International Journal of Grid and Distributed Computing
Vol. 11, No. 3 (2018), pp.139-154
http://dx.doi.org/10.14257/ijgdc.2018.11.3.12

On the Use of Multi-lingual Approach for a Cloud-based Transcription System for the 'Ilonggoish' Dialect

Rowena Alibagon[1], Frank Elijorde[2], Joel De Castro[3] and Yungcheol Byun[4*]

[1]*St. Vincent College of Science and Technology*
Leganes, Iloilo, Philippines
[2,3]*College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines*
[4]*Department of Computer Engineering, Jeju National University, South Korea*
[1]*wheng21_alibagon@yahoo.com,* [2]*felijorde@wvsu.edu.ph,*
[3]*decastrojoel@wvsu.edu.ph,* [4]*ycb@jejunu.ac.kr*

Abstract

The study is aimed at the development of a Transcription System for 'Ilonggoish' Dialect, which is a widely-spoken local language in the Philippines. It is a software that records speech in .wav file format, transcribes speech into text, and generates text file containing the transcribed text. The system has a built in speech recognition that has the capability to recognize pre-recorded speeches spoken in different languages such as English, Filipino, Hiligaynon, and Ilonggoish dialect. Integrated into the system are the recording tool for the input speech data, data storing capability in .wav format, and text storing capability in .txt format. This study presents an approach to extract features of the spoken words by using the Mel Frequency Cepstral Coefficients (MFCC) algorithm from speech signals of isolated spoken words, and Hidden Markov Model (HMM) method in presenting the recognized spoken words in text format. The system uses the Google Cloud's database of words as the baseline for standard words. It was evaluated by linguists specializing in Filipino, English, and Hiligaynon languages, and IT experts in different fields such as the academe and industry.

Keywords*: Hidden Markov Model, Mel Frequency Cepstral Coefficients, Speech Recognition, Transcription System*

1. Introduction

Research in the field of automatic speech and speaker recognition has now spanned more than five decades [1]. Of the many branches of digital signal processing, speech processing is considered to be one of the most important. Automatic speech recognition (ASR) aims to map an audio signal, containing speech, into a text transcription containing a sequence of words. Basically, the goal is to match the transcription as close as possible to the audio message, with no particular understanding of the meaning or scope of what was spoken [2].

Transcription Services use ASR that converts spoken words into text. It has very important applications such as command recognition, dictation, foreign language translation, security control which verifies the identity of the person to allow access to services such as banking by telephone. ASR makes writing on computer applications much easier and faster than using keyboards and could help handicapped people to interact with society.

Received (December 27, 2017), Review Result (February 25, 2018), Accepted (March 5, 2018)
* Corresponding Author

ISSN: 2005-4262 IJGDC

Globally, there is a widespread need for transcription services: minutes of meetings, court reports, medical records, interviews, videos, speeches, and many others. Written text is easier to analyze and store than audio files; and other than this, there are many circumstances one could imagine for the need to transcribe human speech: those who are deaf also need to listen to certain audio files, people with limited ability to type, such as those who are paralyzed or suffer from Carpal Tunnel Syndrome, likewise, need to draft documents, and so on.

Despite important progress, there is currently no universal automatic speech recognition system which is able to transcribe any recording with equivalent performance. In fact, automatic systems can sometimes achieve as good as human annotators, but their performance is strongly dependent in the recording as well as in the quality of the learning phase with respect to the target task. Acoustic and language models can, in most cases, be adapted to new application areas through the integration during the learning stage of additional prior knowledge related to the quality of the recording, the nature of speech, the accent or the lexical field [3].

Though there are considerable amount of work already done in Automatic Speech Recognition (ASR) for English, European, and few Asian languages [4][5], no work has been done previously on the part of Ilonggo language, or specifically 'Ilonggoish' in Iloilo, which is a mixture of the Hiligaynon and English languages. The main reason behind this is the unavailability of linguistic rules and the large annotated corpora for Ilonggo language. Subjecting 'Ilonggoish' to ASR is of special interest due to its broad range of implementations, which can be useful for different applications such as command recognition, dictation, foreign language translation, security control, and many others; thus, this serves as the motivation for the researchers to pursue the study.

2. Related Works

2.1. Feature Extraction Techniques

Feature extraction is the first step in an automatic speech recognition system. It aims to extract features from the speech waveform that are compact and efficient to represent the speech signal. The Linear Predictive Coding presents a compact and precise representation of the spectral magnitude for signals and generates coefficients related to the vocal tract configuration. In LPC, speech sample can be estimated as a linear combination of past samples. Several researches have been performed on Arabic speech recognition using LPC features. In the article Speech Recognition Using Scaly Neural Networks, the authors designed a system using the scaly type architecture neural network for the recognition of speaker dependent isolated words for small vocabularies (11 words). They use LPC features extraction method and get a success rate of 79.5-88 % [6]. Choubassi *et al.* implemented Arabic isolated speech recognition. It uses Modular Recurrent Elman neural networks (MRENN) for recognition and LPC for feature extraction. The recognition rate for 6 Arabic words ranges between 85% and 100% [7]. Linear predictive coding (LPC) has always been a popular feature due to its accurate estimate of the speech parameters and efficient computational model of speech [8]. One main limitation of LPC features is the linear assumption that fails to take into account of the non-linear effects and sensitivity to acoustic environment and background noise [9]. Since the system will record dataset in calm environment, this study will, likewise, use LPC due to its advantages in non-noisy (silent) systems.

2.2. Automatic Speech Summarization

One approach attempted to extract information from such a database by tracking speech by query matching to indexes based on an automatic recognition result which had been synchronized with the speech data [10]. However, users attempting to retrieve information from such a speech database, would and do prefer to access abstracts rather than the whole data, before they decide whether they are going to read or hear the entire data or not. Meeting/conference summarization will become useful if it can be developed to extract relatively important information scattered about in the original speech. In the natural language processing field, recently a sentence compression technique using both text and its abstract has been proposed [11]. However, this technique cannot be directly used for speech summarization. The current issue for automatic speech summarization is how to deal with recognition result including word errors. Handling word errors becomes one fundamental aspect for successfully summarizing transcribed speech. In addition, since most approaches extract information based on each word, approaches based on longer phrase, or compressed sentences are required for extracting messages in speech. In [12] [13], the authors designed a system for recognition of isolated Arabic words by using a combined classifier. A combined classifier is based on a number of Back-Propagation/LVQ neural networks with different parameters and architectures and MFCC features are used. The datasets are records taken from the Holy Quran for many famous reciters. They found that the implemented combined classifier outperforms those traditional classifiers which use the HMM-based speech recognition approaches. The proposed system used only single feature extraction method MFCC and is tested only for 10 words and they restricted the type of words to not have homophone or noise and they used famous Quran reciters not normal speaker. Also, they used manual segmentation to find the boundary of words and when classifying the training data they found some errors.

2.3. Audio Processing

Auditory scene analysis refers to audio processing for extracting and interpreting information in the environment. Bregman gives a comprehensive treatment of this field. Computational auditory scene analysis (CASA) considers computational approaches to understanding the auditory environment, such as automatically separating speech from noise [14]. Bottom-up approaches to analysis, which integrate low level information and forward it to higher level processes corresponds to primitive segregation in Bregman's terms. The speech detector in this thesis uses a bottom-up approach. Top-down approaches, equivalent to Bregman's schema-based segregation, use information at higher level modules to bias lower level perception. Ellis describes a system that uses a predictive model to interpret the auditory scene by resolving expected input with actual received input [15]. Speech detection may be viewed both as a recognition and segmentation problem. Segmenting an audio stream into classes may be performed without recognition. In this work, an audio stream is represented as a sequence of feature vectors, and the distance between successive features is computed. Changes in the audio signal, such as the start or end points of a speech event, should produce peaks in the distance function splitting the audio stream by choosing peaks results in segmentation [16]. Goodwin takes a similar approach, but does use a training phase to find a distance function that reweights the features appropriately to focus on boundaries that distinguish acoustic classes. In addition, the peak picking procedure in their work uses dynamic programming to find the best set of peaks to use as segment boundaries subject to a cost function [17].

3. Transcription System for the 'Ilonggoish' Dialect

3.1. System Architecture

The system presented in this study is a web-based software that converts the input speech signals into word sequences. This system recognizes the recorded wave file. It has a built-in recording tool for the input speech data, data storing capability in .wav format and text storing capability in .txt format. This study presents an approach to extract features of the spoken words by using the Mel Frequency Cepstral Coefficients (MFCC) algorithm from speech signals of isolated spoken words and Hidden Markov Model (HMM) method in presenting the recognized spoken words in text format. The system uses the Google API's database to compare recognized speech audio to existing English or Filipino language texts. The best matched words based on the sound or phonemes are displayed on the screen. The Architecture of the system is shown in Figure 1.

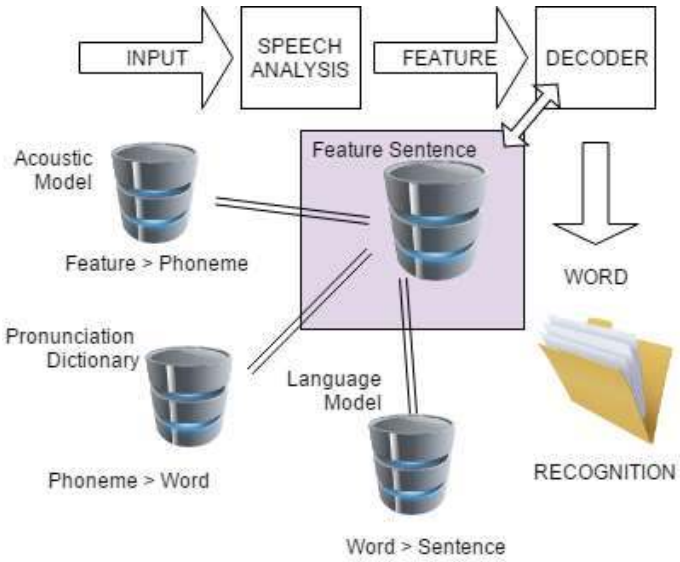

Figure 1. The System Architecture

The Speech Pre-Processing stage follows this methodology: through the built-in recording tool, input speech data is stored in .wav format. Quality of recorded data depends upon the recording device and speed. Segmentation is used to divide raw data speech signal into evenly spaced frames. The frame size is selected in smaller size to capture rapid transitions and achieve sufficient resolution in frequency domain. The frame size is 10 seconds of recording; segmented frames are analyzed to produce feature vector which is useful for speech sound classification. The required measurement parameters for feature analysis are recording channel, environmental noise and speaker variability.

Speech recording and computation of basic recognition features run only on client's side. These features are transferred to the recognition server via Internet. On the recognition server, the recognition features are computed and demanded recognition is executed. Finally, the recognized text is sent back to client's speech application. The Decoder is the part of the system that processes the feature which is the value of the sound. This represents the phoneme of the spoken word. The Decoder's work is to convert each phoneme into text by looking at the most probable text value for the generated feature in the acoustic model, pronunciation dictionary, and language model databases of GoogleTM.

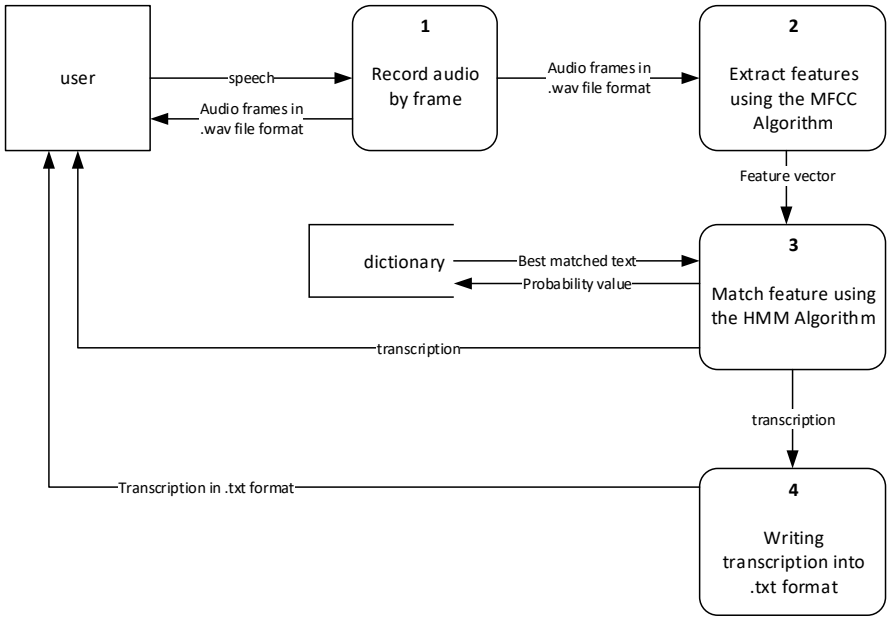

Figure 2. Level 0 Data Flow Diagram of the System

The level 0 of the Data Flow diagram of the system is presented in Figure 2. This contains the much detailed information about the system. The user enters the speech through the use of a microphone. The system then records the speech in .wav format. For longer speech, the speaker may speak continuously. The system simply divides the speech into several chunks of .wav files, each having a timeframe of 10 seconds. This allows the system to easily process the recoded audio. The audio would then be subjected to the MFCC algorithm which divides into frames where the phonemes would be extracted. The phoneme is subjected to the HMM algorithm where the probability of match with the word's sound in the dictionary is computed. The best matched words are displayed to the user, and the system automatically saves the best matched word or words in a text file.

The complete detail of Process 2 which is extracting features using the MFCC algorithm is presented in Figure 3. The first step of process 2 is framing the signal into short frames. The speech that the user has entered in the system is being represented as a continuous raw audio signal that is constantly changing. To calculate the power spectrum of this audio signal, it needs to be divided into small frames where it could be represented as a stationary signal or audio signal that does not change much. The standard size could be within the range of 20-40 milliseconds. The framed signal is then subjected to the Fast Fourier Transform (FFT) where the periodogram estimate of the power spectrum is calculated. The power spectrum is the power of the signal or the energy within the given frequency [18]. The output of the process is the estimated energy or power spectrum present in a frame.

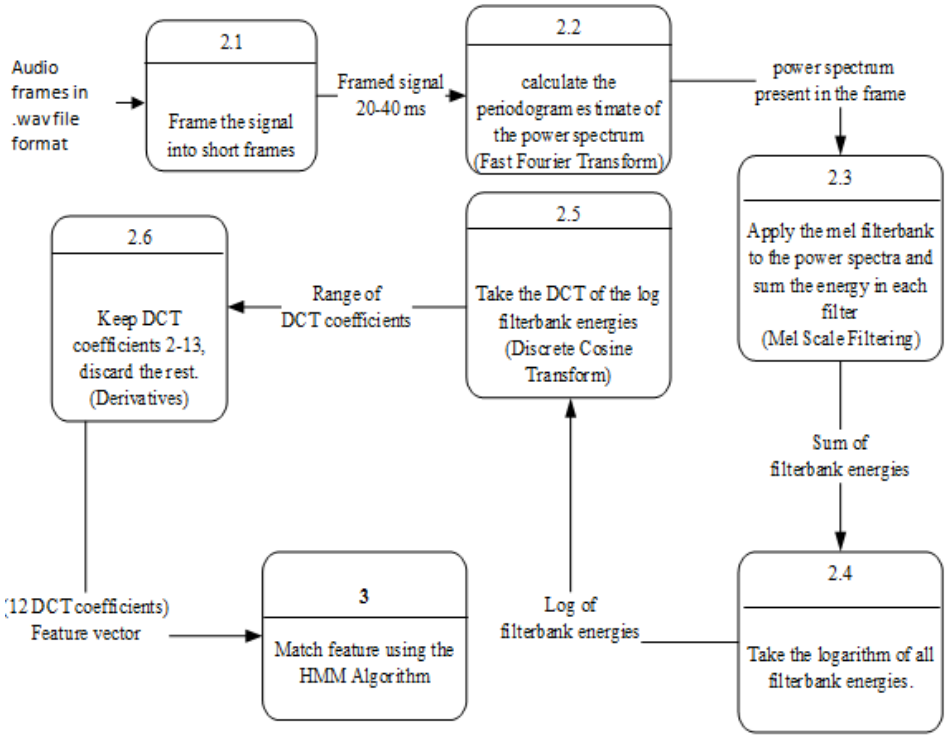

Figure 2. Level 0 Data Flow Diagram of Feature Extraction Process

The power spectrum is then subjected to Mel Scale filtering wherein the perceived frequency derived from the previous process is being related to the actual frequency of the speech. This is done by using the filtering calculations which indicate how the frequencies be spaced depending on its energy. As the frequency gets higher, the spacing gets wider. The result of this process is a frame divided into several filters having a calculated value of energy for each filter. These energies are then summed up per frame. The logarithm of the result is then calculated which is subjected to Discrete Cosine Transform (DCT) whose purpose is to suppress speaker dependent harmonics.

The DCT generates cepstral coefficients which normally range from 13 to 20 coefficients. To represent the changing nature of speech, only the 13 cepstral coefficients are used to derive the final feature vector or the value of the phoneme. This is because, the higher the value of the cepstral coefficients the higher the rate of the changes in the filterbank energies. The final feature vector's value is then passed to the HMM algorithm for matching.

3.2. Establishing the Ilonggoish Phoneme

3.2.1. Hiligaynon Language

Hiligaynon Language is a language mostly found spoken in the Visayan branch of the Central Philippine language family. Approximately 11 million people speaks this language in Western Visayas in the Philippines mostly coming from provinces of Iloilo and Negros Occidental, Capiz, Antique, Aklan and Guimaras [19]. Hiligaynon was widely written based on Spanish orthography consisting of 32 letters called ABECEDARIO:

A B C Ch D E F G H I J K L Ll M N Ng Ñ Ñg/Ñg̃/Ñg O P Q R Rr S T U V W X Y Z

The core alphabet consists of 20 letters used for expressing consonants and vowels in Hiligaynon, each of which comes in an upper case and lower case variety. The Hiligaynon Alphabet and their sounds is shown in Table 1. The researcher has integrated the Google's

Filipino dictionary in the system for it to be able to decode Hiligaynon speeches. For every Hiligaynon word spoken, the system associates the best letter or combination of letters and displays that on the screen. For every word spoken, the system also associates the best letter or combination of letters to the phonemes or sound decoded. The system connects to the database of Google for this association. When the phoneme is not found in the Google database of letters, the system displays an error message. However, the system displays the best match possible if there is.

Table 1. The Hiligaynon Alphabet and Their Sounds [19]

The 1st to 10th letters													
Symbol	A a			B b	K k	D d	E e	G g	H h	I i		L l	M m
Name	a			ba	ka	da	e	ga	ha	i		la	ma
Pronunciation	[a/ə]	[aw]	[aj]	[b]	[k]	[d]	[ɛ/e]	[g]	[h]	[ɪ/i]	[ɪo]	[l]	[m]
in context	a	aw/ao	ay	b	k	d	e	g	h	i	iw/io	l	m

The 11th to 20th letters													
Symbol	N n	Ng ng	O o		P p	R r	S s		T t	U u		W w	Y y
Name	na	nga	o		pa	ra	sa		ta	u		wa	ya
Pronunciation	[n]	[ŋ]	[ɔ/o]	[oj]	[p]	[r]	[s]	[ʃ]	[t]	[ʊ/u]	[w]	[w]	[j]
in context	n	ng	o	oy	p	r	s	sy	t	u	ua	w	y

3.2.2. The Phonemes in the English Language

The English language has 26 letters, however, there are 44 unique sounds, also known as phonemes. The phonemes differentiate one word or meaning from another. These phonemes fall into two major categories namely consonants and vowels.

As illustrated, "A consonant's sound is one in which the air flow is cut off, either partially or completely, when the sound is produced. In contrast, a vowel sound is one in which the air flow is unobstructed when the sound is made. The vowel sounds are the music, or movement, of our language" [20]. Table 2 and 3 show the 44 phonemes which are in line with the International Phonetic Alphabet.

Table 2. List of English Consonants

Sound	Graphemes	Examples
/b/	b, bb	bug, bubble
/d/	d, dd, ed	dad, add, milled
/f/	f, ff, ph, gh, lf, ft	fat, cliff, phone, enough, half, often
/g/	g, gg, gh,gu,gue	gun, egg, ghost, guest, prologue
/h/	h, wh	hop, who
/j/	j, ge, g, dge, di, gg	jam, wage, giraffe, edge, soldier, exaggerate
/k/	k, c, ch, cc, lk, qu ,q(u), ck, x	kit, cat, chris, accent, folk, bouquet, queen, rack, box
/l/	l, ll	live, well
/m/	m, mm, mb, mn, lm	man, summer, comb, column, palm
/n/	n, nn,kn, gn, pn	net, funny, know, gnat, pneumonic
/p/	p, pp	pin, dippy
/r/	r, rr, wr, rh	run, carrot, wrench, rhyme
/s/	s, ss, c, sc, ps, st, ce, se	sit, less, circle, scene, psycho, listen, pace, course
/t/	t, tt, th, ed	tip, matter, thomas, ripped

/v/	v, f, ph, ve	vine, of, stephen, five
/w/	w, wh, u, o	wit, why, quick, choir
/y/	y, i, j	yes, onion, hallelujah
/z/	z, zz, s, ss, x, ze, se	zed, buzz, his, scissors, xylophone, craze

Table 3. List of English Vowels [20]

Sound	Graphemes	Examples
/a/	a, ai, au	cat, plaid, laugh
/ā/	a, ai, eigh, aigh, ay, er, et, ei, au, a_e, ea, ey	bay, maid, weigh, straight, pay, foyer, filet, eight, gauge, mate, break, they
/e/	e, ea, u, ie, ai, a, eo, ei, ae, ay	end, bread, bury, friend, said, many, leopard, heifer, aesthetic, say
/ē/	e, ee, ea, y, ey, oe, ie, i, ei, eo, ay	be, bee, meat, lady, key, phoenix, grief, ski, deceive, people, quay
/i/	i, e, o, u, ui, y, ie	it, england, women, busy, guild, gym, sieve
/ī/	i, y, igh, ie, uy, ye, ai, is, eigh, i_e	spider, sky, night, pie, guy, stye, aisle, island, height, kite
/o/	o, a, ho, au, aw, ough	octopus, swan, honest, maul, slaw, fought
/ō/	o, oa, o_e, oe, ow, ough, eau, oo, ew	open, moat, bone, toe, sow, dough, beau, brooch, sew
/oo/	o, oo, u,ou	wolf, look, bush, would
/u/	u, o, oo, ou	lug, monkey, blood, double
/ū/	o, oo, ew, ue, u_e, oe, ough, ui, oew, ou	who, loon, dew, blue, flute, shoe, through, fruit, manoeuvre, group
/y//ü/	u, you, ew, iew, yu, ul, eue, eau, ieu, eu	unit, you, knew, view, yule, mule, queue, beauty, adieu, feud
/oi/	oi, oy, uoy	join, boy, buoy
/ow/	ow, ou, ough	now, shout, bough
/ə/ (schwa)	a, er, i, ar, our, or, e, ur, re, eur	about, ladder, pencil, dollar, honour, doctor, ticket, augur, centre, chauffeur

This present study uses the lists of consonants and vowels in identifying the phoneme of each word in the speech being entered into the system.

4. Implementation and Evaluation Results

4.1. Implementation

The system accepts speech recorded in real time or speech that was pre-recorded. The system saves the speech in .wav format in a 10-second chunk size recording. For longer speech, several .wav files may be generated. The system transcribes the recorded speech and displays it back to the user. Also, it generates a .text format of the transcribed text and saves that in a folder for reference. The software is capable of transcribing English, Tagalog, Hiligaynon, and Ilonggoish spoken speeches, however, the accuracy of transcription may depend on factors such as the quality of the microphone used and the diction of the speaker.

The main screen is shown in Figure 2, where the system is ready to accept speech input from the user. The “Ready” message signifies the current system state. The red button is the Record button. Once the user starts recording, the red button must be clicked.

Figure 2. The Main Screen Waiting for the Speech Input

Figure 3 is the when the user is currently recording speech input or starting recording. The black button is highlighted and .wav files are automatically saving the recorded file with its sequence of utterances. The speech input is divided into chunks that last for 10 seconds. The longer the speech, the more .wav files are generated.

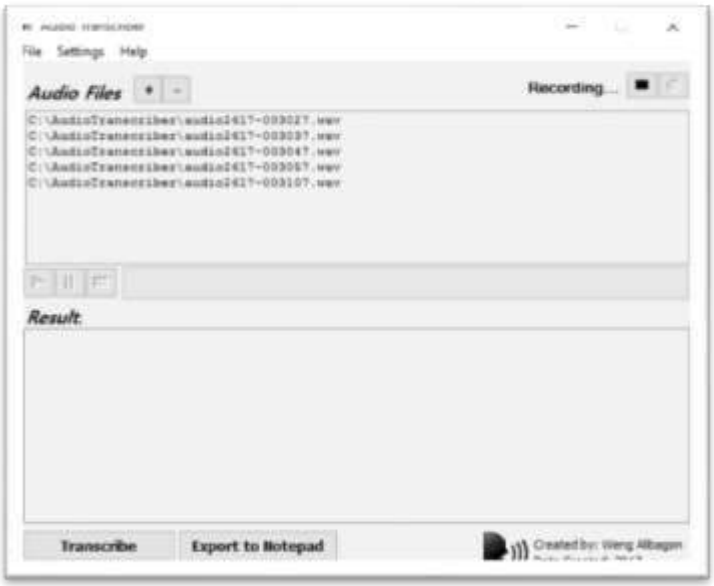

Figure 3. The System While Recording

Figure 4 is the screen when the user chooses to play back a chunk or several chunks of recorded speech. The user has option to playback captured audio. The green-colored status motion is the indication of the frequency coming from the wave signal produced by the microphone from the speaker.

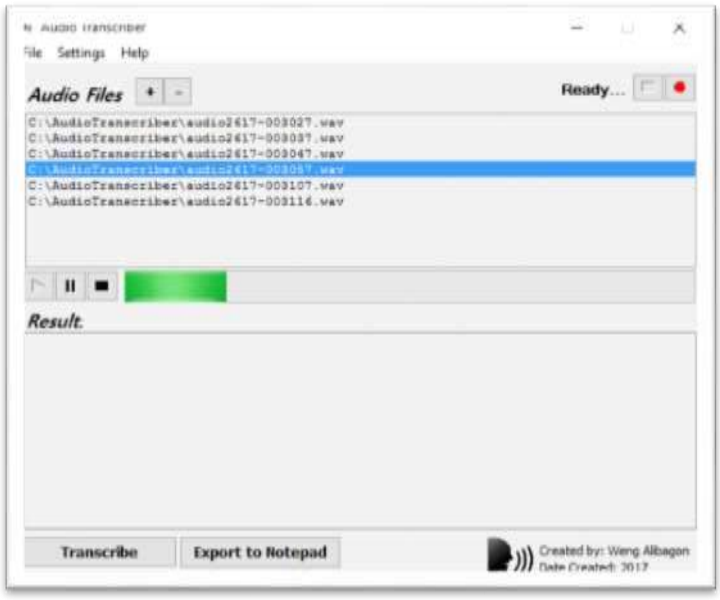

Figure 4. The Play Back Recording Screen

Figure 5 shows the screen when the transcription is currently processing a particular .wav file. To do this, the user must select the desired .wav file and click the Transcribe button. A dialogue box with the display "Processing wave file" appears to let the user know of the current system state.

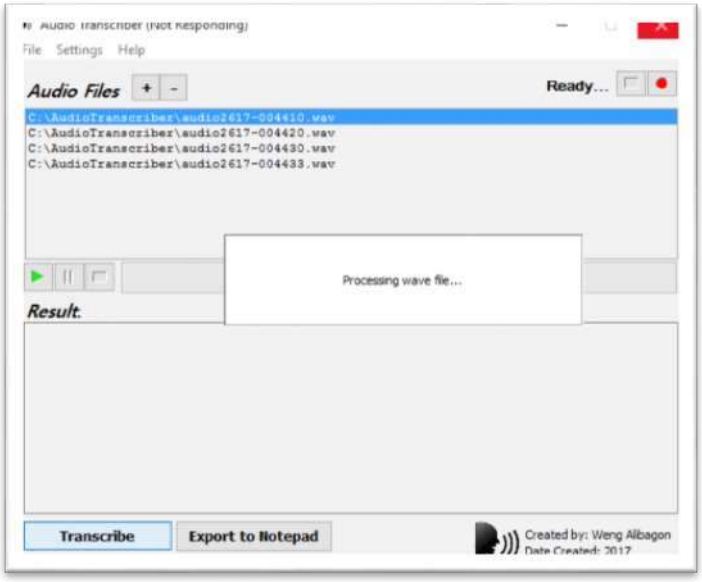

Figure 5. The System during the Transcription Process

Shown in Figure 6 is the screen that displays the transcribed speech input under the Result screen. The system is able to convert the audio format to text format.

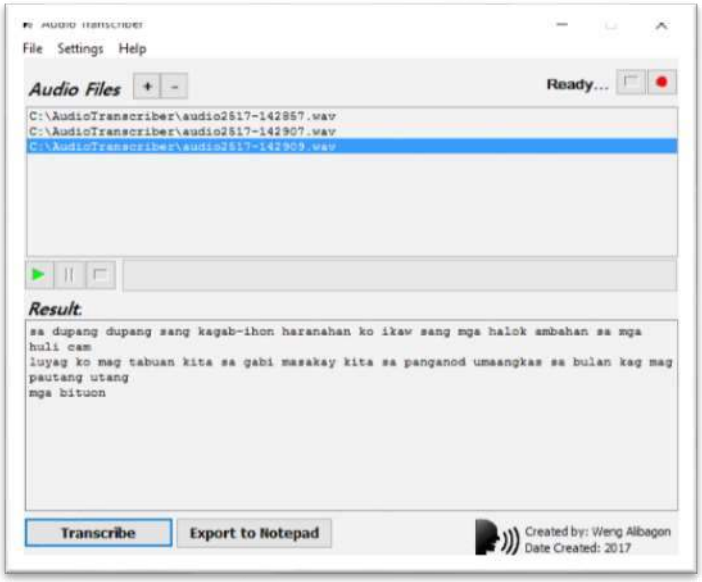

Figure 6. The System Displaying the Transcribed Speech

4.2. Evaluation

The accuracy test of the system was done by six (8) linguists, specifically, two (2) specializing in English Language, two (2) specializing in Filipino language, two (2) in Hiligaynon dialect, and two (2) for Ilonggoish dialect. For each language, three (3) accuracy tests were conducted to see how the system could accurately recognize phonemes. The same microphone was used during the conduct of the study.

English Language. The result, as shown in Table 5, revealed that the system has an accuracy rate of 90% for English language. Specifically, Test 1 has an accuracy rate of 91%, Test 2 has an accuracy rate of 88%, and Test 3 has an accuracy rate of 89%. This result signifies that the system has a "medium-high" accuracy in transcribing English spoken speeches into text format. The result denotes that the system has an average performance when the speech input was spoken using the English language.

Table 5. The Accuracy Test Result for English Language

Test No.	Accuracy Rate Respondent 1	Accuracy Rate Respondent 2	Accuracy Rate Mean	Interpretation
1	93%	89%	91%	High Accuracy
2	85%	91%	88%	Medium-High Accuracy
3	98%	80%	89%	Medium-High Accuracy
Mean	93%	86%	90%	Medium-High Accuracy

Filipino Language. The result, as shown in Table 6, revealed that the accuracy rate for Filipino language is 95%. Test 1 has an accuracy rate of 92%, Test 2 has an accuracy rate of 96%, and Test 3 has an accuracy rate of 95%. The result proves that the system has a high level of accuracy rate in transcribing Filipino spoken speeches into text format. This connotes that the algorithms used in the system have a very efficient performance in finding the best matched words from the Filipino dictionary. Also, this implies that the system can be a very effective transcription tool for Filipino language.

Table 6. The Accuracy Test Result for Filipino Language

Test No.	Accuracy Rate Respondent 1	Accuracy Rate Respondent 2	Accuracy Rate Mean	Interpretation
1	93%	91%	92%	High Accuracy
2	98%	94%	96%	High Accuracy
3	95%	96%	95%	High Accuracy
Mean	96%	94%	95%	High Accuracy

Hiligaynon Dialect. The result, as shown in Table 7, revealed that the system has an accuracy rate of 77% in Hiligaynon speeches. Specifically, Test 1 has an accuracy rate of 76%, Test 2 has an accuracy rate of 82%, and Test 3 has an accuracy rate of 75%. The result means that the system can transcribe Hiligaynon speeches with "medium to high" accuracy. As compared to both English and Filipino, the accuracy rate of Hiligaynon transcription has decreased for the reason that the Google's API does not contain a Hiligaynon dictionary. However, the result implies that the system has the capacity to look up from both Filipino and English dictionary to come up with the best matched Hiligaynon words. Moreover, this implies that the speech recognition capacity of the system is powered by the combination of dictionaries present, thus, it is not limited only to a certain language.

Table 7. The Accuracy Test Result for Hiligaynon Dialect

Test No.	Accuracy Rate Respondent 1	Accuracy Rate Respondent 2	Accuracy Rate Mean	Interpretation
1	70%	83%	76%	Medium - High Accuracy
2	85%	80%	82%	Medium - High Accuracy
3	75%	75%	75%	Medium - High Accuracy
Mean	75%	78%	77%	Medium - High Accuracy

Ilonggoish Dialect. The result, as shown in Table 8, revealed that the system has the accuracy rate of 88%. The result of Test 1 has an accuracy rate of 91%, Test 2 has an accuracy rate of 88%, and Test 3 has an accuracy rate of 86%. The result implies that the system can transcribe Ilonggoish dialect with a "medium to high" accuracy. The result indicates that the system has a considerable performance when it comes to transcribing Iloggoish dialect. It denotes further that it can switch from English dictionay to Filipino dictionary and vice versa, and even combine both to come up with the best matched words.

Table 8. The Accuracy Test Result for Ilonggoish Dialect

Test No.	Accuracy Rate Respondent 1	Accuracy Rate Respondent 2	Accuracy Rate Mean	Interpretation
1	94%	89%	91%	High Accuracy
2	92%	84%	88%	Medium - High Accuracy
3	86%	86%	86%	Medium - High Accuracy
Mean	90%	86%	88%	Medium - High Accuracy

Summary of the Accuracy Test Result. The result, as shown in Table 9, revealed that the system has an over-all accuracy rate of 87%. Specifically, the two highest accuracy ratings were found to be in English and Filipino languages with accuracy rates of 90%and 95% respectively. The lowest accuracy rate was found to be Hiligaynon dialect which is 77%. The Ilonggoish dialect has an 88% accuracy rate.

Table 9. The Summary of the Accuracy Test Result of the System

Language / Dialect	Accuracy Rate	Interpretation
English	90%	Medium to High Accuracy
Filipino	95%	High Accuracy
Hiligaynon	77%	Medium to High Accuracy
Ilonggoish	88%	Medium to High Accuracy
Over-all Accuracy Rate	87%	Medium to High Accuracy

The result means that the system, in general has a "medium to high" accuracy rate. Moreover, the results are understandable because the system uses both English and Filipino dictionaries. For Hiligaynon dialect, the combination of these dictionaries was used which lessened the accuracy rate of the HMM algorithm to find the best matched words. For Ilonggoish dialect, which is a combination of English and Hiligaynon, only the portion of the speech that has the Hiligaynon words are matched with the two dictionaries, consequently, the accuracy rate remains average.

The result establishes that the system having an average performance in transcribing speeches spoken in English, Filipino, Hiligaynon, or combination of these languages can be trusted to transcribe accurately. Nevertheless, the accuracy of the system, regardless of the language used, is largely based on the quality of the microphone used to input speeches, and the diction of the speaker.

5. Conclusion

The accuracy test of the system revealed that the system has an 87% accuracy rate interpreted as "medium to high" which signifies that the system has the capacity to accurately transcribe speeches spoken in English, Filipino, Hiligaynon and Ilonggoish. However, the accuracy of the system is largely dependent on the quality of the recorded speech which may be affected by the quality of the microphone used during the recording and the diction of the speaker. As compared to both English and Filipino, the accuracy rate of Ilonggoish transcription is a little low for the reason that the Google's API does not contain a Hiligaynon dictionary. However, the result implies that the system has the capacity to look up from both Filipino and English dictionary to come up with the best matched Hiligaynon words. Moreover, this implies that the speech recognition capacity of the system is powered by the combination of dictionaries present, thus, it is not limited only to a certain language.

As future work, the length of a recorded audio chunk can be extended to provide longer speech recording. In this current research, the quality of the transcription was affected by the length of the chunk. The challenge to future researchers is to extend the length of the recorded chunk without compromising transcription quality. In addition, the system can be made into a machine learning system wherein the system's database grows and evolves according to new observations, classifications, and training data.

References

[1] S. Furui, "40 Years of Progress in Automatic Speaker Recognition", *Proc. of the ICB 2009 Conf,* **(2009)**, pp. 1050-1059.

[2] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee and F. Pernkopf, "Phase-Aware Signal Processing for Automatic Speech Recognition", 3374-3378. 10.21437/Interspeech, **(2016)**, pp. 2016-823.

[3] S. Renals, "Words: Pronunciations and Language Models", URL https://www.inf.ed.ac.uk/teaching/courses/asr/2011-12/asr-lexlm-nup4.pdf. March 2012, Date retrieved January 15, 2017.

[4] P. Arulmozhi. R.K. Rao and 1. Sobha, "A Hybrid POS Tagger for a Relatively Free Word Order Language", in Proceedings of the Modeling and Shallow Parsing of Indian Language (MSPIL), Bombay, **(2006)**, pp. 79-85.

[5] K. Nagaraj, U. Swant, S. Shelke and P. Bhattacharyya, "Building Feature Rich POS Tagger for Morphologically Rich Languages: Experience in Hindi", in Proceedings of ICON, India, 2007, pp. 201-208.
[6] A. M. Othman and M. H. Riadh, "Speech Recognition Using Scaly Neural Networks", in World Academy of Science, Engineering and Technology, vol. 38, **(2008)**, pp. 187-198.
[7] M. El Choubassi, H. El Khoury, C. Alagha, J. Skaf and M. AlAlaoui, "Arabic Speech Recognition Using Recurrent Neural Networks", IEEE Intl. Symp. Signal Processing and Information Technology ISSPIT, vol. 3, no. 1, **(2003)**, pp. 336-340.
[8] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals", in Report of the 5th Int. Congress on Acoustics, **(1968)**, pp 209-215.
[9] A. M. Toh, "Feature Extraction for Robust Speech Recognition in Hostile Environments", in PhD. thesis, Dept. Elect. Eng., Western Australia Univ.,Australia, **(2007)**.
[10] R. Valenza, T. Robinson, M. Hickey and R. Tucker, "Summarization of Spoken Audiothrough Information Extraction", in Proc. ESCA Workshop on Accessing Information in Spoken Audio, **(1999)**.
[11] K. Knight and D. Marcu, "Statistics-Based Summarization — Step One: Sentence Compression", in Proc. National Conference on Artificial Intelligence (AAAI).
[12] E. Essa, A. Tolba and S. Elmougy, "A Comparison of Combined Classifier Architectures for Arabic Speech Recognition", in the Proceedings of the 2008 IEEE International Conference on Computer Engineering & Systems, Cairo, **(2008)**.
[13] E. Essa, A. Tolba and S. Elmougy, "Combined Classifier Based Arabic Speech Recognition", in the Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008), Cairo, **(2008)**.
[14] A. S. Bregman, "Auditory Scene Analysis: The Perceptual Organization of Sound", The MIT Press, **(1994)**.
[15] D. P. W. Ellis. "Prediction driven computational auditory scene analysis for dense sound mixtures", In ESCA workshop on the Auditory Basis of Speech Perception, Keele, UK, **(1996)**.
[16] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation", in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, **(1999)**.
[17] M. M. Goodwin and J. Laroche, "Audio segmentation by feature-space clustering using linear discriminant analysis and dynamic programming", in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, **(2003)**.
[18] E. W. Weisstein, "Power Spectrum, "From MathWorld--A Wolfram Web Resource", http://mathworld.wolfram.com/PowerSpectrum.html, Date retrieved August 4, 2017.
[19] S. Ager. "Hiligaynon (Ilonggo)", URL http://www.omniglot.com/writing/hiligaynon.htm. Date retrieved August 5, 2017.
[20] "The 44 Phonemes in English", URL http://www.dyslexia-reading-well.com/44-phonemes-in-english.html. N.d. Date retrieved August 5, 2017, 2017.

Authors

Rowena Alibagon, she received her B.S. Degree in Computer Engineering from University of San Agustin, Philippines, in 2001, and M.S. Degree in Computer Science from Iloilo Science and Technology University, Philippines, in 2017. Her 5 years of experience in the industry and affiliation on Social Communication Commission of the Archdiocese of Jaro, Iloilo City led her to this research.

Frank I. Elijorde, he received his B.S. degree in Information Technology and M.S. degree in Computer Science from Western Visayas College of Science and Technology (now Iloilo Science and Technology University), Philippines, in 2003 and 2007 respectively. He received his Ph.D. in Information and Telecommunications Engineering from Kunsan National University, Korea, in 2015. Currently, he is an Associate Professor at the College of ICT in West Visayas State University, Iloilo City, Philippines. His research interests include distributed systems, cloud systems, data mining, ubiquitous computing, and context-aware systems.

Joel T. De Castro, he is currently the Dean and holds a Professor 6 position at the College of ICT in West Visayas State University, Iloilo City, Philippines. He received his B.S. degree in Biology from Cebu Doctors University, in 1981, Master of Science in Computer Science from Western Visayas College of Science and Technology (now Iloilo Science and Technology University) in 2005, and Doctor of Industrial Technology from Iloilo Science and Technology University in 2013. His research interest lies in the area of IT security and Protection, Distributed Systems, and Machine Intelligence and SMS Systems.

Yungcheol Byun, he is a full professor at the Computer Engineering Department (CE) at Jeju National University. His research interests include the areas of Pattern Recognition & Image Processing, Signal Processing & Security, Intelligent Computing, Semantic Web and Ontology, Home Network and Ubiquitous Computing, Healthcare, RFID & IoT Middleware, and Artificial Intelligence. Outside of his research activities, Dr. Byun has been hosting international conferences, CNSI (Computer, Network, Systems, and Industrial Engineering), ICESI (Electric Vehicle, Smart Grid, and Information Technology), and also serving as a conference and workshop chair in various kinds of international conferences and workshops. He received his Ph.D. from Yonsei University in 2001. Before joining the current university, he was a senior researcher of Electronics and Telecommunications Research Institute (ETRI).