Combining lightly-supervised learning and user feedback to construct and improve a statistical parametric speech synthesiser for Malay

Lau Chee Yong,[1] *Oliver Watts*[2], *Simon King*[2]

[1]Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Malaysia
[2] Centre for Speech Technology Research, University of Edinburgh, UK
cylau2@live.utm.my, owatts@staffmail.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

In spite of the learning-from-data used to train the statistical models, the construction of a statistical parametric speech synthesiser involves substantial human effort, especially when using imperfect data or working on a new language. Here, we use lightly-supervised methods for preparing the data and constructing the text-processing front end. This initial system is then iteratively improved using active learning in which feedback from users is used to disambiguate the pronunciation system in our chosen language, Malay. The data are prepared using speaker diarisation and lightly-supervised text-speech alignment. In the front end, grapheme-based units are used. The active learning used small amounts of feedback from a listener to train a classifier. We report evaluations of two systems built from high-quality studio data and lower-quality 'found' data respectively, and show that the intelligibility of each can be improved using active learning.

Index Terms: statistical parametric speech synthesis; active learning; lightly-supervised methods

1. Introduction

Conventional speech synthesis depends heavily on the availability of data: both the transcribed speech data required for waveform generation – whether that is through concatenation or statistical parametric models – as well as additional expert-annotated speech and text data. These additional data are required to train numerous predictors of linguistic features, prosodic structures, and so on (e.g. a pronunciation lexicon and a letter-to-sound module to predict the sequence of phonemes in an utterance). Our previous work has focused on finding ways to speed up the development of systems in new languages by developing techniques for transcribing 'found' data with only light supervision, and by using unsupervised learning to reduce reliance on expert annotated data.

In this paper, we describe the application of these techniques to the training of text-to-speech (TTS) systems for Malay. The existing techniques we have developed in previous work [1] and which are described in Section 2 are used to build a baseline voice. We then go on to describe and evaluate two new innovations to our existing toolset. All systems are based on the statistical parametric ("HMM-based") technique in which the speech corpus is used to estimate acoustic models of context-dependent sub-word units, which are then used to generate novel speech utterances. The contributions of this paper are in the preparation of the data and the construction of the text-processing front end. The estimation of the HMMs from the speech corpus is done in a standard way as described in [2], for example.

First of all, in Section 3 we describe the use of speaker diarisation techniques for extracting a homogenous subset of speech data from a public source of data featuring considerable variation. This technique supplements our existing methods for the lightly-supervised harvesting of data for acoustic model training [3] and is designed to further ease such data collection. The extension of our methods was found to be necessary because in the source of speech data used for the experimental systems of the present work, there is a much looser relation between speech and text than in the audiobooks we have used in previous work.

Then, in Section 4 we describe a method for improving the letter-based synthesis used in the baseline system by exploiting some limited interaction with a native speaker to disambiguate instances of an ambiguous grapheme, but without incurring the cost of compiling a full lexicon. This is an attractive approach for languages with generally regular letter-to-sound correspondences but with a handful of ambiguous letters: Malay is one example of such a language, but we expect the technique would be more widely applicable.

A baseline system built using our existing tools was evaluated for intelligibility and naturalness against systems built with 3 permutations of the innovative techniques. The experiment and results are presented in Section 5.

2. Baseline voice construction

A baseline voice was built from purpose-collected studio data using the procedures described in [1]. The data and techniques used to build this voice will be briefly outlined in this section.

2.1. Speech data collection

The studio data was obtained as described in [4]. The recording script was handcrafted: starting from a word list (from a pre-existing lexicon), text covering this vocabulary was gathered from various sources such as online news in Malay and a Malay language textbook. A recording script consisting of 900 sentences was composed using this text as a starting point. The sentences were not obtained by directly taking sentences from the text; instead, some existing sentences were edited to cover more words in the lexicon, and 250 of the sentences in the script were composed from scratch to further improve coverage of words in the lexicon. A Malay native male speaker was hired to record the resulting script, resulting in around 1 hour of speech material.

2.2. Front End Processing

We adopted the approach to constructing a front end processor described in [5]. The advantage of this method is that it requires no expert knowledge about the target language, such as a pronunciation dictionary, letter to sound rules or a part of speech tagger. Instead, an automatic distributional analysis, involving the construction of Vector Space Models (VSMs) is used to infer linguistic representations from a text corpus. For example, a VSM of letter distributions (i.e., simple counts in various contexts) provides a representation which can take the place of phonetic categories such as vowels, consonants, nasals, etc.

This is a reasonable choice because Malay has a highly regular alphabetic orthography, transparent affixation, simple syllable structure and a straightforward letter-to-sound relationship (with a small number of ambiguities, to which we offer a solution that does not rely on expert knowledge) [6], all of which make it eminently well matched to this approach to front end text processing. For both studio data and found data, the text corpus was manually transcribed and normalised. After normalisation, our text data consisted only of words, whitespace and punctuation symbols, and no abbreviations, numerals and other symbols remain.

2.3. Acoustic model training

Acoustic feature extraction and model training was done in the same way as described in e.g. [2]. Briefly, frames of 60 bark cepstral coefficients extracted from smoothed STRAIGHT [7] spectra were used, together with mel-warped F_0 and STRAIGHT aperiodicity measures in 25 frequency bands with non-linearly varying bandwidths. Dynamic (*delta* and *delta-delta*) coefficients were appended to the static parameters in the normal way.

The model was initialised by training context-independent grapheme models. These monographeme models were then converted into full-context models and retrained. The names of these full-context models encode the textual features described in Section 2.2. To handle problems of data sparsity and to be able to handle unseen contexts at synthesis time, decision trees were used to cluster acoustically similar states which are then tied, after which the tied models are retrained. Two iterations of clustering and retraining were used.

2.4. Synthesis of speech

At run-time, sentences to be synthesised are processed by the front end, resulting in a sequence of context-dependent grapheme labels for each utterance. According to this sequence, the decision trees built during model training are descended and the corresponding acoustic states are concatenated into an HMM of the utterance. A speech parameter generation algorithm [8] is then used to generate spectral and excitation parameters. The STRAIGHT vocoder is then used to generate a speech waveform from these generated parameters.

The two new techniques proposed for incorporation into the basic recipe outlined above are now described.

3. Innovation 1: Speaker diarisation to extract speech from found data

3.1. 'Found' data

The found data for this study is from the website `http://free-islamic-lectures.com` which is a free resource of Islamic teaching including audio recordings. It offers a free download of an audio recording of Al-quran read in intoned Arabic interspersed with a sentence-by-sentence translation into Malay. A total of 60 hours of audio data chunked into 114 files can be found on that website. Excluding the Arabic part leaves approximately 30 hours of Malay speech data. The Malay speech was recorded with an adult male voice, consistent speaking tone and using standard Malay speaking accent suitable for training data. However, the given text does not correspond exactly to the speech with a word accuracy of only approximately 30%. A subset of the text was therefore manually corrected: this is an expensive process and we were only able to obtain 3 hours of Malay speech with corrected transcription.

3.2. Diarisation

To extract Malay speech from the found data, we applied a speaker diarisation technique [9] as it is able to identify speaker homogenous regions throughout the speech data. First, feature extraction is performed on the data using HTK with standard ASR features. A GMM-HMM framework was applied whereby 16 clusters are initialized by dividing the speech frames into 32 uniform parts and using 2 parts (from different points in the data) to initialize each of the 16 GMMs. With these models, segmentation of the entire speech is done using a Viterbi algorithm with a forced minimum constraint of 250 ms. The models are retrained after segmentation followed by a clustering step to merge similar clusters based on the Bayesian Information Criterion (BIC). A penalty factor parameter is not required as the merged models have a complexity equal to sum of the complexity of the models being merged. The iteration terminates at the stopping criterion which is when the BIC scores for all clusters are below 0. Four wav files consist of 7 hours of raw data in total were chunked using diarisation. After the stopping criterion was met, there are 5 to 10 clusters remain in each file. Only the first cluster is the Malay speech that we want. All the other clusters are the intoned Arabic part. As result, 3 hours of Malay speech data were extracted from 7 hours of raw data and were chunked into 564 pieces of smaller data.

3.3. VAD and alignment

To extract transcribed segments of speech suitable for speech synthesizer training, we applied lightly supervised GMM Voice Activity Detection (VAD) [10] and grapheme-based automatic alignment [11] to our speech and text data. The aim of using these techniques is to segment the audio data and confidently match a subset of the resulting chunks with the corresponding text transcription. Our techniques require no prior language-specific knowledge and can be done in a very lightly-supervised manner: the only user input required is to match an initial few utterances to their transcription, which can be done with only some knowledge of the script used. The technique confidently aligned approximately half of the data, resulting in a total of 90 minutes of speech data which were then used for training our TTS systems.

4. Innovation 2: Interactive construction of letter-to-sound rules using active learning

The approach described in [5] and outlined in Section 2 has been shown to be effective for a variety of languages [1]. Because context-dependent models of grapheme-based acoustic units are used, it is at least theoretically the case that any predictable ambiguities in the letter-to-sound mapping (such as single letters that can be pronounced as two or more sounds) can be learned from the examples in the training data. Whilst this is an apparently elegant solution that does not require expert intervention, its performance is limited in practice by two factors. First, the training data are from the speech corpus being used for acoustic model estimation, which limits the number of types and tokens seen. Second, the letter-to-sound model is an integral part of the decision trees used for acoustic model parameter clustering, and these trees may not be the most efficient classifiers for resolving pronunciation ambiguities. Therefore, it may be be necessary to identify and resolve ambiguities using additional measures. Here, we focus on the resolution of one example ambiguity, which we identified by expert listening. Ongoing work is investigating how non-expert listeners could also be used to identify such problems, as well as to resolve them. The method we employ exploits some limited interaction with a native listener, but without incurring the cost – or requiring the expertise – involved in compiling a full pronunciation dictionary.

The relation between graphemes and phonemes in Malay is generally straightforward, but there are a limited number of graphemes with ambiguous pronunciations. One example is the letter <e>, which can correspond either to the phoneme /e/ or to schwa.

The uncertainty sampling approach to active learning was used [12]: several hundred examples of /e/ in different words were collected, and as predictor features neighbouring letters in a 7-letter window were collected. 150 randomly picked examples were initially hand-labelled (using the arbitrary symbols <e> and <é> to disambiguate the pronunciation). A further 200 examples were selected (one at a time) by active learning and presented to the user for labelling. The classifier resulting from the final iteration of active learning was then used to predict the pronunciation of new examples, including in the transcript of speech data for acoustic model training.

The learned classifier achieved an accuracy of 89.83% on the held-out, semantically unpredictable sentences described in Section 5.2.

Note that active learning has been used previously for learning letter-to-sound rules [13]. The difference in the current work is that instead of being asked to supply phonemic transcriptions of lexical items, a possibly technically naive user is instead asked to make a binary choice for each example of a small subset of letter types.

5. Experiment

5.1. Systems built

Four TTS systems were built, covering all combinations of data type (purpose-recorded vs. web-harvested) and letter-to-sound rules type (graphemes with actively learned disambiguation vs. plain graphemes). The four systems are summarised in Table 1. Systems B and D were trained on the 882 utterances of studio data described in Section 2.1, and systems C and E on the 435 utterances of web-harvested data described in Section 3.1. The quantity of data for all systems is comparable (approximately 1 hour). All systems are letter-based: systems B and C are purely letter-based, and systems D and E use the actively learned classifier for predicting the correct pronunciation of <e> described in Section 4.

As a reference 'system', utterances of the natural speech drawn equally from the purpose-recorded and found data and held out of the training set were used. These samples constitute 'system' A.

Table 1: Description of systems built

System	Description
A	Natural Speech
B	Studio data without active learning
C	Found data without active learning
D	Studio data with active learning
E	Found data with active learning

5.2. Evaluation

The synthetic speech from all systems was evaluated in a perceptual test performed by listeners. Two aspects of the speech were evaluated in this study: naturalness and intelligibility. For naturalness, Mean Opinion Score (MOS) was used: listeners were asked to rate the synthetic speech using a 5-point scale (1 for 'completely unnatural' and 5 for 'completely natural'). Natural speech (system A) was included in the naturalness section of the test. To test the intelligibility of the synthetic voices, listeners were asked to transcribe the synthetic speech they heard. In this section, semantically unpredictable sentences (SUS) were generated based on the criteria described in [14]. As naturally spoken SUS were not available, system A was excluded from this part of the evaluation.

5.2.1. Naturalness

For the naturalness part of the evaluation 10 sentence-texts were used; these were provided by the held-out natural utterances of system A. These 10 texts were synthesised with each TTS system, resulting in 50 experimental stimuli.

5.2.2. Intelligibility

10 utterance texts for the intelligibility part of the evaluation were generated for each system from the templates shown in Table 2. These sentences were synthesised by each of systems B–D, resulting in a set of 40 experimental stimuli. The SUS were generated in such a way that they consist mostly of out-of-training-corpus words. Only a few function words appear in both training data and the SUS script (e.g. *dengan*, *dan*, *ini*, *itu*, *ke* and *yang*). SUS were balanced over listeners and systems using a Latin square design. This meant that each listener heard 10 sentences spoken by each system, and heard each sentence text only once.

In short, the complete listening test consisted of 50 sentences to evaluate naturalness and 40 to evaluate intelligibility. In the intelligibility test, listeners were not permitted to listen to a sentence twice; they were exposed to speech synthesized by each system.

28 listeners were hired to listen to the stimuli and rate or transcribe them. The test was conducted via the web: listeners were asked to use headphones to listen to the speech. A subset of 8 listeners were invited into a quiet room to listen to the synthetic speech.

5.3. Result

The result of the naturalness test is shown in Figure 1.

Listeners' transcriptions were aligned with the known reference text of the stimuli, and word error rates (WER) of their transcriptions were computed per system. Figure 2 shows the word error rates (WER) of the systems, illustrating that the active learning improves intelligibility in both the high- and low-quality data scenarios. We performed a Wilcoxon Signed-Rank

Table 2: SUS templates with examples

Sentence type	Template	Example
Intransitive	Noun Det. Verb (intr.) Preposition Adjective Relat. Pronoun Noun	Sistem ini bekerjasama akan penglihatan yang lucu.
Transitive	Noun Adjective Verb (trans) Noun Adjective	Bola sepak tinggi menjamin lumrah keji.
Imperative	Verb (trans.) Noun Conjunction Noun	Dirikan rumah dan rambutan.
Interrogative	Quest. Adv Noun Verb(trans.) Noun Relat. Pronoun Adjective	Bilakah pendingin hawa ambil generasi yang senyap.
Relative	Noun Det. Verb (trans.) Det Noun Relat. Pronoun Preposition Verb(intr.)	Gunung ini menyapu balai bomba yang sungguh pahit.

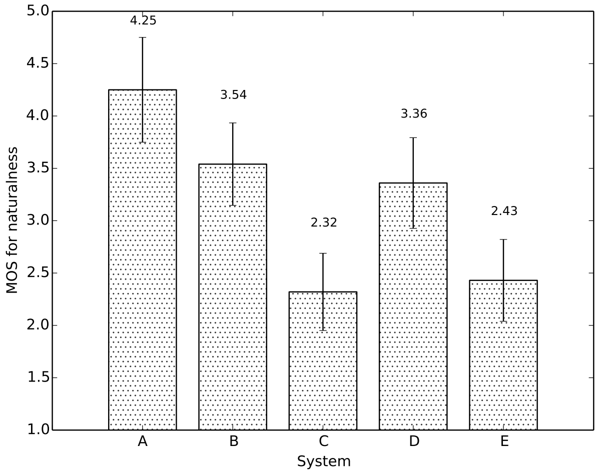

Figure 1: Naturalness of the systems. The use of found data leads to less natural voices, as expected. Active learning, which resolves a letter-to-sound ambiguity, has no effect on naturalness.

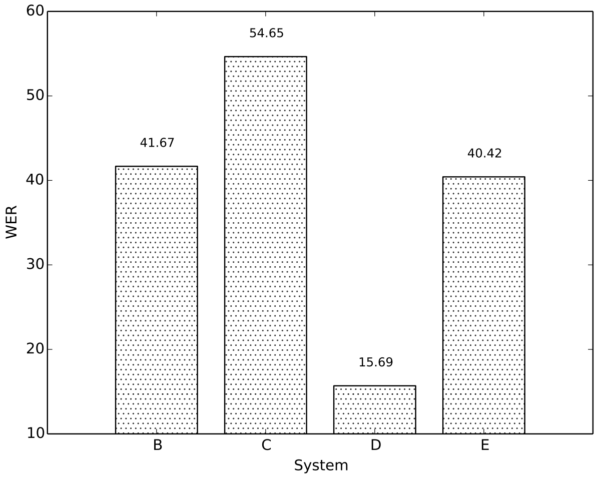

Figure 2: Intelligibility of the systems, expressed as Word Error Rate (lower is better). Active learning, which resolves a letter-to-sound ambiguity has a substantial effect, particularly on the system built from a better quality speech corpus (system D vs system B).

Test ($\alpha = 0.05$) to test the significance of the differences between all pairs of systems. All differences in WER were found to be significant except that between systems B and E, showing that active learning is highly effective in improving intelligibility.

6. Conclusion

We have presented four configurations of a Malay statistical parametric speech synthesizer in this study, which allowed us to evaluate the effect of replacing the standard high-quality studio data with lower-quality 'found' data, and to test a novel active learning technique for pronunciation disambiguation in both data conditions. The Malay language is straightforward in terms of orthography and language structure and in general it poses no major problems in building TTS systems. However, one major difficulty is predicting correct pronunciations for the ambiguous grapheme <e>. In this study, when classifiers trained using the active learning technique were used to predict the correct pronunciation of this letter in words encountered at run-time, we found significant improvements to intelligibility for systems trained on both types of data.

7. Acknowledgements

We thank Mark Sinclair for helping us apply his speaker diarisation tools to our data. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N$^{\underline{o}}$ 287678 (Simple4All).

8. References

[1] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, "Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 121–126.

[2] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS System for Blizzard Challenge," in *Proc. Blizzard Challenge 2010*, Sep. 2010.

[3] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. Clark, J. Yamagishi, and S. King, "TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision," in *Proc. Interspeech*, Lyon, France, August 2013.

[4] L. C. Yong and T. T. Swee, "Low footprint high intelligibility malay speech synthesizer based on statistical data," *Journal of Computer Science*, vol. 10, pp. 316–324, 2014.

[5] J. Lorenzo-Trueba, O. Watts, R. Barra-Chicote, J. Yamagishi, S. King, and J. M. Montero, "Simple4all proposals for the albayzin evaluations in speech synthesis," in *In Proc. Iberspeech*, 2012.

[6] M. J. Yap, S. J. R. Liow, S. B. Jalil, and S. S. B. Faizal, "The malay lexicon project: A database of lexical statistics for 9,592 words," *Behavior research methods*, vol. 42, no. 4, pp. 992–1003, 2010.

[7] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “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 Communication*, vol. 27, pp. 187–207, 1999.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis,” in *In: Proc. ICASSP*, 2000, pp. 1315–1318.

[9] M. Sinclair and S. King, “What are the challenges in speaker diarization?” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[10] Y. Mamiya, J. Yamagishi, O. Watts, R. A. Clark, S. King, and A. Stan, “Lightly supervised gmm vad to use audiobook for speech synthesiser,” in *In Proc. ICASSP*, 2013.

[11] A. Stan, P. Bell, J. Yamagishi, and S. King, “Lightly Supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data,” in *Proc. of Interspeech (accepted)*, 2013.

[12] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*.

[13] K. Dwyer and G. Kondrak, “Reducing the annotation effort for letter-to-phoneme conversion,” in *ACL/IJCNLP*, 2009, pp. 127–135.

[14] C. Benoit, M. Grice, and V. Hazan, “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,” *Speech Communication*, vol. 18, no. 4, pp. 381 – 392, 1996.