PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

HMM-based speech synthesis system with expressive Indonesian speech corpus

Elok ANGGRAYNI; Dhany ARIFIANTO

Institut Teknologi Sepuluh Nopmber, Indonesia

ABSTRACT

In this paper, we present a result of HMM-based speech synthesis system applied to Indonesian expressive speech scorpus. The purpose is to observe speech quality of synthesized speech, conversely. Firstly, we selected expressive Indonesian conversation from movie, novel, and drama transcript. We developed speech database based on phonetically balanced sentence set in which consist of 33 Indonesian phonemes and its IPA symbols and formed 655 sentences. Three expressive styles were applied, namely happiness, sadness, and anger. We hired four professional theater artist to utter the sentences. Segmentation and labeling was performed by manual to create transcription. Variation is given in kind of expressive style and training data amount. The expressive style-dependent decision trees achieve prosodic conversion. The objective and subjective evaluation process are also analyzed. In objective test is using MCD method earn the best score for happiness expressive style with score 4.2 in 82 training data. Then for sadness with score 5.13 in 81 training data and 5.18 for anger in 80 training data. Subjective test with Mean Opinion Score obtain naturalness for happiness, anger, and sadness with score 3.51, 3.38, and 3,0, respectively. The result shown that quality of the synthetic speech is high in term of naturalness.

Keywords: Expressive Indonesian speech scorpus, HMM, Speech synthesis

1. INTRODUCTION

Technology is developed to create tools which can help and provide convenience for people to do their daily activities. Along with the development of technology, people always want to improve the quality and practicality of tools. Therefore, we formed machines which can interact with humans. This technology is called human machine technology (1). Human machine technology aims to create a machine that has the ability to describe information which is spoken by humans, act in accordance with the information, and speak to complete information exchange. In other words, creating a machine with artificial intelligence that can interact with humans through voice. Research about this case is still done to get maximum results.

In Indonesia already developed speech recognition by using neural networks principle. Neural networks principle is based on the processing of an input by following a model, such as Hidden Markove Model (HMM). Input is processed to produce expected output. Furthermore, the training process is needed to identify input data quickly. On development of voice processing techniques, especially speech recognition and speech synthesis have an opportunity to establish natural interaction between machine and human.

In the beginning, input which can be accepted by computer only text input, for example on the internet or telephone. However, machine still can not recognize input of voice. Therefore, it is necessary to develop speech processing technology for input of voice. Until now, Indonesian speech database is still in three types, they are digit isolated, link, and very simple conversation words. Therefore, it is still necessary research and development of Indonesian speech database (2).

This paper describes the procedure of making Indonesian speech database through the recording process for purposes of training and testing of computational algorithms from the program that was created based on Hidden Markove Model (HMM). Sentences that are used to develop Indonesian speech database made based on the principle of phonetically balanced. Phonetically balanced will be achieved if database have the complete Indonesian phonemes. The Indonesian phoneme set consists of 33 phonemes (3). Furthermore, in this paper, we derive a new development of Indonesian natural speech synthesis database based on Hidden Markove Model (HMM).

2. HMM-BASED SPEECH SYNTHESIS SYSTEM APPLIED TO INDONESIAN

2.1 Development Expressive Indonesian Speech Corpus

Indonesian is the official language of Indonesia. Although Indonesian is used by about 263-million people in the world, it is classified as an under-resourced language. Indonesian consists of several components that are systematically arranged according to a certain pattern and form one unity. Indonesian linguistics study consists of several level, namely the level of phonology, morphology, sintax, and lexicon.

Speech corpus is a collection of voice database in the form of audio files and text transcription. The database for this research used Indonesian database. Speech corpus which is used in speech synthesis is the result of voice recording at soundproof room by using Indonesian sentences based on phonetically balanced. Expressive Indonesian speech corpus that are used in has amount to 1-11 words per sentence. Three expressive styles were applied, namely happiness, sadness, and anger. There are two kinds of sound, that are male voiced and female voiced as much as 655 sentences which has met 33 types of phonemes. 33 types of Indonesian phonemes which is used in database development is according to Indonesian phonemes and related english phonemes using International Phonetic Alphabet (IPA) (11).

2.2 Recording Process

Recording process is conducted in the recording studio at Surabaya, Indonesia. Studio has recording room (anechoic chamber) which separated from the operator room. In order to make good speech database for long term project, we hire four professional theater artist from Arts Council of Surabaya and Bandung as the speaker. Two males and two females who were recorded their voice are recorded on different times with same sentence.

The recording process was using microphone condenser (Neuman U87 and Studio Project B1) which be equipped by pop filter to make recording speech having clearer. Speaker asked to sit in front of the microphone with the distance of 2-3 cm. Microphone device is connected to Presonus Studio One and Pro Tools as software editing. This recording process use set up as Figure 1. Figure 2 is shown the recording process.

Figure 1 – Recording setup for male and female speaker

Expressive Indonesian speech corpus was recorded by total four speakers with two male speakers (MDPA and MBAZ) and two female speakers (FYAT and FCIM). FYAT and MDPA were recorded firstly in recording studio (Surabaya, Indonesia). FCIM and MBAZ were recorded in anechoic chamber of Institut Teknologi Bandung, Indonesia. Total duration of recording expressive indonesian speech corpus is 2 hours 53 minute 59 second with male voice for around 1 hours 24 minute 24 second and female voice for around 1 hours 29 minute 35 second. Recording process was under configuration with sampling frequency of 44,1 kHZ, channel input/output mono, 16 bits/sample, and using format ".wav".

2.3 Segmentation and Labeling

Segmentation and labeling are conducted on each recorded sentence. Segmentation and labeling were performed to create transcription of each individual phoneme. There are several methods to do segmentation and labeling such as manual labeling using wavesurfer and also auto labeling using festival, festvox, SphinxTrain, and the other tools. The segmentation and labeling process that used in this research use auto labeling by festival tool and festvox that developed through open source by Alan W. Black.

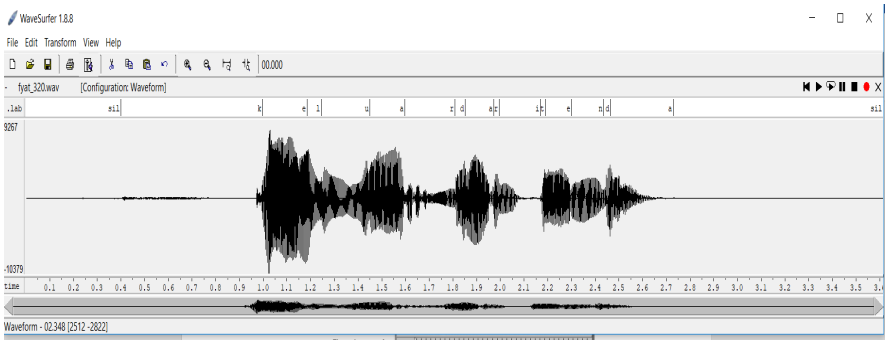


Figure 2 – Wavesurfur for segmentation and labeling process

The process of segmentation and labeling is done by using wavesurfur software. It is conducted by listening the voice and marking the location of each phoneme on the voice. Figure 2 is shown wavesurfur software for segmentation and labeling process. From 655 sentences which is recorded, 82 happiness expressive style, 81 sadness, and 80 anger of male and female voice is taken for samples minimum training. Segmentation and labeling results from 243 sentences is processed statistically and modeling to predict the level of compression sound energy for everyone. This model will be used to estimate level of impairment and in what frequency bands can decrease the level of compression energy. However, hand-labeling was taken so much efforts since the speech database is usually in large number. Otherwise, automatic-labelling will faster but less accuracy.

2.4 Extraction of Parameter F0, Mel-Cepstrum, Delta Cepstrum, and Delta-Delta Cepstrum

At HTS, vector output in HMM consists of two parts, namely spectrum and excitation. Part of the spectrum consists of mel-cepstral coefficients including coefficients to zero, delta and delta-delta coefficients. Furthermore, the excitation part consists of log fundamental frequency (log F0), delta and delta-delta. Regard to recordings result, it can be determined value of spectrum and excitation parameters.Here is the equation which is used to find dynamic features in HMM:

$$\Delta\, c_t = \frac{\partial c_t}{\partial_t} \approx 0.5(c_{t+1} - c_{t-1}) \tag{1}$$

$$\Delta^2\, c_t = \frac{\partial^2 c_t}{\partial_t{}^2} \approx c_{t+1} - 2c_t + c_{t-1} \tag{2}$$

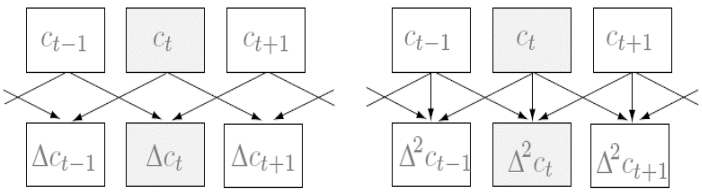


Figure 2 – Dynamic features (4)

2.5 Speech Synthesis by Using HMM-based speech synthesis system (HTS)

HMM-based speech synthesis system (HTS) is a speech parametric technique based on HMM. HMM-based speech synthesis system (HTS) was developed by HTS working group. The advantage

of HTS is able to model and synthesize the speaker's voice, style, and different emotions using only voice data quite a bit. Training part in HTS has been used as modified version of HTK and released as patch code form for Hidden Markov Toolkit (HTK). Patch code is released under free software license.

2.5.1 Training HMM-Based Speech Synthesis System (HTS)

In the training part, spectrum and excitation parameters is taken from speech database and modeled by context dependent HMMs. In the synthesis part, context dependent HMMs is changed in accordance with the text to be synthesized. Then, spectrum and excitation parameters is generated from HMM using voice algorithm parameter [8]. Excitation generation module and filter synthesis module synthesize voice waves which is using excitation and spectrum parameters produced. Extraction of approach is performed on voice characteristic from speech synthesized so that it can be easily changed by changing HMM parameters. However, this process shows that we can change the voice characteristics of speech synthesized by applying speaker adaptation technique, speaker interpolation technique, and Eigen voice technique.

2.5.2 Implementation of HTS on Festival Architecture for Expressive Indonesian Speech Corpus

Implementation process will give several different results in training models. Variation in the kind of training sentences aimed to identify the pattern of the change in the acoustics models parameter. While variations in the training data amount intended to determine the lower limit of the amount of training data that is needed to keep produce the natural synthesized speech. 1529 Indonesian speech database is used in the training process, so better acoustic models will be produced. This is because the distribution of phonemes in speech database will affect the probability of acoustic models formation. Otherwise, with the more and more speech database is used, it will take much time and greater computational.

Variations training of Indonesian speech database are given in the kind of sentences and in the training data amount. In the kind of sentences, variation is done with declarative sentences, question sentences, and the combination of declarative and question sentences. The variation applies to ten speakers. This arrangement can be seen in Table 1.

Table 1 –Variation of training data amount

Kind of training sentences	Minimum training amount	Maximum training amount	Synthesis sentences amount
Happiness	82	227	50
Anger	80	213	50
Sadness	81	215	50

3. EXPERIMENT RESULT

3.1 Statistic of Indonesian Speech Database

To develop Indonesian speech synthesis system based Hidden Markov Model (HMM) need speech database. Speech database appropriate with phonetically balanced concept. From 655 expressive Indonesian sentences that have been made, carried out a statistical process to determine how many consonant phonemes, single vowel, and double vowel in order to know whether the sentence database already meets with 33 types of phonemes. Below is the distribution of phonemes in database system: Figure 3 shown the phoneme distribution of expressive Indonesian speech corpus. From the figure, the largest is phonemes "a" with 4.316 phoneme and the smallest is "sy" with 23 phonemes.

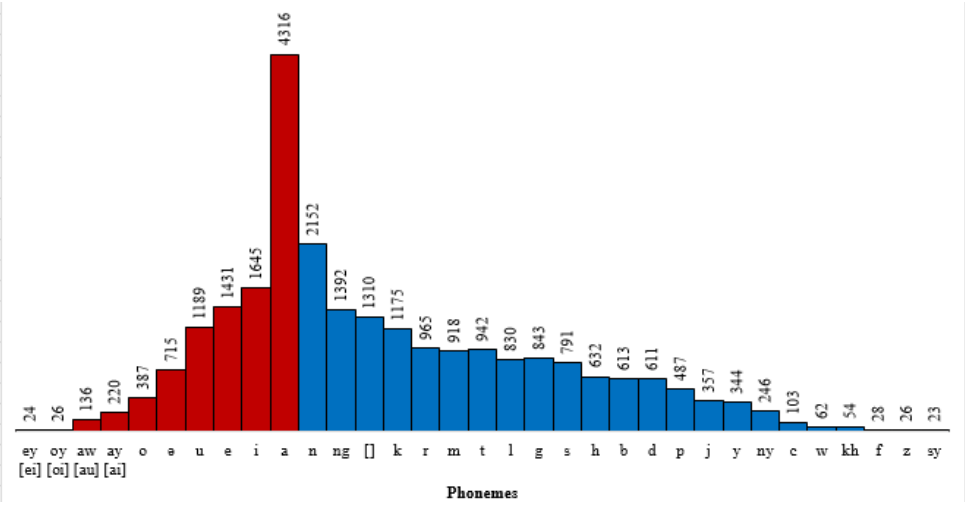


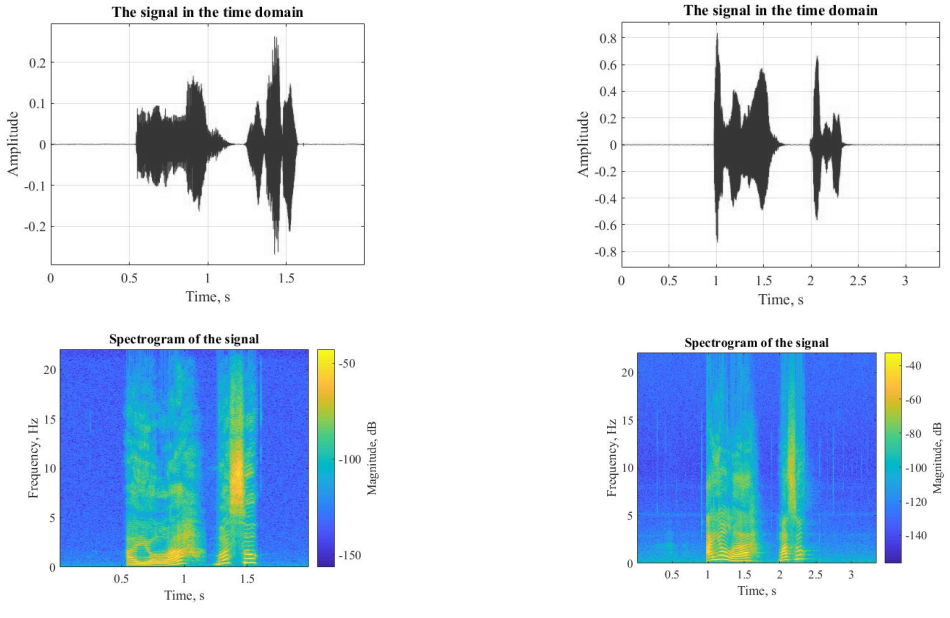
Figure 3 – Phonetically balanced of 655 sentences expressive Indonesian speech database

3.2 Results and Discussion

Recent developments in speech synthesis system is statistic of parametric speech synthesis system based on Hidden Markov Model (HMM). HMM-based speech synthesis system (HTS) was developed to overcome the problem of unit selection speech synthesis system that causes the process to be slow when performing synthesis for very large storage. HTS has ability to synthesize speech with high degree of naturalness that is comparable with speech synthesis unit selection system. This system was first proposed by Tokuda et al (11).

By using HTS implementation on festival architecture, the training process data is using 243 sentences from Indonesian speech synthesis database. Speech signal is sampled on frequency of 16 kHz. These factors were taken from speeches using extraction feature function from Festival speech synthesis system. Running time required core engine consists of 8 modules, decision trees for spectrum, F_0 and duration, spectral distribution, F_0 and duration, a converter that converts the features which have been extracted by Festival into sequence of labels context dependent and synthesizer which is generating waves to be given sequence label.

Figure 4 shows plot waveform and spectrogram of the recorded speech for four speakers. Waveform graphic describe speech signal in time domain and give information of amplitude and time. Spectrogram is used to know the signal representation in frequency domain.



(a) (b)

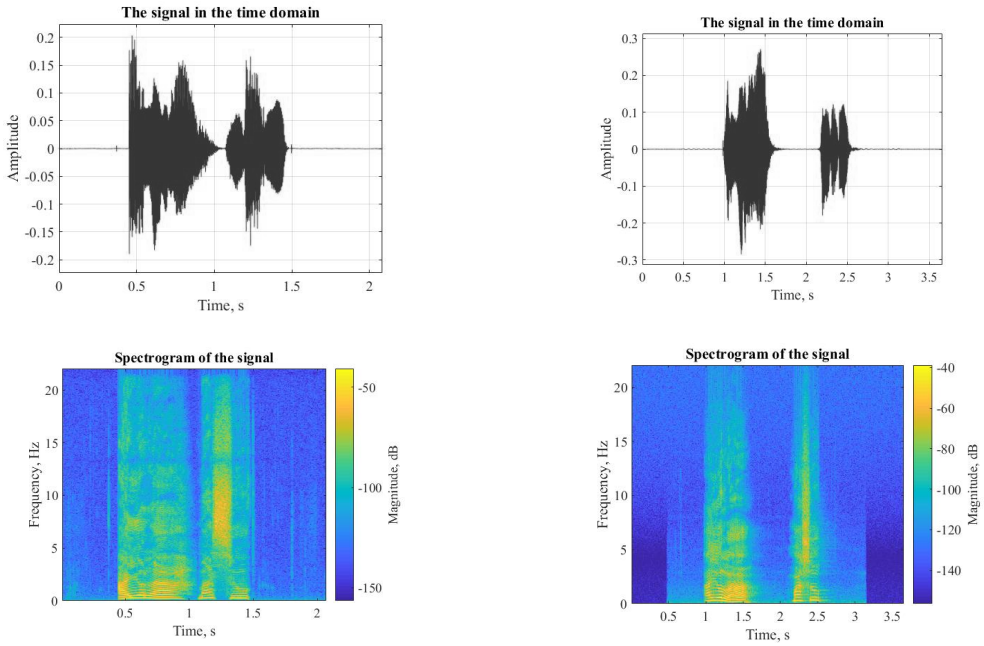


Figure 4 – Plot waveform and spectogram of sentence “ayolah, masuk!” (a) FCIM, (b) FYAT, (c) MBAZ and (d) MDPA

In Figure 5 can be seen the waveform of speech signal followed by comparison among fundamental frequency contour of the synthesized speech. Figure 5 is for female voice FYAT. From the F_0 track, can be seen that there is some distortion between the original speech and the synthesized speech. The distortion is quite big and it is the reason why the synthesized speech still has robotic sound and noise. From the female voice, the F_0 track has almost the same pattern with the original speech.

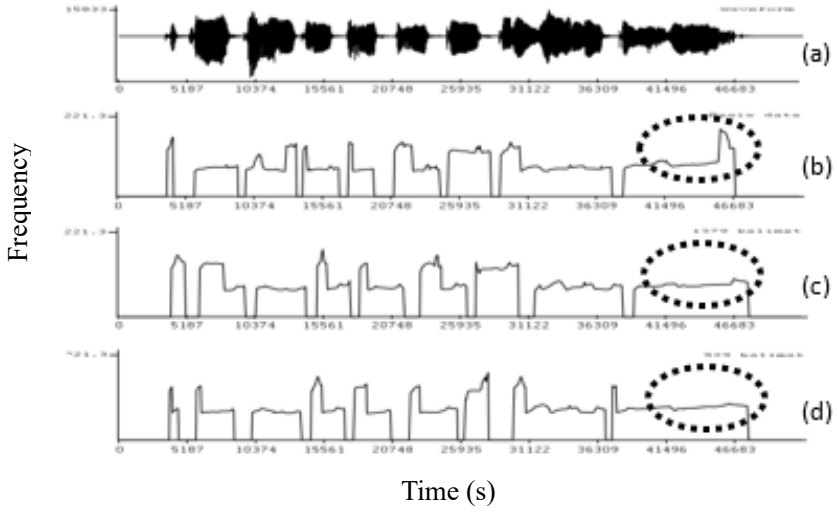


Figure 5 – (a) Plot waveform and spectogram of sentence “alangkah bahagianya dia bisa dekat dengan ayahnya” FYAT speaker (b) original speech, synthesized speech with (c) 82 sentences, (d) 227 sentences

To assess the quality of voice produced after the training process using HTS, tested with MOS (Mean Opinion Score) to see subjective assessment and the objective test using MCD method. In objective test is using MCD method earn the best score for happiness expressive style with score 4.2 in 82 training data. Then for sadness with score 5.13 in 81 training data and 5.18 for anger in 80

training data. Based on the results indicate that the speech quality of synthesized speech is still not enough. Based on these data, can be concluded that the distortion of mel-cepstral will be smaller as the higher number of database which being used. That is because of the more probabilities of the appearance phonemes when using the maximum training data.

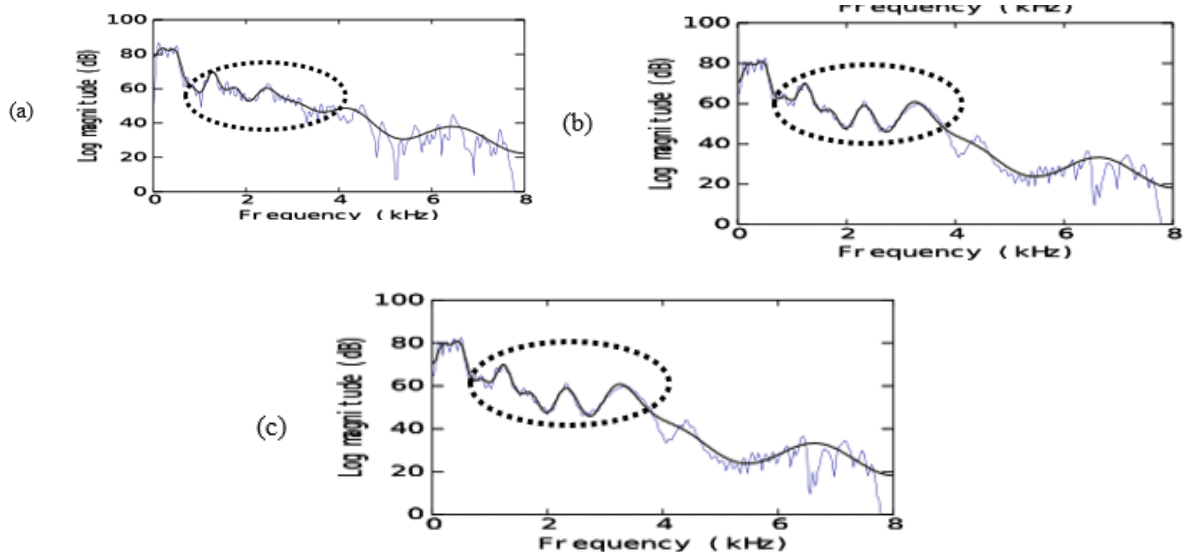


Figure 6 – Mel-cepstrum analysis of female synthesized speech with sentence “ayo bergegas Daeng, kau sudah baikan bukan?” (a) original speech, synthesized speech with (b) 82 sentences, (c) 227 sentences

Subjective test with Mean Opinion Score obtain naturalness for happiness, anger, and sadness with score 3.51, 3.38, and 3,0, respectively. We can see based on test results of voice quality using MOS (Mean Opinion Score) method for testing the results of training expressive Indonesian speech corpus showed that the voice quality is sufficient. Therefore, the resulting synthetic speech can be categorized as good and software can be used to design Indonesian natural speech synthesis.

4. CONCLUSION

In this paper, we develop expressive Indonesian speech database which contain of 655 sentences in accordance with phonetically balanced, which has met 33 types of phonemes. Total duration for four speakers is 2 hours 53 minute 59 second with male voice for around 1 hours 24 minute 24 second and female voice for around 1 hours 29 minute 35 second. The process of segmentation and labeling is done by using wavesurfer software.

The synthesized speech of Indonesian HMM-based speech synthesis system with manual segmentation and labeling is evaluated with MOS (Mean Opinion Score) to see subjective assessment. Based on the test results of voice quality using subjective test, involving 20 respondents. Subjective test with Mean Opinion Score obtain naturalness for happiness, anger, and sadness with score 3.51, 3.38, and 3,0, respectively. In objective test is using MCD method earn the best score for happiness expressive style with score 4.2 in 82 training data. Then for sadness with score 5.13 in 81 training data and 5.18 for anger in 80 training data.Therefore, the resulting synthetic speech can be categorized as good and software can be used to design Indonesian natural speech synthesis.

In the next work, we try to build more voice with different speakers. Furthermore, we try to increase the amount segmentation and labeling dataset by hand-labeling to increase quality of synthesized speech.

ACKNOWLEDGEMENTS

Authors would like to thank for Indonesian Government World Class University Program who support this research under grant No. 1022/PKS/ITS/2019, and also thank for the speakers, Wins recording studio, and Institut Teknologi Bandung who take part in recording process of Indonesian speech corpus.

REFERENCES

1. Tolba, hesham and Douglas O'Shaughnessy, “Speech Recognition by Intelligent Machines”, IEEE Press, 2001.
2. Lestari, Dessi Puji, Nonmember and Sadaoki FURUI, and Fellow, Honorary Member, IEICE TRANS. INF & SYST., VOL.E93-D, NO.9 SEPTEMBER 2010.
3. Suyanto, “An Indonesian Phonetically Balanced Sentence Set for Collecting Speech Database”, Jurnal Teknologi Industri Vol. XI No. 1 Januari 2007: 59-68.
4. Tokuda, Keiichi and Heiga Zen, “Fundamentals and Recent Advances in HMM-Based Speech Synthesis, Nagoya Institute of Technology: Toshiba Research Europe, 2009.
5. A. W. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis”, Proc. EUROSPEECH, pp.601– 604, Sep 1997.
6. Dey, Subhrakanti, “Reduced-Complexity Filtering for Partially Observed Nearly Completely Decomposable Markov Chains”, IEEE Transactions on Signal Processing, Vol. 48, No. 12, Desember, 2000.
7. Dugad, R., and Desai UB., “A Tutorial on Hidden Markov Models”, India: Technical Report, Department of Electrical Engineering, Indian Institute of Technology-Bombay.
8. Evans, S. Jamie and Vikram Krishnamurthy, “HMM State Estimation with Randomly Delayed Observation”, IEEE Transactions on Signal Processing, Vol. 47, No. 8, Agustus, 1999.
9. Fari, Guoliang dan Xia Xiang Gen, “Improved Hidden Markov Models in the Wavelet-Domain”, IEEE Transactions on Signal Processing, Vol. 49, No. 1, Januari, 2001.
10. Fukada, T., K. Tokuda, T. Kobayashi and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” Proc. of ICASSP’92, vol.1, pp.137–140, 1992.
11. Tokuda, K. T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling,” Proc. of ICASSP, 1999.