Proceedings of SIGUL2022 @LREC2022, pages 16–22
Marseille, 24-25 June 2022

Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning

Phat Do[1], Matt Coler[1], Jelske Dijkstra[2], Esther Klabbers[3]
[1]University of Groningen, Campus Fryslân, Leeuwarden, the Netherlands
[2]Fryske Akademy/Mercator Research Centre, Leeuwarden, the Netherlands
[3]ReadSpeaker, Driebergen-Rijsenburg, the Netherlands
{t.p.do, m.coler}@rug.nl, jdijkstra@fryske-akademy.nl, esther.judd@readspeaker.com

Abstract

We propose a new approach for phoneme mapping in cross-lingual transfer learning for text-to-speech (TTS) in under-resourced languages (URLs), using phonological features from the PHOIBLE database and a language-independent mapping rule. This approach was validated through our experiment, in which we pre-trained acoustic models in Dutch, Finnish, French, Japanese, and Spanish, and fine-tuned them with 30 minutes of Frisian training data. The experiment showed an improvement in both naturalness and pronunciation accuracy in the synthesized Frisian speech when our mapping approach was used. Since this improvement also depended on the source language, we then experimented on finding a good criterion for selecting source languages. As an alternative to the traditionally used language family criterion, we tested a novel idea of using Angular Similarity of Phoneme Frequencies (ASPF), which measures the similarity between the phoneme systems of two languages. ASPF was empirically confirmed to be more effective than language family as a criterion for source language selection, and also to affect the phoneme mapping's effectiveness. Thus, a combination of our phoneme mapping approach and the ASPF measure can be beneficially adopted by other studies involving multilingual or cross-lingual TTS for URLs.

Keywords: neural text-to-speech synthesis, under-resourced languages, cross-lingual transfer learning, phoneme mapping, language family

1. Introduction

Research in text-to-speech synthesis (TTS) has seen rapid advancement recently. Since the 2010s, there has been a paradigm shift to neural network-based speech synthesis (neural TTS), which produces much higher output quality in both naturalness and intelligibility compared to previous paradigms such as concatenative synthesis and statistical parametric speech synthesis (Tan et al., 2021).

However, neural TTS requires a large amount of training data. In TTS, training data refers to recordings of human speakers, preferably recorded with high quality (e.g., no or little background noise, good recording equipment, consistent speaking style and pronunciation), have reliable annotations (e.g., split into text-audio pairs with minimal or no discrepancies), and, in regards to quantity: the more the better. For an example, LJSpeech (Ito and Johnson, 2017), a public domain data set recorded by an American English female speaker that is widely used in neural TTS studies, has a duration of nearly 24 hours. Such an amount, though generally not hard to obtain for relatively highly-resourced languages, would likely be problematic for under-resourced languages (URLs).

One solution to address this challenge for URLs is to use cross-lingual transfer learning. This involves pre-training the acoustic model in a different language (called the "source language") that has sufficient training data, before fine-tuning that acoustic model with the limited training data of the URL ("target language"). This helps with the mapping between the input (text or phoneme sequence) and the output (speech features) in the URL, owing to the underlying similarities (e.g., patterns in pronunciation, semantic structures) among the language pair (Tan et al., 2021).

Cross-lingual transfer learning, however, comes with its own challenges. Firstly, there is often a mismatch between the input embeddings of the source and target languages, due to differences in their sets of phonemes or orthographic characters. To overcome this, Chen et al. (2019) proposed a Phonetic Transformation Network, fitted with a preceding automatic speech recognition component, to automatically map input symbols across languages based on their sounds. More recently, Wells and Richmond (2021) experimented between using phonemes and phonological features as input and made use of linguistic expertise (in the source and target languages) to map the embeddings. Notwithstanding these valuable findings, there is yet to be a solution that: a) is simpler but still sufficiently effective, b) can be easily replicated for other languages, and c) does not require specific linguistic expertise in the languages involved. We posit that such qualities are greatly helpful in cross-lingual transfer learning for URLs.

Secondly, numerous previous studies have shown that, for the same target language, transfer learning from different source languages leads to different effects in output quality. This leads to another consideration: by what criterion should the source language be chosen? Traditionally, language family classification has been widely used, with the implication that languages in the same family have more similarities that help in trans-

fer learning (or more generally, in sharing knowledge in a multilingual setting). However, an extensive study by Gutkin and Sproat (2017) found no conclusive evidence for this. In addition, in a meta-analysis of studies involving multilingual and cross-lingual TTS for URLs, Do et al. (2021) also concluded that language family classification was not an effective criterion for selecting source languages.
Accordingly, we aim to make the following contributions in this study:

1) We experiment on using a set of universal phonological features as a guide to map phoneme embeddings across source and target languages. (2.1)

2) We investigate a new criterion for selecting source languages: a measure of cross-lingual phoneme distribution similarity, and compare it with the conventional language family criterion. (2.3)

2. Databases and Proposed Metric

2.1. Phonological Inventory Data

PHOIBLE (Moran and McCloy, 2019) is a database of phonological inventories of 2,186 distinct languages. PHOIBLE uses a fixed set of 37 phonological features to describe all the phonemes in its database and ensures that each phoneme, represented by a unique IPA symbol, has a distinct set of binary attributes from these features. In other words, each IPA symbol representing a phoneme has a unique set of 37 binary attributes (corresponding to the phonological features) associated with that phoneme's pronunciation. This facilitates our proposed method for cross-lingual phoneme mapping, which is described in more detail in 4.2.2.

2.2. Language Classification Data

For language family classification, Ethnologue (Eberhard et al., 2021) is likely the most comprehensive and commonly used reference. It has been used by, e.g., Tan et al. (2019) as the reference for language clustering in their multilingual experiments, and by Do et al. (2021) as a potential factor in the effectiveness of multilingual or cross-lingual TTS models. To enable comparisons, we also use Ethnologue in this study.

2.3. Angular Similarity of Phoneme Frequencies (ASPF)

Cosine similarity (S_C or $\cos(\theta)$) is traditionally used in the field of natural language processing (NLP) to measure similarities between text documents, e.g., by Huang et al. (2011). Recently, a study by Cer et al. (2018) stated that the angular distance (D_θ, calculated from $\cos(\theta)$) performed better. Motivated by this, we experimented with using angular similarity ($S_\theta \coloneqq 1 - D_\theta$) between the vectors of phoneme frequencies of two languages to measure the similarity between their phoneme systems. For language A with phoneme set P_A, we defined a vector of phoneme frequencies PF_A containing frequencies of all phonemes in P_A, calculated from A's data set. To compare languages A and B, we calculated $\cos_\theta$ and then S_θ between PF_A and PF_B (with padding where necessary to avoid size mismatch):[1]

$$S_C(PF_A, PF_B) \coloneqq \cos_\theta = \frac{PF_A \cdot PF_B}{\|PF_A\|\|PF_B\|}$$

$$S_\theta \coloneqq 1 - \frac{2 \cdot \arccos(\cos_\theta)}{\pi}$$

Hereafter we use the name Angular Similarity of Phoneme Frequencies (ASPF) for these S_θ values, which represent the degrees of similarities between the phoneme systems of the two languages from which they are calculated ($0 \leq ASPF \leq 1$).

3. Data Sets and Preparation

3.1. Target Language Data Set

3.1.1. Frisian

Frisian ("Frysk") is the local language of the province of Friesland ("Fryslân"), which is located in the north of the Netherlands. The language has roughly 350,000 native speakers (Gorter, 2003), and has been recognized as the second official language of the Netherlands since 2013. Frisian is formally referred to as West Frisian (to distinguish from North Frisian and East Frisian), but in this study we simply call it Frisian.

3.1.2. Frisian Data Set

Although there are Frisian audio corpora, they were designed for other purposes than TTS. The FAME project (Yilmaz et al., 2016) corpus was designed to study code-switching and the Boarnsterhim Corpus (Sloos et al., 2018) was part of a longitudinal study. As such, they are not ideal for TTS research. Therefore, we created a small single-speaker corpus by using recordings and corresponding texts from a Frisian audiobook. We split the recordings by silence periods and also trimmed the preceding and trailing silences. Following LJSpeech, we further split long excerpts (while still respecting clause boundaries) so that the longest duration was 10 seconds. The corresponding texts had their sentences tokenized, abbreviations and numbers checked and expanded, and were thoroughly inspected to ensure good correspondence between text-audio pairs. From this corpus, we used 30 minutes of recordings (316 utterances) for this study and show their duration histogram in Figure 1.

3.2. Source Language Data Sets

CSS10 (Park and Mulc, 2019) is a publicly available single-speaker data set of 10 languages, consisting of short audio clips cut from audiobooks in the LibriVox project[2]. We chose it for this study since its wide range of languages enables the testing of the language family

[1] This is the formula for when the vectors do not contain negative values, which matches our case.

[2] `https://librivox.org`

factor, and its audio format and structure are similar to what we had for Frisian. From its 10 languages, considering a balance between language family variation and available audio duration, we chose to experiment with the following languages (in alphabetical order): Dutch, Finnish, French, Japanese, and Spanish.

We manually inspected these languages' subsets by listening to the audio files, skimming the paired texts, and remedying (or removing) the mismatches. The most common discrepancies included numbers that were not spelled out in the texts, book and chapter names that were read but not included in the texts, and differences in the audio/text splitting boundaries. To conform to the Frisian data set, we also excluded utterances longer than 10 seconds. Ultimately, each target language had approximately 9 hours of total duration, with similar duration distributions. Figure 1 shows the duration histogram of the Spanish data set as an example.

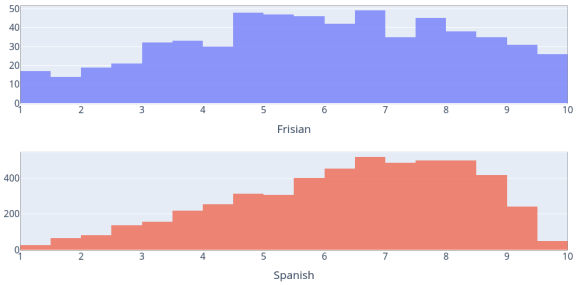

Figure 1: Duration (s) histograms of data sets

3.3. Data Sets Phonemization

We converted all data sets in this study using lexicons (pronunciation dictionaries). The Carnegie Mellon University Pronouncing Dictionary (CMU, 2014) (CMUdict) is a public domain dictionary for American English that is widely used in TTS research. We followed its conventions for phoneme annotations, with the following exceptions: a) we used IPA symbols from the PHOIBLE database instead of the modified ARPABET system in CMUdict, and b) we only included primary stress marks (i.e., secondary stress was treated as unstressed). The latter was in order to accommodate all the source languages involved, since not all of them can be said to have secondary stresses.

We used PHOIBLE to define the phoneme sets of all the languages. For languages that have more than one listed phoneme inventories (i.e., from several "doculects"), we used a union set from all of these, and then removed all the phonemes that were not used (i.e., not present) in the corresponding lexicon.

Frisian: We used the lexicon included as part of the FAME project, modifying it slightly to match the annotation method described above and supplementing it with the corresponding stress information provided by the Fryske Akademy.

Dutch: We used the e-Lex lexicon from the Instituut voor de Nederlandse Taal (INT, 2014), which uses phoneme representations from the Corpus Gesproken Nederlands (CGN) (Oostdijk, 2000) and was thus converted into IPA symbols following its manual. e-Lex includes stress information, and the majority of the entries are already manually checked by the authors.

All the other source languages used lexicons from the *ipa-dict* project (Doherty, 2019), which already uses IPA symbols and thus no conversion was needed.

Finnish: The lexicon readily contains stress information, so we only needed to exclude the secondary stresses.

French: As French does not have lexical stress, the lexicon does not contain stress information. Therefore, we determined the stressed phonemes using the rules from Kelton et al. (2019)[3], with the phrase boundaries predicted from punctuation marks and/or short breaks in the audio. We acknowledge that this is a rudimentary and oversimplifying approach, e.g., compared to that in de Dominicis et al. (2000). Nevertheless, we posited that this would suffice for the current study's purposes.

Japanese: One major challenge was that Japanese texts contain many homographs, which complicates the selection of the right pronunciation from the lexicon. CSS10 dealt with this by including *romaji* annotations (romanized transcriptions) that were post-edited by a native speaker. Although these still contain occasional mistakes, we used them as reference to determine the stressed phonemes. It should be noted that Japanese is not a stress-oriented language (de Dominicis et al., 2000) and instead has pitch (high-low) patterns. However, for the purposes in this study, we treated the vowels in high-pitched morae as stressed. Specifically, we used MeCab (Kudo, 2006) to parse the Japanese orthographic texts, compared them with CSS10's *romaji* annotations for the homographs, and obtained the stress information from a dictionary by javdejong (2022).

Spanish: The lexicon already contains stress marks for accented words. For the others, we followed the guide by Collins (2022) to determine the stress position.

For out-of-vocabulary (OOV) words in all languages, we used OpenNMT (Klein et al., 2017) to train a grapheme-to-phoneme (G2P) model for each language to predict the pronunciations and, to the extent possible, manually inspected and corrected the obvious errors.

4. Training and Evaluation

4.1. Source Language Pre-Training

We chose the FastSpeech 2 architecture (Ren et al., 2020), implemented by Chien et al. (2021) for the acoustic model. Pitch and energy prediction was done at the phoneme-level, following the authors' recommendation. For the vocoder, we used the universal generator of HiFi-GAN V1 (Kong et al., 2020) for all

[3]Available at `https://www.laits.utexas.edu/fi/html/pho/03.html`

source and target language models without fine-tuning, since the duration of the data sets (especially Frisian) was not sufficient for effective fine-tuning.

We trained a separate acoustic model for each source language. As done in the original FastSpeech 2 paper, we used the Montreal Forced Aligner (McAuliffe et al., 2017) to obtain phoneme-level alignments between the annotations and the audio recordings. We then trained each acoustic model for 100K parameter updates, with a batch size of 16 and the Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. To make sure they were trained successfully, for each model, we synthesized the corresponding set of 20 test sentences used in the CSS10 paper (Park and Mulc, 2019). The results had subjectively good quality and can be found online[4].

4.2. Target Language Fine-Tuning

To verify this study's proposed approach to phoneme mapping, we tested two scenarios for each source language: without and with phoneme mapping. We call the corresponding models *separate* and *mapped*, respectively, and describe them below.

4.2.1. Without Phoneme Mapping (*separate*)

In this scenario, we directly fine-tuned the source language model (described in 4.1) on the 30-minute Frisian data set. In other words, for the phonemes that are present in Frisian but not in the source language, the model would have their parameters initialized from scratch and "learn" from the Frisian data.

4.2.2. Phoneme Mapping (*mapped*)

In this scenario, for each phoneme not present in the source language, instead of initializing from scratch, we mapped it to the model parameters of its closest phoneme, which was predicted with a simple rule. The rule is expected to be universal (i.e., independent of the language pairs) and is as follows: for each target language phoneme that needed mapping, we looked for source language phoneme candidates with the most similar sets of PHOIBLE phonological features (represented as a vector of length 37). In case of ties, we compared the cosine similarities (2.3) of the phoneme distributions in the immediately preceding and succeeding positions of the phoneme in question, i.e., the candidate with the most similar adjacent phoneme distributions would be selected. For certain diphthongs and long vowels, no single target phoneme could be found. In that case, the source phonemes were decomposed into unitary vowels, which were subsequently mapped as if they were individual phonemes. All the resulting mapping decisions are reported in Appendix 8.

4.2.3. Model Fine-Tuning

Following the above descriptions, each of the 5 source languages had two separate fine-tuning scenarios: *separate* and *mapped*, both starting from the same pre-trained model. This resulted in a total of 10 fine-tuned models. Each model was trained on the 30-minute Frisian data set for another 100K parameter updates with a batch size of 4 (to better accommodate the small data size), with the other hyperparameters unchanged.

4.3. Evaluation

4.3.1. Test Sentences

We selected a total of 20 unseen test sentences, divided into 5 small sets of 4 sentences each, so that each set: a) contains all phonemes (regardless of frequency) from the Frisian data set[5], b) has a set-wide phoneme distribution as close as possible to that of the Frisian data set, and c) has an average duration of 5 seconds.

4.3.2. Listening Experiment

We used PsyToolkit (Stoet, 2010; Stoet, 2017) for an online listening experiment to obtain subjective evaluation, following the MUSHRA framework (Series, 2014). Each participant was randomly assigned a set of 4 sentences, each with a reference audio sample resynthesized from that sentence's ground-truth mel-spectrogram. The participant was then asked to listen to 10 synthesized samples (from the 10 models described in 4.2.3), together with a hidden resynthesized anchor, before being asked to rate each sample on its naturalness and pronunciation accuracy on a 0-100 scale. We collected answers from 50 participants that fully completed their panels, but had to exclude participants with lower self-rated Frisian proficiency. In the end, we used answers from 46 participants for data analysis ($n = 2024$). The audio samples are available online[4].

5. Results and Discussion

5.1. Phoneme Mapping

The MUSHRA scores are reported in Figure 2. To verify the effect of phoneme mapping, we conducted paired Wilcoxon tests between the scores of the models with and without phoneme mapping (*map* and *sep*). Table 1 reports the effects of phoneme mapping (differences in median scores between *map* and *sep*) and the *p*-values of the corresponding paired Wilcoxon tests, with statistically significant effects in bold.

Despite significantly increasing both naturalness and pronunciation accuracy ratings in the Dutch and Finnish models, phoneme mapping only increased accuracy ratings in the French model, and did not have a significant effect in the Spanish and Japanese models. To investigate this in more detail, we used a linear mixed effect model (Bates et al., 2014), with mapping as the fixed effect, and participants and sentences as random effects (to account for the by-participant and by-sentence variation). For both naturalness and pronunciation accuracy, phoneme mapping did affect the

[4] https://phat-do.github.io/sigul22

[5] This is usually not enforced by other studies, but we believe this would test the phoneme mapping more effectively, despite likely affecting the models' subjective evaluation negatively.

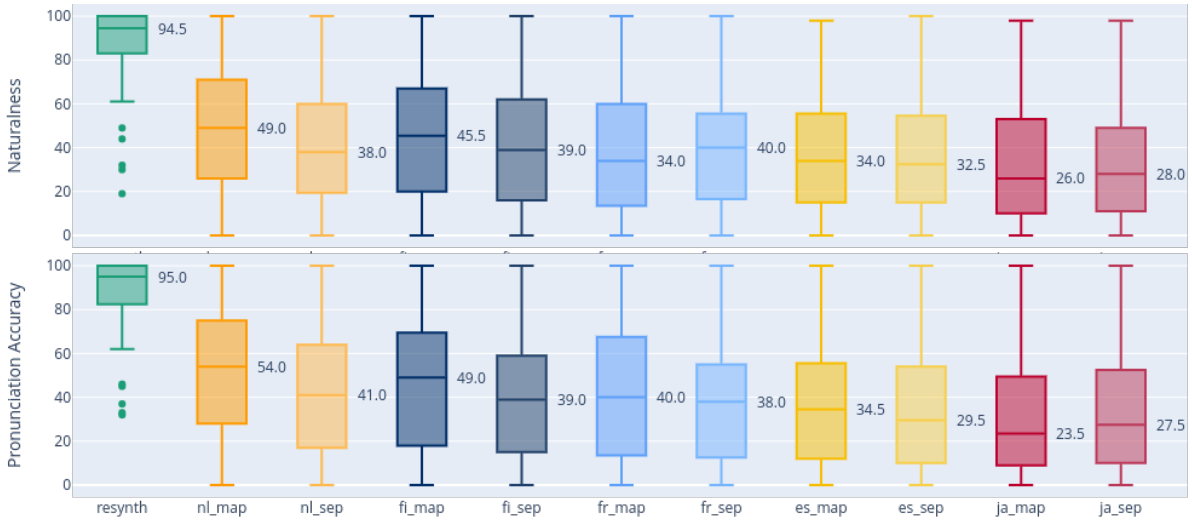

Figure 2: MUSHRA scores in Naturalness and Pronunciation Accuracy (central bars: median scores)

Source language	Naturalness ($M_{map} - M_{sep}$)	Accuracy ($M_{map} - M_{sep}$)
nl (Dutch)	**11 (p <.001)**	**13 (p <.001)**
fi (Finnish)	**6.5 (p = .003)**	**10 (p <.001)**
fr (French)	-6 (p = .82)	**2 (p = .02)**
es (Spanish)	1.5 (p = .17)	5 (p = .21)
ja (Japanese)	-2 (p = .56)	-4 (p = .11)

Table 1: Effect of phoneme mapping

MUSHRA score ($p = .004$ and $p < .001$, respectively), increasing it by 2.42 ($\pm$ 0.85) and 3.79 ($\pm$ 0.88), respectively. This means phoneme mapping did have an overall positive effect, but this effect also depended on the source language. This observation further motivated the analysis in the next stage.

5.2. Source Language Selection Criterion

5.2.1. Language Family

Acknowledging the complexity of measuring in detail the concept of language family distance, similar to Tan et al. (2019), we counted only the first level in the phylogenetic language classification tree (following the terms in Gutkin and Sproat (2017)). Accordingly, Frisian, Dutch, French, and Spanish were considered to be in the same language family (Indo-European), while Finnish (Uralic) and Japanese (Japonic) were not.

5.2.2. ASPF

Following 2.3, we calculated two versions of ASPF: a data set-level ASPF that compares two languages' whole data sets, and a sentence-level ASPF that involves the frequencies of only the phonemes present in each sentence. We posited that the latter was more accurate as a variable, and it also helped alleviate the issue of modeling a continuous variable with very few observed values, as the data set-level ASPF had only 5 values. It was still useful, however, in reaching a recommendation for source language selection criterion.

5.2.3. Results

Linear mixed effect models were used to test the effects of language family and sentence-level ASPF. When tested as the only fixed effect, they both had statistically significant effects on the MUSHRA score. However, since they are collinear by nature (languages in the same family are likely to have similar phoneme characteristics), we wanted to find the true effect that could explain the variation. Therefore, we used likelihood tests between these models and another model with both of them as fixed effects. This showed that language family indeed did not have a significant effect on either naturalness ($p = .56$) or accuracy ($p = .50$), while sentence-level ASPF significantly affected both ($p < .001$), increasing them by 2.93 ($\pm$ 0.36) and 3.66 ($\pm$ 0.37), respectively, for every increase of 10 percentage point in ASPF.

Sentence-level ASPF, however, is not very useful for generalization to other scenarios with other languages. Thus, we also tested for the correlation between data set-level ASPF (reported in Table 2) and the median MUSHRA scores, using the "Spearman" method. This showed that they were significantly correlated, with a coefficient of 1 and $p = .01$, confirming the usefulness of using data set-level ASPF as a criterion for choosing source languages (the higher, the better).

Source language	nl	fi	fr	es	ja
ASPF	0.73	0.47	0.38	0.35	0.33

Table 2: Data set-level ASPF (compared to Frisian)

6. Conclusion

We propose a novel approach for phoneme mapping in cross-lingual transfer learning, using phonological features of the PHOIBLE database and a language-independent mapping rule. We experimented with Dutch, Finnish, French, Japanese, and Spanish as source languages and Frisian as the target language. Listening scores showed that our approach improved both naturalness and pronunciation accuracy compared to without mapping. This effect also depended on the source language, motivating the investigation into a criterion to select source languages.

We then tested the idea of using Angular Similarity of Phoneme Frequencies (ASPF) as a criterion for selecting source languages, and proved through our experiment that it was more effective than the traditional criterion of language family classification.

Future research is intended to expand into experimenting in the setting of a directly multilingual model, with a wider range of languages, and in the scenario of having no available lexicons for the target language.

7. Acknowledgements

We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. We are thankful for the permission to use, for research purposes, the original recordings by Geartsje de Vries of books written by Kookos Tiemersma and published by the publisher Audiofrysk. We would also like to thank the Fryske Akademy (Leeuwarden, the Netherlands) for their support with the Frisian lexicon and the dissemination of the listening experiment.

8. Bibliographical References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder.

Chen, Y.-J., Tu, T., chieh Yeh, C., and Lee, H.-Y. (2019). End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech 2019*, pages 2075–2079.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592. IEEE.

CMU. (2014). cmusphinx/cmudict.

Collins. (2022). Which syllable to stress | Learning Spanish Grammar | Collins Education.

de Dominicis, A., Hirst, D., and Cristo, A. D. (2000). Intonation Systems: A Survey of Twenty Languages. *Language*, 76(2):460.

Do, P., Coler, M., Dijkstra, J., and Klabbers, E. (2021). A Systematic Review and Analysis of Multilingual Data Strategies in Text-to-Speech for Low-Resource Languages. In *Proc. Interspeech 2021*, pages 16–20.

Doherty, L. (2019). ipa-dict - Monolingual wordlists with pronunciation information in IPA.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition.* SIL International.

Gorter, D. (2003). Nederlands en Fries op gespannen voet. *Waar gaat het Nederlands naar toe*.

Gutkin, A. and Sproat, R. (2017). Areal and Phylogenetic Features for Multilingual Speech Synthesis. In *Proc. Interspeech 2017*, pages 2078–2082.

Huang, C.-H., Yin, J., and Hou, F. (2011). A text similarity measurement combining word semantic information with tf-idf method. *Jisuanji Xuebao (Chinese Journal of Computers)*, 34(5):856–864.

INT. (2014). e-Lex.

Ito, K. and Johnson, L. (2017). The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

javdejong. (2022). javdejong/nhk-pronunciation.

Kelton, K., Guilloteau, N., and Blyth, C. (2019). *Français interactif*. Lulu. com.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv:2010.05646 [cs, eess]*.

Kudo, T. (2006). MeCab: Yet Another Part-of-Speech and Morphological Analyzer.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Steven Moran et al., editors. (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Oostdijk, N. (2000). Het corpus gesproken nederlands.

Park, K. and Mulc, T. (2019). CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages. In *Proc. Interspeech 2019*, pages 1566–1570.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Series, B. (2014). Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*.

Sloos, M., Drenth, E., and Heeringa, W. (2018). The Boarnsterhim Corpus: A Bilingual Frisian-Dutch Panel and Trend Study. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4):1096–1104.

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, 44(1):24–31.

Tan, X., Chen, J., He, D., Xia, Y., Qin, T., and Liu, T.-Y. (2019). Multilingual Neural Machine Translation with Language Clustering. *arXiv:1908.09324 [cs]*. arXiv: 1908.09324.

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. (2021). A Survey on Neural Speech Synthesis. *arXiv:2106.15561 [cs, eess]*.

Wells, D. and Richmond, K. (2021). Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis. In *11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 160–165. ISCA.

Yilmaz, E., Andringa, M., Kingma, S., Dijkstra, J., Kuip, F., Velde, H., Kampstra, F., Algra, J., Heuvel, H., and van Leeuwen, D. A. (2016). A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research.

Appendix

Table 3 reports all the mappings resulted from the rule in 4.2.2, with vowels at the top and consonants at the bottom. Frisian phonemes are on the left column (*fy*). An empty cell means no mapping was needed, and a cell with two vowels mean they were converted from either a long vowel or a diphthong.

fy	**nl**	**fi**	**fr**	**es**	**ja**
a		ɑ			
aː	ɑː	ɑː	a a	a a	
ai	aɪ	ɑ i	a i		a i
eː			e e	e e	
ə		e		e	e
ɛ		e		e	e
ɛː	eː	eː	ɛ ɛ	e e	eː
ɛi	ɛ i	ɛ i	ɛ i	ei	e i
i					
iː	eː		i i	i i	
iə	i ə	i e	i ə	i e	i e
ɪ		i		i	i
ɪə	ɪ ə	i e	ɪ ə	i e	i e
o	ɔ				
ø				o	o
oː			o o	o o	
øː	yː		ø ø	o o	o o
œ	ʏ	ø		o	o
oə	ɔ ə	o e	o ə	o e	o e
ou	ɔ u	o u	o u	o u	o o
ɔ		o		o	o
ɔː		oː	ɔ ɔ	o o	oː
ɔu	ʌu	o u	ɔ u	o u	o o
u					o
uː	oː		u u	u u	oː
uə	u ə	u e	uə	u e	o e
ui	u i	u i	u i	oi	o i
y				u	o
yː			y y	u u	o o
yə	y ə	y e	y ə	u e	o e
b					
d					
f					ɸ
g					
h			f	f	
j					
k					
l					ɾ
m					
n					
ɲ		ŋ			ç
ŋ					ɴ
p					
r			l		ɾ
s					
t					
v				β	
x					k
z					

Table 3: Phoneme mapping results