Under-Resourced Natural Bahasa Indonesia HMM-based Text-To-Speech System

Elok Cahyaningtyas[1], Dhany Arifianto[2]
1 Master student, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
2 Associate Professor, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

インドネシアは多民族・多言語国家だが、なかでも世界中で2億6000万人以上が使っている言語がバハサ（Bahasa）。しかし、その言語学的な解析は進んでいない。本稿ではHMMをベースに、この知られざる言語の構造解析に取り組んだ。

Abstract

Although Bahasa Indonesia is used by about 263 million people in the world, it is classified as an under-resourced language. In this paper, we outlined the development of casual sentences of Bahasa Indonesia speech corpus which contains speech database and its transcription. Firstly, we selected casual Bahasa Indonesia sentences from movie and drama transcript and formed 1029 declarative sentences and 500 question sentences. We hired six professional radio news readers to utter the sentences to avoid local dialect in a sound-proof booth. Segmentation and labeling were performed to create transcription including the time label of each individual phoneme. Then, we conducted some experiment to develop text-to-speech (TTS) system in Bahasa Indonesia. We do some variation in the number of sentences and the type of sentences which used in the training part. We use 44, 72, 116, 450, 929 and 1379 training data sentences based on the phonetically balance. The goal is to know the speech quality of Bahasa Indonesia TTS system. Besides that, we also compare the method to build the TTS system, which is using HMM-based text-to-speech system (HTS) and CLUSTERGEN (CLS). In the on-going research, we are developing high quality TTS, namely speaker adaptation and speaker averaging.

Keywords Bahasa Indonesia; under-resourced language; speech corpus; TTS; HMM-based speech synthesis

Introduction

Speech synthesis technique has been developed recently. The unit-selection synthesis is a speech synthesis technique which uses database. In this technique, the sub-word unit will be selected automatically from database given [14]. This technique is able to produce synthesized speech which similar as the original speech from the database. However, this technique requires a lot of database to obtain comprehensive data coverage to build the models. So it makes this technique require huge computing load and lacks the flexibility to be modified.

In 1999, Yoshimura, et al., explained the method of modelling the spectral parameter, excitation parameter and duration simultaneously [15]. Then they sparked a speech synthesis technique based on statistical process known as statistical parametric speech synthesis that then began to grow today [15]; [11]; [16]. This technique uses hidden markov model (HMM) to model the probability distribution of speech and linguistic feature. It is called HMM-based speech synthesis system (HTS). Formation of statistical models gives HTS an advantage in flexibility to modify the acoustic models. Some of the advantages that can make the transformation of character voices, speaking styles, speaking adaptation, and supports multilingual speech synthesis.

HTS has evolved in some countries such as Japan [15],

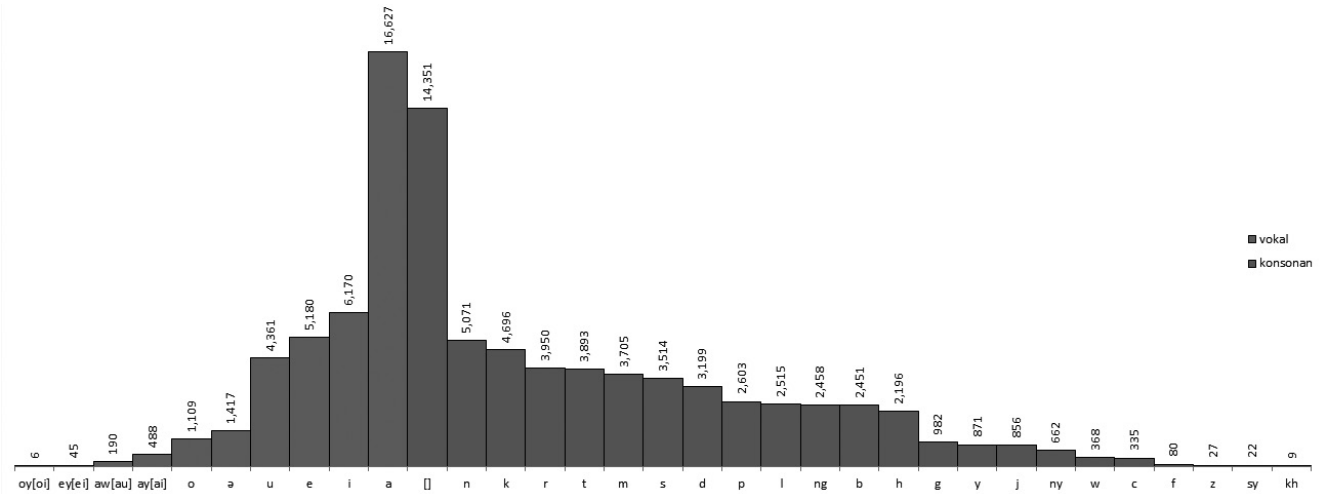

Fig. 1 Phonetical balance of 1529 sentences Bahasa Indonesia speech database [5]

England [3], Chinese (Zen, et al., 2003; Wu and Wang, 2006), Thai (Chomphan and Kobayashi, 2007), Vietnam (Liang, et al., 2008) and other countries. In addition, some modifications for HTS are in style adaptation techniques for speech synthesis using HSMM and features suprasegmental [17], implementation of the algorithm MLLR to sound adaptation of databases bit [19]. While its application in Bahasa Indonesia is still lacking because Bahasa Indonesia is still classified as under-resourced language.

In this paper, we conducted some experiment to develop the speech synthesis system in Bahasa Indonesia. Then we do some variation to know the speech quality and characteristics of Bahasa Indonesia speech synthesis system. Besides that, we also try to compare the method to build the speech synthesis system, which are using HMM-based text to speech system (HTS) and CLUSTERGEN (CLS). To build the speech synthesis system, first we created the Bahasa Indonesia speech corpus [5]. Then applied the HTS demo in Bahasa Indonesia which applied in declarative and question sentences [12]. The implementation of statistical parametric speech synthesis in Bahasa Indonesia by using CLUSTERGEN [8]. Then compare the speech quality of the synthesized speech using subjective and objective measurement [13].

Characteristics of Bahasa Indonesia

Language is an expression of human mind and feeling which using sound as its tool [2]. Every country has a different language with their own characteristics. Bahasa Indonesia is the national language of Indonesia and rooted from the Malay language. Besides Bahasa Indonesia as the main language, most of Indonesians are fluent in their own ethnic language according to the location of their tribe. Some of the ethnic languages such as Javanese, Sundanese, Maduranese, etc. As of 2010 census, Indonesia has 1,340 tribes with each different ethnic language and normally would not be able to understand each other. Then, Bahasa Indonesia is used as national language in order to bridge and bind the Indonesian people together. Bahasa Indonesia has spoken and written system. The spoken system similar to Malay and the written system is referred to Roman alphabet system.

Linguistic studies of Bahasa Indonesia divided in some level, i.e., phonology, morphology, syntax, and lexicon [2]. Phonology explain on how sound produced and its distribution. It is divided in some term i.e. phonetic, phonemic, segmental and suprasegmental sound. Phonetic is a linguistic study of the physical sounds of human speech production. While phonemic is a linguistic study of phonemes and their written representation as the meaning differentiator. A phoneme is the smallest unit of sound which composing a word or phrase. Phonemes is an important role in NLP. Bahasa Indonesia has 32 phonemes and contains of six vocal phonemes, three diphthong phonemes, and 23 consonant phonemes. Table 1 shows the Bahasa Indonesia phonemes based on International Phonetic Alphabet (IPA) excluding a silence character.

Indonesian speech database is the datasets of Indonesian language characteristic in accordance to Indonesian phonology. It

consists of phoneme, speech, and transcription. The database contains of 1529 sentences with 1029 of declarative sentences and 500 of question sentences. The sentences sequence is formed from some literature such as novel, book, newspaper and internet which using Indonesian language. Figure 1 shown the phoneme distribution of Indonesian speech database. From the figure, the largest is phonemes "a" with 16.627 phoneme and the smallest is "oi" with 6 phonemes [5].

No	Indonesian	English	Example
1.	/a/	aa	Father
2.	/e/	ah, ae	Ten
3.	/ê/	ah, ax	Learn
4.	/i/	ih, iy, ix	see, happy
5.	/o/	ow, ao	got, saw
6.	/u/	uh, uw	put, too
7.	/ay/	Ay	Five
8.	/aw/	Aw	Now
9.	/ey/	Ey	Say
10.	/oy/	Oy	Boy
11.	/b/	B	Bad
12.	/c/	Ch	Chain
13.	/d/	d, dx, dh	Did
14.	/f/	f, v	fall, van
15.	/g/	G	Got
16.	/h/	Hh	Hat
17.	/j/	Jh	Jam
18.	/k/	k	Keep
19.	/m/	m	Man
20.	/l/	l	Leg
21.	/N/	n	no
22.	/P/	p	pen
23.	/R/	r	red
24.	/S/	s	so
25.	/T/	t,th	tea
26.	/W/	w	wet
27.	/Y/	y	yes
28.	/Z/	z, zh	zoo
29.	/Kh/	—	—
30.	/Ng/	ng	sing
31.	/Ny/	—	—
32.	/Sy/	—	share

Table 1. Indonesian Phonemes based on International Phonetic Alphabet (IPA)

Bahasa Indonesia speech database was recorded by total six speakers with three male speakers (MMHT, MJRA, MEIA) and three female speakers (FENA, FBAP, FALA). Profile of the speakers is shown in Table 2. The two speakers (MMHT and FENA) were recorded firstly in Japan and the others are recorded in Surabaya, Indonesia. The recording process spent approximately 8-10 hours each speaker. The recorded speech duration is 2-5 second for short sentences and 6-9 second for long sentences. The total duration of all recorded Bahasa Indonesia speech database is 10.65 hours with the male voice for around 5.5 hours and for the female voice for around 5.2 hours. The recorded speech was under configuration with the sampling frequency of 44,1 kHz, channel input/output mono, 16 bits/sample and using ".wav" format.

Speaker	Gender	Age	Profession	Length
MMHT	Male	44	Professional announcer	1 h 43 m 50 s
MJRA	Male	22	Professional announcer	1 h 50 m 34 s
MEIA	Male	32	Professional announcer	1 h 52 m 45 s
FENA	Female	26	Professional announcer	1 h 36 m 27 s
FBAP	Female	20	Professional announcer	1 h 44 m 56 s
FALA	Female	21	Professional announcer	1 h 50 m 33 s
Total Duration				10 h 39 m 5 s

Table 2. Profile of Bahasa Indonesia Speech Corpus's Speaker

Statistical Parametric Synthesis

Statistical parametric synthesis expressed the handicraft of expert from rule based model by statistical model. HMM-based Text to Speech System (HTS) is one of statistical parametric synthesis technique which widely known. In the HMM-based speech synthesis, the speech parameters of a speech unit such as fundamental frequency, phoneme duration and spectrum are statistically modeled and generated by using HMMs based on maximum likelihood criterion [4].

A. HMM-based Text-to-Speech System (HTS)

The HMM-based speech synthesis system consists of two main process, that are training and synthesis part which shown in Figure 2 [15]. In the training part, the HMM model represents the excitation source, i.e., F0, the spectrum, and state duration of the context-dependent speech units. Each HMM model has left-to-right state transition with no skip. Acoustic model in HTS is built from the application of maximum likelihood probabilistic equations in the training process (1) and in the synthesis process (2). The optimal model parameter can obtain with maximizing the likelihood of the training data which given in the following equation,

$$\hat{\lambda} = \arg\max_{\lambda} P(O|T,\lambda) \quad (1)$$

where $\hat{\lambda}$ is the model parameter estimation, O is the training data, T is a word derived from the label (transcription) and λ is a model parameter.

$$\hat{O} = \arg\max_{o} P(O|t,\hat{\lambda}) \quad (2)$$

where $\hat{O}$ is an estimation model speech, o is the speech parameter, t is the word to be synthesized which derived from the phrase labels, and $\hat{\lambda}$ is the estimation model [15].

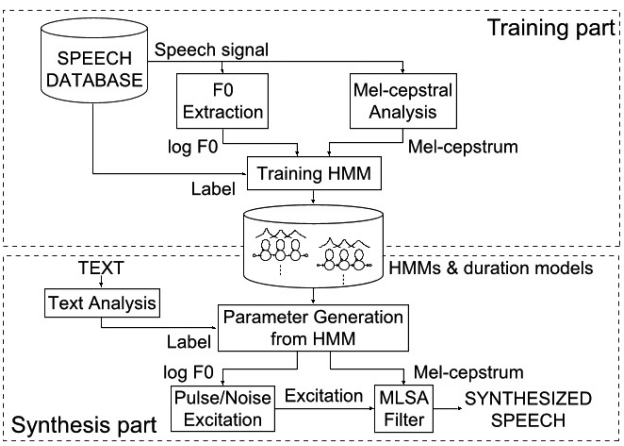

Fig. 2. HMM-based Text-to-Speech System (HTS) [3]

The synthesis part has the inverse operation of speech recognition system. The input system is contextual label sequence of the text which using the same format but different text from the training part. From the context-dependent label of the given text, then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label. After that, the sequence of spectral and excitation parameter is generated by the speech parameter generation algorithm that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the mel log-spectrum approximation (MLSA) filter.

B. CLUSTERGEN

CLUSTERGEN is a method to build synthetic speech with trajectory model. The different way of CLUSTERGEN than HTS model is in trajectory modeling, a setup experiment was set and show trajola, a model trajectory with overlap and add, which is better than other kinds of trajectories model have been build [1]. CLUSTERGEN method predicts vector output in three ways that are previous, current and next. This method is used to get better vector output prediction.

$$S'_t = \frac{S_t - 1 + S_t + S_t + 1}{3} \quad (3)$$

In equation (3), "s" is a set vector arranged in every word which has been training using HMM. This method actually same with decision tree in HTS, but the selection of the acoustic feature for every phoneme is by considering the previous, current and next phoneme for grouping.

The CLUSTERGEN was included in FesVox [http://festvox.org/] build. The newer version of FestVox not only included CLUSTERGEN but also has been included STRAIGHT [9] and moving segment label [2] technique too. STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) is a procedure to manipulate speech signal based on pitch adaptive spectral smoothing and instantaneous-frequency-based F0 extraction.

Experiments

In this paper, we conducted some experiment to build speech synthesis system by using HMM-based speech synthesis system demo for speaker dependent (HTS-demo-CMU-ARCTIC-SLT). The demo program works with some following software, i.e., SPTK, HMM Toolkit (HTK), HDcode, HTS-2.2, hts-engine API-1.05, festival, ActiveTcl and speech tools. All of them is an open source programs on Linux. The HTS demo is available in English. Then we adapting into Bahasa Indonesia with some modification, that are in the speech corpus which contain of speech unit and its context-label, and the question file to build the decision tree according to the phoneme rule of Bahasa Indonesia [12]. All of them will be used for training part to build the parameter generation of HMM model, then it will be used in synthesis part to generate the speech waveform by Mel log spectral approximation (MLSA) filter.

In this section we will describe our experiment to build synthesized speech of Bahasa Indonesia using HMM-based speech synthesis system. These experiments consist of some variation,

first is variation in the number of training sentences, second is variation in the type of sentence using for training process, and third is some comparison of HTS and CLUSTERGEN method.

A. Variation in the Number of Training Sentences

The first experiment is making variation in the number of speech corpus which used in the training part. We are using minimum, maximum and combination number of speech corpus. Such variations are made according to the number of sentences. Declarative sentence has total of 1029 sentences and question sentence has a total 500 sentences. Then we separate the 100 sentences from declarative sentence and 50 sentences from question sentences to be used as synthesized sentences. So we have total 929 and 450 sentences for declarative and question sentences, respectively. This total number of sentences we used as maximum training. While for the minimum training, we construct sentences using the least number of phoneme according to the phonetically balanced of maximum training. So, in the minimum training we have 72 and 44 sentences for declarative and question sentences, respectively. In addition, we also using combination of declarative and question sentences, which was formed from the combination of declarative and question sentences. Thus, obtained combination sentences for minimum and maximum training as many as 116 and 1379 sentences, respectively. The number of speech corpus used in the training can be seen in Table 3. The variation applied to both speaker, mmht and fena.

Training Sentence	Synthesis Sentence	Maximum Training Number	Minimum Training Number	Synthesis Sentences Number
Question	Question	450	44	50
Declarative		929	72	50
Combination		1329	116	50
Question	Declarative	450	44	100
Declarative		929	72	100
Combination		1329	116	100

Table 3. Variation of Training Data Number

Number of Training Data	Time(hours)			
	Question Sentence		Declarative Sentence	
	mmht	*fena*	*mmht*	*fena*
44	1:26:40	0:54:39	0:41:44	0:44:50
72	2:06:26	1:16:33	1:05:16	1:00:05
116	1:25:49	1:25:49	1:16:15	1:11:17
450	2:30:45	2:45:17	2:47:22	2:47:05
929	6:19:38	5:44:14	7:48:42	6:52:31
1329	9:05:05	9:15:15	9:05:05	9:15:15

Table 4. HTS Computation Time

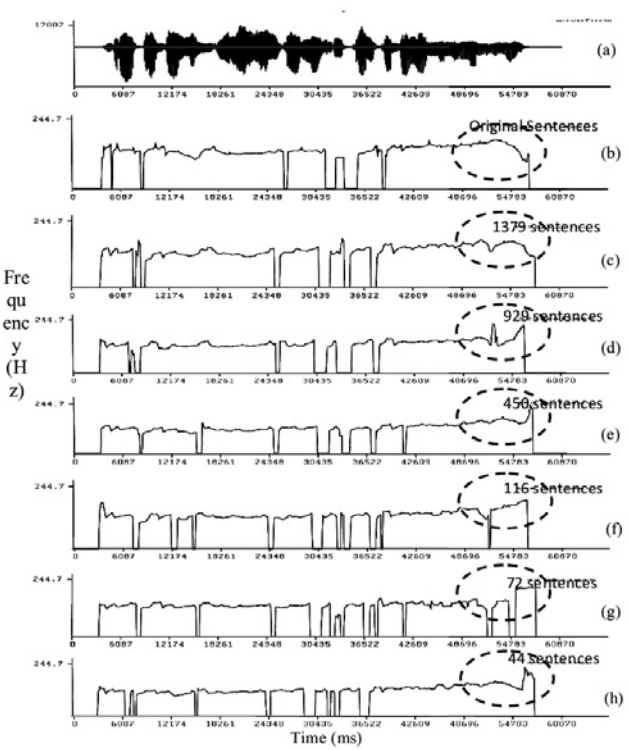

Fig. 3. F_0 Plot of Question Sentence "*Berapa banyak gula yang kau masukkan ke dalam minuman ini?*"

The different training process will give different result in the training models. While variations in the training data number aimed to determine the lower limit of training data to keep produce the natural synthesized speech. The more speech corpus using in the training process, the better acoustic models will be produced. It is because the distribution of phonemes in speech corpus will affect the probability of acoustic model formation. However, with the more speech corpus used in the training part, it will take much computation load and time. In Table 4, shown the computation time while running the HTS demo for Bahasa Indonesia. The computation time varies from the minimum,

maximum and combination sentences for both declarative and question sentences. It shows that the increasing number of training sentences, make the computation time longer.

The synthesis process is proceeded after the formation of model training is completed. The step is to combine the acoustics and linguistic features that have been formed in the training part to be desired synthesized speech. From several variations given, it has different synthesized speech quality which located in the level of naturalness. The combination training sentences produced better synthesized speech than maximum and minimum training sentences. It can be seen from the comparison of fundamental frequency plot (excitation parameter) and mel-cepstral plot (spectral parameter) of speech signal.

The fundamental frequency track show how the pitch of speech signal that show an intonation aspect in a sentence change in time. In Fig.2 show F0 plot of question sentence *"Berapa banyak gula yang kau masukkan ke dalam minuman ini?"* in Bahasa Indonesia, if translated in English become "How much sugar you add in this drink?". From that figure can be seen the waveform of speech signal followed by comparison among fundamental frequency (F0) contour of the synthesized speech and original speech. From the F0 contour can be identified the voiced, unvoiced and silence region [6]. Through the dotted line, can be seen the difference of F0 from each synthesized speech.

The F0 extraction is using *pitch* tools from speech signal processing toolkit (SPTK). It done with condition of sampling frequency in 16000 Hz, frame period 80 point (5 ms), minimum F0 80 Hz and maximum F0 165 Hz. Then using fdrw tools to plot the graph. This pitch extraction results still have some lacks in the F0 extractor. They are the F0 contour shape that obtained is not smooth and have many leaps on its surface. That is because the sound is regarded as noise by the extractor, a voiced region which is considered as unvoiced region, and vice versa. Else is because of pitch halving and pitch doubling.

Aside from fundamental frequency plot for extraction parameter, we can see the spectral parameter by using mel-cepstral plot. It can be obtained by converting the speech signal from the time domain to the frequency domain in logarithmic scale (log FFT). MCEP plot for synthesized speech of mmht with question sentence *"Berapa banyak gula yang kau masukkan ke dalam minuman ini?"* can be seen in in Fig.4. Plot MCEP obtained by us sing Speech Signal Processing Toolkit (SPTK) tools *mcep* with condition of sampling frequency 16000 Hz, frame length 400 points (25 ms), frame period 80 points (5 ms), analysis of order 20, frequency warping parameter FFT size of 0.42 and 512 points, then stored in the file *.mcep*. Afterward plotted in MCEP graph using tools *glogsp* and *mgc2sp*.

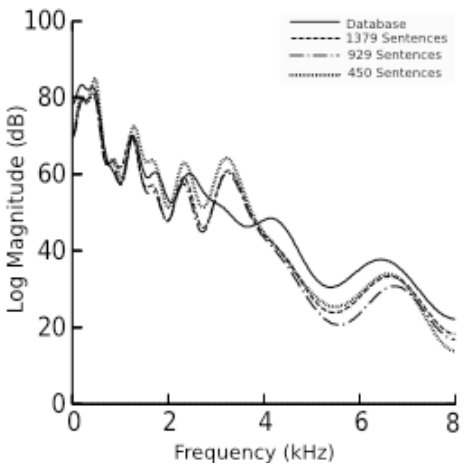

Fig. 4. Mel-cepstral Plot of Question Sentence *"Berapa banyak gula yang kau masukkan ke dalam minuman ini?"*

Mel-cepstrum plot has two main information, that are cepstral and cepstral envelope. Both of which will provide information of location, magnitude, and the characteristics of speech signal including duration, formant frequency (F1-F5), delta (the speed of speech, derived from the difference between the cepstrum peaks), delta-delta (the speech acceleration, derived from derivative of delta cepstrum).

For voiced speech at cepstrum plot will have more energy at lower frequency and have lower energy at high frequency (cepstral tilt). Whereas for the unvoiced speech will have the energy that is almost evenly on each frequency. When compared based on MCEP plot of each synthesized speech both for mmht or fena, and for declarative or question sentences, it appears that between original speech and synthesized speech with maximum training data have less distortion than synthesized speech with minimum training data. This is because of the increasing number of training data that be used, the more acoustics model will be generated. So the probability of the system to generate synthesized speech will be even greater by maximizing the acoustics model of speech which will be synthesized with the acoustics model that generated in training process.

B. Variation in Type of Sentences

The second experiment is to make training using variation in the type of sentences, which are declarative and question sentences. The scheme is we do training using declarative sentences then make the synthesized speech for question sentences, and vice versa. The goal is to see the changes of declarative sentences into question sentence, and vice versa. In the end of question sentence is followed by the rising intonation. While in declarative sentence has flat and decrease intonation in the end of the sentence.

Type of Training Sentence	Number of Training Data	Question		Declarative	
		mmht	*fena*	*mmht*	*fena*
Question	44	100%	100%	40%	20%
Declarative	72	0%	0%	80%	40%
Combination	116	20%	40%	40%	20%
Question	450	100%	100%	0%	20%
Declarative	929	0%	0%	100%	60%
Combination	1329	80%	20%	100%	100%

Table 5. Synthesized Speech Identification

For this purpose, we try to conduct synthesized speech identification with using subjective test evaluation. The evaluation is done with total 20 respondents of male and female whose have healthy hearing. Respondent will be heard the result of synthesized speech randomly, then will try to guess whether the synthesized speech categorized as declarative or question sentence. The result of the test is shown in the Table 5. In the table can be seen the identification result of question and declarative sentences separately both for minimum and maximum training. For the question sentences identification, the respondents were able to identify overall by 50% for mmht voice and 43% for fena voice.

So, to produce synthesized speech with the same training sentences (question to question sentences or declarative to declarative sentences) have higher percentage compared to synthesized speech with combination training sentences. However, the formation of declarative synthesized speech was able to produce transformation to question sentences with a small percentage of between 20 - 40%. But it cannot be achieved from question to declarative synthesized speech. This is because of the original speech that used has different characteristic with synthesized speech. Therefore, it should be given an additional parameter to be able to change the synthesized speech from question to declarative sentences, respectively.

In addition, while looking at the excitation parameter by fundamental frequency shown in Fig. 3 can be seen the difference. The figure shows the variation in training number and also in the type of training sentences to synthesized question sentence. When compared with the original speech, the best result is provided by 1379 and 450 sentences, which have almost the similar pattern. It means that the declarative sentence still not able to produce question sentences, and only can produce better with question and combination sentence.

C. HTS vs CLUSTERGEN

The third experiment aims to compare the HTS and CLUSTERGEN method [13]. In TTS system with CLUSTERGEN for Indonesian language was built by Evan [8], using EHMM [9] as a technique for obtaining label files from each database, it creates label from estimate phoneme based on fully connected state models and forward connected state model of HMM. This technique is shown a number of log likelihood better than technique is use 5 state sequence of HMM. Training part is done with CLUSTERGEN method, this method essentially contains some part, the first step is an extraction of F0 from the audio file in the database with Speech Tools [11]. Then, the next step is combines 24 MFCCs with F0 which have been extracted, as the result is given 25 vectors for every 5ms [11]. The last part of training process is clustering the MFCCs from every sample, this part is used wagon tool which contains in Edinburg Speech Tools CART tree builder [11]. The result of training part is to obtain model parameter which used in synthesis part [8].

The spectral parameter for synthesized speech male and female voice with HTS and CLUSTERGEN shown in Fig. 5. The full line represented the original speech, the dash-line is the synthesized speech of CLUSTERGEN, while the dotted-line is the result of HTS synthesized speech. From the spectral plot, we can see in the male voice shows that the lower frequency of synthesized speech is having almost the similar pattern with the original speech. The CLUSTERGEN give better result than the HTS. While in the female voice do not have significant differences.

In Fig. 6 and Fig 7. can be seen the waveform of speech

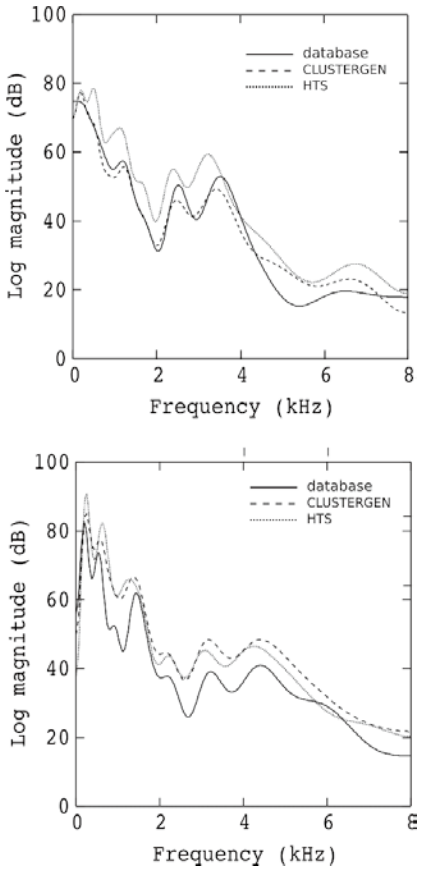

Fig. 5. MCEP plot of male (a), female (b) synthesized speech with HTS and CLS

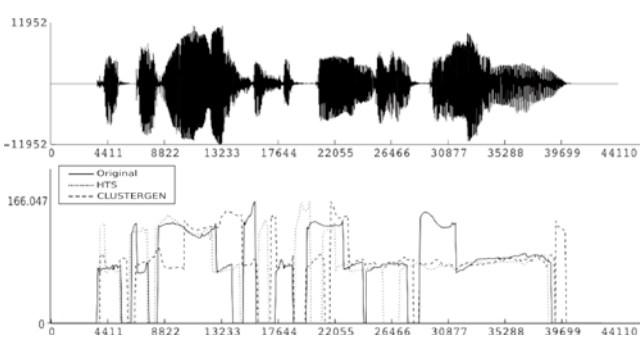

Fig. 6. F_0 Plot of female synthesized speech with HTS and CLS

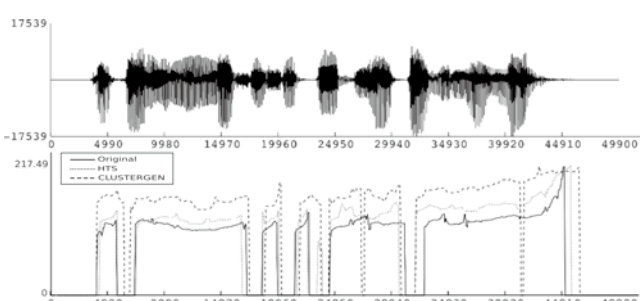

Fig. 7. F_0 Plot of male synthesized speech with HTS and CLS

signal followed by comparison among fundamental frequency contour of the synthesized speech. Fig. 6 is for male voice mmht, and Fig. 7 is for female voice fena. The full line represented the original speech, the dash-line is the synthesized speech of CLUSTERGEN, while the dotted-line is the result of HTS synthesized speech. From the F0 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 male voice, the F0 track has almost the same pattern with the original speech both for the HTS and CLUSTERGEN, but the HTS give the better result. While the result of female voice is far from the original speech for both methods.

Evaluation and Discussion

In this paper, we are using two kinds of test to measure the quality of synthesized speech. First is using objective test, which using mel-cepstral distortion (MCD) method. Second is using subjective test with degradation mean opinion score (DMOS) method.

The objective test is intended to assess the speech quality of the synthesized speech by analyzing mel-cepstrum distortion value from the original speech. The smaller MCD value indicate the closer synthesized speech to produce the natural speech. Fig. 8 and Fig. 9 is the objective test result of synthesized speech for male voice and female voice, respectively.

Based on the results indicate that the speech quality of synthesized speech is still not enough. The smallest distortion value on mmht voice for question sentence is on 450 training data with score 4.32 and for declaration sentence have 5.13 score with 929 training data. 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.

Then the objective result for the comparison of HTS and CLUSTERGEN is shown in Fig. 10. From the graphic, the synthesized speech with CLS give better result than HTS for the male voice, while for female voice the HTS produce better synthesized speech. From the result, we can see that the speech quality is still not enough to produce natural voice, it because

the distortion of the original and the synthesized speech is too big. It probably caused from extraction feature process which not perfect.

The subjective test aimed to measure the naturalness of synthesized speech by using DMOS method. It consists of two parts for each session that are training part and test part. The training part intend to familiarize the respondent to assess but not include in the assessment. Then the test part is a section that will be used as the assessment.

Fig.11 and Fig.12 show the result of subjective test for mmht voice and fena voice, respectively. From this graphic, can be seen that the highest value for mmht voice in declarative and question sentences are obtained with 1379 data training with score 3.53 / 5.00 and 3.36 / 5.00, respectively. Then for fena voice, the highest value for question sentence is obtained in 450 training data with score 2.98 / 5.00 and for declarative sentence is obtained in 1379 training data with score 3.04 / 5.00. That score means that "degradation speech is slightly annoying".

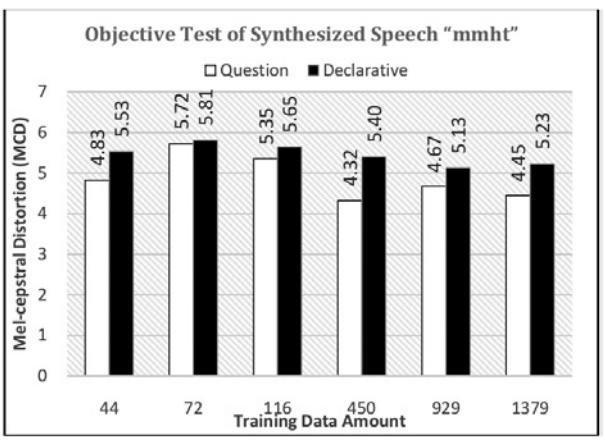

Fig. 8. Objective Test of Synthesized Speech of Male Voice

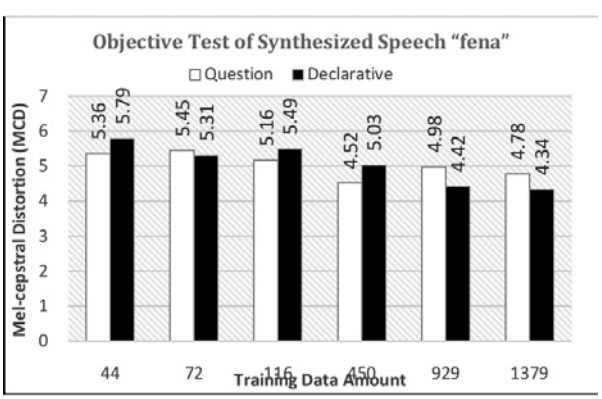

Fig. 9. Objective Test of Synthesized Speech of Female Voice

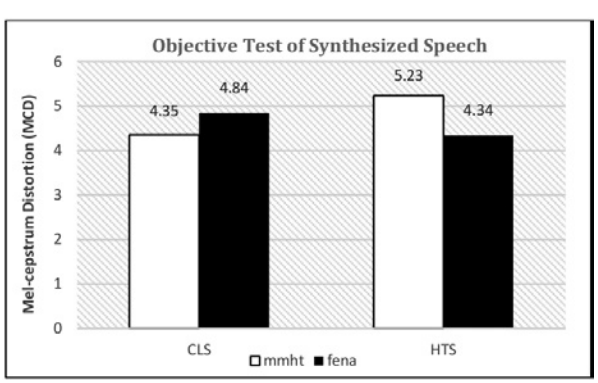

Fig. 10. Objective Test of Synthesized speech with HTS and CLS

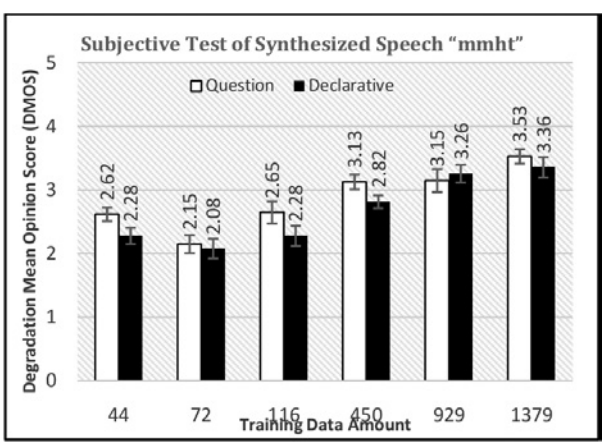

Fig. 11. Subjective Test of Synthesized Speech of Male Voice

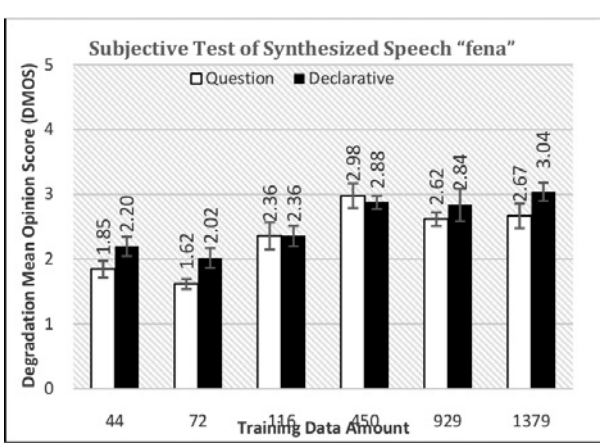

Fig. 12. Subjective Test of Synthesized Speech of Female Voice

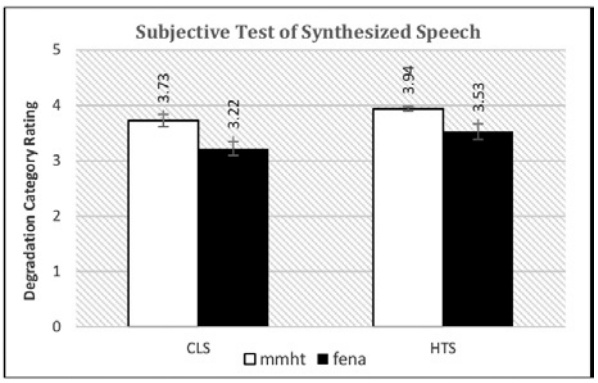

Fig. 13. Subjective Test of Synthesized speech with HTS and CLS

In addition, the comparison result of speech quality in HTS and CLS by subjective evaluation can be seen in Fig.13. It shows that there is no big difference between the synthesized speech of CLUSTERGEN and HTS. But still, HTS who produce the better-synthesized speech with score 3.94/5.00 for mmht and 3.53/5.00 for fena which means that "*degradation speech is slightly annoying*".

Conclusion

Based on the explanation above, it can be concluded that speech synthesis system for Bahasa Indonesia has been built. The system built by statistical parametric speech synthesis system which using statistical model to run the mapping between the speech and linguistic information. Some variation has been applied to the system and achieved the speech quality which measured by objective and subjective test. The speech quality by objective test result is acquire the best value for question sentences is 4.32 using 450 training sentences and for declarative sentences is 5.13 using 929 training sentences. Based on subjective test, acquire the highest value for question and declarative sentences using 1379 training sentences with score 3,53/5.00 and 3,36/5.00 respectively which mean *degradation speech is slightly annoying.*

Beside that also has been compare the HMM based text to speech (HTS) method and the CLUSTERGEN method. This method has difference in the trajectory model which provided by CLUSTERGEN method. From the evaluation acquired that the speech quality result of the synthesized speech by using CLUSTERGEN and HTS are not having big different. The subjective test has shown that both produce the synthesized speech with degradation speech is slightly annoying. The objective test has shown that the synthesized speech still produces big distortion.

That result possibly because of poor F0 estimates. In the future work, should be improved to increase the synthesized speech quality by modifying the system drawback, especially in HTS, by using better vocoder like STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), make better acoustic model and reduce post filtering. The others future work is to build the speaker adaptation of Indonesian TTS with only using small adaptation data, and also build expressive Indonesian TTS.

References

[1] G. K. Anumanchipalli and A. Black, "Adaptation Techniques for Speech Synthesis in Under-resourced", SLTU 2010, Penang, Malaysia, 2010.

[2] A. Black and J. Konimek "Optimizing Segment Label Boundaries for Statistical Speech Synthesis" ICASSP 2009, Taipei, Taiwan. 2009

[3] K. Tokuda, H.Zen and A. BLack "An HMM-based Speech Synthesis System Applied to English", Proc. of 2002 IEEE SSW, Sept. 2002.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis". In Proceeding of ICASSP, p. 1315-1318, 2000.

[5] A. Elok. "Pembuatan Perangkat Basis Data Untuk Sintesis Ucapan (Natural Speech Synthesis) Berbahasa Indonesia Berbasis Hidden Markov Model (HMM)". Jurnal Teknik POMITS Vol. 2, No. 2, (2013) ISSN: 2337-3539, hal. A 443-A 447

[6] K. Tokuda, T. Mausko, N. Miyazaki, T. Kobayashi, "Multi-space probability distribution HMM". IEICE Transactions on Information and Systems, E85-D(3), p. 455-464, 2002.

[7] Muslich, Masnur, "Fonologi Bahasa Indonesia Tinjauan Deskriptif Sistem Bunyi Bahasa Indonesia", Jakarta: PT Bumi Aksara, 2009.

[8] E. Tysmayudanto G and D. Arifianto, "Natural Indonesia Speech Synthesis by using CLUSTERGEN", International Conference on Information, Communication Technology and System, 2014 (ISSN: 978-1-4799-6858-9/14)

[9] K. Prahallad, A. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variation in conversational speech synthesis," in Proceedings of ICASSP 2005, Toulouse, France, 2006

[10] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitchadaptive time-frequency smoothing and an instantaneous frequency based f0 extraction: possible role of a repetitive structure in sounds, "Speech Communications, vol. 27, pp. 187–207, 1999.

[11] A. Black. "CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling", in Interspeech 2006, Pittsburgh,PA., 2006

[12] Cahyaningtyas, E., Arifianto, D., "HMM-based Indonesian Speech Synthesis System with Declarative and Question Sentences Intonation". Proc. IEEE 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) November

9-12, 2015, 1E -7 (pp.153–158).

[13] E. Cahyaningtyas and D. Arifianto, "Synthesized speech quality of Indonesian natural text-to-speech by using HTS and CLUSTERGEN," 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Bali, 2016, pp. 110-115.doi: 10.1109/ICSDA.2016.7918994

[14] A. J. Hunt and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01 (ICASSP '96), Vol. 1. IEEE Computer Society, Washington, DC, USA, 373-376. DOI=http://dx.doi.org/10.1109/ICASSP.1996.541110

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc. of Eurospeech, pp.2347-2350, Sept. 1999.

[16] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, The HMM-based speech synthesis system version 2.0, Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.

[17] Makoto Tachibana, Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. 2006. A Style Adaptation Technique for Speech Synthesis Using HSMM and Suprasegmental Features. IEICE - Trans. Inf. Syst. E89-D, 3 (March 2006), 1092-1099. DOI=http://dx.doi.org/10.1093/ietisy/e89-d.3.1092

[18] ITU-T, "Methods for Objective and Subjective Assessment of Quality", http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[19] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of Speaker Adaptation Algorihms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm, IEEE Audio, Speech, and Language Processing vol.17 issue 1, pp.66-83, January 2009.

[20] E. Cahyaningtyas and D. Arifianto, "Development of under-resourced Bahasa Indonesia speech corpus," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 2017, pp. 1097-1101. doi: 10.1109/APSIPA.2017.8282191