Investigating the Use of Synthetic Speech Data for the Analysis of Spanish-Accented English Pronunciation Patterns in ASR

Margot Masson, Julie Carson-Berndsen

Science Foundation Ireland Centre for Research Training in Digitally Enhanced Reality
School of Computer Science, University College Dublin, Ireland

margot.masson@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

Understanding speech recognition errors, especially those related to accents, is challenging due to the complexity of the models and scarcity of data. This paper addresses this issue by exploring the use of synthetic data to investigate accent-related variations and their impact on Automatic Speech Recognition (ASR) performance. We synthesise Spanish-accented English and compare the speech features captured by synthetic speech with those found in natural speech. We generate speech with phoneme-level variation using Spanish voice synthesis and phoneme-to-speech synthesis and then assess ASR sensitivity to such variations. Our findings show that synthetic data captures phonemic patterns of Spanish well, suggesting its utility, coupled with ASR, in L1-L2 phonemic difference modelling. In contrast, phonotactic patterns are not captured to the same extent by synthetic data. We also show that the variants built from the synthetic data accurately challenge ASR systems, prompting a potential method for testing and enhancing ASR accent robustness and explainability for speech research.

Index Terms: speech synthesis, speech recognition, accents

1. Introduction

It is widely acknowledged that automatic speech recognition (ASR) systems, although demonstrating high performance in accurately transcribing standard English, are much less accurate when handling accents outside their training scope [1]. This accent-related limitation is significant, given the global deployment of these systems, making accent robustness crucial for the improvement of ASR systems. The complexity of the architectures results in a lack of understanding of their error patterns and underlying phonetic phenomena.

While the challenge of ASR explainability has been previously addressed from different perspectives, it remains unclear which ASR errors stem directly from accents compared to other speech features. The scarcity of accented speech data further impedes ASR accent-robustness testing and enhancement. For this reason, synthetic data, which offers flexibility and quantity, seems promising for overcoming this challenge. This is the question we seek to address in this paper.

In this paper, we propose using synthetic data to create variants to investigate the impact of accent-related variation on ASR. Our aim is twofold. Firstly, we investigate how the information captured in synthesised Spanish-accented English speech compares with existing phonological knowledge of Spanish-accented English. Secondly, we assess ASR sensitivity to such accent-specific variations. We explore the phonological features captured by synthetic accented speech data using a text-to-speech (TTS) system to synthesise artificially accented speech and use the resulting ASR confusions to produce synthetic variants incorporating targeted, phoneme-level accent-specific variations.

Our analysis of how ASR copes with both synthetic and natural accented speech shows that synthetic speech can indeed infer Spanish-English phonemic confusion patterns, demonstrating that synthetic data can help in language modeling and variation analysis in under-resourced scenarios, and for ASR accent-robustness test and improvement. While this research highlights the potential of synthetic speech in understanding phonemic confusion patterns, our investigation also reveals that synthetic speech capture of phonotactic patterns is only partial.

The remainder of the paper is structured as follows. Section 2 discusses the related work and Section 3 presents those aspects of Spanish phonology relevant to the analysis. Section 4 describes our method for the generation of variants and Section 5 presents and discusses our results. Section 6 concludes the paper with some pointers to future work.

2. Related Work

This section reviews relevant literature, focusing on ASR challenges with accented speech, previous efforts to improve robustness and explainability, and the use of synthetic data in ASR and speech research more generally.

Improving ASR robustness to accented speech has been a significant research focus. Solutions such as data augmentation [2], model adaptation [3], and transfer learning [4] have been explored to improve ASR robustness to accented speech. These methods show promise but face limitations, especially with less common accents, due to the scarcity of accented speech data and a limited understanding of the learned features in these systems.

To better understand ASR error patterns, several studies have investigated linguistic representations in deep neural networks [5] and recent ASR systems [6, 7]. While these studies do not focus on accents, consistent accent-related error patterns in ASR have been identified for natural data [8]. This, and the fact that ASR has been used for linguistic studies such as data annotation or speech feature analysis [9], suggest that ASR weaknesses can be used to retrieve L1-L2 pronunciation differences. This is the main idea behind our work.

As natural accented speech data can be scarce, particularly for under-resourced languages, it seems natural to turn to speech synthesis to augment data. Synthetic data has been used for data augmentation [10, 11], highlighting its potential for ASR robustness improvement. Synthetic data also serves as a valuable tool for ASR testing due to its flexibility. For example, synthetic speech with word-level variations has been used to test ASR sensitivity [12]. Instead of word-level variations, we propose to look at pronunciation variations, and investigate the extent to

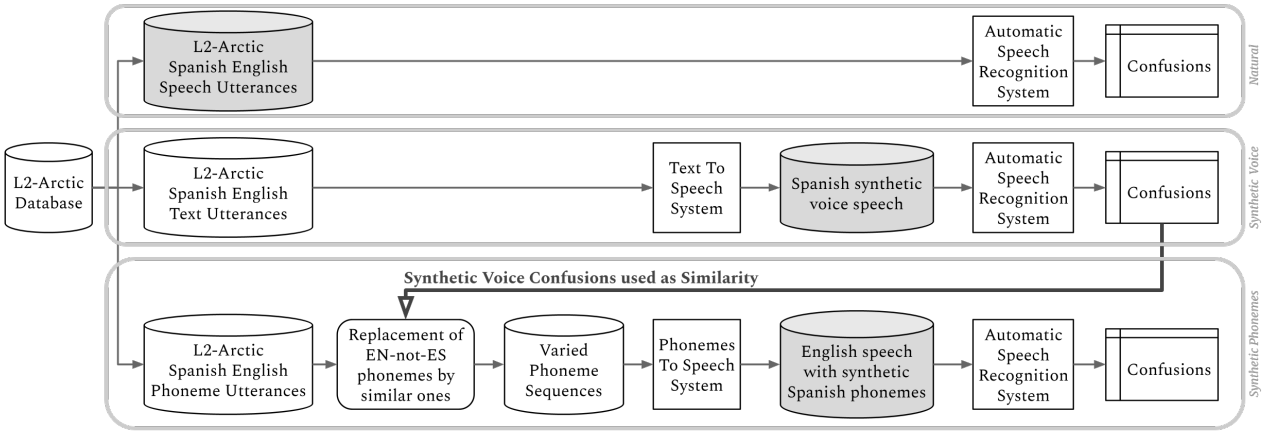


Figure 1: *The three different configurations used to investigate phonemic confusions.*

which we can use synthetic data to infer L1-L2 variations, and to test ASR systems.

This paper builds on previous findings related to ASR explainability and explores the use of synthetic speech for language differences modeling and ASR testing. Although not the primary focus of this work, synthetic data has demonstrated its utility in enhancing ASR systems, suggesting potential applications of the work presented here for ASR accent-robustness improvement. In our previous study [13], we leveraged synthetic French-accented speech to build a French-English similarity matrix and shown that it captured French-English confusion patterns more accurately than a purely knowledge-based approach. This study focused on the paradigmatic aspects of an accent. In our current study, we expand our focus to address non-native accents at both paradigmatic and syntagmatic levels.

3. Spanish Phonology

An accent refers to a variation in phoneme realisation as opposed to a so-called "standard" pronunciation. In the case of non-native accents, these variations are in great part due to the differences between the pronunciation rules of the native language of the speaker (their L1) and the target language (their L2). Thus, non-native speakers usually encounter difficulties producing and perceiving specific segmentals [14] or suprasegmentals [15] that are present in the L2 but absent or different in the L1. Spanish phonology differs from English phonology in several aspects [16, 17, 18, 19, 20]. Firstly, they have different phoneme sets, resulting in phonemic variations. English has twenty-four consonants while Spanish has nineteen. Native speakers of Spanish have difficulty in pronouncing /z/, /v/, /ð/. Also, Spanish has only five pure vowels (/a/, /e/, /i/, /o/, /u/), whereas English has twelve ([20, 18]).

Secondly, the syllable structure differs considerably between the two languages, with English showing a much more diverse range of consonant clusters than Spanish. This leads to phonotactic variations. In Spanish, syllables generally begin with a single consonant, two consonants or a vowel, and end similarly. Conversely, in English, syllable can start with up to three consonants. Additionally, Spanish words end with a vowel most of the time. As a result, Spanish speakers tend to add a vowel - usually a realisation of /e/, which is the default vowel in Spanish - to the beginning of English words with an initial /s/ consonant cluster (termed sC clusters), for example pronouncing "estreet" instead of "street".

4. Accented Speech Data Synthesis

Above we highlighted two types of Spanish variations in English: phonemic variations which arise when English-but-not-Spanish (EN-not-ES) phonemes are encountered, and phonotactic variations which occur when pronouncing EN-not-ES consonant clusters independently from phonemic constraints. In this section, we describe how we leverage synthetic Spanish-accented English data (*synthetic voice*) confusions to generate synthetic variants that capture these variations (*synthetic phonemes / phonotactics*). While we address all phonemic variations, we focus in this paper on a single phonotactic phenomenon, the /e/ epenthesis before sC clusters.

4.1. Experimental Setup

We choose to use L2-Arctic [21] for our experiments. L2-Arctic is a speech corpus of non-native English which contains recordings from twenty-four non-native speakers of English, including Spanish speakers of English. For our experiments, we use the L2-Arctic Spanish-English utterances, which consist of 4401 recordings of four Spanish natives - two males and two females - reading English prompts from CMU's ARCTIC [1], with the corresponding phonetic transcriptions.

For speech recognition, we use the wav2vec 2.0 [22] base model [2], fine-tuned and pretrained on 960 hours of Librispeech [23] for speech recognition. In addition, we also use for phoneme recognition the wav2vec2Phoneme [24] base model [3], which is base wav2vec2 fine-tuned on Common-Voice [25] to recognise phonemes. Finally, we use the Microsoft Azure TTS [4] for speech synthesis. This choice is motivated by the fact that this TTS system accepts both English and non-English phonemes as input.

4.2. Generation of Speech with Phonemic Variations

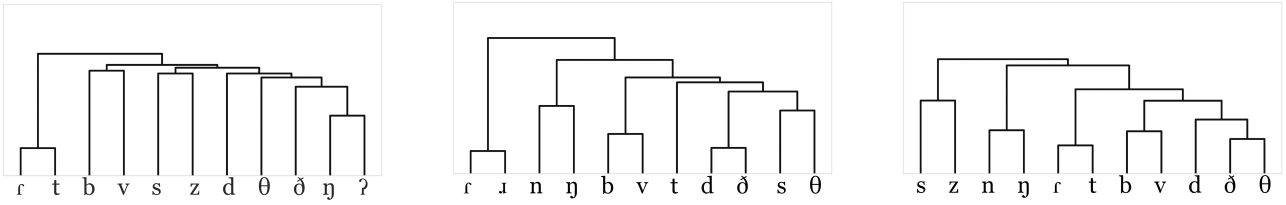
Before assessing the ASR, we generate English speech with Spanish phonemic variations using two different methods. The

[1] http://festvox.org/cmu_arctic/

[2] https://huggingface.co/facebook/wav2vec2-base-960h

[3] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

[4] https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/text-to-speech



(a) *Natural non-native speech dendrogram.* (b) *Synthetic voice speech dendrogram.* (c) *Synthetic phonemes speech dendrogram.*

Figure 2: *Hierarchical view of the ASR confusions for some target consonants.*

first method consists of providing the English text utterances from L2-Arctic Spanish to the TTS system configured to synthesise Spanish. So we input the English text in Azure TTS with its parameter "voice" set to Spanish, resulting in English text read as if it was written in Spanish. This method is intended to represent the worst case scenario, where a Spanish native speaker encounters English for the first time. Thus, we are building on the generation bias to gain insight into the information captured by the TTS and whether it relates to expected Spanish patterns in English. This pipeline is denoted *synthetic voice* and is illustrated in the second box on Figure 1.

Then, we input the *synthetic voice* speech to the ASR and compute the confusion matrix of the transcribed phonemes. In order to verify that synthetic accented speech can be used to produce variants that challenge the ASR, we use this confusion matrix as a similarity measure for the second generation method. This *synthetic similarity matrix* is expected to have captured the phonemic variations of Spanish English and is used in the second method, called *synthetic phonemes*. This method is illustrated in the third box on Figure 1. It consists of replacing the EN-not-ES phonemes by their closest neighbour in the synthetic similarity matrix. First, we use a Spanish-to-English *phonemic mapping* to identify the EN-not-ES phonemes to be replaced. Then, the replacing phonemes are chosen from the *synthetic similarity matrix* and the EN-not-ES phonemes are varied, i.e. replaced by the chosen ones. The Azure TTS is used as a phoneme-to-speech system to generate the varied speech directly from the varied phoneme sequence.

In order to investigate the phonotactic confusion patterns of Spanish English, we filter the sentences from the above mentioned *natural* and *synthetic voice* speech data which contain sC clusters. As a counterpart to *synthetic phonemes*, we produce a *synthetic phonotactic* set, which consists of speech synthesised after adding /e/ before sC clusters in the phoneme sequences.

5. Results

5.1. Phonemic Confusions

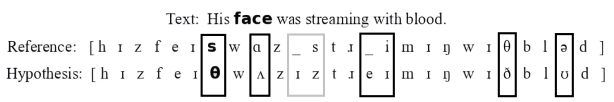


Figure 3: *Alignment of an L2-Arctic speech sample. Phonemic confusions are annotated in black, phonotactic ones in grey.*

For the purpose of investigating the extent to which synthetic speech can be used to model non-native confusion patterns, and to generate variants for ASR robustness assessment, we run the ASR on all three audio sets described in Section 4.2 and illustrated by the shaded areas in Figure 1:

- *natural* Spanish-English from L2-Arctic as baseline;
- *synthetic voice* speech for L1-L2 confusions modelling;
- *synthetic phonemes* speech as variants for ASR testing.

The reference phonemes and transcription from wav2vec 2.0 are force-aligned (see Figure 3 for illustration on a *natural* sample) and the confusions matrices are computed. We visualise the confusions obtained for each set as similarity hierarchies using the Ward clustering method [26]. Figure 2 shows excerpts of these hierarchies, corresponding to each of the three sets. [5]. These hierarchies highlight interesting clusters (cf. Section 3), such as [/s/,/z/], present in *synthetic phonemes* confusions as well as in *natural* Spanish confusions. Similarly, the specifically Spanish confusion [/b/,/v/] is present in all three sets. The phonemes /ð/, /θ/, /t/, /d/ and /ɾ/ are clustered differently across the three sets, but some pattern still emerges with /ð/ and /ɾ/, which are EN-not-ES phonemes, being often confused with and thus regarded as similar to one of the remaining /θ/, /t/ and /d/.

The fact that in *synthetic phonemes* we retrieve Spanish pronunciation patterns suggests that *synthetic voice* accurately captures the pronunciation patterns of Spanish, and that the ASR is sensitive to phoneme-level variations. In the next subsection, we investigate the extent to which we can capture phonotactic differences with synthetic speech.

5.2. Phonotactic Confusions

Again, we run the ASR on the three filtered sets containing sentences with sC clusters. First, we compute the transcribed epenthesis rate, that is the percentage of sC clusters transcribed as vowel + sC cluster by the ASR. In order to have a better understanding of the patterns underlying the recognition of the epenthesis by the ASR, we looked at the last phoneme of the word preceding words with initial sC. Indeed, the sequence of final consonant cluster followed by initial consonant cluster is more problematic for Spanish speakers than the sequence final vowel followed by initial consonant cluster, and induces epenthesis more often. Table 1 shows rates of epenthesis recognition relative to the number of corresponding sequences. The two sequences we examine are: 1) when the previous word ends with a consonant, denoted C_sC, and 2) when the previous word ends with a vowel, denoted V_sC. As expected, the epenthesis is recognised more often in the case of previous consonant than previous vowel. This table also reveals that *synthetic phonotactic* results in more epenthesis being recognised after a vowel than the other methods. That can be explained by the fact that we manually added a /e/, indiscriminately to the previous word, while the human speakers are less likely to epenthesise an /e/ when a vowel is already present.

It can be noted that only a few of the original sentences

[5] Full confusion matrices and hierarchies can be found at https://tinyurl.com/results-phonemic

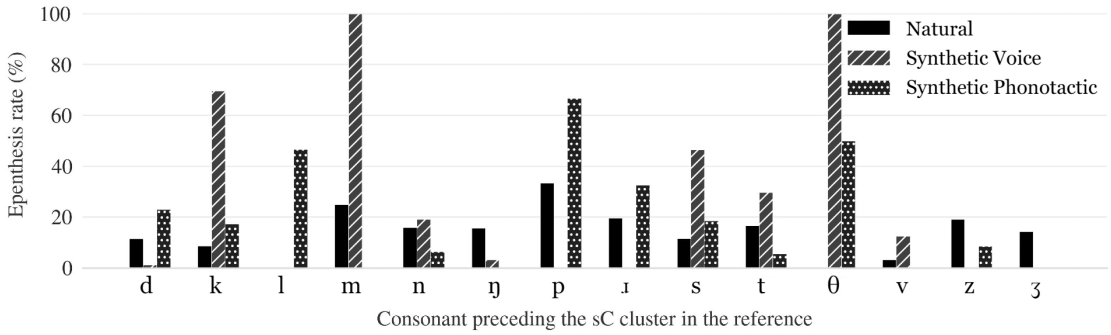


Figure 4: *Capture rate of epenthesis before sC clusters for different preceding phonemes.*

Table 1: *Transcribed epenthesis rate for vowel versus consonant-preceding sC clusters.*

	Cluster	Natural	Syn.Voice	Var.Pho.
epenthesis	C_sC	15.2%	18%	15.2%
	V_sC	0.9%	2.4%	7.3%

where transcribed with an epenthesis. To explore the context in which the epenthesis is recognised, we look into the details of the consonant preceding sC. Figure 4 illustrates emerging rate in the ASR of the epenthesis for the three types of speech and for each preceding consonant. Interestingly, preceding /s/ does not appear to result in much epenthesis by Spanish native speakers. That can be explained, as well as the low rate of epenthesis in the *natural* set, by a hyper-articulation of the difficult sequence [C+sC] that we have confirmed by listening to the *natural* samples. Another interesting pattern lies in the 100% capture rate of the epenthesis in the case of preceding /θ/ or /m/ for *synthetic voice* speech. There are only 16 utterances of [m s C] in the data, and 8 utterances of [θs C], compared to 87 utterances of [d s C] for which the epenthesis was recognised only once.

It is also interesting to note that *synthetic voice* and *synthetic phonotactic* do not follow the same tendencies, suggesting that the /e/-to-speech from the phoneme-to-speech system also depends on the voice parameter used, and therefore on the pronunciation modelling of the system. Indeed, a human investigation of 100 samples of *synthetic phonotactic* revealed that only 56% of the epenthesis were synthesised, and the ASR recognised an epenthesis in 21% of the cases. This reveals a weakness of the TTS, and a relative ability of the ASR to correct this pronunciation pattern.

6. Conclusion and Future Work

In this study, we leveraged TTS generation bias and ASR weaknesses to model non-native confusion. Using these confusions, we created synthetic artificially accented speech data to explore pronunciation patterns of Spanish-accented English. Deploying speech recognition systems on a global scale requires improving the robustness of ASR to different accents to ensure fairness across users of different linguistic backgrounds. In this perspective, this sort of investigation into the error patterns of ASR systems, which reveals a lot about how pronunciation variations are handled, can be very useful. Indeed, our use of synthetic data as a way of systematically analysing the impact of controlled speech variations can be of help 1) as a tool for speech research, 2) to understand ASR learning processes, and 3) for training and fine-tuning ASR.

Firstly, as speech data is time consuming and expensive to obtain and annotate, the work presented here could be of aid for linguistic studies. The comparison of phonemic confusion patterns between synthetic and natural speech suggests that synthetic speech can be used to infer L1-L2 phonemic differences, supporting our earlier findings on French [13]. This implies that synthetic data, coupled with ASR, can be used as a mean to model linguistic differences between languages, and aid language modeling and variation analysis, particularly in under-resourced scenarios. It can be envisaged, then, to train TTS L1-voices on a few unlabelled L1 data to synthesise L2 speech as in *synthetic voice*, and leverage the weaknesses of an ASR to infer L1-L2 differences.

Secondly, we show that these confusion patterns can be used to build variants that accurately challenge ASR systems. From an explainable AI perspective, our approach enables clearer attribution of ASR errors to specific variations in speech. By systematically introducing specific accent-related phonemic variations to speech samples, we can precisely study the impact of these variations on ASR performance, offering a flexible, language-independent and reproducible way to test ASR accent-robustness. This helps identify potential biases in the learning process, and provides insights into the learning processes and speech representation of ASR systems, showing how different phonetic and phonological features are processed and where the system might be failing.

And finally, from a software engineering perspective, the analysis of the system's errors can help improve the performance of ASR. In giving insight into the phonetic variations that cause ASR errors, our method can help design approaches to improve ASR robustness to such variations. Additional training material could be collected, or synthesised, to correct these specific lacks in the original training set. Thus, our variants have applications for the test, development and improvement of ASR towards accent-robust systems.

On the other hand, our investigation into a phonotactic pattern reveals a very partial capture in synthetic speech, particularly compared to phonemic variations, indicating a potential weakness of the TTS for this purpose, and a need for further study. The main limitations of our work include the choice of TTS, ASR, language, and data. The force alignment we use has also shown weaknesses. Future work will address these limitations, as well as further investigate phonotactic patterns in synthetic speech.

7. Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

8. References

[1] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 2024.

[2] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data Augmentation Improves Recognition of Foreign Accented Speech," in *Proc. Interspeech 2018*, 2018, pp. 2409–2413.

[3] M. T. Turan, E. Vincent, and D. Jouvet, "Achieving Multi-Accent ASR via Unsupervised Acoustic Model Adaptation," in *Proc. Interspeech 2020*, 2020, pp. 1286–1290.

[4] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning," in *Proc. Interspeech 2021*, 2021, pp. 1314–1318.

[5] O. Scharenborg, N. van der Gouw, M. Larson, and E. Marchiori, "The representation of speech in deep neural networks," in *Lecture Notes in Computer Science. MultiMedia Modeling: 25th International Conference.* Springer International Publishing, 2019, pp. 194–205.

[6] Y. Belinkov, A. Ali, and J. Glass, "Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 81–85.

[7] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, 2022, pp. 83–91.

[8] E. O'Neill and J. Carson-Berndsen, "Investigating the sensitivity of automatic speech recognition systems to phonetic variation in l2 englishes," *University of Pennsylvania Working Papers in Linguistics*, vol. 29.2, pp. 109–118, 2023.

[9] D. van Esch, B. Foley, and N. San, "Future directions in technological support for language documentation," in *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, vol. 1. Association for Computational Linguistics, 2019, pp. 14–22.

[10] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[11] E. Casanova, C. Shulby, A. Korolev, A. C. Junior, A. da Silva Soares, S. Aluísio, and M. A. Ponti, "ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion," in *Proc. Interspeech 2023*, 2023, pp. 1244–1248.

[12] D. H. Xian Yuen, A. Yong Chen Pang, Z. Yang, C. Y. Chong, M. Kuan Lim, and D. Lo, "Asdf: A differential testing framework for automatic speech recognition systems," in *2023 IEEE Conference on Software Testing, Verification and Validation (ICST)*, 2023, pp. 461–463.

[13] M. Masson and J. Carson-Berndsen, "Investigating phoneme similarity with artificially accented speech," in *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, 2023, pp. 49–57.

[14] M. K. Olsen, "The l2 acquisition of spanish rhotics by l1 english speakers: The effect of l1 articulatory routines and phonetic context for allophonic variation," *Hispania*, vol. 95.1, pp. 65–82, 2012.

[15] P. Trofimovich and W. Baker, "Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech," *Studies in Second Language Acquisition*, vol. 28.1, pp. 1—30, 2006.

[16] M. Swan and B. Smith, *Learner English: A Teacher's Guide to Interference and Other Problems*, 2nd ed., ser. Cambridge Handbooks for Language Teachers. Cambridge University Press, 2001.

[17] S. G. Martinez, "The syllable structure: understanding Spanish speakers pronunciation of English as a L2," *RaeL Revista Electronica de Linguistica Aplicada*, vol. 10, pp. 1–7, 2010.

[18] M. d. l. A. Gomez Gonzalez and T. Sanchez Roura, *English Pronunciation for Speakers of Spanish: From Theory to Practice*. De Gruyter, Inc., 2016.

[19] J. C. Silva Valencia, "A Comparative Linguistic Analysis of English and Spanish Phonological System," *GIST - Education and Learning Research Journal*, vol. 25, pp. 140—-156, 2022.

[20] P. Carr, *English Phonetics and Phonology: An Introduction (2nd edition)*. John Wiley & Sons, 2012.

[21] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in *Proc. Interspeech 2018*, 2018, pp. 2783–2787.

[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, p. 12449–12460.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[24] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," 2021.

[25] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[26] J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58.301, pp. 236–244, 1963.