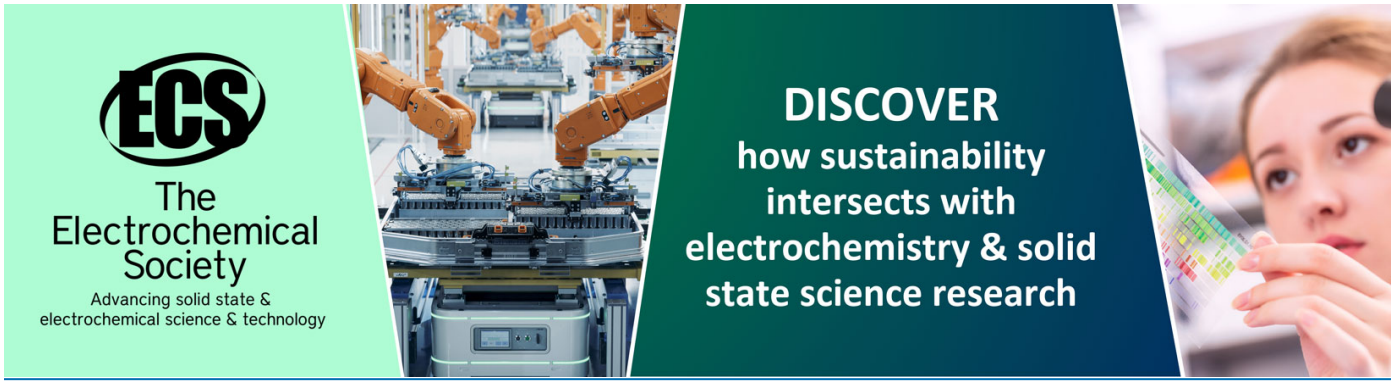
An Indonesian speech synthesis system considering low-power event before vowel-starting-syllable



IOP Conf. Series: Journal of Physics: Conf. Series **1075** (2018) 012034 doi:10.1088/1742-6596/1075/1/012034

An Indonesian speech synthesis system considering low-power event before vowel-starting-syllable

I Setiawan and T Hirai

Arcadia, Inc.

Minoh Sunplaza 1, 7F
6-3-1 Minoh, Minoh-city, Osaka 562-0001, Japan

Email: ivan@arcadia.co.jp

Abstract. In our Indonesian concatenative speech synthesis system, there existed a linking issue between vowel-starting-syllable with its preceding phoneme. Unlike other languages, in Indonesian the power of the speech signal decreases (sometime it is a pause) in this kind of boundary. Phonemes before and after the boundary are uttered separately. For example, the underlined phonemes before and after the boundary in "buah apel (apple fruit)" are not to be uttered continuously, but should be separated. This is not the case in English, where the underlined phonemes in "an apple" are linked. We did not treat this kind of low-power event ("lpow") explicitly, such that the lpow generated indirectly from the syllable boundary information, is sometime too short, resulting in the above linking issue. In this paper, we propose to explicitly treat the lpow. The lpow is treated similarly with phoneme during the model training, so that it is appropriately generated during the synthesis. We confirmed that the synthesized speech is more natural by the introduction of lpow.

1. Introduction

Arcadia, Inc. provides services to deliver disaster prevention/reduction messages through e-mail, FAX, telephone, and public announcement speakers. Thus, speech synthesis system with a high broadcast-quality is required.

An Indonesian language (Indonesian: "Bahasa Indonesia") speech synthesis system has been proposed by Sakti et al [1]. Since this system is a statistical parametric synthesizer that synthesizes speech from feature vector, quality improvement by the adoption of concatenative synthesis method [2, 3], was required.

We have already developed an Indonesian concatenative speech synthesis in [4]. By utilizing a pronunciation dictionary and pronunciation rules, this speech synthesizer can differentiate between the pronunciation of [@] and [e] (X-SAMPA [5]) which are represented with identical "e" character in the text. The system in [4] considered only the phoneme substitution cost and concatenative distortion, but did not include F0 feature when computing target cost during unit selection. This resulted in a rising end -of- sentence intonation in some synthesized speech of declarative sentences (so it sounds like interrogative sentences) which is caused by units with relatively high F0 was selected.

An HMM-based model training method to obtain a model that generates feature vector target based on linguistic information such as phoneme environment (environment where the phoneme appears, such as adjacent phonemes), number of phonemes in phrases, etc., is already released and available for

public use [6]. We incorporated this method to provide a model with a correct F0 target, so that units with appropriate F0 are selected [7]. Rising intonation was solved, however, linking of vowel-starting-syllable/word with the preceding syllable/word during pronunciation occurred.

This paper presents our solution to the linking issue. In section 2, we briefly explain about the concatenative speech synthesis. Then, the main theme in this paper, i.e. the linking issue and its solution by introducing low-power event, are described in section 3. Section 4 describes our Indonesian speech corpus and some narration styles that we found while constructing the corpus. Some experiments to check the effect of the low-power event and narration style classification are shown in section 5, and we provide our summary in section 6.

2. Indonesian concatenative speech synthesis system

This section briefly describes the Indonesian concatenative speech synthesis that we have reported in [4] and [7].

As in many systems, there are 2 parts in the system, that is, the analysis part and the synthesis part. The analysis part, or more widely known as speech corpus building, is done only once during the system construction. On the other hand, the synthesis part is called each time a text needs to be synthesized as a speech waveform.

In the corpus building, speech with transcription (text) is analyzed. Transcription texts are converted to phoneme sequences, then a phoneme acoustic model trained using speech and the phoneme sequences, is generated. A forced alignment technique was used to obtain the start/end point of each phoneme in the speech. When the analysis finished, we obtain the corpus, the acoustic model, the F0 control model, and the duration model. The corpus stores phoneme environment, start/end point, and acoustic features (F0, power, cepstrum) of each phoneme. The model will be used to estimate acoustic features from the phoneme environment.

To synthesize text to a speech, first the input text is converted to phoneme sequence. Then, the acoustic features (target features, F0 etc.) are generated using the trained model. Let s be sequence of unit, then an appropriate unit sequence s^* which has features similar to the target, i.e., low target cost $C_t(s)$, and shows the least concatenative distortion, i.e., low concatenation cost $C_c(s)$, was selected from the database and was concatenated into a waveform. Written formally:

$$s^* = argmin_s\{w_t C_t(s) + w_c C_c(s)\} \quad (1)$$

where w_t and w_c are weights that control the contribution of the target cost and the concatenation cost to the overall cost, respectively.

In the system in [4], the target cost only considered the phoneme environment disagreement cost (phoneme substitution cost) . In the system in [7], the target cost also considered the F0 target cost. Both implementations have their own issues, but in comparison with the statistical parametric synthesis which gives muffled sound, the concatenative synthesis produces a crisp and clearer sound, satisfying the broadcast quality requirement for a public announcement system.

3. Low-power event

When saying Indonesian word or syllable starting with vowel, it is natural to put a brief pause or decrease the speech signal power, before the vowel. For example, phrases like "buah apel (apple fruit)" and "bawah atas (bottom, top)" is better to utter as [buah ap@l] and [bawah atas], instead of [bua hap@l] and [bawa hatas] (a space inside the phoneme sequence, showed inside square brackets "[]", implies a brief pause or decreasing of the speech signal power). The latter case is what we refer as the linking issue, where the vowel (the first phoneme, the vowel [a] of the second word) is linked to the final phoneme of the preceding word (the final phoneme [h] of the first word).

The linking issue occurs not only before vowel-starting-word, but also before vowel-starting-syllable, especially when uttering abbreviations. For example, "SNI (Standar Nasional Indonesia, Indonesian National Standard)" and "PMI (Palang Merah Indonesia, Indonesian Red Cross)" which should utter as [es en i] and [pe em i], might be incorrectly synthesized as [e se ni] and [pe emi] due to the linking issue.

The linking issue may or may not occur in our previous system described in [7], depends on the existence of the phrases in the corpus. If it is part of the training data, then this issue does not happen in our previous system. If it is not, then the linking will happen because the brief stop or the decrease in the speech signal power before vowel -starting-word/syllable was not explicitly treated. In this paper, we propose to explicitly treat this event by introducing a low- power event "lpow" to indicate a brief pause or a decrease in the speech signal power. Beware that lpow is not a pause event like a comma or a period, neither it is a phoneme. Since the lpow event is phonetically meaningful and can be considered during the model training, an lpow event is expected to be generated before the vowel-starting-syllable when using the trained model.

Incorporating the lpow event, "bawah atas", "SNI", and "PMI" will have phoneme sequences of [bawah lpow atas], [es lpow en lpow i], and [pe lpow em lpow i], respectively.

4. Indonesian speech corpus

Indonesian speech data were recorded in 44.1 kHz sampling and 16-bit integer wav-format, from a female native Indonesian speaker who teaches Indonesian language and has been living in Japan over 20 years. Phrases included proper nouns and text samples from Indonesian language learning books, newspaper articles, and TV scripts, which covered almost all possible Indonesian syllables. There were 4,490 phrases lasting a total of 5.1 hours. Pauses were notated in the transcription by listening to the recorded speech.

During corpus checking, we noticed that in our corpus there are 2 types of end-of-sentence's intonation for question (interrogative sentence). These 2 types of intonation were due to the narration styles, i.e., prose (read speech) and dialog (conversational speech). The prose style question ends with a rising intonation as shown in figure 1 (the text is "Siapa anda? (Who are you?)"), while the dialog style question ends with a long and flat intonation as shown in figure 2 (the text is "Wanita yang rambutnya pendek itu siapa? (Who is that short-haired woman?)"). Phoneme boundaries in both figures 1 and 2 are obtained by forced alignment.

Since rising and flat intonations will result in a different F0 feature, we manually classify and assigned different type of question mark for each style in the corpus; and we expect to obtain a better F0 model.

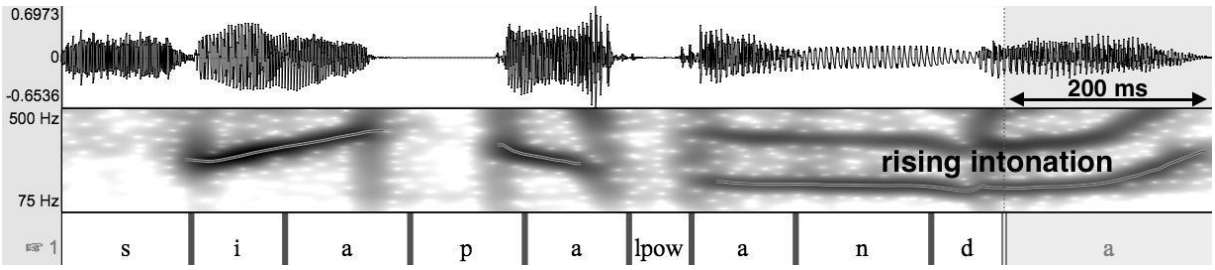


Figure 1. Prose style question "Siapa anda?" in corpus. Top to bottom: waveform, F0 pattern on narrow band spectrum, and phonemes with boundaries. Rising intonation in the shaded part.

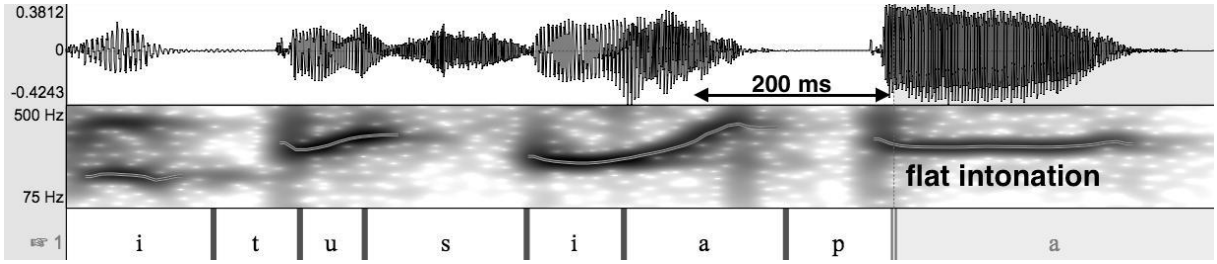


Figure 2. Dialog style question "Wanita yang rambutnya pendek itu siapa?" in corpus. Top to bottom: waveform ("itu siapa?" part), F0 pattern on narrow band spectrum, and phonemes with boundaries. Flat intonation in the shaded part.

5. Experiments and results

Two examples where low-power event improves the naturalness of the synthesized speech are shown below. Then, we also mention the effect of classifying type of question sentence in the corpus.

5.1. Effect of low-power event "lpow" in the case where the lpow exists before a WORD starting with vowel

The input text is "bawah atas" (phoneme sequence: [bawah atas]). A vowel-starting-word "atas (top)" is following the first word "bawah (bottom)". Figure 3 and 4 shows the full waveform of the synthesized "bawah atas" before and after the introduction of lpow event, respectively.

In figure 3, it is clear that the first phoneme [a] of the succeeding word is linked to the preceding word's final phoneme [h]. The speech was sound as [bawahatas]. Actually, the phoneme [h] and [a] were both selected from the continuing units [ha] in the word "kejahatan (crime)" in the corpus, where [h] and [a] are both in the same syllable.

In figure 4, the lpow correctly separates the [bawah] and [atas] such that the speech was clearly sound as [bawah atas]. The lpow here resembles a brief pause.

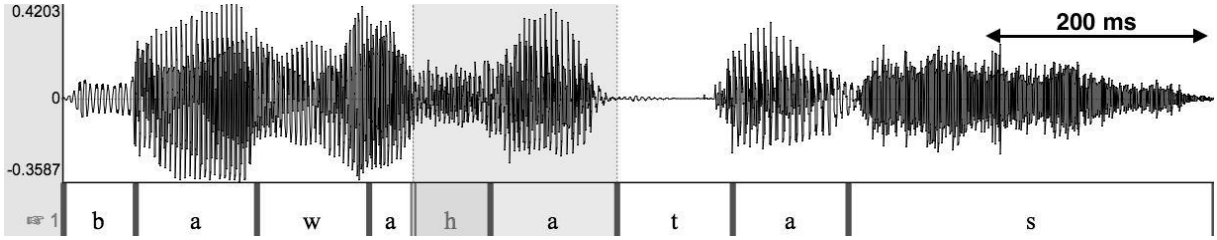


Figure 3. Full waveform of the synthesized "bawah atas" before the introduction of lpow. Top: waveform, bottom: phonemes with boundaries. Linking of [ha] in the shaded part.

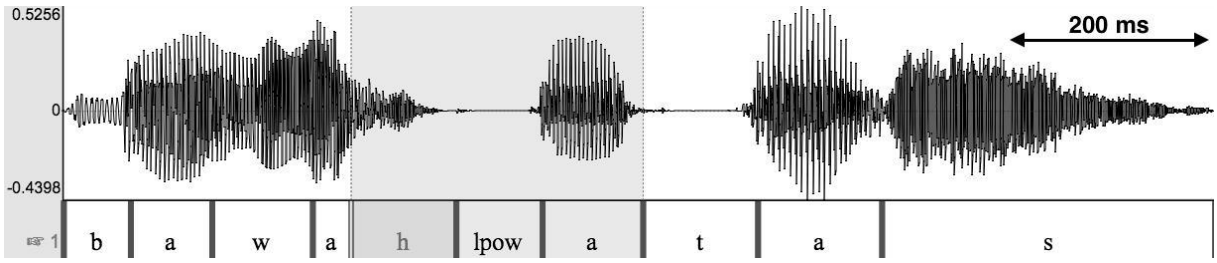


Figure 4. Full waveform of the synthesized "bawah atas" after the introduction of lpow. Top: waveform, bottom: phonemes with boundaries. An appropriate lpow is selected, thus no linking of [h] and [a] in the shaded part.

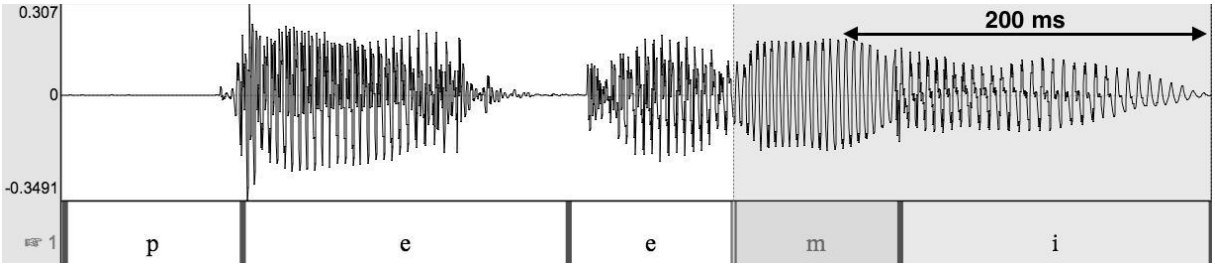


Figure 5. Waveform of the synthesized "PMI" before the introduction of lpow. Top: waveform, bottom: phonemes with boundaries. Linking of [mi] in the shaded part.

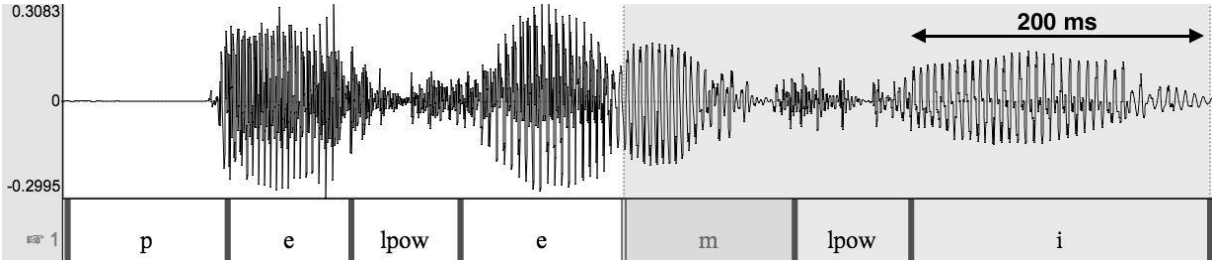


Figure 6. Waveform of the synthesized "PMI" after the introduction of lpow. Top: waveform, bottom: phonemes with boundaries. An appropriate lpow is selected, thus no linking of [m] and [i] in the shaded part.

5.2. Effect of low-power event "lpow" in the case where the lpow exists before a SYLLABLE starting with vowel

The input text is "Palang Merah Indonesia disingkat PMI. (Indonesian Red Cross is abbreviated as PMI.)", and we pay attention to the vowel- starting-syllable [i] in the abbreviation "PMI [pe em i]". Figure 5 and 6 shows the waveform of the synthesized "PMI" before and after the introduction of lpow event, respectively.

In figure 5, the [e] of [pe] and its succeeding [e] of [em] are separated because there is such a unit in the corpus just by coincidence. The [i], however, is linked to the preceding [m], such that the speech was sound as [pe emi]. Similar to the case in the previous example, the phoneme [m] and [i] were both selected from the continuing units [mi] in the word "kami (us)" in the corpus, where [m] and [i] are both in the same syllable.

In figure 6, lpow correctly separates the [em] and the [i] such that the speech was clearly sound as [em i], instead of [emi]. The lpow here is a low-power speech signal, not a brief pause.

5.3. Effect of classifying types of question in the corpus

In section 4, we have described the classification of 2 types of question sentence. However, we have found no good evidence whether this classification really improves the naturalness of the question speech synthesis. This is probably due to the small number of the question sentence in the corpus, which are only 121 (prose style questions: 71, dialog style questions: 51; One phrase includes two question sentences with different type of question.) out of 4,490 phrases. The trained model related to question might be unstable.

6. Summary

We have introduced the low-power event before vowel-starting-word/syllable and showed through examples that its effect is as expected. We also differentiate type of question sentence based on its

end-of-sentence intonation, however no improvement was observed which maybe due to the small number of question sentences in the corpus.

References

[1] Sakti S, Maia R, Sakai S, Shimizu T and Nakamura S 2008 Development of HMM-based Indonesian Speech Synthesis *Oriental COCOSDA*

[2] Kawai H, Toda T, Ni J, Tsuzaki M and Tokuda K 2004 XIMERA: A New TTS from ATR Based on Corpus-Based Technologies *ISCA 5th Speech Synthesis Workshop*

[3] Gonzalvo X, Tazari S, Chan C, Becker M, Gutkin A and Silen H 2016 Recent Advances in Google Real-time HMM-driven Unit Selection Synthesizer *INTERSPEECH*

[4] Hirai T and Setiawan I 2016 An Indonesian Concatenative Speech Synthesis System *5th Joint Meeting of ASA and Acoustical Society of Japan (ASJ)*

[5] https://ja.wikipedia.org/wiki/X-SAMPA

[6] http://hts.sp.nitech.ac.jp/

[7] Hirai T and Setiawan I 2017 An Indonesian Concatenative Speech Synthesis System (in Japanese) *ASJ Autumn Meeting*

Acknowledgment

The authors wish to thank all the members of Arcadia, Inc. for all their supports during this work.