International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4
ISSN: 2088-5334

Novel Statistical Clustering Method for Accurate Characterization of Word Pronunciation

Abdul Rahim Bahari[#], Aminatuzzaharah Musa[&], Mohd Zaki Nuawi[%], Zairi Ismael Rizman[*], Suziana Mat Saad[+]

[#]*Faculty of Mechanical Engineering, Universiti Teknologi MARA, Bukit Besi, Dungun, Terengganu, Malaysia*
E-mail: abdulrahimbahari@tganu.uitm.edu.my, haizuan@tganu.uitm.edu.my

[&]*Faculty of Chemical Engineering, Universiti Teknologi MARA, Bukit Besi, Dungun, Terengganu, Malaysia*
E-mail: aminatuzzaharah@yahoo.com

[%]*Department of Mechanical and Materials Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia*
E-mail: mzn@ukm.edu.my

[*]*Faculty of Electrical Engineering, Universiti Teknologi MARA, Dungun, Terengganu, Malaysia*
E-mail: zairi576@tganu.uitm.edu.my

[+]*Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia*
E-mail: suziana@ukm.edu.my

***Abstract*— This paper discusses the development method to determine the accuracy of pronunciation of the word using global statistical signal analysis parameters. An engineering word that has been chosen is 'leaching'. The pronunciation of the word 'leaching' in the French language has been recorded from 1 native speaker and 4 students. The recording processes use a microphone-laptop system configuration and the signal analyzing processes use MATLAB software. Time and frequency domain plots show a variety of waveforms according to the recorded pronunciation. For data processing, statistical signal analysis parameters involved to extract the signal's features are kurtosis, root mean square and skewness. The mapping process has been performed to cluster each data. The position of the samples from the students is referred to the samples from the native speaker. The result of the accuracy of the pronunciation of words for each student can be evaluated through the comparison of the position of all the samples. In conclusion, the development of mapping and clustering methods are able to characterize the accuracy of the pronunciation of words.**

***Keywords*— speech recognition; kurtosis; clustering; skewness; voice signal**

I. INTRODUCTION

Speech is one of the important things in human such as for communication and learning. For children, they learn about their world by listening and talking to other children and adults. Children with good speech and communication skill tend easily to learn in school and to develop a friendship. The speech is important in a presentation in which it increases confidence and competence in public presentations and helping people to make a decision. Besides, this system used French as a second language. So, as one of the developed country, the second language is very important, especially in communication. Nowadays, learning a second language helps people to communicate across cultures and conduct business to the customers using most comfortable and understanding communicating. Then, second language listening is one of the learning processes to expert in this language. Proficient listening comprehension enables learners to understand the spoken words of the language, which in turn helps the development of other language skills [1].

Hence, this system relates to the speech recognition for the purpose of identifying and converting to a machine-readable format. Next, speech recognition software has a limited vocabulary of words and may only identify these if they are spoken very clearly. Speech recognition is becoming effective that yielding a result of over 90% and above word level accuracy for vocabulary speech

recognition tasks. However, the accuracy of speech recognition may drop to less than 85% with the effect of noise [2]. Nowadays, this system is very popular in technology where it is applied in medicine, telephone network [3], security devices, ATM machines and computers [4]. Moreover, this system is a user-friendly system where it is applied in a handheld device. Since a handheld device can be used anywhere, noise environments should be considered [5]. In addition, the background noise such as air conditioning system, opening and closing doors, fans, footsteps and background conversation give an issue to this system [6]. Thus, the solution is prepared to minimize the effects of background noise in which using a head-mounted close-speaking microphone to record the voice.

The methods for the accuracy of the pronunciation of a word can be divided into a lot of categories such as Mel-Frequency Cepstrum Coefficients (MFCC), Hidden Markov Model (HMM's) and statistical signal analysis. HMM is a system for noise robust of audio-visual speech recognition. It is used for getting noise-robust audio features. This is because an audio-visual speech recognition system is one of the solutions for reliable speech recognition, especially when the audio is corrupted by noise. Thus, audio feature extraction is achieved by training the network to obtain the clean audio which is without noise such as MFFCs. Then, it will be processed with a conventional HMM with a Gaussian mixture observation model (GMM-HMM) to conduct an isolated word recognition task [7]. Not only that, the result can be increased by using MFFC under 10 dB signal to noise ratio. For MFCC, it has been proven to be effective where the results produce good recognition performance without a filter bank in which filter bank analysis is more desirable than Linear Predictive Coding (LPC) analysis [8], [9]. Besides, it also does not use many parameters. Among the methods, MFCC is the most widely used in improving recognition accuracy. However, this method requires a lot of calculations in which will increase the cost and reduce the performance of the hardware speech recognizer. Thus, this experiment focuses on statistical methods for speech and language processing. The knowledge of a speech signal and the language that it expresses is developed through a mathematical and statistical formalism [10].

Not only that, this system also related to the text-to-speech synthesis system where this system has been widely studied for many languages. In [11] had described a text-to-speech synthesis system for statistical parametric synthesis based on hidden Markov models for the Arabic language, since this language has not sufficient progress and it is in the first stage. The experimental results show that the diacritization system can generate a diacritized text with high accuracy in which is composed by two sub-systems. First, for a diacritization system that is designed to restore the missing diacritic mark and second is for a speech synthesis system that transforms the text into speech waveform. Because a speaker is located quite close to microphones in automatic speech recognition system (ASR), an efficient online target-speech-extraction method is used as a pre-processing step for robust ASR. In this case, the required weights for extracting speech and an estimated noise are then determined using an adaptation rule derived from a modified independent component analysis (ICA) cost function. The experimental results show that this method is effective to the system [12].

Previous research has been performed to develop the accuracy of the pronunciation of the word using other methods such as MFCC, HMM's, Gaussian mixture model (GMM) and Parallel model combination (PMC) [13]. According to [4], it introduces a new algorithm for extracting MFCC for speech recognition. It stated that the new algorithm has a recognition accuracy of 92.93% compared to the conventional MFCC extraction algorithm, which has an accuracy of 94.43%. In [7] is focused on a study of connectionist-HMM system for noise robust audio-visual speech recognition. Hence, this study introduces 2 steps in improving the system. First, a deep denoising autoencoder is used for obtaining a noise-robust audio feature. Next is the process of extracting visual features from images of raw mouth area that is known as convolutional neural network (CNN). Not only that, due to previous research, a synthetic speech generated from HMM-based speech synthesis systems is found to be understandable as natural human speech, but it must be in a noiseless condition. It is also to be good in terms of both naturalness and speaker similarity [14].

This research was therefore conducted to develop an alternative speech accuracy recognition system. The statistical signal analysis method was employed. Next, the statistical methods have used in processing real-world signals. On the hand, a measured signal for statistical signal analysis data also consists of variations in amplitude, frequency, phase, and energy.

This paper aims to develop an alternative method to obtain the accuracy of pronunciation. Voice recording experiments were performed to obtain data for analysis and justification.

II. Material and Method

A. Word Pronunciation

The French language is used as a material to start the experiment. The selected language is based on the engineering course, as engineering is famous in France and English is selected as it is a secondary language in Malaysia. The word chosen is 'leaching' as this word is widely used in the chemical engineering discipline and education.

The measurement set up for the present experimental recording work is shown schematically in Fig. 1. It consists of a microphone, a data acquisition system and a laptop. The experiment consists of 1 France native speaker and 4 students as a first learner which is selected by random. The recording process is performed in the low noise level room to avoid any disturbances/distraction. Each speaker pronounced for about ten times and the voice signal was captured periodically.

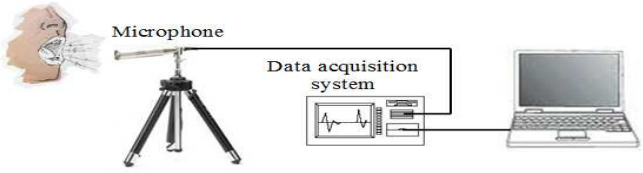

Fig. 1 The components of the measurement system

B. Analyze Signal

The recorded voice signals were filtered using MATLAB to obtain low-noise signal. Voice signal data processes involved the use of application software. The process of filtering used the first time domain graph to get the voice signal without distraction.

Analyzing signal processes of the voice which is without distraction involved the use of MATLAB software [16]. During analyzed signal processes, there are time and frequency domain. The time and frequency domain plots show a variety of waveforms according to the recorded pronunciation. For data processing, statistical signal analysis parameters involved to extract the signal's features are kurtosis, root mean square and skewness.

C. Mapping and Clustering The Data

The mapping process has been performed to cluster each data. The position of the samples from the students is referred to the samples from the native speaker. The result of the accuracy of the pronunciation of words for each student can be evaluated through the comparison of the position of all the samples with the position of the native speaker.

III. Results and Discussion

A. Time and Frequency Domain

Speech signal of the French word 'leaching' is presented in time domain is represented in Fig. 1(a) to Fig. 1(e). These signals are pronounced by 1 French native speaker (as a reference) and 4 beginner students which are selected randomly. It can be seen that the waveform is various depend on the how the word is pronounced by the speaker.

Another commonly used method to describe a characteristic of the speech signal is the frequency domain [15]. This plot shows a frequency-amplitude presentation of a signal. The frequency domain waveform of Fig. 1(a) to Fig. 1(e) is presented in Fig. 2(a) to Fig. 2(e). It is observed that the frequency distribution of the speech for all the speakers within the range of 50 Hz to 1000 Hz. Higher amplitudes were detected in the frequency of 200 Hz to 250 Hz and 450 Hz to 550 Hz.

Fig. 1(a) Time-domain presentation of native speaker

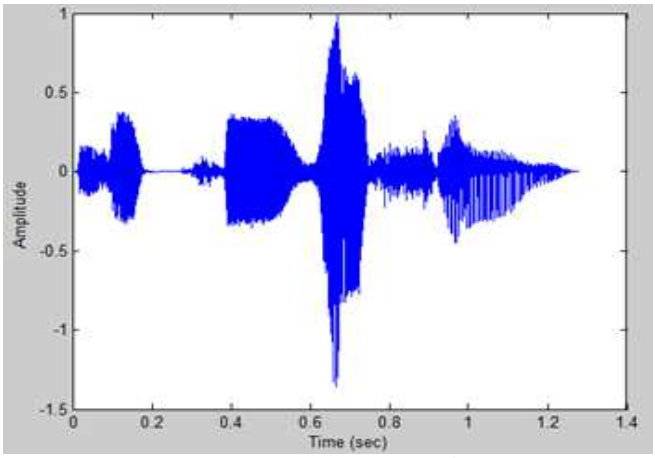

Fig. 1(b) Time-domain presentation of student 1

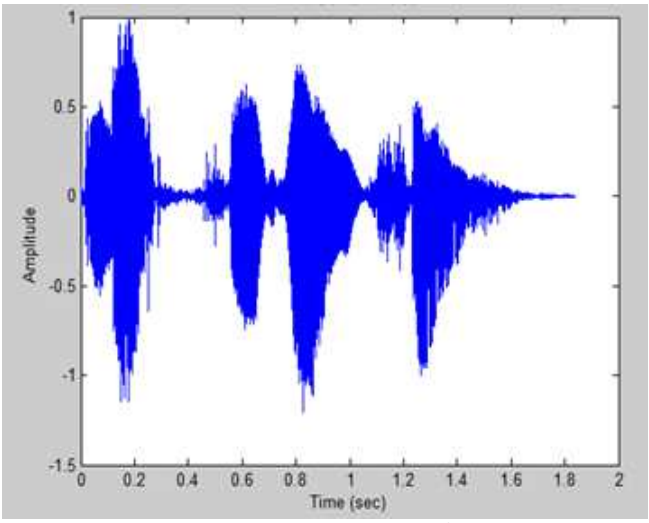

Fig. 1(c) Time-domain presentation of student 2

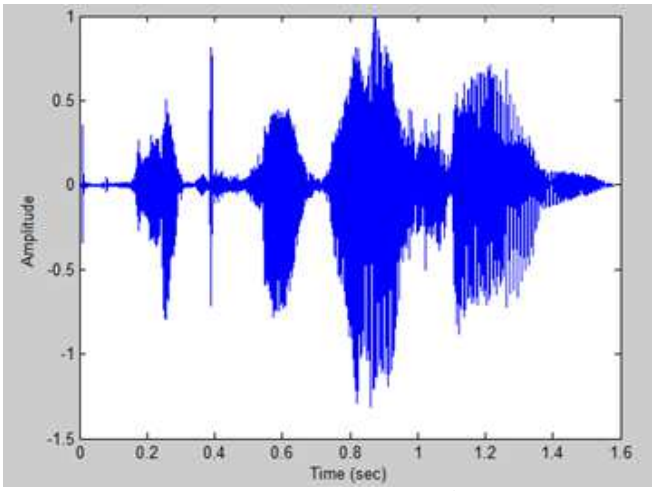

Fig. 1(d) Time-domain presentation of student 3

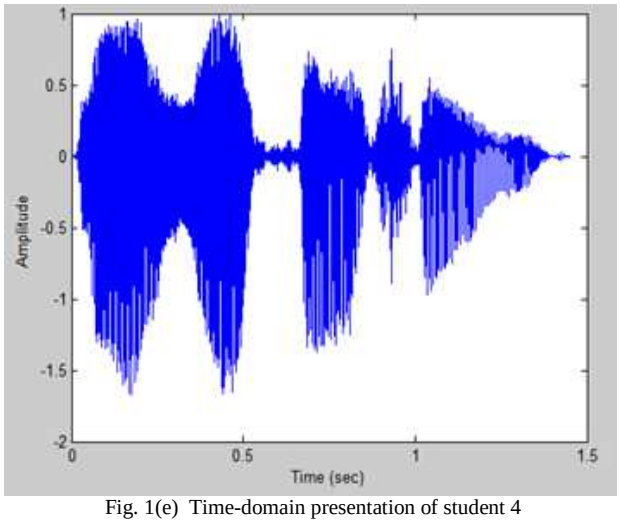

Fig. 1(e) Time-domain presentation of student 4

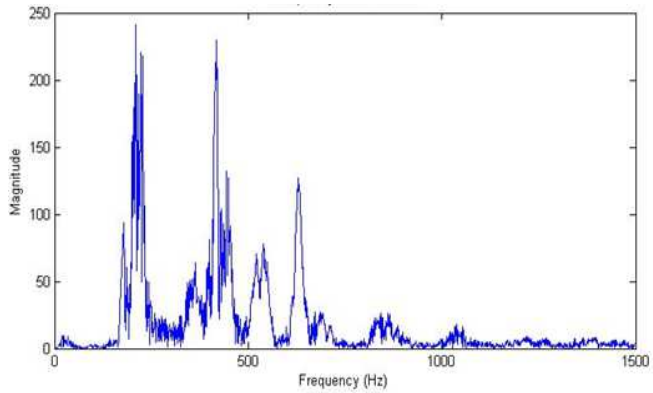

Fig. 2(a) Frequency domain presentation of native speaker

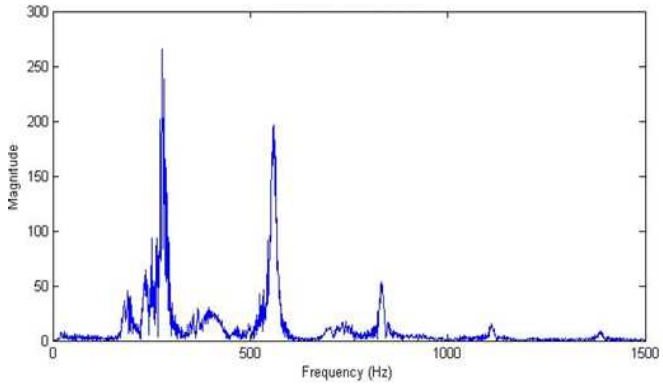

Fig. 2(b) Frequency domain presentation of student 1

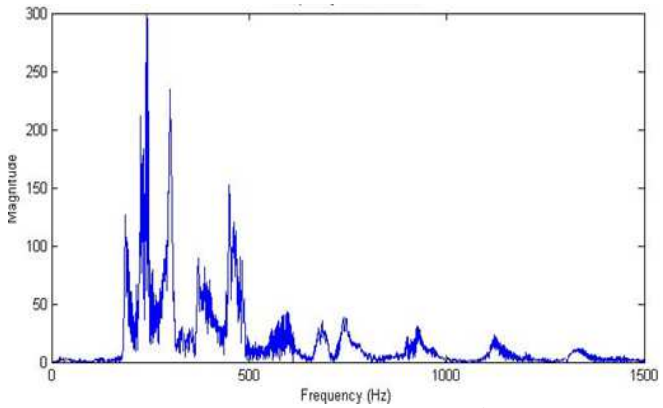

Fig. 2(c) Frequency domain presentation of student 2

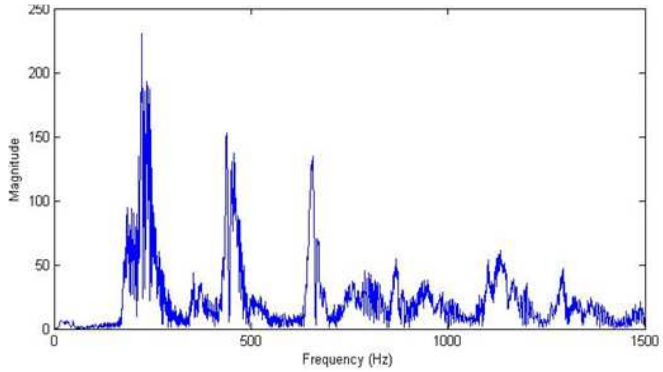

Fig. 2(d) Frequency domain presentation of student 3

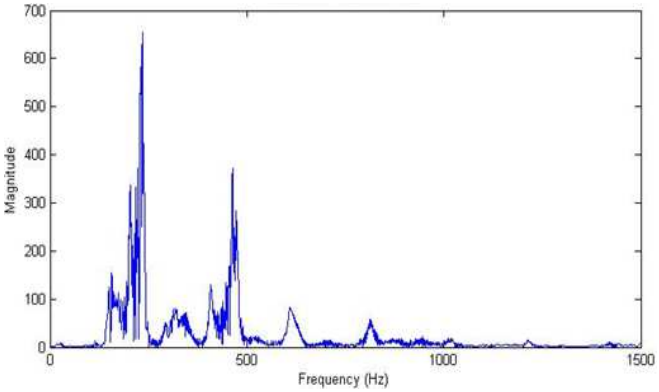

Fig. 2(e) Frequency domain presentation of student 4

Based on the time and frequency domain, the determination of the accuracy of word pronunciation for the 4 students based on the native speaker data almost cannot be identified. Therefore, this work applied an alternative method known as global statistical parameter in analyzing the same signal for determining the accuracy status.

B. Statistical Signal Analysis

The aim of this section is to develop a clustering region representation computed via these 3 parameters in determining the accuracy of pronunciation of words. Clustering has been performed by grouping a set of coefficient parameter from the speaker in one group.

Clustering can be formulated as a multi-objective optimization problem. In this paper, clustering is used to determine the accuracy and precise level of the pronunciation. This process acts as a magnitude spectrum that is physically meaningful to measure the accuracy. Table 1 to Table 5 show the data obtained through the statistical signal analysis processes for each of pronunciation signals. The parameters involved root mean square (rms) value, kurtosis value and skewness value.

TABLE I
DATA STATISTICAL SIGNAL ANALYSIS FROM NATIVE SPEAKER

Data	Kurtosis	Skewness	r.m.s
1	7.4081	-0.6744	0.2583
2	7.1468	-0.6831	0.2501
3	6.5038	-0.7606	0.2587
4	6.3084	-0.7344	0.2771
5	6.2323	-0.6967	0.2791
6	5.729	-0.6643	0.2881
7	5.5256	-0.5943	0.2919
8	6.0393	-0.7044	0.2963
9	5.6986	-0.5632	0.2747
10	6.6175	-0.6727	0.2763

TABLE III
DATA STATISTICAL SIGNAL ANALYSIS FROM STUDENT 1

Data	Kurtosis	Skewness	r.m.s
1	9.9924	-0.4897	0.1807
2	10.535	-0.5045	0.17
3	9.4837	-0.5325	0.1795
4	10.2802	-0.5974	0.1853
5	9.411	-0.5594	0.1844
6	9.3353	-0.4865	0.194
7	10.0484	-0.5059	0.1922
8	9.5457	-0.5928	0.1862
9	9.6524	-0.5341	0.1925
10	12.2998	-0.448	0.1561

TABLE IIIII
DATA STATISTICAL SIGNAL ANALYSIS FROM STUDENT 2

Data	Kurtosis	Skewness	r.m.s
1	8.1121	-0.6098	0.1959
2	8.1345	-0.558	0.1865
3	9.5585	-0.6208	0.1818
4	9.6603	-0.6722	0.1773
5	7.6358	-0.4983	0.2264
6	9.8028	-0.6233	0.1817
7	9.6858	-0.639	0.1892
8	8.0112	-0.6404	0.2046
9	7.7667	-0.6578	0.1867
10	9.3547	-0.5656	0.2031

TABLE IVV
DATA STATISTICAL SIGNAL ANALYSIS FROM STUDENT 3

Data	Kurtosis	Skewness	r.m.s
1	8.3837	-0.8331	0.2027
2	7.5356	-0.9578	0.2721
3	8.0691	-0.9673	0.2334
4	7.9811	-0.8092	0.2353
5	7.5221	-0.9377	0.2564
6	8.6556	-0.8577	0.2352
7	7.9497	-0.8943	0.2272
8	7.8918	-0.7675	0.2026
9	7.3993	-0.9661	0.2684
10	8.1037	-0.9335	0.2641

TABLE V
DATA STATISTICAL SIGNAL ANALYSIS FROM STUDENT 4

Data	Kurtosis	Skewness	r.m.s
1	5.3093	-1.17	0.3774
2	4.9683	-1.1442	0.3703
3	5.0471	-1.1379	0.3799
4	5.0239	-1.0543	0.375
5	4.4395	-1.0042	0.4249
6	5.0544	-1.0984	0.3958
7	5.4998	-1.1117	0.362
8	4.5829	-0.954	0.3683
9	4.8933	-1.0215	0.3666
10	4.0164	-0.8941	0.3958

The identification of the accuracy of pronunciation of the word for each sample between native speaker and the students has been obtained by applying mapping process of the individual data set. The objective of the mapping is to cluster each calculated data within its region. From the region, it can be used to classify and evaluate the accuracy level of pronunciation of word from students. The position of the samples from the students is referred to the samples from the native speaker. The result can be evaluated through the comparison of the position of all samples.

Data values in Table 1 to Table 5 were used to construct the clustering graph. Fig. 4 shows the clustering graph of the data skewness versus rms for all the pronunciation signals [18]. The sample from the French native speaker is used as a reference. It can be observed that student 3 has the highest accuracy of the pronunciation. The sample region of this student is the nearest to the reference region native speaker. The accuracy level follows with student 2, student 1 and student 4. Student 4 has the lowest accuracy level of pronunciation, since the mapping region is the most far to the reference.

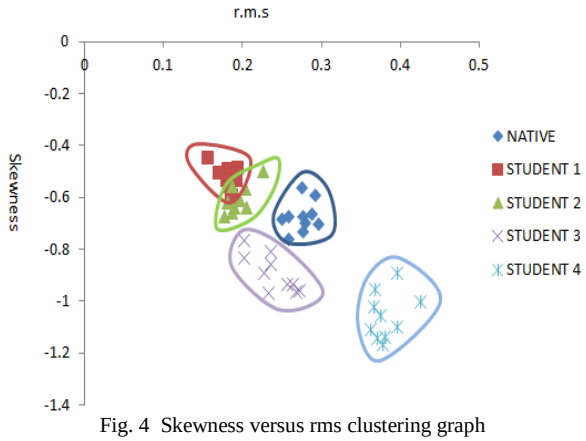

Fig. 4 Skewness versus rms clustering graph

In terms of precise pronunciation of word analysis, it can be seen that group cluster of student 1 has a characteristic of smallest size compare to the other group cluster. It means that data results from student 1 produced high dense areas of the data space. It can be concluded that student 1 has the highest precise pronunciation among them. The statistical distribution of the pronunciation data is almost uniform for each number.

The second pronunciation characterization method is the clustering graph of the data kurtosis versus skewness as shown in Fig. 5. It also can be observed that student 3 has the highest accuracy of the pronunciation. The clustering region of this student is the nearest to the reference region native speaker. The accuracy level follows with student 2, student 1 and student 4. Finally, it can be mentioned that this statistical signal analysis method is capable to reconstruct the speech signal into clearer view to obtain the accuracy for each speaker.

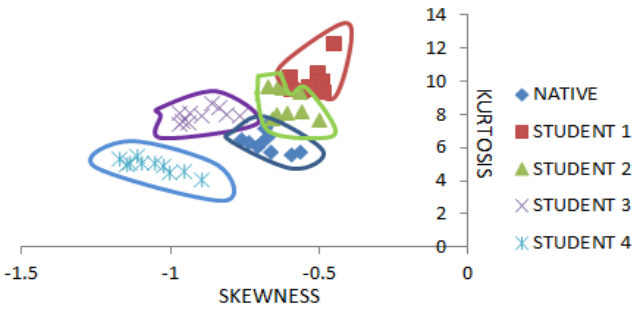

Fig. 5 Kurtosis versus skewness clustering graph

IV. Conclusion

An alternative method for the determination of the accuracy of word vocabulary pronunciation has been presented in this paper. A word 'leaching' in France language spoken by one native speaker and 4 students were used as a subject. The pronunciation has been recorded using data acquisition. Then filtering is applied to obtain low noise signal using MATLAB [17]. The filtered signal has been analyzed using global statistical signal analysis parameter which is skewness, kurtosis, and root mean square. Clustering and mapping method of the parameters. It has been proven that this proposed method is able to characterize the accuracy of the pronunciation of words.

Acknowledgment

The authors express gratitude to the Malaysian Ministry of Education (MOE) and Universiti Teknologi MARA for Research Acculturation Grant Scheme (RAGS) 600-RMI/RAGS 5/3 (165/2014).

References

[1] J. Matthews and J. Cheng, "Recognition of high frequency words from speech as a predictor of L2 listening comprehension," *System*, vol. 52, pp. 1-3, Aug. 2015.

[2] P. Dai, F. Rudzicz, Y. Soon, A. Mihailidis, and H. Ding, "2D Psychoacoustic modelling of equivalent masking for automatic speech recognition," *Signal Processing*, vol. 115, pp. 9-19, Oct. 2015.

[3] S. Swamy and K. V. Ramakrishnan, "An efficient speech recognition system," *Computer Science and Engineering*, vol. 3, pp. 21-27, Aug. 2013.

[4] W. Han, C. F. Chan, C. S. Choy, and K. P. Pun, "An efficient MFCC extraction method in speech recognition," in *Proc. IEEE ISCS'06*, 2006, p. 145.

[5] E. H. Choi, "On compensating the mel-frequency cepstral coefficients for noisy speech recognition," in *Proc. ACSC'06*, 2006, p. 49.

[6] W. Ghai and N. Singh, "Phone based acoustic modelling for automatic speech recognition for Punjabi language," *Journal of Speech Sciences*, vol. 1, pp. 69-83, 2013.

[7] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, pp. 722-737, Jun. 2015.

[8] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. IEEE ICASSP'01*, 2001, p. 73.

[9] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," in *Proc. IEEE ICCI'06*, 2006, p. 1.

[10] L. R. Rabiner and B. H. Juang, *Speech Recognition: Statistical Methods*, ser. Encyclopedia of Language and Linguistics. Amsterdam, Netherlands: Elsevier, 2006.

[11] I. Rebai and Y. BenAyed, "Text-to-speech synthesis system with Arabic diacritic recognition system," *Computer Speech and Language*, vol. 34, pp. 43-60, Nov. 2015.

[12] M. Kim and H. M. Park, "Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition," *Signal Processing*, vol. 117, pp. 126-137, Dec. 2015.

[13] I. Patel and Y. S. Rao, "Speech recognition using HMM with MFCC-An analysis using frequency specral decomposion technique," *Signal and Image Processing*, vol. 1, pp. 101-110, Dec. 2010.

[14] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, pp. 1234-1252, May 2013.

[15] P. L. Chithra and R. Aparna, "Performance analysis of windowing techniques in automatic speech signal segmentation," *Indian Journal of Science and Technology*, vol. 8, pp. 1-7, Nov. 2015.

[16] N. Berahim, S. Besar, M. Z. A. Rahim, S. A. Zulkifli, and Z. I. Rizman, "PID voltage control for DC motor using MATLAB Simulink and Arduino microcontroller," *Journal of Applied Environmental and Biological Sciences*, vol. 5, pp. 166-173, Sep. 2015.

[17] A. Zabidi, N. M. Tahir, I. M. Yassin, and Z. I. Rizman, "The performance of binary artificial bee colony (BABC) in structure selection of polynomial NARX and NARMAX models," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, pp. 373-379, Apr. 2017.

[18] I. M. Yassin, A. Zabidi, R. Jailani, M. S. A. M. Ali, R. Baharom, A. H. A. Hassan, and Z. I. Rizman, "Comparison between cascade forward and multi-layer perceptron neural networks for NARX functional electrical stimulation (FES)-based muscle model," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, pp. 215-221, Feb. 2017.