Vowel Recognition Based on Frequency Ranges Determined by Bandwidth Approach

M.P. Paulraj[1], S. Yaacob[1], S. A. Mohd Yusof[2]
[1] *Universiti Malaysia Perlis, Malaysia*
[2] *Universiti Utara Malaysia, Malaysia*
paul@unimap.edu.my， *s.yaacob@unimap.edu.my shahrulazmi@uum.edu.my*

Abstract

Automatic speech recognition (ASR) has made great strides with the development of digital signal processing hardware and software especially using English as the language of choice. In this paper, a new feature extraction method is presented to identify vowels recorded from 80 Malaysian speakers. The features were obtained from Vocal Tract Model based on Bandwidth (BW) approach. Bandwidth approach identifies frequency bands based on the first peak of vowel frequency responses. Mean and maximum energies were calculated from these Bandwidth frequency bands. Classification results from Bandwidth Approach were compared with the first 3-formant features using Linear Predictive method. A Multi-Layer Perceptron (MLP) and Multinomial Logistic Regression (MLR) were used to classify the vowels. MLR and MLP shows comparable classification results for BW approach of 96.40% and 96.59% respectively. Bandwidth approach obtained 5.49% higher classification rate than 3-formant features using MLP.

1. Introduction

Automatic Speech recognition (ASR) is the process in which an acoustic signal, captured by a microphone, is converted to a set of words by means of a computer program. ASR belongs to the class of digital speech processing technologies that also includes speech synthesis and voice biometrics. In general, their aim is to allow a machine to replicate the ability of a human to hear, identify, and utter natural human spoken language.

The goal of an Automatic Speech Recognition (ASR) system is to transcribe speech to text. As illustrated in Figure 1, the speaker's mind decides the source word sequence W that is delivered through his/her text generator. The source is passed through a noisy communication channel that consists of the speaker's vocal apparatus to produce the speech waveform and the speech signal processing component of the speech recognizer. Finally, the speech decoder aims to decode the acoustic signal X into a word sequence W*, which is hopefully close to the original word sequence W.

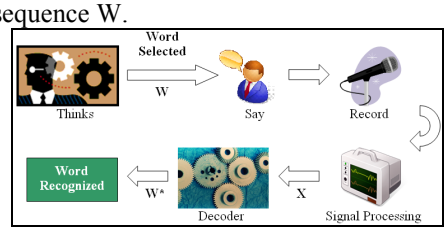

Figure 1: ASR System.

1.1. Linear prediction coding (LPC) model

Linear Prediction is a method which determines the coefficients of a pth-order linear predictor or a finite impulse response filter that predicts the current value of the real-valued time series x based on past samples by minimizing the prediction error in the least squares sense [1].

In linear prediction (LP) analysis, an all-pole filter with transfer function in equation 1 models the vocal tract transfer function.

$$H(z) = \frac{s(z)}{U(z)} = G(1 - \sum_{i=1}^{p} a_i z^{-1}) \quad (1)$$

where ***G*** *is* a constant and ***p*** is the number of poles. ***S(z)*** and ***U(z)*** are obtained by *Z*-transform from output signal ***s(n)*** and input signal ***u(n).*** The linear prediction coefficients a_i are chosen to minimize the mean square prediction error as shown in equation 2:

$$err(n) = x(n) - \hat{x}(n) \quad (2)$$

The filter coefficients' are obtained by using the autocorrelation method of autoregressive (AR) modeling.

1.2. Bandwidth theory

Bandwidth is the difference between the upper and lower cutoff frequencies of a signal spectrum and measured in hertz. In signal processing, the bandwidth is the frequency at which the closed-loop system gain drops 3 dB below peak [2] shown by E_{BW} in equation 3. Refer to Figure 2 for a simple illustration of the theory.

$$E_{BW} = \frac{1}{\sqrt{2}} E_{peak} \quad (3)$$

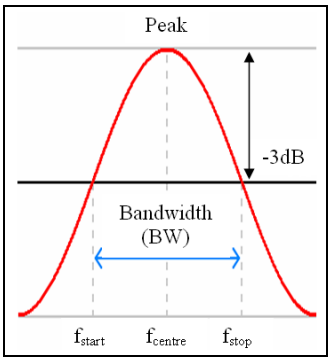

Figure 2: Bandwidth example

2. Speech recognition using vowels

Human speech has strict hierarchical structure. It consists of sentences, which can be divided into words, and they are built by phonemes that are the basic voice construction elements. Vowels could be defined as phonemes with persistent frequency characteristics most expressed. These frequency characteristics represent stable basis for construction of efficient vowel recognizer. It is known from literature [3], [4], and [5] that the spectral properties of male, female and child speech differ in a number of ways especially in terms of average vocal tract lengths (VTL). The VTL of female is about 10% shorter compared to the VTL of male. The VTL of children is even shorter (up to 10%) than that of females.

2.1. Data collection and data preparation

Data collection was done twice taken from 80 individuals consisting students and staffs from Universiti Malaysia Perlis (UniMAP) and Universiti Utara Malaysia (UUM). The recordings were done using a microphone and a laptop computer with a sampling frequency of 8000Hz. The words "KA, KE, KI, KO, KU" were used to represent the five vowels of /a/, /e/, /i/, /o/ and /u/ because vowels have significantly more energy than consonants. Based on [1], [6], [7] and [8], the first three formants for vowels are situated within 4 kHz and so are vowel's main characteristics. For this study, a sampling frequency of 8 kHz was used to sample the vowels. The recordings were done 3 to 4 times per speaker. The summary of the entire Vowel recognition Process is shown in Figure 3 below.

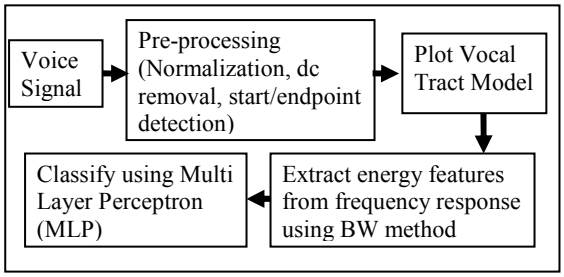

Figure 3: Vowel Recognition Process

A simple program was designed to record utterances samples from the speakers. The details of the data collection are listed in the table 1 below.

Table 1. Data Collection details

Information	1st Data Collection	2nd Data Collection
Sources	40 UniMAP students	40 UUM staffs and students
Recorded utterances	640	445
Sampling Frequency	8000 Hz	8000 Hz
Words uttered	/ka/, /ke/, /ki/, /ko/, /ku/	/ka/, /ke/, /ki/, /ko/, /ku/

2.2. Extracting the vowel from recordings

A simple program was developed to extract the vowel portions of the signal based on energy computation. First, the voice signal was recorded and then normalized. DC components are removed from the signal. The start and endpoint locations were determined using zero crossing and energy method.

2.3. Analyzing the vocal tract model

Formants analysis was used to classify the extracted simple phoneme signal. Formants are essentially the peak energies in the phoneme spectrum. The vibration of the vocal cords produces a simple phoneme. The frequencies in which resonance occurs are the formants. One of the methods to calculate formants is using the autoregressive (AR) model.

The AR model is one of a group of linear prediction formulas that attempt to predict an output y[n] of a system based on the previous outputs and inputs. In Figure 4, all the averaged speakers' frequency response plots were superimposed for each of the vowels. It is easily visible how closely the responses for different speakers match up for any of the vowels.

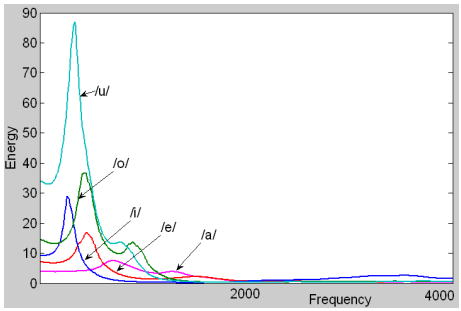

Figure 4: Average Frequency Response Plot of Vowels (linear scaled y-axis)

Based on the observation and analysis of the plotted outputs, significant differences were found between the each of the vowel frequency responses on certain range on the frequency ranges. In terms of differentiating vowels, these energy differences can be used as features that to classify the vowels. Energy parameters such as maximum, mean and minimum values were calculated from these ranges for each speech sample representing each of the vowels.

2.4. Determining the frequency ranges

There are seven ranges of frequency used to extract energy features from the vocal tract model. Five of them were determined using Bandwidth Approach (BW) on first peak of every averaged vowel response. Figure 5 shows the frequency range from the first peak response of vowel /u/

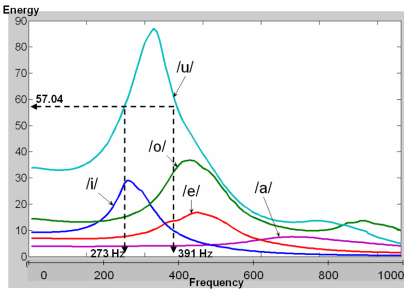

Figure 5: Determining BW frequency ranges for vowel /u/

Two more ranges were determined by comparing each of the average vowel plots. The sixth range is between 1171.88Hz and 1562.50Hz and the seventh range 1796.88Hz and 1953.13Hz as shown in Figure 6.

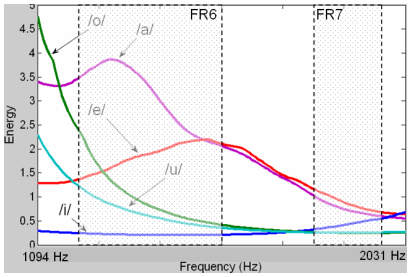

Figure 6: Location of FR6 and FR7

Table 2 shows the ranges of frequency that were used to extract mean and maximum energy features. Fourteen energy features were extracted based on the frequency ranges from Table 2.

Table 2: Frequency Ranges to Extract Energy Features

Frequency Range	f start (Hz)	f stop (Hz)
FR1	578.13	898.44
FR2	382.81	531.25
FR3	234.38	320.31
FR4	367.19	515.63
FR5	273.44	390.63
FR6	1171.88	1562.50
FR7	1796.88	1953.13

2.5. Classifying results using neural network

A Multi-layer Perceptron (MLP) with two hidden layers was used in this study to identify the vowel utterances. There are 14 input neurons representing 14 different parameters obtained from the frequency response ranges, 2 layers of 10 hidden neurons and 3 output neurons. The vowel /a/, /e/, /i/, /o/ and /u/ are represented by the 3-bit output neurons having values of 001, 010, 011, 100 and 101. The network was trained using 70% of the data using learning rate of 0.3 and momentum factor of 0.8. The weights and biases of the MLP were initialized randomly.

For comparison purposes, the first 3-formant values were obtained from the same set of data using the 10th order Linear Predictive Coding (LPC). The formant classification was used to compare with the performance of the bandwidth approach. Table 3 below shows the classification rate of different testing tolerance of method BW and LPC. A testing tolerance

of 0.2 was selected based on its accuracy and its permissible limit of variation.

Table 3: Comparison of BW and LPC (Classification Rate of Different Testing Tolerance)

Testing Tolerance	BW Accuracy	LPC Accuracy	Diff of BW over LPC
0.05	86.59%	35.82%	50.77%
0.1	94.86%	84.94%	9.92%
0.15	96.05%	88.75%	7.30%
0.2	**96.59%**	**91.10%**	**5.49%**
0.25	96.84%	92.39%	4.45%
0.3	97.09%	93.43%	3.66%

Table 4 below shows the summary of the averaged results of the classification based on testing tolerance of 0.2.

Table 4: Classification Results of Different Methods

Methods	Parameters Classified	CR (test tol=0.2)	MSE (averaged)
BW	14 energy	96.59	0.006
LPC	f1, f2 & f3	91.10	0.006

BW obtained a classification accuracy of 96.59%, which is 5.49% better than the 3-formant LPC classification.

The Table 5 shows the improvement in vowel classification rate of BW over LPC method. All the five vowels show improvement using the BW method over the LPC especially for vowel /u/ and /o/ that showed improvement of 8.86% and 11.38%. Figure 7 shows the averaged classification rate of overall vowels for both BW and LPC methods.

Table 5: Classification Rate of Vowels for BW and LPC methods

Vowel	Improvement of BW over LPC
/a/	5.47%
/e/	1.18%
/i/	1.00%
/o/	8.86%
/u/	11.38%

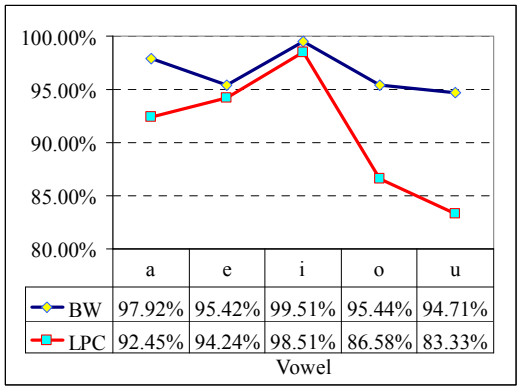

Figure 7: Classification Rate by Vowel (BW vs. LPC)

2.6. Classifying results using Multinomial Logistic Regression (MLR)

Logistic regression is used to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. The multinomial logistic model for the choice probabilities is given by equation 4.

$$\Pr(i \mid C) = \frac{\exp(x_i^{'}\beta)}{\sum_{j=1}^{n}\exp(x_j^{'}\beta)} \qquad (4)$$

where β is a vector of unknown regression parameters.

The features obtained for BW approach for each of vowels are classified again using MLR method. Its results are then compared with classification rates from MLP. The results are shown in Table 6.

Table 6: Classification Rate of BW using MLR and MLP

Classifying Methods	CR %	Train Time
MLR (Full Factorial)	96.40	~0.5 min
MLP (0.2 test tolerance)	96.59	~30 min

The classification rate between MLR and MLP are comparable with each other. In terms of training time, MLR really outperforms MLP. MLR only took a fraction of the time to classify the vowels from BW approach than the MLP method.

Table 7 shows the classification rate of vowels using BW approach based on MLR and MLP. Bold percentage in the table shows greater classification rate.

Table7: Classification Rate of Vowels

	Bandwidth Approach (BW)	
Vowel	MLR	MLP
/a/	97.64%	**97.92%**
/e/	93.33%	**95.42%**
/i/	**100.00%**	99.51%
/o/	**95.58%**	95.44%
/u/	**95.71%**	94.71%

Based on MLR classification, vowel /i/ achieved perfect classification and the rest of the vowels achieved greater than 93% classification rate.

3. Conclusion

In this paper, a new feature extraction method based on parameters of the vocal tract model was presented. This method uses Bandwidth Approach to determine the range of frequency to extract the energy features. Energy features were obtained from specific ranges of frequencies from the frequency response plots that were determined by the Bandwidth Approach.

The data were classified using a using a Back-propagation Neural Network. The training function of this network updates weight and bias values according to gradient descent momentum and an adaptive learning rate. Classification results from different testing tolerance were obtained and the testing tolerance selected was 0.2 based on its accuracy and its permissible limit of variation. The overall classification accuracy of the BW method for testing tolerance of 0.2 was 96.59 %, which is 5.49% higher than the LPC method. Vowel /i/ obtained the best classification rate of 99.51%, followed by /a/, /o/, /e/ and /u/. This result makes this method a good choice to detect the vowels especially the vowel /i/. BW also improves the vowel classification rate of all five vowels over the LPC method especially for vowel /o/ and /u/.

Based on MLR classification, the classification rates achieved for full factorial mode was 96.40%, which is comparable with MLP classification of 96.59%. In terms of training time, MLR only needed less than a minute compared to MLP that requires 30 minutes.

4. References

[1] Rabiner, L. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

[2] G.E.Carlson, *Signal and Linear System Analysis*, Houghton Mifflin Company, 1992

[3] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel identification*," IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-25*, p. 183, 1977.

[4] G. Fant, "Acoustic Theory of Speech Production", *The Hague*, The Netherlands: Mouton, 1960.

[5] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels*," Journal of Acoustic Society of America*, vol. 24, pp. 175–184, 1952.

[6] J. Hillenbrand, L.A. Getty, Clark, M.J., Wheeler, k. "Acoustic Characteristics of American English vowels." *Journal of Acoustic Society of America*, 1995, pg 97, 3099-3111.

[7] V. Vuckovic, M. Stankovic, "Formant analysis and vowel classification methods", *5th International Conference on Telecommunications in Modern Satellite*, Cable and Broadcasting Service (TELSIKS), 2001

[8] X. Huang, et al., *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001.