Analysis of the Indonesian Vowel /e/ For Lip Synchronization Animation

Anung Rachman
Electrical Engineering and Information Technology Department
Universitas Gadjah Mada, Yogyakarta, Indonesia
Institut Seni Indonesia, Surakarta, Indonesia
anung.rachman@mail.ugm.ac.id

Risanuri Hidayat
Electrical Engineering and Information Technology Department
Universitas Gadjah Mada, Yogyakarta, Indonesia
risanuri@ugm.ac.id

Hanung Adi Nugroho
Electrical Engineering and Information Technology Department
Universitas Gadjah Mada, Yogyakarta, Indonesia
adinugroho@ugm.ac.id

***Abstract* - Currently, voice recognition technology is widely used to produce lip sync animation. Vowels take the most dominant roles for lip sync animation as it always exists in every syllable. Therefore, it is necessary to select appropriate vowel traits for the system to be accurate. In general, there are five vowels of Indonesian language, namely /a/ /i/ /u/ /e/ and /o/. However, there are two vowels that contain several different tones: /o/ that are pronounced /o/ and /O/, and /e/ that are pronounced /e/, /ə/, and /ɛ/. The difference in tone can affect the accuracy of voice recognition on the lip sync animation system if it is not specified further. In this paper, the characteristic values of vowel /e/, /ə/, and /ɛ/ are compared and analyzed to find the significance of the difference. The sought characteristic values are the frequency of the formant (F1, F2, and F3) through the Praat software used to extract the features. Comparison is done using a statistical test of t-test. The results show that the three vowel tones /e/ have significant differences for all of F1 and most of F2.**

***Keywords* -** *animation, voice recognition, vowel references, formant frequencies*

I. INTRODUCTION

Dialogue is one of the important things in an animated clip. The dialogue is the result of synchronization between the sound of the actor being recorded and the lip sync of the animated character. In animated lip sync, vowel is the most important factor because it also represents every syllable in the conversation. For example the word "apa", then the form of lip sync is the same for the vowel /a/ and the syllable "pa". Therefore, accurate voice recognition is required to produce animated lip sync that matches the voice of the actor. Previous studies related to vowel recognition for lip sync animation have been widely conducted [1]. But most of the research is done by researchers from America and Japan which means they use vowel references according to their language. Indonesian language contains vowels that have different pronunciation (mouth shapes) from other countries. An application of existing vowel recognition techniques should be carefully considered to be appropriate for Indonesian language.

The vowel reference is used by the system to recognize the vowels on the actor's voice [2]. This reference is derived from the results of feature extraction of the vowels. One of the characteristic extraction results in the frequency value of the formant, especially in F1 (first formant) and F2 (second formant) is a representation of vowel quality. The vowel recognition results are determined by comparing the vowel formant of the voice of the actor and the vowel reference formant. If the comparison is not appropriate, then the introduction of the vowel becomes inaccurate, and finally the animated character lip shapes also become inappropriate.

In general, Indonesian phonemes contain five vowels, namely /a/, /i/, /u/, /e/, and /o/. But among the five vowels, there are two vowels that have more than one tone. The vowel /o/ has two kinds of tones, such as /o/ in different "foto" tones with /O/ on the word "pokok". Similarly, vowel /e/ which has three different tones such as /e/ on the word "sate" that is different from /ə/ in the word "sekolah", or /ɛ/ in the word "bebek". Different tones will of course have different formant values. Additionally, the tone of some of them are also represented in different mouths.

This research paper discusses the level of the significance of the difference in the formant value of the three-tone vowel /e/ in Indonesian phonemes. The results can be used as a reference for making vowel references on animated lip sync systems. By knowing the level of the significance of the difference of the three kinds of tones, the vowel recognition system to determine the lip sync of the characters of the animation become more accurate.

II. INTRODUCTION TO VOWELS

Research related to the introduction of vowels has been done in the languages of various countries with various designations. Azmi [3] analyzed Malay vowel features (Malaysia) using Spectrum Delta method (SpD). The study used four classifiers, the SpD method produces a vowel recognition accuracy rate of 92.42% to 95.11%. Plonkowski [4] presents a simple method for the introduction of Polish vowels. Martin determines three characteristics of the frequency field, namely average, standard deviation, and the maximum value to recognize each vowel. Hindi language is also investigated with regard to the introduction of vowels. The feature extraction was performed by taking three formant frequencies and the cepstral

978-1-5386-0549-3/17/$31.00 ©2017 IEEE

feature of each vowel to improve the performance of the classification under conditions of noise [5].

Hwang [6] conducted a study on the introduction of vowels to lip sync animation. The introduction process consists of input signal, filtering, FFT, and identification. The result is the system is able to recognize the vowels of the actor incoming sound signal in real-time. Hwang displays a block diagram for lip sync animation system using LPC filter (Liner Predictive Coding) as shown in figure 1.

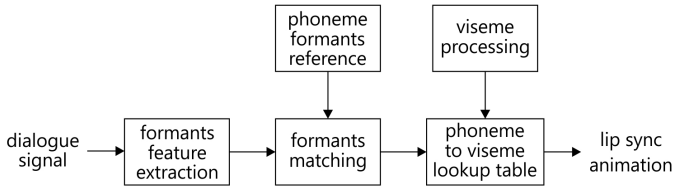

Fig. 1. Block of lip sync animation diagram

To make the lip sync animation system accurate in recognizing the speech signal, the vowel reference of the vowel frequency must be precise as there are differences in vowel precision when the same vowel is pronounced in different vowel locations.

III. INDONESIAN VOWEL CHARACTERISTICS

Single vowel sound referred to as monophthong has differences from consonants in the articulation (obstruction) of the mouth. The vowel sound is produced unimpeded in the vowel cords of the human mouth. Vowels in Indonesian language are very different from vowels in English as a universal language. Indonesian vowels have fewer sound variations and clearer lips shape.

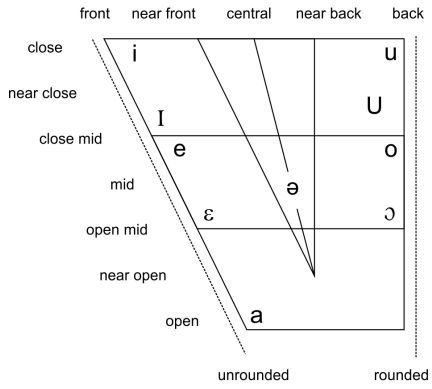

Fig. 2. Map of Indonesian vowels

Indonesian Monophthongs are divided into a round and rounded mouth in their pronunciation. The vowel sound is determined by the shape of the lips and the location and position of the tongue.

[a] location of the tongue in the middle, the position of the lower tongue, neutral lip shape
[i] position of tongue in front, high tongue position, lip shape is not round
[u] the tongue is in the back, the position of the tongue is high, round lip shape
[e] the tongue is in front, the position of the tongue, lip shape is not round
[o] the tongue is in the back, the position of tongue is in the middle, round lip shape

The shapes of the mouth and lips in the vowel pronunciation as shown in figure 3.

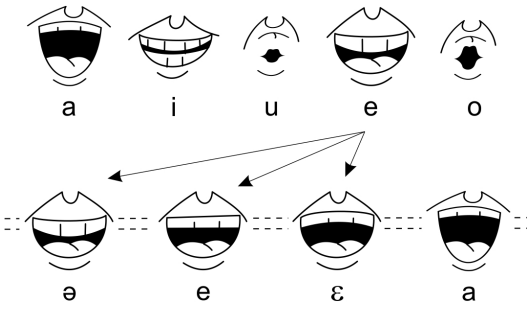

Fig. 3. Mouth shapes of Indonesian vowels

IV. FORMANT

The researchers use a variety of devices to achieve a high degree of accuracy in analyzing the formant. The device is usually equipped with a spectrograph that can be used to analyze the sound quality along with the use of other features. To find the value of the formant including the width of the field, Linear Predictive Coding (LPC) analysis is popularly used. However, sometimes the LPC does not accurately calculate the value of the formant because it detects the location of the vowel incorrectly. The use of LPCs together with a spectrogram can produce vowel formant measurements with a higher degree of accuracy, since the spectrograph can show the investigator the error detection part by the LPC.

Formant is the spectral peak of the sound spectrum. The formant frequency refers to the resonance that lies in the human vowel path [7]. The vowel feature is almost entirely located on the first two forms F1 and F2. For male voices, F1 has a maximum value of up to 1000 Hz, but for women's votes it can reach 1100 Hz (soprano). While F2, F3, and F so on, have maximum frequency ranges of its multiples. Figure 4 shows the frequency of the formant extracted from LPC spectral peak.

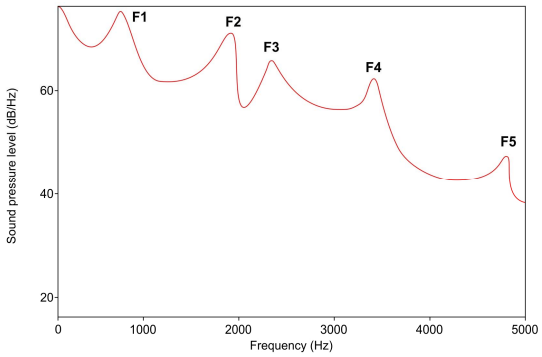

Fig. 4. Formant frequency extracted from LPC spectral peak

V. METHODOLOGY

The research analyzed the significance level of the difference in the vowel formant value /e/ in words. Therefore, only one person was selected as the source of the vote so that the results focus on the same sample of the population. The selected sound source is a woman with the assumption of having a higher frequency range. Spoken words are divided according to the location of the vowel. For vowel 'e', then a word spoken for example 'becak' represents the location of the vowel in front (be), and the word "sore" which represents the location of the vowel at the back (re). Each vowel position is represented by 10 words. Thus, for a tone consists of 20 word samples.

The search for the F1 and F2 formant values is done using Praat software. Praat program created by Paul Boersma and David Weenink from the University of Amsterdam was used to analyze and reconstruct the acoustic conversation signal. Standard Praat settings include: maximum formant frequency 5500 Hz, signal window length 25 milliseconds, and a dynamic distance of 30 dB.

To produce the value of the formant on the recording of the prepared word, the step taken is to extract the formant on the sound recording. If the visual formant has appeared on the spectrogram (dots), then the next step is to select on the vowel part, especially only the middle around (stable range of the formant). This is because the edge of the vowel contains a consonant attached to it. Once selected, the formant data can be retrieved.

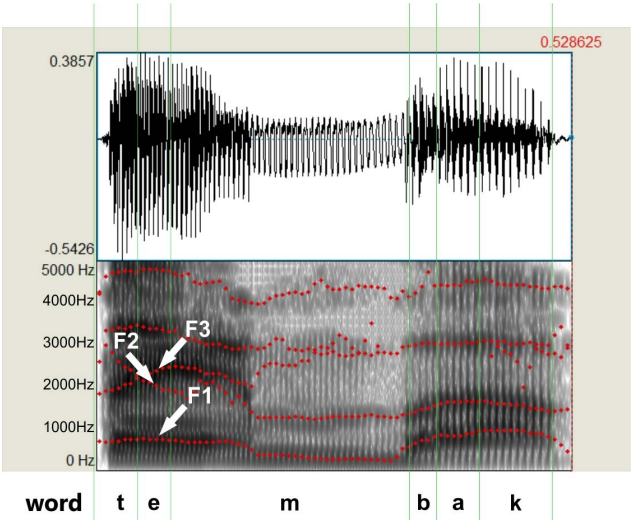

Fig. 5. The formant data retrieved from the “tembak” sound in the middle /e/ vowel

The variance frequency data of all vowels /e/ with various positions are then grouped by the same population to be compared. The comparison method uses the t-test statistics for the same vowels but at different positions on a set of words. If the test value is above 0.05, then the comparable formant value does not differ significantly, meaning it is feasible to serve as a system reference. If a appropriate formant value exists in more than one position, then the selection can be reinforced by looking at the distribution plot. The buildup of the formant values seen in the distribution plot can be interpreted that the value of the forman is not significantly different from each other.

VI. RESULTS AND DISCUSSION

The recorded words used as a research sample for a vowel of 20 are divided into two parts: front and back vowels, with 10 parts of each recording. The observed vowels are three (/e/, /ə/, /ɛ/), so that the total word data is 60. The frequency value of the formant vowel /e/ is shown in Table I.

TABLE I. FREQUENCY OF FORMANT VOWEL /E/

Vowel /e/ on word	Formant (Hz)		
	F1	F2	F3
tembak	689	1977	2403
desa	439	1760	2432
pesta	688	1662	2271
relokasi	692	1975	2545
heran	655	1663	2254
medan	674	1469	2455
becak	439	2382	2758
sendok	635	1475	2205
nego	549	2099	2552
lempar	719	1745	2534
tape	510	2628	2992
sore	562	2403	2897
bale	471	2177	2538
sate	482	2374	2668
jahe	603	2342	2759
kafe	473	2251	2683
kare	511	2417	2667
gawe	458	2217	2673
cabe	405	1949	2597
rame	502	2247	2871

As presented in Table I, the formant values are different for similar F1 and F2, even though all these values represent the vowel /e/. Taking into consideration the standard deviation, little certainty about the magnitude of the formant value is required because this value will be used as a vowel reference for the system reference in order to move the lip sync animation according to the voice of the actor.

To see more about the difference in the value of the formant to the vowel position of the word, the comparison of three populations (/e/, /ə/, /ɛ/) for a vowel can be viewed graphically through the plot distribution.

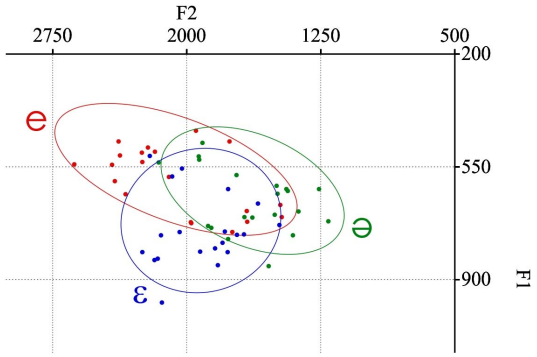

Fig. 6. Format Plot F1 (Hz) and F2 (Hz) show the formant distribution /e/, /ə/, and /ɛ/

It appears in Figure 6 that the three vowels /e/, /ə/, /ɛ/ have close frequency areas. Some frequencies from the sound sample lie in the same area indicated by the stacked area. However, for some words containing these three vowels have different frequency area angles, /e/ tend to lower F1 and high F2, /ə/ tend to F1 and F2 are low, and /ɛ/ tend to high F1 and F2 frequencies .

Although there are parts that accumulate on the plot of the formant distribution for the three vowels in the word, but the difference each /e/ needs to be sought further the level of significance. If one part differs significantly from the others, the mean value of the formant is not feasible for the vowel reference of the system.

Comparison is done for every two sets of words with different vowels /e/, ie /e/ with /ə/, /ə/ with /ɛ/, and /ɛ/ with /e/. It aims to get a more detailed level of differences. To get the significance of the difference, then the analysis uses statistical function called t-test. The results of the t-test can be seen in Table II.

TABLE II. T-TEST OF THE POSITION OF VOWELS ON WORDS

Vowel Comparison		t-test		
		F1	F2	F3
e	ə	0.04517	0.00006	0.72507
ə	ɛ	0.00845	0.00052	0.10919
ɛ	e	0.00004	0.13118	0.05475

In formant F1, the average comparison of vowels /e/ and /ə/, the t-test test value is 0.04517 (if rounded 0.05). It can be interpreted that between the two vowels there are five times of the suitability of every hundred possibilities. The comparison of /ə/ and /ɛ/ has a test value of t-test of 0.00845 which means having a similarity of formant of 845 times on 100000 occurrences of vowels, and /ɛ/ and /e/ of 0.00004, when both appear as 100000, there will be four times the formant pitch. In the t-test, generally the difference is considered significant if the probability is less than 0.05 or the conformity is only one of 20 possibilities. McCall [8] stated that the selection of a significance level of 5% (0.05), it is an agreement among scientists.

With a different visual shape of the mouth, the value of the formant on the test t-test above 0.05 will be a problem if it is used as a reference on the lip sync animation system because it does not differ significantly. In F3 Table II, all t-test values are above 0.05, meaning that all vowels /e/ F3 values can not be used as references in the database. In English vowels, the first two formants (F1 and F2) are frequencies that can be used as a feature [9]. Additionally, Table II proves that the Indonesian vowel /e/ on F3 is also indistinguishable significantly.

On the value of the forman besides F3, there is a test of the value of the t-test is above 0.05 that is the comparison between the vowel /ɛ/ and /e/ on F2. The samples of the vowel /e/ vowel data in the words used in this study consist of two kinds, ie /e/ which exist at the beginning of the words and at the end of the words, as well as vowels /ə/ and /ɛ/.

Figure 7 shows the plot of separate frequency of the forman frequency distribution between the vowel /e/ on the front and the back of the word. The Indonesian vowel /e/ located on the front of the word is labeled e1, and which is at the back of the word labeled e2. While the label ɛ1 is the vowel /ɛ/ which is at the front of the word, and ɛ2 is the vowel /ɛ/ which is at the back of the word. It can be seen in the Figure that e2 has a frequency range of the F2 formant which tends to be equal to ɛ1 and ɛ2. This means that the Indonesian vowel /e/ on the back of a word has a similarity to the vowel /ɛ/ on a word.

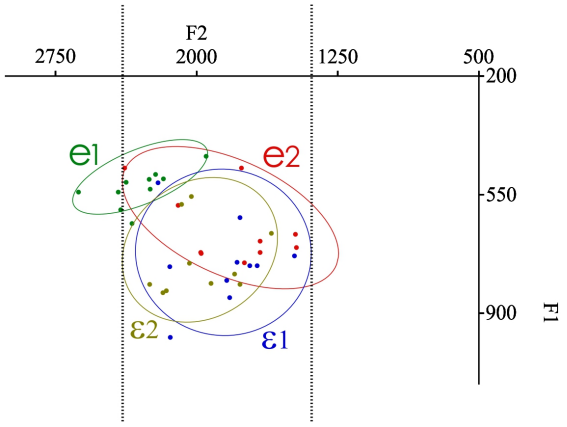

Fig. 7. The range of the forman frequency of F2, between e2 (vowel /e/ which is at the back of a word) has the same tendency compared to ɛ1 and ɛ2 (vowels /ɛ/ which lie either in front of or behind a word)

The fact of similarity the F2 vowel values between /e/ and /ɛ/ as shown in figure 7 can causes the system to be false in recognizing the vowel. Thus, the use of only one formant frequency value as a feature reference for the lip sync animation system can reduce the accuracy of viseme against voice input. Therefore, to make the system work more accurately, it is necessary to use two formant frequency values F1 and F2 of vowel as features reference.

We evaluated the results on the difference of three vowels /e/ using a simple vowel recognition system through Matlab with the output of mouth images. We used the formant features we extracted from LPC and use the Neural Networks to classify. We put seven pictures of mouth from the beginning. And the result is as follows.

TABLE III. COMPARISON BEFORE AND AFTER ADDING THE FORMANT OF /ɜ/ AND /ɛ/ ON VOWEL RECOGNITION

vowel input	a	i	u	e	ə	ɛ	o
Before adding formant /ə/ and /ɛ/ as vowel recognition reference							
After adding formant /ə/ and /ɛ/ as vowel recognition reference							

Table III shows the comparison between before and after adding formant references F1 and F2 for vowels /ə/ and /ɛ/. Adding a reference to the system is done because table II shows the vowel differences /e/, /ə/, and /ɛ/. Classification results show the output of the mouth image becomes more diverse. Viseme diversity will support a series of lip sync animations in more detail.

The formation the mouth shape is an important reference to make sure the accuracy of the shape of the lip sync constructed. The better and more detailed shape of viseme when created, the results will be better [10].

VII. CONCLUSION

Each vowel has characteristics in the first two formants of F1 and F2. If one of the formant values is not significantly different from the other vowel formant, the use of both the F1 and F2 formant values as the system reference database can minimize the vowel recognition error.

A significant difference in the value of the fomant on the vowel to the other vowel signifies that the sampling of the formant in all positions can be used as a reference. However, the best characteristic value to be used as a reference is the value of the formant in a separate area (not overlapping) on the distribution plot.

The Indonesian vowels /e/, /ə/, and /ɛ/ generally have different characteristic values from each other. However, this study proves that there are vowel shapes whose values tend to be the same against other vowels. The vowel position of a word is shown to have an articulation effect on the value of the formant trait.

REFERENCES

[1] S.-M. Hwang, H.-K. Yun, and B.-H. Song, "Automatic lip sync solution for virtual characters in 3D animations," in *International Conference on Convergence Technology 2(1)*, 2013, pp. 432–433.

[2] S.-M. Hwang, H.-K. Yun, and B.-H. Song, "Speaker Dependent Real-Time Vowel Recognition Algorithm for Lip Sync in Digital Contents," in *IT Convergence and Security (ICITCS)*, 2013, pp. 1–4.

[3] M. Y. Shahrul Azmi, "An improved feature extraction method for Malay vowel recognition based on spectrum delta," *Int. J. Softw. Eng. its Appl.*, vol. 8, no. 1, pp. 413–426, Jan. 2014.

[4] M. Płonkowski, "Using bands of frequencies for vowel recognition for Polish language," *Int. J. Speech Technol.*, vol. 18, no. 2, pp. 187–193, 2015.

[5] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition," *WSEAS Trans. Syst.*, vol. 13, pp. 130–143, 2014.

[6] S.-M. Hwang, H.-K. Yun, and B.-H. Song, "Korean Speech Recognition Using Phonemics for Lip-Sync Animation," in *Information Science, Electronics and Electrical Engineering (ISEEE)*, 2014, pp. 1011–1014.

[7] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 1999, pp. 2079–2082.

[8] R. B. McCall, *Fundamental Statistics for Behavioral Sciences*, 8th ed. Wadsworth Publishing, 2000.

[9] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 7th ed. Stamford: Cengage Learning, 2014.

[10] M. B. Nendya and S. Mu'min, "Auto Lip-Sync Pada Karakter Virtual 3 Dimensi Menggunakan Blendshape," *Rekam*, vol. 11, no. 2, pp. 137–144, 2015.