Research Journal of Applied Sciences, Engineering and Technology 8(14): 1691-1694, 2014
DOI:10.19026/rjaset.8.1151
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 14, 2014 Accepted: September 20, 2014 Published: October 10, 2014

Research Article

Improve the Quality of Synthetic Speech Trained with Found Data using Silence Cutter

Lau Chee Yong, Tan Tian Swee and Mohd Nizam Mazenan
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: Using found data as training data in statistical parametric speech synthesis can alleviate various problems in tedious database construction. However, the extra silences resided in found data degrades the quality of synthetic speech. Therefore, in this study, silence cutter was created to eliminate the extra silences in the training data. The motivation is the extra silences would be incorrectly assigned to training script and result in unnatural synthetic speech. Therefore, in this study, a Malay speech synthesis system has been constructed using found data from internet. Silence cutter has been utilized to cut out extra silences. The synthetic speech using found data with and without silence cutter was verified and compared to find out the effect of silence cutter. Result showed that silence cutter has help to improve synthetic speech naturalness and reduce the Word Error Rate (WER) in intelligibility test. In short, using found data can alleviate the problem of preparing high quality training data and silence cutter can be used to refine the found data to generate better quality of synthetic speech.

Keywords: Found data, hidden Markov model, statistical parametric speech synthesis

INTRODUCTION

Statistical parametric speech synthesis (Tokuda *et al*., 2002; Zen *et al*., 2007, 2009) is an approach of synthesizing artificial speech waveform from human recorded speech data. It extracts the spectral and excitation parameters from recorded speech data and re-model it using Hidden Markov Model (HMM) (Ibe, 2013). Therefore, this method can be also named as HMM-based speech synthesis (Chopde and Pushpa, 2014). Conventionally, human speech database was constructed by recording of script reading. However, it requires a good quality recording setting and a lot of human effort from gathering words, making scripts, building lexicon database and recording. In this study, we wish to replace the tremendous work by using alternative source of speech recording which can be found over internet like audiobook, online speeches and many more. This kind of data can be obtained from internet and it is free of charge. Some audiobook audio data possess high recording quality which is suitable to be used as training data. However, the found data might not always be in the correct attribute as a suitable training data. One of the problems is the waveform in found data might contain extra-long initial and end silences. The extra silences would be aligned to text and create inaccurate alignment during training and degrade the synthetic speech quality. Therefore, the found data should be refined before it being used as training data.

In this study, we intended to build a Malay language speech synthesizer using found data from internet. We performed some refinement to eliminate the extra silences of speech waveforms and we compared the effect of the synthetic speech using refined and without refined data. More details are explained in the rest of this study.

METHODOLOGY

Found data: The audio Malay speech data was obtained from the website http://free-islamic-lectures.com which is a free Islamic teaching website. It offers up to 60 h of Malay translation of Al-Quran in Arab language. We manually extracted the Malay part to be our training data. In short, a duration of 1 h of Malay speech data and its script was obtained. However, the found data may not exhibit the correct manner as we desired. There might have extra initial and end silences which could lead to alignment error. Therefore, refinement should be done to eliminate the extra silences.

Silence cutter: The found data may contain extra initial or end silences. These extra silences may complicate the training process by aligning the silences into words

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

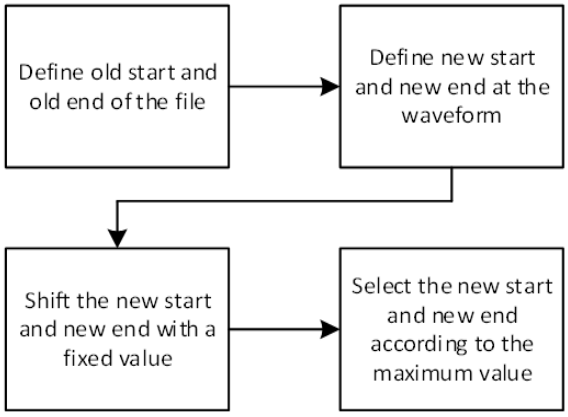

Fig. 1: Block diagram of silence cutter

resulting unnatural synthetic speech. Therefore, removing the extra silences would improve the accuracy of aligning speech into corresponding texts and improve the quality of synthetic speech. The block diagram of silence cutter is shown in Fig. 1. First, the silence cutter defines the starting and ending point of the file as old start and old end. Then it assigns a moving point as new start and new end at the beginning and ending of waveform. The new start and new end were shifted 100 msec to both ends of the file. Then the new start and new end were updated by the maximum value of the time. The waveform is trimmed according to the value of new start and new end.

The result of silence cutter is shown in Fig. 2. It is obvious to see that the initial and end silences have been eliminated.

Silence cutter is a process before front end processing. It only eliminate extra initial and end silence. It does not trim out the silences at the middle of waveforms. The text normalization is done manually and the training text is containing only words, space and punctuation after normalization. No abbreviation, symbol is included.

Training and synthesis: In this study, letter-based statistical parametric speech synthesis method (Watts *et al*., 2010) has been applied. The overall training process can be described as three phases as shown in Fig. 3. At the first phase, the features of the real speech are extracted followed by variance flooring. After that, the letter HMMs are trained with the initial segments of database script and speech using segmental k-means to classify the group of letter and Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977) was used to perform embedded training of the letters. At phase 2, context independent HMMs were converted into context dependent HMMs. Re-estimation was done using embedded training until the parameters were converged. At phase 3, the decision tree clustering (Young *et al*., 1994) is applied for the spectral stream, log f0, band limited aperiodic measures and duration Probability Distribution Functions (PDF). The tied models were further trained until the parameters converged. After phase 3, the context dependent HMMs are untied and the process repeated from phase 2 until the end. After all the phases and iterations were done, the HMM models were converted into HTS engine model. And realignment of HMM was done using Viterbi Algorithm.

For the synthesis process, first, arbitrary sentences or target sentences were created. By using the front end text processing in training stage, the context dependent label sequence was generated. According to the label sequence, the corresponding HMM was concatenated. And the speech parameter generation algorithm (Case 1) (Tokuda *et al*., 2000) was adopted to generate the spectral and excitation parameters. The straight vocoder (Kawahara, 2006) is then rendering the speech waveform using the parameters.

RESULTS AND DISCUSSION

Sixteen listeners were invited to rate the synthetic test in terms of naturalness and intelligibility. In naturalness test, three systems have been constructed and listeners were asked to rate the naturalness using a 5 point scale where 1 represented not natural and 5

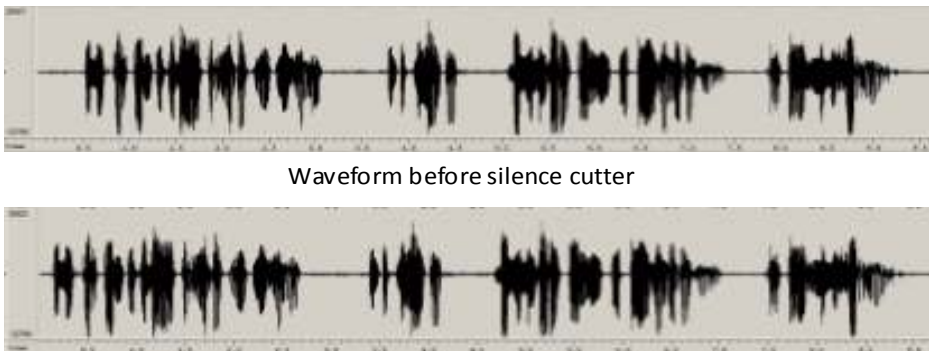

Fig. 2: Training speech waveform before and after silence cutter

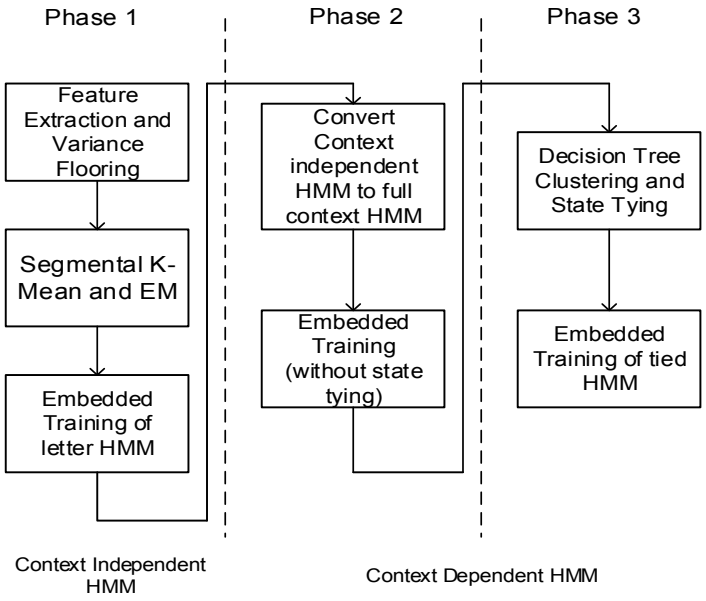

Fig. 3: Process of training

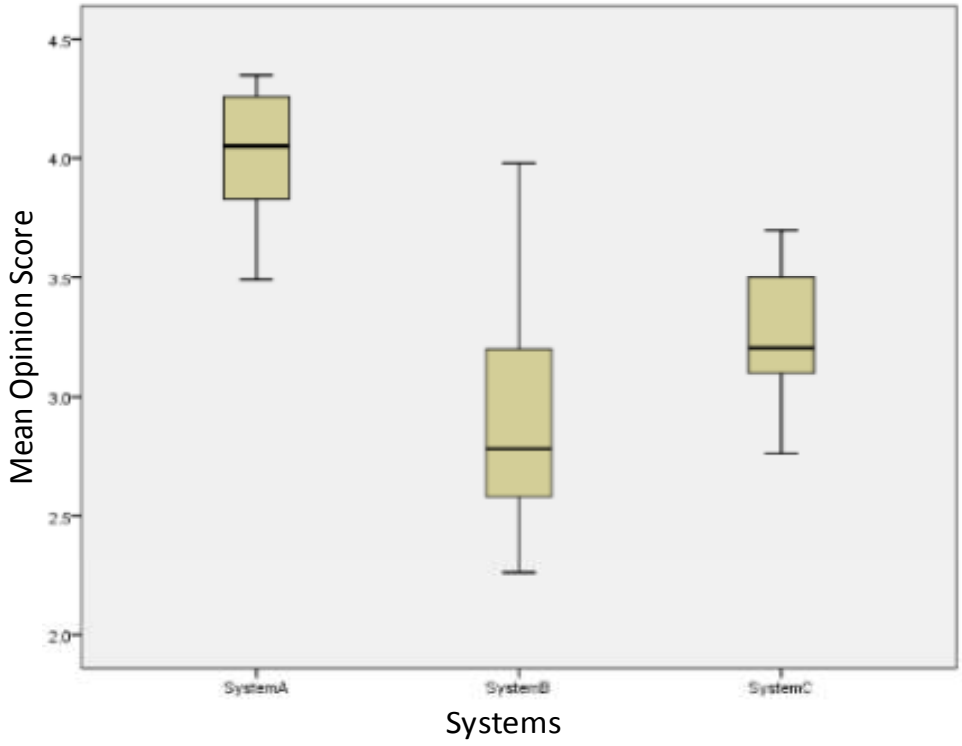

Fig. 4: Graph of naturalness test result

represented very natural. The summary of the systems is shown in Table 1.

The result of naturalness test is shown in Table 2 and illustrated in Fig. 4. The raw unmodified natural data in System A possessed highest naturalness degree among the systems. And apparently System C is more natural compared to System B. This shows that the unwanted silence region would degrade the naturalness of synthetic speech and a silence cutter would help in increase the naturalness of synthetic speech.

In intelligibility test, we asked the listeners to transcribe what they had perceived into text. And we calculated the Word Error Rate based on the output. Four systems have been created to test the effect of

Table 1: Systems created in naturalness test

System	Description
A	Raw found data
B	Synthetic speech without silence cutter
C	Synthetic speech with silence cutter

Table 2: Descriptive result of naturalness test

	System A	System B	System C
Mean	4.020	2.901	3.258
Variance	0.074	0.183	0.069
Standard deviation	0.272	0.428	0.263
Upper bound	4.145	3.114	3.389
Lower bound	3.875	2.688	3.127

Table 3: Systems created in intelligibility test

System	Description
A	Meaningful sentences without silence cutter
B	SUS without silence cutter
C	Meaningful sentences with silence cutter
D	SUS with silence cutter

Table 4: Word Error Rate (WER) of all systems in intelligibility test

System	WER (%)
A	39.41
B	60.89
C	21.94
D	36.43

using silence cutter on training data. Meaningful sentences and Semantically Unpredictable Sentences (SUS) (Benoît *et al*., 1996) were used in this test. The summary of the systems are listed in Table 3.

Each system provides 5 sentences and a listener has to listen 20 sentences in intelligibility test. The order of sentences was shuffled and the listeners would not know the sentence they were listening belongs to which system. The Word Error Rate (WER) was calculated according to the equation below:

$$WER = \frac{S + D + I}{S + D + C} \quad (1)$$

where,

S = Substitution of words
D = Deletion
I = Insertion
C = Correct words

The result of intelligibility test is shown in Table 4.

CONCLUSION

This study has presented a statistical parametric speech synthesis applied to Malay language using found data as training data. The extra initial and end silences in the speech data might lead to alignment error during training. Therefore, the extra silences have been eliminated using the silence cutter presented in this study. After applying silence cutter, the alignment between text and speech can be performed with more accuracy. The synthetic speech generated using found data with and without silence cutter have been compared and verified with 16 listeners. Result showed that after applying silence cutter, the naturalness and intelligibility of synthetic speech have been improved.

ACKNOWLEDGMENT

The authors gratefully acknowledge the research grant provided by Research Management Centre (RMC), sponsored by Ministry of Higher Education (MOHE), Malaysia. Vot: 04H41 and Flagship University Teknologi Malaysia, Johor Bahru, Malaysia.

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.

Chopde, S. and U. Pushpa, 2014. HMM-based speech synthesis. Int. J. Mod. Eng. Res. (IJMER), 3(4): 1894-1899.

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B Met., 39(1): 1-38.

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.

Kawahara, H., 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoust. Sci. Technol., 27(6): 349.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithm for HMM-based speech synthesis. Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00), pp: 1315-1318.

Tokuda, K., Z. Heiga and A.W. Black, 2002. An HMM-based speech synthesis system applied to english. Proceeding of 2002 IEEE Workshop on Speech Synthesis, pp: 227-230.

Watts, O., J. Yamagishi and S. King, 2010. Letter-based speech synthesis. Proceeding of Speech Synthesis Workshop 2010.

Young, S.J., J.J. Odell and P.C. Woodland, 1994. Tree-based state tying for high accuracy acoustic modelling. Proceeding of ARPA Human Language Technology Workshop, pp: 307-312.

Zen, H., T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black and K. Tokuda, 2007. The HMM-based speech synthesis system (HTS) version 2.0. Proceeding of the 6th ISCA Workshop on Speech Synthesis. Bonn, Germany, August 22-24, 2007.

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