SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Teaching American English pronunciation using a TTS service

Jorge Proença, Ganna Raboshchuk, Ângela Costa, Paula Lopez-Otero, Xavier Anguera

ELSA Corp.

{jorge, anna, angela, paula, xavier}@elsanow.io

Abstract

In computer-assisted language learning (CALL) applications students are able to learn/improve a language using automated tools. CALL applications benefit from having spoken examples by native language speakers in order to teach pronunciation. Realistically, this is limited to the pre-defined curricula that the application is teaching. In this work we allow the learner to practice pronunciation on freely input text, where the reference audio is generated using a text-to-speech (TTS) system. Instead of building a TTS system from scratch, we use a high quality external service (Amazon Polly TTS).

In order to successfully use Amazon Polly as a reference for teaching pronunciation, we carefully control the input text normalization and expansion steps and use the visemes information returned by Polly to select the best phonetic transcription out of all the possible transcriptions computed from the text. We show the usefulness of the approach by comparing the pronunciation scores obtained by a native speaker reading some test sentences to scores from the TTS audio on the same sentences. These show that the TTS audio reaches a similar pronunciation score as real audio, and therefore we conclude that it can be used as a reference for pronunciation learning. We also discuss and address issues of transcription and audio mismatch.

Index Terms: computer-assisted language learning, pronunciation assessment, speech synthesis

1. Introduction

Speech technologies are of paramount relevance in computer-assisted language learning (CALL) since they allow the development of automatic tools to assess the level of a learner's speaking proficiency in an L2 language. While the use of automatic speech recognition for pronunciation assessment is quite common in recent years [1, 2, 3, 4], the use of text-to-speech (TTS) technology is steadily appearing in language learning applications. The recent popularity gained by TTS technology for CALL applications is mainly motivated by the development of new speech synthesis paradigms that are able to produce high quality speech such as deep learning approaches based on recurrent neural networks [5], generative adversarial networks [6] or end-to-end strategies [7].

One of the main applications of TTS in CALL is dictation: there are several studies reporting that students achieve better results when the dictation is uttered by a TTS instead of a human [8, 9]. The experiments performed in [8] for Portuguese language suggest that this was probably due to the lower speech rate of the TTS compared to human dictation, which led to less word substitutions and less errors on function words. Analysis of dictation task with TTS versus human speech were performed in [10, 11, 12] in terms of comprehensibility, naturalness, accuracy and intelligibility.

TTS has also been used for prosody evaluation since the results are close enough to those achieved with human speech and the cost of the system is much lower [13]. Another related application is prosody transplantation, which consists in implanting the prosody of a native speaker to the speech of a student as proposed in [14]. Results presented in [15] suggest that self-imitation helps memorize and imitate intonation patterns of native speakers.

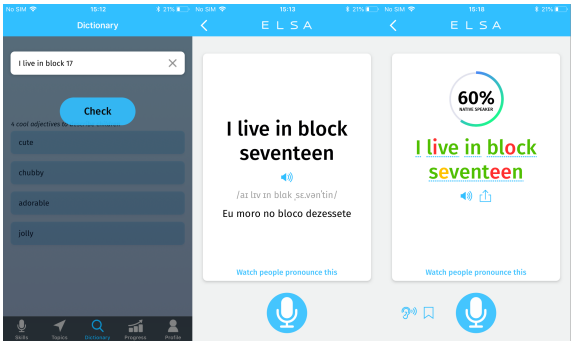

Figure 1: *Three screenshots of Elsa's dictionary mode (text input, pre-practice screen, post-practice screen).*

Besides the aforementioned applications, TTS is a powerful tool to provide feedback in CALL systems. It was used for this purpose in [16] to improve the phonological awareness of final -ed sounds, and in [17] as a corrective feedback resource through minimal-pairs experiments.

One of the problems encountered when using speech synthesis in CALL systems is that commercial TTS approaches do not always fulfill all the requirements of a specific application, and developing a proprietary TTS has a high cost. When assessing pronunciation, it is important that the phonetic transcription of the synthetic speech example matches exactly to the one expected from the user, obtained via a lexicon or a grapheme-to-phoneme (G2P) model. Also, text normalization is still an unsolved issue in both G2P conversion and TTS [18, 19].

This paper addresses the challenge of automatically generating spoken utterances from free text input by the users of a CALL system for the purpose of pronunciation practice and assessment. In this way, users can practice whichever text they would like, e.g., words or sentences that they encounter but do not know how to pronounce. This is achieved using Amazon Polly[1] as the TTS service, chosen for speed, cost and quality. In this work we discuss and propose solutions for the limitations described above when using this TTS service. In particular, the mismatch between expected and TTS phonetic transcriptions is solved by following a strategy that makes use of the visemes output by Amazon Polly to select the best matching phonetic pronunciation. The system described is currently in use in the ELSA app [20] for the free text input mode.

This paper is organized as follows: Section 2 presents the proposed implementation; Section 3 discusses results

[1] https://aws.amazon.com/polly/

10.21437/SLaTE.2019-11

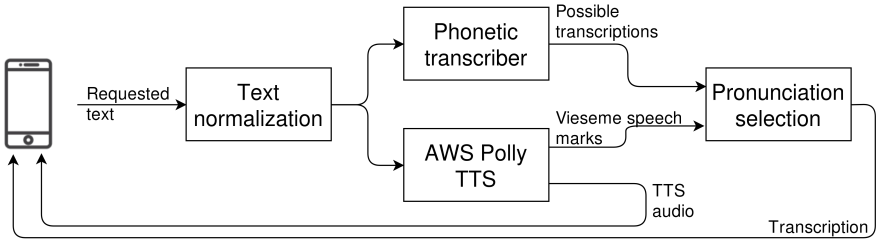

Figure 2: *Diagram of the developed free text input system.*

and issues with the implementation; and Section 4 summarizes some conclusions and future work.

2. System overview

The free text input mode of the ELSA app is called *dictionary* mode. Screenshots of the text input mode and dictionary practice screen are presented in Fig. 1. Separate from the built-in content and lessons of the app, which target specific skills, the *dictionary* mode allows a user to input any text that he/she would like to practice English pronunciation on. After a user inputs the text to be practiced, he/she will be presented with an automatically generated audio (obtained through TTS), the phonetic sequence of the way the sentence or word should be pronounced in the International Phonetic Alphabet (IPA) [21], and a translation of the input text to the user's mother tongue. After this, the user can open the microphone to say the sentence and the app will then provide feedback on the correctness of his/her pronunciation, generated by the pronunciation scoring system of the app. The specifics of the pronunciation scoring system are out of scope for the work presented here.

A requisite of the dictionary system is showing the phonetic transcription of how the text should be pronounced. This transcription can also be made mandatory during the pronunciation check stage, so that a user cannot pronounce a word in a way that would not make sense for this context, e.g., 'live' as the verb /lɪv/ or as the adjective /laɪv/. However, the TTS system we use for reference audio does not provide the sequence of phonemes that correspond to the synthesized pronunciation. Even though a specific pronunciation of our choice can be forced into the TTS service by using Speech Synthesis Markup Language (SSML), initial tests showed that the resulting audio suffered a lot in quality of intonation and fluency, and the pronunciation was often not ideal. The prosody of the synthesized audio seems more natural without forcing a transcription, however, we need a different way to obtain a phonetic transcription that is as close as possible to the TTS audio. This has been done by using the viseme information that the TTS service makes available given any input text.

Fig. 2 shows the diagram of the developed system for free text input: the text that is sent to the system is first normalized; the possible transcriptions for each word are obtained; the text is passed through Polly TTS outputting both the audio file and the speech mark information of visemes; if there are multiple transcriptions available for any of the words, these are compared to the viseme information, and the closest one is chosen; finally, everything is presented back to the user. In the next subsections we describe each step of this process in more detail. In addition, Table 1 shows the result of each of the steps for two example sentences.

2.1. Text normalization

The first step of the system, namely text normalization, is a common first step in TTS systems. The normalization of input text serves two main purposes: a) sanitizing the input by removing irrelevant special characters; and b) expanding the text of compact representations of words (e.g., numbers, measures, dates). With b), we can provide better visual feedback for each character of a word by matching graphemes to actually spoken phonemes instead of having a representation that only allows a large number of phonemes to correspond to one or few characters. For example, the input text '17' is expanded to 'seventeen' so that we can provide feedback for the user's performance on each phoneme, by assigning colors to associated graphemes, in this case /s-ɛ-v-ə-n-t-i-n/ matching to 's-e-v-e-n-t-ee-n' (as exemplified on Fig. 1), whereas for '17' there would be no clear assignment of phonemes. The alignment between graphemes and phonemes is achieved using a weighted finite-state transducer (WFST) model trained on the CMUDict dictionary[2].

Normalization is done using Sparrowhawk[3], Google's open-source implementation of their text normalization system [18]. The verbalization grammars initially taken from the official repository's toy example were augmented by adapting grammars from another source [22, 23], therefore covering more non-standard word expansion cases. While being quite flexible, this system also provides a fast runtime and outperformed other text normalization tools we tested both in terms of accuracy and speed. Currently, the normalization of cardinal and ordinal numbers, measures, date&time, some currencies, float numbers and fractions is supported.

2.2. Phonetic Transcription

At this step, we obtain the phonetic transcription of input text both in ARPAbet and IPA notations, where the former will be used for pronunciation check and the latter is shown to the user. For words with multiple transcriptions available, if possible we assign a preference to the phonetic transcription that is most commonly pronounced by the voice artist that represents the ELSA voice in the app's content. This is done to assure consistency with our content. The way we use these preferences is further discussed in Section 2.3. The process of obtaining a phonetic transcription for the input text is as follows:

First, we obtain the ARPAbet transcription for all the input words. There are several ways in which it can be done:

1. If the word is available in CMUdict, we take the transcription from there. To guarantee the transcription quality, there is an ongoing process in place inter-

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] https://github.com/google/sparrowhawk

Table 1: *Two sentences from input text to final selected transcription.*

Input text	Normalized text	Possible transcriptions (IPA)	Visemes - all transcriptions	Visemes - TTS	Selected transcription (IPA)
I	I	aɪ	a	a	aɪ
live	live	laɪv , lɪv	taf , tif	tif	lɪv
in	in	ɪn	it	it	ɪn
block	block	blɑk	ptak	ptak	blɑk
17	seventeen	ˌsɛ.vənˈtin	sEf@ttit	sEfttit	ˌsɛ.vənˈtin
The	The	ðə, ði	T@, Ti	T@	ðə
€5	five	faɪv	faf	faf	faɪv
	euros	ˈjuˌroʊz	iuros	iuros	ˈjuˌroʊz
will	will	wəl , wɪl	u@t, uit	uit	wɪl
last	last	læst	tast	tast	læst
a	a	ə, eɪ	@, e	@	ə
minute	minute	maɪˈnut, maɪˈnjut, ˈmɪ.nət, ˈmɪ.nɪt	patut, patiut, pit@t, pitit	pit@t	ˈmɪ.nət

nally at ELSA to manually curate the CMUdict transcriptions.

2. Otherwise, we run a grapheme-to-phoneme tool (Phonetisaurus [24]) to obtain the transcription automatically. The G2P model was trained using CMUdict.

3. If no transcription was obtained on Step 1, and if the word only contains letters and numbers, we also treat it as a possible acronym. We provide an additional alternative transcription in which each symbol is spelled out, which will be selected or not in the transcription selection step.

Secondly, IPA transcriptions are obtained from the ARPAbet ones by mapping the symbols accordingly. We also use a script[4] to perform syllabification of the transcriptions in order to be able to mark primary and secondary stresses as well as unstressed syllables.

2.3. Matching transcription to audio

The task of 'language modeling' and context inference to decide which pronunciation should be used is left to the TTS service so, instead of having that step in our system, in here we try to match as close as possible the selected phonetic transcription to the pronunciation decisions made by the TTS.

For the case of the Amazon Polly API, there is no possibility to ask for the sequence of uttered phonemes (which could solve the problem at once). However, a useful piece of information that can be retrieved from the Amazon Polly API is the sequence of visemes. Visemes represent the visual equivalent of phoneme productions [25, 26], i.e. the facial expressions that are made when pronouncing phonemes. This information is normally used to make animated avatars that move their mouth in a natural manner while speaking the corresponding synthesized audio. Since several phonemes can correspond to the same mouth position and therefore to the same viseme, the phoneme-viseme correspondence is a many-to-one mapping. Nevertheless, they are used here to distinguish between multiple pronunciations of words that usually have different visemes, except for some rare cases, so that we can present to the user the same phonetic transcription as the one actually spoken.

For each word of the normalized text, the possible multiple transcriptions obtained from the previous step are converted to visemes according to the TTS service guidelines[5]. Then, the Levenshtein distances between the reference sequence of visemes for a word and the visemes of each transcription are calculated. The transcription that obtains the minimum edit distance is selected to represent the word. If there are multiple transcriptions that are tied for minimum distance, the one that was previously tagged as preferred is chosen. Otherwise, the first one is selected.

In this way, each word will only have one transcription that will represent the phonetic IPA sequence displayed to the user. This transcription will also be forced as the only pronunciation allowed if a user starts a pronunciation check attempt, which implies that speaking any alternative pronunciation of a word will probably result in a pronunciation error even if it is possible as an alternate pronunciation in US English.

3. Evaluation

To evaluate the performance of the system, we first compare the TTS audio to spoken utterances of a native speaker. We selected cases of regular content of the app for which recordings by a native voice artist exist. Specifically, samples of the *assessment test* section are considered, which consists of 13 long sentences about travelling and about speaking fluency in English. Each sentence is also accompanied by a transcription where, for certain words, a specific transcription is forced (manually decided). The audio recordings of the voice artist speaking these sentences are run through our pronunciation score check with the forced transcription, simulating a normal attempt by users in the app. In contrast, the same sentences are run through our *dictionary* system as text input to obtain a TTS audio. Such output, with automatically generated forced transcriptions, are then run through our pronunciation score check.

Table 2 shows the average phonetic pronunciation scores (averaging the total score over all sentences) obtained by running the voice artist audio and the TTS audio on the test sentences with the two versions of forced transcription. For comparison, scores when no transcription is forced (allowing any alternative pronunciation in our lexicon to be chosen) are also shown. Additionally, we calculated the phonetic scores of assessment test samples of 100 randomly selected users and present average scores and the standard deviation amongst all users.

The first result that stands out from not forcing any transcription is that the quality of the TTS audio is very good,

[4] https://github.com/kylebgorman/syllabify

[5] https://docs.aws.amazon.com/polly/latest/dg/ph-table-english-us.html

Table 2: *Average phonetic pronunciation scores (%) from ELSA's pronunciation check (user standard deviation in parentheses).*

	No forced transcription	Manual forced transcription	Dictionary forced transcription
Voice artist	97.94	96.83	94.89
TTS	97.09	96.24	94.06
Users	77.91 (15.27)	77.07 (15.35)	76.44 (14.92)

obtaining 97.09% pronunciation score versus 97.94% score from the native speaker. However, when including their respective forced transcriptions, the result drops slightly more for the TTS audio with its dictionary transcription. As mentioned, the forced transcription of the *dictionary* used with the TTS audio represents the phonetic matches made by the viseme information obtained from the TTS system. The transcription should theoretically match with the audio but the decrease in phonetic score indicates that some mismatches occurred.

By analyzing where the pronunciation score system reported phonetic errors, it could be observed that most cases happened in short words such as 'and', 'the', 'to', 'will'. We find that these words have a common characteristic: they have both strong and weak versions of pronunciation, e.g., 'and' with /ænd/ or /ən/, 'the' with /ði/ or /ðə/, 'to' with /tu/ or /tə/, 'will' with /wɪl/ or /wəl/ (also 'a', 'an', 'at', 'of', 'than', etc.). By forcing the pronunciation that was selected from viseme comparison for these words, our pronunciation check system is often detecting the alternative pronunciation and reporting that an error was made, which lowers the score. This may be due to how the acoustic models that provide phonetic scores were trained, giving preference to one context over the other or often detecting the weak forms on a fast reading speed (or vice versa).

Given the above behavior, it is best to not force a pronunciation in the *dictionary* mode for these short words with strong and weak versions. A user would then be able to get a correct pronunciation on both versions of the words since they both exist in the lexicon. This seems acceptable as they may want to read slower, faster or with a specific emphasis. By redoing the forced transcription for TTS audio in this fashion and running pronunciation check again, the score on the test sentences improves substantially (96.62%), close to the score when the transcription is not forced. The difference in score is due to a few words that also have alternative pronunciations and where the acoustic models considers that the transcription selected from viseme information is wrong.

This result leads to the question: is it necessary to force a transcription at all and going through the trouble of viseme matching? There are still reasons for doing so:

- Homograph words that have different pronunciations should still be forced to not allow the user to speak the pronunciation that gives a different meaning to the word.
- Since we want to show only one IPA sequence to the user in *dictionary* mode, applying the system to all words and obtaining the phonetic sequence that best matches the audio is still an important task.
- Severe mismatches between our obtained transcriptions and the viseme sequence of the TTS service can be logged and analyzed. These mismatches are often related to words that are not in our lexicon and for which the G2P did not provide a similar transcription to what is spoken by the TTS audio. Although most of these cases are user typos or gibberish, some are rare or technical words that we want to add to our lexicon (e.g., 'gingival', 'antiperspirant', 'paroxysm'). This is a process that is already in place to improve the system.

A limitation of the viseme-matching approach is for homograph words whose distinct pronunciations still lead to the same visemes (e.g., 'close', with the verb or noun pronounced as /kloʊz/ and the adjective as /kloʊs/ having the same viseme sequence /ktos/). The only way to correctly evaluate entries with these words would be to depend on context or language modelling information.

From the user scores shown in Table 2, it can be seen that the average user performance is significantly below the voice artist and TTS audio. This shows that the TTS audio can be useful as a reference pronunciation that users can match to improve their performance. The score for users when including the full dictionary forced transcription did not degrade as much as the others. Further analysis is needed to show whether users' scores degrade when users use the TTS as reference versus using the human reference.

4. Conclusions

In this paper, we presented our work on an exercise that lets users practice pronunciation of any word or sentence of their choice. Given an input text, text normalization and non-standard word expansion are applied, and the corresponding transcriptions are obtained. The reference audio is synthesized via the Amazon Polly text-to-speech service and the most appropriate transcription for words with multiple pronunciations is selected by matching viseme information.

Results show that the TTS voice achieves pronunciation scores that are very close to a native speaker of American English. We can conclude that using synthesized speech as a spoken sample can be a good reference for L2 learners that need to improve their pronunciation. Some issues are found from forcing one transcription for all the words of a sentence, specifically with words that have weak and strong versions of pronunciation that frequently result in errors on the pronunciation check module. This leads us to not force a transcription for these common words and to focus our efforts on solving the issue of homograph words.

Even if the viseme matching strategy cannot solve all dubious cases, a process is already in place to check for mismatches and improve the lexicon. Eventually, if pronunciation issues are found with the TTS synthesis itself, it would also be possible to force a specific transcription for a word to be spoken by the synthesizer.

Additionally, for future work we should compare how well students learn and perform when having TTS audio as a reference versus a real native pronunciation, to further confirm that TTS audio can replace real recordings when needed.

5. References

[1] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-natvie spontaneous speech," in *ICASSP*, 1016, pp. 6140–6144.

[2] S. M. McCrocklin, "Pronunciation learner autonomy: the potential of automatic speech recognition," *Syhstem*, vol. 57, pp. 25–42, 2016.

[3] M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of l1 in automatic pronunciation evaluation of L2 speech," in *Interspeech*, 2018, pp. 1636–1640.

[4] F. Eichenberger, P. Bouillon, J. Gerlach, and M. Déjos, "Automatic evaluation of the pronunciation with CALL-SLT, a conversation partner exclusively based on speech recognition," in *Proceedings of EDULEARN18 Conference*, 2018, pp. 6592–6597.

[5] Z. Wu, O. Watts, and O. King, "Merlin: an open source neural network speech synthesis system," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 202–207.

[6] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[7] Y. Wang, R. Skerry-Ryan, D. Santon, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Sourous, "Tacotron: towards end-to-end speech synthesis," in *Interspeech*, 2017, pp. 4006–4010.

[8] T. Pellegrini, A. Costa, and I. Trancoso, "Less errors with TTS? a dictation experiment with foreign language learners," in *Interspeech*, 2012, pp. 1291–1294.

[9] H.-H. Chiang, "A comparison between teacher-led and online text-to-speech dictation for students' vocabulary performance," *English Language Teaching*, vol. 12, no. 3, pp. 77–93, 2019.

[10] T. Bione, J. Grimshaw, and W. Cardoso, "An evaluation of TTS as a pedagogical tool for pronunciation instruction: the 'foreign' language context," in *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017*, K. Borthwick, L. Bradley, and S. Thouesny, Eds. Research-publishing.net, 2017, pp. 56–61.

[11] N. Ní Chiaráin and A. Ní Chasaide, "Effects of educational context on learner's ratings of a synthetic voice," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 47–52.

[12] J. Grimshaw, T. Bione, and W. Cardoso, "Who's got talent? comparing TTS systems for comprehensibility, naturalness, and intelligibility," in *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018*, P. Taalas, J. Jalkanen, L. Bradley, and S. Thouesny, Eds. Research-publishing.net, 2018, pp. 83–88.

[13] Y. Xiao and F. K. Soong, "Proficiency assessment of ESL learner's sentence prosody with TTS synthesized voice as reference," in *Proc. Interspeech*, 2017, pp. 1755–1759.

[14] Y. K., "Imposing native speakers' prosody on non-native speakers' utterances: the technique of cloning prosody," *Journal of the Modern British & American Language & Literature*, vol. 25, no. 4, pp. 197–215, 2007.

[15] E. Pellegrino and D. Vigliano, "Self-imitation in prosody training: a study on Japanese learners of Italian," in *Proc. 6th ISCA Workshop on Speech and Language Technology in Education*, 2015, pp. 149–154.

[16] A. A. de Araújo Gomes, W. Cardoso, and R. Marques de Lucena, "Can TTS help L2 learners develop their phonological awareness?" in *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018*, P. Taalas, J. Jalkanen, L. Bradley, and S. Thouesny, Eds. Research-publishing.net, 2018, pp. 29–34.

[17] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso Payo, "Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 25–29.

[18] P. Ebden and R. Sproat, "The kestrel tts text normalization system," *Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, 2015.

[19] R. Sproat and N. Jaitly, "RNN approaches to text normalization: A challenge," *CoRR*, vol. abs/1611.00068, 2016.

[20] X. Anguera and V. Van, "English Language Speech Assistant," in *Interspeech*, 2016, pp. 1962–1963.

[21] I. P. Association, C. PRESS, and I. P. A. Staff, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999. [Online]. Available: https://books.google.pt/books?id=33BSkFV_8PEC

[22] K. Gorman and R. Sproat, "Minimally supervised number normalization," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 507–519, 2016.

[23] A. H. Ng, K. Gorman, and R. Sproat, "Minimally supervised written-to-spoken text normalization," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 665–670.

[24] R. J. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[25] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp. 796–804, 1968.

[26] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40–67, 2017.