Automatic Arabic Speech Segmentation System

Muhammad Jamil Anwar, M.M.Awais, Shahid Masud, and Shafay Shamail

Computer Science Department, Lahore University of
Management Sciences, Lahore, Pakistan

{jamilgoheer, awais, smasud, sshamail}@lums.edu.pk

Abstract

Emerging growth of information and communication technologies has influenced the research trends to focus on speech technologies. This research explains a basic speech segmentation application developed for Arabic language with the aim to further develop a language tutor. The focus is on Quranic Arabic as there are standards available which help in obtaining better accuracy. The research problem has been formulated in the form of a number of possible cues that could help identify the phoneme boundaries. These cues include zero crossing rate (ZCR), power spectral density (PSD), formant transitions, rhythm of consonants and vowels, intonation pattern also called fundamental frequency and vowel duration. A number of experiments were carried out involving different combinations of cues. The best results for phoneme level segmentation were observed through PSD and ZCR whereas pitch and intensity were helpful in determining pauses. Our system has demonstrated up to 89% accuracy on continuous speech files of eight different speakers which incorporates approximately 14300 phonemes.

Keywords: Speech Segmentation, Power Spectral Density, Zero Crossing Rate, Phoneme Classification.

I. Introduction

In this paper we have investigated the development of automatic Arabic speech segmentation system which would act as a test bed as well as foundation for several speech applications involving Arabic language. Keeping in view the emerging demands of speech solutions a prototype application to teach Quranic Arabic is aimed. The language tutor would help learners in their pronunciation and speech delivery, correcting them where they are wrong. This application includes the automated recognition system for which speech segmentation is an essential part. This speech segmentation system can also be used in speech recognition, speaker identification, summarization of spoken languages, speech documentation, indexing and speech translators.

Speech segmentation involves dividing speech utterances into different chunks which are recognizable and meaningful. This includes segmentation at language level, accent level, sentence level, word level, pause or silence level and phoneme level. Phoneme being the atomic unit in speech and distinguishing itself as a meaningful identity is quite difficult to segment with highly accurate boundaries [1, 4]. A speech signal is composed of several variables including the origin of the speaker, style of speaking and language dependent information which in itself contains variables

phonemic rules, phoneme inventory etc. Segmentation of speech signal at phoneme level has to take into account all of this information. In automated solutions this information is extracted from different properties of speech signal like formant trends, pitch, stress, vowel duration, power spectral density, zero crossing rate, and rhythm or intonation patterns. Using all these parameters and taking weighted information from them may result in some information loss which can be minimized through careful extraction and amalgamation. For different requirements, different combinations of these cues can help generate the desired results.

Section 2 gives a brief description of the segmentation related work found in previous literature. It covers different approaches used by different researchers and the applications where these are useful. Section 3 explains the concepts of phoneme segmentation and phoneme classification whereas Section 4 gives an outline of the cues used during different experiments and their usefulness in developing speech solutions. Section 5 outlines the methodology that we have adopted in phoneme segmentation. Section 6 explains the results and their analysis that are obtained and the way they are useful in analysis. Section 7 gives the conclusion and proposed future work followed by the references.

II. Literature Review

There are different ways of subdividing syllables to get the knowledge about the differences in phoneme inventories across languages and positioning of phoneme in the syllable. Most important is the variation of the syllable structure according to the morphological location of the syllable across languages. Automated language recognition involves the removal of noise and other non speech regions, phoneme separation and segmentation where phoneme boundaries are identified and phonemes are classified into different categories like vowels and consonants broad phoneme level classification where we assign them to a range of classes like affricates, fricatives, nasals, stops, which is further followed by phoneme identification.

Michael R Brent [2] in his paper has discussed in detail the phonology based segmentation and word discovery with the emphasis on English language and segmentation techniques inherently used by children. Shriberg and Stolcke [3] used a prosody based approach to determine segmented text from speech through Hidden Markov and decision tree modeling. Prosody is supposed to convey structural, semantic and functional information and can be extracted even in the absence of language dictionaries or language experts. Indicators for prosody include pausing, changes in pitch range and amplitude, global pitch declination, melody and boundary tone distribution, phone duration and speaking rate variation. Pfeiffer [4] mentioned technique of speech segmentation based on relative silence or pauses only. The only feature used was the perceptual loudness and areas determined as pauses are dependant on minimum duration and maximum interruption constraint.

Demuyunk and Laureys [5] used a hybrid approach where they have used speech signal along with its phonetic transcription. Segmentation results have been compared with manual segmentation in order to calculate the accuracy. Zechner [6] in his paper proposed the importance of speech summarization and discussed it in light of speech recognition and segmentation. Summarization of

spoken languages is a somewhat recent research area grown with the increasing amount of spoken audio databases. Summarization of spoken language may also aid the archiving, indexing, and retrieval of various records of oral communication, such as corporate meetings, sales interactions, customer support or tutoring environment. Wang [7] have used feature set including rate of speech, pause and prosody to do speech segmentation without word or speech recognition.

III. Phoneme Classification as a part of Phoneme Segmentation

Phoneme segmentation and phoneme classification are intertwined if we move by the language properties in the signal. When we talk about segmentation, we are talking about separating them as well as classifying them. Classification includes two main approaches. One of them is the vowel consonant classification in which we segment phonemes and classify them as vowel or consonant. Second is the broad phoneme classification where we classify the phonemes in detail as vowel and consonants and further classify the consonants as stops, fricatives, affricates, approximants and nasals. The second approach is expected to set up significant ground work for phoneme identification.

Broad level phoneme classification requires the knowledge about the properties of different phonemes. This knowledge is traced in the properties of signal to apply automatic classification algorithm. In the production of vowels the articulators do not come very close to each other and the passage of the air stream is relatively unobstructed where as consonants are pronounced when the flow of air or signal incorporates constriction in it which is of variable amount. Complete closure of articulators involved so that the air stream cannot escape through the mouth, specifies a stop. There are two possible types of stops, an oral stop and a nasal stop. If the closure is in the mouth and the nasal tract is blocked then air stream will be completely obstructed signifying an oral stop. On the contrary if the air is stopped in the oral cavity but it can go out from the nose, the sound produced is a nasal stop. Close approximation of two articulators so that the air stream is partially obstructed and turbulent airflow is produced, is called a fricative. An articulation in which one articulator is close to another but without the vocal tract being narrowed to such an extent that a turbulent air stream is produced specifies an approximant. A combination of a stop, immediately followed by a fricative is called an affricate.

An important aspect to be noted is that the line between vowels and consonants cannot be clearly mentioned, as a continuum exists between the two extremes. There are also intermediate instances, such as the semi-vowels and the (frictionless) aspirants. In this paper we have discussed the development of a basic level segmentation system classifying vowel, consonants and non speech regions including pause or noise in a speech signal based on Quranic Arabic.

IV. Features Representing Signal and Phonemic Properties

Properties related to phoneme are embedded in different signal properties which can act as our cues for the phoneme segmentation. There could be different combination of these cues that can generate different results with different accuracy levels. Research is still in progress to find out the best possible combinations suitable for a certain language. These features include the formant pattern, stress pattern, intonation pattern, rhythm, phoneme duration, rate of speech, zero crossing rate (ZCR), power spectral density (PSD) etc. Formant pattern show a typical shape in the start and the end which provides us with information about specific consonants. PSD and ZCR play a major role in segmenting phonemes [1]. Spectrogram and spectrograph are very important in identifying PSD and ZCR respectively. For example figure 1 shows the spectrogram with the energy bursts in vowel and consonant region which reveals that the bursts are very powerful in the vowel region as compared to the consonant region. The spectrograph on the other hand helps in determining the ZCR by evaluating the frequency at which the zero line is crossed by the signal under consideration. Dotted lines in the spectrogram represent the formant trends. They have a typical shape in the start and end which is helpful in broad segmentation of phonemes.

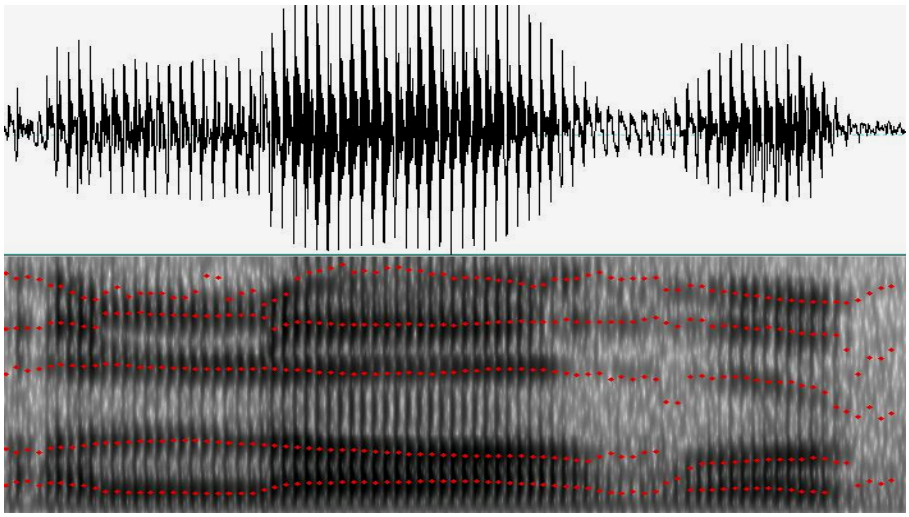

Fig. 1 Spectrogram (below) and spectrograph (above) illustrating different cues

V. Methodology

This section outlines in detail the settings, constraints, algorithm and calculation conducted with reference to the research presented in this paper.

Settings and Constraints

Speech samples are recorded in a constrained environment. Noise free recording room was used and the speakers were highly trained in Quranic recitation according to the “tajweed” rules. This was verified on speech inputs of 8 different speakers who recited “Quranic verses” of approximately 2 minutes each. Praat [12] was used as a speech processing tool. 35 different files were used which made it almost a 1 hour speech sample. Speech segmentation application was written in C++ and was applied on each of the files. Speech files were given as an input to Praat tool, which generated values of formants or zero crossing rate, or power spectral density in numerical format. This numerical representation was the input to the algorithm explained below for processing and generating the phoneme boundaries.

Algorithmic Implementation

Power spectral density, zero crossing rate and bandwidth are calculated from the signal. Through these we separate vowels from consonants using an application developed in C++. Figure 2 shows an abstract level diagram of the speech segmentation system. Non-speech regions are immediately detected and speech regions are further sent as an input to the phoneme segmentation module where basic level segmentation is performed.

We take a wave file for the processing. Wave file is stored in a numerical format in an array. Signal is sampled at 8000 Hz and the windows of 128 samples are taken for further processing. Power spectral density and zero crossing rate is calculated and stored in separate arrays. To proceed further we start with identifying the regions of the speech and ignore the non-speech areas. Non-speech area is determined by scanning the intensity of the speech signal.

Once these areas are determined they are stored in another array, which is further processed for the classification of vowels and consonants. Phonemes are extracted from the preprocessed array. Upon studying the trends of the PSD values, we devised an algorithm to separate vowels and consonants in a speech file. The results obtained, give us 89% accuracy. The results of this function are specified by the starting and ending window in which a particular vowel or consonant is contained. Now we have three cues ready for our further processing, that includes start of the phoneme, end of the phoneme and the classification of the phoneme showing whether it is a vowel or a consonant.

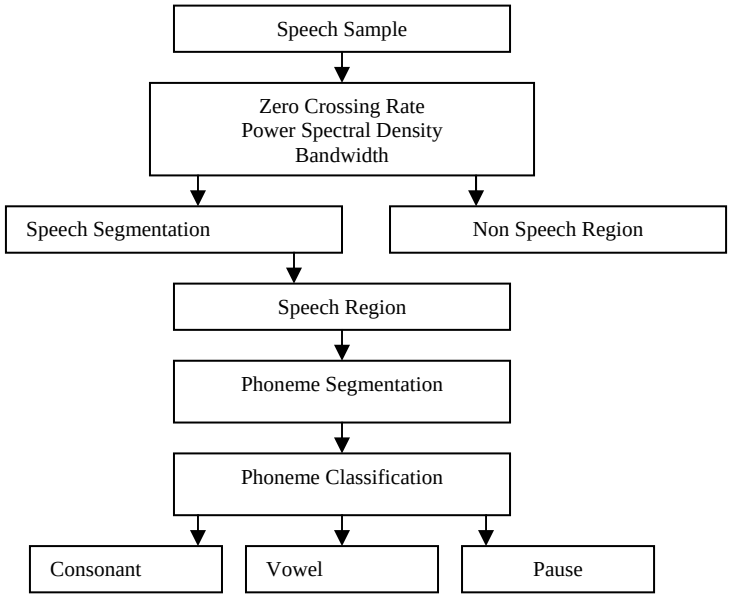

Fig. 2 Abstract level architecture of speech segmentation

Calculation of Cues

This section explains the calculation of the cues in detail. The speech signal was processed for PSD, ZCR and was further divided into speech and non-speech regions. The speech signal was filtered using a high pass filter of 60Hz. In order to calculate the values of PSD and ZCR, algorithms given in Table1 were used. Threshold values of PSD and ZCR were determined while manually scanning the signal for separating vowels and consonants.

Speech and non-speech region were segmented by using intensity of the signal. A typical threshold for PSD was determined to differentiate the speech and non speech regions. Value below the specified threshold was categorized as a non-speech region and above it specified a speech region.

Table 1 Algorithms to calculate PSD and ZCR

Algorithm 1 to calculate PSD	Algorithm 2 to calculate ZCR
1. Divide speech signal into windows. 2. Calculate the square of amplitude of different samples in window. 3. Add all those squared values of amplitude to find PSD of window 4. Calculate normalized PSD of window 5. Repeat until all the windows are finished	1. Divide the speech signal into windows 2. Compare successive samples in the window to find a transition from positive to negative 3. Mark a transition as zero crossing 4. Total number of zero crossings form a ZCR of the window 5. Calculate normalized ZCR 6. Repeat until all the windows are finished

VI. Result and Analysis

The following calculations were performed for the analysis of the results obtained through the application of aforementioned algorithm.

Speech signal was divided into different frames. For each frame the trends of the signal to find the number of consecutive frames specifying a certain class of phonemes i.e. vowel or consonant was checked. As a result each vowel or consonant detected, starting frame, ending frame, starting time, ending time and the proposed classification was obtained.

This methodology was repeated on 35 different files of Quranic Arabic obtained from trained speakers. The starting frame, ending frame, starting time, ending time and proposed classification for each file was calculated. As an example, summarized results generated algorithmically in 8 different files and 8 different speakers are shown in Table 2. The table shows total number of phonemes, total number of vowels and consonants identified through the proposed algorithm.

Table 2 Algorithmically generated results

Files	**Total # of Phonemes**	**Total Vowels**	**Total Consonants**
Speaker1	1967	1280	687
Speaker2	1788	1287	501
Speaker3	1863	1285	578
Speaker4	1831	1282	549
Speaker5	1723	1240	483
Speaker6	1338	924	414
Speaker7	1917	1304	613
Speaker8	1812	1269	543
Mean	**1779**	**1233.8**	**546**

In each file, there were approximately 1800 to 2000 phonemes detected. For each file, the results generated were manually checked by trained resource having in depth knowledge of phonetics, phonology and speech processing. For each file discrepancies were marked against the algorithmically classified vowel and consonants. The overall accuracy for correct segmentation of phonemes into vowels and consonants turned out to be 89%. Table 3 and 4 represent the recall and

precision in vowel and consonants separately. Table 3 represents the total number of vowels identified from algorithm (VAlgo), vowels identified through manual testing (VManual), actual vowels which were termed as consonants by the algorithm (Vas C), and actual consonants which were termed as vowels by the algorithm (Cas V). Formulas for calculating recall and precision values are as follows: -

$$Vowel\ \mathrm{Re}call = \frac{VowelsIdentifiedCorrectly}{VowelsIdentifiedCorrectly + V_{asC}}$$

$$Vowel\ \mathrm{Pr}ecision = \frac{VowelsIdentifiedCorrectly}{VowelsIdentifiedCorrectly + C_{asV}}$$

Precision defines the proportion of the classified phonemes which are actually correct whereas recall depicts the sensitivity, or the proportion of the correct results obtained.

Table 3 Vowel Recall & Precision

Speakers	V_{algo}	V_{manual}	$V_{as\ C}$	$C_{as\ V}$	V_{Recall}	$V_{Precision}$
1	1280	1297	109	92	91	92
2	1287	1309	103	81	92	93
3	1285	1277	89	97	93	92
4	1282	1319	116	79	91	93
5	1240	1245	91	86	92	93
6	924	928	61	57	93	93
7	1304	1320	112	93	91	92
8	1269	1305	121	85	90	93

Table 4 Consonant Recall & Precision

Speakers	C_{Algo}	C_{manual}	$C_{as\ V}$	$V_{as\ C}$	C_{Recall}	$C_{Precision}$
1	687	670	92	109	86	84
2	501	479	81	103	83	79
3	578	586	97	89	83	84
4	549	512	79	116	84	78
5	483	478	86	91	82	81
6	414	410	57	61	86	85
7	613	594	93	112	84	81
8	543	507	85	121	83	77

In Table 4 comparative results for consonants found in manual testing and algorithmic results are shown. In results generated from algorithm, there are possibilities that the actual consonants are wrongly identified as vowels (C as V) or some vowels might be incorrectly marked as consonants (Vas C). Column 4 and 5 show these discrepancies. Consonant recall and precision cater these discrepancies and their calculations are done as follows: -

$$Consonant\ \mathrm{Re}call = \frac{ConsonantsIdentifiedCorrectly}{ConsonantsIdentifiedCorrectly + C_{asV}}$$

$$Consonant\ \mathrm{Pr}ecision = \frac{ConsonantsIdentifiedCorrectly}{ConsonantsIdentifiedCorrectly + V_{asC}}$$

Precision and recall values are used to incorporate error types classified in columns 4 and 5 of Table 3 and 4 respectively. Trends show that mostly the consonants were incorrectly identified. Their precision values are relatively lower than those of vowels. This is due to the overlapping boundaries of vowels and consonants. The selected cues were unable to correctly mark the phoneme boundaries in some regions. Consonants being of very small duration were sometimes missed out. The results reveal that the stops are the main victims because they are formed with the closure of lips and then energy burst showing up the vowel which follows. This might be the reason the consonants are overlooked as they occur for smallest period of time.

VII. Conclusion and Future Work

Power spectral density, zero crossing rate and bandwidth are the cues which play a major role in phoneme level segmentation. Automatic phoneme level segmentation would add significant value to most of the speech solutions. So far most of the research work is done in sentence level or pause level segmentation without aiming towards any specific speech solution. Our system has proved significant level accuracy at phoneme level for Arabic language.

In future besides improving this system for Quranic Arabic we intend to check this for other languages. For Quranic Arabic, we intend to update it for adverse speech conditions where there is noise in the signal. Filters need to be developed for preprocessing of the signal. Moreover, we need to incorporate Quranic speakers from other regions and language backgrounds. A thorough testing is required with the variegated data in order to increase the accuracy. In addition, there is a need to find out the trends in consonants being missed and vowels being incorrectly identified. This would help in adding value to the system to achieve better accuracy.
Further more vowel and consonant duration along with the formant trends may be included to further classify the consonants. Broad level classification needs to be done to get stops, fricatives, affricates, and nasals. Vowels also need to be classified into simple and nasalized vowels.

References

[1] L. Rabiner and B.-H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[2] Michael R. Brent, “Speech segmentation and word discovery: A computational perspective” 1999.

[3] Elizabeth Shriberg, Andreas Stolcke, “Prosody-Based automatic segmentation of speech into sentences and topics ” Speech Communication September 2000

[4] Silvia Pfeiffer, “Pause concepts for audio segmentation at different semantic level” Proceedings of 9th ACM international conference on multimedia 2001

[5] Kris Demuynck, Tom Laureys, “A Comparison of Different Approaches to Automatic Speech Segmentation” 2002

[6] Klaus Zechner, “Summarization of Spoken Language – Challenges Methods and Prospects” January 2002

[7] Dong Wang, Lie Lu, Hong-Jiang Zhong, “Speech segmentation without speech recognition” 2003

[8] Hema A. Murthy, T.Nagarajan, N. Hemalatha, “Automatic segmentation and labeling of continuous speech without bootstrapping” 2004

[9] Steven Bird, Patrick Blackburn, “A Logical Approach to Arabic Phonology” Association of Computational linguistics 1991

[10] Andrea Weber, “A role of phonotactics in the segmentation of native and non native continuous speech” 2000

[11] Mnish Sharma, Richard Mammone, “Blind: Speech Segmentation: Automatic Segmentation of speech without linguistic knowledge” 1996

[12] http://www.fon.hum.uva.nl/praat

[13] Antonio Bonafonte, Albino Nogueiras, Antonio Rodriguez-Garrido, “Explicit Segmentation of Speech using Gaussian Models”

[14] Julie Carson-Berndsen, Michael Walsh, “Generic Techniques for Multilingual Speech Technology Applications”

[15] S. E. Tranter, K. Yu, G. Evermann, P. C. Woodland, “Generating and Evaluating Segmentations for Automatic Speech Recognition of Conversational Telephone Speech” 2004.

Muhammad Jamil Anwar
MS Student
Jamil has an undergraduate and graduate degree in computer science with extensive research in speech and language processing. He is currently involved with a start up company working in speech processing , especially speech searching.

Dr. Shafay Shamail
HOD & Associate Professor
Before joining LUMS Dr Shamail was at SoftNet Systems, where he was responsible for e-commerce technologies, especially those from Microsoft. He was also with Pak-AIMS, involved with various responsibilities concerning the BCS, MIS, and ALCoS programmes including course design and implementation. Dr Shamail was convenor of different committees and patron Pak-AIMS Computer Society.

Dr. MM Awais
Associate Professor
Prior to joining LUMS, Dr Awais conducted European Union research and development projects for a UK based SME. His PhD work related to the development of on-line models for parametric estimation of solid fuel-fired industrial boilers. This involved time series analysis of the systems and realisation of virtual sensors based on neural network- related internal model control schemes. Dr Awais has also conducted research work on a class of iterative methods pertinent to Krylov subspaces for optimisation, such as the oblique projection and implicitly restarted model reduction methodologies. His current research interests include development of application of artificial intelligence techniques to model and control intricate phenomenological processes, image processing, inverse modelling, and data management.

Dr. Shahid Masud
Associate Professor
Dr Masud was a Senior Design Engineer at Amphion Semiconductor Ltd., UK, before joining LUMS in October 2002. He has several years of research and development experience in the design of VLSI systems for image and video coding and computer interfacing. He has published over thirty refereed papers internationally and holds three patents in VLSI design. He is a member of IEEE, IEE and a Chartered Engineer.