Identifying non-native English speech patterns in ASR systems

Margot Masson

SFI Centre for Research Training in Digitally-Enhanced Reality
University College Dublin, Dublin, Ireland
margot.masson@ucdconnect.ie

1. Introduction

Extensively used in commercial products, automatic speech recognition (ASR) systems, which process human speech into sequences of characters, are now part of many human-computer interaction technologies of popular applications. The explosion in their performance over the last decade, due to the paradigm shift from the use of hidden Markov models to the end-to-end deep learning approaches, and due to the increased availability of audio data, has brought them closer to human accuracy, making them crucial tools for today's everyday applications.

However, although they have achieved impressive levels of accuracy, these models still have limitations, particularly when it comes to handling accents [1]. Indeed, ASR systems still perform worse on accented speech, and the ever-increasing number of non-native speakers of English makes the need for accent-robust ASR even more urgent. This issue has been the subject of a lot of research [2, 3], much of which has been limited by the lack of accented speech data with which to train and test models, as well as by the lack of insight into the sources of error in ASR. Indeed, the increasingly deep architectures of ASR systems, combined with the lack of common test frameworks, makes it even more difficult to compare and understand the different models proposed.

Thus, this PhD seeks to gain insight into how current large deep-learning ASR models learn speech representations for identifying non-native English speech patterns and improving the understanding and the evaluation of ASR systems. To do this, I take inspiration from work on artificial speech synthesis [4, 5], and from explainable AI (XAI) [6, 7] techniques by designing speech variants that I use to analyse ASR systems. These variants are aimed to challenge ASR systems while imitating accent-related variations, with the goal of causing errors in the recognition process that will lead, when analysed, to a better understanding of the inner-working of ASR systems.

2. Research Objectives

The main aims of my doctoral research are as follows:

- Building accurate non-native pronunciation approximations by introducing variations in speech.
- Using these variants to challenge ASR systems and analyse their learning patterns.
- Assessing the usability of these variants for testing and improving ASR systems robustness to accented speech.

The first part focuses on defining accents and proposing methods for accurately approximating accents. The variants thus created should be challenging ASR systems - i.e. leading to recognition errors - but they should still mimic L2 speakers of English pronunciation patterns. Therefore, the main challenge for this part is to accurately define accent-related variations. As a first step, I considered that accents are mainly due to the differences between the speaker's L1 set of sounds - or phonemes - versus the English set of phonemes [8, 9]. With that hypothesis, I consider that a French accent can be approximated by replacing English phonemes that do not exist in French, such as the "th" sound - [θ] like in "thing" or [ð] like in "those" - , by similar French phonemes, such as [t], [d], [s] or [z] [10]. Note that this is a very simplified version of the concept of accent [11, 12], which is intended to evolve into a more complete definition later in the PhD.

Once the variants have been created, the goal is to use them for gaining insight into ASR models. XAI techniques, coupled with more traditional methods of transcript analysis - such as word error rate, phoneme error rate, or alignment analysis -, are hoped to help uncover the underlying mechanisms through which speech signals are processed and transcribed, and aim to help identify the specific phonemes and features that contribute to recognition errors in accented speech. The variants will be used for exploring ASR architectures for a better understanding of the model's decision-making process, and identifying areas in the neural network where accents may induce inaccuracies.

Finally, the main goal of this doctoral project is to assess the extent to which the variants designed in the first question can challenge ASR systems sufficiently - while still being coherently close to real speech - for providing a usable feedback to the developers of ASR systems. This feedback would contain information such as broad metrics, but also about the model's learned patterns highlighted in the second research question. The challenge of this last area of research lies in the evaluation of the relevance of these variants for testing ASR systems, for instance the consistency of the test scores or the coverage.

3. Methodology

3.1. Pairwise phoneme substitutions to imitate accents

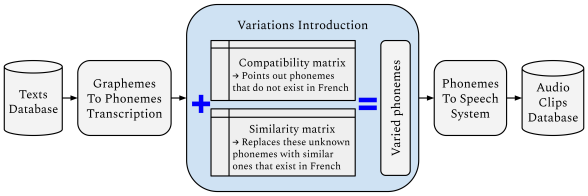

Figure 1: *Overview of the generation of speech with variations.*

Most of my research to date has focused on the generation of artificial non-native variants. The main strategy for generat-

ing variants that mimic the way L1 affects L2-English speech (presented in Figure 1) consists of 1) transforming the input text into phonemes, 2) applying variations to the phoneme sequence according to the target accent, and 3) giving the varied sequence to the TTS system that will generate the audio files. That method offers control over the phonemes that are varied (meaning that the variation should be more easily tracked through the layers and the output of the ASR systems) and does not rely on speech data. However, the definition of the similarity between L1 and L2 phonemes and the assessment of the variants are challenging.

Indeed, while the selection of the phonemes to be varied is quite straightforward - I am using a so-called *compatibility matrix* that is a mapping between the languages and their specific sets of phonemes -, the choice of the replacing phoneme depends highly on the similarity measure chosen. The similarities between the different phonemes are stored into the *similarity matrix*, whose construction is still under investigation as part of my first research question. Possible similarity measures can be ordered onto a scale from the purely knowledge based measures [13] to the purely data driven ones [14, 15, 16]. The next sections delves into more details about the similarities I have implemented.

3.2. Knowledge-based similarity measures

Two different knowledge-based similarity measures have been implemented and tested. The first and most naive one (called KB1) is based entirely on the representation of phonemes as sets of binary features. Thus, the similarity is measured based on the number of features the two phonemes have in common, divided by their total number of features.

The second one (KB2) addresses the fact that, with KB1, very distant features have the same weight. Thus, for KB2, phonemes have been positioned in a 3-dimensional space, representing the features position along three axes corresponding to the *place of articulation*, the *manner of articulation* and the *voicing* [17]. Phonemes are still represented by phonetic features, but the 3D-representation ends up in a weighting of the features along their proximity in the vocal tract, with the similarity being the Euclidean distance.

3.3. Data-driven similarity measures

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, the input English text is to be read by the TTS engine as if it was French. Thus, for method DD1, I used an off-the-shelf TTS system for generating French-accented variants.

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 [18] suggests that phone confusions can be derived directly from raw speech data. This work motivated the development of a second data-driven method (DD2) for generating accented speech. This method consists in running an ASR system on artificial accented speech data for retrieving the non-native confusions. The confusion matrix obtained is then used for generating speech with variations (as per in Figure 1).

4. First Results

For the experiments, we used the TIMIT dataset [19], which contains recordings of 8 major US-English dialects. The ASR system used for conducting these experiments is Wav2Vec 2.0 [20]. The target accent is French, and the reference language is US English. For generating artificial speech, textual sentences from TIMIT dataset were provided to Microsoft Azure TTS, with its parameter *language* set to English and its parameter *voice* set to one of the Azure US voices for generating non accented speech, and one of the Azure French voices for speech with French accent. The similarity matrices were built over 1000 textual sentences out of the 2366 sentences of TIMIT.

The word (and phoneme) error rates (WER and PER) were computed for Wav2Vec2.0 on the 4 different methods I described above, and also on artificial and natural speech without variation. As expected, varied speech obtained higher WER scores $\approx$+0.57 than unaccented speech. This confirms that Wav2Vec 2.0 performs better on speech without variation, as I obtained a drop of more than 50% between accented and non-accented speech recognition accuracy, corresponding to the drop reported in the literature. The PER follows the same tendencies as the WER, indicating that the confusions obtained are due to the *mispronunciations* introduced in the variants.

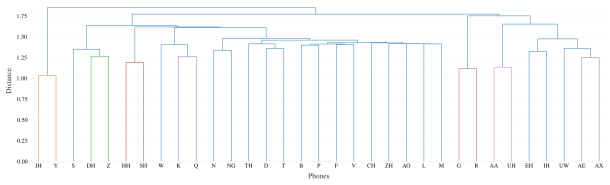

Figure 2: *Hierarchical view of Wav2Vec2 confusions for DD2.*

In order to look at the confusions which emerge from the ASR, I used hierarchical clustering of the output confusions matrices, an example of which can be found in Figure 2. These dendograms highlight some overall interesting patterns in the confusions. Knowledge-based method KB1 exhibits place-of-articulation clusters, which was expected knowing that its similarity matrix was constructed around phonetic features. For data-driven method DD2, [ð] has moved closer to the [s] and [z], corresponding 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. These dendograms still need to be more deeply analysed, particularly to put them into perspective with the variations applied during the variants production stage.

5. Future Work

As a future research plan, I will look more deeply into the confusions obtained with the variants described in this abstract, but I will also address some of the limitations of these approaches. Thus, I will look into other types of variations caused by accents, such as those caused by the differences in phonotactic (which sounds can occur together) constraints between L1 and L2. I will also work on investigating Wav2Vec2.0 layers when recognising my variants. I'm planning to spend the last year of my PhD look at if that work can be applied for assessing and improving ASR systems.

6. 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.

7. References

[1] A. Hinsvark, N. Delworth, M. D. Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin, N. Bhandari, and M. Jette, "Accented speech recognition: A survey," *CoRR*, vol. abs/2104.10747, 2021. [Online]. Available: https://arxiv.org/abs/2104.10747

[2] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," 02 2018.

[3] X. Gong, Y. Lu, Z. Zhou, and Y. Qian, "Layer-Wise Fast Adaptation for End-to-End Multi-Accent Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 1274–1278.

[4] M. H. Asyrofi, Z. Yang, and D. Lo, "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, aug 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3468264.3473124

[5] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 68–75.

[6] O. Scharenborg, N. van der Gouw, M. Larson, and E. Marchiori, "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*. Springer, 2019, pp. 194–205. [Online]. Available: http://homepage.tudelft.nl/f7h35/papers/mmm19.pdf

[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*. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 83–91. [Online]. Available: https://aclanthology.org/2022.sigmorphon-1.9

[8] 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, no. 1, pp. 65–82, 2012. [Online]. Available: http://www.jstor.org/stable/41440363

[9] S. Stefanich and J. Cabrelli, "The effects of l1 english constraints on the acquisition of the l2 spanish alveopalatal nasal," *Frontiers in Psychology*, vol. 12, p. 640354, 2021.

[10] M. Capliez, "Typologie des erreurs de production d'anglais des francophones," Oct 2011. [Online]. Available: https://journals.openedition.org/apliut/1645

[11] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[12] 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, no. 1, p. 1–30, 2006.

[13] T. M. Bailey and U. Hahn, "Phoneme similarity and confusability," *Journal of Memory and Language*, vol. 52, no. 3, pp. 339–362, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749596X0400138X

[14] M. P. Silfverberg, L. Mao, and M. Hulden, "Sound analogies with phoneme embeddings," *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pp. 136–144, 2018. [Online]. Available: https://aclanthology.org/W18-0314.pdf

[15] S. Kolachina and L. Magyar, "What do phone embeddings learn about phonology?" *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2019. [Online]. Available: https://aclanthology.org/W19-4219.pdf

[16] E. O'Neill and J. Carson-Berndsen, "The Effect of Phoneme Distribution on Perceptual Similarity in English," in *Proc. Interspeech 2019*, 2019, pp. 1941–1945. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3042

[17] T. International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[18] M. Kane and J. Carson-Berndsen, "Enhancing Data-Driven Phone Confusions Using Restricted Recognition," in *Proc. Interspeech 2016*, 2016, pp. 3693–3697.

[19] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: https://arxiv.org/pdf/2006.11477.pdf