Interspeech 2024
1-5 September 2024, Kos, Greece

PhoneViz: Exploring Alignments at a Glance

Margot Masson[1,2], *Erfan A. Shams*[1,3], *Iona Gessinger*[1,3], *Julie Carson-Berndsen*[1,2,3]

[1]School of Computer Science, University College Dublin, Ireland
[2]SFI Centre for Research Training in Digitally-Enhanced Reality, Ireland
[3]SFI Research Centre for AI-driven Digital Content Technology, Ireland

margot.masson@ucdconnect.ie, {erfan.shams, iona.gessinger, julie.berndsen}@ucd.ie

Abstract

This Show and Tell presents *PhoneViz*, a phone alignment visualiser which facilitates a deeper analysis of the phone alignments typically used to compare a reference transcription and a concrete speaker pronunciation. *PhoneViz* provides an interactive environment where aligned phones are displayed in the IPA chart helping users to explore phonetic variation beyond symbol substitution, insertion and deletion. We showcase the functionality of the visualiser using samples of Spanish-accented English, where users can see at a glance where there are phonetic similarities between substituted sounds.
Index Terms: phone alignment, visualisation, pattern analysis

1. Introduction

Alignment between a reference and a hypothesis in automatic speech recognition (ASR) is a crucial step in assessing accuracy. It is used to compare the two sequences and results in global accuracy metrics such as the word error rate. While accuracy is often computed at the word level for ASR evaluation, it can also be assessed at the phoneme level, comparing the expected phonemes (reference) with the phones recognised by the system (hypothesis). Beyond evaluating ASR performance, aligning phone sequences is also useful for pronunciation error detection [1] and correction [2].

However, while alignment is widely used in evaluation, there is a notable gap in the availability of tools specifically designed to visualise phone-level alignment [3]. We present a phone alignment visualiser, *PhoneViz*, that provides an interactive environment where users can compare a reference and hypothesis by presenting the aligned phones visually in an interactive IPA chart; this can help users gain insights into the pronunciation patterns of different speech varieties and explore phonetic variation.

For this Show and Tell, we demonstrate *PhoneViz* on the use case of accent pronunciation pattern analysis. A phone-level analysis of the expected versus the produced sounds may be of help, not only to the speaker to realise and overcome these discrepancies, but also to the researcher investigating non-native speech. Here we focus on Spanish-accented English, which is the result of the Spanish phonological system interfering with the English pronunciation. In particular, English and Spanish have different phoneme sets, resulting, for instance, in difficulties pronouncing phones such as [z], [v], [θ], and [ð] like native speakers [4]. The phonotactics of the two languages also differ [5]. For instance, Spanish speakers tend to insert a vowel before English word-initial consonant clusters starting with "s" (sC cluster), pronouncing "estreet" instead of "street". *PhoneViz* can help to further analyse such a phenomenon.

First, we describe the *PhoneViz* interface and its motivation (section 2). Then we present the demonstration setup and an example analysis (section 3).

2. Interface

The *PhoneViz* interface (see Figure 1) displays phone alignments by showing a sequence of *reference* phones (e.g., standard transcription of a target sentence) and *hypothesis* phones (e.g., phonetic transcription of a concrete realisation of this target sentence) one above the other. The interface elements are developed using *matplotlib*.[1] Users can move along the target sentence at the phone level using the *phone* slider, or go to a different target sentence using the *utterance* slider. As the users navigate through the target sentence, the phones encountered in the reference and hypothesis are highlighted in the IPA chart below by colouring the corresponding phone(s). The colours are blue for the reference and red for the hypothesis in the case that they differ, or purple (i.e., the sum of blue and red) in the case that they match. All sounds in the IPA chart are clickable and can be played back[2] by the users. The full input audio can also be played using the play button.

PhoneViz is intended to be used for both research and educational purposes. The motivation for the development of *PhoneViz* was to support the researcher in identifying more information about error patterns than can be explained by symbol substitutions, insertions and deletions. The use case outlined here is to allow the researcher further analyse the phonetic properties of mispronunciations in non-native speech by seeing, at a glance, whether they share manner of articulation (MOA), place of articulation (POA) or voicing features as defined by the IPA consonant chart, or backness, height or rounding features as defined by the IPA vowel chart. *PhoneViz* can also be used to compare different ASR models by exploring and visualising their error patterns. In the context of pronunciation training, students may use it to assess their pronunciation and focus on the specific articulations highlighted by the IPA chart which helps in flagging articulation approximations in the realisation of certain sounds.

3. Demonstration

We use the Spanish-English samples from the L2-ARCTIC [6] speech corpus. This corpus contains recordings of non-native English speakers reading English prompts from the CMU ARCTIC[3] database, including four L1-Spanish speakers of English. At the start of the demonstration pipeline, illustrated in Figure 2, the input speech recording of an L2-ARCTIC sentence

[1]https://matplotlib.org
[2]https://commons.wikimedia.org/wiki/General_phonetics
[3]http://festvox.org/cmu_arctic

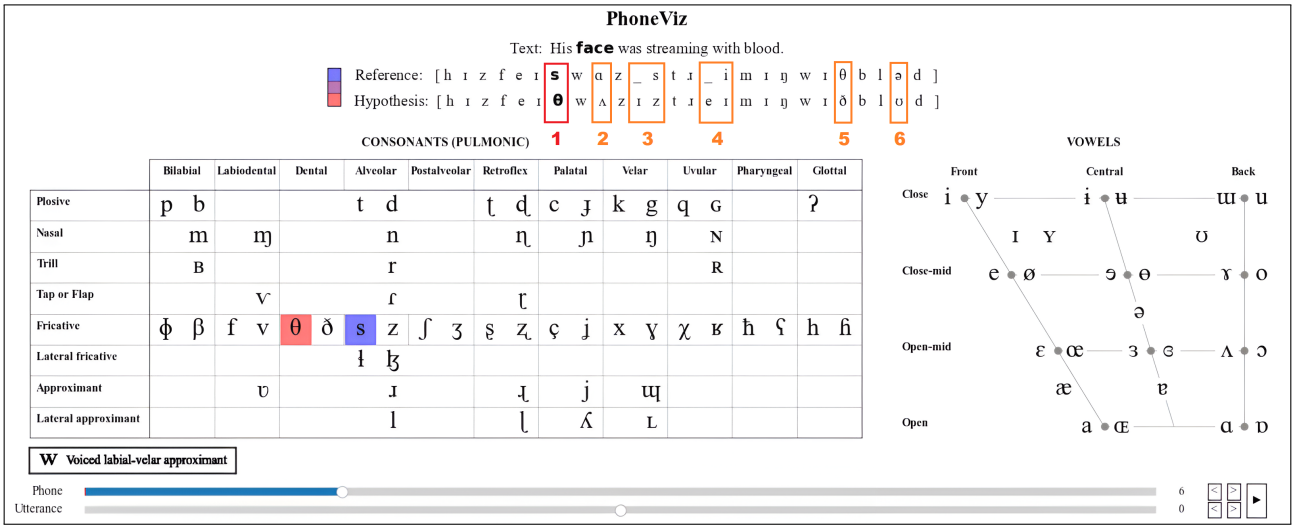

Figure 1: *Screenshot of the PhoneViz interface. The additional annotations (1–6) highlight differences between reference and hypothesis. Example 1 is currently visualised in the interface. The reference ■ is [s], while the hypothesis ■ differs in POA and indicates [θ].*

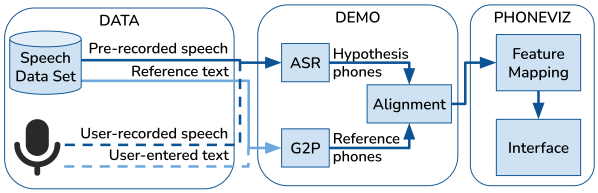

Figure 2: *Demonstration pipeline*

is processed and sent to Wav2Vec2Phoneme [7] for phone recognition (ASR module). This is the base wav2vec 2.0 [8] model fine-tuned for phone recognition, which can be found on HuggingFace.[4] The recognised (hypothesis) phones are then aligned with the reference phones, obtained by transcribing the text prompts into IPA phones using the grapheme-to-phoneme (G2P) Python library *eng_to_ipa*.[5] The alignment is performed using *sclite*, from the NIST Scoring Toolkit (SCTK).[6]

The alignments are then provided to the visualiser itself, which maps the articulatory features (consonants: MOA, POA, voicing; vowels: backness, height, rounding) to the IPA chart. Figure 1 depicts *PhoneViz* output for the target sentence "His face was streaming with blood." from L2-ARCTIC. The *PhoneViz* interface highlights a deviation in POA that resulted in the speaker pronouncing [θ] in place of [s] – MOA and voicing are the same as in the reference (example 1). Examples 3 and 5 show further deviations in the group of dental-alveolar fricatives, here in terms of voicing: [s] → [z] and [θ] → [ð]. Example 3 also sheds light on a vowel epenthesis before an sC cluster [stɪ] → [ɪztɪ], typical for Spanish-accented English. This epenthesis may have lead to the voiced production of the following fricative. Examples 2 and 6 illustrate the difficulty for Spanish native speakers to distinguish vowels outside of the Spanish vowel set, with [ɑ] → [ʌ] and [ə] → [ʊ]. Example 4 shows an English monophthong produced as a diphthong [i] → [eɪ], which may have been induced by the orthography.

The main limitation of *PhoneViz* is its dependence on the performance of the G2P, ASR, and alignment components. However, *PhoneViz* itself is useful for quickly detecting errors in these components. Currently, *PhoneViz* can visualise any pre-recorded speech (e.g., recordings made by users with the respective reference text). A function enabling the comparison of several speakers (hypotheses) side by side is work in progress.

[4]https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft
[5]https://pypi.org/project/eng-to-ipa
[6]https://github.com/usnistgov/SCTK/tree/master/src/sclite

4. Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland (SFI) Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and the ADAPT SFI Research Centre under Grant Agreement No. 13/RC/2106_P2 at University College Dublin.

5. References

[1] J. T. Ratnanather, L. C. Wang, S.-H. Bae, E. R. O'Neill, E. Sagi, and D. J. Tward, "Visualization of speech perception analysis via phoneme alignment: A pilot study," *Frontiers in Neurology*, vol. 12, 2022.

[2] E. O'Neill, R. Young, E. Thiaville, M. MacCarthy, J. Carson-Berndsen, and A. Ventresque, "S-capade: Spelling correction aimed at particularly deviant errors," in *Statistical Language and Speech Processing*, 2020, pp. 85–96.

[3] T. Yousef and S. Janicke, "A survey of text alignment visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1149–1159, 2021.

[4] O. L. Uribe Enciso, S. S. Fuentes Hernandez, K. L. Vargas Pita, and A. S. Rey Pabon, "Problematic Phonemes for Spanish-speakers' Learners of English," *GIST – Education and Learning Research Journal*, vol. 19, p. 215–238, 2019.

[5] S. G. Martinez, "The syllable structure: understanding Spanish speakers pronunciation of English as a L2," *RAEL: revista electrónica de lingüística aplicada*, no. 10, pp. 1–7, 2011.

[6] 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 *Interspeech*, 2018, pp. 2783–2787.

[7] Q. Xu, A. Baevski, and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," in *Interspeech*, 2022, pp. 2113–2117.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.