Investigating Phoneme Similarity with Artificially Accented Speech

Margot Masson, Julie Carson-Berndsen
SFI Centre for Research Training in Digitally-Enhanced Reality (d-real)
School of Computer Science, University College Dublin, Ireland
margot.masson@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

While the deep learning revolution has led to significant performance improvements in speech recognition, accented speech remains a challenge. Current approaches to this challenge typically do not seek to understand and provide explanations for the variations of accented speech, whether they stem from native regional variation or non-native error patterns. This paper seeks to address non-native speaker variations from both a knowledge-based and a data-driven perspective. We propose to approximate non-native accented-speech pronunciation patterns by the means of two approaches: based on phonetic and phonological knowledge on the one hand and inferred from a text-to-speech system on the other. Artificial speech is then generated with a range of variants which have been captured in confusion matrices representing phoneme similarities. We then show that non-native accent confusions actually propagate to the transcription from the ASR, thus suggesting that the inference of accent specific phoneme confusions is achievable from artificial speech.

1 Introduction

Automatic speech recognition (ASR) systems, while achieving high levels of performance on US-accented English, still struggle to handle accents for which they have not been trained (Hinsvark et al., 2021). Thus, accent robustness is an important challenge for the field of speech recognition, especially since such systems have become widespread and are used worldwide.

Various approaches have been tried to build accent-robust ASR systems. The most straightforward one, building accent-specific models, is limited because of the low availability of data for most accents which are mostly not well sourced. The lack of sourced data for training and testing makes the task of recognising accented speech extremely difficult. This lack of data is mainly due to the wide diversity of accents (native and non-native) leading to the complexity of recording enough examples for each, and the difficulty of accurately labelling and transcribing speech data.

Some attempts to overcome both lack of data and accent robustness have been proposed. These include multi-task training (Ghorbani and Hansen, 2018; Yang et al., 2018; Viglino et al., 2019), features adaptation (Gong et al., 2021) or adversarial training (Sun et al., 2018). However, these methods do not completely solve the problem of the lack of data, as data would still be needed for testing. Instead, generating artificial speech data seems promising, as data augmentation has been proven to be efficient for improving the recognition of accented speech (Fukuda et al., 2018), and the use of artificial data has been around for some time (Goronzy et al., 2004; Ueno et al., 2021).

This paper investigates the extent to which artificial speech data can be used to infer accent-related phoneme confusions. We do this by using an off-the-shelf speech synthesis system, in this case Microsoft Azure TTS[1], to synthesise artificially accented speech data and then using the Wav2Vec 2.0 ASR (Baevski et al., 2020), to produce a confusion matrix for this data. This matrix is then been examined and compared to other confusion matrices, in order to evaluate its relevance in representing a particular accent. In this paper, we focus on non-native accents, although the same study could have been applied to native accents.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 describes the process of generating accent related phoneme confusions for artificial accented speech. In sections 4 and 5, we compare the confusions obtained with alternative methods and discuss the extent to which text-to-speech systems can capture accent related phoneme confusions.

[1] https://microsoft/azure/text-to-speech

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 49–57
July 14, 2023 ©2023 Association for Computational Linguistics

2 Related Work

Recent approaches for automatic speech recognition use end-to-end deep neural networks,(e.g. CTC-based, transformer-based and attention-based models) and have been really successful for this task. Commercial options exhibit extremely high performance; however, none of them achieve the same performance on accented speech. Attempts to improve end-to-end ASR performance on accented speech have had mixed results, and rely mainly on the training process. Indeed, the complexity of these architectures makes the understanding of the actual learning process difficult, if not impossible, and leads to an increasing need for explainability. This challenge has been the focus of a number of studies. Scharenborg et al. (2019) highlight the link between linguistic representations of speech and deep learning representation clusters. English et al. (2022) look to investigate in more detail the utility of attention layers, which is used in recent ASR systems. In the test community, Asyrofi et al. (2021) have proposed a testing framework for ASR systems. The work presented in this paper aligns with the goals of these approaches.

Accents are defined as variation in phoneme realisation due to several factors such as geographical location. In the case of non-native accents, which is the focus of this paper, the differences in pronunciation compared to the native language (L1) come mainly from the differences that exist between the phonetic rules of the native language and those of the target language (L2) (Flege, 1995). Thus, many pronunciation difficulties are due to phonological transfer - which involves applying L1 rules to L2 pronunciation - are linked to the non-existence of certain L2 characteristics in the L1, and result from discrepancies between the phonetic systems of the two languages. These challenges may include difficulties in producing and perceiving specific segmentals (Olsen, 2012) - like phonemes, consonant clusters, vowels - or suprasegmentals (Trofimovich and Baker, 2006) - like stress patterns, rhythm and intonation patterns - that are present in the L2 but absent or different in the L1.

Thus, non-native speakers commonly tend to approximate the pronunciation of phonemes which do not exist in their native language, by known ones they perceive as similar, as showed by Stefanich and Cabrelli (2021). For instance, pronouncing the English phoneme [ð] - corresponding to the grapheme sequence “th” as in “those” - as the French phonemes [z] or [d] is common amongst French people when speaking English (Capliez, 2011), since [ð] is not a phoneme of French (International Phonetic Association, 1999). While this is a very simplified version of the concept of accent, which does not include phenomena such as prosodic or phonotactic constraints, we focus in this paper on that definition of an accent, i.e. as the replacement of L2-but-not-L1 phonemes by L1 phonemes. This *paradigmatic* definition is intended to evolve into a more complete definition to include the *syntagmatic* and *suprasegmental* aspects in future work.

In order to understand the way in which non-native speakers switch from a phoneme of the target language (L2) to another phoneme of their native language (L1), we need to characterise phonemes and define what similarity between phonemes means. Several phonetic-based feature systems have been proposed to describe the specific phonemes of a language. Chomsky and Halle (1968) proposed a system to analyse the phonological structure of a language from a generative perspective. They described phonemes through binary features, organised along major features (that distinguish vowels from consonants), place of articulation, manner of articulation and source features (like voicing). Since then, multiple phonological feature sets have been proposed and have been used to capture similarities between phoneme classes.

This description of phonemes with features allow us to calculate their similarity using distance metrics such as Jaccard index (as defined in Equation 1, the Jaccard index between two sets U and V), that can easily be used as a similarity measure between phonemes, assuming that they are represented by their binary features. However, while it is a simple similarity to implement as baseline for the work presented in this paper, this measure is not satisfactory in the sense that all features have the same weight and, therefore, it does not take into account the difference in distance between the phonetic realisation of two features. Furthermore, it is only an a priori knowledge-based similarity, that does not necessarily follow the real-world realisations of phonemes.

$$Jaccard(U,V) = \frac{|U \cap V|}{|U \cup V|} \qquad (1)$$

While Bailey and Hahn (2005) argued that knowledge-based feature based measures are better at predicting similarity, data driven techniques offer new opportunities to identify confusions and similarities. As an example of a data-driven approach, Kane and Carson-Berndsen (2016) built a confusion matrix over the TIMIT (Garofolo et al., 1992) dataset, which contains recordings of 8 major US-English dialects. They created what they call an enhanced confusion matrix, by excluding an acoustic model iteratively, in order to restrict the recognition process and identify what phonemes are recognised in place of those that are missing from the model. This process ends up with a lot more confusions for each phoneme, thus retrieving more similarities. They found that this confusion matrix corresponds better to theoretical expectations. Furthermore, phoneme embeddings have been used as the basis of data-driven similarity, in the context of sound analogies (Silfverberg et al., 2018), for determining allophonic relationships (Kolachina and Magyar, 2019) and for capturing distributional properties (O'Neill and Carson-Berndsen, 2019).

3 Introduction of Non-Native Variations

The overall method presented in this paper for synthesising accented speech consists, broadly, of 1) transforming texts into phoneme sequences, 2) applying variations to the phoneme sequence according the target accent, and 3) synthesising speech from the phoneme sequence using a text-to-speech (TTS) engine. This workflow is illustrated in Figure 1 and is referred in the remainder of the paper as "variation method". The core of this accented speech synthesis lies in the way we choose and apply variations to the phoneme sequence. This is done by 1) selecting the phonemes to vary using a mapping between the phonemes of the different languages - this mapping is called the *phonetic compatibility matrix*, and 2) varying the selected phonemes by replacing them with their nearest neighbour phonemes in terms of similarity. This mimics the way non-native speakers adjust to the target language pronunciation. These replacements could be regarded as *mispronunciations*.

The construction of the *phonetic compatibility matrix* is very straightforward. It is built as a boolean matrix, associating the different languages with their phonemes, the values being 1 if the phoneme exists in the target language, and 0 otherwise. Table 1 shows a sample of a compatibility matrix. For example, it shows that French and Spanish speakers are likely to approximate the [ð] phoneme, while English speakers will probably approximate the [ʁ] phoneme when speaking French. This matrix is based on the IPA handbook (International Phonetic Association, 1999) charts for the different languages.

Phone	English	French	Spanish
d	1	1	1
ð	1	0	0
θ	1	0	1
z	1	1	0
s	1	1	1
t	1	1	1
ʁ	0	1	0

Table 1: Section of the compatibility matrix

When applied, the variation method replaces the incompatible phonemes (i.e. the English phonemes identified in the *phonetic compatibility matrix* as not existing in the target language) with their nearest neighbour (that is with the higher similarity, or smallest distance to the original phoneme) in the *similarity matrix*, amongst the phonemes that exist both in English and in the target language. As we saw in the related work, the similarity between phonemes can be defined in several ways. In this paper, we will briefly introduce three different methods we used for building the *similarity matrix*, with a focus on similarity identification using artificial accented speech data. Thus, the next two subsections explore these methods for defining a *similarity matrix*, which can be separated into two paradigms: knowledge-based and data-driven.

3.1 Knowledge-Based Similarity

As outlined in Section 2, features have been used for describing phonemes and for calculating similarity between them. Thus, a similarity matrix can be constructed based on the Jaccard distance between the phonemes. This method for building a similarity matrix and using it for generating artificially accented speech is referred to as method **KB1** in the remainder of the paper.

However, Jaccard-based similarity does not take into account the difficulty of switching from one articulatory position and manner to another. For instance, switching from [p] to [q] is more counter intuitive than switching from [p] to [m] while they are equally similar along the Jaccard distance (equal to 0.5). Thus, for weighting the features along their physical distance in the mouth, we have positioned

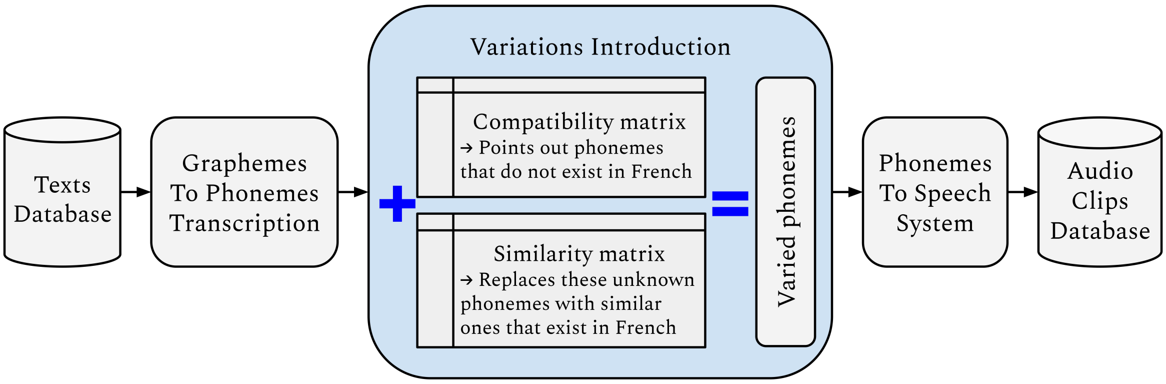
Texts Database
Graphemes To Phonemes Transcription
Variations Introduction
Compatibility matrix
→ Points out phonemes that do not exist in French
+
Similarity matrix
→ Replaces these unknown phonemes with similar ones that exist in French
=
Varied phonemes
Phonemes To Speech System
Audio Clips Database

Figure 1: Overview of the generation of speech with non-native variations.

the phonemes in a three dimensional space (Figure 2), representing the features positioned along three axes corresponding to the *place of articulation*, the *manner of articulation* and the *voicing*; this is used as a measure of phonetic neighbourhood.

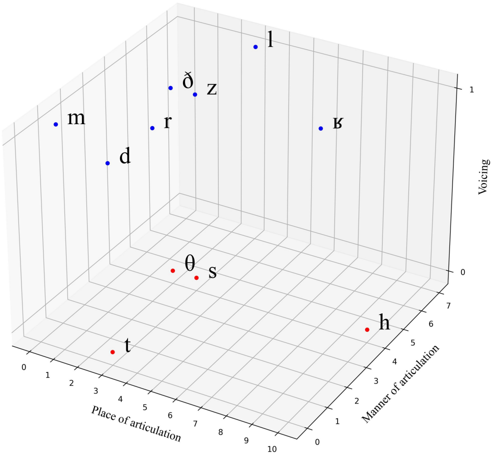

Figure 2: 3D representation of some phonemes

The coordinates of the phonemes in this space (depicted in Table 2) are used to calculate the Euclidean distance between the phonemes, as a similarity measure. For instance, in this space, the coordinates of [ð] are (3,5,1) and the coordinates of [z] are (4,2,1), which results in a Euclidean distance of 3.16 in a space where greatest distance is 13, resulting in a normalised distance of 0.24 (0.76 in similarity). This construction highlights the positional similarity of phonemes. For instance, in this space the distance between [p] and [q] (0.69) is now bigger than that between [p] and [m] (0.11). These distances are stored in the similarity matrix corresponding to that method. This 3-dimensional representation, in addition to its use for building the corresponding similarity and generating artificial accented speech, will now be referred as **KB2**.

The two similarity matrices presented in this subsection, KB1 and KB2, are entirely knowledge-based and do not necessarily highlight other constraints such as phonotactics, pitch or tone. In this sense, the data-driven paradigm presented in the next subsection can be seen as more representative of what may happen in natural accented speech.

3.2 Data-Driven Similarity

One method that has been used previously for synthesizing artificial accented speech is to rely exclusively on deep learning architectures of TTS systems to generate accented speech. This method consists of processing text inputs with a TTS engine, configured with the pronunciation patterns of the target accent. For instance, for generating a French accent in English, we input English text, to be read by the TTS engine as if it was French. We implemented this using an off-the-shelf text-to-speech system (Microsoft Azure TTS) for generating French-accented speech. This method is referred as **DD1** in the remainder of the paper and is explained in more detail in the next section.

However, the above method implies the use of a model that has been trained specifically to synthesize the target language, which brings us back to the problem of lack of data. Besides, the work conducted by Kane and Carson-Berndsen (2016) and presented in Section 2 suggests that phone confusions can be derived directly from speech data. This work motivated the development of our second data-driven method for generating accented speech. This method, denoted **DD2**, consists in running an

	Bilabial	...	Glottal
Plosive	(0,0,0) | (0,0,1)	...	(10,0,0) | (10,0,1)
...	(.., ..,0) | (.., ..,1)	...	(.., ..,0) | (.., ..,1)
Lateral Approximant	(0,7,0) | (0,7,1)	...	(10,7,0) | (10,7,1)

Table 2: Illustration of the construction of the 3D representation of phonemes

ASR system on accented speech data for retrieving the non-native confusions. These confusions can then be used for generating speech with variations as per the method described at the beginning of this section. Given the lack of natural French-accented English data, we decided to look at the recovery of phonetic confusions from artificial data. Section 4 delves into this method in more detail.

4 Artificial Speech Confusions

As introduced in the previous section, DD2 method has three stages: 1) generating artificial French-accented speech by using an off-the-shelf TTS system, 2) generating the recognition confusion matrix using an ASR system, and 3) introducing variations in speech, as per the variation method (see Figure 1), by using the previously obtained confusion matrix as the so-called *similarity matrix* for choosing the phonemes to vary.

The generation of artificial French accented speech is done by providing text inputs (i.e. textual sentences from TIMIT dataset) to the Microsoft Azure TTS, with its two parameters *language* set to English and *voice* set to one of the Azure French voices: *fr-FR-DeniseNeural* or *fr-FR-HenriNeural*. This configuration allows the TTS to synthesize the English sentences with a French pronunciation, that is reading the sentences as if they were written in French. At the end of this process, we end up with a set of artificially accented speech audios.

Then, the second step is the generation of the French confusions. For obtaining that matrix, we use Wav2Vec 2.0 ASR with a subsequent grapheme to phoneme mapping and we align the phoneme sequences with the original ones obtained from TIMIT. The confusion matrix created from these alignments is expected to capture the confusions specifically due to the target French accent.

Lastly, the confusion matrix we just created can be used as the *similarity matrix* described in Section 3 for getting the replacement phonemes for the phonemes that do not exist in French. As for KB1 and KB2, the variation method first selects the English phonemes that do not exist in French, then selects their replacements in the *similarity matrix* and finally a Phoneme-To-Text engine creates the varied speech. This aims to mimic the way French speakers approximate the pronunciation of English.

In the next sections, we evaluate the relevance of the similarity matrix described in this section - i.e. based on artificial non-native confusions - in the context of accented speech generation. This evaluation is done by comparing the results obtained by the ASR on speech generated using the variation method with the above matrix, referred to as **method DD2** in the remainder of the paper, against the other ones described in the paper.

5 Experiments

The experiments aim to evaluate the extent to which it is possible to infer accent-related phonemes confusions from artificially accented speech. For that purpose, we compare the performance of the ASR on the data generated as in Section 4, that is the speech synthesised from artificial confusions, with respect to the other methods described in Section 3, and with respect to speech without variations (artificial and natural native US English speech) as baseline. As a summary, we have the following methods:

- **NV1** is a baseline corresponding to natural US-English speech data from TIMIT.
- **NV2** is a baseline corresponding to artificial US-English speech obtained using Azure TTS.
- **KB1** corresponds to the representation of phonemes as sets of features, and their similarity as Jaccard distance.

- **KB2** corresponds to the representation of phonemes into a 3-dimensional space, and their similarity as Euclidean distance.

- **DD1** corresponds to the use of Azure TTS as a generator of accented speech, with the so-called *voice* parameter set to a French voice.

- **DD2** corresponds to the confusion matrix obtained after running an ASR on the audio files obtained by applying method DD1. This is the main focus of the paper, and has been described in Section 4.

For comparing the different methods, we use three criteria: word error rate (WER), phoneme error rate (PER) and visual inspection of hierarchical similarity clustering in dendogram representations. Global metrics, i.e. WER and PER, are used to consider the impact that variations have on recognition. The hierarchical view of similarity values of some selected phonemes provides an insight into the impact of specific variations on the recognition. That is, it is possible to see if the variation patterns propagate to the output via the confusions.

For the purpose of this paper, we built four similarity matrices, following the methodology described in sections 3 and 4. That is, we built the matrices corresponding to knowledge-based methods KB1 and KB2, as well as the similarity matrices for data driven methods DD1 and DD2. For creating these matrices, we selected 1000 sentences out of the 2366 sentences of TIMIT corpus as a text corpus. The ASR system used for conducting these experiments is Wav2Vec 2.0. The target accent is French, and the reference language is US English.

6 Results and Discussion

6.1 WER and PER

Figures 3 and 4 depict WER and PER values respectively with ASR on the six different methods. As expected, artificial speech with variations obtained higher WER scores ≈+0.57 than speech without variation, thus confirming that Wav2Vec 2.0 performs better on speech without variation. We thus obtained a drop of more than 50% between accented and non-accented speech recognition accuracy, which corresponds to the drop reported in the literature. Unsurprisingly, the PER follows the same tendencies as the WER. This indicates to an extent that the confusions we obtained are due to the difficulties for Wav2Vec2.0 in handling the *mispronunciations* we introduced in our varied speech at the phonemic level.

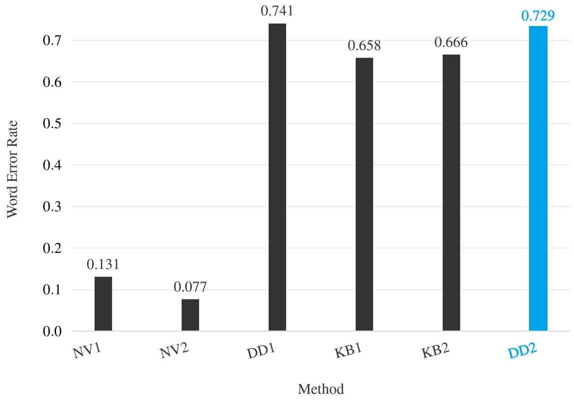

Figure 3: WER scores for DD2 vs other methods.

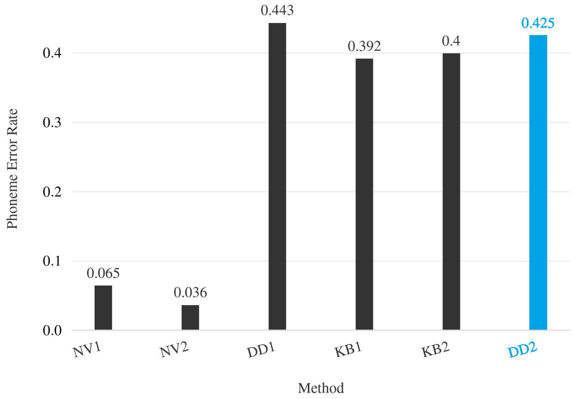

Figure 4: PER scores for DD2 vs other methods.

These results confirm the interest of our variation method for challenging ASR systems, and they are also encouraging for the identification of non-native speech learning patterns. Indeed, we can expect that the drop in accuracy between knowledge-based variation methods and data-driven variation methods is caused by the addition of new variations patterns. While the knowledge-based approaches only apply phoneme substitutions, many more other phenomena are represented by the data-driven approaches, such as phonotactics, coarticulation or prosodic transfer. The low value of the drop, however, could indicate that phoneme substitutions are the main source of errors for ASR systems, but this needs to be investigated further.

6.2 Phoneme Similarities

In order to look at the similarities which emerge from the ASR, we used hierarchical clustering of the output confusions matrices. Dendrograms visu-

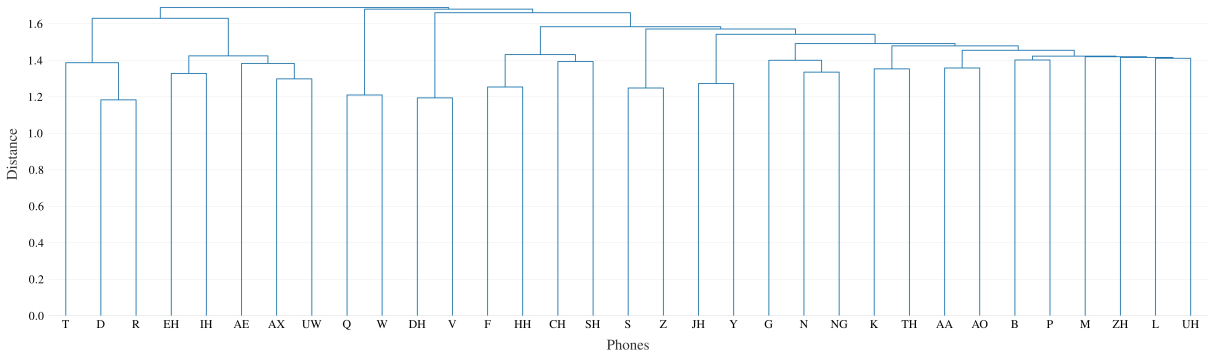

Figure 5: Hierarchical view of the confusions obtained with KB1 method.

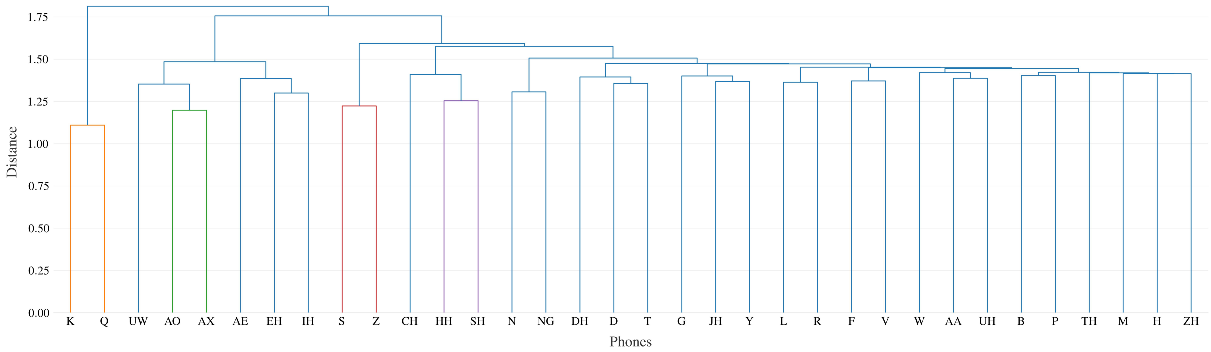

Figure 6: Hierarchical view of the confusions obtained with DD1 method.

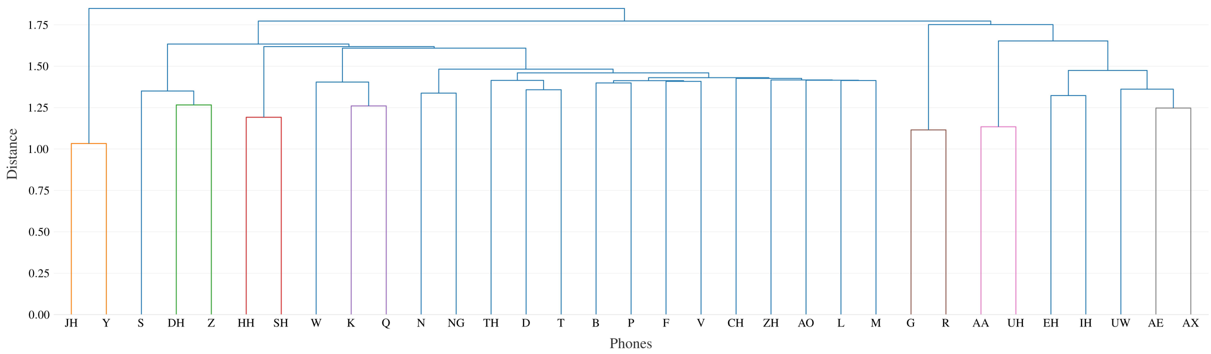

Figure 7: Hierarchical view of the confusions obtained with DD2 method.

alising this clustering can be found in Figures 5, 6 and 7 for KB1, DD1 and DD2 methods respectively [2]. The dendograms highlight some overall interesting patterns in the confusions. KB1 exhibits place-of-articulation clusters (e.g [t], [d], [r] alveolars for instance), which was expected knowing that its similarity matrix was constructed around phonetic features. However, we are looking to investigate whether the variants propagate through the ASR and provide insights into how variants cluster and emerge in a deep learning model. While the differences between the dendograms require further detailed analysis examining the contexts in which the errors occur, it can be seen, for instance, that the [ð] has moved closer to the [s] and [z] in DD2 in Figure 7, and to [d] and [t] in DD1 in Figure 6. These two confusions correspond to typical L1-French pronunciation of the *th* English grapheme. Furthermore, *r* in French is pronounced differently and it can also be seen in DD2 that [r] and [g] now cluster together; this is an indication that these sounds are both articulated further back.

This analysis of phoneme confusions highlighted that Wav2Vec2.0 was not able to correct the vari-

[2] Note that ARPABET rather than the IPA is used in these figures

ations we introduced in the input, and that these variations propagated through the ASR to the transcriptions. Indeed, confusions for KB1 and KB2 relate precisely to the variations we applied. This opens up perspectives for further analysis of the notion of similarity for ASR systems, including for artificial speech.

7 Conclusions

In this paper, we used artificially accented speech for retrieving non-native similarity patterns. We generated accented speech TTS with French voices and were able to use that output for calculating the corresponding confusion matrix. By using this matrix as a representation of similarity for introducing variations in speech, we found that these correspond to actual non-native variations. In the near future, we plan to enhance our knowledge-based methods with other types of variation, in particular phonotactic constraints. In the longer term, there are two motivations for the approach presented in this paper. The first is to investigate and model non-native speech variants as they are captured in deep learning models and the second is to provide a methodology for challenging ASR systems to determine how far a variant can be from the expected phoneme and still be recognised correctly.

Limitations

The speech recognition used was the Wav2Vec 2.0 model. Some of the errors may have been influenced by the fine tuning of the final layers; this could lead to errors being corrected by the language model. Furthermore, Wav2Vec 2.0 produces character output which we transformed to phonemes using a grapheme-to-phoneme tool; this will lead to some loss in the variation. These limitations can be overcome to some extent by using a Wav2Vec 2.0 phoneme model which we plan for our next experiments. We have only worked on French to date, even though we believe that the method is applicable to other languages. Finally, the experiments were done only on TIMIT. While this is a balanced dataset, use of other datasets will likely lead to better insights.

Ethics Statement

We have used existing speech datasets and off-the-shelf tools for speech recognition and synthesis. The use of the existing voices of the native speaker of one language, in this case French, to synthesise artificial non-native English speech is taken as representative of an L2 learner speaking English for the first time. There is much to be learned about speech variation from such artificially generated speech but is should not be regarded as mocking non-native speaker endeavours to learn a language. Indeed the variants learned from such data can provide useful insights for speaker accommodation.

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.

References

Muhammad Hilmi Asyrofi, Zhou Yang, and David Lo. 2021. Crossasr++: a modular differential testing framework for automatic speech recognition. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Todd M. Bailey and Ulrike Hahn. 2005. Phoneme similarity and confusability. *Journal of Memory and Language*, 52(3):339–362.

Marc Capliez. 2011. Typologie des erreurs de production d'anglais des francophones.

Noam. Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper and Row New York.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2022. 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*, pages 83–91, Seattle, Washington. Association for Computational Linguistics.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92:233–277.

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata. 2018. Data Augmentation Improves Recognition of Foreign Accented Speech. In *Proc. Interspeech 2018*, pages 2409–2413.

J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. 1992. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Shahram Ghorbani and John H.L. Hansen. 2018. Leveraging Native Language Information for Improved Accented Speech Recognition. In *Proc. Interspeech 2018*, pages 2449–2453.

Xun Gong, Yizhou Lu, Zhikai Zhou, and Yanmin Qian. 2021. Layer-Wise Fast Adaptation for End-to-End Multi-Accent Speech Recognition. In *Proc. Interspeech 2021*, pages 1274–1278.

Silke Goronzy, Stefan Rapp, and Ralf Kompe. 2004. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42:109–123.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jette. 2021. Accented speech recognition: A survey. *CoRR*, abs/2104.10747.

The International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Mark Kane and Julie Carson-Berndsen. 2016. Enhancing Data-Driven Phone Confusions Using Restricted Recognition. In *Proc. Interspeech 2016*, pages 3693–3697.

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Michael K. Olsen. 2012. The l2 acquisition of spanish rhotics by l1 english speakers: The effect of l1 articulatory routines and phonetic context for allophonic variation. *Hispania*, 95(1):65–82.

Emma O'Neill and Julie Carson-Berndsen. 2019. The Effect of Phoneme Distribution on Perceptual Similarity in English. In *Proc. Interspeech 2019*, pages 1941–1945.

Odette Scharenborg, Nikki van der Gouw, Martha Larson, and Elena Marchiori. 2019. The representation of speech in deep neural networks. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25*, pages 194–205. Springer.

Miikka P Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Sara Stefanich and Jennifer Cabrelli. 2021. The effects of l1 english constraints on the acquisition of the l2 spanish alveopalatal nasal. *Frontiers in Psychology*, 12:640354.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. 2018. Domain adversarial training for accented speech recognition. *CoRR*, abs/1806.02786.

Pavel Trofimovich and Wendy Baker. 2006. Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech. *Studies in Second Language Acquisition*, 28(1):1–30.

Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2021. Data augmentation for asr using tts via a discrete representation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 68–75.

Thibault Viglino, Petr Motlicek, and Milos Cernak. 2019. End-to-End Accented Speech Recognition. In *Proc. Interspeech 2019*, pages 2140–2144.

Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. 2018. Joint modeling of accents and acoustics for multi-accent speech recognition.