Research Journal of Applied Sciences, Engineering and Technology 7(18): 3803-3808, 2014
DOI:10.19026/rjaset.7.737
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: October 31, 2013 Accepted: November 08, 2013 Published: May 10, 2014

Research Article

Investigation of Effects of Different Synthesis Unit to the Quality of Malay Synthetic Speech

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: Synthesis unit of a speech synthesizer directly affects the computational load and output speech quality. Generally, phoneme is the best choice to synthesize high quality speech. But it requires the knowledge of language to precisely draw the segmentation of words into phonemes. And it is expensive to compose an accurate phoneme dictionary. In this study, another type of synthesis unit is introduced which is letter. In Malay language, the unit size of letter is smaller than phoneme. And using letter as the synthesis unit could ease a lot of efforts because the context label can be created in fully automatic manner without the knowledge of the language. Four systems have been created and an investigation was done to find out how synthesis unit could affect the quality of synthetic speech. Forty eight listeners were hired to rate the output speech individually and result showed that no obvious difference between the output speech synthesized using different synthesis units. Listening test showed satisfactory result in terms of similarity, naturalness and intelligibility. Synthetic speech with polyphonic label showed increment in intelligibility compared to synthetic speech without polyphonic label. Using letter as the synthesis unit is recommended because it excludes the dependency of linguist and expands the idea of language independent front end text processing.

Keywords: Hidden Markov model, letter, phoneme, statistical parametric speech synthesis

INTRODUCTION

Speech synthesis is a process of transforming textual representation of speech into waveform (Lim *et al*, 2012). In earlier techniques of speech synthesis, the engine is adopted in a rule based framework. However, this engine is now adopted in data-driven approach using mathematical model like Hidden Markov Model (HMM) (Tokuda *et al*., 1995, 2000) and it is called statistical parametric speech synthesis or HMM Speech Synthesis System (HTS-H Triple-S). This method has gained a lot of recognition and able to synthesize natural and intelligible speech (King and Karaiskos, 2009). The principle of a statistical parametric speech synthesizer is to generate the averages of the similar speech segments. It first extracts the excitation and spectral parameters from real speech and models them using HMM. After several iterative training, a speech waveform can be reconstructed using the parameters from the HMMs in a concatenated manner (Zen *et al*., 2009, 2012). This speech synthesis method is applicable to many kind of language and currently can be found in English, Chinese, Korean, Spanish, Arabic and so on. It also capable to make speaker adaptation and voice conversion available by tuning the acoustic parameters (Watts *et al*., 2010). The flexibility of transforming output speech in different styles is a key advantage and preferred by most of the researchers (Turk and Schroder, 2010).

The selection of synthesis unit of a statistical parametric speech synthesizer is crucial to determine its performance. Conventionally, syllables are appropriate candidates to be speech synthesis unit because they are speech production and perception units. However, the syllable groups might be over excessive. For example, the number of syllables in English exceeds ten thousand (Sagisaka and Sato, 1986) which is impractical to be the synthesis unit. Therefore, the idea of phoneme was introduced (Stan *et al*., 2011). Phoneme is the basic unit of language phonology. It consists of lesser synthesis unit than syllables and it is currently favorable by researchers. However, it requires the knowledge of the linguist to provide a precise lexicon segmentation and phoneme definition. Besides, some phonemes might be

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

inexistence in lexicon database due to its low occurrence. In this study, we proposed another type of synthesis unit which is letter because it is more suitable to be the synthesis unit in term of unit size. We built Malay speech synthesizers using phoneme and letter as synthesis unit. An investigation was conducted to observe the effects of different synthesis unit toward synthetic speech.

MALAY FRONT END TEXT PROCESSING

Text corpus: The script of the lexicon database was collected from newspaper and Malay primary school textbooks to gather daily conversational speech examples. A phoneme selection process has been conducted to collect the entire phoneme in Malay language and include them into database (Tan and Sh-Hussain, 2009). Ten million words have been collected and 988 sentences were composed using the collected Malay words. Eight hundred and thirty eight sentences were used for training while the rest are the text input of the synthesizer.

Recording setup: The speaker who recorded the Malay language database was a native Malay female speaker which could ensure correct pronunciation and good naturalness of the spoken sentences. The recording is done in a quiet room in center of Biomedical Engineering, University Teknologi Malaysia with microphone SM48 Cardioid Dynamic Vocal Microphone from Shure Company. Surrounding noise was measured before recording and only proceeded if the noise was low. The total recording time was around 2 weeks but each recording session only lasts for 1 h to prevent the voice quality being affected after long time speaking. The speaker spoke for few times before the recording session to adapt a constant speaking style and intonation. No any special accent was used and the speaker spokes in Standard Malay. The recordings were made at 96 kHz sampling frequency and 16 bits/sample. The recordings were down sampled to 16 kHz. It is because 16 kHz frequency is substantial to obtain a good quality synthesized speech (Stan *et al*., 2011) and practical for the processing time.

Context labeling based on dictionary and direct mapping letter to sound rule: All the words of lexicon database were extracted. For system using phoneme as synthesis unit, a dictionary was referred to find out the pronunciation of the words. While in letter based system, direct mapping letter to sound rule was applied to the lexicon and all the words were decoded into its letter element. The process is illustrated in Fig. 1.

The difference between these two processes is a dictionary is needed in phoneme based segmentation while the segmentation can be done directly in letter based process without referring to any dictionary. For example, the word pengisytiharannya (his/her announcement) was labeled as /p_e_ng_i_sy_t_i_h_a_r_a_n_ny_a/in phoneme based system while the word was stripped down into letters and labeled as /p_e_n_g_i_s_y_t_i_h_a_r_a_n_n_y_a/in letter based system. The list of synthesis unit for phoneme and letter is shown in the Table 1. For polyphone, special character was used to represent another pronunciation of the letter. For example, 'e' was used to represent 'e' while 'é' was used to represent 'eh'. Total of 38 synthesis unit were used in phoneme synthesis while only 27 synthesis units were used in letter synthesis. The unit size of letter is 28.9% smaller than phoneme.

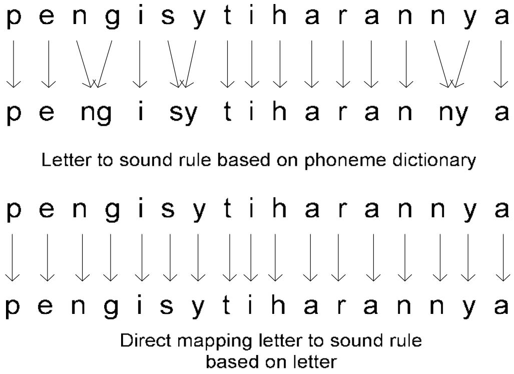
p e n g i s y t i h a r a n n y a
p e ng i sy t i h a r a n ny a
Letter to sound rule based on phoneme dictionary
p e n g i s y t i h a r a n n y a
p e n g i s y t i h a r a n n y a
Direct mapping letter to sound rule
based on letter

Fig. 1: Letter to sound rule based on phoneme and letter

Table 1: List of synthesis unit (phoneme and letter) in Malay language

Phoneme units	Letter units
/a/, /b/, /c/, /d/, /e/, /eh/, /f/, /g/, /h/, /i/, /j/, /k/, /l/, /m/, /n/, /o/, /p/, /q/, /r/, /s/, /t/, /u/, /v/, /w/, /x/, /y/, /z/, /sy/, /kh/, /gh/, /sh/, /ng/, /ny/, /ai/, /au/, /oi/, /ua/, /ia/	/a/, /b/, /c/, /d/, /e/, /é/, /f/, /g/, /h/, /i/, /j/, /k/, /l/, /m/, /n/, /o/, /p/, /q/, /r/, /s/, /t/, /u/, /v/, /w/, /x/, /y/, /z/

Polyphonic label: The letter 'e' is a polyphone in Malay language. It can be pronounced as 'e' and also 'eh'. To differentiate these two pronunciations, special character 'é' was used to represent 'eh' in lexicon database. This approach is only required in letter based synthesizer. For phoneme based system, the pronunciations were correctly defined according to dictionary.

HMM BASED SPEECH SYNTHESIZER SETUP

Overview of speech synthesis system in this study: In this study, statistical parametric speech synthesis method was adopted as in Blizzard Challenge 2010 (Oura *et al*., 2010) which use Hidden Semi Markov Model (HSMM) (Zen *et al*., 2004) as acoustic model with parameter generation algorithm considering Global Variance (GV) (Toda and Tokuda, 2005) to drive a high quality vocoder which is Speech Transformation and Representation using Adaptive Interpolation of weighted spectrum (STRAIGHT) (Kawahara *et al*., 1999). The process can be categorized into two parts which are training and synthesis. At

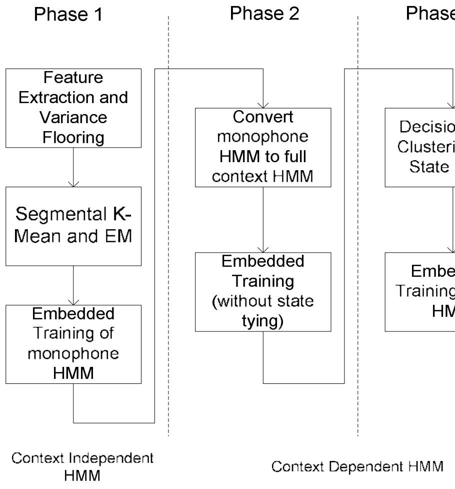
Phase 1
Phase 2
Phase 3
Feature Extraction and Variance Flooring
Segmental K-Mean and EM
Embedded Training of monophone HMM
Convert monophone HMM to full context HMM
Embedded Training (without state tying)
Decision Tree Clustering and State Tying
Embedded Training of tied HMM
Context Independent HMM
Context Dependent HMM

Fig. 2: Block diagram of training process

training stage, the statistical features of the real speech which are excitation, spectral parameters and fundamental frequency are extracted from database and trained into acoustic model using Maximum Likelihood (ML) criterion. To estimate the optimal model parameter, the likelihood for a given training data is maximized as:

$$\lambda_{ML} = \arg\max_{\lambda} \{p(O \mid \omega, \lambda)\} \tag{1}$$

where, λ is a set of model parameters, O is a set of training data and ω is a set of word sequences corresponding to O. The model parameter λ_{ML} is difficult to obtain analytically. So EM algorithm is applied to estimate the model parameter. We then generate speech parameters sequence, $O = [O_1^T, O_2^T, \ldots, O_T^T]^T$, for a given word sequence to be synthesized, w, from the set of estimated models, λ_{ML}, to maximize their output probabilities as:

$$o_{ML} = \arg\max_{o} \{p(o \mid w, \lambda_{ML})\} \tag{2}$$

And finally, the waveform is built from the generated speech parameters. This training and synthesis framework is language independent. Specific context-dependent label for different language is required to drive this framework. So we used the context label from Malay front end text processing to generate Malay speech.

Feature extraction setting: STRAIGHT mel-cepstral analysis was used to extract 39th order of mel-cepstral coefficient at 5 ms frame shift. For F0 estimation, voting was done to select the best estimation method among F0 trackers which are Instantaneous Frequency Amplitude Spectrum (IFAS), fixed point analysis (TEMPO) and ESPS get_f0 (Yamagishi *et al*., 2009).

Table 2: Systems created for listening test

System	Description
A	Natural speech
B	Synthetic speech using phoneme as synthesis unit
C	Synthetic speech using letter as synthesis unit
D	Synthetic speech using letter with polyphonic label as synthesis unit

Training and synthesis: The overall training process can be described as three phases as in Fig. 2. At the first phase, the features of the real speech are extracted followed by variance flooring. After that, the monophone HMMs are trained with the initial segments of database script and speech using segmental k-means to classify the group of synthesis unit (in this study: phoneme and letter) and Expectation-Maximization (EM) algorithm (Dempster *et al*., 1977) was used to perform embedded training of the monophone. At phase 2, monophone 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 is then rendering the speech waveform using the parameters.

EXPERIMENT

Four systems have been created in listening test as listed in Table 2.

Objective method of evaluating quality of synthetic speech is available but is only useful in certain situations. Judgment about naturalness and intelligibility is more reliable if conducted in subjective manner. To test the robustness and vulnerability of the speech synthesizer, the output speech was evaluated by 48 listeners. The listeners were asked to listen to the synthetic voice with headphone in a quiet room. Different tests were given to listeners based on different aspect as below.

Table 3: Semantically Unpredictable Sentences (SUS) design for intelligibility test

Structure	Example sentence
Intransitive (Det + Noun + Verb (intr.) + Preposition + Det + Adjective + Noun)	Penganiayaanitubertandinganuntukkesempatan yang utama
Interrogative (Quest. Adv + Aux + Det + Noun + Verb (trans.) + Det + Adjective + Noun)	Mengapakahcerminmenolaktelevisyen yang licin
Relative (Det + Noun + Verb (trans.) + Det + Noun + Relat. Pronoun + Verb (intr.))	Tindakanmenewaskanperlanconganitu yang berkeliruan

Similarity: Listeners were presented speeches from all the systems after listened to original voice. They were asked to rate from 1 (Totally not similar to original voice) to 5 (Very similar to original voice) based on their opinion about the synthetic voice. There are total of 40 sentences in this test.

Naturalness: All the voices were shuffled randomly so listener would not know the order of the voices from all the systems. They rated the voices from 1 (Totally unnatural) to 5 (Very natural). The scores were made without taking intelligibility into consideration. Only focus on the naturalness of the voice. There are 40 sentences in this test.

Intelligibility: Measured by the ability of listeners to correctly identify aspects of a speech stimulus like phonemes, syllables, words, or sentences, given that the language is known to the listeners and the syntax are correct (Childers and Wu, 1990). Semantically Unpredictable Sentences (SUS) (Benoît *et al*., 1996) were composed in this test. The sentences design is concluded in Table 3.

Every sentence in this section is less than 8 words to prevent memory loss during listening and only can repeat once in this test. Listeners were asked to transcribe the speech into words and intelligibility was calculated using Word Error Rate (WER). Total of 40 sentences were prepared in this test. For System B to D, different set of synthetic speech samples were prepared. Listeners would not listen to same sentence but they were all listened to synthetic speech from all the systems.

RESULTS AND DISCUSSION

The result of listening test is shown in Fig. 3. The graph shows the mean scores with 95% of Confidence Level (CL) for the synthetic speech of both synthesizers. From the graph, no significant variation in similarity and naturalness was observed from system B to D and their mean scores are ranging from 3.48 to 3.71. For intelligibility, Word Error Rate (WER) was measured. The formula of WER is given by the equation below and the detail result of the listening test is shown in Table 4 and 5:

$$WER = \frac{S + D + I}{S + D + C} \tag{3}$$

where, S is the number of substitution, D is the number of deletion, I is the number of insertion and C is the number of correct words. The WER was computed based on the number of substitution, insertion, deletion and correct word from the transcription by listeners. Spelling error is not considered as substitution.

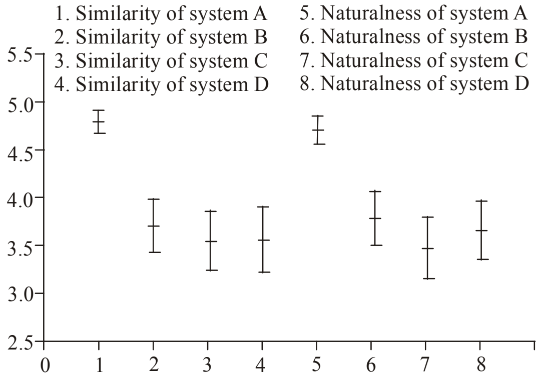

Fig. 3: Graph of listening test result with 95% of confidence level

Table 4: Result of listening test

	System			
	A	B	C	D
Similarity	4.79	3.71	3.54	3.56
Naturalness	4.71	3.77	3.48	3.67

Table 5: Word Error Rate (WER) for system A to D

System	S	D	I	C	WER (%)
A	51	17	0	2860	2.3
B	87	151	1	2690	8.2
C	188	231	2	2509	14.4
D	109	176	1	2643	9.8

From the result, no significant variation was observed in similarity and naturalness for system B to D. While for intelligibility, polyphonic label helped to improve intelligibility of synthetic speech. Even though no clear difference between these two results, but we recommended synthesizer using letter as synthesis unit because it offers several advantages in building Malay synthetic speeches:

- It excludes the dependency of linguist in defining phoneme groups. Since letter is the synthesis unit for synthesizer, the segmentation can be done in an automatic manner without referring to phoneme dictionary. In this process, the knowledge of linguist in Malay language is not required. Or in another word, human effort is omitted at front end processing.
- Basically the speech synthesizer framework (training + synthesis) is language independent. The

knowledge of language is only needed at front end text processing in defining the boundary of synthesis unit. However if letter is used as synthesis unit, front end text processing also can be done in language independent approach since the knowledge of language is not required. A fully language independent speech synthesizer can be created.

CONCLUSION AND RECOMMENDATIONS

We have presented the first statistical parametric speech synthesis system in Malay language and evaluated the effect of different synthesis unit towards the quality of output speech. Synthesizers were built using phoneme and letter as synthesis unit. Similarity, naturalness and intelligibility of the output speeches were rated by 48 listeners. From the result obtained, no clear difference was observed for output speeches in terms of similarity and naturalness. For intelligibility test, Word Error Rate (WER) was computed and satisfactory result was obtained. Synthesizer with polyphonic label helps to improve intelligibility of synthetic speech. Thus we concluded that synthesizer with either phoneme or letter as synthesis unit both give acceptable result and no clear difference was observed. However, we recommended letter as synthesis unit because it excludes the dependency of linguist, bypass the problem of insufficient training sample and expand the idea of a fully language independent speech synthesizer. In this study, polyphone was lightly treated manually to identify correct pronunciation in different words. For future work, classifiers can be trained to automatically predict the pronunciation of polyphone in different words.

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 RUG grant VOT 00G70.

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.

Childers, D.G. and K. Wu, 1990. Quality of speech produced by analysis-synthesis. Speech Commun., 9(2): 97-117.

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

Kawahara, H., I. Masuda-Katsuse and A. De Cheveigné, 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun., 27(3-4): 187-207.

King, S. and V. Karaiskos, 2009. The blizzard challenge 2009. Proceeding of the Blizzard Challenge Workshop. Edinburgh, U.K.

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.

Oura, K., K. Hashimoto, S. Shiota and K. Tokuda, 2010. Overview of NIT HMM-based Speech Synthesis System for Blizzard Challenge 2010. Retrieved from: http://festvox.org/blizzard/bc2010/NITECH_Blizzard2010.pdf.

Sagisaka, Y. and H. Sato, 1986. Composite phoneme units for the speech synthesis of Japanese. Speech Commun., 5(2): 217-223.

Stan, A., J. Yamagishi, S. King and M. Aylett, 2011. The Romanian Speech Synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Commun., 53(3): 442-450.

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

Toda, T. and K. Tokuda, 2005. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. Proceeding of the Interspeech, pp: 2801-2804.

Tokuda, K., T. Masuko and T. Yamada, 1995. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. Proceeding of the Eurospeech.

Tokuda, K., K. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceeding of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), pp: 1315-1318.

Turk, O. and M. Schroder, 2010. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. IEEE T. Audio Speech, 18(5): 965-973.

Watts, O., J. Yamagishi, S. King and K. Berkling, 2010. Synthesis of child speech with HMM adaptation and voice conversion. IEEE T. Audio Speech, 18(5): 1005-1016.

Yamagishi, J., T. Nose, H. Zen, L. Zhen-Hua, T. Toda, K. Tokuda, S. King and S. Renals, 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE T. Audio Speech, 17(6): 1208-1230.

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

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

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 2004. Hidden semi-Markov model based speech synthesis. Proceeding of the ICSLP, 2: 1397-1400.

Zen, H., N. Braunschweiler, S. Buchholz, M.J.F. Gales, K. Knill, S. Krstulovic and J. Latorre, 2012. Statistical parametric speech synthesis based on speaker and language factorization. IEEE T. Audio Speech, 20(6): 1713-1724.