Development of Feature Set, Classification Implementation and Applications for Vowel Migration/Modification in Sung Filipino (Tagalog) Texts and Perceived Intelligibility

Virginia B. Bustos[1], Triah Joyce G. Dela Cruz[1], Ramon Maria G. Acoymo[2] and Rowena Cristina L. Guevara [1*]

[1]Electrical and Electronics Engineering Institute, College of Engineering , University of the Philippines
Diliman 1101 Quezon City

[2]Voice and Music Theater/Dance Department, College of Music University of the Philippines,
Diliman 1101 Quezon City

*Corresponding author: gev@eee.upd.edu.ph

Received: 5 November 2009; Revised: 18 February 2010; Accepted: 24 June 2010

ABSTRACT

With the emergence of research on real-time visual feedback to supplement vocal pedagogy, the utilization of technology in the world of music is now seen to accelerate skills learning and enhance cognitive development. The researchers of this project aim to further analyze vowel intelligibility and develop software applications intended to be used not only by professional singers but also by individuals who wish to improve their singing capability.

Data in the form of sung vowels and song pieces were obtained from 46 singers. A Listening Test was then conducted on these samples to obtain the ground truth for vowel classification based on human perception. Simulation of the human auditory perception of sung Filipino vowels was performed using formant frequencies and Mel-frequency cepstral coefficients as feature vector inputs to a two-stage Discriminant Analysis classifier. The setup resulted in an over-all Training Set accuracy of 89.4% and an over-all Test Set accuracy of 90.9%. The accuracy of the classifier, measured in terms of the correspondence of vowel classifications obtained from the classifier with the results of the Listening Test, reached 92.3%. Using information obtained from the classifier, offline and online/real-time software applications were developed.

The main application features include the display of the spectral envelope and spectrogram, pitch and vibrato analysis and direct feedback on the classification of the sung vowel. These features were recommended by singers who were surveyed and were incorporated in the applications to aid singers to adjust formant locations, directly determine listener's perception of sung vowels, perform modeling effectively and carry out vowel migration.

Keywords: Filipino, Vowel Migration, Intelligibility

INTRODUCTION

One of the objectives of the singers of bel canto is the development of a vocal scale without interruption throughout its length. (Appelman, 1986) This demands vowel modification in the upper notes to preserve the true vowel sound and to prevent notes from becoming disagreeable or harsh. The technique has become a means of transition to the upper voice for many centuries. The singer, when modifying a vowel, is actually causing the vowel to migrate in the direction of another recognizable vowel. Vowel migration causes the vowel to lose its integrity for the sake of enhancing musical aesthetics. Thus, intelligibility, the degree or level to which the intended vowel of the singer is perceived correctly by the listeners, is diminished. For example, when the vowel /a/ sung by a singer in its unmigrated or unmodified form is correctly perceived as /a/ by all the listeners, the intelligibility is 100%. However, when the same vowel is migrated to another vowel, fewer listeners might perceive the vowel correctly. For example, when 52 out of 100 listeners correctly perceive the intended vowel, resulting intelligibility is 52%.

A more quantitative and objective measure of these vowel modifications can be achieved through software applications that offer real-time visual feedback on vowel intelligibility, together with the analysis of pitch and frequency spectra. The assessment is helpful in determining the extent with which the acoustical demands can be met with minimal compromise in intelligibility.

The classification of sung vowels is usually performed by listeners. Research in vowel classification generally applies to spoken vowels through the use of Automatic Speech Recognition (ASR) systems. These systems can be implemented using numerous feature extraction algorithms and classification models as shown by different studies presented in Table 1.

Table 1. Studies in Vowel Classification for Spoken Vowels

Researcher/s	Data (number of speech samples)	Features used	Classifi-cation Model	Achieved Accuracy
Merkx and Miles	17213	MFCCs	Single Layer Feed-forward ANN	91.50%
Dumitri and Gavat	145	MFCCs	3-Layer MLP	96.4% for male speakers 77.9%% for female speakers
Schmid and Barnard	58268	Formant Features and MFCCs	MLP	73.40%

The study made by Merkx and Miles (2005) utilized thirteen Mel-Frequency Cepstral Coefficients (MFCCs) as feature vectors. MFCCs are coefficients derived from a logarithmic scaling of audio frequencies using Mel filterbanks. For pattern classification, the study used a feed-forward Artificial Neural Network (ANN) with 28 internal nodes. ANN is an adaptive, often nonlinear, system that is trained to perform an input/output mapping using a given input and a target data. The classifier reached a recognition accuracy of 91.5% on a subset of 5 vowel phonemes. Another study made by Dumitru and Gavat (2007) used twelve MFCCs, a 3-layer Multilayer Perceptron (MLP) classifier and 145 speech samples as input data. MLP is a feedforward ANN that uses three or more layers of nodes with nonlinear activation functions. Vowel recognition rates reached 96.4% for male speakers and 77.9% for female speakers. Schmid and Barnard (1997) tested the efficiencies of cepstral-based features, MFCCs and formant features including formant trajectory, amplitude, bandwidth, pitch and segment duration in vowel classification. Formants are resonating frequencies that show up as peaks in the sound spectrum. Results showed that formant features alone reached an accuracy rate of 71.8%, while MFCCs alone reached an accuracy rate of 71.6%. The combination of the two features proved optimal, reaching an accuracy rate of 73.4%.

This project is part of a research track at the UP Digital Signal Processing Laboratory. The preceding study (Dimaculangan & Felias, 2008) applied strategies used in ASRs to the classification of sung vowels and determined the relationship between vowel migration in sung Filipino text and perceived intelligibility from the perspective of the audience. Several parameters including formants, spectral envelopes, vowel triangles and MFCCs were investigated to identify the parameter that is most appropriate in simulating vowel perception. Results showed that nine MFCCs, extracted from each of the 281 sung unmigrated vowel samples, best discriminated the sung vowels. These features were used as inputs to a Linear Discriminant Analysis (LDA) classifier that proved optimal compared to an ANN. LDA computes a linear predictor from two sets of normally distributed data to allow for classification of new observations. The accuracies achieved were 68.1% for the Training Set and 52.9% for the Test set, based on the correspondence of the vowel classification determined by the classifier and results of the Intelligibility Tests that were conducted. In the Intelligibility Tests the ground truth for vowel classification was based on the perception of 45 listeners.

Wilson et al. (2008) studied the effects of real-time visual feedback on teaching pitch accuracy in singing. They investigated whether the style of feedback affects the amount of learning achieved and whether the provision of concurrent visual feedback hampers the simultaneous performance of the singing task. Through the implementation of a baseline-intervention-post test between-groups design, it was determined that real-time visual feedback to the learner promotes the acquisition of the neuromuscular skills underlying the task of singing the correct pitch. Moreover, different styles of visual feedback did not produce differences in the amount and rate of learning.

In this paper, the objective of the researchers is to develop software applications that provide an objective assessment of the intelligibility of sung vowels, based on the perception of listeners, through real-time visual feedback. Furthermore, the applications should help singers readily assimilate feedback and improve their singing ability. The novelty of this project is the application of ASR system techniques on sung vowels in unmigrated and migrated forms with emphasis on intelligibility based on the perception of listeners.

METHODOLOGY

This section contains a description of the data, a discussion of the various extracted features, the pattern classification methods, the application development and testing of these applications.

Data

Audio recording of the data was held inside a WhisperRoom®, a sound isolation enclosure. The recording equipment includes TASCAM DV-RA1000 High definition Audio Master Recorder, Sennheiser Ew100 G2 wireless microphone and Behringer Ultravoice XM8500 wired microphone. All audio data samples are recorded in stereo wav format with the following attributes: PCM signed 24bit, 44.1 kHz sampling rate and 2116 kbps bit rate. Video recordings were also done using a Sony Handycam SR220 to document the singers' mouth shapes.

Data in the form of sung vowels and song pieces were obtained from 46 singers. The audio recording consists of two parts, vocalization and singing of folk songs. Vocalization was divided into six sets: unmigrated, migrated to [a], migrated to [e], migrated to [i], migrated to [o] and migrated to [u]. Each set comprised vocalizes for each of the five Filipino vowels in increasing pitch, spanning the singer's stable range of frequency. The second part of the recording involved singing of three folk songs: "Si Pilemon", "Lubi-Lubi" and "Neneng at Nonoy". Each of the songs was sung unmigrated, migrated to [a], migrated to [e], migrated to [i], migrated to [o] and migrated to [u].

A survey regarding application design was conducted among 38 of the singers who participated in the project. Singers were asked about software features that would be helpful in improving their singing ability and how they visualize these features to appear in the applications. Information gathered in this stage served as the basis for the design of the applications that were developed.

An Intelligibility/Listening Test was done to establish the ground truth for the intelligibility of the recorded vocalizes. 100 listeners composed of 51 non-music major students and 49 students from the UP College of Music participated in this test. A total of 1350 sliced vowels were included in the test, utilizing the vowel preceding the last vowel note for singers who chose to vocalize within an octave, while the second to the last vowel note was used for singers whose vocalises exceeded an octave. Data obtained from the Intelligibility Test served as the ground truth for vowel classification based on human perception.

Feature Extraction

Before extracting the features from the data, signal pre-processing was done to remove the DC offset of the signal and normalize the average audio volume levels to -18 dB. Vowel detection was implemented by computing short-time energy and short-time zero crossing rate per signal frame then comparing these parameters to set thresholds.

Several features have been extracted from the vowels and were used as inputs to different pattern classifiers. Below is a summary of the extracted features from the vowels.

- **Mel-Frequency Cepstral Coefficients (MFCC)** is a representation of cepstral coefficients, taken from the Fourier transform of the decibel spectrum, wherein the analysis is done on a non-linear frequency scale known as the Mel scale. The MFCC extraction algorithm used was developed by Slaney (1998). The computation process is as follows.

The signal is divided into short time windows, where the Discrete Fourier transform (DFT) of each time window for the discrete-time signal *x(n)* with length *N* is computed using

$$X(k)=\sum_{n=0}^{N-1} w(n)x(n)\exp(-j2\pi kn/N) \qquad (1)$$

for *k = 0, 1, . . . ,N − 1,* where *k* corresponds to the frequency $f(k) = kf_s/N$, f_s is the sampling frequency in Hertz and *w(n)* is a Hamming window, given by

$$w(n)=0.54-0.46\cos(\pi n/N). \qquad (2)$$

The magnitude spectrum |*X(k)*| is scaled in both frequency and magnitude. The frequency is scaled logarithmically using the Mel filter bank *H(k,m)* using

$$X'(m)=\ln\left(\sum_{k=0}^{N-1}|X(k)|\cdot H(k,m)\right) \qquad (3)$$

for *m = 1, 2, . . . ,M*, where *M* is the number of filter banks.

The Mel filter bank is a collection of triangular filters defined by the center frequencies $f_c(m)$, as shown in Eq. 4.

$$H(k,m)=\begin{cases} 0 & for\ f(k)<f_c(m-1) \\ \dfrac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & for\ f_c(m-1)\le f(k)<f_c(m) \\ \dfrac{f(k)-f_c(m-1)}{f_c(m)-f_c(m+1)} & for\ f_c(m)\le f(k)<f_c(m+1) \\ 0 & for\ f(k)\ge f_c(m+1) \end{cases} \qquad (4)$$

The center frequencies of the filter banks are computed by approximating the Mel scale using Eq. 5.

$$\phi=2595\log_{10}\left(\frac{f}{100}+1\right). \qquad (5)$$

A fixed frequency resolution in the Mel scale is computed, corresponding to a logarithmic scaling of the repetition frequency, using Eq. 6, where ϕ_{max} is the highest frequency of the filter bank on the Mel scale, ϕ_{min} is the lowest frequency in Mel scale.

$$\Delta\phi=(\phi_{max}-\phi_{min})/(M+1) \qquad (6)$$

The center frequencies on the Mel scale are given by Eq. 7 for *m = 1, 2, . . . ,M*. The center frequencies in Hertz can be obtained using Eq. 8.

$$\phi_c(m)=m\cdot\Delta\phi \qquad (7)$$

$$f_c(m)=700\left(10^{\frac{\phi_c(m)}{2595}}-1\right) \qquad (8)$$

The MFCCs are obtained by computing the Discrete Cosine Transform (DCT) of *X'(m)* from Eq. 3 using Eq. 9 for *l = 1, 2, . . . ,M*, where *c(l)* is the *lth* MFCC.

$$c(l)=\sum_{m=1}^{M} X'(m)\cos\left(l\frac{\pi}{m}m-1\right) \qquad (9)$$

- **F1+F2+MFCC** is a combination of the first and second formants, *F1* and *F2,* of the vowels and the extracted MFCCs. Formant frequencies characterize the acoustic structure of each vowel, enabling listeners to perceptually identify the vowel. The vowel formants were appended to the MFCCs in two ways: on a formants per frame basis ($F1_f$+$F2_f$+MFCC) as shown in Figure 1 and on a formants per vowel basis ($F1_v$+$F2_v$+MFCC) as shown in Figure 2.

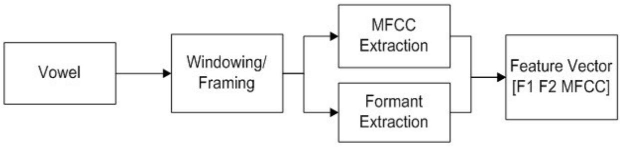


Figure 1. Formants computed per frame appended to MFCCs

The extraction of formant frequencies involves the determination of resonance peaks from the filter coefficients obtained through Linear Prediction Coding (LPC) analysis of the signal. (Makhoul,

1972) Once the prediction polynomial A(z), shown in Eq. 10, has been calculated, the formant parameters are determined by solving for the roots of the equation A(z) = 0.

$$A(z)=1+a_1 x^{-1}+a_2 z^{-2}+...+a_p z^{-p} \qquad (10)$$

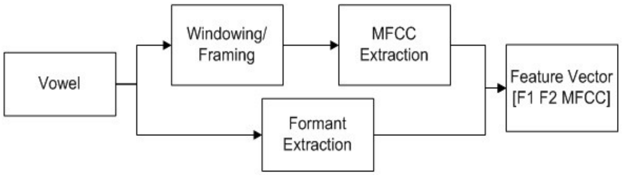


Figure 2. Formants computed per vowel appended to MFCCs

- **MFCC+ΔMFCCs** is a combination of the MFCCs and its derivatives calculated using a simple linear slope.

- **Line Spectral Frequencies (LSF) or Line Spectral Pairs** is a representation of LPC coefficients that represents glottal activity.

To determine the LSFs, the Linear Prediction polynomial shown in Eq. 10 is decomposed into *P(z)* and *Q(z)* as shown in Eq. 11 where P(z) corresponds to the vocal tract with the glottis closed and Q(z) with the glottis open. The roots of both polynomials represent the Line Spectral Pairs.

$$\begin{aligned} P(z)&=A(z)+z^{-(p+1)}A(z^{-1}) \\ Q(z)&=A(z)-z^{-(p+1)}A(z^{-1}) \end{aligned} \qquad (11)$$

- **MFCC-RASTA** is the addition of a RASTA (Relative Spectral) filter block after MFCC extraction. The RASTA filter was approximated by a simple fourth order Butterworth bandpass filter. (Slaney, 1998)

Pattern Classification

The extracted features from the audio signals were then used as feature vector inputs to pattern classifiers. Below is a summary of the pattern classification methods that were implemented.

- **Discriminant Analysis (DA)** is used to find a number of projection directions that are efficient in separating the features into classes. The process involves maximizing the ratio of between-class variance to within-class variance so that adequate class discrimination is obtained. The different types of DA that were implemented were Linear (LDA), Quadratic (QDA), Mahalanobis (MDA), Diagonal Linear and Diagonal Quadratic.

- **Feedforward Backpropagation Artificial Neural Network** is a type of ANN that is trained using input vectors and corresponding target output vectors until it can approximate a function and associate input vectors with specific output vectors. This is done by reducing calculated errors between the input and output data and consequently adjusting the weights of the network's forward-connected layers.

- **Classification Tree (CT)** is a type of machine learning algorithm used for non-parametric data classification. A classification tree is a structural mapping of binary decisions that lead to a decision about the class (interpretation) of an object.

- **Support Vector Machines (SVM)** are decision-based prediction algorithms which can classify data into two groups. The training data is mapped to a higher dimensional space and separated by a plane defining the two classes of data. Input data are classified based on the side of the plane they fall on.

Cross Validation

A comparison of the performance of the classifiers was done to determine the optimal sung Filipino vowel discriminator. The classifiers were compared using the Training Set, consisting of sung vowels from 80% of the singers, and the Test Set, consisting of sung vowels from the remaining 20%. Singers belonging to each set were randomly chosen. The Listening Test vowels were also tested to determine the perception accuracy.

Development of Applications

In this project, two software applications have been developed using Matlab®, a command-line software development program. One software application runs offline and processes audio files while the other software application runs in real-time/online. Screen display interface was developed for both software applications. The following features were incorporated in the applications based on the recommendation of the 38 singers who were surveyed.

a. Pitch Detection

Two algorithms, average magnitude difference function - autocorrelation function (AMDF-ACF) and the correlogram model of pitch perception (Slaney, 1998), were tested to develop an accurate pitch estimator. The AMDF-ACF algorithm extracts the pitch period of signals from the short-term autocorrelation of computed AMDF values as shown in Eq. 12.

$$R(k)=\sum_{n=0}^{N-k-1} x(n)x(n+k) \tag{12}$$

The correlogram model of pitch perception uses the largest peak from a summarized autocorrelation plot, a plot of sample autocorrelations versus time lags, as the pitch estimate.

b. Calculation of Vibrato Parameters

Vibrato was represented by three parameters: intonation, rate and extent. The intonation curve was obtained by passing the instantaneous pitch frequency curve or pitch contour of a signal through a moving average filter. Vibrato rate was estimated as the reciprocal of the maximum period of the pitch contour. Moreover, vibrato extent was estimated as the mean amplitude difference between the pitch and intonation contours.

Testing of Developed Software Applications

For preliminary testing, Dean Ramon Acoymo of the College of Music assessed the performance and usefulness of the application's features. Further testing was implemented using vocalises of singers. Vocalises were recorded while singers tested the Online Application. The recorded vocalises were then run on the Offline Application to match the results of both applications. Singers were also asked to fill out a questionnaire assessing the performance and usefulness of the application.

RESULTS AND ANALYSIS

Optimal Unmigrated Vowel Classifier

In order to develop an effective vowel classifier, initial tests on unmigrated vowels were conducted. The samples were assumed to be perceived correctly by the listeners, thus accuracy of the tests depended on the intended vowel classification. Several feature extraction and pattern classification methods were tested on the vowels to determine the optimal vowel classifier elements. The features that were extracted included: MFCCs, $F1_f$+$F2_f$+MFCC, $F1_v$+$F2_v$+MFCC and MFCC+ΔMFCCs. Moreover, the implemented pattern classification methods were Discriminant Analysis, Classification Tree and Feedforward Backpropagation Neural Network. This phase was done in parallel with the Listening Test. Table 2 shows the top 5 classifiers ranked based on Training Set and Test Set accuracies computed as the average accuracy per vowel.

Table 2. Top 5 classifiers based on Training Set and Test Set accuracies

Number of Coeffi-cients	Features Extracted	Pattern Classifi-cation Method	Training Set Accuracy	Test Set Accuracy
15	$F1_f$+$F2_f$+ MFCC	QDA	85.90%	89.40%
21	$F1_f$+$F2_f$+ MFCC	QDA	87.80%	86.50%
15	$F1_v$+$F2_v$+ MFCC	QDA	86.40%	89.10%
21	$F1_v$+$F2_v$+ MFCC	QDA	87.80%	88.60%
21	$F1_v$+$F2_v$+ MFCC	CT	100.00%	84.90%

Development of Data Set based on Listener Perception

The Listening Test that was conducted resulted in 1041 consistently perceived vowels out of the 1350 sung vowels that were presented to the listeners. A consistently perceived vowel has the same classification for 60% or more of the listeners in the Listening Test. To increase the data used in the development of the classifier based on listener perception, k-means clustering based on formants was implemented on consistently perceived vowels. K-means clustering is a partitioning method that finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. The resulting clusters of the vowels were labeled with vowel classifications similar to the consistently perceived vowels from the Listening Test and included in the data set. Unless stated otherwise, this is the developed data set cited in this paper.

The number of vowels that are included in the Training Set and Test Set using the data set is shown in Figures 3 and 4, respectively.

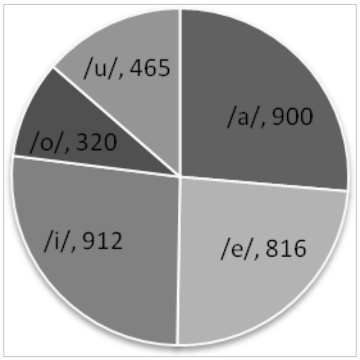


Figure 3. Vowel distribution of training set.

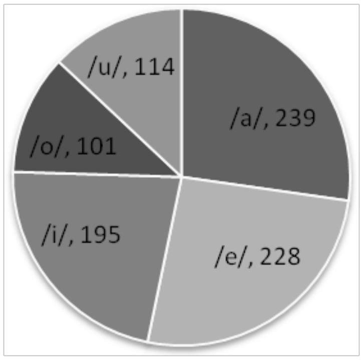


Figure 4. Vowel distribution of test set.

Test on Project 1 Listening Test Data

To determine the optimal classifier based on listener perception, the top 5 unmigrated vowel classifiers from Table 2 were tested on the Listening Test data gathered from Project 1 (Dimaculangan & Felias, 2008). Project 1 data were recorded in a large hall. The classifiers were trained using data from 80% of the singers included in the developed data set based on listener perception. The results are summarized in Table 3. The trained classifier with the highest accuracy is the 21 $F1_v$+$F2_v$+MFCC QDA.

Classifier Optimizations

The trained classifier with the highest accuracy, 21 $F1_v$+$F2_v$+MFCC QDA, was used to test the remaining data from 20% of the singers in the developed data set. The resulting accuracies, computed based on the correspondence of the classifier outputs to the labeled vowel classifications based on listener perception, are shown for each vowel in Table 4.

Table 4. Accuracies for each Vowel using the Developed Data Set

Vowel	Training Set Accuracy	Test Set Accuracy	Perception Accuracy
/a/	80.00%	91.60%	87.20%
/e/	91.40%	90.40%	94.70%
/i/	93.10%	96.90%	94.00%
/o/	81.60%	63.40%	84.30%
/u/	86.90%	89.50%	86.60%

To address the speed constraints of the software applications, minimization of the Training Set size was employed. Feature vectors from the fifteen middle frames of each Training Set vowels were taken to form a new Training Set. The minimization has reduced classification time from 1245.5ms to 131.5ms. Furthermore, accuracies have improved as shown in Table 5. However, using the new Training Set, the Perception accuracies for the vowel /o/ have decreased due to misclassification of the vowel /a/.

Table 5. Accuracies for the Minimized Training Set

Vowel	Training Set Accuracy	Test Set Accuracy	Perception Accuracy
/a/	82.30%	92.10%	88.90%
/e/	91.90%	92.50%	94.70%
/i/	95.50%	98.00%	96.50%
/o/	80.60%	68.30%	81.90%
/u/	89.50%	90.40%	91.10%

A second stage classifier was developed to address the confusion between the vowels /a/ and /o/, with the goal of increasing the accuracies for both vowels. The features that were extracted included LSF, Formants, MFCC and MFCC-Rasta. Moreover, the pattern classification methods that were implemented were Discriminant Analysis, Support Vector Machine and Classification Tree. Comparing the results of testing the combination of features and pattern classification methods showed that the $F1_v$+$F2_v$+MFCC and LDA combination had the highest accuracy in classifying the vowels. The second-stage classifier takes the first-stage classifier /a/ and /o/ outputs whenever the classification accuracy is below 50% for /a/ and below 80% for

/o/, following computed optimal thresholds. The resulting accuracies for vowels /a/ and /o/ after adding the second stage classifier are shown in Table 6, the perception accuracy has improved for both vowels.

Table 6. Accuracies for Vowels /a/ and /o/ After Adding the Second Stage Classifier

Vowel	Training Set	Test Set	Perception Accuracy
/a/	84.40%	92.50%	89.20%
/o/	79.70%	70.30%	83.10%

Final Classifier

The final vowel classifier based on listener perception consists of a first-stage 21 $F1_V+F2_V$+MFCC QDA classifier with a second-stage classifier 21 $F1_V+F2_V$+MFCC LDA classifier for the vowels /a/ and /o/. The Training Set size was minimized by using only the 15 middle frames of the Training Set vowels.

Confusion matrices of the final classifier for the Training Set, Test Set and Listening Test vowels used to compute the perception accuracy are shown in Tables 7, 8 and 9, respectively. The diagonal elements of the confusion matrix represent the correctly classified vowels. Off-diagonal elements denote the confusion of one vowel for another vowel – each row of the confusion matrix sums up to 100%. The first row in Table 7 is interpreted as follows: the vowel /a/ in the Training Set is classified as /a/ 84.4% of the time and classified as /e/ 2.4% of the time, as /i/ 0.3% of the time, as /o/ 11.8% of the time and as /u/ 1% of the time. It can be observed that there is confusion among the back and central vowels /a/, /o/ and /u/ which are mainly caused by the overlapping formant values (*F1* and *F2*) of the three vowels. Moreover, among all the vowels, /i/ is the most accurately classified vowel.

Table 7. Confusion Matrices for the Training Set

Training Set					
	/a/	**/e/**	**/i/**	**/o/**	**/u/**
/a/	**84.40%**	2.40%	0.30%	11.80%	1.00%
/e/	2.80%	**91.90%**	4.70%	0.10%	0.50%
/i/	0.20%	3.60%	**95.50%**	0.00%	0.70%
/o/	14.40%	0.00%	0.00%	**79.70%**	5.90%
/u/	1.70%	0.70%	2.80%	5.40%	**89.50%**

Table 8. Confusion Matrices for the Test Set

Test Set					
	/a/	**/e/**	**/i/**	**/o/**	**/u/**
/a/	**92.50%**	1.30%	0.00%	5.40%	0.80%
/e/	0.40%	**92.50%**	7.00%	0.00%	0.00%
/i/	0.50%	1.50%	**98.00%**	0.00%	0.00%
/o/	19.80%	1.00%	0.00%	**70.30%**	8.90%
/u/	4.40%	1.80%	1.80%	1.80%	**90.40%**

Table 9. Confusion Matrices for the Listening Test Vowels

Listening Test Vowels					
	/a/	**/e/**	**/i/**	**/o/**	**/u/**
/a/	**89.20%**	2.00%	0.70%	8.10%	0.00%
/e/	1.90%	**94.70%**	3.00%	0.00%	0.40%
/i/	0.00%	2.80%	**96.50%**	0.00%	0.70%
/o/	10.80%	0.00%	0.00%	**83.10%**	6.00%
/u/	0.00%	0.90%	4.50%	3.60%	**91.10%**

The final classifier was used in the developed offline application. The online application, on the other hand, used a further minimized classifier that utilizes only the five middle frames of the Training Set vowels to address the greater need for classification speed.

Overall accuracies for the offline application vowel classifier are 89.4% for the Training Set and 90.9% for the Test Set. Overall accuracies for the online application vowel classifier are 89.4% for the Training Set and 89.7% for the Test Set. The overall perception accuracy for both classifiers is 92.3%.

It was observed that the Test Set had a higher accuracy than the Training Set. The same result was observed after changing the singers included in both sets.

The improvement over Project 1 baseline accuracy (Dimaculangan & Felias, 2008) for vowel classification is 21.3% for the Training Set and 38.0% for the Test Set using the offline classifier. The improvement for the online classifier is 21.3% for the Training Set and 36.8% for the Test Set. The objective of improving the accuracy of the classifier was achieved.

Pitch Estimation Algorithms

Two algorithms, AMDF-ACF and the correlogram model of pitch perception, were tested to develop an accurate pitch estimator. Sinusoids with fundamental frequencies ranging from 65 to 932 Hz (C2 to A#5) were used as inputs to the pitch estimation algorithms.

For the AMDF-ACF algorithm, large deviations, averaging 23.86 Hz, of the pitch estimates from the test fundamental frequencies were observed especially in high frequencies. On the other hand, the correlogram model resulted in more accurate pitch estimates with smaller deviations, averaging 4.95 Hz, from the fundamental frequencies occurring only at very high frequencies. From this result, the correlogram model of pitch perception was set as the pitch estimation algorithm for the software applications.

Vibrato Rate and Extent Estimation

The performance of the vibrato rate and extent estimation was tested using synthesized vowels with duration of 2 seconds and formant frequencies of 300 and 870 Hz. The synthesized vowels had pitch ranging from 65 to 835 Hz. Vibrato extent was set to be 3% of the pitch value of the vowel and the vibrato rates ranged from 4 Hz to 7 Hz.

The average deviation of the estimated vibrato rates from the set vibrato rates is 0.42 Hz while average deviation of estimated vibrato extents from the values computed as 3% of the pitch frequencies, is 4.77 Hz. Inconsistencies in the estimation, especially at high frequencies, are attributed to deviations in the pitch estimate and ripple effects from the pitch interpolation and moving average filter.

Offline Application

The screen display of the offline application with the corresponding feature labels is shown in Figure 5.

The offline application takes transcribed audio files as input and displays the following features: spectral envelope, pitch contour, intonation contour, and estimates for pitch, vibrato rate and vibrato extent, formant frequencies, spectrogram and vowel classification. The vowel classification uses the 21 $F1_V$+$F2_V$+MFCC QDA-LDA classifier with 15 frames per Training Set vowel. Options for vowel detection using short-time energy, transcription file reading and vowel synthesis are also included in the application.

Online/Real-time Application

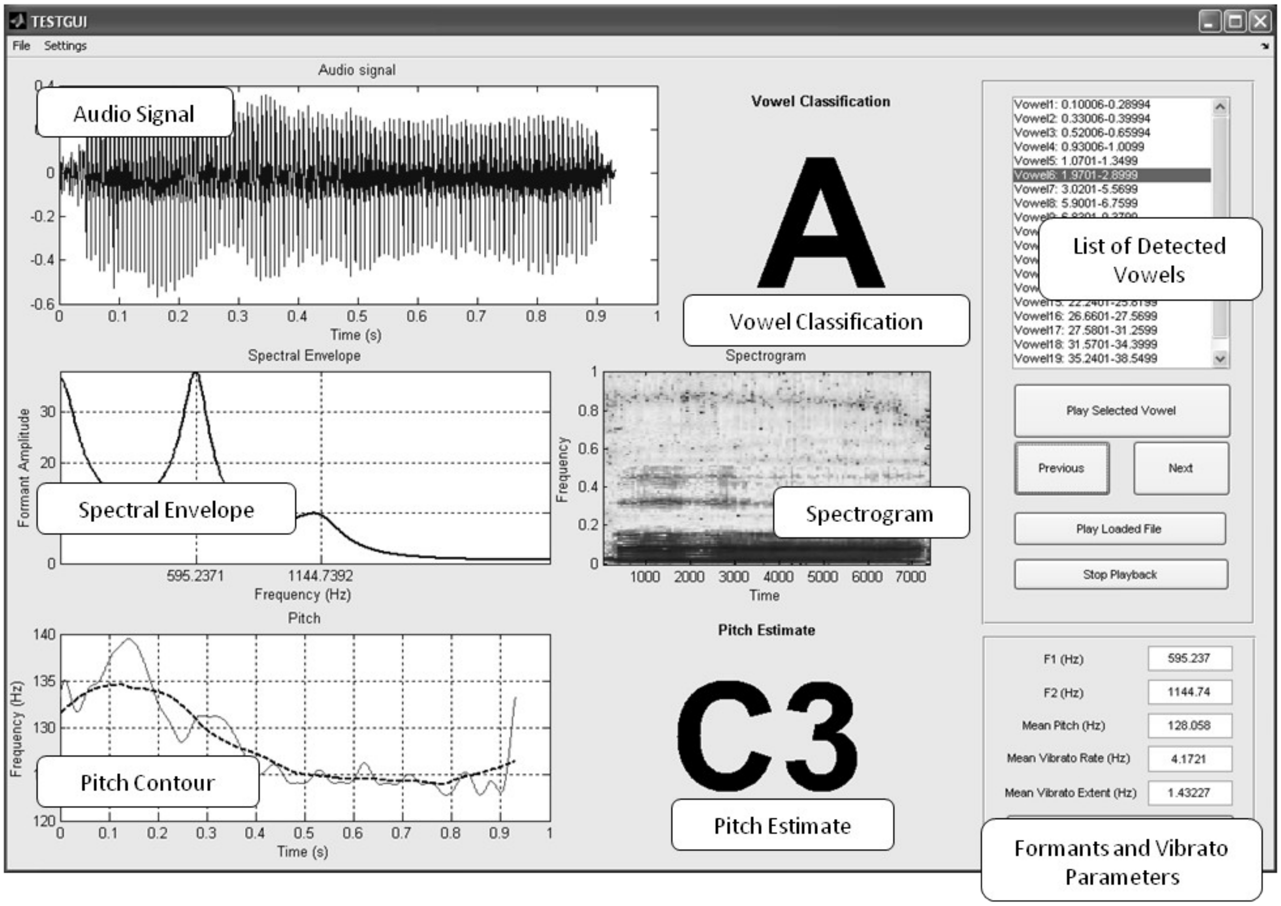
The screen display of the online application with the corresponding feature labels is shown in Figure 6.

The application continuously takes 250ms of audio input from a microphone. Vowels are detected using short-time energy and ZCR. The following features are included in the application and are displayed: spectral envelope, pitch contour, pitch estimate, formants, vibrato rate, vibrato extent, spectrogram and vowel classification. The vowel classification uses the 21 $F1_V$+$F2_V$+MFCC QDA-LDA classifier utilizing 5 frames per Training Set vowel. Options for audio logging, audio exporting and accompaniment playback are also included in the application. The average processing time for each detected vowel was computed to be 193ms.

Testing of Software Applications

Testing with Dean Ramon Acoymo of the UP College of Music was conducted to gauge the performance and usefulness of the features included in the applications.

According to Dean Acoymo, the vowel classifications made by the classifier were consistent with human perception. The human ear usually has trouble discriminating among the vowels /a/, /o/ and /u/ due to the intersection of the formant values of these vowels. The display of the vowel classification and spectral envelope was advantageous in the assessment of the trade-off between vocal color and quality. Moreover, the features are applicable in singing pedagogy wherein preserving both vocal color and quality is important. The display of the pitch and vibrato parameters, on the other hand, is especially useful for the singers who perform different styles of singing.



Vowel1: 0.10006-0.28994
Vowel2: 0.33006-0.39994
Vowel3: 0.52006-0.65994
Vowel4: 0.93006-1.0099
Vowel5: 1.0701-1.3499
Vowel6: 1.9701-2.8999
Vowel7: 3.0201-5.5699
Vowel8: 5.9001-6.7599
Vowel9: 6.8301-9.3799
Vowel15: 22.2401-25.8199
Vowel16: 26.6601-27.5699
Vowel17: 27.5801-31.2599
Vowel18: 31.5701-34.3999
Vowel19: 35.2401-38.5499

F1 (Hz) 595.237
F2 (Hz) 1144.74
Mean Pitch (Hz) 128.058
Mean Vibrato Rate (Hz) 4.1721
Mean Vibrato Extent (Hz) 1.43227

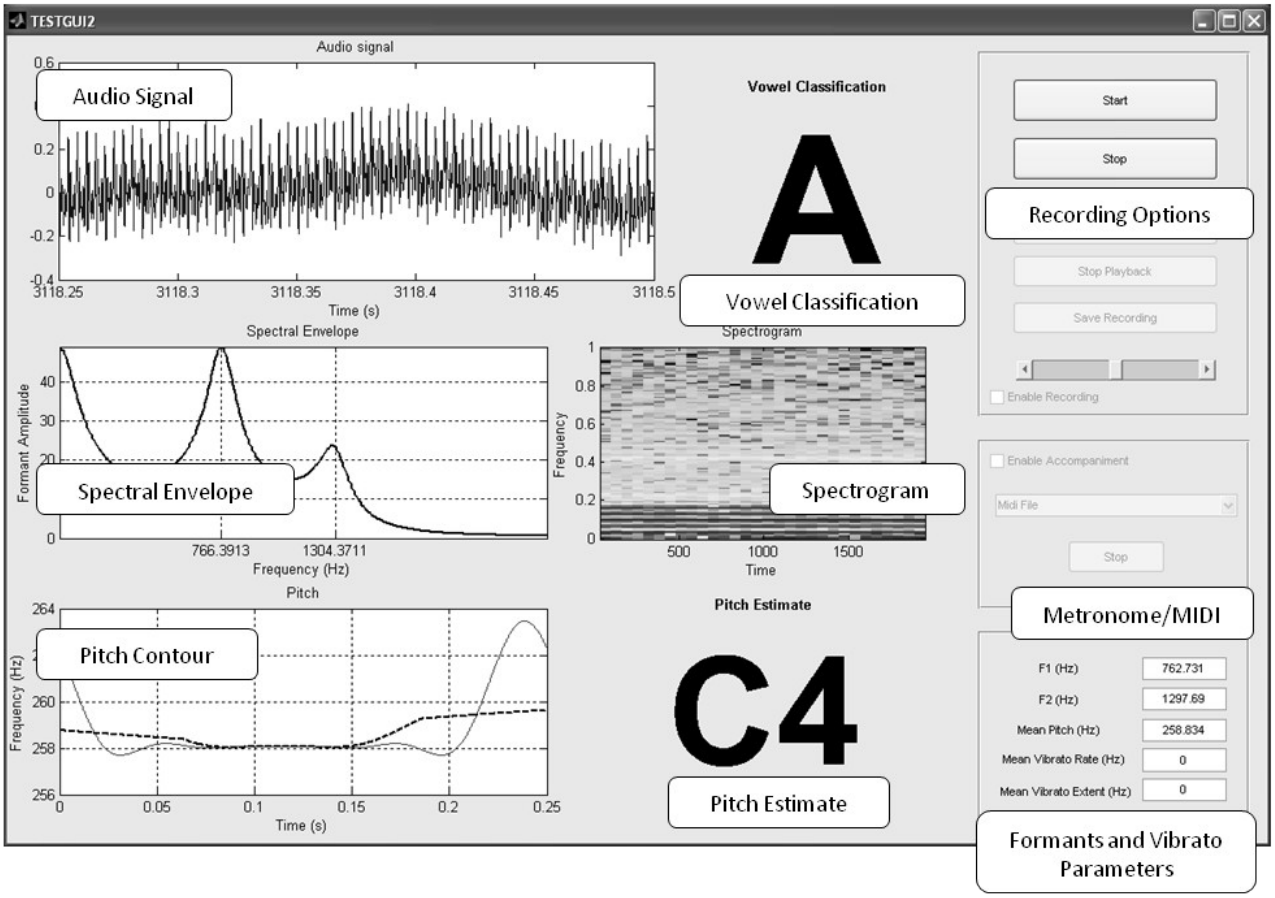
Figure 5. Graphical user interface of the offline application.

Three other singers tested the software using vocalises. The correspondence between the intended vowels of the singers and the vowel classifications made by the online application was observed before and after the singer had a chance to use the software. The results showed an average increase of 7.0% in the correspondence between the intended vowels of the singers and the vowel classifications made by the online application. Moreover, the singers gave positive feedback in the usefulness, ease of use, layout and performance of the online application.

CONCLUSION

The developed vowel classifier based on listener perception utilizes the formant frequencies, F1 and F2, and MFCCs as features. Vowel classification is made by a first-stage QDA classifier and a second-stage LDA classifier for the vowels /a/ and /o/. The classifier was optimized for speeds applicable to the offline and online applications through the reduction of the Training Set size while preserving the integrity of the data. Resulting overall accuracies for the classifier used in the offline application are 89.4% and 90.9% for the Training Set and Test Set, respectively. On the other hand, overall accuracies for the classifier used in the online application are 89.4% and 89.7% for the Training Set and Test Set, respectively. Using the Listening Test vowels as inputs to the classifiers, the overall perception accuracy for both offline and online classifiers is 92.3%.

Aside from vowel classification assessing the intelligibility of sung vowels, additional features have been incorporated in the developed software applications. The added features are spectral envelope and spectrogram displays, pitch estimation and computation of vibrato parameters; these parameters were chosen based on the suggestions of the singers who were recorded. Vowel classification and spectral envelope displays help singers assess



F1 (Hz) 762.731
F2 (Hz) 1297.69
Mean Pitch (Hz) 258.834
Mean Vibrato Rate (Hz) 0
Mean Vibrato Extent (Hz) 0

Figure 6. Graphical user interface of the online application.

their vocal color and quality which are important elements in singing and given significant consideration in vowel pedagogy. Moreover, the display of pitch, vibrato parameters and spectrogram will help singers improve their vocal tonalities and assess their adherence to the musical style that they are performing. Other features such as voice recording and accompaniment playback were included in the applications to further enhance usage.

In conclusion, the researchers have been able to develop a novel approach for assessing the intelligibility of sung vowels which performs with an accuracy exceeding 89% and effectively emulates the human auditory vowel perception. The implementation of the algorithm was based on the vowel classification made by 100 listeners. Moreover, software applications were developed based on this algorithm. Initial tests of these software applications show them to have potential use in vocal pedagogy and have been enhanced with features that prove to be beneficial to the intended users.

RECOMMENDATIONS

The inclusion of a lip-shape detector and online video feedback of the user's lips should be part of the next version of the software. It would be also interesting to study the pedagogical impact of the software on both students at the College of Music and pop singers. It is expected that an increase in the database of recorded sung vowels and song pieces, as well as listeners in the Listening Test, will lead to a higher accuracy in the vowel classifier.

ACKNOWLEDGEMENT

This research project is funded by the Office of the Vice-Chancellor for Research and Development Open Grant of the University of the Philippines Diliman and the Office of the Vice-President for

Academic Affairs Emerging Fields Grant, and is part of Interdisciplinary Signal Processing for Pinoys (ISIP) Program.

REFERENCES

D.R. Appelman. 1986. The Science of Vocal Pedagogy (Theory and Application), Indiana University Press.

J. Dimaculangan, and R. Felias. 2008.“Vowel Migration in Sung Filipino Text and Perceived Intelligibility,” Undergraduate Student Project, Department of Electrical and Electronics Engineering, University of the Philippines, Diliman.

C. Dumitru, and I. Gavat. 2007. “Vowel, Digit and Continuous Speech Recognition based on Statistical, Neural and Hybrid Modelling by using ASRS_RL”, *EUROCON, The International Conference on "Computer as a Tool*, Warsaw, Poland, 856-863.

P. Merkx, and J. Miles. 2005. Automatic Vowel Classification in Speech, Duke Project Paper in Math 196S, Duke University, Durham, NC, USA.

P. Schmid, and E. Barnard. 1997. “Explicit, N-Best Formant Features for Vowel Classsification”, *Proc. Intl. Conf. On Acoustics, Speech, and Signal Processing, Munich, Germany*, 991-994.

M. Slaney. 1998. Auditory Toolbox for Matlab Technical Report, Interval Research Technical Report.

P. Wilson, K. Lee, J. Callaghan, and C. W. Thorpe. 2008. “Learning to sing in tune: Does real-time visual feedback help?”. *Journal of Interdisciplinary Music Studies* 2(12):157-172.

J. Makhoul. 1972. “Linear prediction: A tutorial review,” in *Proceedings of the IEEE,* pp. 1973–1986.