Utilizing Indonesian Allophones and Intraword Glottalization Handling to Improve Performance of Indonesian Text-To-Speech

Mohammad Teduh Uliniansyah*, Lyla Ruslana Aini, Elvira Nurfadhilah, Siska Pebiana, Harnum Annisa Prafitia, Made Gunawan, Agung Santosa, Asril Jarin, Gunarso, Fara Ayuningtyas, Hammam Riza
Center of Information and Communication Technology
Agency for the Assesment and Application of Technology (BPPT)
Jakarta, Indonesia
*e-mail : teduh.uliniansyah@bppt.go.id

Abstract

An allophone is a phoneme variant based on the position within a word, for instance, the first phoneme e in "pendekar" is pronounced differently from the second phoneme e. According to Badan Pengembangan dan Pembinaan Bahasa (Language Development and Fostering Agency), Bahasa Indonesia has five vowels, and 22 consonants, with 6 of them have allophones. There are only allophones of the phoneme e that can change the meaning of a word. Meanwhile, the allophones of the other five phonemes are not changing the words' meanings. Therefore, most researches/projects on developing an Indonesian text-to-speech (TTS) system focus only on the allophones of the phoneme e. This paper reports our new experiment and compares the results with our previous work. In our latest work, there was an additional of 5 new allophones and a different objective evaluation method, which is PESQ (Perceptual Evaluation of Speech Quality). The new experiment shows better results compared to the previous one.

Keywords

Indonesian language, Allophone, Glottalization, PESQ, Text-to-speech, Natural language processing, Deep neural network.

1. Introduction

The existence of free trade policies in Asia in recent years has made higher demands for

human resources with proper knowledge, technical, and communication skills. However, there are still obstacles for Indonesians to communicate well using foreign languages such as English. Therefore BPPT, as one of the research institute in Indonesia, is concerned to resolve this issue by developing a speech-to-speech translation [A. Santosa et al., 2016] (S2ST) system. The speech-to-speech translation is a system that includes several components, and one of those is a text-to-speech system.

There was some research for developing an Indonesian text-to-speech system. [Tritoasmoro, 2006] used a diphone concatenation technique, and [Jangtjik et al., 2014] used a parametric statistics technique. BPPT began doing the text-to-speech research in 2009, started with a study with diphone concatenation technique [Charpentier et al., 1986], followed by a parametric statistical technique ["HMM/DNN-based speech synthesis system", 2018] and finally use the Deep Neural Network (DNN) with end-to-end method [Wang et al., 2017].

In our previous studies [Uliniansyah et al., 2018], we used the DNN approach with the addition of allophones using Tacotron, an end-to-end method. The result showed better results than DNN without allophone, but it was still not so natural and clear. A possible solution is to use more allophones than the ones that the Language Development and Fostering Agency of Indonesia introduced. Therefore, we implemented a set of new allophones of phonemes b, d, and g at the end of words, for instance, the words *adab, bab, wahid, sujud, gudeg, budeg,* etc. We also exploit the insertion of allophone *w* and *y* between two vowels in a root word like *buah, buih, kuah, diam, kuali*, etc.

In addition, we implemented a new symbol for representing glottalization occurred within a word such as pronouncing following words: "*diindahkan*", "*kuamankan*", "*Qur'an*", and "*bapak*". Glottalization occurs between two vowels in Indonesian words, borrowed words from a foreign language or words ending with the letter k. The addition of new allophones or glottalization was based on [Muslich M., 2008] and [Alwi H. et al., 2003].

Differs from previous research, we did subjective testing with new testers. Members of the tester team are those not involved in the development of the S2ST system so that the results of the experiment can be more objective. We used a different objective evaluation method, which is the perceptual evaluation of speech quality (PESQ) by [Beerends et al., 2005]. The PESQ method is suitable to measure the level of the audio quality produced,

which is closely related to the clarity and naturalness of the resulting pronunciation.

This paper is organized into the following sections. Section 2 describes rules to identify allophones, handling intraword glottalization, and the experiment to validate the proposed method. The experimental results are presented and discussed in section 3. Section 4 concludes this paper.

2. Methodology

2.1. Rules to identify allophones

Based on examples given by Language Development and Fostering Agency [D. Sugono et al., 2004], we defined simple rules to identify allophones in a word by looking at the position of the allophone in the word. Table 1 summarizes the rules to identify allophones.

The rule notations were written using Perl regular expression notation. There are several borrowed words from Arabic language having glottalization also, but the glottalizations are not between two vowel phonemes such as *jumat*. In our lexicon, these words are listed manually since their number is not many.

In addition, we also introduced new rules to identify glottalization between two vowel phonemes found in words such as *keadaan, diumpamakan, saat*, etc., and the rule was /[aiueo][aiueo]/. An "X" is used as a symbol to represent intraword glottalizations.

We found that there are two situations where two adjacent vowel phonemes create glottalizations :

- Two vowel phonemes are the same, such as in the following words: "*kenyataan*", "*cemooh*", etc.
- Two vowel phonemes are not the same, but the word is a derivative created by concatenating prefixes "*ke*", "*di*", "se*",* or clitics "*ku*", or "*kau*", such as in words "*keadaan*", "*diumumkan*", "*seukuran*", "*kuamati*", "*kauamati*", etc.

Phoneme	Allophone	Word examples	Pronunciation notations	Rule	Pronunciation symbol
a	a	hutan jatah hijau	[hutan] [jatah] [hijau]		a

Phoneme	Allophone	Word examples	Pronunciation notations	Rule	Pronunciation symbol
	[ʷa]	**kuat** **buat** **peluang**	**[kuʷat]** **[buʷat]** **[peluʷaŋ]**	**/ua/**	**W**
e	[ɛ]	enak merah sore	[enaʔ] [merah] [sore]		e
	[ə]	empat beri ritme	[əmpat] [bəri] [ritmə]		B
	[ɛ]	**kakek** **nenek** **bebek**	**[kakɛʔ]** **[nɛnɛʔ]** **[bɛbɛk]**	**/e[^aiueo][^aiueo]*e[^aiueo]/**	**E**
h	glottal	hutan jatah hijau	[hutan] [jatah] [hijau]		h
	glottal stop	**tahu** **tahun**	**[tau]** **[taun]**	**/ahu[^aiueo]*$/**	**H**
i	[i]	itu babi teliti	[itu] [babi] [təliti]		i
	[ɪ]	**batik** **bangkit** **culik**	**[batɪk]** **[baŋkɪt]** **[culɪk]**	**/[aiueo][^aiueo]+i[kt]$/**	**L**
	[iʸ]	**riuh** **diam** **siar**	**[riʸuh]** **[diʸam]** **[siʸar]**	**/i[aueo]/**	**Y**
k	glottal	kita makan ombak	[kita] [makan] [ombak]		k
	glottal stop	**bapak** **tidak** **lunak**	**[bapaʔ]** **[tidaʔ]** **[lunaʔ]**	**/[aiueo]k$/**	**q**

Phoneme	Allophone	Word examples	Pronunciation notations	Rule	Pronunciation symbol
o	[o]	toko roda olahraga	[toko] [roda] [olahraga]		o
	[ɔ]	**bonus** **odol** **bodoh**	**[bɔnus]** **[ɔdɔl** **[bɔdɔh]**	**/o[^aiueo][^aiueo]*o[^aiueo]$/**	**0**
u	[u]	ubah baru susu	[ubah] [baru] [susu]		u
	[ʊ]	**sahur** **rambut** **lembur**	**[sahʊr]** **[rambʊt]** **[ləmbʊr]**	**/[aiueo][^aiueo][^aiueo]*u[^aiueo]$/**	**v**
b	[b]	buku baru butuh	[buku] [baru] [butuh]		b
	[p˃]	**sebab** **jilbab** **adab**	**[səbap˃]** **[jilbap˃]** **[adap˃]**	**/b$/**	**P**
d	[d]	dunia aduh dua	[dunia] [aduh] [dua]		d
	[t˃]	**abad** **wahid** **jilid**	**[abat˃]** **[wahit˃]** **[jilit˃]**	**/d$/**	**D**
g	[g]	gelap gurita gaduh	[gelap] [gurita] [gaduh]		g
	[k˃]	**gudeg** **bedug** **ajeg**	**[gudek˃]** **[beduk˃]** **[ajek˃]**	**/g$/**	**C**

Table 1: Rules to identify allophones

In developing an Indonesian TTS system, we have been maintaining a lexicon consisting of words and their pronunciations. Rows with bold fonts in Table 1 shows the allophones not yet recorded in the lexicon. Up to now, there are around 254 thousand data in the lexicon.

To apply the newly introduced allophones, new symbols which meet the following two requirements were used:

- Can illustrate the represented allophone, and
- Not used in the existing lexicon.

After building the rules for identifying allophones, pronunciation symbols in the lexicon file were changed based on the rules.

2.2. Handling intraword glottalization

Concatenating a root word (with a vowel as its first character) with prefixes "*di*", "*ke*", or "*se*", or clitics "*ku*" or "*kau*" creates glottalizations between the first pair of vowels in the word. For instance "*diampuni*", "*keadaan*", "*seandainya*", etc. However, there are many Indonesian root words with "*di*", "*ke*", "*se*", "*ku*", or "*kau*" as the first characters in the word, such as "*diam*", "*kuasa*", "*kaum*", etc., and there are no glottalizations between the first vowel pair of those words.

We used a morphological analysis program [Uliniansyah et al., 2004] that we developed earlier to identify whether a word begins with "*di*", "*ke*", "*se*", "*ku*", or "*kau*" followed by a vowel is a derivative or not.

It is worth noted that in pronouncing those derivatives, whether glottalizations exist or not between the first pair of vowels will not change the words' meanings. Here, the glottalization functions as an emphasis on the meaning of the word.

2.3. Training Process

The modeling of text to speech synthesis was done by using the implementation of Tacotron model developed by Mozilla ["Deep learning for Text to Speech", 2019] while the previous work used Tacotron model developed by Keithito ["A TensorFlow implementation", 2018]. The training process needed speech corpus and metadata as the training data. The difference with our previous TTS system is that the training data are generated with enhanced rules, which include the five new allophones.

Table 2 shows summary of the training process comparing previous and current work for female TTS model.

Parameter	With Allophone and intraword glottalization	
	Previous Experiment	**Current Experiment**
Number of Utterances	15,140	15,354
Sample rate of the wav file	22,050 Hz	16,000 Hz
Precision of the wav file	16-bit	16-bit
Batch size	32	32
Step number	1 - 83000	1-168000
Validation Loss	0.08106	Totalloss : 0.10821, PostnetLoss : 0.05027, DecoderLoss : 0.05235, Stoploss : 0.00559
Training duration	2 – 3 days	5 - 6 days

Table 2: Summary of the training process comparing previous and current work

3. Results and discussion

In evaluating this TTS system (female voice model), we have carried out subjective and objective evaluation as done before to previous TTS system [A. Santosa et al., 2018]. Subjective evaluation was done by using semantically unpredictable sentences (SUS) [Benoit et al., 1996] and mean opinion score (MOS) methods to assess the performance of the TTS regarding of intelligibility and naturalness aspects.

The evaluation using SUS method was performed by generating synthesized speeches of several meaningless but syntactically correct sentences, whereas the evaluation with MOS method used the speech synthesis results from several sentences taken from an Indonesian language newspaper site.

The purpose of making SUS random sentences is to make sure that evaluators were unable to guess words based on the sentence context.

The following are some examples of the SUS sentences:

- *siang itu kakek menghimbau malaikat yang ibu diadakan ini*

- *wahai abjad klausul itu menjuntai bagaikan guntai*
- *assalamu'alaikum murid sangkaan ini menjuntai dengan hijau*
- *kuingat ketuaan atau ayo otentik menjuntai*
- *kuambil keutamaan karena sujud jalan gemulai menjuntai*

The set of SUS sentences used in the subjective evaluation is the same as the one used in the previous work. However, we inserted words with the new five allophones into the SUS sentences set.

Evaluators listened to the synthesized speeches and wrote down what he/she heard, which then compared to the SUS sentences. The evaluation was done by counting the word error rate (WER), and the results are showed in Table 3. In that table, the test result from our previous TTS is also represented as the comparison.

TTS model	**Number of Evaluators**	**Results**
Female voice; previous experiment	10	TOTAL Words: 760 Correct: 611 (80.39%) Errors: 155 (20.39%) **Accuracy : 79.61%** Insertions: 6 Deletions: 13 Substitutions: 136
Female voice; current experiment	10	TOTAL Words: 840 Correct: 712 (84.76%) Errors: 135 (16.07%) **Accuracy : 83.93%** Insertions: 7 Deletions: 18 Substitutions: 110

Table 3: Results of subjective evaluation (Intelligibility Aspect) with SUS method

Similar to [Uliniansyah et al., 2018], we did the same subjective evaluation by using MOS method. Evaluators agreed that there is a slight improvement in naturalness aspect. The MOS evaluation score for the latest experiment is 4.2, while the score for previous work is 4.1.

From the evaluation results, it can be noted the following:

- As shown in Table 3, the TTS models of our current work showed better intelligibility compared with the previous TTS model, indicated by the lower error rate (16.07% against 20.39%).

- The insertion value indicates the number of words that cannot be captured (understood) at all by the evaluators. Our previous work shows that there are 6 words (out of 760 words) or nearly 0.79% while our recent system shows that there are 7 words (out of 840 words) or nearly 0.83%.
- The deletion value indicates the total number of words that are completely wrong captured or understood. Our previous work shows that there are 13 words (out of 760 words) or nearly 1.7% while our recent system shows that there are 18 words (out of 840 words) or nearly 2.1%.
- The substitution value shows the number of words that are almost correctly understood by the evaluators (but still considered as wrong). Our previous work shows that there are 136 words (out of 760 words) or nearly 17% while our recent system shows that there are 110 words (out of 840 words) or nearly 13.1%.
- Although the number of insertion and deletions of the current system is slightly higher than that of the previous work, the overall number of errors of the current system is considerably better.

Based on the above subjective test results, utilizing the new five allophones improved the intelligibility aspects of our TTS system.

In objective evaluation, we used PESQ (perceptual evaluation of speech quality) method [ITU-T P.862, 2001]. The PESQ method is an objective method designed for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. It is standardized as ITU-T recommendation P.862. We used PESQ method developed by British Telecommunications ["ITU-T_pesq", 2019]. The evaluation result is shown in Table 4.

TTS model	PESQ MOS	PESQ MOS LQO
Female previous experiment	0.489	1.124
Female current experiment	1.000	1.201

Table 4: Results of objective evaluation with PESQ method

There were two parameters in the objective evaluation using the PESQ method: PESQ MOS [ITU-T P.862, 2001] with scale of -0.5 to 4.5 and PESQ MOS LQO (listening quality objective) [ITU T P.862.1, 2003] with scale of 1 to 5 (higher value means better). PESQ MOS LQO is a score of transforming the value of MOS using a mapping function.

As shown in Table 4, the value of parameters in the current TTS system is better than the previous one.

4. Conclusion

This paper presented experimental (female voice) TTS Systems for Indonesia language that have utilized allophones and intraword glottalization. The subjective evaluations using SUS and MOS methods showed that this TTS system was more intelligible and more natural than our previous TTS system. This was indicated by less number of word errors. Both TTS system implemented allophones and intraword glottalization, but the new system introduced five new allophones. Objective evaluation of this TTS system using PESQ method shows that the new system is better than the previous one.

5. References

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, A. M. (2003). Tata Bahasa Baku Bahasa Indonesia Edisi Ketiga. Jakarta: Balai Pustaka.

A. Santosa et al., "Utilizing Indonesian Data Resources for Text-To-Speech Using End-To-End Method,". (2018). In The 21st Conference of the Oriental COCOSDA.

"Arti kata alofon - Kamus Besar Bahasa Indonesia (KBBI) Online." [Online]. Available: https://www.kbbi.web.id/alofon. [Accessed: 20-Jul-2018].

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication, 18(4), 381-392.

Beerends, J. G., van Wijngaarden, S., & van Buuren, R. (2005). Extension of ITU-T recommendation P. 862 PESQ towards measuring speech intelligibility with vocoders. TNO TELECOM DELFT (NETHERLANDS).

Charpentier, F., & Stella, M. (1986, April). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 11, pp. 2015-2018). IEEE.

D. Sugono et al. (2004). Lentera Indonesia Pemula Penerang untuk Memahami Masyarakat dan Budaya Indonesia. Pusat Bahasa.

"Home - HMM/DNN-based speech synthesis system (HTS)." [Online]. Available: http://hts.sp.nitech.ac.jp/. [Accessed: 20-Jul-2018].

ITU-T P.862 (02/2001) - "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs."

ITU T P.862.1 (2003), Mapping function for transforming P.862 raw result scores to MOS-LQO.

"ITU-T_pesq" [Online]. Available: https://github.com/dennisguse/ITU-T_pesq. [Accessed: 18-Jul-2019].

Jangtjik, K. A., & Lestari, D. F. (2014, November). The Indonesian Language speech synthesizer based on the Hidden Markov Model. In 2014 International Conference on Electrical Engineering and Computer Science (ICEECS) (pp. 12-16). IEEE.

Keithito, "A TensorFlow implementation of Google's Tacotron speech synthesis with pre-trained model." [Online]. Available: https://github.com/keithito/tacotron. [Accessed: 23-Jul-2018].

Mozilla, "Deep learning for Text to Speech" [Online]. Available: https://github.com/mozilla/TTS. [Accessed: 25-Jul-2019].

Muslich, M. (2008). Fonologi Bahasa Indonesia: Tinjauan Deskriptif Sistem Bunyi Bahasa Indonesia. Bumi Aksara.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001, May). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 2, pp. 749-752). IEEE.

Tritoasmoro, I. I. (2006, November). Text to Speech Bahasa Indonesia Menggunakan Concatenation Synthesizer Berbasis Fonem. In Seminar Nasional Sistem dan Informatika (pp. 171-176).

Uliniansyah, M. T., Ishizaki, S., & Uchiyama, K. (2004). Solving Ambiguities in Indonesian Words by Morphological Analysis Using Minimum Connectivity Cost. Journal of Natural Language Processing, 11(1), 3-20.

Uliniansyah, M. T., Nurfadhilah, E., Annisa, H., Gunawan, M., Aini, L. R., Santosa, A., ... & Riza, H. (2018, November). Utilizing Indonesian Allophones and Intraword Short Pauses Handling to Improve Performance of Indonesian Text-To-Speech. In 2018 International Conference on Asian Language Processing (IALP) (pp. 143-146). IEEE.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Le, Q. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.