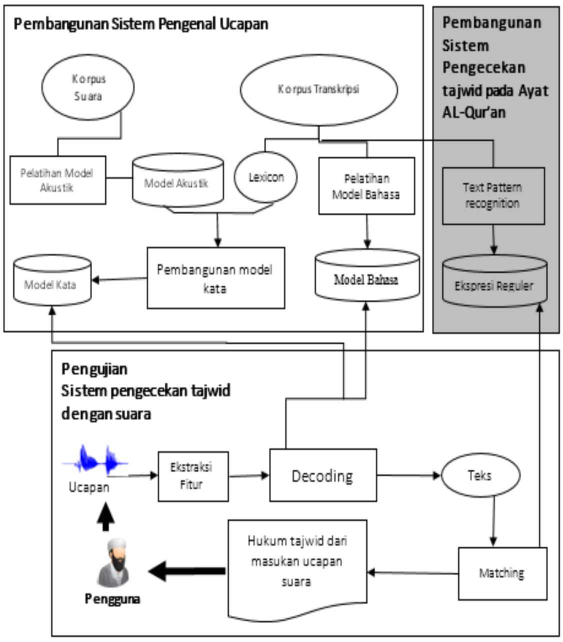


Figure 4. Spoken Query Architecture for checking Tajweed

From the system architecture in Figure 4, the process of getting a transcription text is done by using a speech recognition system. When the query has been in the form of writing, a matching process is performed using regular expressions to recognize specific patterns.

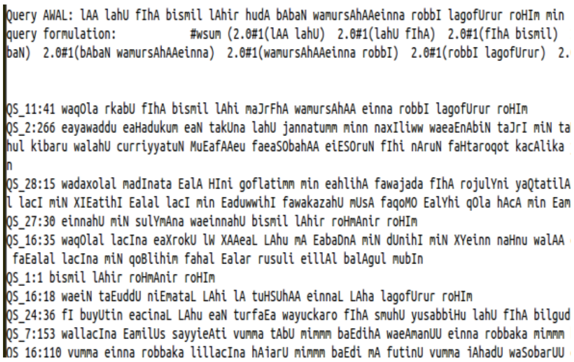
Research Results

Results The testing of the back-and-forth system in the Qur'an which uses spoken queries with Inference Networks was found to be the effect of the WER value on speech recognition and query formulation systems.

The effect of the WER value is that the greater the error value in the speech recognition system, the lower the value of the meeting-system ability to find relevant documents can be seen from the Mean Reciprocal Rank which shows the order in which a related document is found which decreases.

The results of the use of query formulations on transcription text originating from the correct text do not show significant results, but in the ASR scheme that has transcription errors shows substantial results. Query formulations carried out overlapping on ASR transcription can improve search system performance by still finding relevant documents even though the system is wrong in transcription. The best results achieved are by using an overlapping scheme on a phrase consisting of 2 words which show the MRR value of 0.9714 for WER of 5.55%, then 0.9239 for WER of 15.96%, and 0.8582 for WER of 33.93%.

The results of the introduction of speech and verse search can be seen in Figure 5.



```
Query AWAL: lAA lahU fIhA bismil lAhir hudA bAbaN wamursAhAAeinna robbI lagofUrur roHIm min
query formulation:          #wsum (2.0#1(lAA lahU)  2.0#1(lahU fIhA)  2.0#1(fIhA bismil)
baN)  2.0#1(bAbaN wamursAhAAeinna)  2.0#1(wamursAhAAeinna robbI)  2.0#1(robbI lagofUrur)  2.


QS_11:41 waqOla rkabU fIhA bismil lAhi maJrFhA wamursAhAA einna robbI lagofUrur roHIm
QS_2:266 eayawaddu eaHadukum eaN takUna lahU jannatumm minn naxIliww waeaEnAbiN taJrI miN ta
hul kibaru walahU curriyyatuN MuEafAAeu faeaSObahAA eiESOruN fIhi nAruN faHtaroqot kacAlika
n
QS_28:15 wadaxolal madInata EalA HIni goflatimm min eahlihA fawajada fIhA rojulYni yaQtatilA
l lacI miN XIEatihI Ealal lacI min EaduwwihI fawakazahU mUsA faqoMO EalYhi qOla hAcA min Eam
QS_27:30 einnahU miN sulYmAna waeinnahU bismil lAhir roHmAnir roHIm
QS_16:35 waqOlal lacIna eaXrokU lW XAAeaL LAhu mA EabaDnA miN dUnihI miN XYeinn naHnu walAA
 faEalal lacIna miN qoBlihim fahal Ealar rusuli eillAl balAgul mubIn
QS_1:1 bismil lAhir roHmAnir roHIm
QS_16:18 waeiN taEuddU niEmataL LAhi lA tuHSUhAA einnaL LAha lagofUrur roHIm
QS_24:36 fI buyUtin eacinaL LAhu eaN turfaEa wayuckaro fIhA smuhU yusabbiHu lahU fIhA bilgud
QS_7:153 wallacIna EamilUs sayyieAti vumma tAbU mimmm baEdihA waeAmanUU einna robbaka mimmm
QS 16:110 vumma einna robbaka lillacIna hAiarU mimmm baEdi mA futinU vumma iAhadU waSobarUU
```

Figure 5. Test results for spoken queries originating from sound files

The results of the introduction in Figure V.9 are found:

- Correct text: waqOla rkabU fIhA bismil lAhi maJrFhA *** wamursAhAA einna robbI lagofUrur roHIm
- ASR Results: lAA lahU fIhA bAlL hArr hAA hAA hArAA hAAA eAna robbI lagOF roHIm min
- WER: 6/11 * 100% = 54.54%

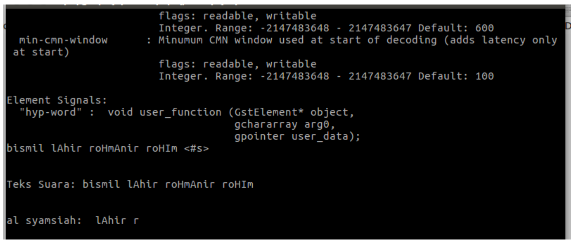
In testing, the WER value was found to be 55.54%. Although some words are wrong in the recognition process, it turns out that the system can still perform the relevant document retrieval process at the top.

The test results of the Tajwid checking system show that the system can recognize the regular expression patterns that have been compiled. Table 1 shows the regular expressions of the recitation patterns that have been collected.

Table 1. Regular and tajwid expression patterns

No	Law	Pattern
1	mad lin	*Other letters + V* **OR** *other letters + Q*
2	Mad	*Other letters + A* **OR** *other Letters + I* **OR** *other Letters + U*
3	Mad Lazim Mutsaqqal Harfi	*Other letters + Mad Thobii+ Mad Thobii+ Mad Thobii*
4	Mad wajib Muttashil	*(AA* **OR** *OO* **OR** *II* **OR** *UU) + e*
5	Mad wajib Muttashil	*(AA* **OR** *OO* **OR** *II* **OR** *UU) + space + e*
6	alif lam syamsiyyah	*Two letters that same in initial word*
7	alif lam qomariyah	*Initial letter + l+ word*
8	Lafal Jalallah	*Letter + LAh*
9	idgham bilaghunnah	*ll* **OR** *r+r*
10	idgham bighunnah	*m mm* **OR** *y yy* **OR** *w ww* **OR** *n nn*
11	Iqlab	*mmm b*
12	Ikhfa	*N + letter*
13	Idzhar	*n + letter*

Figure 6 shows an example of the introduction of the recitation of recitation law.



```
                        flags: readable, writable
                        Integer. Range: -2147483648 - 2147483647 Default: 600
  min-cmn-window      : Minumum CMN window used at start of decoding (adds latency only
 at start)
                        flags: readable, writable
                        Integer. Range: -2147483648 - 2147483647 Default: 100

Element Signals:
  "hyp-word" :  void user_function (GstElement* object,
                                    gchararray arg0,
                                    gpointer user_data);
bismil lAhir roHmAnir roHIm <#s>


Teks Suara: bismil lAhir roHmAnir roHIm


al syamsiah:  lAhir r
```

Figure 6. Test results for the introduction of Tajwid

Figure 6 shows the transcription text that was successfully identified by the system according to the actual speech sounds. From Figure 6 it can also be seen the introduction of the law of alif lam syahsiah which was successfully carried out by the system. This shows the network has succeeded in matching the rules of recitation that were built.

From the test, it was found that the system has been able to recognize the correct recitation law according to the applicable rules. However, when the speech recognition system incorrectly identifies (transcribes) some of the sound words being tested, the system cannot know some of the laws of the word in question.

From the results of this test, it can be concluded that the testing process of the tajwid checking system will depend on the speech recognition process. This is because the system input comes from another system, namely the speech recognition system. The higher the accuracy of the speech recognition system, the better the Tajweed checking system identifies the recitation of the voice of the speaker being tested.

Conclusion

The results of spoken query research found the effect of the accuracy value of the speech recognition system on the back-to-back system that uses spoken queries. The greater the amount of the error in the speech recognition system, the lower the value of the system meeting-back ability to find relevant documents.

In the next stage can be spoken queries that have been built using a more interactive user interface and looking for ways to improve the accuracy of the speech recognition system.

In further research, we can find ways to improve the accuracy of the speech recognition system in testing. This can help other methods that will use spoken queries to enhance their performance.

References

Abdel-Hamid, Ossama., Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, & Dong Yu (2014): Convolutional Neural Networks for Speech Recognitio,. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, october 2014.

Aslam Muhammad, Zia ul Qayyum, Waqar Mirza M., Saad Tanveer, Martinez-Enriquez A.M., & Afraz Z. Syed. (2012). E-Hafiz: Intelligent System to Help Muslims in Recitation and Memorization of Quran, *Life Science Journal*.

Anwar, Agus Sofiyan., Zainal Abidin, & Ririen Kusumawati (2010). *Mesin Pencari Ayat Al Quran Menggunakan Inexact String Matching*. UIN Maliki Malang.

Dahl, George Edward (2015): *Deep Learning Approaches to Problems in Speech Recognition, Computational Chemistry, and Natural Language Text Processing*, Disertasi Program Doctor, University of Toronto

Hammo B, Sleit A, El-Haj M. (2007). Effectiveness of *query* expansion in searching the Holy Quran. *Proceedings of the Second International Conference on*

Arabic Language Processing, CITALA. 07.

Istiadi, Muhammad Abrar (2012). *Sistem Pencarian Ayat Al-Qur'an Berbasis Kemiripan Fonetis*, Skripsi Program Sarjana, Institut Pertanian Bogor.

Noor Jamaliah Ibrahim, Mohd Yamani Idna Idris, Zaidi Razak, & Noor Naemah Abdul Rahman. (2010). *Automated Tajweed Checking Rules Engine for Quranic Learning*, University of Malaya.

Rabiner, Lawrence R., & Ronald W. Schafer. (2007). *Introduction to Digital Speech Processing,* Boston: now Publishers.

Rafe, Vahid., & Morteza Nozari. (2014). An Efficient Indexing Approach to Find Quranic Symbols in Large Texts, *Indian Journal of Science and Technology,* Vol 7(10), 1643–1649, October 2014.

Ridwan, Taufik & Lestari, Dessi Puji. 2017. Spoken Query for Qur'anic Verse Information Retrieval. *Proceeding of The Oriental Chapter of International Committee for The Co-ordination and Standardization of Speech Database and Assesment Techniques (O-COCOSDA) 2017.*

Ridwan, Taufik & Majid, Nuur Wachid Abdul. (2018). Development System for Recognize Tajweed in Qur'an using Automatic Speech Recognition. *Proceeding of International Conference on Science and Technology for Internet Of Thing (ICSTI) 2018*

Siegler, Matthew A. (1999). Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance, Disertasi Program Doctor, Carnegie Mellon University

Tarawneh, Montherm., & Emad Al-Shawakfa. (2014). A Hybrid Approach for Indexing and Searching The Holy Quran, *Jordanian Journal of Computers and Information Technology (JJCIT),* Vol. 1, No. 1, December 2015.

Ta'a, Azman., Qusay Abdullah Abed, Bashah Mat Ali, & Muhammad Ahmad (2016): Ontology-Based Approach for Knowledge Retrieval in Al-Qur'an Holy Book, *International Journal of Computational Engineering Research (IJCER)* Volume 06, 2016.

Yusuf, Rahmatullah. (2016). *Transformasi model akustik spesifik terhadap Pembicara pada sistem pengenal ucapan Al-Quran dengan MAP dan CMLLR*. Tugas Akhir. Institut Teknologi Bandung.

Yuwan, Rahmi., dan Dessi Puji Lestari. (2015): Pengembangan Sistem Pengenalan Bacaan Al-Qur'an Memanfaatkan Phonetically Rich and *Balanced*d Corpus. *Konferensi Nasional Informatika* ITB.