ROBUST FEATURE EXTRACTION BASED ON SPECTRAL AND PROSODIC FEATURES FOR CLASSICAL ARABIC ACCENTS RECOGNITION

Noor Jamaliah Ibrahim[1*], Mohd Yamani Idna Idris[2*], Mohd Yakub @ Zulkifli Mohd Yusoff[3], Noor Naemah Abdul Rahman[4] and Mawil Izzi Dien[5]

[1,2]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[3,4]Academy of Islamic Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia

[5]Faculty of Humanities and Performing Arts, University of Wales Trinity Saint David, Lampeter, UK

Email: noor3184.um@gmail.com[1]*(corresponding author), yamani@um.edu.my[2*](corresponding author), zulkifliy@um.edu.my[3], naemah@um.edu.my[4], izzidien@hotmail.com[5]

DOI: https://doi.org/10.22452/mjcs.sp2019no3.4

ABSTRACT

The variability of speech patterns produced by individuals is unique. The uniqueness is due to the accent influenced by the individual's native dialect. Modeling individual variation of spoken language is a challenge under the Automatic Speech Recognition (ASR) field. The individual differences concerning of accent revealed the critical issues in Classical Arabic (CA) recitation among Malay speakers. This problem is caused by the misarticulate phonemes, which affected by the Malay colloquial dialect and native language. Most of ASR researchers are unable to understand the behavior of phonemes and speech patterns in CA, thus degrading the ASR performance. This paper focuses on identifying the accent of Malay speakers on the recitation of Sūrah Al-Fātiḥah with 7 Quranic accents, using the proposed feature extraction technique. In this work, the technique presented is a combination of spectral and prosodic features, which are mainly designed for accent in ASR. Differed with current conventional method, where the spectral feature alone has been applied for feature extraction in many ASR research. The prosodic elements in CA such as pitch, energy and spectral-tilt need to be taken into consideration, thus a significant variety of features for each phoneme able to help in distinguishing one accent from another. Meanwhile, the spectral representation of Mel-Frequency Cepstral Coefficients (MFCC) is utilized for the decorrelating property of the cepstrum. At present, Gaussian Mixture Models (GMM) has been applied for the classification stage. From experimental results, the system performance is the best when the prosodic is integrated with MFCC, alongside the GMM with 81.7%-89.6% of accuracy. It was 5.5%-7.3% increment as compared to MFCC alone.

Keywords: Quranic accents, spectral, prosodic, MFCC, GMM, Malay speakers, ASR

1.0 INTRODUCTION

The speech signal has a variety of information embedded in it. Various acoustic and linguistic information stored within the signal makes it unique. An appropriate parametric representation of the speech signal is required that can extract the relevant information statistically, which is necessary with the desired goals. Feature extraction is the crucial function of a speech processing front-end, and it plays the critical procedure in extracting and recognizing the prominent properties of speech data, due to facilitate its use for further processing in Automatic Speech Recognition (ASR). Conventionally, most ASR research has extracted the phonological info from the speech waveform through the spectral feature, which is considered a time-domain feature extraction method. However, not all phonological details can be extracted from the speech waveform in the time-domain. The variability of speech patterns and the behavior of each phoneme were considered a critical issue to be dealt with the feature extraction method. The variations in individuals' patterns of speech make ASR development a challenging task. The accuracy of the ASR system could be degraded and reduced when the variability of speech pattern varies [1]. The variations in individuals' patterns of speech are due to the accent influenced by the individual's native dialect, which is an aspect of speaker variability [2]. In sociolinguistics, accent

refers to a manner of pronunciation peculiar to a specific person, location, and ethnicity. It typically differs in the voice's quality, pronunciation, distinction of vowels and consonants, as well as stress and prosody.

For the Arabic language, the variability is focused on inter-speaker and intra-speaker variability of Classical Arabic (CA), which is due to the accent, particularly the method of recitation. The accent in CA is referred to the Quranic accents (known as '*Qiraat*'). It belongs to a group that is described as stress-timed languages. This criterion described the unique characteristic in the Arabic language, where the pharyngeal and emphatic phonemes had more lexical stress systems compared to any other language [3]. It differed to the Malay language, where the sounds and phonemes produced have light stress and no stress at all [4]. Thus, mistakes and mispronunciation errors while reciting the Quranic Arabic have potentially occurred among Malay speakers. In this case, the common pronunciation errors were regularly being made towards the specific Quranic Arabic phonemes (see Table 1) [5], which often confused with the Malay colloquial dialect. A few researches [5],[6],[7] have revealed the issues related to the misarticulate phonemes among Malay speakers, and these problems also been noticed by the Quranic Arabic expertise [8],[9]. The difficulties have arisen due to the lack of ability to pronounce strictly according to the point of articulation (*makhraj*) of the Quranic Arabic phonemes, that have shaped the student's tongue and set their speaking preferences. The lexical stress characteristic, such as energy, pitch, spectral-tilt, syllable duration, and other traits [10], were considered as the elements of prosody and crucial for determining the variability of accents. The speech traits often ignored and less concern by the ASR researchers, even though the research community has witnessed an improvement in the performance of the ASR systems.

As mentioned earlier, most of the ASR researchers were more concerned about the spectral feature and disregarded the existence of the prosodic feature which carried the attributes of accent in languages. Less exploration in Quranic accents makes them unaware of the issues highlighted, especially among the Muslim and Malay community. It is probably because of the lack and unavailability of a speech database for Quranic accents, which leads to the major hurdle in conducting this kind of research. Research conducted previously by [11],[12],[13],[14] has used the recitation audio of the Quranic accents available on the web. Noises, echo, and reverberation effects might influence the downloaded audio samples from the web and thus disputed the overall performance results. Spectral feature is referred to short-term cepstral feature that reflects the voice parameters and signal characteristics of the speakers in the frequency time-domain. The algorithm of spectral features, known as Linear Predictive Coding Coefficient (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) have been carried out by [11], whereas MFCC only been implemented by [12],[13],[14] for the feature extraction phase. Meanwhile, the Quranic Arabic experiment without accent assessment has been conducted by [15],[16], using the local in-house database constructed by non-Arabs-male certified speakers, and audio in the web for Arabs-male certified speakers, respectively. Both kinds of research also used a conventional MFCC algorithm for the feature extraction stage. Last, for Arabic recognition research, only [17] used a different spectral feature, unlike the others, known as Multi-Directional Local Features (MDLF), but the research conducted by [18] still using the conventional MFCC algorithm as a spectral feature. All the ASR research presented here was related to the Arabic or CA language, which concentrated solely on spectral based-features. Limitation issues regardless of accent and behavioral patterns existed in the Quranic Arabic need to be considered by previous researchers for improvement. Our enthusiasm in this research is to scrutinize the accentual prominences on the phonetics and its prosodic effects towards the accent recognition process. Here, integrating prosodic and spectral features for the feature extraction approach was highly significant due to improve the recognition results. Via technology, efforts in preserving the *Sūrah Al-Fātiḥah* in Quranic Arabic from distortion have motivated us, to rectify any issues related to accent for a better understanding and overcome the misarticulate problems while performing prayers either in the congregation or alone [8],[19],[20],[21]. The proposed feature extraction technique used the prosody elements such as pitch, energy, and spectral tilts, meanwhile MFCC applied for the spectral feature. For classification, the Gaussian Mixture Model (GMM) was exploited to complete the recognition process from the proposed front-end processing technique. The performance results based on accuracy and the Equal Error Rate (EER) in regards of Quranic accents are compared with other previous research.

The entire paper is arranged as follows. In section 2, we go over some related literature on the linguistic properties of the Quranic accents, as well as previous implementation related to ASR, accent and dialect identification. In section 3, the methodology of the proposed technique of front-end processing is highlighted. The performance of the proposed feature extraction evaluation results is presented in section 4. Last, the conclusions and future works are drawn in section 5.

2.0 RELATED LITERATURE REVIEW

The Quranic Arabic, as well as Quranic accents (*Qiraat*) is the holy book for Muslims that is written and recited in the Arabic language, the language in which it was revealed [22],[23]. It is considered an ancient Arabic language, notorious as Classical Arabic (CA). CA differs from the other two types of Arabic language known as Modern Standard Arabic (MSA) and Dialectal Arabic (DA) because both languages intended to change over time and often flexible with dialect. However, the Quranic Arabic is protected, since the reciters recite it orally from one person to another for preserving its correct and manner of articulations, even in different modes of recitations. Appropriately reciting Quranic Arabic is crucial for all Muslims and indispensable in Islamic worship, such as prayers. Mistakes while reciting it are considered forbidden and not permissible in Islam.

Based on *Ibn al-Jazari* (751H), the science of the Quranic accents (*Qiraat*) is the discipline that studies the recitation of the words in the Quranic Arabic and its differences, from one reciter to another who heard it orally [24]. According to the Prophets Muhammad PBUH words (*hadīth*), there are seven facet or modes of Quranic recitation, in which permissible by Allah to His servants, which are; *'An-Nafi,' Ibn Kathir, Abu' Amru, Ibn' Amir, 'Aasim, Hamzah* and *Al-Kisa'ei*. These seven Quranic accents are from *Shatebiah's* way. Meanwhile, additional three modes of recitation were agreed by the majority of Muslim scholars upon the commentary of *Imam Al-Jazari*, which includes, *Abu Ja'far, Ya'akub al-Yamani* and *Khalaf* [25], which are from *Al-Durrah* way. Each Muslim leader (*Imam*) has two narrators (*Riwayah*). Any modes of recitation are allowed to be used because all the modes of Quranic Arabic recitation are considered correct. It is because Allah SWT has revealed Quranic Arabic through Prophet Muhammad PBUH in seven modes of recitation, and there is no contradiction between one another. Utmost importantly, all these varieties of Quranic accents had shown the mercy of Allah SWT towards His servant, by giving flexibility towards Muslim people (with varieties of dialects) to recite the Quranic Arabic, in such a way that conveniently applied by reciters and suited with individual dialect and accent. However, Muslim people still make mistakes in preserving the correct articulations while reciting the Quranic Arabic. Without proper learning and training of Quranic recitation, most speakers keep making pronunciation errors, especially when pronouncing the misarticulate phonemes. Different speakers from different regions and dialects make mistakes differently depending on the phonemes. It is due to the pronunciation habits acquired from the speakers' mother tongue language.

Based on the population in Malaysia, most Muslims were Malay natives that were considered as non-Arabs speakers. The Malay language itself is obtained initially from the English and Arabic languages. However, Arabic is a morphologically rich language compared to English and Malay languages [26]. Here, the majority of Muslims recited Quranic Arabic using the narrator of *Hafs* reading style from *Imam 'Aasim*. Based on Muslim scholars and experts [8],[9], learning Quranic accents are considered mandatory and understandings the Quranic accents also crucial as different narrators narrate it differently. Reading the Quranic accents, either in prayer (alone or congregation) or outside the prayer, is a matter directly related to the accent. Awareness towards the Quranic accents has to be extended to Muslim audiences, for example (in Malaysia), so that the chaos and misconception towards other Muslims can be avoided [8],[20],[21]. Moreover, the Quranic accents have a strong relationship between knowledge of Islamic jurisprudence (*fiqh*) in sectarian diversity, knowledge interpretation of the Quranic Arabic that led to various interpretations.

Several phonemes of Arabic letters have been identified as misarticulate phonemes by Malay speakers (see Table 1), which are regularly pronounced by Malay colloquial dialect. These phonemes of letters have closely articulated phonemes and similar sound with the Malay language, but in fact, have different articulation points. Meanwhile, [27] have reported that, there are seven consonants (/q/-[ق], /ð̲/-[ظ], /x/-[خ], /ɣ/-[غ], /H/-[ح], /ʔ/-[ع] and /h/-[ه]), unable to give the promising results based on inappropriate values of formant frequencies measured using spectrogram. Thus, it is considered as difficult Arabic consonants to be articulated properly among Malaysian primary school students. Experts and researchers had identified those phonemes after years of teaching the Malay students from all levels of educations [5],[8],[9],[27], and those misarticulate phonemes have been summarized in Table 1.

Table 1: Misarticulate phonemes (Malay speakers) [5],[8],[9],[27],[28]

Group 1			Group 2			Group 3		
Hijaiyah Letter	**Phoneme**	**IPA**	**Hijaiyah Letter**	**Phoneme**	**IPA**	**Hijaiyah Letter**	**Phoneme**	**IPA**
ص	Sad	S	س	Seen	s	ث	Theh	θ
ذ	Thal	ð	ز	Zai	z	ظ	Thah	ð̲
ت	Teh	t	ط	Tah	T			
د	Dal	d	ذ	Thal	ð			
س	Seen	s	ش	Sheen	ʃ			
أ	Aa	E	ع	Ain	?			
ك	Kaf	k	ق	Qaf	q			
ث	Theh	θ	س	Seen	s			
ت	Teh	t	ث	Theh	θ			
د	Dal	d	ض	Dad	D			
ح	Hah	H	ه	Heh	h			
خ	Khah	x	غ	Gheyn	γ			

Based on Table 1, the phoneme of Sad /S/-[ص], often confused and misarticulate as the phonemes of Seen /s/-[س], and often confused with the phoneme of Theh /θ/-[ث]. Meanwhile, other phonemes such as Thal /ð/-[ذ] was often confused and misarticulate as the phoneme of Zai /z/-[ز], as well as Thah /ð̲/-[ظ] and so on, as have been reported by researchers in the Table 1 [5],[8],[9],[27]. Each of phoneme is represented in the term of IPA (International Phonetic Association) for Arabic [28]. In this paper, the experiment has been tested on *Sūrah Al-Fātiḥah*, the first chapter of Quranic Arabic. This chapter has a crucial function in Islamic prayer [29] and is represented as the mother of Quranic Arabic. Here, the phonemes of the consonant letters (/S/-[ص], /s/-[س], and /z/-[ز]) from the word "*Ssiraatha"* (صِرَاطَ) for verse 6 and 7 in *Sūrah Al-Fātiḥah* have been choosing for assessment, since the differentiation of specific words related to accent were located in those words and clearly observable (see Section 4.1). Properties of these accent words that represented through these 3 phonemes was indicated as 'Context-sensitive phonetic', and the errors while pronouncing these 3 phonemes are predictable, not just limited to the Malay speakers on behalf of non-Arabs speakers, but also the Arabs speakers. Means that, the majority of Muslim in the world has a problem to articulate these phonemes, because of the close articulation sounds related with accent [3],[30],[31]. The variants in individuals' patterns of speech are special and unique. It is due to the accent influenced by the native dialect. The accent is the subset of the dialect, and dialect itself is an element of speaker variability [2]. Based on the socio-linguistics definition, accent describes as a manner of pronunciation peculiar to a particular person, depending on ethnicity, location, and others. Accent typically differed in terms of quality of the voice or vocal, pronunciation, as well as the distinction of vowels, consonants, stress, and prosody. Accent features in particular languages can be important cues for profiling. Phonotactic, spectral, and prosodic features within the speech signal offered adequate information about the native language of the speakers [32]. A couple of works have been performed for the recognition of the spoken dialect. The majority of them are based on the phonotactic research study [33],[34],[35]. Some works based upon acoustic features have been carried out for automated identification of spoken dialect for languages in western countries. Accent variations stretch out not only in phonetic characteristics but also in prosodic features of speakers [32]. Accent classification scheme based on the Hidden Markov Model (HMM) for the performance analysis of Linear Predictive Coding Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and formants were introduced in [36]. Formants for three primary English accents, such as British, American, and Australian were examined through [37], where the second formant is the most compelling and prominent formant for dialect classification. Lastly, research by [38] applied the pitch and formant contours, as well as the stochastic trajectory nodes in distinguishing between Americans, Chinese, and Turkish accents.

Arabic language, as well as other languages such as English and Russian, belongs to the group of 'stress-times' languages [18]. Prosodic features are essential for the intelligibility and proficiency of the stress-timed language. Moreover, the Arabic language has more lexical stress systems compared to any other language [3], including the Malay accent. As per agreed by the linguistics experts, there is no underlying stress in the Malay language [4], compared to the Arabic language. As referring to Classical Arabic (CA), the recitation of Quranic Arabic and its accents have lexical stress and pitch accented on designated consonants [3]. In other words, the Quranic Arabic language has specialized

roles for stress placement. The classical grammarians are mutually agreed that CA is a word stress language, but not justified clearly towards the Arabic language. However, [39] still has the same perspective towards the Arabic language as a word stress language [40]. Moreover, the characters used in the CA language are well-defined and well-categorized, including its articulatory discipline, as well as its acoustic representation. Although the research community has witnessed an improvement in the performance of the ASR systems, most past ASR researchers were still less concerned and often ignored the speech pattern and behavior of phonemes in the Quranic Arabic language. The traits in phonemes also include as the elements of prosody, which are the lexical stress characteristics of energy, pitch, and spectral-tilt, and it was merely crucial for determining the accent. Besides, the prosodic traits also crucial in determining the additional features and special traits belong to Arabic and Quranic Arabic.

Currently, most of the ASR research only explored the mono-language, and concentrated solely on spectral based-features, that derived through short-time spectral analysis of the speech signal. The size of the short-time spectrum encrypts the information concerning the shape of the speakers' vocal tract [41],[42],[43]. Therefore, spectral features are extensively used for speaker modeling. The primary components of speaker recognition are front-end processing and speaker modeling. Front-end processing converts the input signal into a suitable feature space that is fed to the modeling part. Feature extraction is a very crucial component in ASR. Most feature extraction techniques fall into two categories, which are modeling human voice production and modeling the peripheral auditory hearing. In the first category, the most popular feature is Linear Prediction Cepstral Coefficients (LPCC), meanwhile in the second category, the most popular features are Mel-Frequency Cepstral Coefficient (MFCC) and RASTA-Perceptual Linear Predictive (RASTA-PLP). Nevertheless, all of these feature extraction algorithms were difficult to extract the specific characteristics and traits, regardless of the accent language.

As highlighted by [1], the accuracy values of the ASR system are reduced when the accent of speakers varies. Furthermore, robust feature extraction is considered a complex issue when dealing with the variability of speech patterns. It can be proven when the result of recognition accuracy is decreased, after implementing Mel-Frequency Cepstral Coefficients (MFCC) or Linear Predictive Coding Coefficient (LPCC) alone as spectral features [13],[15],[17],[18], using the Arabic or Quranic Arabic as data samples. Here, the development of the Quranic Arabic recitation recognizer, conducted using Maximum-likelihood (ML) and Minimum Phone Error (MPE) was presented by [15]. The feature extraction process has been computed using MFCC, which is limited to the *Hafs* style of pronunciation only. The same process goes to the research conducted by [18], where a phoneme recognizer is developed to identify and transcribe the Arabic phonemes using a data-driven approach, and it also solely depends on spectral features. In another hand, both LPCC and MFCC algorithm were tested on the Quranic audio-based accents available online on the web by [2], using Probabilistic Principal Component Analysis (PPCA). Beside MFCC and LPCC, another alternative of the spectral feature is known as Multi-Dimensional of Local Feature (MDLF), which mathematically computed using four primal Local Features (LF). This LF feature has been extracted by a three-point Linear Regression (LR) along time, frequency and time-frequency axes, which resemble to capture the acoustic evidence of rising and falling sound, spectral peaks in steady sound, formant transitions and onset/offset. Here, the research highlighted the previous study in ASR that did not count the aspect of prosodic attributes, which existed in the Quranic Arabic recitation and its accents. Constraint issues related to the accent of speakers (data samples), as well as the behavior and traits of speech pattern in Quranic accents were two crucial elements that required consideration by the previous researchers for enhancement.

As stated earlier, prosodic features are crucial for certain languages, especially stress-times languages. This feature deals with acoustic qualities of sound. These features give information about utterance, as well as the speaker. Pitch, energy, and duration are well-proven prosodic features. Prosodic features are essential for the intelligibility and proficiency of stress-timed languages such as Arabic and English [44]. The accent in the prosodic features, play a significant role in language communication and belongs to the concept of auditory perceptive category. From the perspective of the listener, the accent mainly indicates those words that sound more prominent than the word or word rounding [45]. According to [46], prosody is also denoted as 'Acoustic-Phonetics'. It is a study of the acoustic characteristics of speech, including analysis and description of speech in terms of its physical properties. The different properties that can be extracted from the speech are formants, intensity, pitch, duration, and voiced/unvoiced. The specific interaction between pitch variations, intensity (energy), duration, and spectral-tilt, play an essential role in determining the prosody. The literature available highlights that the parameters representing all those elements are found to be the most discriminative features for identifying possible accent. Here, the first parameter involved is pitch or also known as fundamental frequency (F0). Pitch or F0 is intrinsic in the periodic signal and embodies the perceived pitch in human speech. The temporal characteristics of pitch throughout a signal of speech impart the related information. Numerous

research studies have indicated the significance of prosodic features, including the word accent, as well as the phrase tonality in human speech processing. However, only a few make use of fundamental frequency (F0) or pitch in combination along with other acoustic features [47],[48]. Then, the second parameter is intensity or frame energy. In this part, the element of lexical stress is computed, where the amount of energy that helps in determining the voiced and unvoiced part of the speech will be represented the stress pattern of speakers. The energy of each overlapping frames of segmented speech is acquired through summing the squared amplitude of each sample. Next, the duration parameter, and identical to it is the rhythm that is directed by the speaker's native dialect that has been exemplified. The length of the individual speaking style could be varied because of the dialectal effects involving the vowel duration used in the segment [49]. Lastly, the spectral tilt of speech which originates from the glottal flow generated by vibrating vocal folds. Spectral has been analyzed, for instance, in studying linguistic stress [50], phonation type [51], vocal effort [52], and speech intelligibility improvement [53].
Our intention in this research is to preserve Quranic accents from distortion and incorrect interpretation, especially in *Surah Al-Fātiḥah*, where it is crucial while performing prayer [19]. Better understanding and awareness about the Quranic accents in *Surah Al-Fātiḥah* ultimately urgent, due to avoid any misconception and misunderstanding among Muslim communities with a different locality, dialects and school of thought (*maḏhab*) [8],[20],[21].

3.0 METHODOLOGY

In this study, the feature extraction process for various parameters is examined and investigated. The experiment was conducted and tested on spectral features of Mel-Frequency Cepstral Coefficient (MFCC), as well as the prosodic features. Both MFCC and prosodic features are employed to determine the technique that could produce the better results of speech recognition performance and reduce the error rate, mainly for recognition of Quranic accents recitation. The performance results are compared between the experiments that implemented the spectral feature of MFCC only with the experiment that conducted the proposed integration of MFCC algorithm and prosodic features. Figure 2 illustrates the block diagram of the proposed approach for the feature extraction technique. MATLAB tool is considered a programming language for modeling time-series data and appropriately used in this research study.

3.1 Data Preparation

Audio speech data are collected for further processing in a speech recognition machine. In this research, the lack of a speech database for the Quranic accents is the major hurdle for this language. Hence, a local database is generated by conducting an in-house recording. Here, the collection of speech samples was obtained from 14 certified reciters (*Huffaz*) of Malay speakers, where 7 of them are females, whereas other 7 speakers are males. All those reciters are mastered in Quranic accents and have a diploma in Quranic accents (*Qiraat*) from *Darul Quran, JAKIM*[1]. Before the assessment and evaluation take action, the audio samples needs to be validated by the Quranic expertise. In this work, the data samples from *Sūrah Al-Fātiḥah* in 7 different styles will be generated through the recording process in a controlled environment. The first chapter of *Sūrah Al-Fātiḥah* has been prioritized to be used for evaluation, because this chapter is mandatory while performing prayer and considered as a pillar in Muslim prayer. Mistakes while reciting this chapter during the prayer are not permissible in Islam, which led to the invalid of prayer [19],[20],[29].

During the recording process, speakers need to recite in the moderate tone, with proper recitation and correct articulation of the phonemes, as well as followed the rules of pronunciation (*Tajwīd*). The recitation of the complete verse of *Sūrah Al-Fātiḥah* with 7 Quranic accents was made twice per speaker. Each sentence in between were paused, and the audio data with a duration of 5-8 minutes per speaker were able to generate. These speech samples were recorded in a controlled environment and performed using a digital voice recorder (OLYMPUS WS650S) at a 44kHz sampling rate. After the recitation completely performed by speakers, those recitation samples were then manually segmented into sentences and words, using '*Audacity'* sound editor software. The audio data was then converted into '.wav' format, downsample into 16 bits and 16kHz sampling rate.

[1] Darul-Quran, JAKIM - Quranic learning and memorizing institute, under Department of Islamic Development, Malaysia

3.1.1 Misarticulate Phonemes (All verses)

Generally, there are several phonemes, which considered as misarticulate phonemes in *Sūrah Al-Fātiḥah*. Those phonemes were notable as closely articulated phonemes which commonly mispronunced among Malay reciters (see Table 1) and compared with Table 2(a)). The phonemes on those verses are not related to differentiation regardless of Quranic accents (*Qiraat*) and any mistakes while articulating those Quranic Arabic phonemes are not permissible in Islam.

Table 2(a): *Sūrah Al-Fātiḥah* (Verse 1, 2, 3 and 5) – All Qiraat

Verse No.	Verse	Phonetic
Verse 1	بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ	<B I S M I – L> <L A H I – R> <R AH HH M AA N I – R> <R AH HH II II II M>
Verse 2	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ	<AE A L HH A M D U> <L I L L A H I><R AH B B I – L> <EA AA L A M II II II N>
Verse 3	ٱلرَّحْمَٰنِ ٱلرَّحِيمِ	<AE A R R AH HH M AA N I – R> <R AH HH II II II M>
Verse 5	إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ	<AE I J J AA K A><N A EA B U D U><W A AE I J J AA K A> <N A S T A EA II II II N>

* Highlight in blue – misarticulate phonemes among Malay speakers (see Table 1)

3.1.2 Misarticulate Phonemes (Quranic accents)

On the other hand, the specific phonemes which classified as misarticulate phonemes were occurred at the places where the accent took place within the verse, and is underlined below (see Table 2(b), 2(c), 2(d) and 2(e)). Those letters located under a specific word of verses 4, 6, 7-part 1, and 7-part 2 in *Sūrah Al-Fātiḥah*, and can be differentiate based on Quranic accents (*Qiraat*). Focused on these verses, any changes occurred within the recitation, which regards to accent is acceptable in Islam. Otherwise, if the articulation were mistakenly recited with a wrong phoneme, it is forbidden in Islam.

Table 2(b): *Sūrah Al-Fātiḥah* (Verse 4)

Quranic accents	Verse	Phonetic
Imam ‘Aasim (Hafs) Imam Ya’akob (Ruwais)	مَٰلِكِ يَوْمِ ٱلدِّينِ	<M AA L I K I> <J AW M I-D> <D II II II N>
Imam Hamzah (Khalad) Imam Hamzah (Khallaf) – facet 1 Imam Hamzah (Khallaf) – facet 2 Imam Ibn Kathir (Al-Bazzi) Imam Ibn Kathir (Qunbul)	مَلِكِ يَوْمِ ٱلدِّينِ	<M A L I K I> <J AW M I-D> <D II II II N>

Table 2(c): *Sūrah Al-Fātiḥah* (Verse 6)

Quranic accents	Letters	Verse	Phonetic
Imam ‘Aasim (Hafs) Imam Ya’akob (Bazzi)	(ص)	ٱهۡدِنَا ٱلصِّرَٰطَ ٱلۡمُسۡتَقِيمَ	<AE I H D I N AA-SS> <SS I R AA TH A-L> <M U S T A Q II II II M>
Imam Hamzah (Khalad) Imam Hamzah (Khallaf) – facet 1	(ز)	ٱهۡدِنَا ٱلزِّرَاطَ ٱلۡمُسۡتَقِيمَ	<AE I H D I N AA-Z> <Z I R AA TH A-L> <M U S T A Q II II II M>
Imam Hamzah (Khallaf) – facet 2 Imam Ibn Kathir Qunbul) Imam Ibn Kathir (Ruwais)	(س)	ٱهۡدِنَا ٱلسِّرَاطَ ٱلۡمُسۡتَقِيمَ	<AE I H D I N AA-S> <S I R AA TH A-L> <M U S T A Q II II II M>

Table 2(d): *Sūrah Al-Fātiḥah* (Verse 7, Part 1)

Quranic accents	Letters	Verse	Phonetic
Imam ‘Aasim (Hafs) Imam Ya’akob (Bazzi)	(ص)	صِرَٰطَ ٱلَّذِينَ أَنۡعَمۡتَ عَلَيۡهِمۡ	<SS I R AA TH A – L> <L A DH II N A> <AE A N EA A M T A> <EA A L AJ H I M>
Imam Hamzah (Khalad)	(ص)	صِرَٰطَ ٱلَّذِينَ أَنۡعَمۡتَ عَلَيۡهِمۡ	<SS I R AA TH A – L> <L A DH II N A> <AE A N EA A M T A> <EA A L AJ H U M>
Imam Hamzah (Khallaf) – facet 1	(ز)	زِرَاطَ ٱلَّذِينَ أَنۡعَمۡتَ عَلَيۡهِمۡ	<Z I R AA TH A – L> <L A DH II N A> <AE A N EA A M T A> <EA A L AJ H U M>
Imam Ibn Kathir (Ruwais)	(س)	سِرَاطَ ٱلَّذِينَ أَنۡعَمۡتَ عَلَيۡهِمۡ	<S I R AA TH A – L> <L A DH II N A> <AE A N EA A M T A> <EA A L AJ H U M>
Imam Hamzah (Khallaf) – facet 2 Imam Ibn Kathir (Qunbul)	(س)	سِرَاطَ ٱلَّذِينَ أَنۡعَمۡتَ عَلَيۡهِمُو	<S I R AA TH A – L> <L A DH II N A> <AE A N EA A M T A> <EA A L AJ H I M UU>

Table 2(e): *Sūrah Al-Fātiḥah* (Verse 7, Part 2)

Quranic accents	Verse	Phonetic*
Imam ‘Aasim (Hafs)	غَيۡرِ ٱلۡمَغۡضُوبِ عَلَيۡهِمۡ وَلَا ٱلضَّآلِّينَ	<GH AJ R I – L> <M A GH DD UU B I> <EA A L AJ H I M> <W A L AA – DD> <DD AA AA AA L L II II II N>
Imam Ya’akob (Khalad) Imam Hamzah (Khallaf) – facet 1 Imam Hamzah (Ruwais)	غَيۡرِ ٱلۡمَغۡضُوبِ عَلَيۡهِمۡ وَلَا ٱلضَّآلِّينَ	<GH AJ R I – L> <M A GH DD UU B I> <EA A L AJ H U M> <W A L AA – DD> <DD AA AA AA L L II II II N>
Imam Hamzah (Khalaf) – facet 2 Imam Ibn Kathir (Bazzi) Imam Ibn Kathir (Qunbul)	غَيۡرِ ٱلۡمَغۡضُوبِ عَلَيۡهِمُو وَلَا ٱلضَّآلِّينَ	<GH AJ R I – L> <M A GH DD UU B I> <EA A L AJ H I M UU> <W A L AA – DD> <DD AA AA AA L L II II II N>

* Word in English (phonetic level transcription)

Properties of the accent’s phonemes (see Table 2(c) and 2(d)) was represented by phoneme of letters *Sad* /S/-(ص), *Seen* /s/-(س), and *Zai* /z/-(ز). These three phonemes are multiple letters that construct the word “*Ssiraatha”* (صِرَاطَ) for verse 6 and 7 in *Sūrah Al-Fātiḥah*. Although these three phonemes were produced from the same articulation point, the way those phonemes were pronounced was different. It is because each of this phoneme represents its features and traits, known as ‘*sifat.’* These three phonemes have been chosen for assessment because the differentiation-related to the accents are located in those words. The manner of articulations and common mistakes of these 3 letters of phonemes are shown below.

*Table 3: Properties and features for letters ***Sad* (ص)**, ***Seen* (س)**, and ***Zai* (ز)**

Features	**Hijaiyah Letters**		
	Sad (ص) - /S/	**Seen (س) - /s/**	**Zai (ز) - /z/**
Consonant type	Fricative		
Phonation	Unvoiced - Pharyngealized	Unvoiced (no vibration of the vocal cords)	Voiced (vocal cords vibrate)
Manner of articulation	**Pharyngealization** the – secondary articulation of consonants (pharynx or epiglottis is constricted)	• produced by tightening the air movement to produce a disturbance	
Place of articulation	**Articulation point –** Alveo-dental Alveolar consonants		
	• **The tongue –** emitted by placing the top pointer of the tongue on the plates of both front lower incisors, in between the upper and lower incisors • A little space left in between the top pointer of the tongue and the plates of the teeth. • "plate" - refers to the long axis of the tooth. The long axis is on the inner side, rather than the outer side of the teeth. • Also known as 'whistle' letters (associated whistle type sound heard when correctly emitted)		
	• Letter **Sad (ص)** - the *tafkhiim* (heavy) letter, and has the characteristic of sticking. • Need to pronounce with bold sound	• Letter **Seen (س)** - the thin voice (light) letter • Need to pronounce lightly	• Letter **Zai (ز)** - the stressed letter • Need to pronounce with stress sound
Distinctive Phonetic Features (DPF)	• Described as **/sˤ/** [3]. • Consonant – Fricative – Unvoiced – Emphatic • The phoneme /sˤ/ - Sad (ص) – found **ONLY in Arabic language (MSA)**. Thus, it is considered unique.	• Describes as **/s/** [3]. • Consonant – Fricative – Unvoiced – Non-Emphatic • The phoneme /s/ - Seen (س) - found in Arabic language (MSA) and English, as well as most of languages in the world	• Described as **/z/** [3]. • Consonant – Fricative – Voiced – Non-Emphatic • The phoneme /z/ - Zai (ز) - found in Arabic language (MSA), English, and other languages in the world
Common Mistakes	• If the **Sad (ص)** is not articulate heavy enough, it sounds just like, or very close to a **Seen (س)**	• If the **Seen (س)** is articulate in heavy way, it sounds just like, or very close to a **Sad (ص)**	
	• Lack of 'whistle (الصّفير)' vibration. • This problem occurred due to a strong overbite.		
Solutions	• Speakers with overbite problem still can learn to pronounce these letters correctly by making compensation in the lower mouth (jawline). • The lower jawline should be protruded until it aligns with the upper jaw while articulate, and proper 'whistle' sound can be heard clearly. The tongue also should not press up against the plates of the teeth or the sound will be improperly locked up when trying to articulate these letters.		

* Based on Arabic and English consonants [3],[30],[31].

According to [5],[8],[9],[27],[54], the phoneme of letters *Seen* /s/-(س) and *Theh* /θ/-(ث) has a similar sound (see Table 1), but in fact the articulation point is different. The articulation points of consonants *Theh* /θ/-(ث) is placed at the inter-dental area, which differed with *Seen* /s/-(س) (see Table 3). It is frequently pronounced like a 'whistle' group (*Seen* /s/-(س), *Zai* /z/-(ز) & *Sad* /S/-(ص)), instead of their unique articulation point by mistake. The top pointer of the tongue needs to collide or slip up along with the bottom edges of both top front incisors and not the inner side of the plates. This kind of mistake can occur by both Arabs and non-Arabs.

3.2 Prosodic Features

Prosody deals with acoustic qualities of sound. This feature is essential for the intelligibility and proficiency of stress-timed languages such as Arabic [44]. The lexical stress is phonetically realized through the manipulation of three acoustic-phonetic variables: (1) the signal intensity (energy), (2) the fundamental frequency (F0) (or pitch), and (3) syllable duration [10], as well as, an additional feature of (4) spectral-tilt proposed by [55]. The consonants encapsulate to the acoustic features were thoroughly analyzed in this study. A significant variation of features for each phoneme helped in distinguishing one accent from the other. It will focus on multiple Quranic accents, which tested on Malay reciters.

In this section, the analysis on verse 6 and 7 in *Surah Al-Fātiḥah* for 3 words, which are "*Siraatha*" for phoneme /s/, "*Ziraatha*" of phoneme /z/ and "*Ssiraatha*" of phoneme /S/ have been conducted. Only these 3 phonemes were selected for evaluation, in order to see how the prosodic features differed from one consonant to another based on Quranic accents. The Praat tool [56] and MATLAB Voice Source [57],[58] has been used for development, purposely for evaluating the potential acoustic features correlates of the accent, such as fundamental frequency (F0)/pitch, energy and duration. Here, F0 measurement was estimated using Praat and Straight tools [59]. Both tools are very flexible for speech analysis in phonetics, which provide a variety of standard and non-standard procedures, that include spectrographic analysis, articulatory synthesis and neural networks. F0 values were estimated using the Praat and Straight algorithm [60], over the selected words in verse 6 and 7. For each target word, polynomial fitting was computed over the F0 values, due to smooth the raw contour and to make sure the detection of minimum/maximum F0 is more precise and robust. In this study, each word was manually checked to make sure the fitting of F0 values was accurately represented. The weighted least square error criterion, E_a has been used in optimization of the $a_i's$ through equation (1).

$$E_a = \sum \left(y(n) - \sum_{i=0}^{N} a_i P_i(n) \right)^2 \cdot W(n) \quad (1)$$

Meanwhile, the energy measures were calculated by an adaptive window size, particularly for the effects of F0. The window size at the particular point in time was set to the pitch periods, as determine by the Straight and Praat measurement. Lastly, the duration of each stress consonants and vowels were obtained from the manual segmentation. Here, the onset and offset time were distinguished, as well as the evidence of syllables closure or release, such as the abrupt fall in signal amplitude. The results of the analysis were categorized based on gender of the speakers. It is due to the acoustical differences between male and female speakers, thus attributed to the gender differences [61].

3.3 Spectral Features

The majority of current speaker and language recognition systems rely on the spectral features, that derived through short-time spectral analysis of the speech signal. The spectral envelope was responsible for identifying the shape of the vocal tract during the production of sentences. In this research, a widely used feature extraction of Mel-Frequency Cepstral Coefficient (MFCC) has been proposed to represent the spectral features in the front-end processing stage.

3.3.1 Pre-processing of Speech Signal

The original sampling frequency of the speech samples is 44.1kHz. For processing purposes, each of the samples is down sampled to 16kHz. The extraction and recognition processes were implemented in MATLAB. First, the sample speech $x(n)$ is pre-processes by applying the pre-emphasis with a first-order highpass filter. The concept of the pre-emphasis is to amplify the high-frequency components obtained by the highpass filter and remove the DC components in the speech signal. Equation (2) exemplified the filtered output:

$$H(z) = 1 - \alpha * z^{-1}, \qquad 0.9 \leq \alpha \leq 1.0 \tag{2}$$

Afterward, $x'(n)$ need to be degraded by multiplying the following Hamming window function, due to compensate the overlap between the neighboring frames and to minimize the signal discontinuities at the beginning and end of the frames.

$$x_t(n) = x'(n) \cdot \left\{0.54 - 0.46 \cdot cos\left(\frac{2\pi n}{N-1}\right)\right\} \tag{3}$$

The matter of fact, the speech signal can be frame into 20-30ms, and in this experiment, the successive overlapping frames used is 20ms or 25ms, with an overlap rate of 10ms. Thus, the different size of the window is not considered a big issue to be highlighted in this case, although there might be some differences in the recognition result.

3.3.2 Parameterization of Speech

Speech parameterization is an important step in speech recognition systems. It is used to extract relevant information, such as voices (phonemes) from the audio signal [62]. The speech parameterization technique investigated in this study is known as Mel-Frequency Cepstral Coefficient (MFCC), that used to extract speech info within the speech signal. The spectral envelope identifies the shape of the vocal tract during the production of sentences. Mel-Frequency Cepstral Coefficient (MFCC) is used in state-of-the-art speech processing systems [48],[63],[64],[65], and it is proven to be among the most effective spectral feature representatives in speech-related tasks. Besides, the MFCC algorithm also shown to work exceptionally well for the Arabic ASR application [18]. MFCC successfully exploits human auditory principles and the decorrelating property of the cepstrum [66]. Due to this characteristic, the MFCC system also widely used as a state-of-the-art in the Language Identification (LID) system.

In the previous discussion (see Section 2.0), the MFCC algorithm also shown to work very well for the Arabic ASR application. The computational steps of MFCC feature extraction are illustrated in Figure 1, which extract MFCC vectors for a sound frame x. In order to analyze $x_t(n)$ in the frequency domain, firstly, the N-point Fast Fourier transform (FFT) applies to compute spectral coefficients from each frame and convert into frequencies. The result is further weighted by a series of triangular filters to gain Mel spectral coefficients. The placements of these filters are equally spaced together with the Mel frequency, which is mathematically associated with the signal linear frequency of f by the following formula [67].

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

The following step is to apply the Discrete Cosine Transform (DCT) on the log energies of the triangular bandpass filters to obtain the Mel-scale cepstral coefficients. By taking the DCT, the greater coefficients are eliminated, and simultaneously, the spectral shape of the signal is taken. Based on Figure 1, basic MFCC computational steps are computed till Inverse Discrete Cosine Transform (IDCT) or energy process. However, it is up to research necessities to expand or add the features related to the changes in cepstral features (the slopes) over time. The matter of fact, if Fourier transforms outputs are straight to be used as a final feature vectors, the process itself generally incapable of providing such as a better result in returning the essential details in lower coefficients. Typically, the lower-order coefficients represents the spectral form, while the higher-order coefficients are noise-like features. Additionally, in speech analysis, the certain amplitude at different frequencies is much less critical than the general shape of the spectrum.

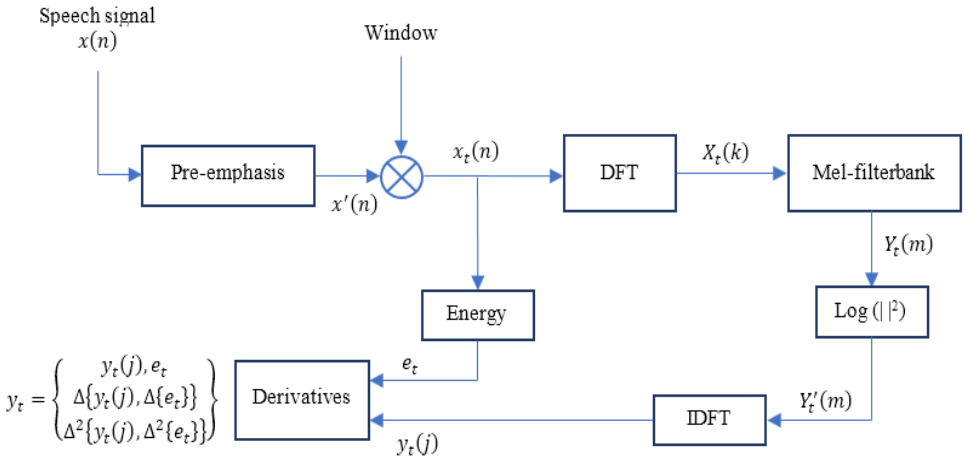

Fig. 1: MFCC Computational Steps for Feature Extraction

Finally, it is often helpful to compute the time derivatives in MFCC as new features, which ideally shows the velocity and acceleration of each MFCC feature vectors. These additional feature vectors are known as delta and double delta features. The speech features, which are the time derivatives of the spectrum-based speech features, are known as dynamic speech features. The features described previously have not captured the dynamic of the spectral changes (slopes). Thus, the derivative is essential to obtain temporal information. [68] have shown that, the system performance might be enhanced by adding time derivatives to the static speech parameters [69]. The first and second-order dynamic features (Delta and delta-delta cepstral coefficients) are usually appended and tested based on the MFCC acoustic vector [70],[71],[72].

3.4 Front-end Configuration (Spectral and Prosodic Features)

One of the crucial issues in ASR is the order in which the feature processing components appear in recognition. Figure 2 demonstrates how feature processing executed in this research is performed in ASR and accent recognition systems. Integrating spectral feature and prosodic features are considered our proposed implementation in this research. The spectral feature will be concentrated on MFCC as a feature extraction, meanwhile, prosodic features are derived from pitch, energy, and spectral tilt. The research in spoken language recognition has the primary focus, that concentrates on spectral information using the acoustic property of sound units (refer as acoustic phonetics) and their sequencing (refer as phonotactics)[73],[74]. Language and speaker recognition systems based on spectral features are performed well in favorable acoustic conditions, but their performance could deteriorate, due to the noise and unmatched acoustic conditions [43]. Prosodic features acquired from the pitch, energy, and duration are relatively less affected by the channel variations and noise [75]. Although the systems based upon the spectral features outperform the prosody-based systems, their integrated performance might offer the necessary improvement for recognition systems.

In this section, the proposed procedure which involved the integration of prosodic feature and spectral feature have been presented. The same sample data as computed previously with a similar computational process of MFCC were primarily performed in this experiment as the core feature extraction algorithm. Firstly, the speech data were downsamples from 44.1kHz to 16kHz during the pre-processing stage. Here, frame-level of MFCCs, and frame-level pitch and energy are estimated from the waveform during the segmentation process (see Figure 2). Three prosodic features alongside the complete spectrum of the speech signal (MFCC) were then computed in 25ms windows, with 10ms overlap. F0 contour for the voiced segment was extracted from each utterance, and due to preserve the temporal continuity of the F0 contour, the unvoiced segments contour was generated from neighboring voiced of F0 values through linear interpolation. The envelope of energy was computed as;

$$E = \sum_{n=n_1}^{n_2} |x[n]|^2 \tag{5}$$

Where, the x is the speech waveform and n_1, n_2 stand for beginning and the end of the Hamming window, respectively. The MFCC values were obtained after computing the logarithmic value of Mel-scale, and performing the Discrete Cosine Transform (DCT) on the resulting spectrum, later on. From the final results of MFCC, the spectral tilt was represented by the first MFCC coefficients, while the full band of spectrum was represented using the 12 first MFCC coefficients.

3.5 Classification

This section explains the fundamental theory of the Gaussian Mixture Model (GMM). The input for the classifiers is the set of feature vectors discussed above. Conventional validation was performed with 70% of dataset used for training and 30% of dataset used for testing. GMM is a well-studied statistical method used to model speakers or languages/accents. In pattern recognition, GMM is used to generate speaker models with different accents, as well as to match different patterns against the trained models. Multi-variate Gaussian distribution is represented as:

$$P(x|\theta_m) = \frac{1}{(2\pi)^{pm/2} det(\Sigma_m)^{1/2}} exp\left(-\frac{1}{2}(x-\mu_m)^S \sum_m^{-1}(x-\mu_m)\right) \quad (6)$$

In creating models for accents, a set of observations (features vector) used in this research has to be statistically described by GMM. Prior information about statistical distribution, a likelihood function is computed.

$$L(X;\theta) = \prod_{i=0}^{N} p(x_i;\theta) \quad (7)$$

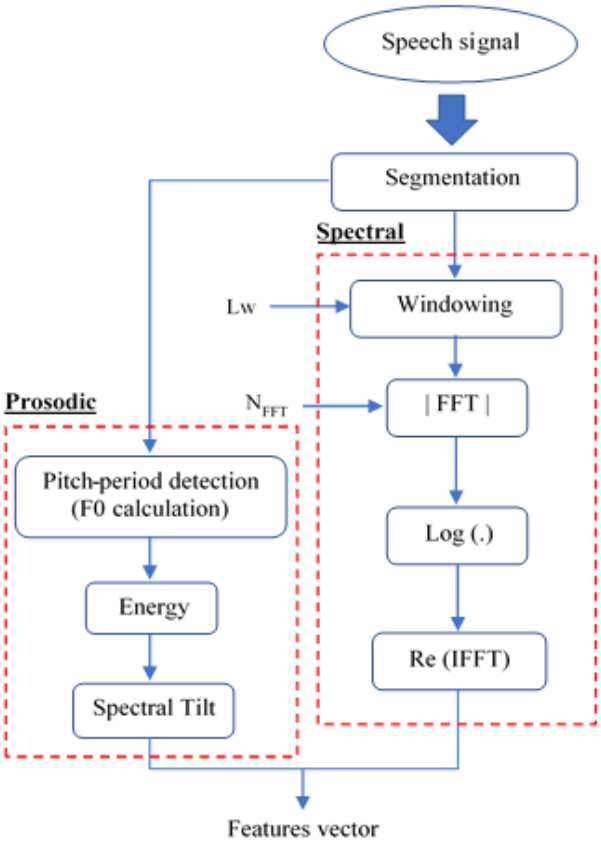

Fig. 2: Block diagram of the proposed spectral and prosodic features in the speech signal

4.0 RESULTS AND DISCUSSION

4.1 Prosodic Evaluations

In this part, the prosodic features for a few letters where its phonemes are differed, based on Quranic accents (*Qiraat*) are shown in Table 4. Those phonemes have been used in the verse 6 in *Sūrah Al-Fātiḥah,* which constructed the word "*Ssiraatha*" (صِرَاطَ).

Table 4: Verse 6 – Comparison of the median, mean-energy, pitch and time-frame for consonant /S/, /z/, and /s/

Accents	Phone-me	Male				Female			
		Median (dB)	Mean-E.(dB)	Pitch /F0 (Hz)	Time (sec)	Median (dB)	Mean-E.(dB)	Pitch/F0 (Hz)	Time (sec)
'Aasim (Hafs)	**/S/ -** (ص)	55.566	**58.416**	**156.78**	0.2203	55.932	**57.752**	**266.99**	0.2180
Hamzah (Khalad)	**/z/ -** (ز)	58.108	**59.736**	**137.13**	0.2230	60.025	**61.580**	**222.18**	0.2161
Hamzah (Khallaf) - facet 1	**/z/ -** (ز)	57.194	59.936	141.27	0.2230	59.629	61.279	223.37	0.2168
Hamzah (Khallaf) - facet 2	**/s/ -** (س)	51.8634	**55.616**	**171.52**	0.2705	53.968	**56.108**	**276.25**	0.2403
Ibn Kathir (Bazzi)	**/S/ -** (ص)	55.188	58.614	161.29	0.2364	56.758	58.458	270.02	0.2170
Ibn Kathir (Qunbul)	**/s/ -** (س)	51.1994	55.376	170.96	0.2755	54.384	56.108	269.19	0.2397
Yakob (Ruwais)	**/s/ -** (س)	54.0954	56.174	171.36	0.2775	53.658	55.544	269.46	0.2379

***Legend:** /S/ - Sad (ص), /z/ - Zai (ز) and /s/ - Seen (س), based International Phonetic Association (IPA) for Arabic [28]

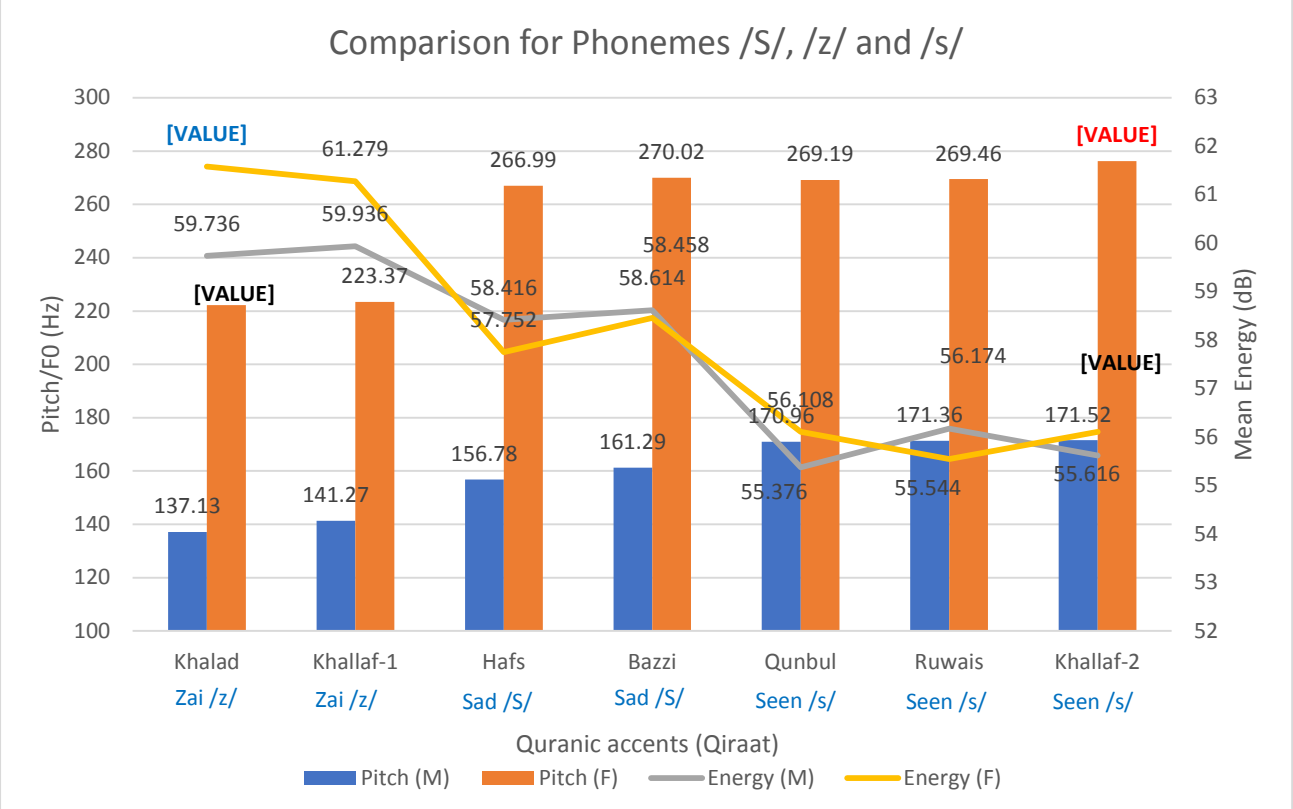

Fig. 3: Comparison graph of Pitch (F0) and Energy for phonemes Sad /S/, Zai /z/ and Seen /s/

Based on the results listed in Table 4 and graph in Figure 3, most speakers achieved a higher fundamental frequency (F0) values of pitch accented for word "*Siraatha*" with a letter of phoneme /s/-(س), compared to other words of "*Ziraatha*" of phoneme /z/-(ز) and "*Ssiraatha*" of phoneme /S/-(ص). In other words, phoneme /s/-(س) achieved the most high pitch value compared to others, but gained the lowest energy. Meanwhile, the phoneme /z/-(ز) obtained the highest energy value, but the pitch value was the lowest. The results indicated that, each phoneme has its own cues and traits that distinguish one accent to another. As expected, all the three words indicated a higher value of pitch (F0) from the female speakers rather than the males. However, the energy (median & mean) values for the phoneme /z/-(ز) is preferably higher compared to the phoneme /s/-(س) and phoneme /S/-(ص) (less energy), for both male and female. This indication described a more abrupt closure of the consonant and vocal folds on stressed syllables, which significantly agreed with [50]. Contradict with pitch (F0) values, if the energy value for a particular substitute has a lower energy, thus the pitch value is considered higher. The value of pitch is relatively more dominant towards the consonants Seen /s/-(س) for both male and female. The time duration for all phonemes are the same within 0.2 seconds.

Figure 4(a), 4(b) and 4(c) have provide a better illustration on those letters, where the analysis on the recorded words of verse 6 had shown a general view on how the prosodic features differed from one consonant to another, based on Quranic accents. The analysis was carried out using MATLAB coding [57],[58], and the value of Fundamental Frequency (F0)/pitch was estimated using Praat and Straight algorithms [59]. In "*Siraatha*", the correlation in between the fundamental frequency using Praat (pF0) and the fundamental frequency using Straight (strF0) achieved a good result for pitch accent, with the similarity of contour shape and maximum value of the peak in pitch obtained from all the speakers. However, the value of energy that represented the stress syllables of "*Si*" showed the minimum value for both male and female speakers. This result has totally differed when compared to the stress syllable of "*Zi*" from the word of "*Ziraatha*", with a maximum value of energy and minimum value of pitch (pF0 and strF0).

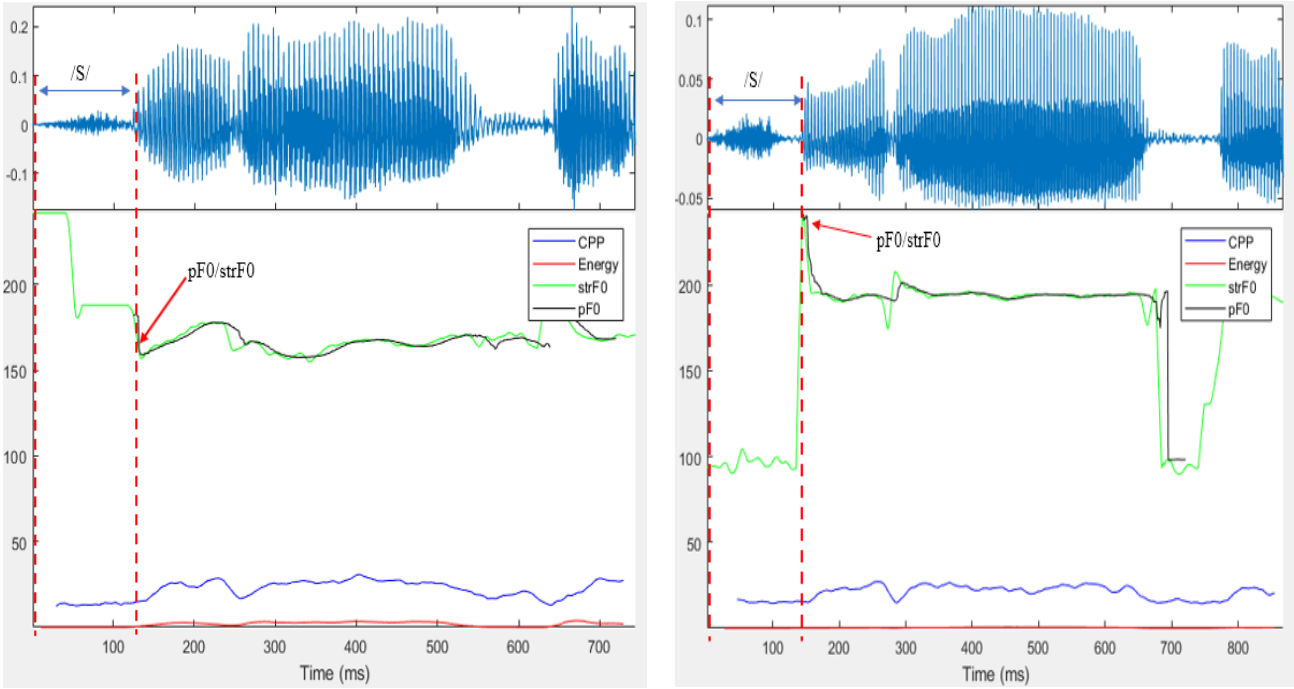

Fig. 4(a): Graph analysis for phoneme /S/-(ص) in word *Ssiraatha* for Male (Left), and Female (Right)

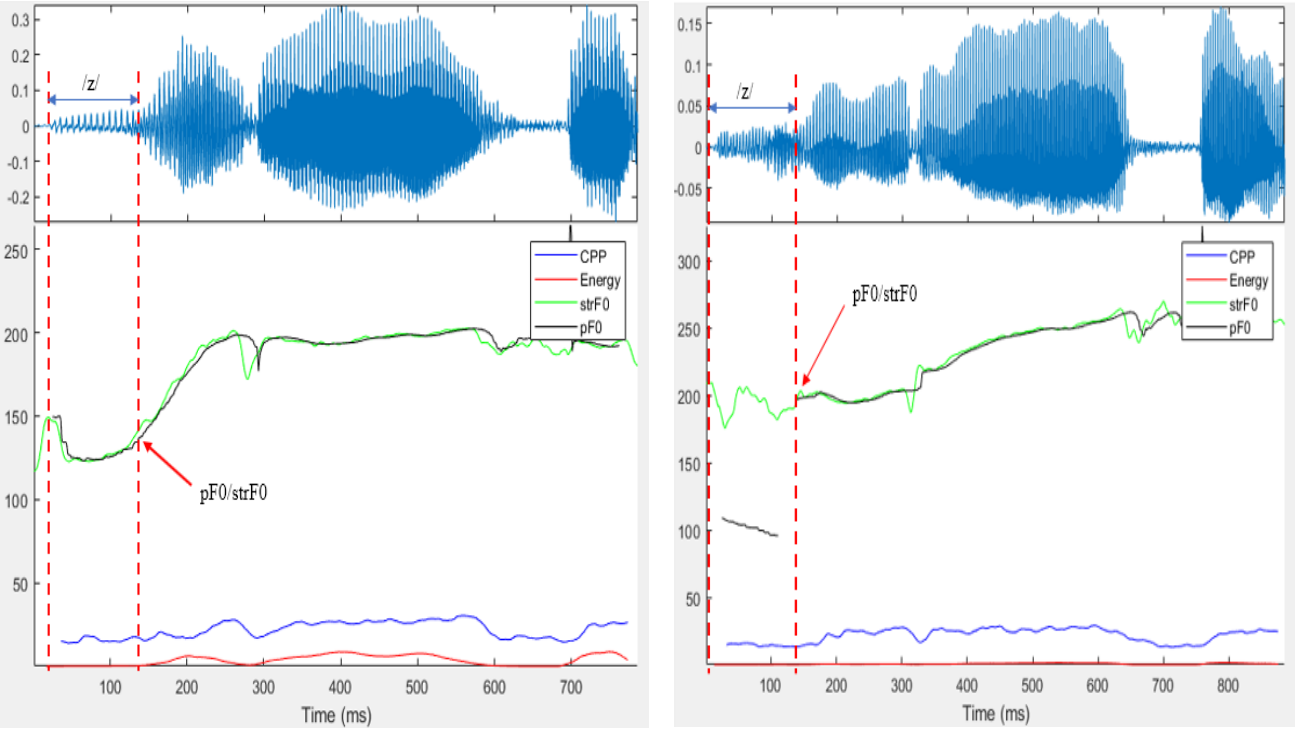

Fig. 4(b): Graph analysis for phoneme /z/-(ز) in word *Ziraatha* for Male (Left) and Female (Right)

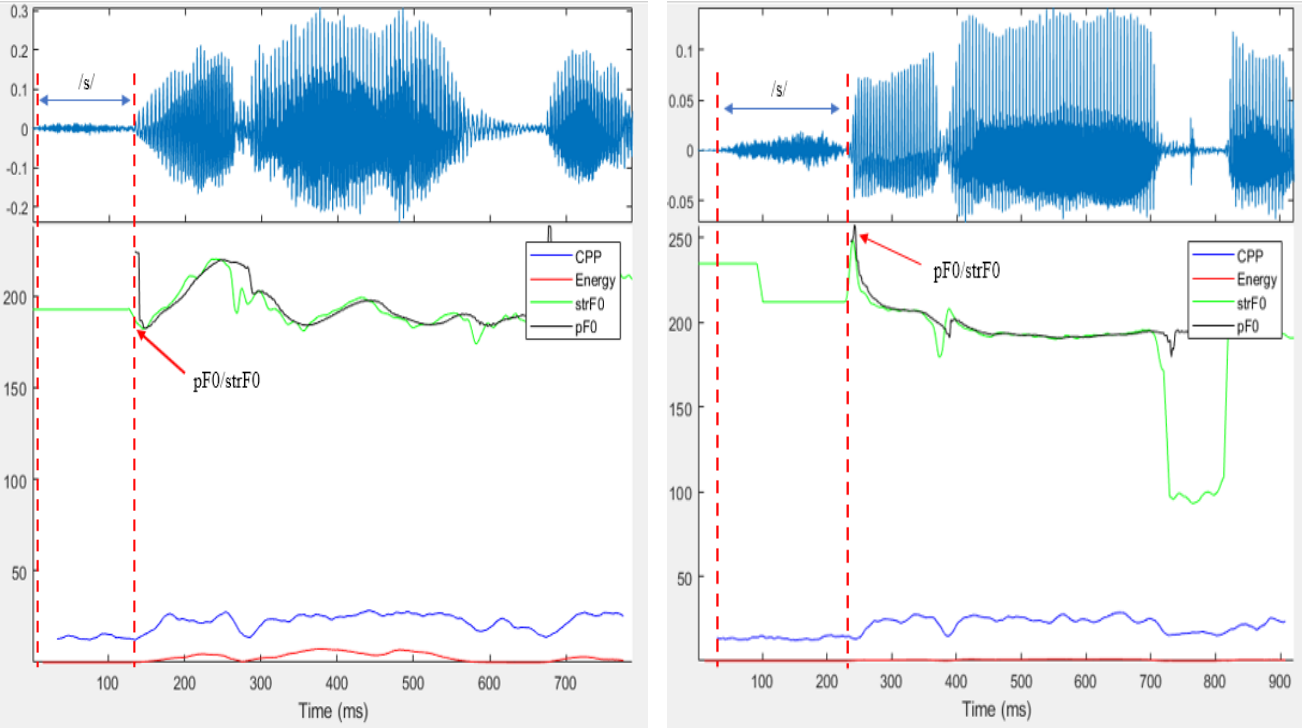

Fig. 4(c): Graph analysis for phoneme /s/-(س) in word *Siraatha* for Male (Left) and Female (Right)

4.2 Spectral Evaluation

The most widely known feature extraction is based on MFCC, as it is usually associated with a high recognition accuracy [47]. MFCC feature extraction has been used as a parameterization of choice for many speech recognition applications, including in this experiment. Technically, the parameterization from MFCC gives good discrimination and lends themselves to several manipulations.

The recorded data-set was divided into two subsets, the train-set and the test-set. The samples data (audio) were collected in the form of sentences (verses) and words. For word level stage, the total samples (audio data) are 1568 sentences (verses), which equivalent to 5,684 words, where 70% of the data samples were used for modeling, while 30% were used for the testing phase. Data samples belonging to each class are equally distributed. Each column adds up to give the number of samples (testing) belonging to each class. Here, 116 (14%) samples belong to each class namely; *Hafs 'Asim, Khalad An-Hamzah, Khallaf An-Hamzah* for facet (1) & (2), *Al-Bazzar Ibn Kathir, Qunbul Ibn Kathir* and *Ruwais An-Ya'akob.* The reason is to evaluate the difference of reading Quranic accents between each of Imam.

Acoustic models of words were estimated using the Generative Training approach through the Maximum Likelihood (ML) estimation method. Once the recognizer got trained, recognition accuracies were observed for both training and test data. Table 5 shows the benchmark system and the spectral features are used for speech based on accents. The system is evaluated one by one using MFCC based feature extraction and Gaussian Mixture Model (GMM). Here, the results of recognition accuracies for Quranic accents of *Sūrah Al-Fātiḥah* using MFCC as a feature extraction alone are tabulated in Table 5 below. For recognizing speech utterances, the conventional ASR system based on GMMs and k-Nearest Neighbor (kNN) are utilized. Later, both classifiers were compared due to find the best technique used for modeling and recognition purposes. Performance results of front-end processing were described stages by stages, due to provide a clear view of the significant improvement for each and individual features applied. For GMM, the experiments were conducted using a single number of Gaussian mixtures model and only left-to-right transitions without skips were allowed. Different values of dimensional features have been compared in this experiment, due to determine the optimal feature that able to give a higher accuracy of recognition as the finest results.

Table 5: The percentage accuracies of recognition based on spectral features using MFCC & GMM

No.	Feature Extraction	Classification	Train-set		Test-set	
			Accuracy	EER	Accuracy	EER
1.	MFCC (25-dimensional features) [14]	Gaussian Mixture Model (GMM)	73.66	26.34	66.07	33.93
2.	MFCC (39 dimensional features) [12],[13],[15],[16],[18]	Gaussian Mixture Model (GMM)	84.22	15.78	74.41	25.59

The 39-dimensional features vector of MFCC using 13 MFCC with 13Δ and 13ΔΔ features of acceleration coefficients, showed a better result compared to the 25-dimensional features vector, that using 13 MFCC with 12Δ without absolute Energy, and Cepstral Mean Normalization (CMN). The same 39-dimensional features vector of MFCC, also been conducted previously by [12],[13],[15],[16],[18], in their research. The only different is limited to the certain parameter related to the window size of 20ms and 25ms, implemented by [15],[18], respectively. In fact, both sizes are still acceptable to be implemented in MFCC.

Other than GMM, other classification for modeling and testing also been executed, known as k-Nearest Neighbors (k-NN). In this case, using 39-dimensional features of MFCCs as a spectral features algorithm, along with k-NN as a classifier, representing a better result compared by using GMM as classification. The results are shown below in Table 6.

Table 6: The percentage accuracies of recognition based on spectral features using MFCC & k-NN

No.	Feature Extraction	Classification	Train-set		Test-set	
			Accuracy	EER	Accuracy	EER
1.	MFCC	K-Nearest Neighbors (k-NN)	85.26	14.74	75.45	24.55

4.3 Prosodic and Spectral Features Evaluations

In this sub-section, the feature extraction used in this experiment involved the combination of ASR and language identification in Quranic accents. It correlates between the language and stress detection model's features over lexical, based on duration, pitch, energy, spectral tilt and MFCC-based measurement across the consonant-vowel or syllable nuclei. These different values obtained are significantly varied for each accent, which shows the mismatch of consonants and vowels. This can be used as an essential cue to the classifier for decision-making, especially for accent identification and classification. A novel aspect of our system is the successful integration of spectral information (MFCCs and spectral tilt) and prosodic (duration, pitch and energy) information. For the phonetic feature extraction, the spectral implementation using the MFCC alone or the combination of spectral (MFCC) and prosodic (pitch, energy, spectral tilt) were evaluated based on the final results of recognition accuracy at the testing phase.

Acoustic models of phonemes were first estimated using Generative Training known as the Maximum Likelihood (ML) estimation approach. It is based on the values of iterations and log-likelihood, which are gained from the Expectation Maximization-Gaussian Mixture Model (EM-GMM). Once the recognizer got trained, the recognition accuracies were monitored and tested for both train-set and test-set using the GMM. Here, the algorithm and approach that demonstrate the better extraction process-based accent, and the classification model that is able to present the better generalization capability of unseen speech data will be used for implementation.

Table 7: The percentage accuracies of recognition based on spectral and prosodic features using GMM

No.	Feature Extraction	Classification	Train-set		Test-set	
			Accuracy	EER	Accuracy	EER
1.	MFCC + Pitch	Gaussian Mixture Model (GMM)	87.72	12.28	77.890	22.11
2.	MFCC + Pitch + Energy	Gaussian Mixture Model (GMM)	89.061	10.939	80.113	19.887
3.	MFCC + Pitch + Energy + Tilt	Gaussian Mixture Model (GMM)	**89.697**	10.303	**81.713**	18.287

The average recognition accuracy for both train-set and test-set is depicted in Table 7. In this experiment, the results of MFCC were tested and presented in various stages and level, started from the combination of MFCC as a spectral feature and prosody of pitch, then added with another prosody element of energy, and lastly, computed with the spectral-tilt values. Here, the value of accuracy and the Equal Error Rate (EER), using 10-folds cross-validation techniques were calculated. Using the test-set data, results from Table 5 and Table 7 have shown significant improvement by +7.303% using the integration of spectral and prosodic features, better than implementing the MFCC alone. Meanwhile, the performance gained from train-set able to achieve a better result of +5.477 enhancement. It is proved that, the proposed feature extraction technique able to compete and gave a better recognition results, compared to the previous research that implemented the spectral features only [12],[13],[15],[16],[18]. Since the integration of spectral and prosodic is performed and able to give a better result as compared to others, thus, the confusion matrix based-Quranic accents are highlighted in Table 8(a) and Table 8(b) below. The results of test-set and train-set were reported separately, since the testing samples were obtained from two different group of people.

Table 8(a): Confusion matrix based on spectral & prosodic features (MFCC, Pitch, Energy & Tilt) and GMM (Test-set)

Quranic Accents	(MFCC + Pitch + Energy + Tilt) and GMM						
	Hafs	**Khalad**	**Khallaf_1**	**Khallaf_2**	**Bazzi**	**Qunbul**	**Ruwais**
Hafs	**83.192**	2.4327	2.9095	0.2291	1.9998	4.6068	4.6298
Khalad	2.1761	**79.417**	2.2826	0	4.9388	2.2317	8.9537
Khallaf_1	3.5633	1.6447	**80.4592**	3.0281	5.3011	0.50767	5.4960
Khallaf_2	8.2479	0.8891	4.0084	**81.6890**	0.2899	1.9327	2.9431
Bazzi	6.6958	0	0.6556	1.0165	**82.957**	3.9643	4.7103
Qunbul	6.1037	0.8198	0.6424	6.8084	2.2555	**80.5154**	2.8549
Ruwais	5.7390	0	4.2956	0	1.6757	4.5265	**83.7632**

Table 8(b): Confusion matrix based on spectral & prosodic features (MFCC, Pitch, Energy & Tilt) and GMM (Train-set)

Quranic Accents	(MFCC + Pitch + Energy + Tilt) and GMM						
	Hafs	**Khalad**	**Khallaf_1**	**Khallaf_2**	**Bazzi**	**Qunbul**	**Ruwais**
Hafs	**90.629**	4.8707	3.5152	1.2665	0	0.3277	0
Khalad	1.5102	**90.739**	3.2048	0	0.8277	2.512	1.2092
Khallaf_1	0.6954	0.9032	**93.579**	1.1741	0.2954	1.7818	1.5801
Khallaf_2	0.1278	0	1.3366	**91.059**	2.2908	2.5913	2.6035
Bazzi	1.3723	2.3887	1.2409	2.2168	**86.999**	2.4317	3.3596
Qunbul	0	1.2943	3.3307	1.1905	2.7682	**88.91**	2.5063
Ruwais	10.1159	2.3042	0	0	1.6189	0	**85.961**

Similar to the previous experiment using GMM, the overall results obtained in Table 10 was quite alike and parallel. As compared between GMM and k-NN classifier, the accuracy values gained by k-NN classifier using the MFCC alone (see Table 10) is slightly higher than GMM. The results have showed that, the performance (based accuracy) were much

better if the MFCC and pitch were combined altogether and applied with k-NN classification. However, the values of accuracy started to reduced and gradually regenerate after the energy elements is computed with the feature extraction process. In this case, the implementation of MFCC, pitch, energy and spectral-tilt as the feature extraction, alongside the GMM classification as for modeling and testing phase outperform k-NN classification for both phases.

Table 9: The percentage accuracies of recognition based on spectral features (MFCC, Pitch, Energy & Tilt) & k-NN

No.	Feature Extraction	Classification	Train-set		Test-set	
			Accuracy	EER	Accuracy	EER
1.	MFCC + Pitch	K-Nearest Neighbors (k-NN)	88.323	11.677	78.513	21.487
2.	MFCC + Pitch + Energy	K-Nearest Neighbors (k-NN)	88.397	11.603	79.470	20.53
3.	MFCC + Pitch + Energy + Tilt	K-Nearest Neighbors (k-NN)	**88.397**	11.603	**79.470**	20.53

Table 10: Overall results of recognition between GMM and k-NN Classification

No.	Feature Extraction	Classification (Testing-phase)				
		Algorithm	Train-set (%)	Error Rate (EER)	Test-set (%)	Error Rate (EER)
1.	MFCC [12],[13],[15],[16],[18]	Gaussian Mixture Model (GMM)	**84.22**	15.78	**74.41**	25.59
		k-Nearest Neighbor (k-NN)	85.26	14.74	75.45	24.55
2.	MFCC + Pitch	Gaussian Mixture Model (GMM)	87.72	12.28	77.890	22.11
		k-Nearest Neighbor (k-NN)	88.323	11.677	78.513	21.487
3.	MFCC + Pitch + Energy	Gaussian Mixture Model (GMM)	89.061	10.939	80.113	19.887
		k-Nearest Neighbor (k-NN)	88.397	11.603	79.470	20.53
4.	MFCC + Pitch + Energy + Spectral Tilt	Gaussian Mixture Model (GMM)	**89.697**	10.303	**81.713**	18.287
		k-Nearest Neighbor (k-NN)	88.397	11.603	79.470	20.53

Based on experimental results (see Table 10), the use of spectral (MFCC) and prosodic features (pitch, energy, spectral tilt) for feature extraction have aided the GMM recognition performance of the classifiers. The improvement for GMM is 2.243% higher than the k-NN classifier. The percentage of statistical measures using the integration process with GMM for classification is within the range of 81%-90%, with an accuracy rate of 81.7%. For other classifiers with spectral features alone, it falls by 1%-7%. Thus, accordance with accuracy and EER values, as well as recognition rate based on confusions matrices, it can be justified that integration of spectral (MFCC) and prosodic, alongside EM-GMM classifier outperforms other algorithms that using solely on the spectral features [12],[13],[15],[16],[18]. The comparison graph of feature extraction applied, using both GMM and k-NN classification are presented below for both test-set and train-set (see Figure 5(a) and 5(b)).

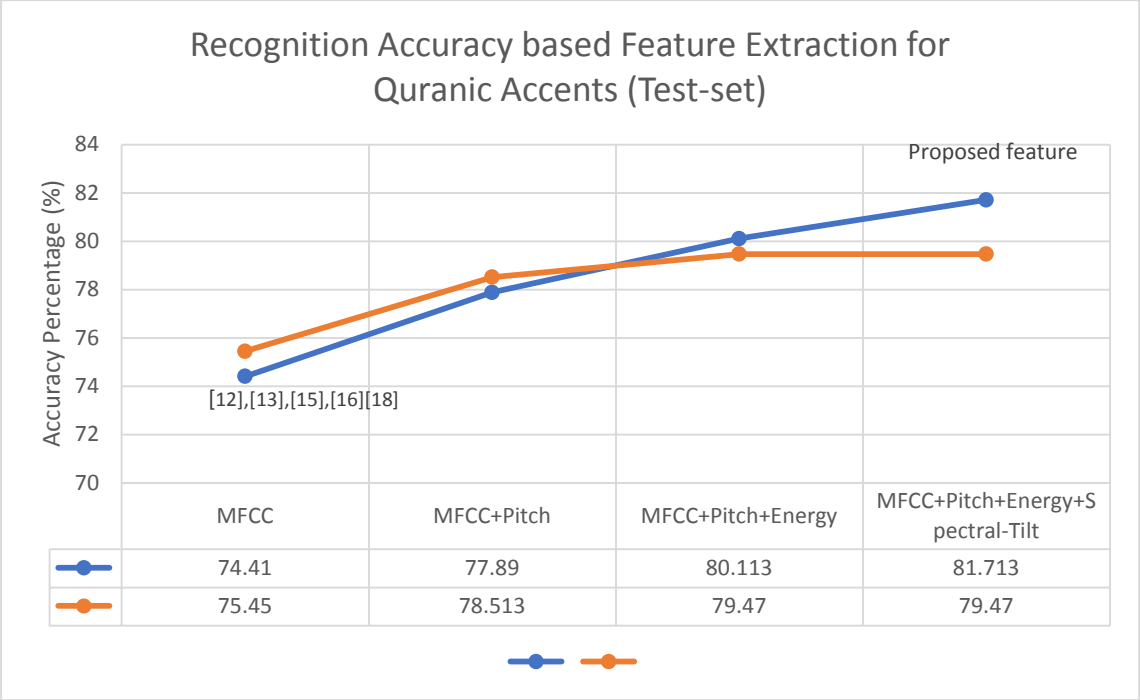

Fig.5(a): Comparison graph of the feature extraction implementation between GMM & kNN (Test-set)

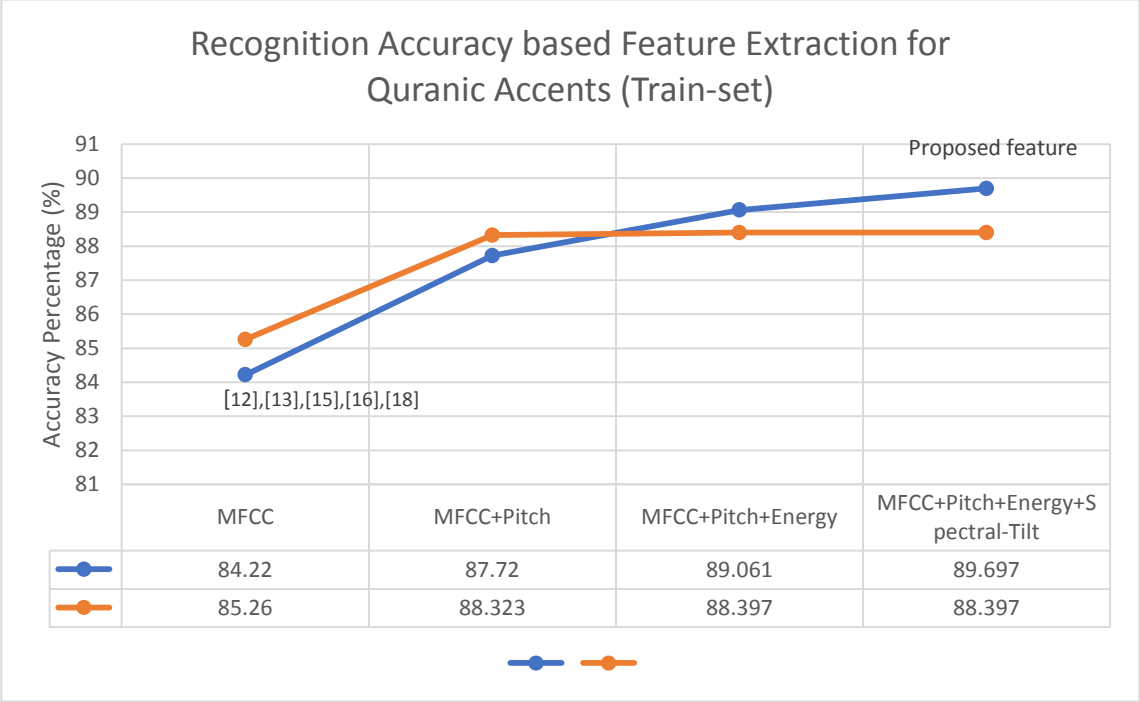

Fig.5(b): Comparison graph of the feature extraction implementation between GMM & kNN (Train-set)

5.0 CONCLUSION

Conclusions can be drawn through the recognition evaluation from the results of accuracy and Equal Error Rate (EER), specifically for both spectral (MFCC) and prosodic features, as well as spectral (MFCC) features alone. The accuracy and error values gained from the evaluation showed that, the integration of MFCC and prosodic features is more robust than MFCCs alone. The robustness of the ASR could depend on the better choices of the front-end processing technique applied after ML estimation. Thus, the better quality of features vector obtained from the feature extraction process may bring the better possibility and chances in terms of accuracy and performance at the recognition part.

In future work, we plan to extend our proposed design of feature extraction to be optimized with a better classification technique, mainly for Quranic accents recognition. The better design and the best option of feature extraction and classification algorithms implemented, might increase the accuracy and performance of recognition, as well as reduce the mismatch/errors in ASR process. Also, we have the intention to cover other chapters of Quranic Arabic accents.

REFERENCES

[1] L. E. A. Mingkuan, "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling", in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100),* 2000, 2, pp. II1025-II1028 vol.2.

[2] N. F. Chen, et al., "Characterizing Phonetic Transformations and Acoustic Differences Across English Dialects*". IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22 No. 1, 2014. pp. 110-124.

[3] Y. Alotaibi and A. Meftah, "Review of distinctive phonetic features and the Arabic share in related modern research*". Turkish Journal of Electrical Engineering & Computer Sciences*, Vol. 21 No. 5, 2013. pp. 1426-1439.

[4] Z. Mohd Don, et al., "How words can be misleading: a study of syllable timing and “stress” in Malay*". Linguistics Journal*, Vol. 3 No. 2, 2008.

[5] M. N. M. Hasbullah, *Analisis kesilapan sebutan bahasa Al-Quran di kalangan pelajar sekolah menengah*. 2001, Universiti Malaya.

[6] A. Mannan, *Optimization of Arabic Speech Recognition for Non-Native Speakers using Diverse Training Corpora*, in *Faculty of Computer Science and Information Technology*. 2013, University of Malaya.

[7] H. B. Dahan and A. Mannan, "Arabic Speech Pronunciation Recognition and Correction Using Automatic Speech Recognizer (ASR)", in *INTED2012: 6th International Technology, Education and Development Conference,* Valencia, Spain, 5-7 March, 2012, pp. 4009-4016.

[8] I. Abdullah, *Contemporary Issues of Al-Quran and Qiraat Studies in Malaysia*, in *Head of Department Al-Quran & Qiraat, Darul Quran JAKIM, Malaysia*, N. J. Ibrahim, Editor. 2017.

[9] M. L. Ibrahim, *Assessment and Evaluation for Quranic Recitation in National and International Competition*, in *Judge/Lecturer at Faculty of Quranic & Sunnah Studies, USIM*, N. J. Ibrahim, Editor. 2015.

[10] M. A. Shahin, et al., "Automatic classification of unequal lexical stress patterns using machine learning algorithms", in *2012 IEEE Spoken Language Technology Workshop (SLT),* pp. 388-391.

[11] N. Kamarudin, et al., "Feature extraction using Spectral Centroid and Mel Frequency Cepstral Coefficient for Quranic Accent Automatic Identification", in *2014 IEEE Student Conference on Research and Development,* 16-17 Dec. 2014, pp. 1-6.

[12] N. Kamarudin, et al., "Acoustic echo cancellation using adaptive filtering algorithms for Quranic accents (Qiraat) identification". *International Journal of Speech Technology*, Vol. 19 No. 2, 2016. pp. 393-405.

[13] N. Kamarudin, et al., "Analysis on Mel Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients as Feature Extraction on Automatic Accents Identification". *International Journal of Applied Engineering Research*, Vol. 11 No. 11, 2016. pp. 7301-7307.

[14] M. S. Abdo and A. H. Kandil, "Semi-Automatic Segmentation System for Syllables Extraction from Continuous Arabic Audio Signal". *International Journal of Advanced Computer Science and Applications*, Vol. 7 No. 1, 2016. pp. 535-540.

[15] M. Baig, et al., "Discriminative Training for Phonetic Recognition of the Holy Quran". *Arabian Journal for Science and Engineering*, Vol. 40 No. 9, 2015. pp. 2629-2640.

[16] M. O. Khelifa, et al., "Strategies for implementing an optimal ASR system for quranic recitation recognition". *International Journal of Computer Applications*, Vol. 172 No. 9, 2017. pp. 35-41.

[17] A. Mahmood, et al., "Automatic Speaker Recognition Using Multi-Directional Local Features (MDLF)". *Arabian Journal for Science and Engineering*, Vol. 39 No. 5, 2014. pp. 3799-3811.

[18] K. Nahar, et al., "Arabic Phonemes Transcription using Data Driven Approach". *International Arab Journal of Information Technology (IAJIT)*, Vol. 12 No. 3, 2015.

[19] A.-K. L. Al-Fatawi. *Hukum Imam Membaca Al-Fatihah atau Ayat-ayat yang lain dengan Qiraat yang Pelbagai (Law of Imam while Reciting Al-Fatihah or other verses in Different Qiraat (Accents) Style)*. Al-Kali li al-Fatawi: Mufti of Federal Territory of Malaysia 2019 [cited 2019 11 March 2019]; Available from: http://muftiwp.gov.my/en/artikel/al-kafi-li-al-fatawi/3186-al-kafi-1144-hukum-imam-membaca-al-fatihah-atau-ayat-ayat-yang-lain-dengan-qiraat-yang-pelbagai?highlight=WyJmYXRpaGFoIl0=.

[20] S. M. Zaini and N. C. Noh, *Imam tertinggal sepotong ayat ketika solat, makmum diminta qada solat Isyak (The Imam abandon 1 verse while prayer, and the followers were asked to replace the Isyak prayer)*, in *Berita Harian Online*. 2019, New Straits Times Press (M) Bhd.: Malaysia.

[21] S. M. Zaini, *Solat Isyak di Masjid Puncak Alam Isnin lalu tak sah (The Isyak Prayer at Puncak Alam Mosque Last Monday is Void)*, in *MyMetro*. 2019, New Straits Times Press (M) Bhd.: Malaysia.

[22] F. Barkatulla. *The Importance of Tajweed*. 2009; Available from: http://www.islamicawakening.com/viewarticle.php?articleID=934.

[23] A. Mohammed and M. S. Sunar, "Verification of Quranic Verses in Audio Files using Speech Recognition Techniques", in *1st International Conference of Recent Trends in Information and Communication Technologies (IRICT 2014),* Universiti Teknologi Malaysia, Johor, Malaysia, 12th -14th September, 2014, pp. 370-381.

[24] A. Z. S. A. Hadi, et al., "Enhancing Teaching and Learning Methodology with Computing Visualization in Studies of Qiraat (Malaysia)". *International Journal of Academic Research in Business and Social Sciences*, Vol. 8 No. 2, 2018. pp. 823-835.

[25] M. O. Alqahtany, et al., "Analyzing the Seventh Vowel of Classical Arabic", in *International Conference on Natural Language Processing and Knowledge Engineering, 2009 (NLP-KE 2009),* Dalian, pp. 1-7.

[26] M. Maamouri, et al., "Diacritization: A Challenge to Arabic Treebank Annotation and Parsing", in *Proceedings of the British Computer Society Arabic NLP/MT Conference,*

[27] N.-A. Abdul-Kadir and R. Sudirman, "Difficulties of standard arabic phonemes spoken by non-arab primary school children based on formant frequencies". *Journal of Computer Science*, Vol. 7 No. 7, 2011. pp. 1003.

[28] M. E. Ahmed, "Towards An Arabic Text-To-Speech System". *The Arabian Journal for Science and Engineering*, Vol. 16 No. 4B, 1991. pp. 565-583.

[29] M. Al-Bukhari, *Six major Hadith collections*. 9th century.

[30] *Manners when reading the Qur'an*. 2007 25 May 2017]; Available from: https://www.abouttajweed.com/tajweed-basics/quran-reading-manners.

[31] *Pronunciation of Letters*. 2014 [cited 25 May 2017; Available from: http://www.quranreading.com/Tajweed-Quran.

[32] J. C. Wells, *Accents of English*. Vol. 2. Cambridge University Press, New York, 1982.

[33] M. Mehrabani, et al., "Dialect distance assessment method based on comparison of pitch pattern statistical models", in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing,* 14-19 March 2010, pp. 5158-5161.

[34] S. Sinha, et al., "Dialectal influences on acoustic duration of Hindi phonemes", in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE),* 25-27 Nov. 2013, pp. 1-5.

[35] D. Mishra and K. Bali, "A Comparative Phonological Study of the Dialects of Hindi", in *ICPhS,* pp. 1390-1393.

[36] B. Gajic and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms", in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221),* 1, pp. 85-88.

[37] Q. Yan and S. Vaseghi, "Analysis, modelling and synthesis of formants of British, American and Australian accents", in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* 1, pp. I-I.

[38] S. Gray and J. H. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system", in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.,* pp. 35-40.

[39] J. C. Watson, *The phonology and morphology of Arabic*. Oxford University Press on Demand, 2002.

[40] S. R. H. Alkumet, *The identification of voiced versus voiceless consonants in English by advanced Iraqi learners of English*. 2013.

[41] S. Furui, "Cepstral analysis technique for automatic speaker verification". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 29 No. 2, 1981. pp. 254-272.

[42] J. Makhoul, "Spectral linear prediction: Properties and applications". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 23 No. 3, 1975. pp. 283-296.

[43] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models". *IEEE transactions on speech and audio processing*, Vol. 3 No. 1, 1995. pp. 72-83.

[44] M. A. Shahin, et al., "Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning", in *INTERSPEECH,* pp. 175-179.

[45] Z. Yunxue, et al., "Chinese accent detection research based on features structured". *International Journal of Hybrid Information Technology*, Vol. 8 No. 5, 2015. pp. 303-316.

[46] D. K. Phull and G. B. Kumar, "Consonant Analysis for Indian English". *International Journal of Control Theory and Applications*, Vol. 9 No. 51, 2016. pp. 241-248.

[47] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28 No. 4, 1980. pp. 357-366.

[48] S. Sinha, et al., *Speech processing for Hindi dialect recognition*, in *Advances in Signal Processing and Intelligent Recognition Systems*. 2014, Springer International Publishing. p. 161–169.

[49] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*. 2011, Columbia University.

[50] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress". *The Journal of the Acoustical society of America*, Vol. 100 No. 4, 1996. pp. 2471-2485.

[51] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception". *the Journal of the Acoustical Society of America*, Vol. 90 No. 5, 1991. pp. 2394-2410.

[52] W. V. Summers, et al., "Effects of noise on speech production: Acoustic and perceptual analyses". *The Journal of the Acoustical Society of America*, Vol. 84 No. 3, 1988. pp. 917-928.

[53] E. Jokinen, et al., "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech". *Computer Speech & Language*, Vol. 28 No. 2, 2014. pp. 619-628.

[54] *Similar Sound Letters*. 2014 25 May 2017]; Available from: http://www.quranreading.com/Tajweed-Quran.

[55] L. Ferrer, et al., "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems". *Speech Communication*, Vol. 69 2015. pp. 31-45.

[56] B. P and W. D. *Praat: Doing phonetics by computer*. 2014 5 May 2014]; Available from: http://www.praat.org.

[57] Y.-L. Shue, et al., "Pitch accent versus lexical stress: Quantifying acoustic measures related to the voice source", in *Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH 2007),* Belgium, 27-31 August 2007, pp. 2625-2628.

[58] Y.-L. Shue, et al., "Effects of intonational phrase boundaries on pitch-accented syllables in American English ", in *Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH 2008),* 2008, pp. -.

[59] H. Kawahara, et al., "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised Tempo in the Straight-Suite", in *Fifth International Conference on Spoken Language Processing (ICSLP 98),* Sydney, Australia, 30 Nov-4 Dec 1998, pp. -.

[60] H. Kawahara, et al., "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite", in *Fifth International Conference on Spoken Language Processing,*

[61] M. Iseli, et al., "Age, sex, and vowel dependencies of acoustic measures related to the voice source". *The Journal of the Acoustical Society of America*, Vol. 121 No. 4, 2007. pp. 2283-2295.

[62] R. Gajšek and F. Mihelič, "Comparison of speech parameterization techniques for Slovenian language", in *In 9th International PhD workshop on systems and control: young generation viewpoint,*

[63] R. K. Aggarwal and M. Dave, "Integration of multiple acoustic and language models for improved Hindi speech recognition system". *International Journal of Speech Technology*, Vol. 15 No. 2, 2012. pp. 165–180.

[64] R. K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system". *Telecommunication Systems*, Vol. 52 No. 3, 2013. pp. 1457-1466.

[65] K. S. Rao, "Role of neural network models for developing speech systems". *Sadhana - Academy Proceedings in Engineering Sciences*, Vol. 36 No. 5, 2011. pp. 783-836.

[66] L. Mary and B. Yegnanarayana, "Auto-associative neural network models for language identification", in *Proceedings: International Conference on Intelligent Sensing and Information Processing,* pp. 317–320.

[67] O. C. Ai, et al., "Classification of speech dysfluencies with MFCC and LPCC features". *Expert Systems with Applications*, Vol. 39 No. 2, 2012. pp. 2157-2165.

[68] S. Memon, et al., "Speaker verification based on different vector quantization techniques with gaussian mixture models", in *Third International Conference on Network and System Security, 2009,* pp. 403 – 408.

[69] H. S. Jayanna and S. R. M. Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions", in *TENCON 2008 - IEEE Region 10 Conference ,2008,* pp. 1 - 4.

[70] J. Chen, et al., "Robust MFCCs derived from differentiated power spectrum", in *Eurospeech 2001,* Scandinavia, 2001, pp. -.

[71] C. Wang, et al., "Differential mfcc and vector quantization used for real-time speaker recognition system", in *2008 Congress on Image and Signal Processing,* 5, pp. 319-323.

[72] S. Sakti, et al., *Incorporating knowledge sources into statistical speech recognition*. Vol. 42. Springer Science & Business Media, 2009.

[73] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech". *IEEE Transactions on speech and audio processing*, Vol. 4 No. 1, 1996. pp. 31.

[74] J. Navratil, "Spoken language recognition-a step toward multilinguality in speech processing". *IEEE Transactions on Speech and Audio Processing*, Vol. 9 No. 6, 2001. pp. 678-685.

[75] A. E. Thymé-Gobbel and S. E. Hutchins, "On using prosodic cues in automatic language identification", in *Fourth International Conference on Spoken Language Processing (ICSLP 96),* Philadelphia, PA, USA, 3-6 October 1996, pp. -.