

HIDDEN MARKOV MODELS BASED INDONESIAN VISEME MODEL FOR NATURAL SPEECH WITH AFFECTION

^aEndang Setyati, ^bJoan Santoso, ^cSurya Sumpeno, ^dMauridhi Hery Purnomo

^{a,b,c,d}Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

^{a,b}Informatics Department, Sekolah Tinggi Teknik Surabaya, Surabaya, Indonesia

E-Mail: ^aendang@stts.edu

Abstrak

Dalam komunikasi menggunakan teks input, viseme (visual fonem) berasal dari kelompok fonem yang memiliki penampilan visual sama. Untuk pengenalan emosi ucapan, HMM dilatih untuk setiap emosi dan sampel yang tidak diketahui diklasifikasikan sesuai dengan model yang mengilustrasikan urutan fitur yang terbaik. Penelitian ini didasarkan pada model viseme Indonesia, yang berasal dari pemetaan fonem-viseme untuk Bahasa Indonesia berdasarkan animasi blend shape, yang merupakan hasil penelitian penulis sebelumnya. Indonesia viseme model berasal dari 49 fonem dan menghasilkan 12 viseme, termasuk diam. Sampai saat ini belum ada peneliti lain yang tertarik menggunakan input teks dari kalimat Indonesia afektif. Oleh karena itu, dalam penelitian ini akan didapatkan ucapan alami dari urutan viseme yang dapat menerima parameter ekspresi berdasarkan input teks kalimat Indonesia. Kami menggunakan HMM untuk urutan viseme alami. Ada dua proses utama yang digunakan dalam penelitian ini. Proses pertama adalah penandaan model dengan trigram HMM untuk pemisahan data pelatihan dari input teks kalimat Indonesia berafeksi. Proses kedua adalah proses decoding menggunakan algoritma Viterbi untuk mendapatkan urutan viseme yang digunakan sebagai sinkronisasi ucapan dari bentuk mulut dan gerakan bibir. Akurasi dari hasil eksperimen sekitar 83,73%.

Kata kunci: Hidden Markov Model, Indonesian viseme, Ucapan alami dengan afeksi

Abstract

In a communication using text input, viseme (visual phoneme) is derived from a group of phonemes having similar visual appearances. For speech emotion recognition, a HMM is trained for each emotion and an unknown sample is classified according to the model which illustrate the derived feature sequence best. This study was based on Indonesian viseme models, which are derived from the phoneme-viseme mapping for Indonesian Language based on blend shape animation, which are resulted of previous author's research. They come from 49 phonemes and 12 Indonesian visemes have been produced, including silent. Until now there has been no other researcher who is interested in using text input from affective Indonesian sentences. Therefore, in this paper we will get a viseme sequence for natural speech that can accept a parameter expressions based on text input Indonesian sentence. We used a HMM for natural viseme sequence. There are two main processes in this study. The first process is tagging model with trigram HMM for the separation of training data from text input Indonesian sentence with affection. The second process is decoding process using viterbi algorithm to obtain a viseme sequence that is used as a synchronization by speech from mouth shape and lip movements. The accuracy is about 83.73%.

Keywords: Hidden Markov Model, Indonesian viseme, Natural speech with affection

INTRODUCTION

Speech is one of the most fundamental and natural communication means of human beings. As part of the effort to realize face animation, we explored the mouth shape sequence synthesis techniques for natural speech. The synchronization of speech playing and facial mouth shape is one of the difficult problems in human face animation. If the mouth shape are timely inconsistent, a false feeling will be communicated, and the effect of human-computer interaction will be disturbed. Hence, research on how to mouth shape and natural speech, are of real meaning and practical value [1].

The perception of speech depends not only on acoustic cues, but also on visual cues such as lips movements and mouth shapes [2]. Lips movements when speaking give a visual direction about the things spoken.

A challenge in visual speech animation is that there is variation in the realisation of visemes during the production of natural speech. This event is termed coarticulation, which is the influence of surrounding visemes upon the current viseme [3].

The basic unit of human speech is phoneme [3]. Phonemes and visemes have high correlation [4]. Visemes can be derived using phonemes-to-viseme mapping. The mapping has to be a many-to-one map, because many phonemes can not be distinguished using only visual cues. [4].

The correspondence among speech, mouth shape, lip movements, viseme and phoneme spoken is needed to produce a realistic text-to-audiovisual [5]. A realistic lips animation requires synchronization of viseme with the spoken phonemes [6].

A viseme based on HMM is used in acoustic domain to detect viseme segments from speech. Since emotions play a crucial role in human communication and most of them are expressed through the face, a system based on HMM was built for the synthesis of emotional facial expressions during speech. The HMMs were trained on a set of emotion examples with different intensities [7].

This paper proposes research to build a visualization of the Indonesian visemes. This final system is to get viseme sequence for

Indonesian sentence natural speech with affection.

The paper is organized as follows: we introduce literature review of Indonesian viseme model, Affectional Definition, Introduction of HMM. The next section will discuss HMM Terminology, which aims to better understand the basics of HMM. In this section, will be explained about the definition and the three problems in the HMM. In research methodology section, will be explanation about Viterbi algorithm and Tagging with HMM. The last section will described about experimental results and and concludes the paper.

LITERATURE REVIEW

The literature review consist of Indonesian viseme model, Affectional Definition and Introduction of HMM. Especially for Indonesian Viseme Model is aimed to make the natural speech of a visualisation which can be applied to the mouth shapes. The construction of Indonesian viseme model is expected to be a reference for our future research in the research of making Indonesian Language Talking Head System.

Indonesian Viseme Model

In this research, the Indonesian Viseme models are based on Endang, et al. in 2015 [8]. We acquired this Indonesian Viseme model by phoneme-viseme mapping for Indonesian Language based on blend shape animation [8]. Phoneme is a smallest element of a language that can differentiate a meaning. Viseme is derived from a group of phonemes having similar visual appearances. Viseme classes are defined through linguistic knowledge and grouped by the same visual appearance. The approach used in Indonesian phoneme-to-viseme mapping is based on linguistic data, and then validated through a survey. [8]

The Indonesian language is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. Indonesian is the official language used by almost more than 250 million people in 34 provinces of the Republic of Indonesia. In Indonesian words, there are lots of absorbed words from other languages. They come from vernacular or foreign languages, such as Arabic and English.

Every sound of a language, if proven to be able to differentiate a meaning, can be considered as a phoneme. Phoneme is the smallest unit of sound which becomes a basis for building a human speech. Viseme is derived from a group of phonemes having similar visual appearances, the equivalent unit in the visual domain that models a speech recognition system audio-visually.

Every language sound has an equal chance to become a phoneme, but not all language sounds must become a phoneme. Usually the number of phonemes in a language is fewer than that of language sounds.

Using the letter, visemes and phonemes are correlated through phoneme-to-viseme mapping. It has to be a many-to-one mapping, because many phonemes can not be distinguished using visual signals. A phoneme-to-viseme mapping can help to create the appropriate lips movements of a speech.

In Indonesian words, there are lots of absorbed words from other language, such as Arabic and English. A word can consist of one, two or more phonemes. Indonesian phoneme set based on consonants and vowels classification. There are 12 Indonesian viseme (including silent) generated from 49 Indonesian phonemes, including 13 vowels and 36 consonants, monophthong (single letters) and diphthong (double letters).

The result of phoneme-to-viseme mapping for Indonesian language is based on linguistic approach and validated through a survey. Based on [8] Indonesian visemes are classified into 12 classes. This classification is completed with articulation area for each mouth shape can be seen on Table 1. It is possible to recognize a vowel and determine the consonant analysis. For our future research, we plan to use this outcome to explore other geometric parameters to develop a more refined class of Indonesian Viseme.

Viseme class of "P" is determined based one of set of phonemes /b/, /m/ and /p/. In Indonesian Language, 2 of 3 number of different phonemes, namely /b/ and /p/ have the sound of letter are same with "P", such as "sab-tu" and "sap-tu" (Saturday).

Viseme class of "F" is determined based one of set of phonemes /f/, /v/, /w/ and /ph/. In Indonesian Language, 3 from 4 number of different phonemes, namely /f/, /v/ and /ph/

have the majority sound of letter are same with "F", such as "ak-tif" and "ak-tiv" (active).

Viseme class of "T" is determined based one of set of phonemes /d/, /dh/, /dl/, /dz/, /l/, /n/, /t/, and /th/. In Indonesian Language, 6 of 8 number of different phonemes, namely /d/, /dh/, /dl/, /dz/, /t/, /th/ have the majority sound of letter are same with "T", such as "jum-at", "jum-ad" (Friday) or "a-bad" and "a-bat" (century).

Like Viseme class of "S", 6 from 12 number of different phonemes, namely /ps/, /s/, /sh/, /ts/, /sy/, and /z/ have the sound of letter are same with "S", such as "ber-sa-rat", "ber-sha-rat" or "ber-sya-rat" (conditional).

While Viseme class of "G", 5 from 8 number of different phonemes, namely /g/, /gh/, /k/, /kh/, and /q/ have the sound of letter are same with "G", such as "gu-bug", "gu-buk" or "gu-buq" (hut).

Affectional Definition

Affection is usually identified with emotion, but actually this is very different phenomena although closely related. The fundamental difference between emotion and affection is that the emotion is something that takes place inside the person, while the affection is something that flows and moves from one person to another, producing some emotion.

Detecting emotional dimensions [9] in speech is an area of great reaserch interest as a means of improving human computer interaction. Considering the use that we make of the word 'affection' in every day's life, it can be inferred that affection is something that can be given to others. We say that we "give affection" or we "receive affection". This way, it seems that affection may be something that we can provide and receive. On the contrary, emotions are neither given nor taken, they are only experienced by oneself without the requirement of any other person.

Psychologists have tried to explain the human emotions for more two decades. However, they have not yet agreed upon a set of basic human emotions [11] as shown in Table 2. They disagree on the exact number of affects, i.e. basic emotions.

From Table 2, it can be seen that some form of general agreement has been reached as regards the definition of at least 4 key emotions, anger, fear, sadness, and happiness.

So, there exists a relationship between facial expression and emotional state.

Table 1. Indonesian Viseme Classification [8]













Code	Viseme Class	Number of Phoneme	Set of Phoneme	Mouth Shape	Articulation Area
V1	P	3	/b/, /m/, /p/		<i>Bilabial</i> , a consonant produced by two lips, lower lip closing to upper lip
V2	F	4	/f/, /v/, /w/, /ph/		<i>Labio-dental</i> , a consonant produced by upper teeth and lower lip, upper teeth closing to lower lip
V3	T	8	/d/, /dh/, /dl/, /dz/, /l/, /n/, /t/, /th/		<i>Dental / alveolar</i> , a consonant produced by the tip of tongue attached to gum, a coarse area behind upper teeth
V4	R	1	/r/		<i>Alveolar-semi vowel</i> , a consonant produced by vibrating tip of tongue attached to gum, a coarse area behind upper teeth
V5	S	12	/c/, /j/, /ks/, /ps/, /s/, /sh/, /ts/, /sy/, /x/, /y/, /z/, /ny/		<i>Palatal alveolar</i> , a sound produced by tongue touching hard palate
V6	G	8	/g/, /gh/, /h/, /k/, /kh/, /ky/, /q/, /ng/		<i>Velar</i> , a sound produced by papillae of tongue touching soft palate
V7	A	1	/a/		Front vowel, open and low
V8	I	2	/i/, /I/		Vowel between front and mid, closed, and high
V9	E	4	/ə/, /e/, /ɛ/, /ai/		Vowel between front and mid, between semi closed dan semi open, medium
V10	O	3	/o/, /O/, /oi/		Rear vowel, between semi closed and semi open, medium
V11	U	3	/u/, /U/, /au/		Vowel between rear and mid, closed, and high
V12	-	0	silent		Upper lip and lower lip closing together, constant.
Total		49			

Table 2. List of Emotional Categories [10]

Psychologist	Emotion
Lazarus	Anger, Fear, Sadness, happiness, Anxiety, Disgust, Pride, Same, Guilt
Ekman	Anger, Fear, Sadness, happiness, Disgust, Pride, Same, Guilt
Buck	Anger, Fear, Sadness, happiness, Anxiety, Disgust, Pride, Same, Guilt
Lewis & Haviland	Anger, Fear, Sadness, happiness, Anxiety, Disgust, Pride, Same, Guilt
Banse & Scherer	Anger, Fear, Sadness, happiness, Anxiety, Disgust, Same
Cowie, et al.	Anger, Fear, Sadness, happiness, Anxiety

[11], [12] believed there exist a relationship between facial expression and emotional state. The proponents of the basic emotions view, assume that there is a small set of basic emotions that can be expressed distinctively from one another by facial expressions. To match a facial expression with an emotion implies knowledge of the categories of human emotions into which expressions can be assigned. The most robust categories are discussed in the following paragraphs.

In this research, we used eight basic emotions, namely happy, sad, fear, angry, surprise, disgust, shame and neutral. The description on the facial expressions of basic emotions can be found in Tabel 3.

Introduction of Hidden Markov Model (HMM)

Unlike other template image, viseme has an identity in two different media. The first is audio domain, where viseme is often related to phoneme as linguistic unit. The second is image domain, where viseme is defined by the image of human articulator.

Viseme is an equivalent unit in visual domain, but there are some who disagree with what a viseme is and how many viseme models are proper for a speech recognition system in audio-visual.

Based on [15] several methods are combined to improve the accuracy of HMM based Part-of-Speech (POS) tagger for Indonesian Language. POS tagging is the process of assigning part-of-speech tags to words in a text. A part-of speech tag is a grammatical category such as verbs, nouns, adjectives, adverbs, and so on. POS tagger is an essential tool in many natural language processing applications such as word parsing, question answering, and machine translation.

[3] Phonemes are the standard modelling unit in HMM-based continuous speech recognition systems. Visemes are the equivalent unit in the visual domain. Many authors have demonstrated that the incorporation of visual information into speech recognition systems as a pattern recognition task, the most common solution is a HMM-based system. Phonemes are the typical model unit for continuous speech.

Table 3. Textual Description of Facial Expression [11, 12]

No	Basic Emotion	Textual Description of Facial Expression
1	Neutral	The eyebrows and eyelids are relaxed. The upper and lower lips slightly closed.
2	Sad	The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is relaxed.
3	Angry	The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are pressed against each other or opened to expose the teeth.
4	Happy	The eyebrows are relaxed. The mouth is open and the mouth corners pulled back toward the ears.
5	Fear	The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.
6	Disgust	The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.
7	Surprise	The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is opened.
8	Shame	Shame is universal expression that is seen when the eyes are turned downward with a sad or worried look.

This research presents an approach to extract visemes from the two domains above. In the image domain, mouth shape is represented by nine feature points in the lips contour, extracted using face tracking and mouth image analysis. In the acoustic domain, viseme segmentation is obtained automatically by harmonizing phonemes string in an audio signal using Viterbi algorithm, Hidden Markov Model for the training.

Viterbi algorithm is used for guessing the most possible state sequence of observable states and an HMM. Besides, this algorithm can also be used for calculating the matching probability of the observable state and the HMM.

To obtain a natural viseme sequence, Hidden Markov Model method is used. Two main reasons for this choice are: (1) HMM provides time series data. A human speech is based on time unit function, both the observable state as well as the hidden state. This matches the HMM model's characteristics well; (2) HMM is a stochastic model. This character also matches the reality in which a human speech from different persons has different mouth shapes and different lips movements although the words spoken are the same.

HMM TERMINOLOGY

Before discussing the research methodology, will be explained about a HMM terminology. Purpose of discussion of this section is get to the basic knowledge of HMM for the better.

Hidden Markov Model (HMM) is used for modeling stochastic processes and sequences in various applications. Some of these applications are found in natural language modeling, handwriting recognition, and especially in voice signal processing. The theory was developed and was widely used for many learning problems, specifically for speech processing.

Hidden Markov Model (HMM) observe a probabilistic function of the states. The stochastic process is not observable, it can only be observed by another stochastic process. In this model, every state omits an observation according to some distribution function over all observations. This distribution is unique for every state.

In HMM Terminology, will be described briefly about the HMM definition and Three problems in HMM. The explanation in this section is expected to support steps of research to be more focused on HMM.

HMM Definition

We mark the observation sequence as $\bar{o} = o_1, o_2, \dots, o_T$, where $o_i \in \{O_1, O_2, \dots, O_M\}$. The probability that state S_i will omit observation O_j is $P(O_j/S_i)$, where $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, M$. Formally, the HMM is defined by:

1. N , the number of possible states.
2. M , the number of possible observations.
3. A , the state transitions probability matrix, $A_{i,j} = P(S_j/S_i)$.
4. B , the observations probability matrix, $B_{i,j} = P(O_j/S_i)$.
5. π , the initial state distribution vector, $\pi_i = P(S_i)$.

The model parameters of HMM is shown in Equation (1).

$$\lambda = (A, B, \pi) \quad (1)$$

So, HMM is always stated in three parts: vector π , transition matrix A , and observation matrix B . The definition of each part will be described below.

Vector π shows the initial probability of the first observation data at a particular stage. The vector size is N , which is the number of states. The total number of vector values must be 1, can be seen in Equation (2).

$$\pi = \{\pi_1, \dots, \pi_N\} \text{ and } \sum_{i=1}^N \pi_i = 1 \quad (2)$$

Matrix A shows the probability of state displacement. Every element a_{ij} from this matrix shows the probability if the observation is currently at state- i and the next observation is at state- j . According to first order Markov assumption, the probability value does not depend on observation time. The value depends only on one preceding observation. The size of matrix A is $N \times N$, where N is the number of states. The total value of a line in this matrix must be 1.0 is written in Equation (3).

$$A = \{a_{ij}\}, 1 \leq i, j \leq N \text{ and } \sum_{j=1}^N a_{ij} = 1 \quad (3)$$

Matrix B shows the probability of an observation symbol at a particular state. The size of matrix B is $N \times M$, where N is the number of states and M is the number of types of unique symbols in an observation data. Every element of matrix b_{ij} shows the probability of symbol j at state- i , can be written as Equation (4).

$$B = \{b_i(v_j)\}, 1 \leq i \leq N, 1 \leq j \leq M \text{ and } \sum_{k=1}^M b_i(v_k) = 1 \quad (4)$$

Three Problems in HMM

There are three basic problems associated with HMM, which are: (1) evaluation problem, (2) decoding problem and (3) training problem. Description of the three problems in HMM is written at the following paragraphs.

Evaluation problem appears if there are several HMM and a sequence of observable states. The question is, from those HMMs, which one is the most suitable for the available observable states? The most suitable one is the HMM which is the most likely to produce a sequence of states similar to the observable states.

Decoding problem appears if there is an HMM and a sequence of observable states. The question to be answered is, which hidden state is the most suitable for the HMM pair and the available observable states?

Training problem is the most complex one. If there is a set of observable state sequences and a set of hidden state sequences, the problem is, find the most suitable HMM for both sets. Based on the data he has, he has to find an HMM which can model the relationship between those data.

To solve the above three problems in HMM, some HMM algorithms have been developed. Forward, Backward, and Viterbi algorithms can be used for evaluation problems. For decoding problems, only Viterbi algorithm can be used. For training problems, Viterbi, Baum-Welch or Segmental k-Means algorithm can be used.

RESEARCH METHODOLOGY

In research methodology will be described briefly about the Viterbi algorithm and Tagging model with HMM. This Viterbi Algorithm was used for calculating the matching probability of the observable state. In speech recognition, it is useful to associate an optimal sequence of states to a sequence of observation, given the parameters of a model, which state allows to locate the word boundaries across time. Especially for Section of Tagging model with HMM in this research methodology, there is an explanation of a system has been constructed to generate the a natural viseme sequence.

Tagging model with HMM was a system of implementation of HMM. There are two processes in the implementation, separation of training data and decoding process with viterbi algorithm. Further details of the implementation of HMM, will presented in Tagging model with HMM section.

Viterbi Algorithms in HMM

In this research, the problems are recognition and training, so the discussion will be focused on Viterbi only. The input-output diagram of this Viterbi algorithm can be seen as Figure 1.

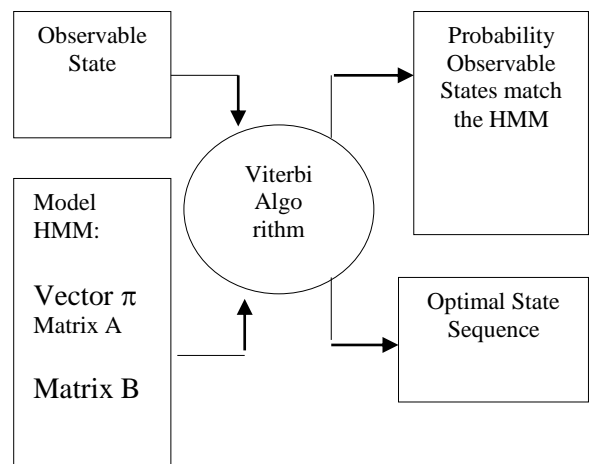


Figure 1. Input-Output Diagram of Viterbi Algorithm

A reasonable optimality criterion, consists of choosing the state sequence (or path) that has the maximum likelihood with respect to a given model. This sequence can be determined recursively via the Viterbi algorithm.

The Viterbi algorithm is used to produce the most probable sequence of states given a sequence of observations and given the model as Equation (5):

$$\arg \max_S P(S | O, A, B, \pi) \quad (5)$$

At each time frame in equation (5), the data structures are updated by finding the path with the highest probability to each state at this time. When the entire observation sequence has been consumed, a backtrace recovers the state sequence.

We now turn to the problem of finding the optimal state sequence using Viterbi algorithm. The Viterbi algorithm stage are described below:

(1) Initialization Stage

Calculate the probability of O_1 at a particular state. This probability is calculated by multiplying vector π with matrix B. The column selected from B must match the symbol at O_1 .

$$O_1(i) = \pi_i \cdot B_{i,S_1} \quad i = 1, 2, \dots, N_S \quad (6)$$

where π_i is the prior probability of being in state S at the time $n = 1$.

(2) Induction or Recursion

Calculate the probabilities of O_2, O_3, \dots, O_T at some particular states. The probabilities are calculated from the multiplication of matrices A and B with the previous probabilities. The state to visit is determined from the result of the multiplication which gives the largest value. Hence, the best path is found by induction.

$$O_n(j) = \max_{1 \leq i \leq N_S} (O_{n-1}(i) \cdot A_{i,j} \cdot B_{j,S_n}) \quad (7)$$

$$1 \leq i < N_S, 1 \leq j \leq N_S, 2 \leq n \leq N$$

(3) Termination

The optimal state sequence is the one visited at the previous steps. The value of the final probability is also treated as an output. Find the best likelihood when the end of the observation sequence is reached.

$$P(S | A, B, \pi) = \max_{1 \leq i \leq N_S} O_N(i) \quad (8)$$

For details of the Viterbi algorithm stages in the research methodology was described in section Tagging model with HMM below.

Tagging Model with HMM

Tagging model with HMM, which used Viterbi algorithm, consists of two main processes, namely the separation of training data and the process of decoding using Viterbi algorithm.

This implementation of HMM, which used viterbi algorithm was performed for generation of natural viseme sequences. The Input of viseme sequence generation is the affective Indonesian sentences. This sentence as an input to get the best state of viseme frame's length.

The following stage will presented a system was built as the implementation of HMM using viterbi algorithm consists of two main processes, namely the process of separation of training data and the process of decoding using viterbi algorithm.

Separation of Training Data

The first step in the process of Viterbi algorithm implementation is the separation of training data. This separation is based on the input from an Indonesian sentence and a sequence of its state. Separation of training data will depend on the input affection's parameters.

The amount of training data have been used in the proposed research are 50 Indonesian sentences. In the 50 Indonesian sentences, which were used as this training data, has represented all of 12 Indonesian viseme classifications (included silent), which consists of 49 Indonesian phonemes, as listed in Table 1.

The training data for each parameter of affection are different. We used 8 affection, namely happy, sad, fear, angry, surprise, disgust, shame and neutral.

So, the amount of separation of training data for first main process was multiply 50 Indonesian sentences with 8 parameter expressions different. So, total of training data were 400 training data.

The training data will be split into four kinds, namely (1) the amount of Trigrams, (2) the amount of Bigram, (3) the amount of

viseme state, and (4) the number of successful states.

The format of the training data has been stored in a text file. This format will be divided into two lines. The first line is an Indonesian sentence of training data, and the second line is the correct state order for a sentence on the first line.

Decoding Process Using Viterbi Algorithm

After the training data separation process is completed, the next stage is decoding process using Viterbi algorithm. This process will calculate the probability of a state based on the training data, which have been processed from previous processes.

The decoding process uses two HMM matrix, matrix A represents the transition matrix, while matrix B represents the matrix emissions. Both matrices will be filled by different data, namely data of HMM Trigram.

Viterbi algorithm for decoding process as following:

1. Initialization matrix A, matrix B, and vector π .
2. Matrix A is filled of the training data by use of Equation (9).

$$a(z | x, y) = \frac{c(x, y, z)}{c(x, y)} \quad (9)$$

3. B Matrix is filled of the training data by use of Equation (10)

$$b(z | S) = \frac{c(S \rightarrow z)}{c(S)} \quad (10)$$

4. Calculate probability of each the state is using Equation (11)

$$V_{t,k} = \max_{x \in S} [P(y_t | k) \cdot a_{x,k} \cdot V_{t-1,x}] \quad (11)$$

5. Save the best state to the next state calculation is using the above step (4).
6. Repeat steps 4-5 of the above process until all of the data calculation is completed.

Viterbi algorithm calculation can be described by equation (11), where $V_{t,k}$ is the

probability calculation of the most possible sequence of states for $-t$ observation with k as the possibility of the state.

Equation (11) will calculate the maximum value of each calculation $P(Y_t / k)$ will calculate the probability for observation y to $-t$, when given a state k . $P(Y_t / k)$ can be found in matrix B or matrix of emission.

$a_{x,k}$ will calculate the transition probability from state x , that is the state prior to the current state k . $a_{x,k}$ can be found in matrix A or transition matrix.

$V_{t-1,x}$ is the probability calculation of observation $t-1$ or earlier observations on state x , where the maximum value of these calculations will be selected, and state k will be selected as the final state for observation $-t$.

In order to achieve the calculation of (11) in the form of HMM trigram, it is necessary that other parameters are calculated specifically for an HMM trigram form. These parameters will consist of two kinds of matrices, namely matrix A and matrix B specific to trigram HMM.

Matrix A of HMM trigram is still a transition matrix, but the transition in this matrix is a transition from bigram state to trigram state. The calculation of the value of matrix A of trigram HMM can be described by equation (9), where $a(z / x, y)$ is a calculation of the value of matrix A which belongs to HMM trigram.

This *trigram* is calculated during the transition from bigram state x, y towards trigram state x, y, z . The transition will be calculated by dividing the sum of trigram states of the training data described by equation $c(x, y, z)$ by the sum of bigram states of the training data described by equation $c(x, y)$.

This calculation will not count the entire matrix A, but only the values required at the time when the certain state and observation values are given. In addition to the value of matrix A, there is also the value of matrix B that will be calculated.

The calculation of the value of matrix B or HMM trigram emission matrix can be described by equation (10), where $b(z / S)$ is one of the calculations of the value of matrix B which belongs HMM trigram when the observation data z and a state S are given. These calculations will be done by dividing the sum of the states of the observation data described by equation $c(S \rightarrow z)$ by the sum of

the states of training data described by equation $c(S)$.

Similar to the previous equation, this calculation will not count the entire matrix B, but only the values required at the time when the certain state and observation values are given.

With the existence of these equations, the equation of the Viterbi algorithm can be rewritten using Equation (12).

$$V_{t,k} = \max_{x \in S} [a(z | x, y) b(z | S) \cdot V_{t-1,x}] \quad (12)$$

where $V_{t,k}$ is the probability calculation of the most possible state sequence for observation -t which has k as the possibility of its state, as in equation (12).

Equation (12) will also keep calculating the maximum value of each possible state, with $\max_{x \in S}$ that will maximize each x, the observation data of member S, which is the possible state sequence.

$a(z | x, y)$ is the equation calculating the value of the parameter of matrix A or the transition matrix as in Equation (9).

$b(z | S)$ is the equation calculating the value of the parameter of matrix B or emission matrix as in Equation (10).

$V_{t-1,x}$ is the probability calculation of observation -t-1 or earlier observations at state x, where the maximum value will be selected from those calculations, and a state k will be selected as the final state for observation -t as in equation (11).

RESULT AND DISCUSSION

The results from this experiments are described into natural speech test with Viterbi algorithm. This natural speech test has been done on the available 50 Indonesian sentences of testing data for each of 8 affection. So, the number of testing data are 400 testing data. In the 50 Indonesian sentences as the testing data, has represented all of 12 Indonesian viseme classifications (included silent), which consists of 49 Indonesian phonemes..

The text input Indonesian sentences, and then comparing the states obtained from the results of the program execution upon the available states in the training data. Each state

of each viseme from the input sentence is used as a parameter for calculating the accuracy.

If there is more than 1 training data for a sentence, the state of each training data will be compared with the states which have been obtained. Then, the average of all available training data will be calculated. The calculations of accuracy and errors in each affection are done by using similar Indonesian sentences in each affection.

Natural Speech Test

The accuracy calculation on this testing data can be illustrated using Equation (13):

$$\left(\frac{V - s}{V} \right) \times 100\% \quad (13)$$

where V is the number of constructed visemes, and s is the number of predictions from incorrect viseme states. From this calculation, the accuracy percentage of the state to be tested will be obtained. The average of the results from this accuracy calculation is calculated based on each available affection.

For example, the given text input is “*Saya Sakit Kepala*” (I have a headache) with the visemes’s output and state’s output as follow:

Texts Input : *Saya Sakit Kepala*
 Output (O) : S, A, S, A, $_$, S, A, G, I, T, $_$, G, E, P, A, T, A
 Output of State : 2, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3
 Length of Viseme : 17

Suppose there is a similar sentence in six different training data, i.e., training data 1 (DT1), training data 2 (DT2), training data 3 (DT3) until training data 6 (DT6) with their states as follow:

DT1 : S, A, S, A, $_$, S, A, G, I, T, $_$, G, E, P, A, T, A
 State : 2, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 2, 2, 2, 3, 2, 4
 DT2 : S, A, S, A, $_$, S, A, G, I, T, $_$, G, E, P, A, T, A
 State : 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2, 4
 DT3 : S, A, S, A, $_$, S, A, G, I, T, $_$, G, E, P, A, T, A

State : 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2,
2

DT4 : S, A, S, A, S, A, G, I, T, G, E, P, A,
T, A

State : 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2,
3

DT5 : S, A, S, A, S, A, G, I, T, G, E, P, A,
T, A

State : 3, 2, 2, 3, 3, 2, 2, 2, 3, 2, 3, 2, 2, 2, 3, 3,
3

DT6 : S, A, S, A, S, A, G, I, T, G, E, P, A,
T, A

State : 3, 3, 2, 3, 3, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2,
4

The illustration of the accuracy calculation of each output natural speech, as shown in Equation (13) on each training datum can be seen in these following steps:

State of DT1 : 2, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 2, 3,
2, 3, 2, 3

Output_DT1 : 2, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 2, 2,
2, 3, 2, 4

Output Accuracy of DT1: ($V = 17$ and $s = 2$)

$$Accuracy_DT1 = \left(\frac{17-2}{17}\right) \times 100\% = 88.23\%$$

State of DT2 : 2, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 2, 3,
2, 3, 2, 3

Output_DT2 : 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 3, 2, 4

Output Accuracy of DT2: ($V = 17$ and $s = 7$)

$$Accuracy_DT2 = \left(\frac{17-7}{17}\right) \times 100\% = 58.82\%$$

Calculation of output accuracy from DT3 to DT6 using Equation (10), with the same $V = 17$ and respectively DT3 until DT6 are $s_3 = 6$, $s_4 = 4$, $s_5 = 5$ and $s_6 = 3$ can be written in below.

$$Accuracy_DT3 = \left(\frac{17-6}{17}\right) \times 100\% = 64.70\%$$

$$Accuracy_DT4 = \left(\frac{17-4}{17}\right) \times 100\% = 76.47\%$$

$$Accuracy_DT5 = \left(\frac{17-5}{17}\right) \times 100\% = 70.58\%$$

$$Accuracy_DT6 = \left(\frac{17-3}{17}\right) \times 100\% = 82.35\%$$

Therefore, the average of combined accuracy is obtained as follows:

$$Rate_of_Accuracy_DT1,DT2,\dots,DT6 = \frac{88.23\% + 58.82\% + 64.70\% + 76.47\% + 70.58\% + 82.35\%}{6} = 73.51\%$$

Mean Square Error

To evaluate reliability of the propose approach, it is necessary to compare with the other methods, namely Mean Square Error (MSE). MSE is used to measure the performance of State between Output of State and Training Data of State.

MSE is found based on the difference between the value of State observation and expectation. The lower the value of MSE is the better the performance.

Suppose there is a similar sentence in six diferent training data, i.e., DT1 to DT6. For example of sentence was used the same training data with previous.

Table 4. Mean Square Error of 6 DT

Text Input : SAYA SAKIT KEPALA		Output : SASA_SAGIT_GEPATA					
Length of State : 17		Training Data of State					
Output of State	DT1	DT2	DT3	DT4	DT5	DT6	
2	2	2	2	2	3	3	
3	3	2	2	3	2	3	
2	2	2	2	2	2	2	
3	3	3	2	2	3	3	
3	3	2	2	3	3	3	
2	2	2	2	2	2	2	
3	3	2	2	2	2	2	
2	2	2	2	2	2	2	
3	3	2	2	2	3	2	
2	2	2	2	2	2	2	
3	3	3	3	3	3	3	
2	2	2	2	2	2	2	
3	2	2	2	2	2	2	
2	3	2	2	2	2	2	
3	3	3	3	3	3	3	
2	2	2	2	2	3	2	
3	4	4	2	3	3	4	
MSE	0.17	0.35	0.41	0.23	0.29	0.29	
	65	3	18	53	41	41	
Average of MSE = 0.294117647							

MSE can be measured by Equation (14).

$$MSE = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N (S_{i_obs} - S_{j_exp})^2 \quad (14)$$

where S_{i_obs} is the output of state or the observation state, S_{j_exp} is the training data of state or the expectation state, $i = 1, 2, \dots, N$, $j = 1, 2, \dots, N$ is the length of state, M is the number of all training data.

The results of this natural speech test on the available 50 Indonesian sentences in the testing data and comparing 50 Indonesian sentences in the training data for each 8 affection, can be seen in Table 5. The data testing of Indonesian sentences in Table 5 has been represented of 12 Indonesian viseme classifications (included silent), which consists of 49 Indonesian phonemes.

The highest accuracy is found in neutral affection with an accuracy average of 100%, while the lowest accuracy is found in disgusted affection with an accuracy average of 75.33%. And accuracy average of 50 testing data was 83.73%.

REFERENCES

- [1] Ming Xu, Ruimin Hu, "Mouth Shape Sequence Recognition Based on Speech Phoneme Recognition," in *Proceedings of the First International Conference IEEE on Communications and Networking in China 2006 (ChinaCom'06)*, 25-27 Oct, 2006, Beijing, pp. 1-5.
- [2] K. Nielsen, "Segmental Differences In The Visual Contribution to Speech Intelligibility," in *Proceeding ISCA-INTERSPEECH 2004 (ICSLP)*, pp. 1-4, October 4-8, 2004.
- [3] Luca Cappelletta and Naomi Harte, "Phoneme-to Viseme Mapping for Visual Speech Recognition," in *Proceeding of the 2012 International Conference on Pattern Recognition Applications and Methods (ICPRAM 2012)*, February 7, 2012.
- [4] Elif Bozkurt, Cigdem Eroglu Erdem, Engin Erzincan, Tanju Erdem, and Mehmet Ozkan, "Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation," in *Proceeding of Signal Processing and Communications Applications of the 2007 IEEE International Conference*, May 7, 2007, pp. 1-4.
- [5] Arifin, Surya Sumpeno, Mochamad Hariadi, Hanny Haryanto, "A Text-to-Audiovisual Synthesizer for Indonesian by Morphing Viseme", *International Review on Computers and Software (IRECOS)*, Vol.10, No.11, November 2015.
- [6] Arifin, Mulyono, Surya Sumpeno, Mochamad Hariadi, Towards Building Indonesian Viseme : A Clustering-Based Approach, *CYBERNETICSCOM 2013 IEEE International Conference on Computational Intelligence and Cybernetics*, pp 57-61, December 2013.

CONCLUSION

To generate a viseme sequence natural speech, each text input need to go through the process of separation of training data and decoding process using Viterbi Algorithm in Hidden Markov Models. HMM based 12 Indonesian viseme model have success to produce the natural lip movements using Indonesian sentence text input.

The average accuracy of each expression has percentage above 75% and total of the accuracy of all the testing data is 83.73%. These scores indicate that each viseme sequence has produced natural lip movement in each transition between viseme is quite smooth and natural.

ACKNOWLEDGMENT

The authors wish to convey their gratitude to *Ditjen Dikti Kemendikbud RI* (Indonesian Higher Education General Director, Ministry of Education and Culture, RI) for its financial support in the form of scholarships under the *BPP-DN*.

- [7] A. Nogueiras, A. Moreno, A. Bonafonte, and Mario: "Speech Emotion Recognition Using Hidden Markov Models", in *Proceedings Eurospeech*, Scandinavia, 2001, pp. 2679-2682,
- [8] Endang Setyati, Surya Sumpeno, Mauridhi Hery Purnomo, Koji Mikami, Masanori Kakimoto, and Kunio Kondo, "Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation," *IAENG International Journal of Computer Science*, Vol. 42, No. 3, pp. 233-244, 2015.
- [9] M. Schroeder, "Speech and Emotion Research An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis, in Faculty of Philosophy, 2014, Universit"at des Saarlandes, pp. 288-298.
- [10] C. Cullen, B. Vaughan, S. Kousidis, Wang Yi, C. McDonnell, and D. Campbell, "Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction," *Digital Media Centre, Dublin Institute of Technology, Ireland*, pp. 1-6.
- [11] Surya Sumpeno, Mochamad Hariadi, and Mauridhi Hery Purnomo, "Facial Emotional Expressions of Life-like Character Based on Text Classifier and Fuzzy Logic," *IAENG International Journal of Computer Science*, Vol. 38, no. 2, pp. 122-133, 2011.
- [12] Endang Setyati, Yoyon K. Suprpto, and Mauridhi Hery Purnomo, "Facial Emotional Expressions Recognition Based on Active Shape Model and Radial Basis Function Network," in *Proceeding of 2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSAs 2012)*, July, 2-4, 2012, pp. 41-46.
- [13] Alfian Farizki Wicaksono, Ayu Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", On *Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop*, 2nd August 2010.

Table 5 Results of Natural speech Test Calculations

No	Indonesian Sentences	Length of Viseme	Accuration for Each Affection (%)							
			Neutral	Sad	Angry	Happy	Fear	Disgust	Surprise	Shame
1	<i>Fotocopi Xerox</i> (Xerox photocopy)	14	100.00	80,71	77,14	76,43	86,43	82,14	97,14	87,86
2	<i>Saya Minum Susu</i> (I drink milk)	15	100.00	70	75.56	73.33	83.33	70.00	96.67	80.00
3	<i>Kucing Saya Mati</i> (My cat is dead)	15	100.00	73,33	71.11	66.67	83.33	70.00	96.67	80.00
4	<i>Anjing Saya Mati</i> (My dog is dead)	15	100.00	63,33	73.33	63.33	83.33	70.00	96.67	80.00
5	<i>Aku Cinta Padamu</i> (I love you)	16	100.00	62,5	72.92	62.50	84.38	68.75	96.88	81.25
6	<i>Saya Sakit Perut</i> (I have a stomach ache)	16	100.00	75	77.08	75.00	84.38	71.88	96.88	81.25
7	<i>Dia Suka Bermain</i> (He likes playing)	16	100.00	71,88	62.50	71.88	84.38	68.75	96.88	81.25
8	<i>Mereka Bawa Kayu</i> (They are carrying wood)	16	100.00	68,75	72.92	71.88	84.38	68.75	96.88	81.25
9	<i>Foto hitam putih</i> (Black and white photo)	16	100.00	81,88	80	80,63	88,13	78,75	96,25	91,15
10	<i>Sayur mayur hijau</i> (Green vegetables)	16	100.00	76,25	81,88	76,88	85,63	79,38	95,63	83,75
11	<i>Aku Capek Sekali</i> (I am very tired)	17	100.00	71,88	72.92	62.50	84.38	68.75	96.88	81.25
12	<i>Rumah Makan Padang</i> (Padang Restaurant)	17	100.00	74,71	79,41	79,41	88,82	80,59	97,06	86,47
13	<i>Saya Sakit Kepala</i> (I have a headache)	17	100.00	73,53	74.51	73.53	85.29	70.59	97.06	82.35
14	<i>Jeihan sudah insyaf</i> (Jeihan already repented)	18	100.00	80,56	81,67	78,33	83,33	76,67	96,67	85
15	<i>Pulang Pergi Sekolah</i> (Going to and from school)	19	100.00	76,32	75.44	73.68	86.84	73.68	97.37	92.11
16	<i>Saya Tidak Mau Tahu</i> (I do not want to know)	19	100.00	68,42	75.44	71.05	86.84	68.42	97.37	84.21
17	<i>Muadzin adzan di surau</i> (Muezzin call to prayer in the mosque)	19	100.00	75,79	78,42	78,42	82,63	78,42	95,26	84,74
18	<i>Hari Ini Malam Minggu</i> (This is Saturday night)	20	100.00	72,5	70.00	72.50	87.50	70.00	97.50	85.00

Table 5 Results of Natural speech Test Calculations (Continued)

No	Indonesian Sentences	Length of Viseme	Accuration for Each Affection (%)							
			Neutral	Sad	Angry	Happy	Fear	Disgust	Surprise	Shame
19	<i>Segeralah Kamu Pulang</i> (Go home soon)	20	100.00	75	70.00	72.50	87.50	70.00	97.50	85.00
20	<i>Pasar itu sangat ramai</i> (The market was very crowded)	20	100.00	77,5	82,5	81	85,5	79,5	99	85,5
21	<i>Amboi, pulau itu indah</i> (Wow, the island was beautiful)	21	100.00	77,62	78,57	82,38	81,35	78,57	97,14	85,32
22	<i>Bersyukur kepada Tuhan</i> (Thank God)	21	100.00	80,95	79,52	78,1	83,81	76,19	95,24	88,1
23	<i>Saya Tidak Naik Kelas</i> (I fail pass this grade)	21	100.00	71,43	73.02	73.81	88.10	71.43	97.62	85.71
24	<i>Pelangi itu sangat indah</i> (It was a beautiful rainbow)	22	100.00	79,09	79,09	75,91	83,71	78,64	97,27	84,47
25	<i>Kebun binatang Surabaya</i> (Surabaya Zoo)	22	100.00	81,36	80,91	86,36	88,64	76,82	98,18	83,64
26	<i>Singa Itu Berlari Kencang</i> (That lion runs fast)	23	100.00	73,91	75.36	73.91	89.13	71.74	97.83	86.96
27	<i>Nyonya Endang sedang menyanyi</i> (Mrs Endang is singing)	23	100.00	80,71	77,14	76,43	86,43	82,14	97,14	87,86
28	<i>Tips disayang oleh orang tua</i> (Tips loved by parents)	25	100.00	70	75.56	73.33	83.33	70.00	96.67	80.00
29	<i>Antar Saya Pergi Ke Kantor</i> (Give me a ride to the office)	26	100.00	73.08	78.21	73.08	90.38	73.08	98.08	61.54
30	<i>Kami Suka Makan Nasi Padang</i> (I like eating Padang rice)	26	100.00	71.15	78.21	73.08	90.38	71.15	98.08	88.46
31	<i>Jangan Ajak Aku Pergi Dahulu</i> (Do not ask me to go yet)	26	100.00	72.22	76.54	70.37	90.74	70.37	98.15	88.89
32	<i>Bintang kecil di atas langit</i> (Little star in the sky)	26	100.00	79.23	76.54	79.23	89.74	80.77	97.69	83.33
33	<i>Kelelawar Masuk Universitas</i> (A bat enters a university campus)	27	100.00	74.07	72.84	74.07	90.74	74.07	98.15	88.89
34	<i>Rujak Cingur Ini Sangat Pedas</i> (This Indonesian fruit salad is very spicy)	27	100.00	72.22	74.07	68.52	90.74	72.22	98.15	88.89
35	<i>Vas bunga itu berwarna merah</i> (Vase of flowers are red)	27	100.00	78.89	82.59	81,11	86.11	78.89	96.3	88.15
36	<i>Kaum dhuafa berdzikir di masjid</i> (Dhuafas dhikr in the mosque)	28	100.00	81.07	84.29	79.64	84.29	77.5	97.5	85.00

Table 5 Results of Natural speech Test Calculations (Continued)

No	Indonesian Sentences	Length of Viseme	Accuration for Each Affection (%)							
			Neutral	Sad	Angry	Happy	Fear	Disgust	Surprise	Shame
37	<i>Monyet bergelantungan di pohon</i> (Monkey was hanging in a tree)	28	100.00	78.57	76.43	81.43	88.69	77.5	97.14	87.14
38	<i>Ilmu anthropologi dan psikologi</i> (Anthropology and psychology)	29	100.00	80.34	86.55	83.45	86.21	80.34	96.69	88.28
39	<i>Hari Ini Saya Tidak Bisa Hadir</i> (Today I cannot be present)	30	100.00	70.00	75.56	70.00	91.67	70.00	98.33	90.00
40	<i>Gelombang tsunami sedang menerjang</i> (Tsunami waves were crashing)	30	100.00	82.33	79.67	79.00	80.67	80.33	99.33	86.67
41	<i>Tolong Buatkan Minum Untuk Tamu</i> (Please make a drink for the guests)	30	100.00	73.33	77.78	70.00	91.67	73.33	98.33	90.00
42	<i>Puisi ini karangan Chairil Anwar</i> (This poem essay Chairil Anwar)	30	100.00	79.67	77.33	86.33	80.33	78.33	76.00	86.06
43	<i>Marsya sedang belajar teori graph</i> (Marsya is learning graph theory)	30	100.00	76.67	78.33	79.33	81.67	82.00	76.67	83.61
44	<i>Sholat Maghrib pada bulan ramadhan</i> (Maghrib prayer in Ramadan)	31	100.00	77.10	82.9	83.55	83.87	77.10	77.1	87.1
45	<i>Jangan Kau Pecundangi Diriku Sayang</i> (Do not outwit me, dear)	32	100.00	71.88	73.96	71.88	92.19	71.88	98.44	90.63
46	<i>Jeni sedang bermain pasir di pantai</i> (Jeni was playing sand at the beach)	33	100.00	78.18	78.79	81.52	81.82	81.82	70.91	88.18
47	<i>Sang Khaliq adalah penentu akhir zaman</i> (Khaliq is a determinant of the end times)	36	100.00	74.44	77.78	80.56	84.26	81.67	74.44	85.42
48	<i>Membaca sebuah buku teks</i> (Reading a textbook)	38	100.00	78.42	81.58	83.42	89.47	83.16	75.53	82.11
49	<i>Saya Sangat Menyesal Dengan Perbuatan Saya</i> (I am really sorry for what I have done)	38	100.00	78.21	79.49	74.36	91.03	73.08	98.72	91.03
50	<i>Gadis cantik itu mempunyai rambut keriting</i> (The pretty girl has curly hair)	39	100.00	78.97	88.21	85.13	87.69	79.74	77.69	87.95
<i>Accuracy Average</i>			100.00	75.43	77.26	75.98	86.27	75.33	94.25	85.35
<i>Accuracy Average of Expression</i>						83,73375				