Learning pronunciation from a foreign language in speech synthesis networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although there are more than 65,000 languages in the world, the pronunciations of many phonemes sound similar across the languages. When people learn a foreign language, their pronunciation often reflect their native language's characteristics. That motivates us to investigate how the speech synthesis network learns the pronunciation when multi-lingual dataset is given. In this study, we train the speech synthesis network bilingually in English and Korean, and analyze how the network learns the relations of phoneme pronunciation between the languages. Our experimental result shows that the learned phoneme embedding vectors are located closer if their pronunciations are similar across the languages. Based on the result, we also show that it is possible to train networks that synthesize English speaker's Korean speech and vice versa. In another experiment, we train the network with limited amount of English dataset and large Korean dataset, and analyze the required amount of dataset to train a resource-poor language with the help of resource-rich languages.

1 Introduction

Among many languages in the world, some of the languages have phonemes with the same pronunciation. Conceptually, if the intersection of the pronunciations of two languages is large enough, the burden of learning the pronunciation of the one language after acquiring the other language will be lowered. We can also see people tends to show accent of their first language when speaking in their second language. This may tell us that people learn new pronunciation based on their first language. Motivated by this, we decided to investigate how the multilingual dataset can help training and generalizing of neural speech synthesis network.

Recent advances in Text-To-Speech (TTS) using deep neural network allow us to build end-to-end TTS models that can generate natural-sounding speech [12]. The end-to-end approach requires us only a little amount of prior knowledge about language, and TTS models can be built easily if we have enough text-speech pair data. Some researchers were interested in transferring voice across languages [6], yet they only focused on transferring voice color, not pronunciations of languages.

In this study, we are interested in analyzing end-to-end TTS models in cross-lingual setting. We train bilingual-TTS model using monolingual speakers' speech database. We use both English and Korean speakers' speeches to synthesize Korean speaker's English speech and visa versa. According to [3], Tacotron requires a large amount of speech data (>10 hours) from one speaker to obtain good generation quality. However, it is not easy to record hours of speech as well as to find a person who can speak multiple languages. Fortunately, given enough amount of multiple speakers' speech data, we can build a TTS system despite the amount of each speaker is relatively small [4]. Inspired by this result, we trained a multi-speaker cross-lingual TTS model using text-speech pairs of English and Korean. Interestingly, although we did not have any speaker who speaks both English and Korean,

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

every speaker in the trained model could speak both English and Korean fluently. We investigated whether the phoneme embeddings from different languages are learned meaningful representations. We found phonemes with similar pronunciation tend to stay closer than the others even across the different languages. From these results, we thought that the cross-lingual model would be possible to generalize for a language with scarce amount of data when there is another language with abundant data. We trained cross-lingual TTS models while differing the amount of data for a resource-scarce language. Then we computed and compared character error rate (CER) of generated speeches from each model by crowd-sourced human dictation.

To summarize, the contributions of this study are as follows:

1. We successfully trained a cross-lingual multi-speaker TTS model using English and Korean data in which no bilingual speaker is included.
2. We found two learned phoneme embeddings from different languages are to located close if pronunciation of the two phonemes are similar.
3. We showed how much data of a language is required to train a TTS model when we have a large amount of data from another language.

2 Model

We use simplified version [11] of Tacotron [12] for the TTS model, but we use the original Tacotron style of Post-processing net and Griffin-Lim algorithm [5] for conversion of linear-scale spectrogram to waveform. A sequence of phonemes are converted to phoneme embeddings, then fed to the encoder as input. Note that, the phoneme embeddings are normalized to have same norm. Since we need multi-speaker TTS model, we adopt Deep Voice 2 [4] style speaker embedding network. One-hot speaker identity vector is converted to a 32-dimensional speaker embedding vector by the speaker embedding network. Unlike Deep Voice 2, we did not used speaker embedding vector in the encoder module. In Deep Voice 2, the speaker embedding vector is worked as a bias in the every layer of the encoder module, this makes phonemes to be represented differently for each speaker. Since we want TTS model to utilize common information of pronunciation between speakers and language, we choose to avoid the bias introduced by the speaker embedding vector in the encoder module. Unless stated otherwise, we used same hyperparameter settings with [11].

3 Experiments and results

3.1 Dataset

In the subsection 3.2, we used two public datasets of English and Korean. For English dataset, we used the entire VCTK dataset, which included 109 speakers. The amount of VCTK dataset was 28.3 hours after removing silence. For Korean dataset, we used randomly selected subset of Zeroth-Korean dataset [7], which includes 50 speakers. The amount of selected Zeroth-Korean dataset was 18.2 hours after removing silence. In the subsection 3.3, we used 2013 Blizzard Challenge dataset which amounts 10 hours after removing silence. For Korean dataset, we used the entire Zeroth-Korean dataset, which includes 115 speakers and contains 46.2 hours after removing silence.

We used grapheme-to-phoneme (G2P) libraries to convert text to the corresponding phoneme sequence. For English text, we used a G2P library [8] which extends CMUdict[10] phoneme dictionary by predicting phonemes of out-of-vocabulary words with neural network. For Korean text, we used KoG2P [2] which is commonly used for Korean text preprocessing.

3.2 Training a multi-speaker bilingual TTS

We investigated how the phoneme representation is learned when the network is trained with the dataset containing both English and Korean. The model was trained to minimize L1 losses between ground-truth spectrogram and predicted spectrogram for both linear-scale spectrogram and mel-scale spectrogram.

It is known that the pronunciation of different languages can be described by one unified alphabetic system, International Phonetic Alphabet (IPA) [1][1]. The conversion tables of IPA-CMUdict and IPA-KoG2P can be found in [9] and Appendix A respectively. Although one-to-one correspondence does not hold between English phoneme set and Korean phoneme set in terms of IPA, we could carefully choose subset of each set to include only the phonemes that have the common pronunciation with the opposite language. The chosen subsets can be found in Appendix B.

For each phoneme in the chosen subsets, say anchor phoneme, we computed cosine distance between each of the other phonemes. Then, we listed the 5-nearest phoneme embedding of the opposite language to analyze the learned phoneme representation. The results are shown in Table 1 and Appendix C. The numbers 0, 1, and 2 after the English phonemes denote "No stress", "Primary stress", and "Secondary stress" respectively. While CMUdict distinguished stressed pronunciation, we could not find corresponding IPA representation, so we reported all stress types. The results show that most of the anchor phonemes' corresponding phonemes in the opposite language were found in the 5-nearest phoneme embeddings. As shown in Table 2, the 5-nearest phonemes among 70 phonemes (57 phonemes in Korean) include the corresponding phonemes and similar pronunciations, which implies that the phoneme embeddings learned the relation of pronunciations across the languages.

Table 1: The 5-nearest English phonemes of each Korean phoneme whose pronunciation exists in English, IPA symbols are written in the parentheses

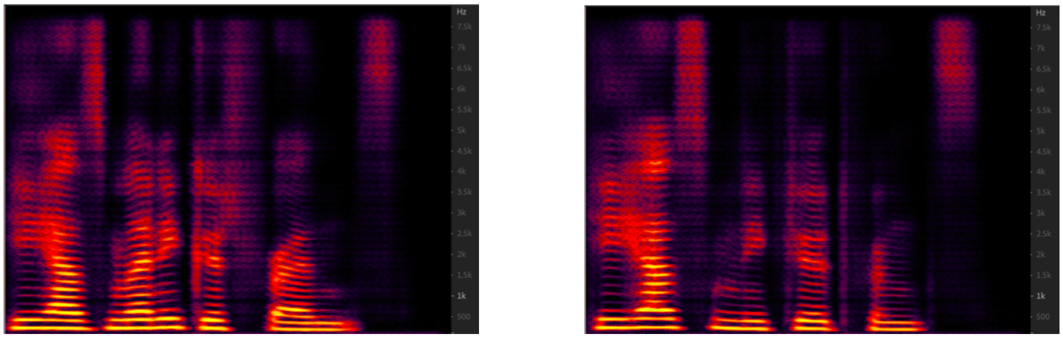
Korean phoneme	1st	2nd	3rd	4th	5th
ii (i)	IY2 (i)	Y (j)	IY1 (i)	IY0 (i)	EY0 (eɪ)
k0 (g)	G (g)	B (b)	DH (ð)	W (w)	UH2 (ʊ)
kf (k)	NG (ŋ)	K (k)	UH2 (ʊ)	T (t)	P (p)
ll (l)	ER2 (ɝ)	ER0 (ɝ)	R (ɹ)	UH0 (ʊ)	OY0 (ɔɪ)
p0 (b)	B (b)	P (p)	M (m)	DH (ð)	AA0 (ɑ)
pf (p)	W (w)	T (t)	B (b)	M (m)	G (g)
qq (ɛ)	EY0 (eɪ)	EY1 (eɪ)	EY2 (eɪ)	EH1 (ɛ)	AY0 (aɪ)
rr (ɾ)	D (d)	UH2 (ʊ)	ER0 (ɝ)	AH0 (ʌ)	ER2 (ɝ)
t0 (d)	D (d)	TH (θ)	EH2 (ɛ)	V (v)	G (g)
tf (t)	T (t)	HH (h)	AO0 (ɔ)	N (n)	SH (ʃ)
uu (u)	OY2 (ɔɪ)	UW2 (u)	OY1 (ɔɪ)	UH2 (ʊ)	AO0 (ɔ)
vv (ʌ)	AO0 (ɔ)	AA1 (ɑ)	AA2 (ɑ)	AO1 (ɔ)	AA0 (ɑ)

The 4-nearest phonemes of the whole phonemes, see Appendix E and Appendix D. Although not all of them have the corresponding pronunciation in the other language, pronunciation of the nearest phonemes sounded similar pronunciation. To describe their similarity, we first generated speech of a sentence "He has many good friends." in English phoneme sequence[2] (Figure 1-(a)). Then, we substituted each phoneme with the nearest phoneme in Korean[3] (Figure 1-(b)). When we compared the two speeches, they were sounded similarly. Figure 1 shows an example spectrogram pair of the generated speeches. More examples are posted in our demo page.[4]

3.3 Data requirements for foreign language

We observed that the learned phoneme embedding space can represent the relation between pronunciations. With this learning behavior, we wanted to check how much data is required to train one language when we have an abundant amount of data for another language. In this experiment, we fixed amount of a large Korean dataset while varying amount of English data. We could compare the relationship between the amount of resource-poor data and generation quality. To quantify the generation quality, we obtained transcribed sentences of generated speeches in a held-out test set

[1] The pronunciation of each symbol can be heard from: https://en.wikipedia.org/wiki/Help:IPA
[2] HH, IY1, , HH, AE1, Z, , M, EH1, N, IY0, , G, UH1, D, , F, R, EH1, N, D, Z
[3] h0, wi, , h0, ya, s0, , mf, ye, nf, ii, , kk, yo, tt, , ph, ks, ye, nf, tt, s0
[4] https://x02kdkhjs8.github.io/index.html



(a) Speech of the original English phonemes

(b) Speech of the nearest Korean phonemes

Figure 1: Spectrogram of generated speech using English phonemes and the nearest Korean phonemes

using crowd-sourcing platform, and the average CER of the transcriptions is reported in Table 2. Also, the generated samples from each model are posted in the demo page.

Table 2: Change of character error rate by amount of English dataset

	Hours of English data used			
CER	10	5	3	1
English test set read by Korean speaker	22.5%	26.9%	35.6%	39.9%
Korean test set read by English speaker	12.5%	12.9%	7.9%	11.1%

Since we are interested only in the generalization capability of the cross-lingual TTS model, we reported only the CER of each test set read by the opposite language speakers. As the amount of the English data increases, the CER of the English test set decreased, but the CER of Korean test set did not change much with at the low level. We think it is because the amount of Korean data was large enough so that the quality for Korean was good even though there was only a small amount of English data. From this result, we concluded that the absolute amount of data for each language is the key factor of the generation quality.

It is worth noting that the same TTS model could not be trained at all (cannot learn attention alignment) if we used only 1 hour of the Blizzard dataset. Though the reported CER in Table 2 is not much appealing, the cross-lingual model could utilize information in Korean language for learning English, and the model showed better generation quality. From this fact, we could see that the model could generalize the pronunciation of Korean to English. If we can obtain similar results for other language pairs, we may train a TTS model of a resource-poor language with the help of a resource-rich language, such as English. Given that TTS models require a large amount of data and that acquiring annotated speech data is expensive, it was difficult to train a TTS model for resource-poor languages. We believe our finding can help to produce TTS models in various languages.

4 Conclusion

In this work, we have trained a cross-lingual TTS model using two monolingual datasets. By investigating the distance between learned phoneme embeddings, we have experimentally shown that the embeddings represent the relation of pronunciations across the different languages. We also checked the quality of the cross-lingual TTS model that is trained with a resource-poor language and a resource-rich language.

We are interested in using cross-lingual TTS model for data augmentation of automatic speech recognition (ASR) model. Like TTS, ASR also requires large amount of training data. For resource-poor languages, we may train a TTS model as proposed in this work. Since TTS model can generate speech of arbitrary sentence in various speakers' voice, it will improve robustness of the ASR model. We hope to investigate how the cross-lingual TTS can help training of the ASR model in the future.

References

[1] I. P. Association, C. PRESS, and D. Decker. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. A Regents publication. Cambridge University Press, 1999. ISBN 9780521637510.

[2] Y. Cho. Kog2p. `https://github.com/scarletcho/KoG2P`, 2017.

[3] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. *ArXiv e-prints*, 2018.

[4] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in Neural Information Processing Systems 30*, pages 2962–2970, 2017.

[5] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83, Boston, Massachusetts, USA, April 14-16, 1983*, pages 804–807, 1983.

[6] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*, 2018.

[7] L. Jo and W. Lee. Zeroth-korean corpus. `http://www.openslr.org/40/`, 2018.

[8] K. Park and J. Kim. g2p-en. `https://github.com/Kyubyong/g2p`, 2018.

[9] L. Rice. Hardware and software for speech synthesis. *Dr. Dobbs J. of Computer Calisthenics & Orthodontia*, 1(4), 1976.

[10] A. I. Rudnicky. The cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, 2015.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783, 2018.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*, pages 4006–4010, 2017.

Appendices

A Conversion table of IPA and KoG2P phoneme set

IPA	KoG2P phoneme	IPA	KoG2P phoneme
b	p0	ts̤	cc
d	t0	ts^{h}	ch
dz	c0	tɕ	c0
dʑ	c0	tɕ̤	cc
g	k0	tɕ^{h}	ch
h	h0	w	wa, wq, wv, we, wi
ɦ	h0	z	s0
j	ya, yq, yv, ye, yo, yu	ʑ	s0
k	kf	a	aa
k̤	kk	aː	aa
k^{h}	kh	e	ee
l	ll	eː	ee
m	mf, mm	ɛ	qq
n	nn, nf	ɛː	qq
ŋ	ng	i	ii
p	pf	iː	ii
p̤	pp	o	oo
p^{h}	ph	oː	oo
ɾ	rr	ø	wo
s	s0	øː	wo
s̤	ss	u	uu
ɕ	s0	uː	uu
ɕ̤	ss	ʌ	vv
t	tf	ɔː	vv
t̤	tt	ɯ	xx
t^{h}	th	ɯː	xx
ts	c0	ɯi	xi

We made this table by referring [1] and [2]. Note that, some of the KoG2P phonemes (ks, nc, nh, lk, lm, lb, ls, lt, lp, lh, ps) are not included in the table, because we could not find exact match for those phonemes.

B Common phonemes between English and Korean by IPA

IPA	ENG	KOR
b	B	p0
d	D	t0
ɛ	EH	qq
g	G	k0
h	HH, H	h0
i	IY	ii
j	Y	ya, yq, yv, ye, yo, yu
k	K	kf
l	L	ll
m	M	mf, mm
n	N	nn, nf
ŋ	NX, NG	ng
p	P	pf
ɾ	DX	rr
s	S	s0
t	T	tf
u	UW	uu
ʌ	AH	vv
w	W	wa, wq, wv, we, wi
z	Z	s0

The IPA symbols in this table appear in both English and Korean. The phoneme symbols in the second and third columns are corresponding phonemes in English and Korean.

C The 5-nearest Korean phonemes of each English phoneme whose pronunciation exists in Korean

English phoneme	1st	2nd	3rd	4th	5th
AH0 (ʌ)	lh (-)	rr (ɾ)	xx (ɯ)	h0 (h)	ll (l)
AH1 (ʌ)	ya (j)	vv (ʌ)	oo (o)	yv (j)	lm (-)
AH2 (ʌ)	wv (w)	wa (w)	yv (j)	ya (j)	vv (ʌ)
B (b)	pp (p̤)	mm (m)	tt (t̤)	ph (p^h)	p0 (b)
D (d)	tt (t̤)	t0 (d)	c0 (dz)	nc (-)	lh (-)
EH0 (ɛ)	ya (j)	ye (j)	qq (ɛ)	wo (ø)	lm (-)
EH1 (ɛ)	ye (j)	lh (-)	yq (j)	nc (-)	ya (j)
EH2 (ɛ)	h0 (h)	t0 (d)	ya (j)	wo (ø)	lk (-)
G (g)	kk (k̤)	k0 (g)	ks (-)	tt (t̤)	t0 (d)
IY0 (i)	ii (i)	wi (w)	lt (-)	xi (ɯi)	yu (-)
IY1 (i)	wi (w)	ls (-)	yq (j)	ii (i)	xi (ɯi)
IY2 (i)	ii (i)	wi (w)	lt (-)	yq (j)	xi (ɯi)
K (k)	kh (k^h)	kk (k̤)	th (t^h)	kf (k)	ps (-)
L (l)	oo (o)	nn (n)	nh (-)	nf (n)	h0 (h)
P (p)	pp (p̤)	ph (p^h)	kk (k̤)	th (t^h)	kh (k^h)
T (t)	ch (t̤s^h)	c0 (dz)	th (t^h)	lp (-)	cc (ts̤)
UW (u)	aa (a)	xx (ɯ)	lb (-)	xi (ɯi)	yo (-)
UW0 (u)	yu (-)	yo (-)	uu (u)	wq (w)	mm (m)
UW1 (u)	yq (j)	yu (-)	yo (-)	ye (j)	lk (-)
UW2 (u)	uu (u)	wi (w)	ps (-)	yu (-)	yo (-)

The symbols in the parentheses are IPA symbols. The symbol '-' in the parentheses denotes that there was no IPA symbol for that phoneme.

D The 4-nearest English phonemes for each Korean phoneme

KOR-phoneme	1st	2nd	3rd	4th	KOR-phoneme	1st	2nd	3rd	4th
aa (a)	AE2 (æ)	AY0 (aɪ)	AW0 (aʊ)	AW1 (aʊ)	pf (p)	W (w)	T (t)	B (b)	M (m)
c0 (dz)	T (t)	D (d)	SH (ʃ)	ZH (ʒ)	ph (p^{h})	P (p)	F (f)	TH (θ)	B (b)
cc (ts̤)	JH (ʤ)	CH (tʃ)	T (t)	Z (z)	pp (p̤)	P (p)	B (b)	M (m)	D (d)
ch (ts^{h})	CH (tʃ)	T (t)	S (s)	JH (ʤ)	ps (-)	UW2 (u)	ER2 (ɝ)	OY2 (ɔɪ)	AW1 (aʊ)
ee (e)	EY0 (eɪ)	UH0 (ʊ)	AW2 (aʊ)	EH1 (ɛ)	qq (ɛ)	EY0 (eɪ)	EY1 (eɪ)	EY2 (eɪ)	EH1 (ɛ)
h0 (h)	HH (h)	EH2 (ɛ)	AE0 (æ)	AA2 (ɑ)	rr (ɾ)	D (d)	UH2 (ʊ)	ER0 (ɝ)	AH0 (ʌ)
ii (i)	IY2 (i)	Y (j)	IY1 (i)	IY0 (i)	s0 (s)	Z (z)	S (s)	SH (ʃ)	TH (θ)
k0 (g)	G (g)	B (b)	DH (ð)	W (w)	ss (s̤)	SH (ʃ)	S (s)	TH (θ)	F (f)
kf (k)	NG (ŋ)	K (k)	UH2 (ʊ)	T (t)	t0 (d)	D (d)	TH (θ)	EH2 (ɛ)	V (v)
kh (k^{h})	K (k)	CH (tʃ)	P (p)	HH (h)	tf (t)	T (t)	HH (h)	AO0 (ɔ)	N (n)
kk (k̤)	K (k)	G (g)	P (p)	AO0 (ɔ)	th (t^{h})	TH (θ)	T (t)	K (k)	P (p)
ks (-)	AO0 (ɔ)	NG (ŋ)	R (ɹ)	G (g)	tt (t̤)	TH (θ)	D (d)	P (p)	B (b)
lb (-)	V (v)	UW (u)	Y (j)	UW2 (u)	uu (u)	OY2 (ɔɪ)	UW2 (u)	OY1 (ɔɪ)	UH2 (ʊ)
lh (-)	EH1 (ɛ)	AA1 (ɑ)	AH0 (ʌ)	D (d)	vv (ʌ)	AO0 (ɔ)	AA1 (ɑ)	AA2 (ɑ)	AO1 (ɔ)
lk (-)	SH (ʃ)	HH (h)	OW2 (oʊ)	CH (tʃ)	wa (w)	OY1 (ɔɪ)	AO0 (ɔ)	OY2 (ɔɪ)	AH2 (ʌ)
ll (l)	ER2 (ɝ)	ER0 (ɝ)	R (ɹ)	UH0 (ʊ)	we (w)	EY1 (eɪ)	OY1 (ɔɪ)	EY2 (eɪ)	IH1 (ɪ)
lm (-)	EY1 (eɪ)	AA0 (ɑ)	AA2 (ɑ)	EH0 (ɛ)	wi (w)	IY1 (i)	UW2 (u)	IY2 (i)	UH2 (ʊ)
lp (-)	T (t)	UH2 (ʊ)	AO0 (ɔ)	K (k)	wo (ø)	EY1 (eɪ)	ER1 (ɝ)	ER2 (ɝ)	OY2 (ɔɪ)
ls (-)	IY1 (i)	EY1 (eɪ)	ER2 (ɝ)	IY2 (i)	wq (w)	EY1 (eɪ)	IH1 (ɪ)	EY0 (eɪ)	OY1 (ɔɪ)
lt (-)	EY0 (eɪ)	IY2 (i)	AY2 (aɪ)	IY0 (i)	wv (w)	OY1 (ɔɪ)	AA1 (ɑ)	AH2 (ʌ)	OY2 (ɔɪ)
mf (m)	M (m)	N (n)	NG (ŋ)	P (p)	xi (ɯi)	EY0 (eɪ)	EY2 (eɪ)	IY1 (i)	IY2 (i)
mm (m)	M (m)	B (b)	AO0 (ɔ)	ER0 (ɝ)	xx (ɯ)	UW (u)	AY2 (aɪ)	UH2 (ʊ)	OY2 (ɔɪ)
nc (-)	EH1 (ɛ)	AA2 (ɑ)	D (d)	DH (ð)	ya (j)	AE2 (æ)	AE0 (æ)	AW1 (aʊ)	AW0 (aʊ)
nf (n)	N (n)	NG (ŋ)	ER2 (ɝ)	M (m)	ye (j)	EY1 (eɪ)	EY0 (eɪ)	EH1 (ɛ)	ER1 (ɝ)
ng (ŋ)	NG (ŋ)	M (m)	AA0 (ɑ)	N (n)	yo (-)	OW1 (oʊ)	OW0 (oʊ)	UW2 (u)	OW2 (oʊ)
nh (-)	T (t)	AO2 (ɔ)	L (l)	B (b)	yq (j)	EY1 (eɪ)	IY1 (i)	EH1 (ɛ)	IH1 (ɪ)
nn (n)	N (n)	AO0 (ɔ)	L (l)	DH (ð)	yu (-)	UW0 (u)	ER2 (ɝ)	UW2 (u)	UW1 (u)
oo (o)	OY1 (ɔɪ)	OY2 (ɔɪ)	AO1 (ɔ)	UH2 (ʊ)	yv (j)	AW2 (aʊ)	OW2 (oʊ)	AH2 (ʌ)	AA1 (ɑ)
p0 (b)	B (b)	P (p)	M (m)	DH (ð)					

The symbols in the parentheses are IPA symbols. The symbol '-' in the parentheses denotes that there was no IPA symbol for that phoneme.

E The 4-nearest Korean phonemes for each English phoneme

ENG-phoneme	1st	2nd	3rd	4th	ENG-phoneme	1st	2nd	3rd	4th
AA0 (ɑ)	vv (ʌ)	wv (w)	aa (a)	ya (j)	IH1 (ɪ)	wq (w)	yq (j)	we (w)	ye (j)
AA1 (ɑ)	aa (a)	vv (ʌ)	wv (w)	wa (w)	IH2 (ɪ)	ee (e)	cc (ts̤)	yq (j)	c0 (dz)
AA2 (ɑ)	vv (ʌ)	aa (a)	nc (-)	yv (j)	IY0 (i)	ii (i)	wi (w)	lt (-)	xi (ɯi)
AE0 (æ)	ya (j)	ee (e)	h0 (h)	yq (j)	IY1 (i)	wi (w)	ls (-)	yq (j)	ii (i)
AE1 (æ)	ya (j)	aa (a)	ye (j)	lm (-)	IY2 (i)	ii (i)	wi (w)	lt (-)	yq (j)
AE2 (æ)	ya (j)	aa (a)	wa (w)	qq (ɛ)	JH (ʤ)	cc (ts̤)	ch (tsʰ)	c0 (dz)	lm (-)
AH0 (ʌ)	lh (-)	rr (ɾ)	xx (ɯ)	h0 (h)	K (k)	kh (kʰ)	kk (k̤)	th (tʰ)	kf (k)
AH1 (ʌ)	ya (j)	vv (ʌ)	oo (o)	yv (j)	L (l)	oo (o)	nn (n)	nh (-)	nf (n)
AH2 (ʌ)	wv (w)	wa (w)	yv (j)	ya (j)	M (m)	mf (m)	mm (m)	ng (ŋ)	pp (p̤)
AO0 (ɔ)	vv (ʌ)	ks (-)	wa (w)	oo (o)	N (n)	nf (n)	nn (n)	mf (m)	tt (t̤)
AO1 (ɔ)	oo (o)	vv (ʌ)	wv (w)	yv (j)	NG (ŋ)	ng (ŋ)	kf (k)	ks (-)	nf (n)
AO2 (ɔ)	oo (o)	vv (ʌ)	wv (w)	yo (-)	OW0 (oʊ)	yo (-)	oo (o)	qq (ɛ)	wv (w)
AW0 (aʊ)	aa (a)	ya (j)	vv (ʌ)	wa (w)	OW1 (oʊ)	yo (-)	we (w)	wo (ø)	wv (w)
AW1 (aʊ)	aa (a)	ya (j)	wa (w)	yq (j)	OW2 (oʊ)	yv (j)	yo (-)	wq (w)	lk (-)
AW2 (aʊ)	aa (a)	yv (j)	ee (e)	vv (ʌ)	OY0 (ɔɪ)	vv (ʌ)	wv (w)	wa (w)	lh (-)
AY0 (aɪ)	aa (a)	qq (ɛ)	lt (-)	wa (w)	OY1 (ɔɪ)	oo (o)	we (w)	wv (w)	uu (u)
AY1 (aɪ)	aa (a)	wa (w)	we (w)	vv (ʌ)	OY2 (ɔɪ)	oo (o)	uu (u)	wv (w)	wa (w)
AY2 (aɪ)	aa (a)	lt (-)	xx (ɯ)	wa (w)	P (p)	pp (p̤)	ph (pʰ)	kk (k̤)	th (tʰ)
B (b)	pp (p̤)	mm (m)	tt (t̤)	ph (pʰ)	R (ɹ)	ks (-)	ll (l)	we (w)	rr (ɾ)
CH (tʃ)	ch (tsʰ)	cc (ts̤)	kh (kʰ)	ks (-)	S (s)	ss (s̤)	ch (tsʰ)	s0 (s)	lp (-)
D (d)	tt (t̤)	t0 (d)	c0 (dz)	nc (-)	SH (ʃ)	ss (s̤)	c0 (dz)	s0 (s)	lk (-)
DH (ð)	th (tʰ)	t0 (d)	nc (-)	lm (-)	T (t)	ch (tsʰ)	c0 (dz)	th (tʰ)	lp (-)
EH0 (ɛ)	ya (j)	ye (j)	qq (ɛ)	wo (ø)	TH (θ)	th (tʰ)	tt (t̤)	ss (s̤)	ph (pʰ)
EH1 (ɛ)	ye (j)	lh (-)	yq (j)	nc (-)	UH0 (ʊ)	ee (e)	wi (w)	ll (l)	nn (n)
EH2 (ɛ)	h0 (h)	t0 (d)	ya (j)	wo (ø)	UH1 (ʊ)	yo (-)	qq (ɛ)	oo (o)	lm (-)
ER0 (ɝ)	ll (l)	rr (ɾ)	yv (j)	mm (m)	UH2 (ʊ)	oo (o)	wi (w)	uu (u)	yo (-)
ER1 (ɝ)	wo (ø)	wq (w)	ye (j)	aa (a)	UW (u)	aa (a)	xx (ɯ)	lb (-)	xi (ɯi)
ER2 (ɝ)	ll (l)	yu (-)	wo (ø)	ls (-)	UW0 (u)	yu (-)	yo (-)	uu (u)	wq (w)
EY0 (eɪ)	ye (j)	qq (ɛ)	xi (ɯi)	ee (e)	UW1 (u)	yq (j)	yu (-)	yo (-)	ye (j)
EY1 (eɪ)	we (w)	ye (j)	wq (w)	yq (j)	UW2 (u)	uu (u)	wi (w)	ps (-)	yu (-)
EY2 (eɪ)	we (w)	xi (ɯi)	wq (w)	qq (ɛ)	V (v)	lb (-)	t0 (d)	s0 (s)	mm (m)
F (f)	ph (pʰ)	ss (s̤)	kh (kʰ)	lk (-)	W (w)	oo (o)	pf (p)	wv (w)	lm (-)
G (g)	kk (k̤)	k0 (g)	ks (-)	tt (t̤)	Y (j)	ii (i)	yq (j)	lb (-)	wi (w)
HH (h)	h0 (h)	kh (kʰ)	lk (-)	ks (-)	Z (z)	s0 (s)	cc (ts̤)	ss (s̤)	t0 (d)
IH0 (ɪ)	ee (e)	h0 (h)	nf (n)	ii (i)	ZH (ʒ)	c0 (dz)	we (w)	s0 (s)	lt (-)

The symbols in the parentheses are IPA symbols. The symbol '-' in the parentheses denotes that there was no IPA symbol for that phoneme.