ISSN No. 0976-5697

Volume 5, No. 6, July-August 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Prosodic Models of Indonesian Language: State of the Art

Arif B. Putra N.[1] , Kuspriyanto[2], Tito Waluyo Purboyo[3]
[1,2,3]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia
[1]Fakultas Teknik, Universitas Tanjungpura, Pontianak, Indonesia
[3]Electrical Engineering Faculty, Telkom University, Bandung, Indonesia

Abstract: Text to Speech (TTS) is a system that synthesizes speech from text. The quality of TTS can be judged from intelligibility and naturally. Prosody is one of the parameter that can improve the quality of TTS. This study will develop a model of prosody based on information of Indonesian syntax category. Categories syntax is a word or combination of words that can be categorized as a subject, predicate, object or complement in a sentence. Prosody models developed in this study using chunking method to determine the syntax phrase category, and hidden Markov models (HMM) to predict the curve templates that match the input range of syntax phrase category. The hidden state of HMM declared by the template type pitch curve, and state the type of observation is expressed by the phrase syntax. Template pitch curve is developed by combining theory and models of prosodic pitch contours Fujisaki. Prosody generation method to convert the template pitch curve into phoneme codes, duration and pitch values for each input text sentence pitch contour representing speech.

Keywords: Phrosody; Text to Speech; phrase; pitch; Hidden Markov Model

I. INTRODUCTION

Quality text to speech system (TTS) is determined by the intelligibility and naturally. Is the synthesis of speech intelligibility can be understood, naturally is the synthesis of natural sounding speech can be. According Malfrere and Dutoit [1, 2], that there are three basic problems that must be solved to improve the quality of TTS , ie writing graphem into phonemes, prosody generation, and digital speech synthesis. Van Santen et. all [3], saying that prosody has three roles in human communication. First, the acoustic characteristics of prosody can be used to distinguish words, eg pitch in a tone language, a language that the duration of the speech distinguish long and short vowels or double along with their characteristics. Essentially as an additional characteristic that produces different pressures on words containing the same phoneme. Second, the nature of the acoustic prosody is used to structure the shape of the phrases and sayings to express the relationship between phrases and sayings, such as speech that lists the different pronunciations in the prosodic structure of the speech utterance containing inserts. Third, prosody is used to focus attention on a particular word for many uses, for example emphasis on certain words to attract attention.

According to Vincent et. all [4] in the science of linguistics, prosody is the rhythm, stress, and intonation of connected (smaller than the element syllable) when speaking. According Nababan [5], the Indonesian prosody is often known as the intonation, the melody in melafazkan words or sentences. Intonation can be a song or a sentence accuracy of sentence presentation of high and low tones. TTS components that generate the greeting direpresen - invest the prosody in the target, usually a phoneme segment in ms duration, and pitch in Hz. [6, 1-3, 5]. So it can be said that prosody is the change in pitch over time for each phoneme pronunciation is raised, where the values of these changes are expressed in prosody models. To synthesize speech TTS system is to calculate the acoustic prosodic characteristics (such as pitch, duration) of a prosodic marking text (such as syllable stress) pitch contours corresponding to syllable stress tagging text. Prosody is dependence on speaker and language, it is necessary for a particular language model. Various researches related to the modeling of prosody that use foreign languages, such as English language has a lot to do, yet still rare for Indonesian .

One of the prosody models are widely applied by researchers is the prosody models Fujisaki. Fujisaki [18], assumes the presence of components of the phrase and accent components in the model prosodinya. At TTS system components phrases associated with pitch curve and access components associated with pressure on a syllable utterance .

Hidden Markov models (Hidden Markov Model/HMM) is a model that can predict the state will come in and he separated from the state in the past, of the current input. HMM can be used to solve problems where the observations can not be made against a state but we can calculate the probability of a process in a particular state. Several studies using HMM for the TTS system [8, 11, 12, 15, 19-21] . Hidden Markov models require very large corpus of data in the training process .

Based on the above, this research will improve the quality of Indonesian TTS systems by developing a model of prosodic information using the syntax category by combining the theory of pitch contours, prosodic models Fujisaki, chunking method and hidden Markov models. Truncated form text input sentence phrases that are syntactically categorized using chunking method. Speech signal corresponding to the syntactic category of the phrase analyzed its pitch contour method Fujisaki receipts, and text phrases form sentences strung jedanya type characterized by indexes Tobi. Type the phrase, the type of pitch contour and type of pause is inserted into modeling for hidden Markov trained. HMM training results are used to predict the type of curves and the type of input lag on the type of phrase syntax category. Furthermore, the HMM prosody prediction processed to form a raised pitch contour that corresponds to the text input.

II. AN OVERVIEW OF PROSODY MODELS

A. *Prosody Curve*

TTS is a system that turns text into speech. Sentence is represented by a network of letter symbols, and speech can be represented in a network of symbols sounds or phonemes .

If a sentence is expressed as k, and the sequence of phonemes as fn and the end result is a speech then takes two functions to perform the transformation, namely fttp (function text to phonemes) and fpts (phoneme-to-speech function). Relationships text, code and speech phonemes can be expressed in equation (1) and (2).

fn = fttf (k) ..... (1)
u = fftu (fn) ..... (2)
where:
k = sentence
fn = the sequence of phonemes
u = speech
fttp = function text to phoneme
fpts = function phoneme to speech

Word of k and fn phoneme code can be expressed as a series of component constituents respectively. Sentence is a group of characters ks, this can be a letter symbol, the symbol spaces or other symbols. Phoneme code sequence, fn can be expressed as characters ranging from kf, that phoneme cod , both of these relationships can be expressed in equation (3) and (4).

k = s1 , s2 , ... , sn ..... (3)
fn = kf1 , kf2 , ... , KFM ..... (4)
where:
kf = phoneme code
s = character (symbol characters , spaces , or other)
n = number of symbols in a sentence
m = the number of symbols in the phoneme

To produce the desired speech sounds, each phoneme code should come with duration and pitch. As fn' can be expressed as :

fn' = (kf1df1p1) , (kf2df2p2) , ... , (kfmdfmpm) .... (5)
where:
kf = phoneme code
df = duration, constant
p = pitch , constant
m = the number of symbols in a phoneme in speech produced

Based on the equation (5) we can see that fn' characters are made up of three tuple, ie kf , df and p. Df and p value of this is normally determined by a prosody model. If the df and p is zero then the resulting speech synthesis hear unnatural flat where a short-term fixed phoneme. The scale length and pitch is then formed which represent prosody pitch curves as shown in Figure 2.

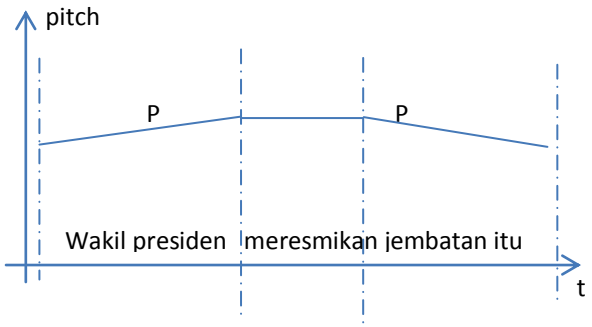

Figure 2. Prosody curve of speech "wakil presiden meresmikan jembatan itu"

Prosody models developed in this study will be generated on the TTS system. TTS system in this study using Indonesian diphone database which has been developed in other studies [6, 7, 13, 14].

B. *Fujisaki Model*

Fujisaki distinguish two types of discrete events in terms phrases and accents are modeled by using the pulse function and a step function. These commands are then controlling the second order filter that generates Fo curves. Fujisaki assumes that prosody can be represented by the phrase component and access component.

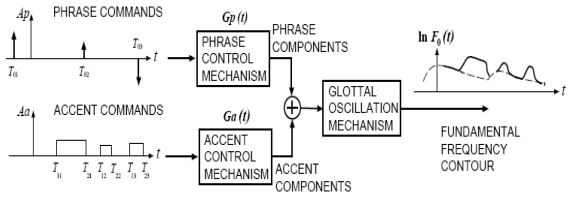

Figure 3. Fujisaki Prosody Model [16]

Some important things from the Fujisaki model are as follows:

a. Curve is superimposed on the prosodic phrase and accent components.
b. The number of phrases and commands two parameters for each command are the amplitude and time.
c. Total orders accents and three parameters for each command is amplitude, time of onset and offset time.

This model has been used to generate prosody for some of the other Indonesian languages developed by Arman [6, 7].

C. *Hidden Markov Model (HMM)*

Hidden Markov Model (HMM) is a statistical model of the development of the Markov chain, where the state can not be observed directly, but can only observe by another set of state. State that can not be observed or hidden state of this research is the set of pitch contour and pause, while an observation set is a set of PoS. The elements of the HMM are:

- N the number of states with state space S = {s1, s2, s3, ..., sN} and the state at time t is expressed by Qt. The research problem is the pitch contour pattern and pause.
- M the number of observations in each state, the observation space V = {v1, v2, v3, ... vm}. The research problem is the type set of PoS.
- A=[aij], is a matrix of transition probability
- B=[bjm], is a matrix of conditional probability of observation vm if the process is in state j, where bjm=bj(Ot)=P(Ot=vm|Qt=sj), $1 \le j \le N$ and $1 \le j \le M$
- π_i , is a distribution of initial state.

So that HMM can be written in the notation $\lambda = (A, B, \pi)$. If given N, M, A, B, and π, the HMM can be used as an observation sequence generator: O = O1, O2, O3, ... OT.

If given a series of sentences, and then represented in a series of PoS, then the chance of the pitch curve to characterize PoS-t (Qt) with the possibility of ascending curve {s1 = s2 = curve decreases, s3} = horizontal curve can only be obtained by observation Qt, with qt = NN, or Qt = VBT, or Qt = MD, and so on.

To predict the type of pitch and pause curve based on observations of type PoS, then used the size of the probability called likelihood (L):

$$L\left(Q_1 = s_{i_1},...,Q_T = s_{i_T} \middle| O_1 = v_{m_1},...,O_T = v_{m_T}\right)$$

$$= \prod_{t=1}^{T} P\left(O_t = v_{m_t} \middle| Q_t = s_{i_t}\right) \cdot P\left(Q_1 = s_{i_1}\right) \cdot \prod_{t=2}^{T} P\left(Q_t = s_{i_t} \middle| Q_{t-1} = s_{i_{t-1}}\right)$$

In order HMM can be applied to various real problems, there are three fundamental HMM problems that can be solved:
- Evaluation problem, namely searching for P (O|λ) or opportunities of observation sequence O = {O1=VM1, VM2=O2, ..., OT=VMT} if given HMM, λ (A, B, π) . These opportunities can be determined by inductively by using the forward algorithm.
- Decoding Problem, which is looking for the optimal state sequence Q* = {Q*1, Q*2, ..., Q*T}. If given the observation sequence O = {O1, O2, O3, ..., OT} and a model λ = (A, B, π). Rows of the best state to be determined in the form of a single path that connects from t = 1, 2, ..., T. To resolve this problem can be solved by the Viterbi algorithm .
- Learning problem, if given the HMM and the observation sequence O=O1, O2, O3, ..., OT, then how to set the parameters of the model λ = (A, B, π) so that P(O|λ) maximum. To resolve this problem using the Baum-Welch algorithm.

N-gram models used in this study is the bigram and trigram. Markings made on the type of pitch curve and the type of input types PoS pause. Marking pitch curve type and the type of pause is represented by hidden Markov models, while the PoS types represented by the state observer. Opportunities transition depends on the state, which is a partner tagging. Probability of emission depends only on the current marking. In order to find the tagging sequence type pitch curve in the case of bigram models is given by equation (6) follows.

$$t_{1-n} = \arg max_{t_1 \ldots t_n} P(t_1)\, x \prod_{i=2}^{n} P(t_i|t_{i-1}) x \prod_{i=1}^{n} P(pos_i|t_i) \quad \ldots. (6)$$

For the case of trigram models is given by the equation (7).

$$t_{1-n} = \arg max_{t_1 \ldots t_n} P(t_1)\, x\, P(t_2|t_1) x \prod_{i=3}^{n} P(t_i|t_{i-1}t_{i-2}) x \prod_{i=1}^{n} P(pos_i|t_i) \quad \ldots. (7)$$

t1-n is the order of the series marking the best curve type for input type PoS, pos,.., posn is the order of the post-type input circuit, and t1, .., tn are elements of the set tagging. P (t1) and P (t2 | t1) is the first order in a series of tagging. For the added designation "<starttag>" bigram model and "<starttag> <starttag>" trigram model at the beginning of the post type. So that P (t1) in equation (6) is calculated using P (t1 | StartTag). Similarly to the model trigram P (t1) and P (t2 | t1) in equation (7) is calculated using P (t1 | StartTag, StartTag) and P (t2 | t1, StartTag). Opportunities transitions and emissions are calculated from the corpus tagging curve type and the type of post. In HMM used method of maximum likelihood odds (MLE) to calculate the probability of transitions and emissions.

III. State of The Art of Indonesian Language Prosody Models

Various research related to the modeling of prosody that use foreign languages, such as English language has a lot to do, yet still rare for Indonesian.

Table I. State of The Art of Indonesian Language Prosody Models

Methods	Focus	Results		
		Superiority	Drawback	Examination
(Amran H, 1984) [13]	Characteristics Analysis 0f Indonesian Speech Prosody		Stage of Analysis	Not specified
Model Fujisaki (Arman AA, dkk ,2001) [7]	Implementing of Fujisaki Methods for Indonesian Language	Produce a short sentence prosody	Less dynamic for long sentences	Not specified
Fujisaki Model and Pitch Contour Theory (Arman AA, 2004) [4]	Sentence segmentation based on linguistic and the linguistic information in the form of words and punctuation marks in the sentence.	Can generate dynamic prosody for a long sentence.	Less dynamic in a sentence that does not have a collection of linguistic and linguistic information in the system	MOS: engine quality in generating prosody: 4:15, Prosody quality compared to human speech 3.65
MNRR (Multi rate Recurrent Neural Network) (Tritoasmoro. Iwan I., Tjondronegoro Suhartono, 2007) [9]	Control prosody by predicting Fo trained with a number of speech patterns.	Pronunciation of words pretty well and still be understood	Prosody is still less natural.	MOS Integibility= 3,45 Fluidity=2,99 Naturalness =2,33
Tagging Analysis Using Speech Filling System (Novianti D, 2009) [16]	The analysis of interrogative sentence is based on information of words "what, how, where, why, when, who"	Interrogative sentence prosody models.	The model is not implemented in the TTS	Not specified
HMM (Sakriani sakti, 2009) [17]	Implement HMM models to synthesize speech in Indonesian	Produce a good inteligibility with small data	Require large data in order to sound natural	MOS = 2,78 SUS word = 90,48% SUS sentence = 54,67%
HMM (Vania C., and Adriani M., 2011) [10]	Adding stress on syllables and words	improving the quality of speech sounds natural, MOS 67.3%	Require large data in order to sound natural	MOS=2.85 SUS word = 85% SUS sentence=64 %

The research publication of Indonesian speech prosody modeling obtained by researchers are shown in Table I.

IV. Conclusions

In accordance with the "Pitch Contour Theory", Indonesian sentence constituent segments, each of which is derived from a set of finite number of curve segment. Information in the form of linguistic categories that are speaker dependent syntax can be used to sort out the Indonesian sentences into segments corresponding sentence with "Pitch Contour Theory".

Each segment can be formed from several sub-segments, each of which can be implemented with the theory of "Fujisaki Model". Fuzisaki models assume that a prosodic curve may consist of two components, namely the phrase and accent

components. Where in use, considers the phrase component is a linear component for each segment, and the accent component is a component of the positive pulse occurs at each syllable to (n-1) on each word. Pitch curve for a sub-segment of the components are super- impose the phrase and accent.

Future works will include generating new algorithms to generate Indonesian Speech using various prosody models, where the sentence segmentation is done based on the information obtained in the process of syntax parsing.

V. REFERENCES

[1] F. Malfrere and T. Dutoit, "Speech synthesis for text-to-speech alignment and prosodic feature extraction," in Proceedings of the 1997 IEEE International Symposium on Circuits and Systems, ISCAS'97. Part 4 (of 4), 4 ed. Anon, Ed. Hong Kong, Hong Kong: IEEE, 1997, pp. 2637-2640.

[2] F. Malfrere and T. Dutoit, "High Quality Speech Synthesis for Phonetic Speech Segmentation," 1997, pp. 2631-2634.

[3] J. v. Santen, T. Mishra, and E. Klabbers, "Prosodic Processing," 2008, pp. 471-488.

[4] J. Vincent, H. Van, and E. V. Zanten, "Prosody in Indonesian Languages," Leiden University Centre for Linguistics, 2007, pp. 471-488.

[5] Nababan, Intisari Bahasa Indonesia untuk SMA PT Kawan Jakarta, 2008.

[6] A. A. Arman, "Pengembangan Model Prosodi Bahasa Indonesia dan Sistem Text to Speech Bahasa Indonesia." Disertasi, Institut Teknologi Bandung, 2004.

[7] A. A. Arman, Soemintapoera, Kudrat, A. S. Ahmad, T. R. Mengko, "Prosody Model for Indonesia Language," APCC 2001 Proceeding Tokyo, 2001.

[8] K. M. Alan, J. S. William, G. L. Eldon, and M. Ronald, "Pitch Countour Generation in Speech Synthesis, A Junction Grammar Approach," Signal Processing, 1997.

[9] Tritoasmoro, Iwan, I. & T., Suhartono, "Text To Speech Bahasa Indonesia Dengan Pembangkitan Prosodi Menggunakan Metoda Multirate Recurrent Neural Network," Jurnal Penelitian dan Pengembangan Telekomunikasi, vol. 12 No 2, pp 132-139, 2007.

[10] C. Vania and M. Adriani, "The effect of syllable and word stress on the quality of Indonesian HMM-based speech synthesis system," in 2011 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2011 Jakarta: 2011, pp. 413-417.

[11] L. Cheng-Yuan, H. Chien-Hung, and K. Chih-Chung, "A simple and effective pitch re-estimation method for rich prosody and speaking styles in HMM-based speech synthesis," 2012, pp. 286-290.

[12] C. Y. Yeh and S. H. Hwang, "Efficient text analyser with prosody generator-driven approach for Mandarin text-to-speech," Vision, Image and Signal Processing, IEE Proceedings -, vol. 152, no. 6, pp. 793-799, Dec.2005.

[13] Amran Halim, "Intonasi dalam hubungannya dengan sintaksis bahasa Indonesia," Penerbit Djambatan, Jakarta, 1984.

[14] Vincent, J., Heuven, van and Zanten, E.V., "Prosody in Indonesian Languages", Leiden University Centre for Linguistics, 2007.

[15] Wicaksono, Alfan Farizki dan Ayu Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," The 4th International Malindo Wokshop, Jakarta, 2010.

[16] D. Novianti, "Analisis Model Prosodi Pada Pesintesa Suara Bahasa Indonesia Menggunakan Speech filing System," Tesis, Universitas Gunadarma, 2009.

[17] Sakriani Sakti S. Sakti, S. Sakai, R. Isotani, H. Kawai, S. Nakamura, "Quality and Intelligibility Assessment of Indonesian HMM-Basaed Speech Synthesis System," in Proc. MALINDO, pp. 51-57, Jakarta, Indonesia, August 2010.

[18] H. Fujisaki and S. Ohno, "Prosodic parameterization of spoken Japanese based on a model of the generation process of F0 contours," 4 ed 1996, pp. 2439-3442.

[19] Y. Li, S. Pan, and J. Tao, "HMM-based expressive speech synthesis with a flexible Mandarin stress adaptation model," Qinghua Daxue Xuebao, vol. 51, no. 9, pp. 1171-1175, 2011.

[20] X. Gonzalvo, I. Iriondo, J. C. Socor+¦, F. Alias, and C. Monzo, "Mixing HMM-based spanish speech synthesis with a CBR for prosody estimation,", 4885 LNAI ed Paris: 2007, pp. 78-85.

[21] K. Tokuda, Z. Heiga, and A. W. Black, "An HMM-based speech synthesis system applied to English," 2002, pp. 227-230.