pISSN 2005-8063
eISSN 2586-5854
2021. 9. 30.
Vol.13 No.3
pp. 65-70

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2021.13.3.065

Designing a large recording script for open-domain English speech synthesis*

Sunhee Kim[1,**] · Hojeong Kim[2] · Yooseop Lee[1] · Boryoung Kim[1] · Yongkook Won[3] · Bongwan Kim[4]

[1]*Department of French Language Education, Seoul National University, Seoul, Korea*
[2]*Department of Foreign Language Education, Seoul National University, Seoul, Korea*
[3]*Center for Educational Research, Seoul National University, Seoul, Korea*
[4]*Kakao Enterprise Corp., Seongnam, Korea*

Abstract

This paper proposes a method for designing a large recording script for open domain English speech synthesis. For read-aloud style text, 12 domains and 294 sub-domains were designed using text contained in five different news media publications. For conversational style text, 4 domains and 36 sub-domains were designed using movie subtitles. The final script consists of 43,013 sentences, 27,085 read-aloud style sentences, and 15,928 conversational style sentences, consisting of 549,683 tokens and 38,356 types. The completed script is analyzed using four criteria: word coverage (type coverage and token coverage), high-frequency vocabulary coverage, phonetic coverage (diphone coverage and triphone coverage), and readability. The type coverage of our script reaches 36.86% despite its low token coverage of 2.97%. The high-frequency vocabulary coverage of the script is 73.82%, and the diphone coverage and triphone coverage of the whole script is 86.70% and 38.92%, respectively. The average readability of whole sentences is 9.03. The results of analysis show that the proposed method is effective in producing a large recording script for English speech synthesis, demonstrating good coverage in terms of unique words, high-frequency vocabulary, phonetic units, and readability.

Keywords: recording script, speech synthesis, English, word coverage, phonetic coverage, readability

1. Introduction

It is generally known that the recording script, which is a major part of speech corpora (Chevelu & Lolive, 2015; Möbius, 2000), greatly affects the quality of synthesized sentences generated by speech synthesis systems. Previous studies mostly focus on efficiency in the script design based on phonetic coverage and propose to use greedy algorithms to select a minimal number of sentences with maximal phonetic coverage (Bozkurt et al., 2003; Kominek & Black, 2003, 2004; Matoušek et al., 2001; Torres et al., 2019; Van Santen & Buchsbaum, 1997). Apart from phonetic coverage, prosodic coverage is also considered from a linguistic perspective. Tonal

* This work was supported by the Kakao Enterprise Corporation.
** sunhkim@snu.ac.kr, Corresponding author
Received 31 July 2021; Revised 9 September 2021; Accepted 9 September 2021

information is employed based on syllables in a tonal language like Chinese languages (Tao et al., 2008; Zhu et al., 2002). And lexical stress information is considered in stress-timed languages like English (Dong et al., 2009). To deal with different prosodic realizations, Chevelu & Lolive (2015) and Bonafonte et al. (2005) recommend keeping at least ten realizations of each concatenation unit in the script. However, phrasing information cannot be included directly during the script designing stage, because phrasing information cannot be predicted from the text. Kawai et al. (2000) proposes to select sentences using a phrasing prediction module, the front-end part of the speech synthesis engine.

Word coverage and readability are also proposed to be considered in Dong et al. (2009). They used Token Coverage Rate (TCR) and Corpus Coverage Rate (CCR) to improve the unique word coverage in the given corpus for which readability is measured using the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL) (Klare, 1974-1975). In addition to these word coverage measurements, Kim et al. (2013) suggests to consider the entropy of each unique word. To improve readability, it is also proposed to select a group of sentences which together contain only 10,000 high-frequency words (Honnet et al., 2017) although such selection could, however, result in excluding low-frequency words characterizing certain domains.

The state-of-the-art high-quality speech synthesis systems are based on Deep Neural Network techniques approach (Arık et al., 2017; Purwins et al., 2019; Van den Oord et al., 2016; Wang et al., 2017) and unit selection and concatenation approach (King, 2014). For both approaches, a large amount of data is indispensable, but few studies address the issue of constructing a large speech corpus. Furthermore, recent research also reports on the use of found data for Text-to-Speech as in Gallegos et al. (2020), Kuo et al. (2019), Park & Mulc (2019), Prahallad & Black (2011), Watts et al. (2013), and Zen et al. (2019). These studies focus on judging the quality of the acoustic characteristics of the sound, but the methods and standards to apply in selecting sentences from the data has rarely been dealt with.

The goal of this paper is to propose a method of designing a large recording script for open domain English speech synthesis. The completed script will be analyzed by using four criteria, word coverage, high-frequency vocabulary coverage, phonetic coverage, and readability. Here, a "large" recording script means a script consisting of at least 500,000 words, which would correspond to about 50 hours of recording. The results of this study is also expected to serve as a guide for selecting sentences from found data.

This paper is organized as follows. Section 2 describes the script design and the process of selecting sentences. In Section 3, the statistics of the final script are presented, and in Section 4, the final script is analyzed in terms of four factors proposed in Section 2. Section 5 concludes the paper including discussion.

2. Methods

2.1. Script design

The text corpus is composed of both read-aloud style sentences and conversational style sentences, and the ratio of the number of words contained in each is set at 7 to 3 respectively.

2.1.1. Design of read-aloud style script

For collecting read-aloud style text, five American daily newspapers covering different regions of the United States are used. The subjects of news articles were classified into 12 domains by referring to the sections of each newspaper, such as politics, world, business, technology & science, sports, education, humanity, culture, lifestyle, accidents, climate & environment, and health. Major domains are classified into 294 sub-domains. This design of categories is crucial for collecting sentences containing various unique words specific to each domain, which will lead to high word coverage. Collecting at least 20 sentences per sub-domain is recommended to maintain the balance of different domains. And only up to 10 sentences are recommended to be collected in each article to ensure various vocabulary of corresponding topics within a sub-domain.

The length of sentences is also considered for different prosodic realizations, such as stress, rhythm and intonation. The sentences are thus divided into phrases, short sentences (5–14 words), medium sentences (15–24 words), and long sentences (25–34 words) with their composition ratio of 1:10:3:2.

2.1.2. Design of conversational style script

For collecting sentences of conversational style text, movie subtitles provided by Subscene (https://subscene.com/) were used. American movies were mainly selected and classified according to their subject matters into 4 major domains: professionals, specialty, fantasy, and daily life. Major domains were in turn further classified into 36 sub-domains. The domain or the sub-domain of each film was judged based on the content of the movie review along with personal experience and knowledge of the experts. At least 5 movie subtitles were collected for each sub-domain.

The length of sentence is not considered in conversational style because most of the conversational style sentences contain less than fifteen words. Instead, the traditional sentence types are considered to provide prosodic diversity. Declarative sentences, yes/no questions, wh-questions, and imperative sentences are selected at their compositional ratio of 10:2:2:1.

2.2. Selecting sentences

Three linguistic experts participated in selecting and analyzing sentences of both styles.

2.2.1. Selecting sentences of read-aloud style

Each news article selected is downloaded from its URL and it is decomposed into sentences using Natural Language Toolkit (NLTK) (Bird et al., 2009). Then, the sentences are sorted according to the number of words in ascending order. The basic data frame used for collecting sentences include 5 columns: domain, sub-domain, sentence, length of sentence, and URL. A collection status table is also provided at the right corner of each working page, so that the experts can check and comply with the collection ratio of sentences. In sum, the corpus is constructed with sentences selected in 294 sub-domains, considering unique word corresponding to the given sub-domain and the sentence length.

2.2.2. Selecting sentences of conversational style

A total of 237 films are selected and their subtitles are downloaded through Subscene (https://subscene.com/), and all the subtitles are integrated into one document. While the corpus of read-aloud style text is collected by selecting sentences one-by-one, the sentences of conversational style are constructed based on this integrated corpus. A certain number of conversational style sentences

appear in almost all sub-categories with high frequency, such as “I don’t know,” or “What are you doing?” We call these types of sentences ‘basic conversational style sentences,’ and they are extracted from the integrated corpus. However, sentences consisting of swear words or names were excluded despite their high frequency appearance.

The remaining sentences are selected from each of 36 sub-domains. For each sub-domain, sentences containing high-frequency words are selected. In addition, the sentences uttered before and after these selected sentences are reviewed and those comprising domain-specific words are also selected. Then, sentences containing high-frequency N-grams are also selected, retaining collocations, which is important in natural sounding. The tokenization of each sentence is performed using NLTK (Bird et al., 2009).

2.3. Analysis

The final script completed as described above is analyzed based on the following four criteria:

- Word coverage
- High-frequency vocabulary coverage
- Phonetic coverage
- Readability

For word coverage, two measurements are used, the type (or unique word) coverage (UC) and the token coverage (TC). The type coverage indicates the ratio of unique type occurrences of the script and the token coverage indicates the ratio of token occurrences of the script. These measurements are analogous to those proposed in Dong et al. (2009). Supposing X is a part or the whole of our script, and Y is the test corpus, the type coverage is calculated as

$$UC(X) = \frac{U(X)}{U(Y)} \times 100 \quad (1)$$

where U(x) is the number of unique words in the corpus x. Similarly, the token coverage is calculated as

$$TC(X) = \frac{T(X)}{T(Y)} \times 100 \quad (2)$$

where T(x) is the total number of tokens in the corpus x.

In order to calculate the high-frequency vocabulary coverage, we use the top 10,000 words selected in the BNC/COCA headword lists from the Victoria University of Wellington's Vocabulary lists (Nation, n.d.).

For the phonetic coverage, the diphone coverage and the triphone coverage are extracted after the script is converted into phone sequences using the CMU Pronouncing Dictionary, which is available online (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), and the G2P (Park & Kim, 2019).

To evaluate the readability of recording scripts, the FKGL (Klare, 1974–1975) is used. As shown in (3) below, FKGL is based on a formula that includes the average number of words per sentence (AWS) and the average number of syllables per word (ASW). As FKGL values vary from 0 to 18, FKGL of lower than 0 was merged into level 0 and that of greater than 18, level 18.

$$FKGL = (0.39 \times AWS) + (11.8 \times ASW) - 15.59 \quad (3)$$

3. Results

In order to avoid excessive collection of overlapping words, we periodically reviewed the status of new words at every phase when about 5,000 sentences are collected. Table 1 shows the increasing trend in the number of words for each phase. During the five phases, the number of tokens and types in read-aloud style sentences show a constant increase, with more than 70,000 tokens and 5,000 types for each phase. Similarly, in the case of conversational style sentences, the trend is relatively constant, increasing by 50,000 tokens and 3,000 types during the 3 phases.

Table 1. Word increasing rate for each phase

	Read-aloud style		Conversational style	
Phase	Token	Type	Token	Type
1	75,209	12,988	54,934	7,097
2	150,951	19,939	100,202	10,447
3	226,200	24,987	146,725	13,063
4	302,383	29,393		
5	402,958	34,743		

When all phases are completed, the final script obtained consists of 43,013 sentences, 27,085 read-aloud style sentences and 15,928 conversational style sentences, which amount to 549,683 tokens and 38,356 types. Table 2 and Table 3 provide detailed statistics for each style of sentences. The average number of tokens, which indicates the average length of each type of sentences, and the average number of types are provided for each style of script. The average numbers are calculated by dividing the total number by the number of sentences. The average number of tokens is 14.88 and the average number of types 1.28 in the read-aloud text, while the average number of tokens is 9.21 and the average number of types 0.82 in the conversational text.

Table 2. The numbers of sentences, tokens and types of read-aloud style sentences

	Phrase	Short	Medium	Long	Total
Sentence	1,678	16,870	5,163	3,374	**27,085**
Tokens	13,467	187,127	102,545	99,819	**402,958**
Types	5,114	23,646	16,620	15,944	**34,743**
Avg. tokens /sentence	8.03	11.09	19.86	29.58	14.88
Avg. types /sentence	3.05	1.40	3.22	4.72	1.28

Avg, average; Short, 5–14 words; Medium, 15–24 words; Long, 25–34 words.

Table 3. The numbers of sentences, tokens and types of conversational style sentences

	Dec.	Int. (Y/N)	Int. (WH)	Imp.	Total
Sentence	10,521	2,096	2,035	1,276	**15,928**
Tokens	106,296	16,911	15,405	8,113	**146,725**
Types	11,584	3,095	2,507	1,940	**13,063**
Avg. tokens /sentence	10.10	8.07	7.57	6.36	9.21
Avg. types /sentence	1.10	1.48	1.23	1.52	0.82

Dec, declarative; Int, interrogative; Imp, imperative.

Figure 1 shows the percentage of unique words (type) inclusion for each domain of read-aloud style sentences. The average ratio of unique words in each domain is 22.71%. This indicates that various words are selected almost evenly from domain to domain as planned.

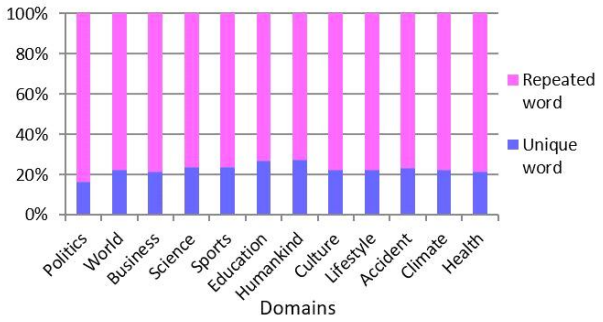

Figure 1. Ratio of unique words in each domain of read-aloud style sentences

Table 4 shows the ratio of overlap between read-aloud style sentences and conversational style sentences in terms of unique words and N-grams. As shown in Table 4, 65.9% of all types are included only in read-aloud style sentences, and 9.4% of all types belong to conversational style sentences, while 24.6% overlap between the two. On the other hand, the overlap ratio of N-grams between two styles is 4.3% (2-grams) and 1.9% (3-grams), which indicates that collocations differ significantly depending on the style of the sentence.

Table 4. Overlap of unique words and N-grams between each style

	Read-aloud style only (%)	Overlap (%)	Conversational style only (%)
Types	25,293 (65.9)	**9,450 (24.6)**	3,613 (9.4)
2-grams	6,292 (76.9)	**355 (4.3)**	1,536 (18.8)
3-grams	497 (71.0)	**13 (1.9)**	190 (27.1)

4. Analysis

4.1. Word coverage

Because the read-aloud style sentences are not selected using a specific corpus, part of Kaggle's News Articles dataset (2018), which consists of 18,506,913 tokens and 73,931 types, is used as the test corpus. For the conversational style sentences, the original corpus consisting of 237 movies is used as the test corpus. Table 5 presents token coverage and type coverage of the whole script and those of each style. The type coverage of the read-aloud style script is 33.86%, that of the conversational style script 29.39%, and that of the whole script 36.86%. These numbers should be significant, given that the number of tokens of our script is 2.97% of that of the test corpus extracted from News Articles dataset, and that of the conversational style script is 5.06% of the original corpus.

Table 5. Word coverage of the script

	Token coverage (%)	Type coverage (%)	Test corpus
Whole script	2.97	36.86	News Articles
Read-aloud style	2.18	33.86	News Articles
Conversational style	5.06	29.39	Movie corpus

4.2. High-frequency vocabulary coverage

Two corpora were created to be compared to our script, using only CNN data from News Articles dataset in Kaggle (2018). CNN_549K has the similar volume of tokens, 549,849 words, to that of the whole script. CNN_402K contains 403,183 tokens similar to read-aloud style sentences.

Table 6 shows the high-frequency vocabulary coverage of the script and the test corpora. Comparing the whole script and CNN_549K, we see that the whole script contains more various high-frequency vocabulary, 73.82%, than 61.53% in CNN_549K despite the similar total number of tokens.

Table 6. High-frequency vocabulary coverage of corpora

	Token	Type	Vocabulary coverage (%)
Whole script	549,683	38,356	**73.82**
Read-aloud style	402,958	34,743	**68.37**
Conversational style	146,725	13,063	**44.03**
CNN_549K	549,849	28,920	**61.53**
CNN_402K	403,183	25,031	**56.73**

4.3. Phonetic coverage

For phonetic coverage analysis, a total of 42 units are used, composed of the CMU Pronouncing Dictionary's 39 phonemes, plus two silence symbols (before sentence, SIL1; after sentence, SIL2) and one pause symbol (PAU). In the process of combining diphone set and triphone set, meaningless combinations such as (SIL1+PAU), (SIL1+SIL2), (PAU+SIL2) and (PAU+PAU) are eliminated.

Table 7 presents the phonetic coverage for each style of sentences. For each style, diphone coverage and triphone coverage are provided with and without word stress. If the stress is not included, the diphone coverage and the triphone coverage of the whole script are 86.70% and 38.92%, respectively.

4.4. Readability

Figure 2 shows the frequency of read-aloud style and convert-sational style sentences per FKGL. The average FKGL of read-aloud style sentences is 11.18 and that of conversational style sentences 5.36. The FKGL distribution of read-aloud style sentences

Table 7. Phonetic coverage of the script

	Read-aloud style				Conversational style				Whole script
	Included	Not included	Total	Coverage (%)	Included	Not included	Total	Coverage (%)	Coverage (%)
Diphone	3,112	1,925	5,037	**61.78**	2,620	2,417	5,037	**52.02**	**62.68**
Diphone (w/o stress)	1,448	229	1,677	**86.34**	1,353	324	1,677	**80.68**	**86.70**
Triphone	47,314	305,276	352,590	**13.42**	28,510	324,080	352,590	**8.09**	**13.96**
Triphone (w/o stress)	25,316	41,764	67,080	**37.74**	18,975	48,105	67,080	**28.29**	**38.92**

leans towards higher levels, while that of conversational style sentences leans towards lower levels. In total, the average FKGL is 9.03. The frequency distribution of read-aloud style and conversational style sentences shows that the selected script has fairly good coverage of readability levels.

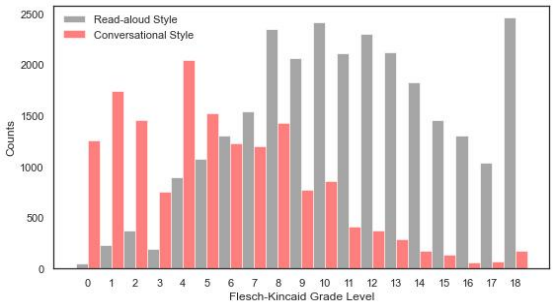

Figure 2. Sentence frequency per Flesch-Kincaid grade level

5. Discussion and Conclusion

This paper proposes a method of designing a large recording script for open domain English speech synthesis. The final script consists of 43,013 sentences, which are composed of 549,683 tokens and 38,356 types. The read-aloud style sentences are collected manually based on a classification of domains and sub-domains, while the conversational style sentences are selected from the integrated corpus. For the conversational style sentences, high-frequency sentences, sentences with high-frequency words and high-frequency N-grams, are selected using simple codes, while domain-specific expressions are collected manually. The resulting high type coverage should be due to this meticulous manual classification of domains and sub-domains as well as manual selection work.

The completed script is analyzed using four criteria, word coverage, high-frequency vocabulary coverage, phonetic coverage, and readability. Comparing the completed script to the test corpus, the type coverage of our script is 36.86%, while its token coverage is only 2.97%, which appears to be quite significant. As for the high- frequency vocabulary coverage, our script shows 73.82% compared to that of 61.53% in the test corpus even though the two contain a similar number of tokens.

For each style, the phonetic coverage is analyzed with and without word stress. The diphone coverage and the triphone coverage of the whole script are 86.70% and 38.92%, respectively. In comparison to earlier studies on text design, which mainly focus on selecting minimum sentences with maximum phonetic coverage, this study primarily focuses on a manual design of selecting sentences based on word coverage, which produce results showing higher phonetic coverage than earlier studies.

For future work, it would be possible to transform the manual part of the selection process into an automatized process based on the collected word list and the sentence length of each domain. Also, the four criteria used for analysis in this study can be utilized as the corpus selection criteria for a large recording script.

References

Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., ⋯ Shoeybi, M. (2017, August). Deep voice: Real-time neural text-to-speech. *Proceedings of the 34th International Conference on Machine Learning*, *PMLR 70* (pp. 195-204). Sydney, Australia.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Beijing, China: O'Reilly Media.

Bonafonte, A., Höge, H., Tropf, H. S., Moreno, A., van der Heuvel, H., Sündermann, D., ⋯ Kiss, I. (2005). TTS baselines and specifications (Report No. FP6-506738). Retrieved from https://docsbay.net/tc-star-projectdeliverable-no-d8title-tts-baselines-specifications

Bozkurt, B., Ozturk, O., & Dutoit, T. (2003, September). Text design for TTS speech corpus building using a modified greedy selection. *Proceedings of the Eurospeech 2003* (pp. 277-280). Geneva, Switzerland.

Chevelu, J., & Lolive, D. (2015, September). Do not build your TTS training corpus randomly. *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)* (pp. 350-354). Nice, France.

Dong, M., Cen, L., Chan, P., & Li, H. (2009). Readability consideration in speech synthesis recording script selection. *International Journal on Asian Language Processing, 19*(2), 45-54.

Gallegos, P. O., Williams, J., Rownicka, J., & King, S. (2020, October). An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets. *Proceedings of the Interspeech 2020* (pp. 1758-1762). Shanghai, China.

Honnet, P. E., Lazaridis, A., Garner, P. N., & Yamagishi, J. (2017). The SIWIS French speech synthesis database – Design and recording of a high quality French database for speech synthesis. Retrieved from https://infoscience.epfl.ch/record/225946

Kawai, H., Yamamoto, S., Higuchi, N., & Shimizu, T. (2000, October). A design method of speech corpus for text-to-speech synthesis taking account of prosody. *Proceedings of the 6th International Conference on Spoken Language Processing* (pp. 420-425). Beijing, China.

Kim, S., Kim, J., Kim, S., & Kim, H. (2013, November). Recording script design for speech corpus of English news reading TTS. *Proceedings of the 2013 Autumn Conference of Acoustical Society of Korea* (pp. 49-52). Jeju, Korea.

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens, 1*(1), e006.

Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly, 10*(1), 62-102.

Kominek, J., & Black, A. W. (2003). CMU Arctic database for speech synthesis (Report No. CMU-LTI-03-177). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6699C4E348169581A2EED5E3041C1C81?doi=10.1.1.64.8827&rep=rep1&type=pdf

Kominek, J., & Black, A. W. (2004, June). The CMU Arctic speech databases. *Proceedings of the 5th ISCA ITRW Speech Synthesis* (pp. 223-224). Pittsburgh, PA.

Kuo, F. Y., Ouyang, I. C., Aryal, S., & Lanchantin, P. (2019, September). Selection and training schemes for improving TTS voice built on found data. *Proceedings of the Interspeech 2019* (pp. 1516-1520). Graz, Austria.

Matoušek, J., Psutka, J., & Krůta, J. (2001, September). Design of speech corpus for text-to-speech synthesis. *Proceedings of the Eurospeech 2001* (pp. 2047-2050). Aalborg, Denmark.

Möbius, B. (2000). Corpus-based speech synthesis: Methods and challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, 6*(4), 87-116.

Nation, P. (n.d.). Vocabulary lists. Retrieved from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

News Articles [dataset] (2018, May). Retrieved from https://www.kaggle.com/harishcscode/all-news-articles-from-home-page-media-house/version/1

Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., … Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. Retrieved from https://arxiv.org/abs/1609.03499.

Park, K., & Mulc, T. (2019, September). CSS10: A collection of single speaker speech datasets for 10 languages. *Proceedings of the Interspeech 2019* (pp. 1566-1570). Graz, Austria.

Park, K., & Kim, J. (2019). g2pE: A simple Python module for English grapheme to phoneme conversion. Retrieved from https://github.com/Kyubyong/g2p

Prahallad, K., & Black, A. W. (2011, July). Segmentation of monologues in audio books for building synthetic voices. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1444-1449.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing, 13*(2), 206-219.

Santen, J. V., & Buchsbaum, A. (1997, September). Methods for optimal text selection. *Proceedings of the 5th European Conference on Speech Communication and Technology* (pp. 553-556). Rhodes, Greece.

Tao, J., Liu, F., Zhang, M., & Jia, H. (2008, October). Design of speech corpus for mandarin text to speech. *Proceedings of the Blizzard Challenge 2008 Workshop* (pp. 1-4). Brisbane, Australia.

Torres, H. M., Gurlekian, J. A., Evin, D. A., & Mercado, C. G. C. (2019). Emilia: a speech corpus for Argentine Spanish text to speech synthesis. *Language Resources and Evaluation, 53*(3), 419-447.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., … Saurous, R. A. (2017, August). Tacotron: Towards end-to-end speech synthesis. *Proceedings of the Interspeech 2017* (pp. 4006-4010). Stockholm, Sweden.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., & King, S. (2013, September). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. *Proceedings of the 8th ISCA Speech Synthesis Workshop* (pp. 101-106). Barcelona, Spain.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019, September). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *Proceedings of the Interspeech 2019* (pp. 1526-1530). Graz, Austria.

Zhu, W., Zhang, W., Shi, Q., Chen, F., Li, H., Ma, X., & Shen, L. (2002, September). Corpus building for data-driven TTS systems. *Proceedings of the 2002 IEEE Workshop on Speech Synthesis* (pp. 199-202). Santa Monica, CA.

• **Sunhee Kim,** Corresponding author
Professor, Dept. of French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7693
Email: sunhkim@snu.ac.kr
Fields of interest: French, Phonetics, Speech synthesis

• **Hojeong Kim**
M.A. candidate, French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7690
Email: hojeong43@snu.ac.kr
Fields of interest: L2 acquisition

• **Yooseop Lee**
Undergraduate, Dept. of French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7690
Email: lyooseop@snu.ac.kr
Fields of interest: L2 acquisition

• **Boryoung Kim**
Undergraduate, Dept. of French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7690
Email: jadebr@snu.ac.kr
Fields of interest: L2 acquisition

• **Yongkook Won**
Visiting researcher, Center for Educational Research
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-5721
Email: purgatorio@snu.ac.kr
Fields of interest: Language assessment, Natural language processing

• **Bongwan Kim**
Text to Speech Part Leader, Kakao Enterprise Corp.
235, Pangyoyeok-ro, Bundang-gu, Seongnam 13494, Korea
Email: montae.k@kakaoenterprise.com
Fields of interest: Text to Speech