STUDY OF NOISE ROBUSTNESS OF FIRST FORMANT BANDWIDTH (F1BW) METHOD

Shahrul Azmi M.Y[1], Fadzilah Siraj[2], S.Yaacob[3] , Paulraj M.P[4] and Ahmad Nazri[5]

[1] *Universiti Utara Malaysia, Malaysia, shahrulazmi@uum.edu.my*
[2] *Universiti Utara Malaysia, Malaysia, fad173@uum.edu.my*
[3] *Universiti Malaysia Perlis, Malaysia, s.yaacob@unimap.edu.my*
[4] *Universiti Malaysia Perlis, Malaysia, paul@unimap.edu.my*
[5] *Universiti Malaysia Perlis, Malaysia, ahmadnazri@unimap.edu.my*

ABSTRACT. The performance of speech recognition application under adverse noisy condition often becomes the topic of researchers regardless of the language used. Applications that use vowel phonemes require high degree of Standard Malay vowel recognition capability. In Malaysia, researches in vowel recognition is still lacking especially in the usage of Malay vowels, independent speaker systems, recognition robustness and algorithm speed and accuracy. This paper presents a noise robustness study on an improved vowel feature extraction method called First Formant Bandwidth (F1BW) on three classifiers of Multinomial Logistic Regression (MLR), K-Nearest Neighbors (k-NN) and Linear Discriminant Analysis (LDA). Results show that LDA performs best in overall vowel classification compared to MLR and KNN in terms of robustness capability.

Keywords: Malay Vowel, Spectrum Envelope, Speech Recognition, Noise Robustness.

INTRODUCTION

Normally, human listeners are capable of recognizing speech when input signals are corrupted by low level of noise. According to Devore & Shinn-Cunningham (2003), human listeners can select and follow another speaker's voice (Devore & Shinn-Cunningham, 2003). Even in more adverse scenarios such as at packed football stadium, listeners can select and follow the voice of another speaker as long as the signal-to-noise ratio (SNR) is not too low. In terms of speech recognizers, most of these applications are affected by adverse environmental conditions. According to (Uhl & Lieb (2001), it is important to suppress additive noise before the feature extraction stage of any speech recogniser (Uhl & Lieb, 2001). Invariance to background noise, channel conditions and variations of speaker and accent are the main issues in noise robust applications (Al-Haddad, Samad, Hussain, & Ishak, 2008; Huang, Acero, & Hon, 2001). Development of signal enhancement techniques is an effort to remove the noise prior to the recognition process but this may cause the speech spectral characteristics to be altered. This may cause the speech signal to be unsuitable to be used in the already designed acoustic models of the recognizer thus deteriorating the performance of the recognizer (Kyriakou, Bakamidis, Dologlou, & Carayannis, 2001). This justifies the efforts of developing a robust speech recognizer modeled from robust speech features.

This paper will present a robustness study on First Formant Bandwidth (F1BW) method introduced by Shahrul Azmi (2010) (Shahrul Azmi, Siraj, Yaacob, Paulraj, & Nazri, 2010) which is an improved formant method based on single framed analysis on isolated utterances.

LITERATURE REVIEW

There are many researches on the topic of vowel recognition. Features such as formant features of formant frequency, bandwidth, and intensity were used to classify accents conversions between British, Americans and Australian speakers (Yan & Vaseghi, 2003). Formant Amplitude and 2-dimensional formant Euclidean were also used for vowel classification (Carlson & Glass, 1992; Vuckovic & Stankovic, 2001). The first three formant values of F1, F2, and F3 using Praat's linear predictive coding algorithm were used to study formant characteristics of vowels produced by mandarin esophageal speakers (Liu & Ng, 2009).

According to Hillenbrand and Houde (2003), majority of vowel identification models assumed that the recognition process is driven by either the formant frequency pattern of the vowel (with or without a normalizing factor of fundamental frequency) or by the gross shape of the smoothed spectral envelope (Hillenbrand & Houde, 2003). Several other researchers have made excellent reviews of this literature. The main idea underlying formant representations is the notion that the recognition of vowel identity is controlled not by the detailed shape of the spectrum but rather by the distribution of formant frequencies, mainly the three lowest formants (F1, F2 and F3).

In terms of robustness analysis, Luo (2008) proposed a method to sharpens the power spectrum of the signal in both the frequency domain and the time domain by integrating simultaneous masking, forward masking and temporal integration effects into traditional mel-frequency cepstral coefficients (MFCC) feature extraction algorithm (Luo, Soon, & Yeo, 2008). Yeganeh (2008) proposes a set of noise-robust features based on conventional MFCC feature extraction method based on a weight parameter (Yeganeh, Ahadi, & Ziaei, 2008). Rajnoha (2007) uses white noise and car noise to study the classification robustness of MFCC and PLP features (Rajnoha & Pollak, 2007). Gajic (2006) investigated how dominant-frequency information can be used in speech feature extraction to increase the robustness of automatic speech recognition against additive background noise (Gajic & Paliwal, 2006). In Malaysia, Al-Haddad (2009), proposed an algorithm for noise cancellation by using recursive least square (RLS) and pattern recognition by using fusion method of Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) (Al-Haddad, Samad, Hussain, Ishak, & Noor, 2009). He collected Malay number speech data from 60 speakers.

METHODOLOGY

Vowel Recognition Process

Vowel Recognition process starts with the Data Acquisition process followed by filtering, pre-processing, frame selection, Auto-regressive modelling, and feature extraction process. These processes are shown in Fig.1 and their details will be explained in the rest of this paper. Data Collection process was taken from a total of 80 individuals consisting of students and staff from Universiti Malaysia Perlis (UniMAP) and Universiti Utara Malaysia (UUM). The speakers consist of individuals from both male and female genders. They are from the three main races of Malaysia which are Malay, Chinese and Indians. The details of the data collection are explained in (Shahrul Azmi et al., 2010).

Improved Vowel Feature Extraction Method

In order to train the data, two features were extracted from each recorded vowel during data collection. The first feature was extracted based on the energy of the first formant (F1) peak and denoted by $F1BW_1$. The second feature was extracted from the valley between the first (F1) and the second formant (F2) peaks and denoted by $F1BW_2$. Mean intensity of $F1BW_1$ *and* $F1BW_2$ were calculated using equation (2) where *SI* is the spectrum intensity.

$$F1BW_x(vowel) = \frac{1}{N}\sum_{f=F_{low}}^{f=F_{high}} SI(f) \quad (2)$$

Six Malay vowels were represented by a total of twelve features of *F1BW1a*, *F1BW2a*, *F1BW1e*, *F1BW2e*, *F1BW1i*, *F1BW2i*, *F1BW1o*, *F1BW2o*, *F1BW1u*, *F1BW2u*, *F1BW1ə* and *F1BW2ə*. The details of the method can be found in (Shahrul Azmi et al., 2010).

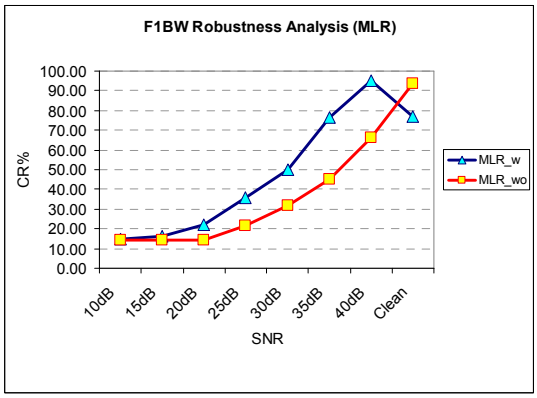
Classification Techniques Used

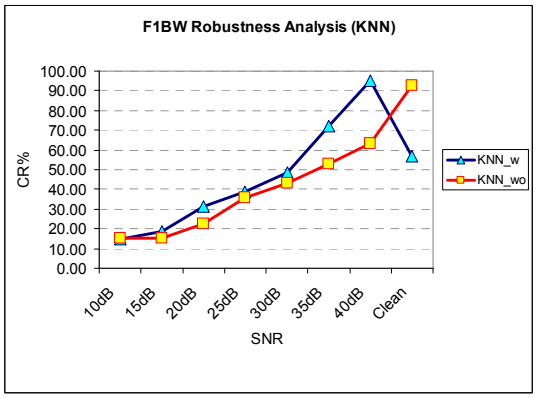
In this study, two non-linear classifiers which are K-Nearest Neighbours (KNN), Multinomial Logistic Regression (MLR) and a linear classifier which is Linear Discriminant Analysis (LDA) will be used to classify all the features in this study. These classifiers were chosen based on their popularities in speech recognition researches. All the features in this paper are classified using MATLAB built-in functions for all the four classifiers.

NOISE ROBUST ANALYSIS

A robustness analysis was done to study the robustness of the proposed features of First Formant Bandwidth and compare the results with the single frame Mel-Frequency Cepstrum Coefficients. White Gaussian noise was used to proof robustness. Seven signal-to-noise (SNR) levels of 10dB, 15dB, 20dB, 25dB, 30dB, 35dB and 40dB were used in this experiment in addition to the clean signal. These experiments were done on three of classifiers which are Multinomial Logistic Regression (MLR), K-Nearest Neighbors and Linear Discriminant Analysis (LDA). In the rest of the figures in this paper, the abbreviation "_w" means that the classifier model was trained with noise and "_wo" means that classifier model was trained without noise. The analysis was based on cross validation testing where the original data is randomized and split into 70% training set and 30% testing set (unseen input).

In Figure 1, blue line represents the overall vowel classification rate of F1BW features trained with noise and tested with different SNR level data. The red line represents the overall vowel classification rate of F1BW features trained with data from raw signal only and tested with different SNR level data. For the overall vowel classification trained with only clean, classification rate increases as SNR increases as shown by the plotted red lines in Fig. 6.1. Optimum overall vowel classification rates obtained for MLR, KNN and LDA were 93.78%, 92.50% and 90.19% respectively. For the overall vowel classification trained with noise, MLR and KNN overall vowel classification rates were better for SNR of 40dB and lower compared to the features trained with only clean data. As for LDA, for the overall vowel classification trained with noise, the optimum overall vowel classification rate were obtained at SNR of 30dB which is better compared to both MLR and KNN. For all classifiers, for the classification rate results trained with noisy data, "over trained" behavior was observed.





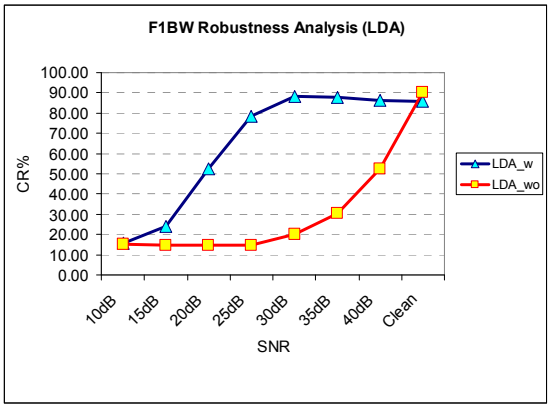


Figure 1. Overall F1BW Classification Rate by Different SNR level

In terms of classification rate trained with noisy data, LDA classifier performs the best among the three classifiers because as SNR increases, the classification rate approaches optimum faster at less than 30dB SNR which was better than MLR and KNN suggesting it to be the most noise robust. Furthermore, LDA shows less "over trained" effect when compared to KNN and MLR.

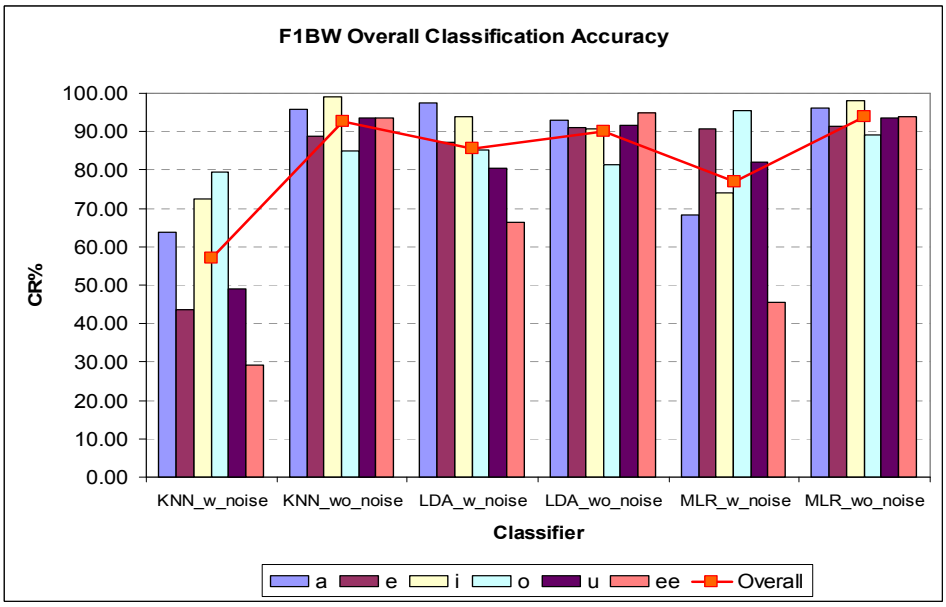


Figure 2. Overall F1BW Classification Rate of Vowels based on Classifiers and Training Conditions using Clean Training Data

Figure 2 shows the detailed overall classification result of F1BW features classified with MLR, LDA and KNN classifiers trained using only clean data. In figure 2 and table 1, the abbreviation "_w_noise" means that the clean trained classifier model was tested with noisy unseen data "_wo_noise" means that the clean trained classifier model was tested with raw unseen data. Based on overall vowel classification, MLR classifier gave the best result of 93.78% when tested with clean data with vowel /i/ giving the best classification accuracy. This is shown in Table 1.

Table 1. Overall Classification Rate of Vowels on F1BW features using Clean Training Data (Tabulated Results)

Classifiers	Testing Data	a	e	i	o	u	ə	Overall Vowel CR%
KNN	With noise	63.67	43.50	72.55	79.41	49.09	29.02	57.07
KNN	Without noise	95.87	88.85	98.98	84.84	93.48	93.66	92.50
LDA	With noise	97.50	87.30	93.92	85.21	80.32	66.30	85.65
LDA	Without noise	92.81	91.11	90.69	81.54	91.57	94.82	90.19
MLR	With noise	68.18	90.71	74.14	95.54	82.12	45.58	76.98
MLR	Without noise	96.26	91.50	97.96	89.26	93.55	94.06	93.78

MLR tested with data with noise gave only 76.98% with /o/ giving the highest classification rate. This difference in vowel recognition performance between classifier model trained with and without noise may be caused by how well the classifier model adapt to the noisy data. For the model which is trained with noisy data, LDA obtained the highest overall classification rate of 85.65% followed by MLR with 76.98% and KNN with a low classification rate of only 57.07%.

CONCLUSION

This paper presents a noise robustness study on a new improved vowel feature extraction method of First Formant Bandwidth based on formant and spectrum envelope called First Formant Bandwidth (F1BW). It was observed that LDA performs best in overall vowel classification compared to MLR and KNN in terms of robustness capability with less "over trained" effect. It also performs better compared to MLR and KNN in the robustness category especially for SNR above 20dB. The worst robust performed feature is F1BW for LDA clean trained model.

REFERENCES

Al-Haddad, S., Samad, S., Hussain, A., Ishak, K., & Noor, A. (2009). Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering. *American Journal of Applied Sciences, 6*(2), 290-295.

Carlson, R., & Glass, J. (1992). *Vowel Classification based on analysis-by-synthesis*. Paper presented at the 2nd Internaional Conference on Spoken Language Processing (ICSLP 92).

Devore, S., & Shinn-Cunningham, B. G. (2003, 6-9 July). *Perceptual consequences of including reverberation in spatial auditory displays.* 2003 International Conference on Auditory Display, Boston, MA, USA.

Gajic, B., & Paliwal, K. K. (2006). Robust speech recognition in noisy environments based on subband spectral centroid histograms. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(2), 600-608.

Hillenbrand, J., & Houde, R. (2003). A narrow band pattern-matching model of vowel perception. *The Journal of the Acoustical Society of America, 113*, 1044-1055.

Kyriakou, C., Bakamidis, S., Dologlou, I., & Carayannis, G. (2001, January 14-17). *Robust Continuous Speech Recognition in the Presence of Coloured Noise.* Proceedings of 4th European Conference on Noise Control (EURONOISE2001), Patra.

Liu, H., & Ng, M. L. (2009). Formant Characteristics of Vowels Produced by Mandarin Esophageal Speakers. *Journal of voice, 23*(2), 255-260.

Luo, X., Soon, Y., & Yeo, C. K. (2008). *An auditory model for robust speech recognition.* International Conference on Audio, Language and Image Processing, 2008. ICALIP 2008. , Shanghai.

Rajnoha, J., & Pollak, P. (2007). *Modified Feature Extraction Methods in Robust Speech Recognition.* 17th International Conference of Radioelektronika, 2007, Brno.

Shahrul Azmi, M. Y., Siraj, F., Yaacob, S., Paulraj, M. P., & Nazri, A. (2010). *Improved Malay Vowel Feature Extraction Method Based on First and Second Formants.* 2nd International Conference on Computational Intelligence, Modelling and Simulation (CIMSIM 2011), Bali, Indonesia.

Uhl, C., & Lieb, M. (2001). *Experiments with an extended adaptive SVD enhancement scheme forspeech recognition in noise.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 01), Salt Lake City, UT, USA.

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

Yan, Q., & Vaseghi, S. (2003). *Analysis, modelling and synthesis of formants of British, American and Australian accents.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003).

Yeganeh, H., Ahadi, S. M., & Ziaei, A. (2008). *A new MFCC improvement method for robust ASR.* 9th International Conference on Signal Processing (ICSP 2008) Beijing, China.