Research Journal of Applied Sciences, Engineering and Technology 7(24): 5143-5147, 2014
DOI:10.19026/rjaset.7.910
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 28, 2014 Accepted: February 10, 2014 Published: June 25, 2014

Research Article

Statistical Parametric Speech Synthesis of Malay Language using Found Training Data

Lau Chee Yong and Tan Tian Swee
Medical Implant Technology Group (MediTEG), Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA), Faculty of Biosciences and Medical Engineering (FBME), Universiti Teknologi Malaysia, Malaysia

Abstract: The preparation of training data for statistical parametric speech synthesis can be sophisticated. To ensure the good quality of synthetic speech, high quality low noise recording must be prepared. The preparation of recording script can be also tremendous from words collection, words selection and sentences design. It requires tremendous human effort and takes a lot of time. In this study, we used alternative free source of recording and text such as audio-book, clean speech and so on as the training data. Some of the free source can provide high quality recording with low noise which is suitable to become training data. Statistical parametric speech synthesis method applying Hidden Markov Model (HMM) has been used. To test the reliability of synthetic speech, perceptual test has been conducted. The result of naturalness test is fairly reasonable. The intelligibility test showed encouraging result. The Word Error Rate (WER) for normal synthetic sentences is below 15% while for Semantically Unpredictable Sentences (SUS) is averagely in 30%. In short, using free and ready source as training data can leverage the process of preparing training data while obtaining motivating synthetic result.

Keywords: Hidden Markov Model (HMM), letter to sound rule, statistical parametric speech synthesis

INTRODUCTION

Speech synthesis is a process of converting text representation of speech into waveform that can be heard by listeners (Ekpenyong *et al*., 2014). Statistical parametric speech synthesis (Zen *et al*., 2009) is a method of using natural speeches and texts as training data, the input training data is transformed into intermediate label data and the speech synthesizer uses the intermediate label data to synthesize speech. This method is using famous mathematical model which is Hidden Markov Model (HMM) (Ibe, 2013) that can be applied in various area such as pattern recognition, signal processing and so on. The quality of the synthetic speech is affected by the quality of the training data. Therefore, the preparation of input training data is crucial and requires thorough design of script and good quality of recording. However, the process of preparing input training data is not an easy task. The selection of script requires tremendous human effort in collecting words and designing sentences (Tan and Salleh, 2009). The recording setup must be good to reduce noise and able to record clean speech.

In this study, we have built a Malay language speech synthesizer using alternative sources such as audio-books, educational storytelling audio data, clean speech and so on. Those data can be obtained online for free. We have taken the free speech online and segmented only the clean portion and prepared the corresponding script to be the input training data. The synthetic speech using free source has been compared to the synthetic speech using specially designed and recorded training data. More details are explained in later section.

Statistical parametric speech synthesis using Hidden Markov Model (HMM): Statistical parametric speech synthesis is a speech synthesis method which generates average sets of similar sounding speech segment instead of using real speech segment like in unit selection method (Lim *et al*., 2012). Typically, it uses mathematic model such as Hidden Markov Model (HMM) to model the spectral and excitation parameters extracted from a real speech database. Model parameters are usually estimated using Maximum Likelihood (ML) criterion as:

$$\hat{\lambda} = \arg\max_{\lambda} \{p(O \mid W, \lambda)\} \quad (1)$$

where, λ is set of model parameters, O is set of training data and W is set of word sequences corresponding to O. When we want to generate desired speech, first the sentences is composed, then follow the equation below:

Corresponding Author: Tan Tian Swee, Medical Implant Technology Group (MediTEG), Cardiovascular Engineering Center, Material Manufacturing Research Alliance (MMRA), Faculty of Biosciences and Medical Engineering (FBME), Universiti Teknologi Malaysia, Malaysia

$$\hat{o} = \arg\max_{o}\{p(o \mid w, \hat{\lambda})\} \quad (2)$$

where, o is the speech parameters we want to generate, w is the given word sequence and $\hat{\lambda}$ is the set of estimated models. These parameters are then used to generate speech waveform. Any generative model can be used but HMM is most widely used model in this approach because of its memory-less ability to reduce complexity during process. It is commonly known as HMM-based speech synthesis (Yoshimura *et al*., 1999).

METHODOLOGY

Database preparation: The found data for this study is from the website http://free-islamic-lectures.com which is a free resource providing Islamic teaching recording. It offers free download of audio recording of Al-quran reading in Arabic language with translation of Malay language. We manually segmented the Malay speech portion out and prepared the corresponding script. In short, we obtained 1 h of Malay speech from this free source.

The training data text script that is specially designed and recorded were obtained from (Yong and Swee, 2014). However, this set of text script was recorded by a male native adult speaker. In short, 1 h of Malay recorded speech was obtained to become training data.

Front end processing using direct mapping letter to sound rule: Unlike conventional speech synthesizer which uses phoneme as the basic synthesis unit, we used letter to be the basic synthesis unit instead. The difference between using phoneme or letter as the training unit is: a dictionary is required to find out the precise phoneme boundary for every phoneme but it is not required to segment the lexicon into letters. Decode the lexicons into letter is much simpler than in phoneme and requires no knowledge from language experts. Figure 1 shows how the direct mapping letter to sound rule is defined.

Speech training: The process of training can be categorized into 3 phases.

Phase 1: The features of the original training speech were extracted and variance flooring was applied. Then the Hidden Markov Model (HMM) was initialized using K-mean clustering and re-estimated using Expectation-Maximization (EM) algorithm. After that, the HMMs were converted into context dependent models.

Phase 2: Embedded training of context dependent models without parameter tying was conducted. Then, the models were compressed and decision tree clustering was applied. After the models were tied, embedded training was applied again to tied models. And the parameters were untied after the embedded training.

Fig. 1: Direct mapping letter to sound rule

Phase 3: Convert trained HMM into HTS-engine models. Viterbi algorithm is then applied to re-align HMMs.

The training process is illustrated in Fig. 2.

Synthesis of speech: The desired synthetic sentences were formed and labeled like in training stage, resulting in a sequence of context-dependent phone labels for each utterance. Then, acoustic models were joined based on the synthetic sentence. And the speech parameter generation algorithm (Case 1) (Tokuda *et al*., 2000) was adopted to generate the spectral and excitation parameters. The STRAIGHT vocoder is then generates the speech waveform using the parameters.

Evaluation: Five systems (System A to E) have been created to test the reliability of synthetic speech which uses found data as training data. We used the original training speech from both recorded data and found data as standard reference. And we designed some normal sentences which is meaningful and intelligible and Semantically Unpredictable Sentences (SUS) (Benoît *et al*., 1996) for both recorded data and found data. The summary is listed in Table 1.

The SUS design was based on the following structures (Table 2).

Perceptual test was conducted by 17 listeners to evaluate the quality of synthetic speech in terms of naturalness and intelligibility. All the listeners are native Malay speaker. Even though there are some objective methods to test the quality of synthetic speech, but only perceptual test is able to effectively evaluate the naturalness and intelligibility of synthetic speech (Ekpenyong *et al*., 2014). For naturalness test, listeners were presented the synthetic speeches from all the systems. They were asked to rate the speech based on their opinion about its naturalness using a range

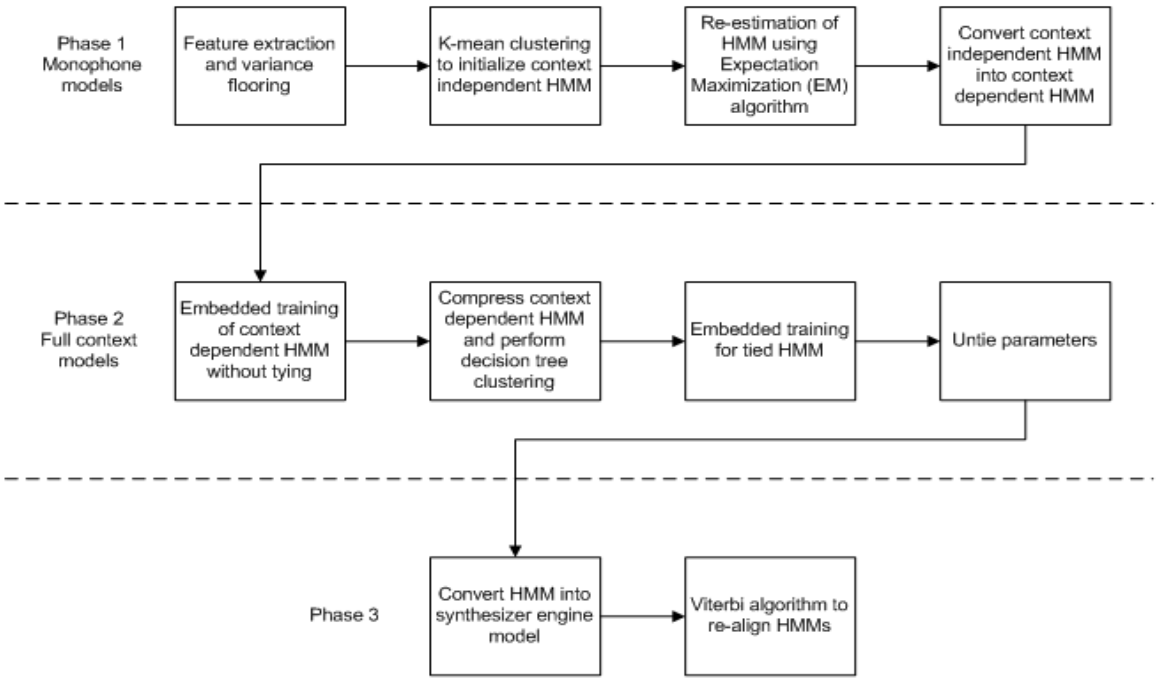

Fig. 2: Block diagram of training process

Table 1: Systems created for listening test

System	Detail
A	Original speech from recorded data and found data
B	Synthetic speech from recorded data using normal sentences
C	Synthetic speech from found data using normal sentences
D	Synthetic speech from recorded data using SUS
E	Synthetic speech from found data using SUS

Table 2: SUS structure and its example

Structure	Example
Intransitive (noun+det+verb (intr.) +preposition+noun+det+adjective)	Kangkung ini bersambilan dengan pendengaran yang besar.
Transitive (noun+adjective+verb (trans) +noun+det)	Almari rendah melayan beg itu.
Interrogative (quest. adv+noun+ det+verb(trans.) +noun+det+adjective)	Manakah orang itu menolak lampu yang bising?

Table 3: Naturalness test result

System	A	B	C	D	E
Naturalness	4.5965±0.1904	4.2188±0.5836	4.0488±0.6622	3.5276±0.4295	3.5612±0.2552

Table 4: Word Error Rate (WER) of each system

System	A	B	C	D	E
WER	9.08	11.87	19.61	36.16	53.84

from 1 to 5. Five represents very natural while 1 represents least natural. For intelligibility test, listeners were asked to transcribe the perceived synthetic speeches into texts. This listening test was conducted in a quiet room in Universiti Teknologi Malaysia. Headphone was used for every listening test. Each listening test lasts around 40 min as they have to listen to 50 sentences for naturalness test and 50 sentences for intelligibility test.

RESULTS

The result of naturalness test is shown in Table 3 and Fig. 3.

Listeners were asked to transcribe the sentences into text. From the response of listeners in this test, we calculated the Word Error Rate (WER) according to the equation below:

$$WER = \frac{S + D + I}{S + D + C} \quad (3)$$

where, S is substitution of words, D is deletion of words, I is insertion of words and C is correct words. Table 4 shows the Word Error Rate (WER) of all systems.

DISCUSSION

Using both recorded data and found data as training data, the naturalness of synthetic speech of

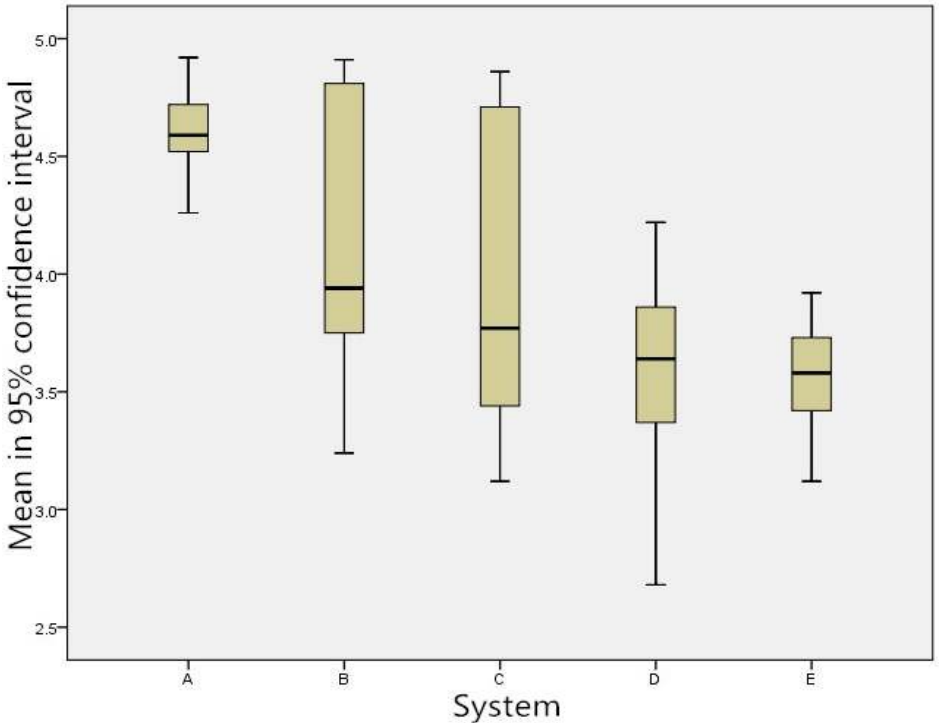

Fig. 3: Result of naturalness test

normal sentences is close to the original recorded speech. And the synthetic speech of SUS is slightly lower than normal sentences. But the naturalness is similar for both synthetic speeches using recorded and found data. The slightly decrease of naturalness in SUS may due to the understanding of the sentences. Listeners may find it unnatural since it is not intelligible and meaningful in terms of sentence content. On the other hand, similar trend happened in intelligibility test. The WERs of normal sentences is close to the WER of original speech. And the WER of SUS is similar for both speeches trained by recorded data and found data. However, there is a noticeable increase in WER of SUS compared to normal sentences. It may due to the random placement of words in the SUS due to the nature of SUS so listeners were feeling difficult to perceive the correct words.

In this study, the naturalness and intelligibility of synthetic speech trained by found data is satisfactory and listeners were able to perceive the meaning of normal sentences. This is a great ease of training data collection process because recording database and constructing recording script is tremendous and requires good quality of recording setup. However, there are a lot of free source like educational audio-book, storytelling book, speech and so on can be found online. The quality of the recording of the free source can be good enough to be the training data. Manually segmentation can be done to select only clean and clear speech to be the input data.

CONCLUSION

We have presented a Malay language speech synthesizer in this study. We compared the synthetic speech trained by recorded data and found data. Recorded data were obtained from a series of procedure from words collection, sentence design and recording under good quality of recording setup while the found data was obtained from free source like audio-book, speech and so on. The listening test result showed no significant difference between synthetic speeches trained by recorded data and found data. It is an encouraging result to show that alternative source of training data is able to become training data while a lot of human efforts were bypassed in preparing the training data.

To mention future work, different accent of free speech source can be used to synthesize speeches in different accent. Automatic segmentation of clean speech like diarization of speech can be conducted to reduce more human effort.

ACKNOWLEDGMENT

The authors would like to thank IJN for their professional opinions and involvement, Ministry of Higher Education (MOHE), Universiti Teknologi Malaysia (UTM) and UTM Research Management Centre (RMC) for supporting this research project under grant code 04h41.

REFERENCES

Benoît, C., M. Grice and V. Hazan, 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Commun., 18(4): 381-392.

Ekpenyong, M., E.A. Urua, O. Watts, S. King and J. Yamagishi, 2014. Statistical parametric speech synthesis for Ibibio. Speech Commun., 56: 243-251.

Ibe, O.C., 2013. 14-hidden Markov Models. In: Ibe, O.C. (Ed.), Markov Processes for Stochastic Modeling. 2nd Edn., Elsevier, Oxford, pp: 417-451.

Lim, Y.C., T.S. Tan, S.H. Shaikh Salleh and D.K. Ling, 2012. Application of genetic algorithm in unit selection for Malay speech synthesis system. Expert Syst. Appl., 39(5): 5376-5383.

Tan, T.S. and S.H.S. Salleh, 2009. Corpus design for Malay corpus-based speech synthesis system. Am. J. Appl. Sci., 6(4): 696-702.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithm for HMM-based speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00). Istanbul, 3: 315-1318.

Yong, L.C. and T.T. Swee, 2014. Low footprint high intelligibility Malay speech synthesizer based on statistical data. J. Comput. Sci., 10(2): 316-324.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of the Eurospeech, 1999.

Zen, H., K. Tokuda and A.W. Black, 2009. Statistical parametric speech synthesis. Speech Commun., 51(11): 1039-1064.