Visualising Vowel Dynamics for the Training of Text to Speech Systems

Henry An, *Jesin James*, *Catherine Watson*, *Binu Abeysinghe*

Department of Electrical, Computer and Software Engineering, University of Auckland

han285@aucklanduni.ac.nz, jesin.james@auckland.ac.nz, c.watson@auckland.ac.nz, babe269@aucklanduni.ac.nz

Abstract

Currently, the main tool used for visualising text to speech training are learning curves. Recent research examined the use of linguistic features for the visualisation of the training of speech synthesis systems by plotting the monophthongs of the model by F1 and F2 measurements at the midpoint and confirmed its applicability. This paper investigates a proposed approach for improving linguistic based visualisation for the training of text to speech systems by representing formant dynamics in synthetic speech vowels. A model was fine tuned from American English to New Zealand English and this process was used to test two visualisation designs.

Index Terms: speech synthesis, machine learning, formant trajectory

1. Introduction

The field of text to speech has advanced to a stage where there exists synthetic speech such that the difference in naturalness between synthetic and natural speech is statistically insignificant [1]. This advancement is in part driven by the adoption of neural network based synthetic speech approaches. Neural network based text to speech (TTS) are a type of machine learning based speech synthesis where a deep neural network model is trained on a dataset and then can be fed input text to produce speech [2].

As the technology of text to speech matures, its applications have become more widespread and, for example are used in assistant systems like Siri or Alexa [3] but also for purposes such as screen readers or to give pronunciation examples for language education [4]. As such, for the purposes of equity, it is important that text to speech systems are available in a variety of languages and dialects and thus important to be able to effectively train TTS models for a variety of languages and dialects.

In machine learning, a major form of training visualisation used are learning curves. Learning curves plot the progress of a metric over the course of the training process. A typical metric would be the loss function, which represents the error between the predictions made by the model compared to the actual data provided. These learning curves can be used to visualise the progress of the training, representing, for example, how well the model is learning the data. However, in the case of TTS, this is not fully sufficient. In TTS training we not only need to see how well the model is learning the information of a dataset through the visualisation of metrics such as loss, but also what the synthesised speech of the model sounds like to humans. To this end using a linguistic based visualisation can be helpful [5].

Vowels are one of the types of sound which are found in languages. They are the sounds which are formed by an unimpeded flow of air through the vocal tract and have identifying features such as formants. The formants of a vowel are the resonances of the vowel at certain frequencies, of which formant 1 (F1) and formant 2 (F2) are considered the most important for identifying the vowel [6]. The vowels of a language can be said to form a vowel space which is the space occupied by the set of vowels of said language in terms of a two dimensional F1, F2 space [7]. By having access to the vowel space, one can visually identify the pronunciation of vowels in a sample of speech. A common way of plotting a vowel space is by taking a measurement of F1 and F2 at the midpoint of the vowel duration. These points can then be plotted on a two dimensional frequency graph.

In the study by Abeysinghe et al, a model was trained from American English to New Zealand English. It was observed that the vowel space shifted from one resembling the original American English to one resembling the desired New Zealand accent [5]. If samples of speech from the American and New Zealand English models are synthesised and then the vowel spaces are plotted, one will be able to see the differences between the two models visually by comparing the vowel spaces and thus see the training progress of the model in terms of pronunciation. This study focused on monophthongs and thus relied on measurements of the vowel formants taken at the midpoint only. However, this does not form a complete picture of the vowel pronunciation as formant values can change over the duration of a vowel, even for monophthongs, and can be important for identifying the vowel [8].

This research examines how to improve this visualisation, mainly by extending the visualisation to contain dynamic aspects of the vowel. Previous work [5] established the validity of a linguistics based approach in the visualisation of TTS model training by examining the effectiveness of plotting the F1 and F2 at the midpoints of the vowels. By only taking formant measurements at the midpoint, only a snapshot of the vowel is captured and used to represent the vowel. This can be termed the *static visualisation of the vowels*. But the formants of the vowel do not necessarily stay the same for the whole duration of the vowel and this change, termed *formant dynamics*, has been used to analyse the historic sound change of vowels, including monophthongs [9] [10]. In some cases, such as for diphthongs, this movement is an identifying feature of the vowel. Thus, in order to more fully visualise the vowels of a model, in particular said diphthongs, this paper explores two ways of plotting formant dynamics for a TTS training context. To guide the investigation into improving this visualisation, we developed the following research questions:

RQ1: How best can formant dynamics be represented for the training of TTS systems?

RQ2: How useful is formant dynamic information to the person training TTS models?

In the rest of this paper, we will explain our methodology and discuss the results. Finally, a conclusion will be drawn, and future work discussed.

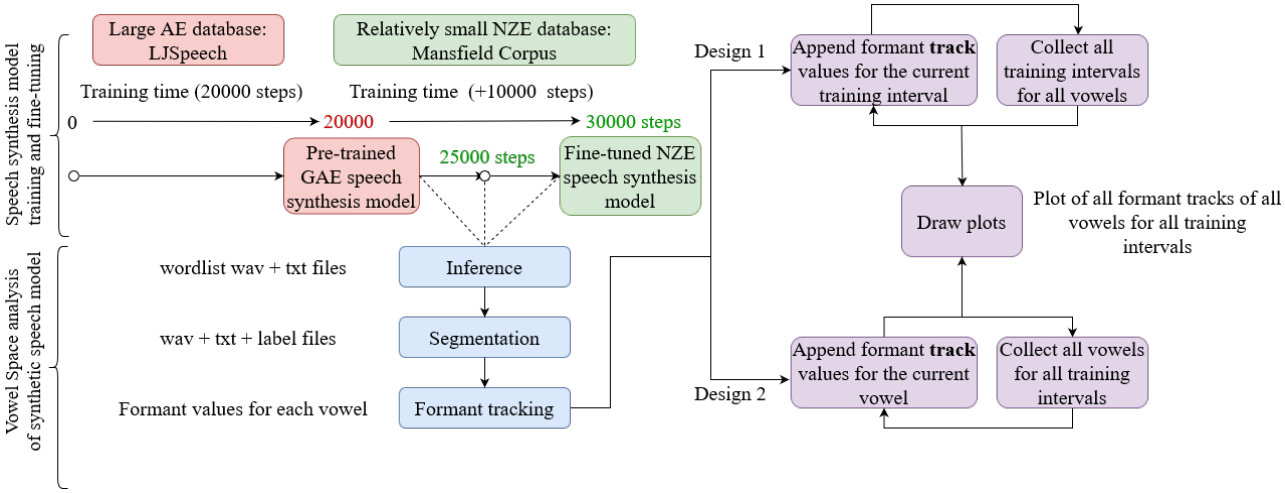


Figure 1: *Flow chart of the visualisation implementation across training. Adapted from [5] .*

Table 1. *Table of diphthongs of New Zealand English*

Lexical Set [11]	FACE	PRICE	CHOICE	MOUTH	NEAR	GOAT
Synthesised Word Used	aid	hide	lloyd	how'd	hare	oat
Truth Word Used	date	hide	voice	about	air	code
IPA Transcription	æɪ	ɑe	oɪ	æɔ	eː	əʉ

2. Methodology and Implementation

This research, based on the previous static formant visualisation, focused on modifying the system to include dynamic formant information and then investigating the merits of different visualisations. To answer our research questions, we first trained a model on American English for 20k steps before further training on New Zealand English for 10k total steps for a total of 30k. We then analysed the formants of the vowels of the models and produced two designs for visualising the formants. This process of training the model and drawing plots at select stages can be seen in Figure 1.

In linguistics, one way formant dynamics are represented are formant trajectories [12]. Formant trajectories are made by taking measurements of the F1 and F2 values of a vowel at multiple fixed points in the vowel duration before plotting the trajectories in a F1, F2 frequency graph. The designs of the systems were inspired by the usage of formant trajectory plots which are used to visualise formant dynamics in the field of linguistics such as in [12].

While the previous approach visualised all of the vowel spaces of all of the training stages in a single plot, this was not the method taken in our approach. This was because our study investigated visualising the vowel dynamics instead of the vowel spaces and as such there was more information to be represented. Thus, two designs were created which presented the vowel trajectories in different ways and compared.

The two designs were investigated to understand how best to visualise the dynamic formant information and were compared, hereafter referred to as design 1 seen in Figure 2 and design 2 seen in Figure 3. Design 1 attempted to group the vowels by phoneme and plotted each vowel phoneme as an individual plot with each curve representing the vowel of an individual model while design 2 was to plot the vowel trajectories of each training stage and juxtapose the plots next to each other, with each curve representing an individual diphthong in the vowel space of a model.

The investigation began by training a model to experiment on. The FastSpeech2 [13] system was used for this purpose. In order to train a model on FastSpeech2, as seen in figure 1, text grids with the phonetic alignments need to be generated of the speech data. This was done automatically using the Montreal forced aligner system [14]. The data used to train the model consisted of two sets. One was the LJSpeech corpus [15], a dataset of a female speaker of American English, which was used to train the base model and the second was a sample of a female speaker of New Zealand English from the Mansfield Corpus [16] which was used to fine tune.

The implementation of FastSpeech2 used was based on an implementation made using PyTorch [4] which itself was based on [17]. The proposed idea was to interrupt the training process at fixed intervals of steps, such as when a model was saved, and create a visualisation of the vowel space of the TTS model at each saved step with the formant trajectory of the vowels included. Samples of speech from the model were generated using words of a 'h' vowel 'd' format (hVd words) as the input when possible and fed into a WEBMAUS [18] forced aligner API to generate phonetic alignment text grids. Then the generated alignments were processed to extract the vowels, for which each vowel, for each training stage had the formant data from the beginning, middle, and end extracted. Then the data is assembled into a Numpy data frame [19] which is used to plot the graph.

The words used to generate the formant plots as well the ground truth, the words of the New Zealand English speech used to train the model, are listed in Table 1 in addition to their IPA transcriptions. Not all the words are hVd words, as in the case of the synthesised speech, hVd words could not be reliably generated, and in the case of the ground truth, was not available.

Figure 2: *Visualisation design type 1, where each curve represents the trajectory of a vowel in an individual model. (Splines are visualisation aids, only the 3 indicated points are real measurements)*

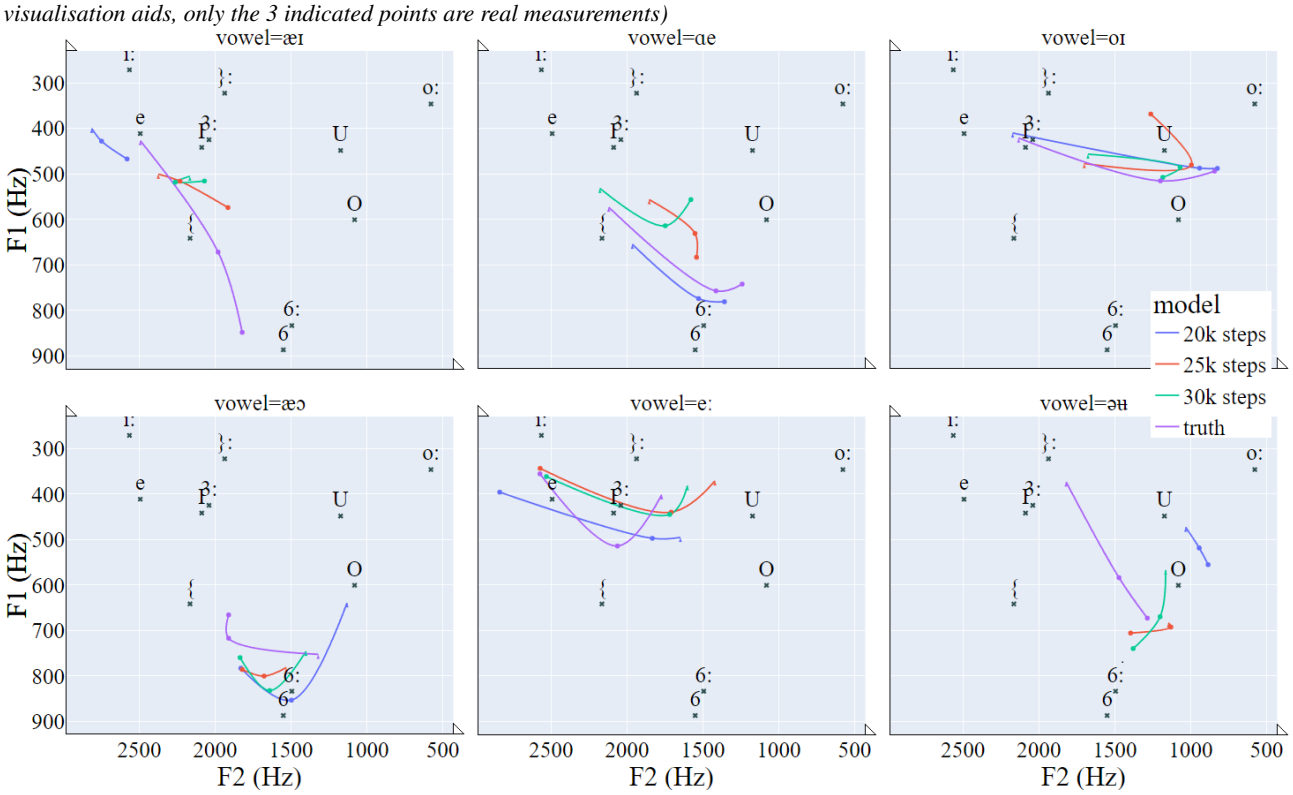
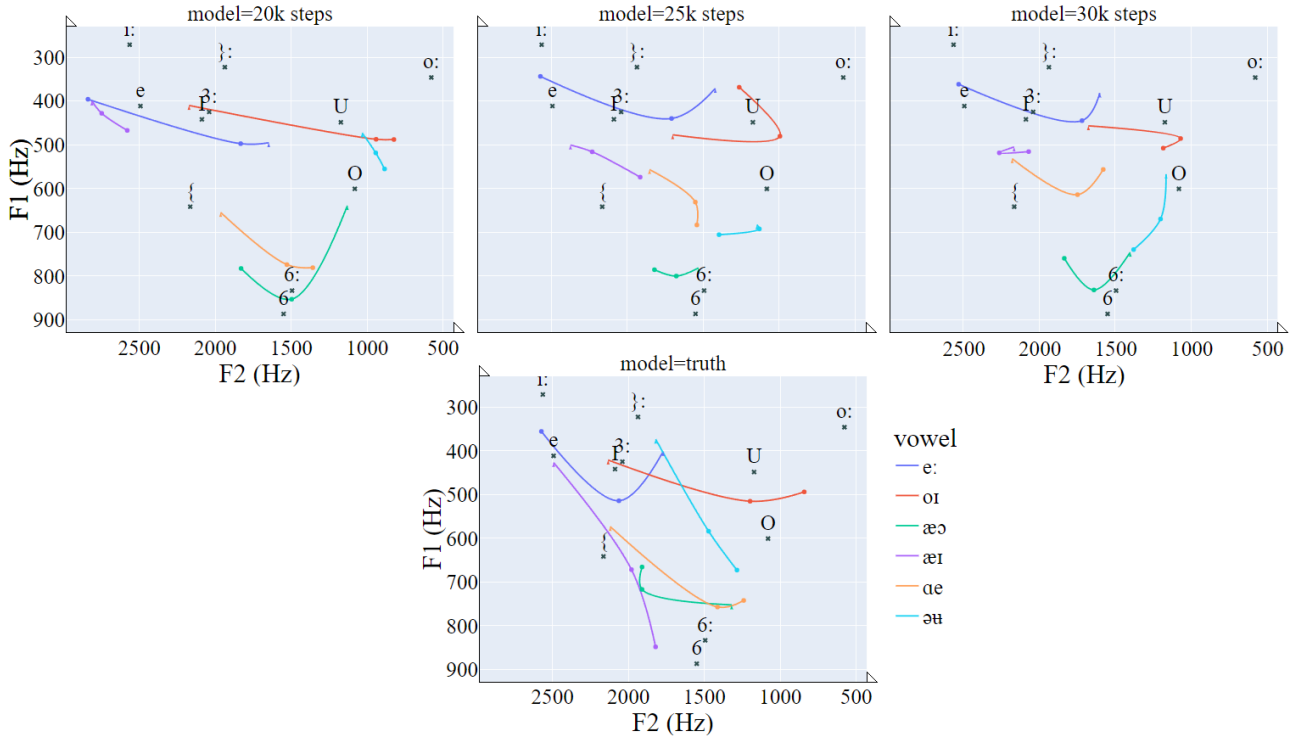


Figure 3: *Visualisation design type 2, where each curve represents the trajectory of an individual vowel in the vowel space of a particular model. (Splines are visualisation aids, only the 3 indicated points are real measurements)*



The lexical set of the word from [11] is also given.

The two designs plot the formant trajectory of each vowel in a F1, F2 space. The formants are plotted by frequency and the formant trajectories are created using frequency measurements from the start, middle and end of the vowel duration. Measurements for the start and end are taken at 20 percent from the end and beginning to mitigate the coarticulatory effects at the edges of the vowel while the measurement of the middle is taken at the exact midpoint of the vowel duration. The system was created with Python using the Parselmouth [20] library which is an implementation of Praat [21] functions within Python. Parselmouth is used to extract the formant data which is then plotted using Plotly [22].

3. Results and Discussion

Here in Figures 2 and 3 we can see a model being fine tuned from 20k steps trained American base model for 10k steps to a 30k steps total trained New Zealand English model. We can see that both designs consist of a series of F1, F2 plots of the formant trajectories of the vowels analysed for all of the training stages analysed. The two designs are differentiated by how they group the vowel trajectories in different ways. Design 1 groups the trajectories by the vowel phoneme while design 2 groups the trajectories by training stages of the model. The differences between the set of vowel trajectories of each training stage is emphasised on the training stage plots while the progression of the vowels over the training process is emphasised on the vowel phoneme graphs. In both designs, the vowel is represented by 3 points which have been connected with a line where the end representing the beginning of the vowel is unmarked and the side representing the end of the vowel is marked with an arrowhead. The lines were smoothed using a spline to give an approximation of the formant trajectory. The spline serves only an illustrative purpose and is not a prediction of the formant values.

By plotting the formants trajectories of each vowel rather than representing vowels as F1, F2 frequencies taken at their midpoint, the change in F1 and F2 over the duration of the vowel is also visualised. This can help improve the visualisation of the changes to vowel production that occur when training a TTS. An example can be seen in the vowel æɪ in 3. The midpoints of the 25k steps and 30k steps models occur near each other which would make them appear very similar in a single point representation, with a distance of only 32Hz, but when the start and end points are added, the difference between the two models becomes more clear with the star points being 165Hz apart and the end points being 208Hz apart. The greater the distance, the greater the difference in vowel quality, in effect indicating how similar two vowels sound. When comparing to the 'truth' this would indicate the closeness of the model to the desired voice.

As formant dynamics are useful for the analysis of the historic change in vowels [10], it stands to reason that this can also be applied to analysis of change in vowels for TTS training. The two designs can also be considered to constitute two different approaches to visualisation, with design 1 being an more engineering focused perspective, examining the change in vowels, while design 2 could be considered a more linguistic based perspective, examining the shape of the vowel space as a whole. A possible usage would be to examine design 2 for a holistic understanding of the change in vowel pronunciation and to refer to design 1 when desiring further information on specific vowels.

A limitation of this approach is that it relies on automated systems to generate the text grids which cannot be hand corrected. Studies [23] have shown that while forced aligners can be highly accurate and reliable, for non American speech they still generally perform at a level lower than humans, especially for highly divergent local varieties. Thus, the visualisation cannot be seen as a guaranteed completely accurate representation of the vowel space and it would be more appropriate be used to direct the researcher to the relevant locations for manual inspection.

4. Conclusion and Future Work

As TTS has become more widespread and important, so too does the training of TTS systems. Currently, the usage of linguistic based visualisation of the training of TTS systems is in its early stages with its viability being demonstrated with static formant measurements. This paper examined two proposed approaches to extend this visualisation to formant dynamics and suggests some possible uses for such a system.

While work has been done into the investigation of formant dynamic based visualisation for the training of text to speech systems, further testing is required to draw concrete conclusions. An evaluation of the system has yet to be conducted and has been planned to consist of a survey directed at the intended users of the system. Testing of the extent to which formant dynamics are important for synthetic speech could also be conducted.

In terms of future work for TTS training visualisation, further investigation of linguistic feature visualisation could be conducted such as visualisation of higher formants, fundamental frequency or prosody.

5. References

[1] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., and others,, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Zen, H., Senior, A., and Schuster, M., "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*, pp. 7962–7966, IEEE, 2013.

[3] Hoy, M. B., "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[4] Pine, A., Wells, D., Brinklow, N., Littell, P., and Richmond, K., "Requirements and motivations of low-resource speech synthesis for language revitalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7346–7359, Association for Computational Linguistics, 2022.

[5] Abeysinghe, B. N., James, J., Watson, C., and Marattukalam, F., "Visualising model training via vowel space for text-to-speech systems," in *Proc. Interspeech 2022*, pp. 511–515, 2022.

[6] Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J., "An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns," *Word*, vol. 8, no. 3, pp. 195–210, 1952.

[7] Klein, W., Plomp, R., Pols, L. C., and others,, "Vowel spectra, vowel spaces, and vowel identification," *Journal of the Acoustical Society of America*, vol. 48, no. 4, pp. 999–1009, 1970.

[8] Watson, C. I. and Harrington, J., "Acoustic evidence for dynamic formant trajectories in australian english vowels," *The Journal of the acoustical society of America*, vol. 106, no. 1, pp. 458–468, 1999.

[9] Winn, M. B. and Wright, R. A., "Reconsidering commonly used stimuli in speech perception experiments," *The Journal of the Acoustical Society of America*, vol. 152, no. 3, pp. 1394–1403, 2022.

[10] Cox, F., Penney, J., and Palethorpe, S., "Australian english monophthong change across 50 years: Static versus dynamic measures," *Languages*, vol. 9, no. 3, p. 99, 2024.

[11] Wells, J. C., *Accents of English: Volume 1*, vol. 1. Cambridge University Press, 1982.

[12] Renwick, M. E. and Stanley, J. A., "Modeling dynamic trajectories of front vowels in the american south," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 579–595, 2020.

[13] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y., "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[14] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M., "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, vol. 2017, pp. 498–502, 2017.

[15] Ito, K. and Johnson, L., "The lj speech dataset." `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[16] Watson, C. I. and Marchi, A., "Resources created for building new zealand english voices," in *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, pp. 92–95, 2014.

[17] Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., and Lee, H.-y., "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8588–8592, 2021.

[18] Kisler, T., Reichel, U., and Schiel, F., "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[19] Oliphant, T. E. and others,, *Guide to numpy*, vol. 1. Trelgol Publishing USA, 2006.

[20] Jadoul, Y., Thompson, B., and de Boer, B., "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] Boersma, P. and Van Heuven, V., "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[22] Inc., P. T., "Collaborative data science," 2015.

[23] MacKenzie, L. and Turton, D., "Assessing the accuracy of existing forced alignment software on varieties of british english," *Linguistics Vanguard*, vol. 6, no. s1, p. 20180061, 2020.