Development of Language Identification using Line Spectral Frequencies and Learning Vector Quantization Networks

Teddy Surya Gunawan[1], Mira Kartiwi[2], and Nor Hazima Ardzemi[1]
[1]*Electrical and Computer Engineering Department, Kulliyyah of Engineering, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia*
[2]*Information Systems Department, Kulliyyah of Information and Communication Technology International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia*
tsgunawan@iium.edu.my

***Abstract*—Language identification system has become a very active research nowadays due to the need of intercultural human communication. This paper proposed a Language Identification System using Line Spectral Frequencies (LSF) and Linear Vector Quantization (LVQ) network. LSF was used due to its robustness compared to normal linear predictor coefficients (LPC), while LVQ was used due to its low complexity. Three languages, i.e. Arabic, Malay, and Thai, for both native male and female speakers were recorded at IIUM Recording Studio. Several experiments have been conducted to find the optimum parameters, i.e. sampling frequency (8000 Hz), LPC order (18), number of hidden layers (300), and learning rate (0.01). Results show that our proposed system is able to recognize the trained languages with the recognition rate of 73.8%. Further research could be conducted to improve the performance using different features, classifiers, or using deep learning neural network.**

***Index Terms*—Language Identification; Learning Vector Quantization Networks; Line Spectral Frequencies.**

I. INTRODUCTION

Human communication is a process of interaction among people and it is divided into linguistic communication and non-linguistic communication. Linguistic communication requires language as a medium where in non-linguistic communication require a signs or gestures as the medium to communicate. Basically, non-linguistic communication cannot identify the identity of people because same signs are used to express same feelings regardless country. For example, smile or laugh is used to express happiness or pleasure and clench the fists to express anger. However, linguistic communication can somehow identify the speakers through the language that being used. There are about 7105 living languages owned by 6.7 billion populations in this world [1] and these languages definitely differ from each other.

Language identification (LID) system is a mechanism that use to identify the spoken language between people, either from speech audio or recording. Besides, it also widely used in several applications such as to route the incoming phone call to the human auditory operator in the equivalent language and to use it as multi language translation system [2]. Due to technologies expansions, language identification is very important since we need it for multi-language communication and can accommodate people from different nations in their own native language. In addition, it also important for society communication in business where the interaction between people are in different language.

Many researches have been conducted in the area of LID. A tutorial on LID has been presented in [2] in which syntactic, morphological, and acoustic, phonetic, phonotactic, and prosodic level information have been discussed in details. Around 87 prosodic features has been used for LID system in [3] which provides better recognition performance, while [4] utilizes visual features with error rate less than 10%. In [5], a highly accurate and computationally efficient framework of i-vector presentation is proposed for rapid language identification. A hierarchical LID framework is proposed in [6], in which a series of classification decisions is performed at multiple levels with individual languages identified only at the final level. Recent researches showed the potential of deep neural network with better recognition rate using limited training data [7-9]. Moreover, learning vector quantization (LVQ) has been identified as smart alternatives with low complexity and computational costs compared to deep learning and support vector machine as highlighted in [10].

Although many researches have been conducted on LID, but most of the researches are only identifying around two languages. Therefore, in this paper, three languages including Malay, Arabic, Thai, spoken by both male and female speakers will be recorded and analyzed. Robust line spectral frequencies features [11, 12] and LVQ network classifier will be utilized. The rest of the paper is organized as follows. Next section will discuss typical LID system followed by the proposed system. Results, discussion, and conclusion will be presented.

II. REVIEW OF LANGUAGE IDENTIFICATION SYSTEM

Figure 1 illustrates the basic block diagram of language identification system. Preprocessing is a process of speech signal refinement. The raw speech signal that we obtained is not proper to use directly as input. The weak signal that we obtained has to be amplified, removed the longer silence, and also extracted the background noise or music for further processing. There are many feature extractions that can be used for LID system, in order to extract the speech signal from each different speaker of different language, for example Mel-Frequency Cepstral Coefficients (MFCC),

Shifted Delta Cepstra (SDC), Perceptual Linear Prediction (PLP), Dynamic Time Warping (DTW), and Bark Frequency Cepstral Coefficients (BFCC). There are a few classifiers that can used, including Vector Quantization (VQ), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Ergodic Hidden Markov Model (HMM), K-Means Clustering Algorithm and Artificial Neural Network (ANN) [2].

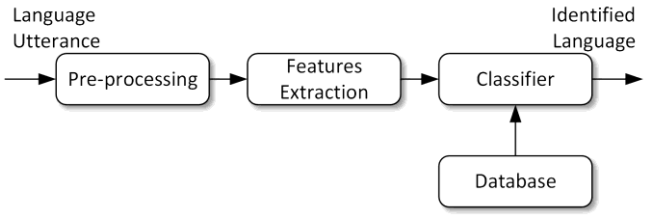


Figure 1: Typical Language Identification System

In [13], the authors have applied hybrid feature extraction technique to identify a particular language. There are two hybrid feature extraction methods are used which are BFCC and RPLP; obtained from combination of PLP and MFCC method. Moreover, there are seven different speakers with ten different of languages spoken. Thus, the characteristic of database will be 16 KHZ of sampling frequency and 16 bit resolutions. Besides, there are two classifiers were used which are Vector Quantization (VQ) for training and Dynamic Time Warping (DTW) for classification. From the overall results, BFCC performs better than MFCC and RPLP perform better than PLP. However, the results for all feature extraction methods BFCC, MFCC, RPLP with GMM and PLP shows that the performance is higher with increment in length of test utterance. In addition, RPLP is shown to be the excellent feature extraction method for LID system and BFCC is better than MFCC in experimental result.

Language identification using warping feature and shifted delta cepstrum (SDC) has been proposed in [14] which links together with the warping PLP to get the warped feature Moreover, when the feature is extracted, all the classification is done by using GMM, where it used for maximization the expectation algorithm. Thus, the result shows that MFCC and SDC will give a good performance over the LID system and it higher capability to match with GMM classifier.

Louradour kernel with GMM method running at the background has been proposed in [15]. In this research, 8Khz sampling rate is used for 5 languages and 50 speakers for each language was trained. Besides, Mel Frequency Cepstral Coefficient (MFCC) is used for feature extraction method. In addition, the performance is better when the mixture of GMM is increasing along with Louradour kernel used in SVM. So, it produces a good result than basic GMM and GLDS technique.

Table 1 shows the summary of various LID system along with its method, advantages, and disadvantages. The acoustic front-end, audio features, and classifier could be selected. The main function of acoustic front-end is to capture the essential difference between languages by modeling the distribution spectral directly. Moreover, it is also important to compact and make the speech waveform efficient by including all the aspect of the speech characteristic without multiple information. There are four steps of front-end system which are preprocessing, feature parameterization, append derived temporal information and robustness against noise [2].

Table 1
Summary of Various LID Systems

Method	Advantages	Disadvantages
Used BFCC and RPLP methods in hybrid feature extraction for LID system. Thus, it was obtained from MFCC and PLP [13]	Give better identification performance for noisy environment as well as clean environment.	Sometime have noise robustness.
Used warping and Shifted Delta Cepstrum (SDC) method for LID system [14]	Warping feature method shows a better result over the system using MFCC and SDC	No improvement to the LID accuracy if warping to the SDC value.
Used Support Vector Machine (SVM) with Louradour kernel [15]	SVM with Louradour kernel give a better performance than GMM	Do not support variable length of input.

III. Proposed Language Identification System

Figure 2 shows our proposed language identification system. For feature extraction, LSF was selected due to its robustness compared to LPC, while LVQ network was selected as classifier. Basically, there are two main parts as shown in the Figure, i.e. feature extraction and classifier. LSF and LVQ will be explained in more details.

Pre-emphasis is the basic operation needed in analyzing speech to remove the unwanted noise in the speech and to emphasize the higher frequencies. Pre-emphasize is implemented as a fixed-coefficient filter or as an adaptive one, where the coefficient *a* is adjusted with time according to the autocorrelation values of the speech. The pre-emphasize has the effect of spectral flattening which renders the signal less susceptible to finite precision effects (such as overflow and underflow)in any subsequent processing of the signal. The selected value for *a* in our work was 0.97. Eq. (1) shows the first-order FIR pre-emphasis filter used.

$$H(z) = 1 - az^{-1}, \quad 0.9 \le a < 1.0 \tag{1}$$

Since the vocal tract moves mechanically slowly, speech can be assumed to be a random process with slowly varying properties. Hence, the speech is divided into overlapping frames of 50ms every 20ms. The speech signal is assumed to be stationary over each frame and this property will prove useful in the following steps. Each frame has been windowed to increase the correlation of the linear predictive coding (LPC) spectral estimates between consecutive frames in order to minimize the discontinuity of the signal at the beginning and end of each frame. A typical window is the Hamming window as shown in Eq. (2).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \tag{2}$$

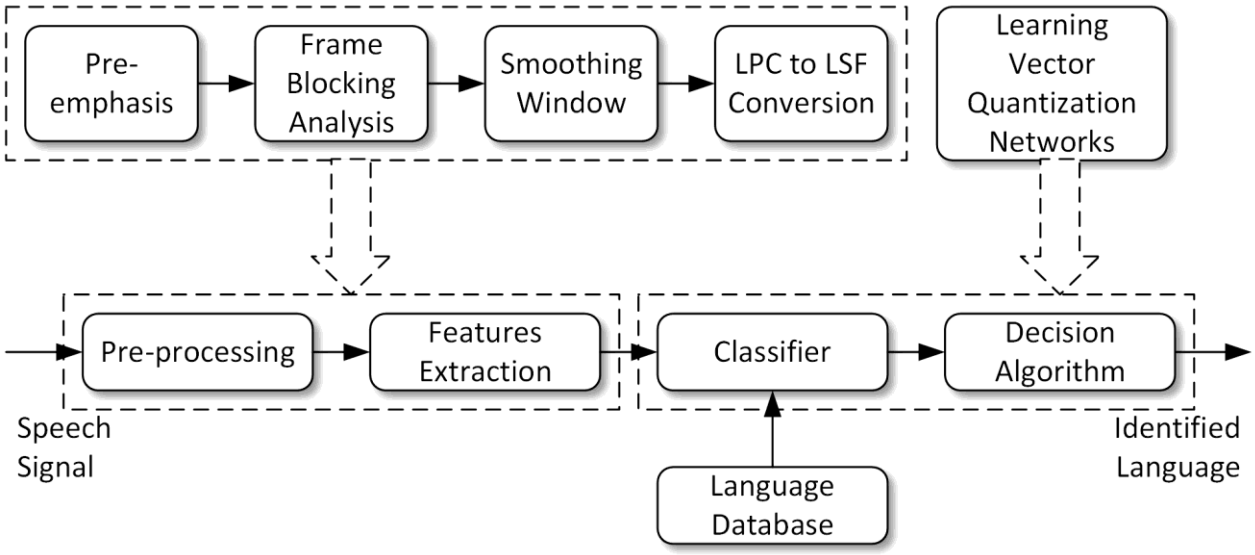


Figure 2: Proposed Language Identification System

A. Line Spectral Frequencies (LSF) Feature Extraction

A widely used source-filter model of speech is the linear prediction coefficient (LPC) model. LPC models are used for speech coding, recognition and enhancement. A LPC model with order *p* can be expressed as shown in Eq. (3).

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + e(n) \tag{3}$$

where $x(n)$ is speech signal, a_k is the LP parameters and $e(n)$ is speech excitation. Note that the coefficients a_k model the correlation of each sample with the previous p samples whereas $e(n)$ models the part of speech that cannot be predicted from the past p samples.

The line spectral frequencies (LSF) is an alternative representation of linear prediction parameters. LSFs are used in speech coding, and in the interpolation and extrapolations of LP model parameters, for their good interpolation and quantization properties. LSFs are derived as the roots of the following two polynomials as shown in Eq. (4) and (5).

$$\begin{aligned} P(z) &= A(z) + z^{-(P+1)} A(z-1) \\ P(z) &= 1 - (a_1 - a_p) z^{-1} - (a_2 - a_{p-1}) z^{-2} - \ldots \\ &\quad - (a_p - a_1) z^{-P} + z^{-(P+1)} \end{aligned} \tag{4}$$

$$\begin{aligned} Q(z) &= A(z) + z^{-(P+1)} A(z-1) \\ Q(z) &= 1 - (a_1 - a_p) z^{-1} - (a_2 - a_{p-1}) z^{-2} - \ldots \\ &\quad - (a_p - a_1) z^{-P} + z^{-(P+1)} \end{aligned} \tag{5}$$

where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-P}$ is the inverse liner predictor filter and $A(z) = \frac{1}{2}[P(z) + Q(z)]$. The polynomial equations (Eq. (4) and (5)) can be rewritten in the factorized form as shown in Eq. (6) and (7).

$$P(z) = \prod_{i=1,3,5,\ldots} \left(1 - 2\cos\omega_i z^{-1} + z^{-2}\right) \tag{6}$$

$$Q(z) = \prod_{i=2,4,6,\ldots} \left(1 - 2\cos\omega_i z^{-1} + z^{-2}\right) \tag{7}$$

where ω_i are the LSF parameters. It can be shown that all the roots of the two polynomials have a magnitude of one and they are located on the unit circle and alternate each other. Hence, in LSF representation, the liner predictor coefficients $[a_1, a_2, \ldots, a_p]$ is converted to LSF vector $[\omega_1, \omega_2, \ldots, \omega_p]$. Matlab implementation function `lpc()` and `poly2lsf()` were used for this purpose.

B. Linear Vector Quantization (LVQ) Networks

Learning vector quantization (LVQ) has been very popular due to its intuitive prototype based learning algorithms with application ranging from telecommunications to robotics [16]. Basic algorithms as proposed by Kohonen is based on Hebbian learning. Assume data $\xi_i \in \Re^n$ with $i = 1, \ldots, N$ are labeled y_i where labels stem from a finite number of different classes. A generalized LVQ network is characterized by m prototypes $w_j \in \Re^n$ with priory fixed labels $c(w_j)$. Classification takes place by a winner takes all scheme as shown in Eq. (8).

$$\xi \mapsto c(w_j) \text{ where } d(\xi, w_j) \text{ is minimum} \tag{8}$$

with squared Euclidian distance $d(\xi, w_j) = \|\xi - w_j\|^2$, breaking ties arbitrarily.

For training purposes, it is usually assumed that the number and classes of prototypes are fixed. Training aims at finding positions of the prototypes such that the classification accuracy of the training set is optimized. The cost function used is shown in Eq. (9).

$$\sum_i \frac{d(\xi_i, w^+) - d(\xi_i, w^-)}{d(\xi_i, w^+) + d(\xi_i, w^-)} \tag{9}$$

where w^+ is the closest prototype with the same label as ξ_i and w^- is the closest prototype with different label than ξ_i. Training takes place by a simple stochastic gradient descent. Given a data point ξ_i, adaptation takes place as shown in Eq. (10) and (11). We can characterize LVQ as classifier which classification rule is based on a number of quantities

as shown in Eq. (12).

$$\Delta w^{+} \sim -\frac{2d\left(\xi_i, w^{-}\right)}{\left(d\left(\xi_i, w^{+}\right)+d\left(\xi_i, w^{-}\right)\right)^2} \cdot \frac{\partial d\left(\xi_i, w^{+}\right)}{\partial w^{+}} \quad (10)$$

$$\Delta w^{-} \sim -\frac{2d\left(\xi_i, w^{+}\right)}{\left(d\left(\xi_i, w^{+}\right)+d\left(\xi_i, w^{-}\right)\right)^2} \cdot \frac{\partial d\left(\xi_i, w^{-}\right)}{\partial w^{-}} \quad (11)$$

$$D(\xi, w) := \left(d\left(\xi_i, w_j\right)\right), \; i = 1, \ldots, N; \; j = 1, \ldots, m \quad (12)$$

Training aims at an optimization of a cost function of the form $f(D(\xi, w))$ by means of the gradients as shown in Eq. (13).

$$\frac{\partial f\left(D(\xi, w)\right)}{\partial w_j} = \sum_{i=1}^{m} \frac{\partial f\left(D(\xi, w)\right)}{\partial d\left(\xi_i, w_j\right)} \cdot \frac{\partial d\left(\xi_i, w_j\right)}{\partial w_j} \quad (13)$$

with respect to the prototypes w_j or the corresponding stochastic gradients for one point ξ_i. Matlab implementation function `lvqnet()`, `configure()`, and `train()` were used for this purpose. There are two parameters on `lvqnet()` function which can be optimized, i.e. number of hidden layer (N_{hidden}) and the learning rate (L_{rate}).

IV. Results and Discussion

This section will discuss the language database preparation, experimental setup, various experiments to find optimum parameters, and the performance evaluation of the proposed LID system.

A. Experimental Setup and Language Database

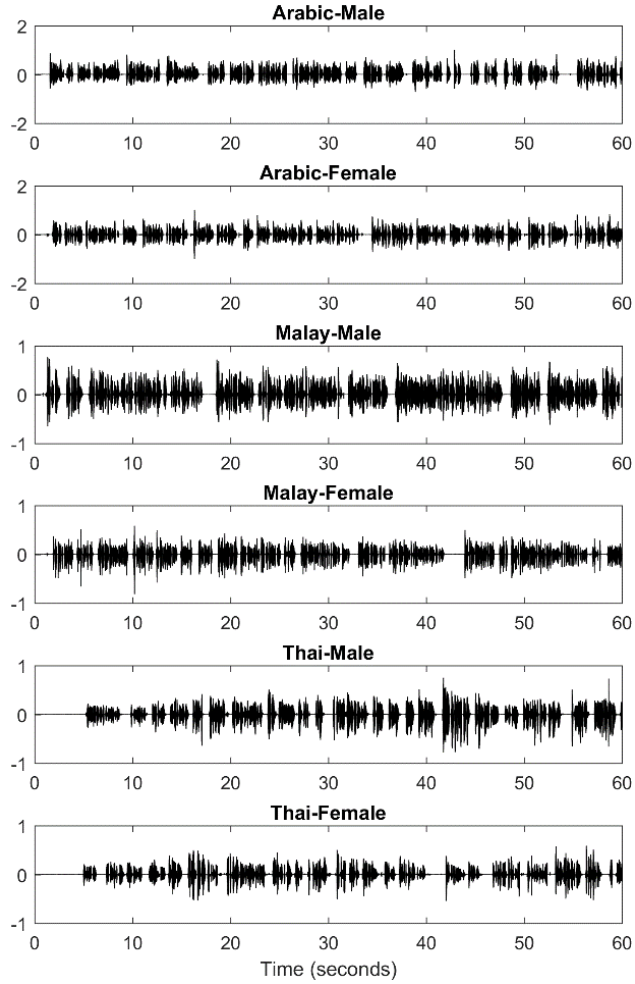
A high performance system was used for processing, i.e. a multicore system with Intel Core i7 6700 K 4.00 GHz (4 cores with 8 threads), 32 GBytes RAM, and 2 TBytes hard disk, installed with Windows 10 operating system and Matlab 2017a with Signal Processing and Neural Network Toolboxes.

The languages have been recorded in the IIUM Recording Studio in CELPAD-IIUM which has soundproof room for recording. The equipment used for the recording are Sony Sound Forge Version 9 Software, high quality microphone and Soundcraft Professional Audio Mixer.

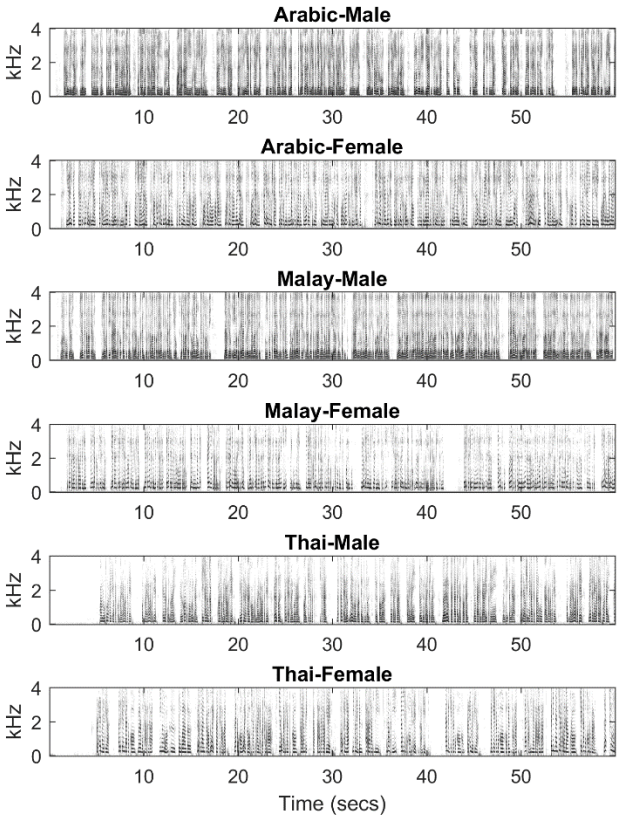
There are three languages to be recorded, i.e. Arabic, Malay, and Thai, and it will be spoken by both male and female speaker. Each language has been recorded for 30 minutes at sampling rate of 44100 Hz stereo signals. Six native speakers for respective language have participated. The speaker is reading popular book in their own language. Table 2 shows the list of recorded audio files with the actual length. Figure 3(a) shows the time domain signals and Figure 3(b) shows the spectrogram of each audio files for the first 60 seconds of the recorded audio files downsampled to 8000 Hz.

Table 2
Recorded Language Audio Files

Speech Signals	Duration (second)
Arabic-Male.wav	1818 (30:18)
Arabic-Female.wav	1940 (32:20)
Malay-Male.wav	1834 (30:34)
Malay-Female.wav	2004 (33:24)
Thai-Male.wav	1574 (26:14)
Thai-Female.wav	1600 (26:40)



(a) Time Domain Audio Signals



(b) Spectrogram of Audio Signals

Figure 3: Recorded Language Audio Files Shown In The Time And Frequency Domains

B. Experiments on Optimum Parameters

There are many parameters could be optimized in order to achieve the highest performance, in terms of recognition rate. Several important parameters will be analyzed, including *Fs* (sampling frequency), *p* (LPC order), N_{hidden} (number of hidden layer), and L_{rate} (learning rate). All experiments will be conducted by fixing other three variables while changing one parameter to be optimized. For this purpose, only the first 60 seconds of the audio files will be used to avoid overfitting.

Various research have reported that sampling frequency has an effect on the recognition rate. Therefore, the first experiment will vary the sampling frequency, i.e. 8000 Hz and 16000 Hz. The signal is obtained by downsampled the original audio signal which was sampled at 44100 Hz. For this experiment, the other three parameters were fixed as follows, p = 12, N_{hidden} = 6, and L_{rate} = 0.01. The number of training epochs is also fixed at 100.

Table 3
Experimental Result on Varying Sampling Frequency

Fs	Training Time (second)	Recognition Rate (%)
8000	832	62.43
16000	1672	62.95

Table 3 shows the recognition rate versus training time for two sampling frequency, i.e. 8000 and 16000 Hz. The training time for 16000 Hz sampling frequency was twice that of 8000 Hz with negligible recognition rate difference. In this case, sampling frequency of 8000 Hz will be selected due to its lower computational requirement.

The LPC order (subsequently LSF order), *p*, has significant effect on the quality of the predicted signal. The higher the order, the better the prediction. The second experiment will vary LPC order *p* from 8 to 20 with the interval of 2. For this experiment, the other three parameters were fixed as follows, *Fs* = 8000 Hz, N_{hidden} = 6, and L_{rate} = 0.01. The number of training epochs is also fixed at 100.

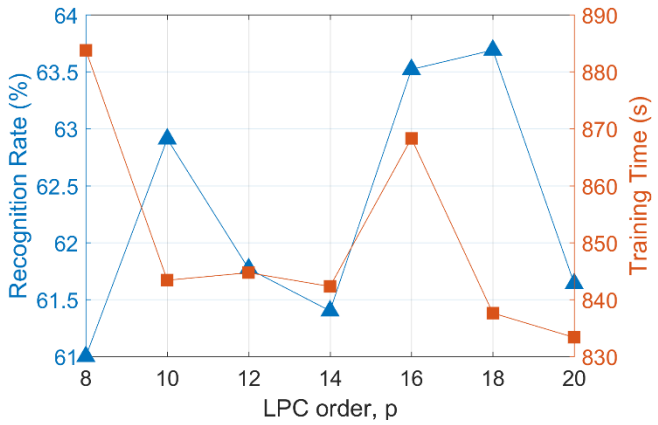


Figure 4: Variation of LPC order, *p*

Figure 4 shows the recognition rate versus training time for various LPC order *p*. It can been seen that the training time is almost the same. As we focus on the higher recognition rate, the order of *p*=18 was selected as the optimum LPC order.

Using any neural network, the number of hidden layer is one of the most important parameter in order to obtain higher performance. Normally, the higher the number of hidden layer, the better the recognition rate. The third experiment will vary the number of hidden layer from 3 to 24 with the interval of 3. Later on, it was further varied from 27 to 60 with interval of 3, and from 60 to 100 with interval of 10. Finally, it was further extended by including 150, 200, 300, 400, and 500 number of hidden layers. For this experiment, the other three parameters were fixed as follows, *Fs* = 8000 Hz, *p* = 18, and L_{rate} = 0.01. The number of training epochs is also fixed at 100.

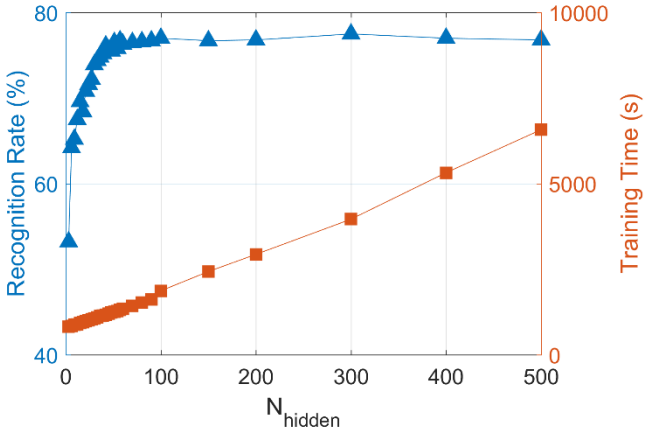


Figure 5: Variation of Number of Hidden Layer, N_{hidden}

Figure 5 shows the recognition rate versus training time for various number of hidden layer. It can been seen that after N_{hidden}=100 the recognition rate is almost the same. As we focus on the higher recognition rate, the N_{hidden}=300 which has recognition rate of 77.5% was selected as the optimum number of hidden layer. If the training time is a constraint, then we could select N_{hidden} = 42 as it has comparable recognition rate of 76.1%.

The learning rate is another important parameter in the training of neural network. It configures the neural network on how fast the change of weights and biases. The fourth experiment will vary the L_{rate} to 0.01, 0.05, 0.01, 0.05, 0.1, and 0.5. For this experiment, the other three parameters were fixed as follows, *Fs* = 8000 Hz, *p* = 18, and N_{hidden} = 42 for faster training time. The number of training epochs is also fixed at 100.

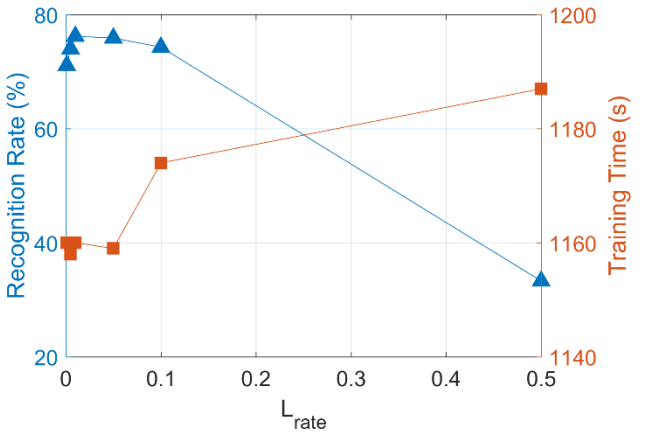


Figure 6: Variation of Learning Rate, L_{rate}

Figure 6 shows the recognition rate versus training time for various number of learning rate. It can been seen that after N_{hidden}=100 the recognition rate is almost the same. As we focus on the higher recognition rate, the learning rate L_{rate}=0.01 which has recognition rate of 76.2% was selected

as the optimum learning rate. This experiment has also shown that the training time is comparable for various learning rate value.

C. Performance Evaluation

From previous section, the optimum parameters obtained are Fs = 8000 Hz, p = 18, N_{hidden} = 300, and L_{rate} = 0.01. The final LVQ network, the training time, and its training performance is shown in Figure 7. For comparison, the N_{hidden} = 42 was also simulated. It was found that the training time was 3851 seconds versus 18264 seconds (almost 4.7 times difference) for N_{hidden} of 42 and 300, respectively. While the recognition rate was 69.3% versus 73.8% (around 4.5% difference) for N_{hidden} of 42 and 300, respectively. In terms of memory usage, it uses 4 Gb versus 5 Gb memory when it was monitored using Task Manager in Windows 10 for N_{hidden} of 42 and 300, respectively. As the current simulation system has memory of 32 Gb, we could tradeoff between the training time and the recognition rate performance.

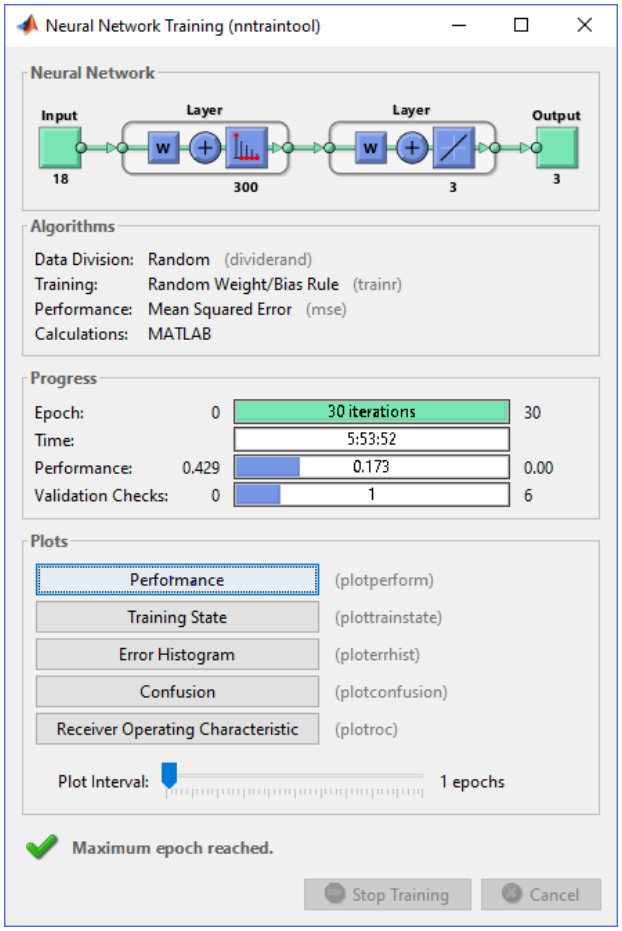


Figure 7: LVQ Neural Network Configuration And Its Training Performance

All the speech signals listed in Table 2 is processed in a frame-by-frame processing with a Frame Size of 256 samples (31.25 ms). The total feature vector for training was 18x336659 representing around 3 hours speech signals. The training data was further divided into 60% for training, 20% for validation, and 20% for testing with random selection. Figure 8 and Figure 9 shows the confusion matrix and ROC for training, validation, testing, and overall.

Further research could be conducted to improve the language identification system. More languages and more audio samples could be trained to improve the recognition rate or identify more languages. As mentioned in [17], the voiced speech tends to be more dominant compared to the unvoiced speech (in the case of English and Arabic languages). It is predicted that the voiced sound carry more significant features compared to the unvoiced sound. Therefore, we could add Voiced Activity Detector (VAD) to separate speech signals into voiced, unvoiced, and silence, and use that information to train only the voiced speech, or voiced and unvoiced signals, as silence carries no information and should be removed from the training phase. Finally, deep learning neural network, although requires longer training time and bigger data, could be used for feature extraction and classifier.

Figure 8: Confusion Matrix For Training, Validation, Testing, And Overall

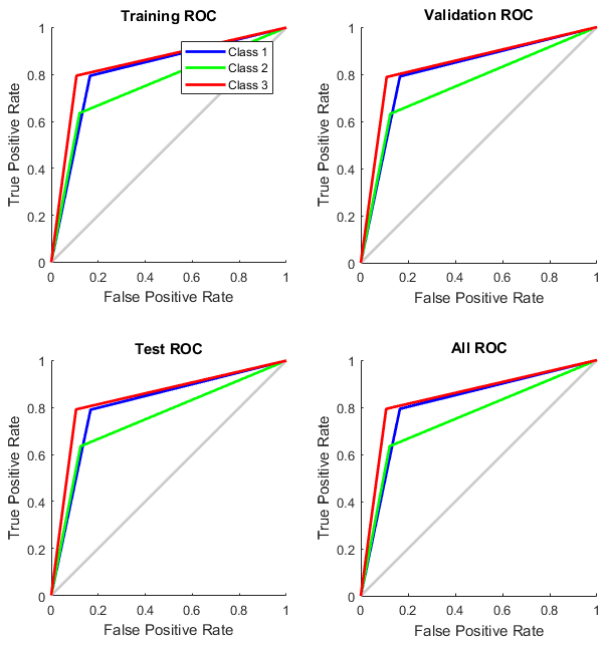


Figure 9: ROC for Training, Validation, Testing, And Overall

V. CONCLUSION

The language identification system has been presented. Line Spectral Frequencies (LSF) features has been used as the input for Learning Vector Quantization (LVQ) neural network. Systematic experiments have been conducted to find the optimum parameters, i.e. sampling frequency, linear predictor order, number of hidden layer, and learning rate. The optimum parameters obtained were Fs = 8000, p = 18, N_{hidden} = 300, and L_{rate} = 0.01. The best performance obtained was 73.8% recognition rate. Results showed that our proposed LID system which combine LSF and LVQ could be used to identify languages. Further research includes using deep neural network for feature extraction and classifier, or use different audio features and different classifiers.

ACKNOWLEDGMENT

We would like to thank Azwan from IIUM Recording Studio for his help in recording, Alaa (Arabic Female Voice), Haleemah (Thai Female Voice), Shah Abdul Hafiz (Malay Male Voice), and Adha (Thai Male Voice). This research has been supported by Ministry of Higher Education Malaysia Research Fund, FRGS15-194-0435.

REFERENCES

[1] M. P. Lewis, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the world*, vol. 16, SIL international Dallas, TX, 2009.

[2] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, pp. 82-108, 2011.

[3] R. W. Ng, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Spoken Language Recognition With Prosodic Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1841-1853, 2013.

[4] J. L. Newman and S. J. Cox, "Language identification using visual features," *IEEE Transactions on audio, speech, and language processing*, vol. 20, pp. 1936-1947, 2012.

[5] M. Van Segbroeck, R. Travadi, and S. S. Narayanan, "Rapid language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1118-1129, 2015.

[6] S. Irtza, V. Sethu, H. Bavattichalil, E. Ambikairajah, and H. Li, "A hierarchical framework for language identification," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 5820-5824, 2016.

[7] Y. Song, R. Cui, X. Hong, I. Mcloughlin, J. Shi, and L. Dai, "Improved language identification using deep bottleneck network," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4200-4204, 2015.

[8] S. Ranjan, C. Yu, C. Zhang, F. Kelly, and J. H. Hansen, "Language recognition using deep neural networks with very limited training data," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 5830-5834, 2016.

[9] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Study of senone-based deep neural network approaches for spoken language recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, pp. 105-116, 2016.

[10] T. Villmann, A. Bohnsack, and M. Kaden, "Can Learning Vector Quantization be an Alternative to SVM and Deep Learning? - Recent Trends and Advanced Variants of Learning Vector Quantization for Classification Learning," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, pp. 65-81, 2017.

[11] I. V. McLoughlin and S. Thambipillai, "LSP parameter interpretation for speech classification," in Electronics, Circuits and Systems, 1999. Proceedings of ICECS'99. The 6th IEEE International Conference on, pp. 419-422, 1999.

[12] J. J. Parry, I. S. Burnett, and J. F. Chicharo, "Linguistic mapping in LSF space for low-bit rate coding," in Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, pp. 653-656, 1999.

[13] P. Kumar, A. Biswas, A. Mishra, and M. Chandra, "Spoken language identification using hybrid feature extraction methods," *Journal of Telecommunication*, vol. 1, pp. 11-15, 2010.

[14] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in Multimedia Signal Processing, 2005 IEEE 7th Workshop on, pp. 1-4, 2005.

[15] W. Zhang, B. Li, D. Qu, and B. Wang, "Automatic language identification using support vector machines," in Signal Processing, 2006 8th International Conference on, pp., 2006.

[16] T. Kohonen, *Self-Organizing Maps, 3rd Edition*, Springer, 2000.

[17] T. S. Gunawan and M. Kartiwi, "On the Characteristics of Various Quranic Recitation for Lossless Audio Coding Application," in Computer and Communication Engineering (ICCCE), 2016 International Conference on, pp. 121-125, 2016.