FORMANT ANALYSIS FOR KISWAHILI VOWELS

[1]YY Sungita, and [2]EE Mhamilawa
[1]Tanzania Atomic Energy Commission, P. O. Box 743, Arusha
[2]Department of Physics, University of Dar-es- Salaam, P. O. Box 35063, Dar es Salaam

ABSTRACT

Vowels' spectral characteristics in a language have been studied for suitability in speech recognition by using formants analysis technique. Other techniques do mostly require large computer memories for speech processing and analysis. In this paper, the formant analysis for Kiswahili vowels has been presented. The spectrographs for each vowel and their respective average formant frequencies are tabled. The distribution of formants for the vowels modelled in the form of an articulatory model is shown. The results show that there is a big separation of formant frequencies among the Kiswahili vowels that signify the suitability for automatic speech recognition.

INTRODUCTION

The automatic speech recognition and speech synthesis is one of the most recent technologies with a growing market demand as people are becoming comfortable with hi-tech equipment. It can be argued that speech being the natural mode of communication between humans should also be used in man-machine communication. There are already some voice recognition products in the market for various international languages like English, French, Italian, Spanish, German and Arabic (Davis et al. 1952, Rebecca et al 1998). None has been done to utilize Kiswahili language in automatic speech recognition technology. Therefore, the study of speech signals for Kiswahili vowels is vital for the exploration of their characteristics and utilization in speech synthesis and recognition.

The formants are the natural frequencies or resonance of the vocal tract when the human is uttering. Acoustic energy transfers from the excitation source to the output of the sound production system results into generation of formants. The human voice has formant regions determined by the size and shape of the nasal, oral and pharyngeal cavities (vocal tract) (Fig. 1), which permit the production of different vowels and voiced consonants (Parsons 1987, Shuzo 1992, Rabiner and Juang 1993).

Therefore the formants are the most immediate source of the articulation information because the vowels have well defined spectral representation that lead to best recognition rate. Formants have long been regarded as one of the most compact and descriptively powerful parameter-sets for voiced speech sounds, with important correlates in both the auditory- perceptual and articulatory domains (Akira et al. 1973, Keller 1995, Zolfaghari 1996). Formant based representation is found to be appropriate for study of static vowels or synthetic speech due to difficulty in accurate and reliable estimation of formants information on continuous speech.

The discrete Fourier transform (DFT) serves as a basis for the formant analysis of speech, since it directly contains the formant information in its magnitude spectrum (Mills 1996, Zolfaghari 1996, Mokhlari 2000, Milan 2001). There are several techniques that can be used to identify the formant frequencies from the speech uttered. In this paper the estimation of formants for Kiswahili vowels were made using formant based speech analysis employing short time Fourier transform analysis (stft). In this technique, the spectrographs for Kiswahili

digits were obtained and those regions representing the vowels identified. The darkest bands in the spectrographs indicated the location of formants. A primary motivation of spectrograph representation is to discover how the power spectrum of a signal changes over time. The spectrographs are plotted with their frequency in linear scale as this makes the formants clearly identifiable.

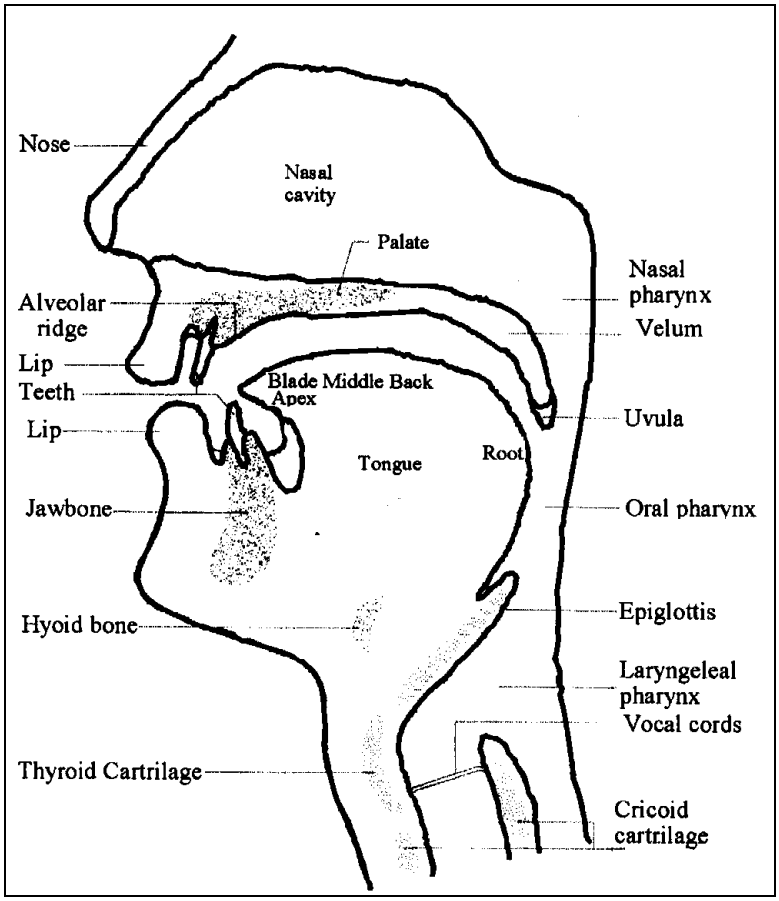

Figure 1: The principal organs of speech production (articulatory model) (Parsons 1987)

METHODS

The utterances from the speaker for ten isolated Kiswahili digits were recorded using an omni-direction (hypercadiot model) microphone. Ten samples of speech sounds for each ten Kiswahili digits were captured. The processing and editing of these sound samples were done using PC Dell computer, Pentium II, 64.0MB processor. The editing procedure was done to mark the beginning and end of the signal under processing.

There were some steps taken during sound recording to reduce the effects of acoustic variability of speech signals. First, the recording was done in acoustically conditioned audio recording room of a radio studio. Second, the same microphone was used to capture speech signals during

recording for all samples. Third, the same male speaker uttered the predetermined words and did it at the same sitting. The changes in the recording of environment, position and characteristics of the transducers (microphones) and the speaker's physical and emotional state, speaking rate or voice quality cause the acoustic variability of speech signal.

Endpoint Detection

The correct location of the beginning and end of an utterance minimizes the amount of subsequent processing and has been found important in improving the accuracies of representation of isolated words. To detect the start and end points of a word the power of the incoming signal was constantly monitored. Once the signal goes above the threshold, the wave was recorded until after the signal goes below the end threshold. The silences before the beginning and after the end of the signal were chopped off respectively. This procedure reduced the errors due to the incorrect locations of the beginning and end of the speech signal.

The edited speech sounds were stored in the computer as raw data, in the WAVE format using pulse-coded modulation (PCM). The analogue sound signals were digitised by analogue to digital converter (ADC) at the interface sound card. The speech signals were band-limited to 200 - 4000Hz. The sampling rate of 8 KHz and a 16 bits resolution were applied.

Determination of Formants by Spectrograph

The LabView software with joint time-frequency analysis (JTFA) add-on software is the graphical design software that makes use of virtual instruments programming for designing and performing some functions. The system implemented to determine the formant frequencies comprised of three main parts namely; data acquisition, windowing, signal analysis and display of data (Fig. 2).

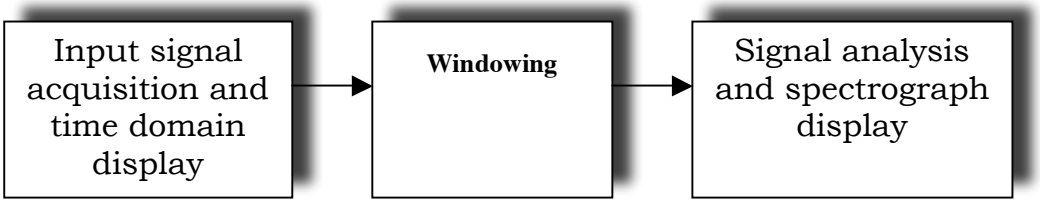

Figure 2: Block diagram for short time Fourier transforms (STFT) spectrograph analysis

In data acquisition, the 'analogue input (AI) acquire waveform' VI (Fig. 3) was used to acquire data (input signal) via sound card VI. The VI acquires the specified number of samples at a specific scan rate and returns all the data acquired in units of volts. This VI calls the 'AI CONFIG' VI and 'AI SCAN' VI from the analogue input palette, with the specified parameters such as device number, number of samples, sample rates and channel number. Device specification identifies the number of the plug-in data acquisition board. In this paper the device (1) corresponds to the National Instruments Data Acquisition, NI-DAQ (AT-MIO-16E-2) board. The number of samples and sample rates were identified because they specify the number of samples VI acquires before the acquisition is complete and the number of samples per second to acquire respectively. The channel number specifies the analogue input channel to acquire the data from. According to the configuration of the adapter sound card used the 'channel (0)' was set. The captured speech signal is fed to the input of 'windowing and stft spectrograph analysis' VI (Fig. 4).

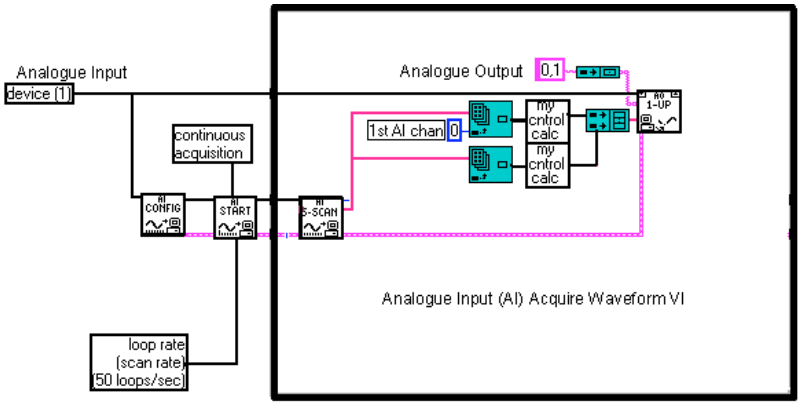

Figure 3: The 'analogue input acquire waveform VI' that reads and charts the input speech.

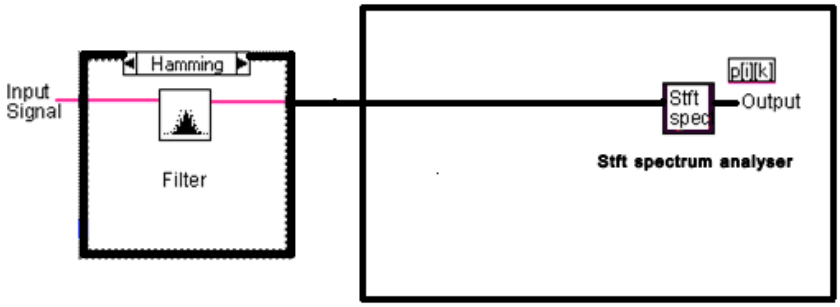

Figure 4: The 'windowing and stft spectrograph analysis VI'.

The signal is windowed by the 'Hamming windowing VI' (cosine window) that attenuates the signal towards the edges to minimize the signal discontinuities that might arise at the beginning and the end of each frame. The main concepts were to minimize the spectral distortion by using the window to soften the edges of the signal by tapering the signal to zero at the ends of the signal. The duration of the analysis window was 32 msec (N = 256 samples) that was proposed to give good frequency resolution (Rabiner and Juang 1993). Note that the multiplication of the signal by a window function in the time domain is the same as convolving the signal in the frequency domain. Thereafter, it is fed to input terminal r(i) at the 'stft spect. analysis VI' to be analysed. The output, p(i)(k) displays the spectrographs in which the formants were estimated.

RESULTS AND DISCUSSION

The spectrographs and the time domain representation of Kiswahili digits extracted from their respective speech signals are shown in figures 5 - 13. The presence of vowels is characterised by the evenly spaced harmonics of a periodic voicing as well as their downward diagonal movements as the pitch falls. These harmonics are darker when they are in frequency region of a formant peak, since they have high dB level. Thus, the dark bands in the spectrographs show the

location of the formant frequencies for the vowels in the digits. The consonants have aperiodic sounds that do not have discrete harmonics. Nevertheless, the vertical position has haphazard fluctuations in amplitude, indicating that the sound is voiceless frication and the source should be a noise.

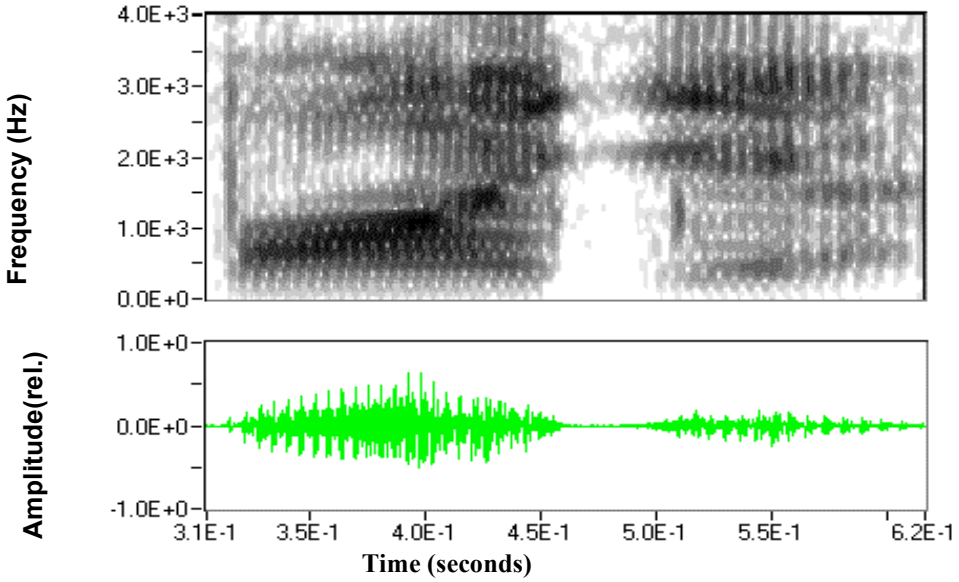

Figure 5: The spectrograph for digit 'moja'.

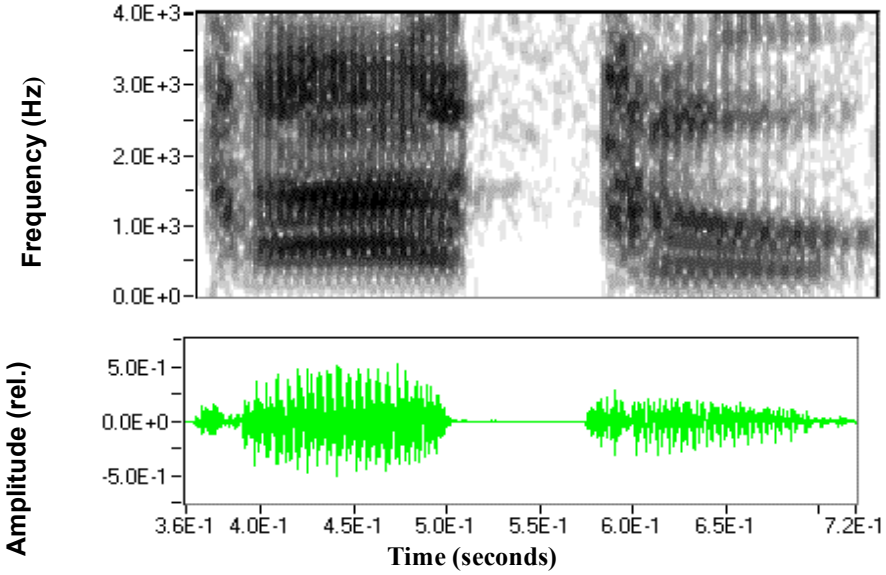

Figure 6: The spectrograph for digit 'tatu".

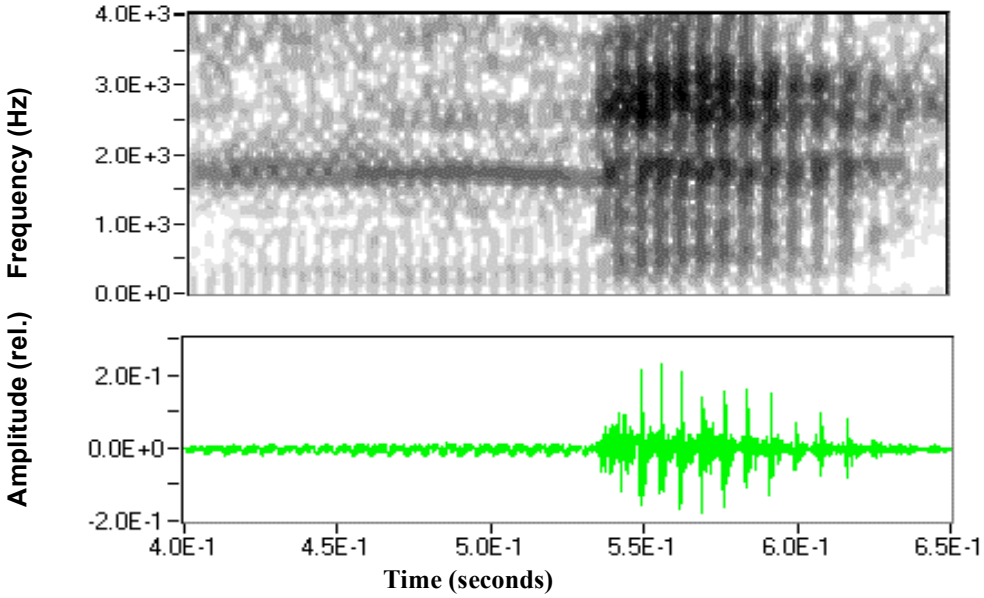

Figure 7: The spectrograph for digit 'nne'.

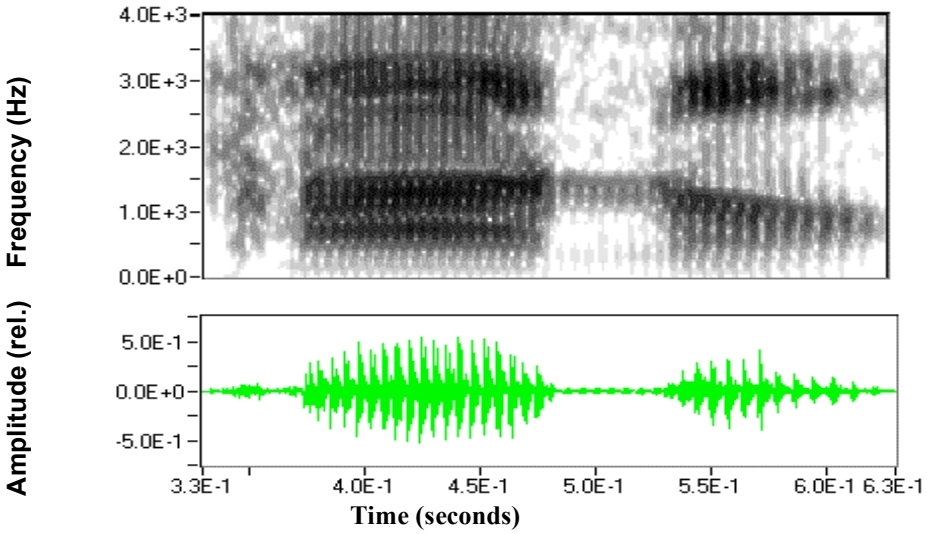

Figure 8: The spectrograph for digit 'tano'.

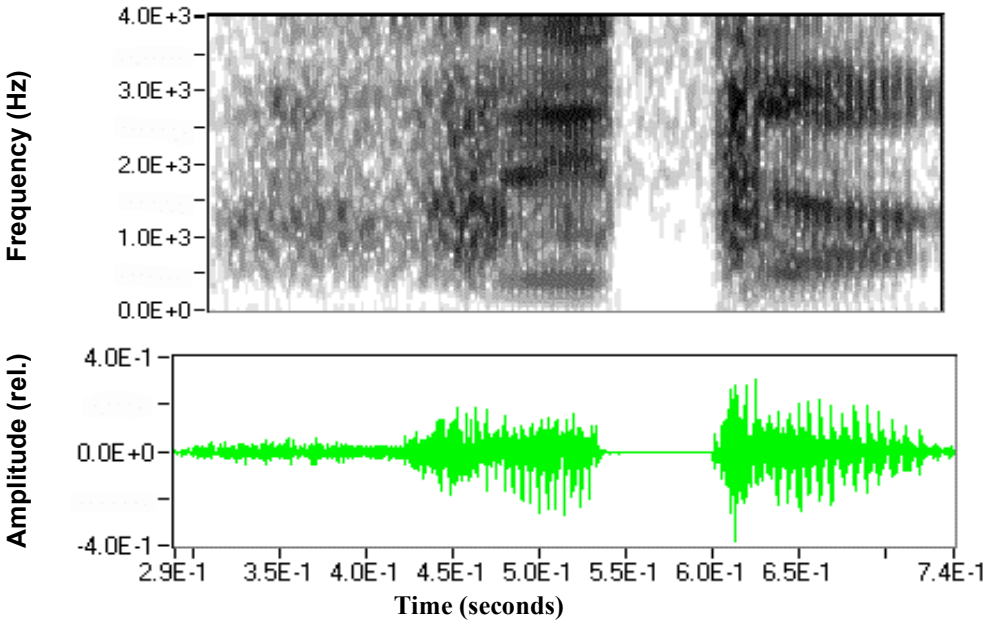

Figure 9: The spectrograph for digit 'sita'.

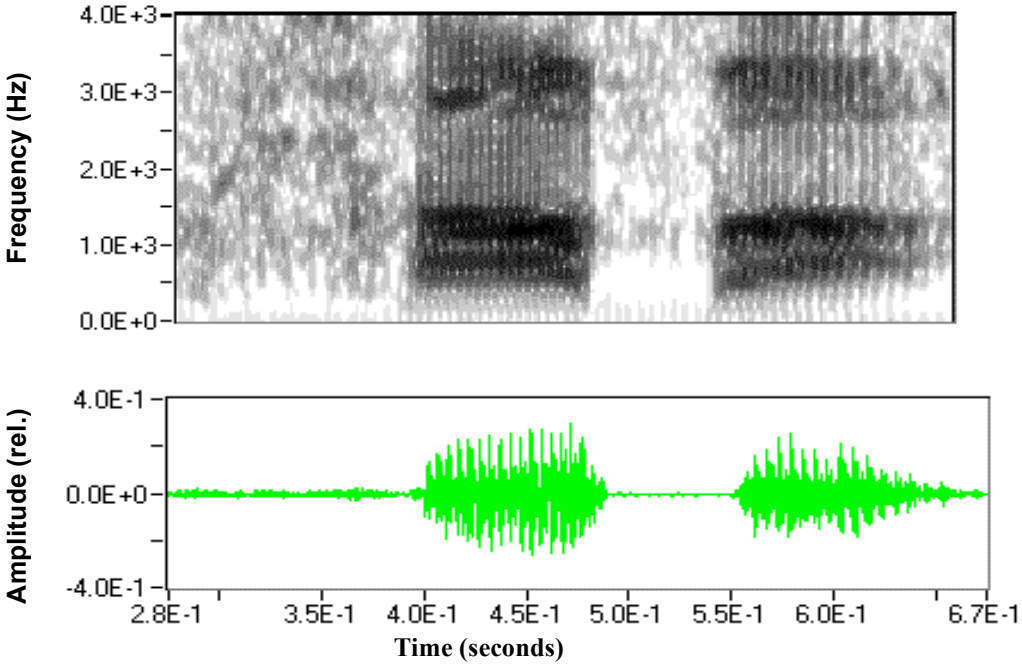

Figure 10: The spectrograph for digit 'saba'.

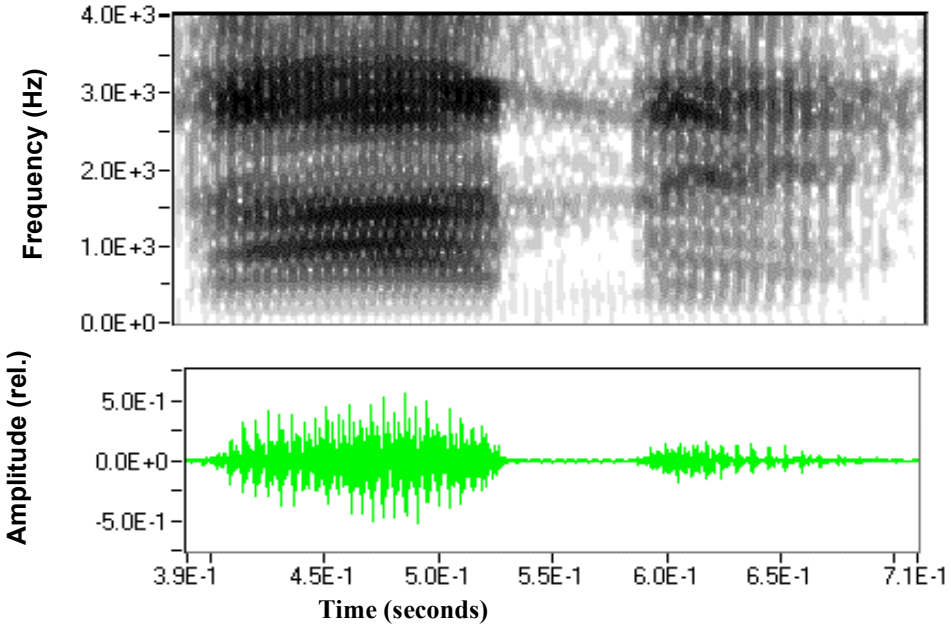

Figure 11: The spectrograph for digit 'nane'.

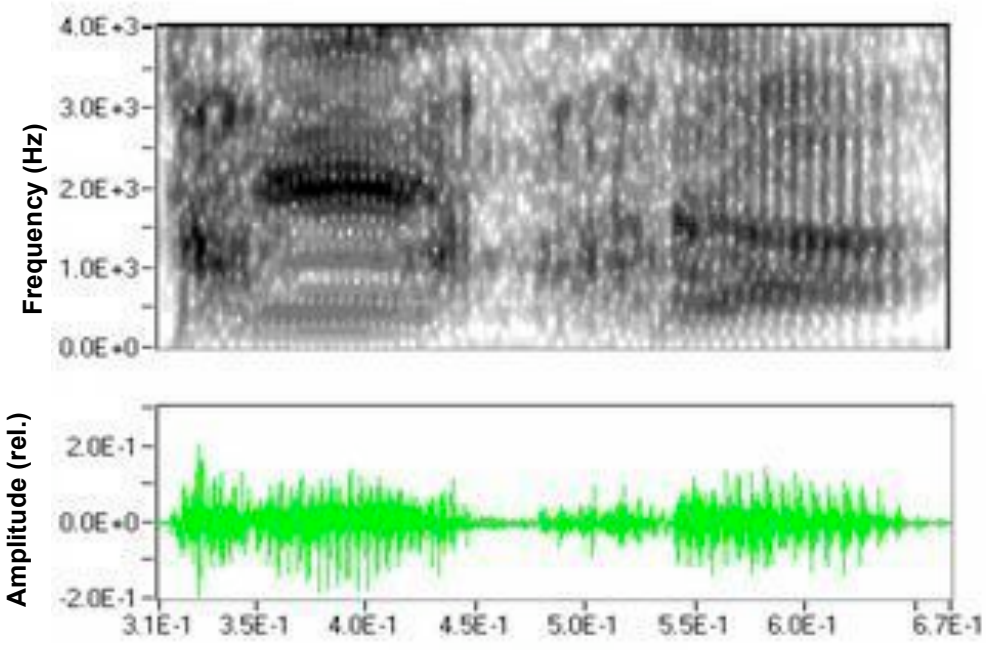

Figure 12: The spectrograph for digit 'tisa'.

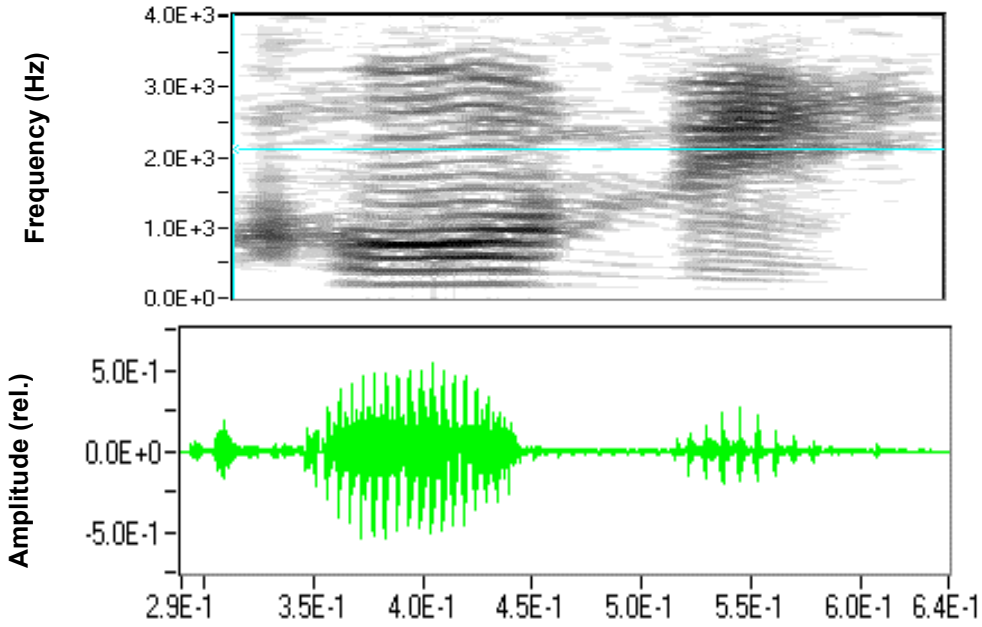

Figure 13: The spectrograph for digit 'kumi'.

Table 1: Mean formant frequencies for the Kiswahili vowels.

DIGIT VOWEL		FORMANTS FREQUENCIES (Hertz)			DURATION seconds
		F1	F2	F3	
moja	o	625	1000	2800	0.31
	a	812	1480	3000	
tatu	a	812	1440	2800	0.36
	u	500	1000	2310	
nne	e	500	1720	2630	0.25
tano	a	840	1380	2800	0.30
	o	500	900	2910	
sita	i	400	1880	2470	0.45
	a	725	1280	2800	
saba	a	719	1250	2750	0.39
	a	719	1250	2750	
nane	a	844	1380	2880	0.32
	e	562	1750	2780	
tisa	i	400	2000	2870	0.36
	a	740	1380	2910	
kumi	u	562	844	2440	0.35
	i	320	2290	2840	

Table 1 indicates an estimation of formant frequencies for vowels from male speaker utterances as seen on the spectrographs.

If we set up a coordinate system using the first formant frequency, F1 and the second formant frequency, F2 as a basis, vowels lie at specific regions. Fig. 14 shows the distribution of formant frequencies of the vowels extracted from Kiswahili digits.

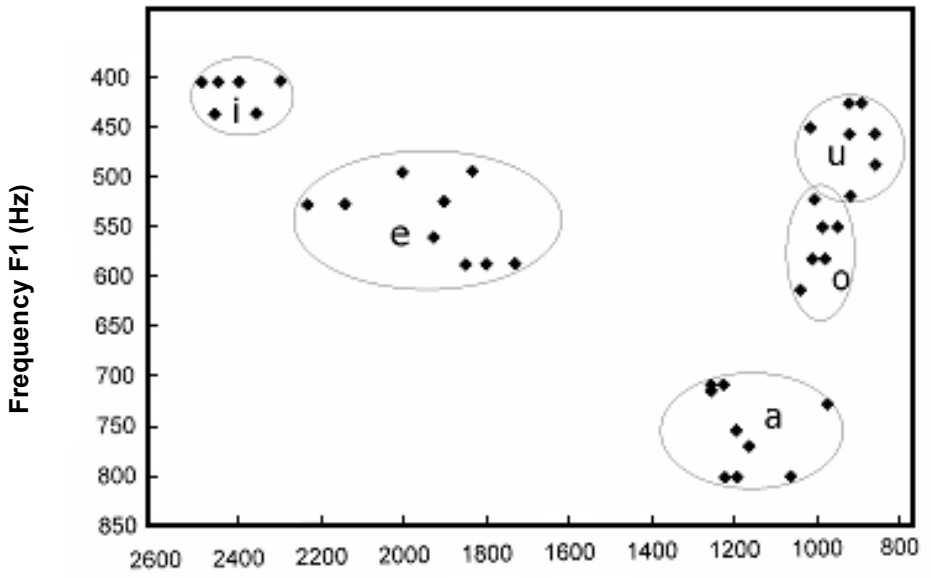

Figure 14: Measured frequencies of first and second formants for Kiswahili vowels.

Setting F2 at x-axis and F1 at y-axis and reversing direction (Fig. 14), the vowel loci correspond roughly with the position assigned to these vowels in the articulatory vowel location (Fig 1). In articulatory model, the vowels /i/ is classified as high-front, /e/ as middle and /a, o, u/ as back vowels. Further classification shows that vowel /a/ is low-back, /o/ is medium-back and /u/ is high-back vowel. The exact partitioning of the F1-F2 space varies with the age, sex and language also from one talker to another, but the overall pattern does not vary. This correspondence between the vowel sounds and the formant frequencies is expected because changing the shape of vocal tract produces different vowel sounds.

The F1-F2 plots of the formant frequencies extracted from the Kiswahili vowels indicated that there is big separation among the vowels. Since the formants for the vowels are nicely separated, then, recognition of Kiswahili digits by using these parameters is expected to be high. The vowels have well defined spectral waveform such that they influence the recognition rate of the speech in which they occurred contrary to the consonants.

Examining the formants location from spectrograph of each digit shown in figures 5 - 13 leads to possible citing of some problems that are expected to bring about some confusion in recognition. The influence of vowels in determining the

spectral waveforms of speech and hence the speech recognition rate can be explained by some examples below. The digits 'tatu' and 'tano' show similar spectrographs. This may be due to the fact that both words start with the same 'click' voiced phoneme /t/ followed by the voiced phoneme /a/. It is observed that the formants, F1 and F2, for both phonemes /o/ and /u/ at the syllables /no/ and /tu/ in their respective word occupy very close frequency bands. Therefore, most likely it can cause the confusion in recognition.

We can deduce from the spectrographs for digits 'sita' and 'saba' that there are some similarities that are expected to cause some confusion. Both spectrographs show some long haphazard fluctuation noises at the beginning of the signals. This is because these words start with unvoiced phoneme, /s/ as seen from syllables /si/ at digit 'sita' and /sa/ at digit 'saba'. After the silences we can see some dark bands indicating the location of formants for phoneme /i/ and /a/ being located at different frequencies. Second part of these digits consists of syllables /ta/ and /ba/ respectively. Similar patterns are seen on their spectrographs because both words end with the same voiced phoneme /a/. Therefore the different locations of the formant for vowels lead to dissimilarities between these words.

The digits 'sita' and 'tisa' have similar distribution of vowels. Since vowels have a big influence in spectral representation of speech signals, it is convincing that some confusion might arise to recognise those digits. However, from the spectrographs there are large differences, particularly at the beginning of each word. The digit 'sita' starts with a long silence (unvoiced sound), /s/, while the digit 'tisa' starts with click voiced phoneme. Also the duration for these two words is far different.

The digits 'nne' and 'nane' were also among the combinations that were expected to have some recognition problems. They have the same beginnings and similar endings. The duration of digit 'nne' is very short relative to other digits, such that it may lead to some recognition problems. But, according to spectrograph presentation, the presence and influence of two vowels in digit 'nane' caused dissimilarity.

CONCLUSION

The formant analysis of Kiswahili vowels has been performed. The use of spectrographic representation of speech enabled the visual inspection of the energy distribution in a spectrum that led to the location of the formants for vowels. The use of formants to predict the articulatory vowels information for uttered digits has been justified. There is clear separation of formants distribution among the Kiswahili vowels that influenced the speech recognition rate. Also some possible confusion as consequences of close occurrence of formants for some vowels that may arise in automatic speech recognition is explained. Conclusively, formants being one of speech parameters indicated that Kiswahili words could be nicely recognized by automatic speech recognition device.

REFERENCES

Itchikawa A, Nakano Y and Nakata 1973 "Evaluation of Various Parameter sets in Spoken Digits Recognition" *IEEE Trans.on Audio and Electro-acoustic,* AU-**21(3)**.

Davis KH, Biddulph R and Balashek S 1952 "Automatic Recognition of Spoken Digits", *J. Acoust. Soc. AM,* **24(6)**, 647-642.

Keller E 1995 "Fundamentals of Speech Synthesis and Speech Recognition" Basics Concepts State-of-Art of Future Challenges; by John Wiley & Sons ltd.

Sigmund S 2001 "Estimation of speaker Characteristics by Average Long-time Spectrum". Brno Univ. of Techn. Czeck Republic.

Patrick MM 1996 *"Fuzzy Speech* Recognition*",* MSc (thesis), Univ. of South Carolina.

Mokhlari, P. and Tanaka, K, "A 2000 Corpus Of Japanese Vowel Format Patterns".

Parsons, T. W., 1987 "Voice and Speech Processing", McGraw-Hill Series in Electrical Engineering.

Rabiner L and Juang, B 1993 "Fundamentals of Speech Recognition", PTR Prentice-Hall, Inc.

Rebecca BB and Paul KS 1998 "Voice Recognition for Embedded Systems", *Proc. ICDCSP,* UK.

Shuzo S, 1992 "Speech Science and Technology", 3-1 Kanda Nishiki-cho.

Zolfaghari P and Robinson T 1996."Formant Analysi Using Mixtures of Gaussian", Cambridge Univ. UK,