MALAYSIAN ENGLISH LARGE VOCABULARY CONTINUOUS SPEECH RECOGNIZER: AN IMPROVEMENT USING ACOUSTIC MODEL ADAPTATION

(Date received: 13.01.2022/Date accepted: 09.05.2022)

Kah Chung Yoong[1], Kai Sze Hong[1]

*[1] Department of Electrical & Electronic Engineering, Faculty of Engineering & Technology,
Tunku Abdul Rahman University College, Kuala Lumpur, Malaysia.*

**Corresponding author: yoongkc-wg17@student.tarc.edu.my, hongks@tarc.edu.my*

ABSTRACT

This research project aims to develop Malaysian English Continuous Speech Recognition system by adapting US English acoustic model with Malaysian English speech corpus using Maximum a posteriori reasoning (MAP) and Maximum Likelihood Linear Regression (MLLR). During feature extraction stage, the Mel-Frequency Cepstral Coefficients (MFCC) technique was used. The Hidden Markov Model was used as the back end pattern comparison technique. For the purpose of implementation, the CMU Sphinx toolkit, which includes Pocketsphinx and Sphinxtrain as well as an acoustic model, was used to develop a speech recognition system for Malaysian English. Malaysian English speech samples were recorded and transcribed to produce the training database required for acoustic model adaptation. The adaptation speech corpus were collected from a number of speakers. The outcome of this research could increase the application of Malaysian English speech recognition in Malaysia due to accent problem. As a result, speech recognition systems that have gone through the MAP adaptation had the best performance. Its average word error rate achieved was 32.84%. Average word recognition rate was 72.52% and average sentence error rate was 78.89%.

Keywords: Speech Recognition, Acoustic Model, MAP, MLLR, Pocketsphinx

1.0 INTRODUCTION

Speech is a form of language-based individual vocalization. The audio within each linguistic vocabulary is formed by pronunciation arrangements of vowel and consonant tones. While using several terms in their linguistic context as vocabulary in a linguistic dictionary in accordance with the grammatical restrictions that regulate the role of lexical parts of speech (Houghton Mifflin, 2015). Speakers may use pronunciation, inflection, volume of voice, rhythm, and several other non-representational or vocabulary elements of vocalization to express meaning in several specific deliberate speech acts, such as telling, announcing, questioning, convincing, and guiding. Speakers unwittingly express many facets of their social status in their speeches, including gender, aged, point of birth, cognitive abilities, psychological condition, physio-psychic condition, background, or knowledge, etc. Several components of speech are investigated by researchers, including speech processing and voice detection (Houghton Mifflin, 2015). Voice duplication, voice defects, and the failure to translate hearing speaking terms into the vocalizations required to reproduce them are all examples of these. This is important for kid's vocabulary development and mind development in various fields. Sociology, theoretical physics, information science, sociology, software engineering, forensic linguistics, ophthalmology, and sound systems are all fields that research speech. Speech is related to linguistic knowledge, which can distinguish from spoken language in terms of pronunciation, grammar, and phonology, a disorder known as diglossia (Houghton Mifflin, 2015; NIDCD, 2021).

The conversion of individual voice signals into vocabulary or commands is known as speech recognition. Speech recognition is dependent on the sound of a person's voice. That is a subdivision of information processing and a significant research path in speech signal processing. Software engineering, machine learning, digital signal processing, information processing, sound systems, linguistics, and cognitive science all play a role in speech recognition study. It's an interdisciplinary, all-encompassing field of study. One of the research objectives are to improve the speech recognition accuracies. Besides, speech recognition systems are designed to work in either a constrained environment or an opened environment. Different applications have different requirements of specific type of speech recognizer. Based on research objectives and restrictions, numerous research areas have arisen. These fields could be separated into isolated words, connected words, and continuous speech recognition systems, depending upon the needs of the presenter's style of communicating. The above fields could be classified into speech recognition systems for individuals and non - specific persons depending on the level of reliance on the person speaking (Science Direct, 2021). These could be categorized into small vocabulary, medium vocabulary, big vocabulary, and infinite vocabulary speech recognition systems based also on scale of their pronunciation. The concept of the voice recognition model is predicated on information processing, according to the speech recognition model. That aim of speech recognition is to use phonology and textual data to convert a received signal feature vector pattern into a series of text. A full speech recognition system involves information extraction technique, the acoustic template, a language model, and a search method, as according to configuration of a speech recognition system. A multifaceted information processing system is exactly what a speech recognition system is. Researchers use different recognition methodologies for specific speech recognition systems. However, the simple concepts are the same. Function abstraction is applied to the obtained speech recognition. The template database system collects and processes the speech information received by the system. The voice information retrieval module recognizes speech samples based on the template database, and then calculates the segmentation results (Science Direct, 2021).

The overall organization of this paper is described here. Firstly, section 2 describes the literature review of Malay LVCSR system. Next, section 3 elaborates the research methodology. Section 4 shows the results and discussions. Finally, section 5 concludes the whole research work.

2.0 LITERATURE REVIEW

Speech recognition systems allow a computer to carry out human-spoken commands, perform automatic translation, and generate print-ready dictation. The microphone receives the sound signal, which is then transformed into a digital signal by the system hardware. The analysis tool of a speech recognition device uses the produced digital signal as input to extract and recognise the differentiated phoneme. Several of the lowest units of sound is the phoneme, which distinguishes the sound of one word from that of another. Nevertheless, most of the voice of the phrases are similar. As a result, the software must depend on context to distinguish the right punctuation among all these similar sounding words (Abhang *et al.*, 2016).

2.1 Process of Speech Recognition

Those two most popular strategies to speech recognition could be classified as "template matching" and "feature analysis." Template matching is the most efficient method with the highest precision, but it often would have the most drawbacks. That individual should first say a word or statement into a mic, just like in any technique to speech recognition. An "analogue-to-digital (A/D) converter" transforms the electric signals from the receiver into digital form, which is then loaded into memory. This approach is somewhat similar to using a button to enter commands. That software includes an input framework and provides a simple constraint declaration to try and adapt the model to the real input (HITL, 2021). The overview of process of speech recognition general framework is showed in Figure 1:

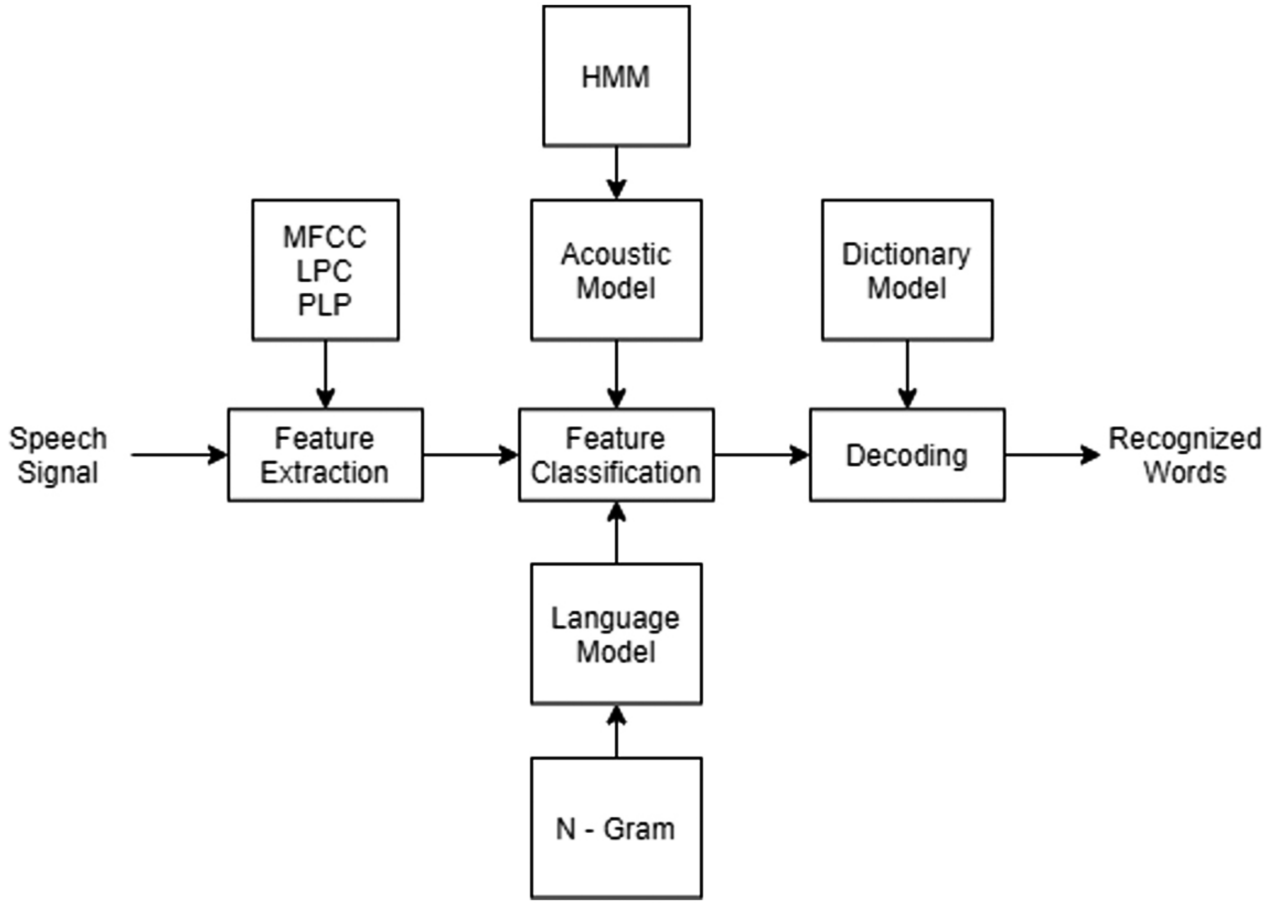


Figure 1: Process of Speech Recognition

Figure 1 depicts the general framework of the speech recognition process. Speech signal features are extracted using feature extraction techniques like MFCC, LPC and PLP. Next, features that have been extracted will be classified using acoustic modelling and language modelling. Finally, during the recognition stage, the features are decoded with the aid of a dictionary model to produce the recognized words.

2.2 Feature Extraction

The key component of a speech recognition program is feature extraction. These are regarded as the program's beating heart. The main objective is to extract features through the source voice signal which will aid the device in recognising the user. That intensity of the source signals is compressed by feature extraction without affecting the voice signal's strength. Furthermore, it attempts to minimise the loss of data kept among the terms during this point. That aids throughout the consistency comparison of its acoustic model's distributional assumption.

Mel frequency cepstral coefficients (MFCC) were first proposed for recognizing idiomatic phrases in consistently voiced statements, but not for determining the user. The MFCC algorithm is really a simulation of the living thing listening process that serves to theoretically enforce the ear's working theory, assuming also that living person ear is really an effective voice recognition system. That MFCC models are based on a known difference between the essential bandwidths of the living person ear and frequency filtering distributed sequentially at lower frequency. That pronunciation critical characteristics of the voice signal were preserved by doing this logarithmically at higher frequency. Sounds of various frequencies are generally used in voice signal, with every voice having its own frequency. The Mel measurement is used to calculate arbitrary pitch. Approximately 1000 Hz, that Mel-frequency level contains linear frequency spaced, and over 1000 Hz, these have linear interpolation frequency width. The 1000 meals was its pitch of even a 1 kHz voice at 40 dB over the perceived detectable level, which is applied as a level of comparison (Alim, 2018; Vergin, R. *et al.,* 1999; X. Zhou, D. Garcia-Romero, R. Duraiswami *et al.*, 2011; Muda, L *et al.*, 2021).

That MFCC algorithm depends upon signal dissolution using a wiener filter. That MFCC generates a discrete cosine transform (DCT) of even a particular logarithm of its simple terms power also on Mel frequency range. Under safety purposes, the MFCC is being used to classify travel arrangements, contact information, and speech recognition. With improved robustness, several improvements to the simple MFCC methodology are being suggested. Besides instance, while implementing the DCT, raise the log-mel-amplitudes to a suitable capacity. That alone lessens the detrimental effects of low-energy materials (Alim, 2018; Vergin *et al.*, 1999; Zhou *et al.*, 2011; Muda *et al.*, 2021).

The Mel Frequency Cepstrum parameters of MFCC are constructed from a distorted frequency range focusing on individual perceptual experience. Its first stage in MFCC processing is windowing the voice signal which divides these into layers. Although the high frequency formants method would have a lower amplitude than the low frequency formants, the high frequencies are stressed in order to achieve a comparable amplification for both formants. The energy spectrum within each frame is determined using the Fast Fourier Transform (FFT) upon windowing. Following that, Mel-scale filter system work is performed on the energy spectrum. In order to determine MFCC parameters, the DCT is added to the voice signal once the energy spectrum being converted to log field (Alim, 2018; Vergin *et al.*, 1999; Zhou *et al.*, 2011; Muda *et al.*, 2021).

$$mel(f) = 2595\ x\ log_{10}\ (\ 1 + \frac{f}{700}) \qquad \textit{Eq. 1}$$

Where Mel (f) represents the mel scale of frequency and f represents the frequency (Hz).

The below formula is used to measure the MFCCs.

$$C_n = \sum_{n=1}^{k} (\log S_k) \cos\ [n\left(k - \frac{1}{2}\right)\frac{\pi}{k}] \qquad \textit{Eq. 2}$$

The k defines the number of Mel Cepstrum parameters S the filter bank production, and C the last MFCC approximation.

Figure 2 shows the block diagram including its MFCC module. That low frequency area could be essentially denoted by MFCC, while the high frequency field can indeed be efficiently denoted by MFCC. This one can approximate and define vocal folds resonances using formants throughout the low frequency region. That is widely acknowledged as a front-end technique for popular Speech Recognition implementations. These have minimised noise disruption uncertainty, provides minute’s process instability, and is simple to develop. Often, whenever the useful in the evaluation are balanced and coherent expression, it is a good demonstration for voices. It could also derive data through processed signals with a peak frequency of 5 kHz. This covers the bulk of the power found in human-made voices (Alim, 2018; Vergin *et al.*, 1999; Zhou *et al.*, 2011; Muda *et al.*, 2021).

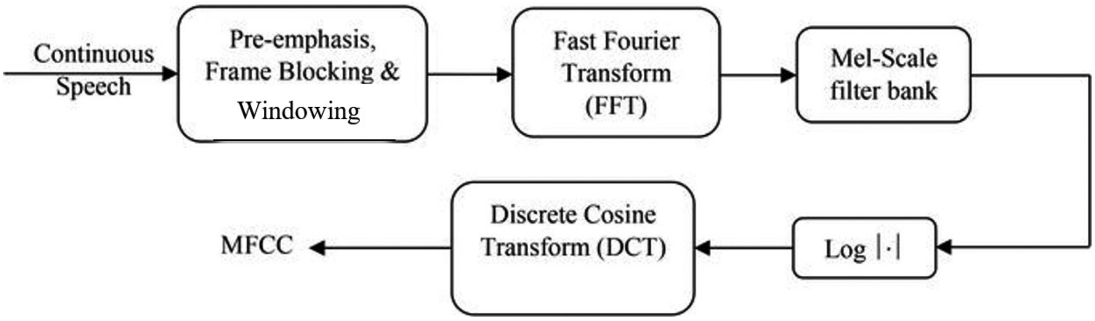

Continuous Speech
Pre-emphasis, Frame Blocking & Windowing
Fast Fourier Transform (FFT)
Mel-Scale filter bank
Log |·|
Discrete Cosine Transform (DCT)
MFCC

Figure 2: Block Diagram of MFCC

In several pattern recognition difficulties concerning living person speech, cepstral coefficients are shown to be correct. They are commonly utilized in numerous speech recognition and presenter detection. Those certain formants should be over 1 kHz, and indeed the wide filter width throughout the high frequency region does not effectively account for them. Throughout the presence of ambient noise, MFCC functions may not always be reliable, and they are not very well equipped for generalisation (Alim, 2018; Vergin *et al.*, 1999; Zhou *et al.*, 2011; Muda *et al.*, 2021).

2.3 Feature Classification

Within various conditions, a classification model is being used to determine the right voice speaking style. The strong pattern classification scheme is being trained through appropriately labelled illustrations in speech recognition.

The Hidden Markov Model (HMM) is a mathematical method that might be used to explain how measurable phenomena evolve over time (Rabiner and Juang, 1986). The observable element influencing the occurrence is referred to as a 'state,' whereas the observable phenomenon is referred to as a 'symbol'. The unidentifiable procedure of unidentified conditions and the transparent methodology of measurable signs are the two stochastic systems that make up an HMM. That invisible conditions construct a Markov chain, and perhaps the detected sign's probability distribution will be determined by the corresponding system's probability distribution. Mostly as result, a doubly-embedded random variable is another name about an HMM (Rabiner and Juang, 1986).

It is sufficient to define findings within those 2 phases, one transparent and the other hidden. Although several real-world problems include categorising unstructured data into a set of classification or membership functions. That matters to them as well. Remember the issue of voice recognition, where this HMMs had already long been shown. Forecasting the spoken phrase from such a registered voice signal will be the aim of speech recognition. According to interpretations, the voice recognition system attempts to locate the series of consonant conditions that resulted in the real spoken voice. Although real transcription can vary greatly, the basic phonetic symbols cannot be detected explicitly and must be expected (Yoon, 2009; Cuiling, 2016; Xue, 2018).

There are also other research work related to Hidden Markov Models by other researchers (Aymen *et al.*, 2011; Gales, 2009; Abushariah *et al.*, 2010).

2.4 N - Gram

The N-gram is a texting-gram structure in which the "N" objects from a provided series or series are continuously arranged. Text mining, communication theory, text probability, and data compression are only a few of the areas where it is used. When assigning with varying probabilities, the N-gram is useful. That is because it assists in deciding which N-grams have the greatest chance of chunking together to create a single entity (Gadag and Sagar, 2016; Takahashi and Morimoto, 2012).

The N-gram analysis shows the appearance of an interpretive structural on the appearance of N-1 preceding terms. Unigram (N), bigram (N=2), trigram (N=3), and so on are all examples of N-grams. Bigram model N=2 predicts a term occurring based on the previous single word (N-1) and bigram model N=3 forecasts a term phenomenon based on the previous two terms (N-2) (N-2) (Ito and Kohda, 1996; Hatami, Akbari and Nasersharif, 2013).

2.5 Maximum Likelihood Linear Regression (MLLR)

Maximum Likelihood Linear Regression (MLLR) would be the first adaptation algorithm inspired. Through analysing that acoustic model's mean and variance, MLLR can predict and determine that optimum probability distribution and characteristics. Maximum A-Posteriori (MAP) has been the second adaptation approach applied. MAP performs nearly the similar purpose as MLLR, with the exception that it considers the prior throughout the prediction. The mixture weight and the transition matrices are also observed by MAP in addition to the acoustic model's variance and means. MAP updates the characteristics in the acoustic model, contrasting MLLR, which just generates a matrix that could be passed to the system upon decoding. Furthermore, while MAP appears to become a preferable adaptation approach, that alone necessitated a large amount of data for adaptation in attaining the desired prediction performance. Although MLLR proved possible to provide a noticeable enhancement in recognition rate with small dataset, two adaption methods were used in this study, one simultaneously and one individually, to examine respective impact on speech recognition application efficiency. Research works related to MLLR have been done by some researchers (Lestari and Irfani, 2015; Oh *et al.*, 2007).

2.6 Novelty of this Research Work

There were no work previously done in the area of adapting Malaysian English to English LVCSR. The reason is most speech recognition research works are focusing on other languages. Thus this research project helps to study the possible accuracies when Malaysian English is adapted to English LVCSR.

3.0 METHODOLOGY

3.1 Speech Recognition Process

Figure 3 depicts the flow chart for the overall system process. This overall system process is separated into 2 stages including as adapting stage and decoding stage. For adapting stage, 50 individual phrases of utterances by five separate speakers (Anonymous FYP students) may also being used to adjust the Malaysian English language acoustic model utilising two adaptation strategies, MLLR and MAP. At first, all the speech sample will be collected and convert them into Mel Frequency Cepstral Coefficients (MFCC) format by undergoes the feature extraction. This feature extraction process will extract the main speech information and critical characteristic that is required for the system. So, all the MFCC files with extracted feature are produced. These MFCC files, list of transcripts of the adaption speech samples, list of speech sample filenames, and acoustic model were used in the adaptation.

Prior beginning the decoding stage, those speech samples utilized to analyse the results were first feature extracted by MFCC. This speech recognition system then uses the created MFCC files, including comprise all extracted feature, customised acoustic model, MLLR matrices, language model, and dictionary model, to conduct speech recognition. After the decoding stage,

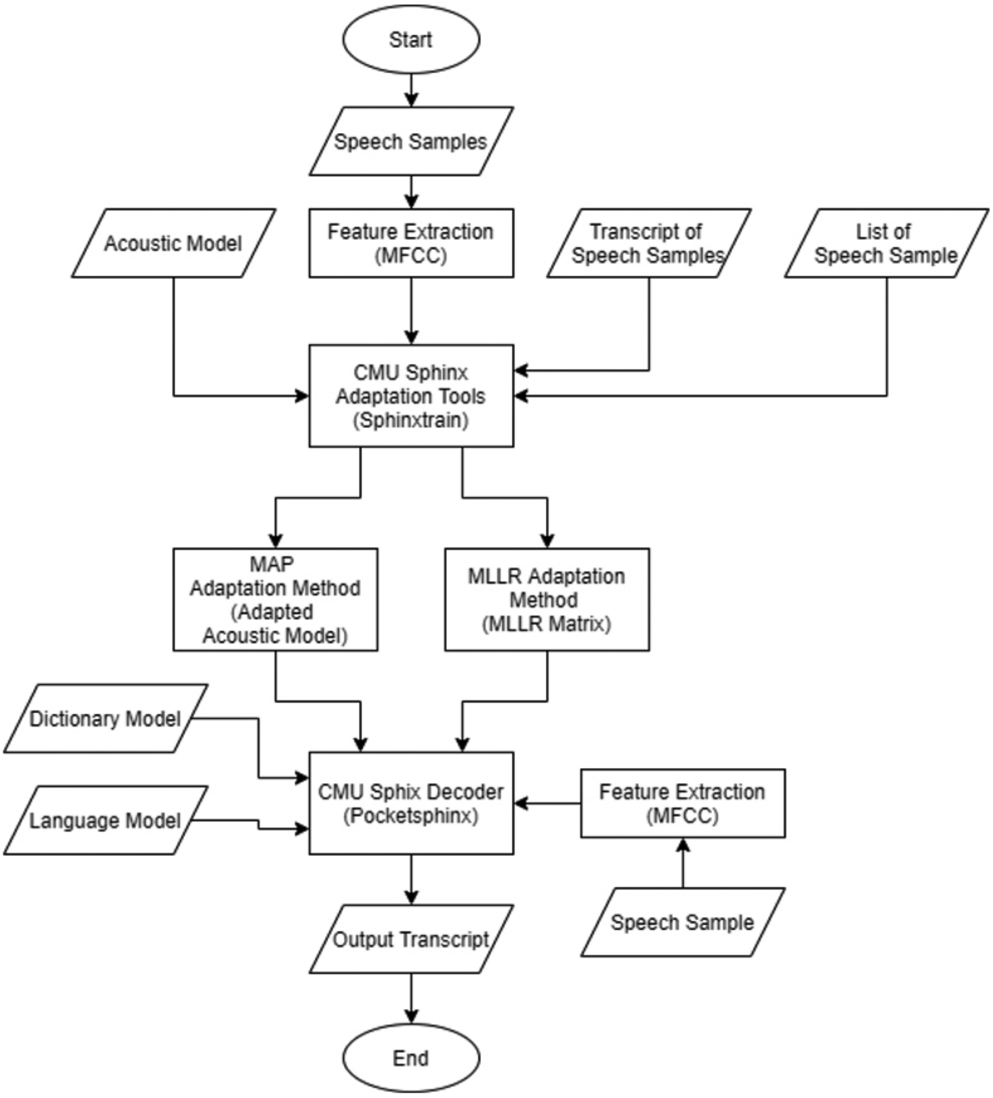

Start
Speech Samples
Acoustic Model
Feature Extraction (MFCC)
Transcript of Speech Samples
List of Speech Sample
CMU Sphinx Adaptation Tools (Sphinxtrain)
MAP Adaptation Method (Adapted Acoustic Model)
MLLR Adaptation Method (MLLR Matrix)
Dictionary Model
Language Model
CMU Sphix Decoder (Pocketsphinx)
Feature Extraction (MFCC)
Speech Sample
Output Transcript
End

Figure 3: Flow Chart for the overall system process

all the result will be analysed by Sphinxtrain modules with Perl to measure the Word Error Rate (WER), Word Recognition Rate (WRR) and Sentence Error Rate (SER). Regarding comparison purposes, the system's functionality is evaluated under various conditions, such as without adaptation, with MLLR adaptation only, with MAP adaptation solely, and with combined MAP and MLLR adaptation.

3.2 Model Adaptation

Adaptation was employed throughout this research to increase the program's speech recognition performance. 5 separate people's voice samples and one internet-based conversation containing 50 different sentences were used in this research project. For adaption, many of these voice samples were obtained from typical conversations with five different people, as well as a presentation for one online speech. That work was completed using Sphinxtrain and Sphinxbase. Maximum Likelihood Linear Regression (MLLR) and Maximum A-Posteriori (MAP) were really the two adaption strategies used.

Typically implemented to the means through one or perhaps more Gaussian Mixture Models, this MLLR method determines a linear transformation. Pertaining to such models, this also maximises the probability of a sample of information. On the other hand, MAP is a sort of enhancement in which the functional form represents the likelihood of an outcome based (or symbol, or sequence) on prior information. That would be the likelihood with that occurrence, consistent with past prediction with that event's possibility (prior probability distribution) and one or perhaps more occurrences. Furthermore, MLLR adaptation is convenient and ideal for limited sample adaption, but MAP adaptation needs additional sample to achieve the desired precision. These MLLR matrix and customised acoustic models are again incorporated and modified correspondingly only with initial acoustic model.

3.3 Software Required

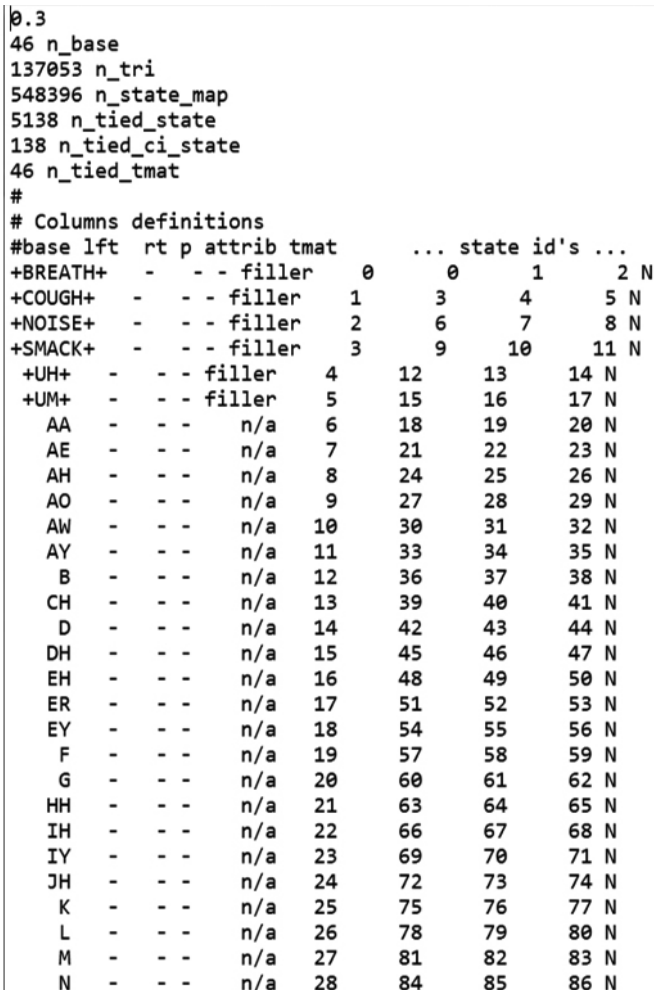
This research had utilized GoldWave (Goldwave, 2021) as the main recording tool to capture voices from Malaysian English speakers. Besides, CMU Sphinx (CMU Sphinx, 2021) had been used to build speech recognition software decoders. This is a cutting-edge large-vocabulary speech recognition framework introduced at CMU. Continuous Hidden Markov Models (HMMs) with optimised Gaussian Mixture Model (GMM) computations are used to decode conversation. Microsoft Visual Studio (Microsoft, 2021) provides the integrated programming environment (IDE) developed by Microsoft for various sorts of programming production, which including software applications. Implementation programs, compilers, and several other capabilities are also included to make overall program construction procedure easier. In this project, the PyCharm IDE (PyCharm, 2021) is used to construct the Graphical User Interface (GUI) for this project.

3.4 Method/ Tools/ Technologies Involved

3.4.1 Acoustic Model

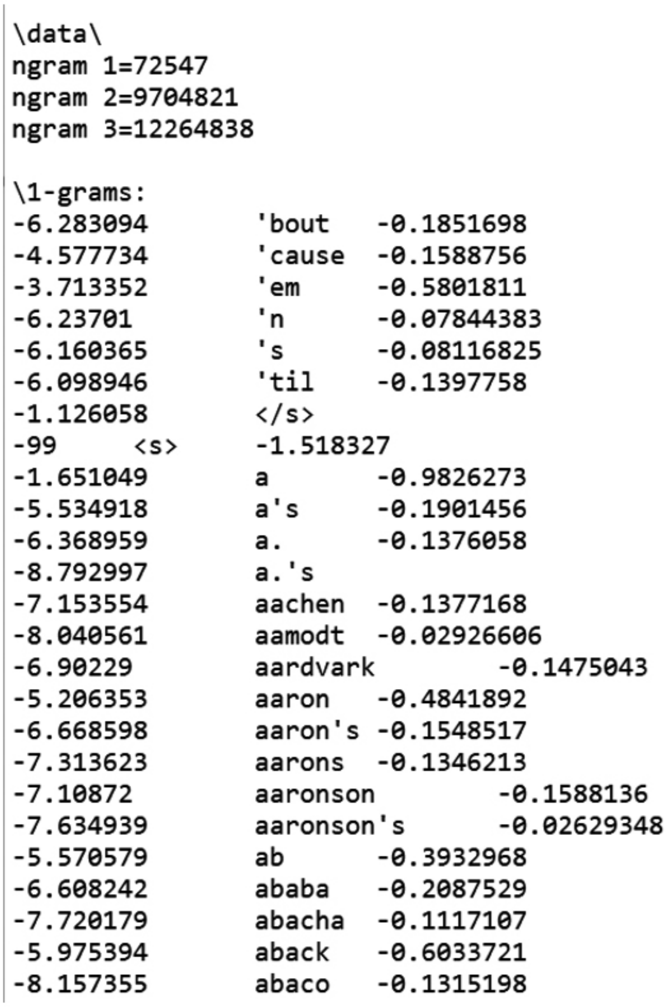
The quantitative interpretations of different sounds that make up a phrase are contained in the acoustic model. Furthermore, these statistical interpretations of voice are referred to as phonemes. Phonemes were created by training Hidden Markov Models (HMM) on a huge array of voice databases, and each phoneme will have its own hmm value. Throughout decoding, the decoder will compare the hmm value of certain phonemes to the adjusted distinct sound to evaluate whether phrase has been spoken.

Figure 4 depicts the acoustic model raw file used to describe the phonemes and their respective tied states. The acoustic model is important to calculate the probabilities of tri-phone sequences.



```
0.3
46 n_base
137053 n_tri
548396 n_state_map
5138 n_tied_state
138 n_tied_ci_state
46 n_tied_tmat
#
# Columns definitions
#base lft  rt p attrib tmat      ... state id's ...
+BREATH+   -   - - filler    0      0      1      2 N
+COUGH+   -    - - filler    1      3      4      5 N
+NOISE+   -    - - filler    2      6      7      8 N
+SMACK+   -    - - filler    3      9     10     11 N
 +UH+   -    - - filler    4     12     13     14 N
 +UM+   -    - - filler    5     15     16     17 N
   AA   -    - -    n/a    6     18     19     20 N
   AE   -    - -    n/a    7     21     22     23 N
   AH   -    - -    n/a    8     24     25     26 N
   AO   -    - -    n/a    9     27     28     29 N
   AW   -    - -    n/a   10     30     31     32 N
   AY   -    - -    n/a   11     33     34     35 N
    B   -    - -    n/a   12     36     37     38 N
   CH   -    - -    n/a   13     39     40     41 N
    D   -    - -    n/a   14     42     43     44 N
   DH   -    - -    n/a   15     45     46     47 N
   EH   -    - -    n/a   16     48     49     50 N
   ER   -    - -    n/a   17     51     52     53 N
   EY   -    - -    n/a   18     54     55     56 N
    F   -    - -    n/a   19     57     58     59 N
    G   -    - -    n/a   20     60     61     62 N
   HH   -    - -    n/a   21     63     64     65 N
   IH   -    - -    n/a   22     66     67     68 N
   IY   -    - -    n/a   23     69     70     71 N
   JH   -    - -    n/a   24     72     73     74 N
    K   -    - -    n/a   25     75     76     77 N
    L   -    - -    n/a   26     78     79     80 N
    M   -    - -    n/a   27     81     82     83 N
    N   -    - -    n/a   28     84     85     86 N
```

Figure 4: Acoustic Model that describes the features of the sounds



```
\data\
ngram 1=72547
ngram 2=9704821
ngram 3=12264838

\1-grams:
-6.283094       'bout   -0.1851698
-4.577734       'cause  -0.1588756
-3.713352       'em     -0.5801811
-6.23701        'n      -0.07844383
-6.160365       's      -0.08116825
-6.098946       'til    -0.1397758
-1.126058       </s>
-99     <s>     -1.518327
-1.651049       a       -0.9826273
-5.534918       a's     -0.1901456
-6.368959       a.      -0.1376058
-8.792997       a.'s
-7.153554       aachen  -0.1377168
-8.040561       aamodt  -0.02926606
-6.90229        aardvark        -0.1475043
-5.206353       aaron   -0.4841892
-6.668598       aaron's -0.1548517
-7.313623       aarons  -0.1346213
-7.10872        aaronson        -0.1588136
-7.634939       aaronson's      -0.02629348
-5.570579       ab      -0.3932968
-6.608242       ababa   -0.2087529
-7.720179       abacha  -0.1117107
-5.975394       aback   -0.6033721
-8.157355       abaco   -0.1315198
```

Figure 5: N-gram Modeling which predicts the sequence of words

3.4.2 Language Model

The N-gram language model was included in this research. Three-gram, unigram, bigram, and trigram are among the N-grams included in this research. Various probabilities were allocated to the phrases throughout this N-gram. That sequence probabilities of a phrase appearing in a conversation are represented by the probabilities allocated to all of it. In other words, using all these probabilities, the voice recognition software can correctly determine which term will occur next and differentiate separating two words with quite comparable pronunciations.

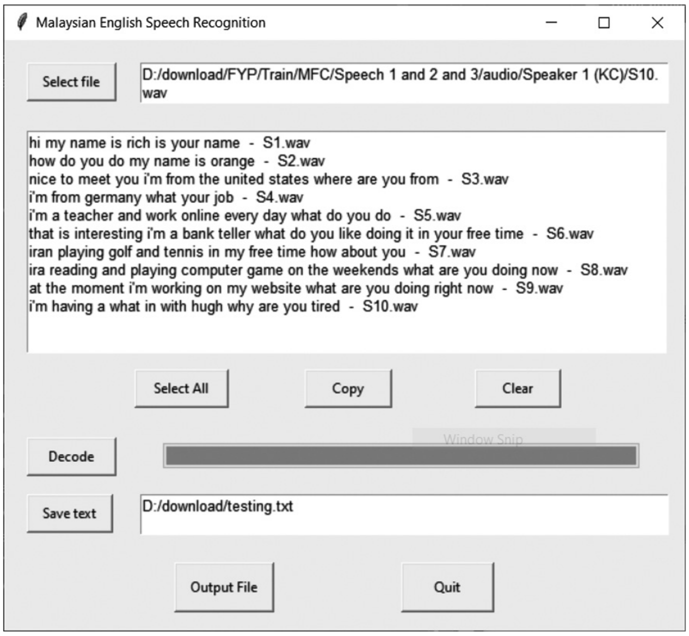
Figure 5 depicts the language model used in this speech research project. The first column shows the log probabilities of the n-gram while the last column shows the back-off probabilities.

3.4.3 Dictionary Model

The dictionary would be a document which includes all or most of the vocabulary words which can be identified mostly by software, within both vocabulary and phonology transcription. None with the voiced phrases would be identified without the dictionary.

3.5 PyCharm Community Edition Graphical User Interface Design

This graphical user interface for the speech recognition system has been devised and provided with the opportunity for the average consumer to be used. For the Malaysian English Continuous Speech Recognition System, PyCharm Community Edition has been applied to construct and establish the graphical user interface.



Malaysian English Speech Recognition

D:/download/FYP/Train/MFC/Speech 1 and 2 and 3/audio/Speaker 1 (KC)/S10.wav

hi my name is rich is your name - S1.wav
how do you do my name is orange - S2.wav
nice to meet you i'm from the united states where are you from - S3.wav
i'm from germany what your job - S4.wav
i'm a teacher and work online every day what do you do - S5.wav
that is interesting i'm a bank teller what do you like doing it in your free time - S6.wav
iran playing golf and tennis in my free time how about you - S7.wav
ira reading and playing computer game on the weekends what are you doing now - S8.wav
at the moment i'm working on my website what are you doing right now - S9.wav
i'm having a what in with hugh why are you tired - S10.wav

D:/download/testing.txt

Figure 6: Graphical user interface for Malaysian English Speech Recognition System

The graphical user interface platform designs, as depicted in Figure 6, were created using Window Form App (.Net Framework). The developed graphical user interface GUI is shown in Figure 6. The buttons have been written with their specific roles and implemented to ensure that they functioned as intended.

4.0 RESULT AND DISCUSSION

Overall effectiveness of speech recognition is compared with a popular measurement defined as Word Error Rate (WER) to determine the functionality including its produced speech recognition system. WER is calculated through using words of speech, as stated with in equation following.

$$Word\ Error\ Rate\ (WER) = \frac{Substitutions\ (S) + Insertions\ (I) + Deletions\ (D)}{Number\ of\ Word\ in\ the\ reference\ (N)}$$

Eq. 3

Word Recognition Rate (WER) is calculated by dividing the overall amount of words throughout the comparison transcript through the amount of matching terms. Sentences Error Rate (SER) is calculated by dividing the quantity of incorrect sentences throughout the comparison transcript through the entire quantity of sentences throughout the transcript.

This would calculate the amount of words which are substituted, deleted from its transcription, and unsaid words which are already inserted to evaluate the effectiveness about an automatic speech recognition system. Word Recognition Rate (WRR) and Sentences Error Rate (SER) are two metrics that can be used to evaluate the effectiveness about an ASR system. Several metrics should be applied to evaluate how many words inside the comparison translation fit and how many phrases are erroneous.

The model adaptation will be trained and tested by 1 online presenter and 5 speakers. These speakers will help to adapt the acoustics models and testing the adapted models. For all adapted transcription will used 50 sentences to adapt acoustics model and 30 sentences to test the adapted model.

4.1 Speech Recognition System Performance Test by using Adaptation Transcription

For the MAP average value, the WER is 15.57%, the WRR is 86.41% and the SER is 64.97%. By comparison between the MAP and MLLR, the WER of MAP is lesser than the MLLR with 31.39%, the WRR of MAP is higher than MLLR with 22.29% and the SER of MAP is lower than MLLR with 32.47%. This comparison showed that WER and SER of MAP have much smaller compare with MLLR and the WRR have very large improvement compare with MLLR. The MAP is better than MLLR in improve accuracy around 20% and reduce mistake for the speech recognition system around 30%.

For both adaptation with MAP and MLLR average value, the WER is 17.94%, the WRR is 88.84% and the SER is 70.83%. By comparison between both adaptation and MAP, the WER of both adaptation is more than the MAP with 2.37%, the WRR of both adaptation is higher than MAP with 2.44% and the SER of both adaptation is higher than MAP with 5.86%. This comparison showed that both adaptations have improved the accuracy of the speech recognition system, but it also increases the error rate of the speech recognition system by comparing with MAP. The increase of SER is much higher compare WRR by comparing both adaptation and MAP.

In conclusion, the MAP adaptation method is best choice for the speech recognition system, and it is highest improvement in accuracy with lowest error rate for the speech recognition system. Even through, both adaptations method has the highest WRR but the WER and SER also increased. The increase of the error rate is much higher compare with recognition rate for both adaptation method. Table 1 tabulates the word error rate, word recognition rate and sentence error rate when adaptation transcription was used.

Table 1: Error Rate and Recognition Rate with Adaptation Transcription

	WER	WRR	SER
Original	55.10%	58.25%	97.44%
MLLR	46.96%	64.12%	97.44%
MAP	15.57%	86.41%	64.97%
MLLR and MAP	17.94%	88.84%	70.83%

4.2 Speech Recognition System Performance Test by using Test Transcription

For the MAP average value, the WER is 32.84%, the WRR is 72.52% and the SER is 78.89%. By comparison between the MAP and MLLR, the WER of MAP is lesser than the MLLR with 18.55%, the Word Recognition Rate (WRR) of MAP is higher than MLLR with 12.79% and the SER of MAP is lower than MLLR with 18.33%. This comparison showed that WER and SER of MAP have much smaller compare with MLLR and the Word Recognition Rate (WRR) have very large improvement compare with MLLR. The MAP is better than MLLR in improve accuracy around 12% and reduce mistake for the speech recognition system around 18%.

For both adaptation with MAP and MLLR average value, the WER is 48.01%, the Word Recognition Rate (WRR) is 61.97% and the SER is 81.67%. By comparison between both adaptation and MAP, the WER of both adaptation is more than the MAP with 15.17%, the Word Recognition Rate (WRR) of both adaptation is lower than MAP with 10.55% and the SER of both adaptation is higher than MAP with 2.78%. This comparison showed that both adaptations have reduced the accuracy of the speech recognition system and it also further increases the error rate of the speech recognition system by comparing with MAP.

In conclusion, the MAP adaptation method is also best choice for the speech recognition system with test transcription, and it have highest improvement in accuracy with lowest error rate for the speech recognition system. The accuracy of both adaptation method is decreased compare with MAP by using test transcription and this case is not similar compare with adaptation transcription. By using adaptation transcription, both adaptation method is showed the improvement of the accuracy, but it also increases the error rate. That adaptation method showed the accuracy decrease and further improve the WER by using test transcription. Table 2 tabulates the word error rate, word recognition rate and sentence error rate when test transcription was used.

One of the possible reason for MAP to outperform MLLR using both adaptation and test transcription is due to the small amount of adaptation data. Thus, in future research, it is advisable to record more adaptation data in different settings so that the performances of MLLR could be better.

Table 2: Error Rate and Recognition Rate with Test Transcription

	WER	WRR	SER
Original	58.55%	54.36%	97.22%
MLLR	51.39%	59.73%	97.22%
MAP	32.84%	72.52%	78.89%
MLLR and MAP	48.01%	61.97%	81.67%

5.0 CONCLUSION

In conclusion, adaptation and test scripts were used to evaluate the effectiveness of the constructed speech recognition system. For testing this system by using adaptation transcription, the MAP adaptation method has showed the high accuracy with lowest error rate. The MAP has the lowest WER and lowest SER with 15.57% and 64.97%. The MAP also has high Word Recognition Rate (WRR) with 86.41%. But the Word Recognition Rate (WRR) of MAP adaptation method is lower than both adaptation method and the different of them is 2.44%. Both adaptation method also will increase the WER and SER compare with MAP adaptation method. The improvement accuracy of both adaptation method is lower compare with increment of error rate. The MLLR adaptation method is showed the worst improvement in accuracy and error rate compare with MAP and both adaptation method. So, the MAP is best choice for adaptation transcription.

For testing transcription, the MAP adaptation method is showed the highest accuracy and lowest error rate for speech recognition system. The WER of MAP adaptation method is 32.84%, the Word Recognition Rate (WRR) is 72.52% and the SER is 78.89%. Moreover, both adaptation method has second higher accuracy and second lower error rate. Both adaptation with test transcription have showed the decrease accuracy and increase error rate compare with adaptation transcription. The MLLR adaptation method is showed the worst improvement in accuracy and error rate compare with MAP and both adaptation method.

Finally, the MAP adaptation method is suitable for adaptation and test transcription. The MAP adaptation method is best adaptation method, and it will be implemented into speech recognition system. This Malaysian English speech recognition system's weakness is that it may function poorly if somehow the voice stream includes a lot of distortion and noise. Whereas if surroundings are noisy, this system may not achieve sufficient performance.

This research work found that MAP adaptation approach outperformed MLLR adaptation approach in all settings. It may be due to the limited adaptation data recorded. In future, more adaptation data should be recorded in different environment so that the results could be improved. ■

REFERENCES

[1] Abhang P, Gawali B and Mehrotra S (2016) Introduction to EEG- and speech-based emotion recognition. Academic Press.

[2] Abushariah AAM, Gunawan TS, Khalifa OO and Abushariah MAM (2010) English digits speech recognition system based on Hidden Markov Models. International Conference on Computer and Communication Engineering (ICCCE'10), Kuala Lumpur, Malaysia, 2010, pp. 1-5, doi: 10.1109/ICCCE.2010.5556819.

[3] Alim S (2018) Some Commonly Used Speech Feature Extraction Algorithms, From Natural to Artificial Intelligence - Algorithms and Applications. Ricardo Lopez-Ruiz, IntechOpen.

[4] Aymen M, Abdelaziz A, Halim S and Maaref H (2011) Hidden Markov Models for automatic speech recognition. 2011 International Conference on Communications, Computing and Control Applications (CCCA), Hammamet, Tunisia, 2011, pp. 1-6.

[5] CMU Sphinx (2021) CMU Sphinx website. Available at: https://cmusphinx.github.io/ (Accessed: 11 September 2021).

[6] Cuiling L (2016) English Speech Recognition Method Based on Hidden Markov Model. 2016 International Conference on Smart Grid and Electrical Automation (ICSGEA), Zhangjiajie, China, 2016, pp. 94-97.

[7] Gadag A and Sagar DBM (2016) N-gram based paraphrase generator from large text document. 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016, pp. 91-94.

[8] Gales MJF (2009) Acoustic modelling for speech recognition: Hidden Markov models and beyond?. 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, Moreno, Italy, 2009, pp. 44.

[9] Goldwave (2021) Goldwave website. Available at: https://www.goldwave.com/ (Accessed: 11 September 2021).

[10] Hatami A, Akbari A and Nasersharif B (2013) N-gram adaptation using Dirichlet class language model based on part-of-speech for speech recognition. 2013 21st Iranian Conference on Electrical Engineering (ICEE), Mashhad, Iran, 2013, pp. 1-5, doi: 10.1109/IranianCEE.2013.6599642.

[11] HITL (2021) Voice Recognition. Available at: <http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/I.D.2.d.VoiceRecognition.html> (Accessed 19 March 2021).

[12] Houghton Mifflin (2015) The American Heritage Dictionary of the English Language. Boston.

[13] Ito A and Kohda M (1996) Language modeling by string pattern N-gram for Japanese speech recognition. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Philadelphia, PA, USA, 1996, pp. 490-493 vol.1, doi: 10.1109 /ICSLP.1996.607161. ISBN: 978-0-12-804490-2.

[14] Lestari DP and Irfani A (2015) Acoustic and language models adaptation for Indonesian spontaneous speech recognition. 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2015, pp. 1-5.

[15] Microsoft (2021) Microsoft Studio website. Available at: https://visualstudio.microsoft.com/ (Accessed: 11 September 2021).

[16] Muda L, Begam M and Elamvazuthi I (2021) Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Available at: <https://arxiv.org/abs/1003.4083> (Accessed 20 March 2021).

[17] NIDCD (2021) What Is Voice? What Is Speech? What Is Language? Available at: <https://www.nidcd.nih.gov/health/what-is-voice-speech-language> (Accessed 17 March 2021).

[18] Oh Y, Yoon J and Kim H (2007) Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Communication, 49(1), pp.59-70.

[19] Rabiner L and Juang B (1986) An introduction to hidden Markov model. IEEE ASSP Magazine, Vol. 3 Issue 1, pp. 4-6.

[20] PyCharm (2021) PyCharm website. Available at: https://www.jetbrains.com/pycharm/ (Accessed: 11 September 2021).

[21] Science Direct (2021) Speech Recognition - an overview ScienceDirect Topics. Available at: <https://www.sciencedirect.com/topics/engineering/speech-recognition> (Accessed 17 March 2021).

[22] Takahashi S and Morimoto T (2012) N-gram Language Model Based on Multi-Word Expressions in Web Documents for Speech Recognition and Closed-Captioning. 2012 International Conference on Asian Language Processing, Hanoi, Vietnam, 2012, pp. 225-228, doi: 10.1109/IALP.2012.55.

[23] Vergin R, O'Shaughnessy D and Farhat A (1999) Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. IEEE Transactions on Speech and Audio Processing, 7(5), pp.525-532.

[24] Xue C (2018) A Novel English Speech Recognition Approach Based on Hidden Markov Model. 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Hunan, China, 2018, pp. 1-4.

[25] Yoon B (2009) Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics, 10(6), pp.402-415.

[26] Yoong, KC and Hong KS (2021) Development Of Malaysian English Large Vocabulary Continuous Speech Recognizer Using Acoustic Model Adaptation. International Conference on Digital Transformation and Applications (ICDXA), 25-26 October 2021, pp. 36-48.

[27] Zhou X, Garcia-Romero D, Duraiswami R, Espy-Wilson C and Shamma S (2011) Linear versus mel frequency cepstral coefficients for speaker recognition. IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 2011, pp. 559-564.

PROFILES

YOONG KAH CHUNG is a bachelor degree graduate of Faculty of Engineering and Technology, Tunku Abdul Rahman University College, Kuala Lumpur. His specialization and area of interest are speech recognition and electronic technology.
Email address: yoongkc-wg17@student.tarc.edu.my

DR HONG KAI SZE is a senior lecturer of Faculty of Engineering and Technology, Tunku Abdul Rahman University College, Kuala Lumpur. He received his PhD from Universiti Sains Malaysia. His specialization and area of interest are speech recognition and artificial intelligence.
Email address: hongks@tarc.edu.my