TTSVowelViz: A Tool for Visualising Text-to-Speech Model Training via Vowel Spaces

Pasindu Udawatta [1,*], **Jesin James** [2,*], **B. T. Balamurali** [†], **Catherine Watson** [3,*], **Ake Nicholas** [4,*], **Binu Abeysinghe**[*]
[*]University of Auckland, New Zealand
[†]Singapore University of Technology and Design, Singapore
[1]puda360@aucklanduni.ac.nz, {[2]jesin.james, [3]c.watson, [4]ake.nicholas}@auckland.ac.nz

Abstract

In text-to-speech (TTS) model training, the saturation of the loss curve indicates how well a model learns the characteristics of the training dataset. But it does not reveal the linguistic properties learned by the model. Existing TTS approaches miss the potential to incorporate linguistic insights into model training. We introduce *TTSVowelViz*, a novel tool that visualises static and dynamic vowel spaces during model training, bridging linguistic knowledge and TTS model development. It helps identify which vowel sounds are accurately learned and how the vowel spaces are evolved during training. To assess *TTSVowelViz*, we fine-tuned a TTS model from General American English to New Zealand English and conducted a perception test. Our results show that the formants of specific vowels in the vowel spaces generated by *TTSVowelViz* align with human perception, effectively visualising the perceived accent shift. This work highlights vowel space visualisation as a valuable interpretability tool for TTS training.

Keywords: text-to-speech, evaluation, vowel space, visualisation tool, linguistics, interpretability

1. Introduction

Deep neural network (DNN)-based text-to-speech (TTS) systems produce high-quality synthetic speech (Kumar et al., 2023), and are evaluated using subjective and objective methods *after* the TTS (speech) model is trained (Loizou, 2011). If evaluations reveal issues with the synthetic voice, the model is typically retrained with more diverse data, which is resource-intensive. During speech model training, the saturation of the loss curve to the minimum error helps determine when to stop training, as it indicates how well the model captures training data characteristics. However, the loss curve does not pinpoint specific sounds that the model struggles with (Abeysinghe et al., 2022). To overcome this, it is crucial to identify which sounds are accurately learned ***during*** training and which sounds are not.

Current TTS training approaches overlook the extensive linguistic knowledge available for many languages, even those with limited speech data for speech technology development. These methods are data-driven (Gladston and Pradeep, 2023), missing the potential opportunity to incorporate linguistic insights into model training.

Vowels constitute a significant portion of the voiced sections in speech and are distinguished by changes in first (F1) and second (F2) formants (Hillenbrand et al., 1995). Variations in vowel pronunciations are the primary differences among English accents (Wells, 1982a, p. 181). The vowel space maps vowel sounds onto a 2D space bound by F1 and F2 (Sandoval et al., 2013). It is widely used to study language characteristics (Sandoval et al., 2013; Jacewicz et al., 2007; Jongman et al., 1989; Carey, 2004) across regions (Watson et al., 2016) and time (Watson et al., 2018), and to provide language learners with visual feedback for accurate vowel production (Watson et al., 2017; Chao et al., 2020; Brett, 2004).

Recently, vowel spaces have been explored as a potential post-training measure for evaluating the naturalness of synthetic speech (Albrecht et al., 2022). Abeysinghe et al. (2022) visualised the evolution of monophthongs in the vowel space during speech model training, and An et al. (2024) investigated how diphthongs evolve during training. However, both studies (Abeysinghe et al., 2022; An et al., 2024) were conducted as proof-of-concept experiments in which vowel spaces were generated after training using model checkpoints saved at selected training steps. Despite this limitation, their findings suggest that observing vowel spaces during model training can help evaluate how well a model learns vowel sounds at each training step. *Automating vowel space generation at training steps would allow such evaluations to be performed* ***during*** *training*.

We introduce *TTSVowelViz*, a novel tool to visualise static and dynamic vowel spaces *during* speech model training, enabling real-time visualisation of model learning at each training step. Static vowel space represents a vowel sound using a single (F1, F2) pair (GUI block in Figure 1), averaged across all occurrences of a vowel in a dataset. Dy-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4778–4786
11–16 May 2026.

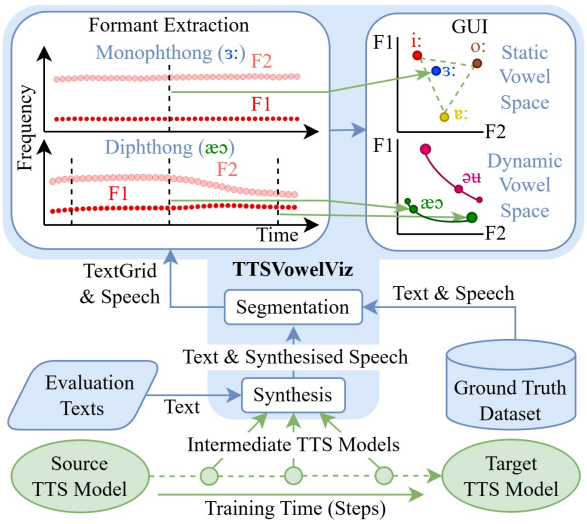


Figure 1: Workflow diagram of *TTSVowelViz*.

namic vowel space captures temporal changes in formants during vowel production. The average F1-F2 trajectory (Figure 1) is calculated per vowel category. Monophthongs (e.g., ɜː) show minimal change in F1 and F2, while diphthongs (e.g., æɔ) undergo formant transitions (Formant Extraction block in Figure 1). Therefore, monophthongs are commonly represented using the static vowel space (Peterson and Barney, 1952), whereas diphthongs are visualised within the dynamic vowel space (Renwick and Stanley, 2020). However, when temporal changes of monophthongs are of interest, they can be incorporated into the dynamic vowel space for visualisation (Watson and Harrington, 1999; Cox et al., 2024).

Our interactive GUI-based tool enables users to analyse vowel spaces and formant progressions during training, examine vowel space shapes of the learned language, and compare vowel spaces across training steps. We fine-tuned a speech model and conducted a perception test to assess whether changes in the vowel spaces generated by *TTSVowelViz* align with human perception of synthesised vowels. While not replacing perception tests, *TTSVowelViz* provides real-time assessment, helping optimise training parameters and refine datasets, making it the first tool of its kind for speech model evaluation.

In summary, the key contributions of this study are as follows:

- A fully functional, interactive, and framework-independent tool for visualising static and dynamic vowel spaces ***during*** the training of any state-of-the-art TTS model. Unlike prior work that visualised vowel spaces only after training using model checkpoints saved at selected steps, our method enables interpretable, real-time visualisation throughout model training.
- Formant extraction at any number of time points across the duration of a vowel, allowing for more granular vowel space representations, addressing the limitations of previous methods that relied on fixed time points.
- A vowel-based analysis demonstrating that the formant shifts observed in the generated vowel spaces align with human perception, supported by both visual analysis and perception test responses.

For broader accessibility, *TTSVowelViz* has been made publicly available as a Python package on the Python Package Index (PyPI) (`https://pypi.org/project/ttsvowelviz/`). An interactive demonstration is also accessible at `https://pasindu-ud.github.io/ttsvowelviz/`, allowing users to explore its features.

2. TTSVowelViz

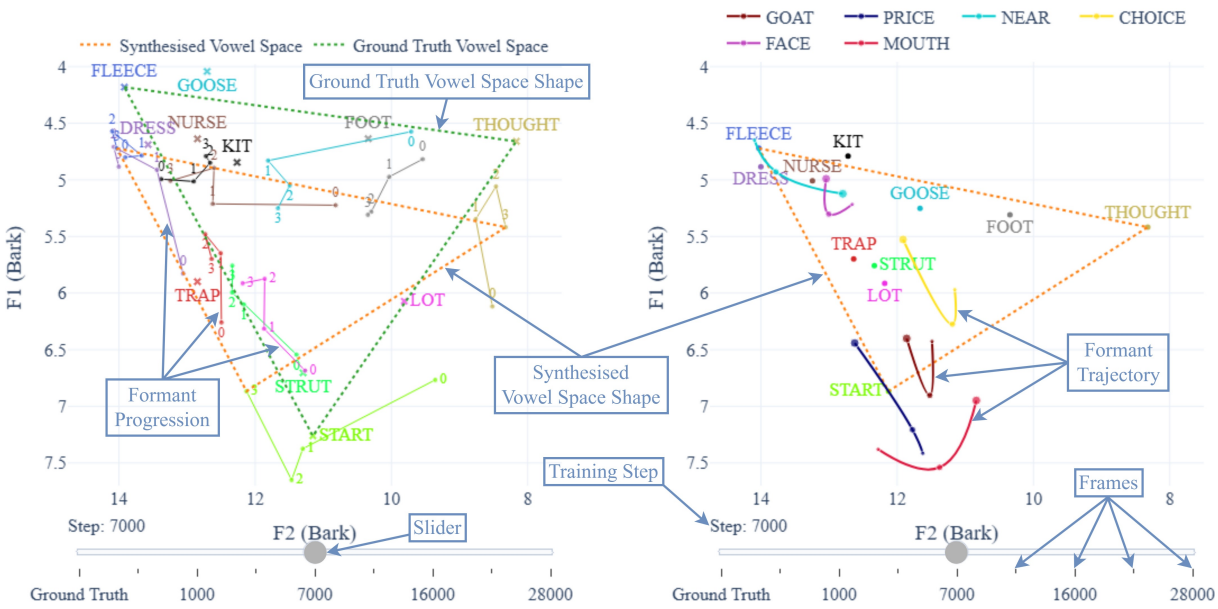
As shown in Figure 1, for any TTS model that is being trained, *TTSVowelViz* workflow is as below:

Synthesis: Speech is synthesised from evaluation texts at the training steps using intermediate speech models (i.e., models trained up to a specific training step). By requiring only the synthesis logic of the model being trained, the tool remains TTS model-independent. Integration with any model is straightforward by aligning with the technical specifications of *TTSVowelViz*.

Segmentation: Speech is segmented to identify phoneme boundaries using WebMAUS (Schiel, 1999), aligning it with text transcriptions to produce TextGrids with timestamped phonemes.

Formant Extraction: Formants are extracted using the Parselmouth (Praat) (Jadoul et al., 2018) library. For static vowel space, a single (F1, F2) pair is obtained by either taking formants at a single time point or averaging formants across multiple time points in each occurrence (a user-selectable option) and then averaging across all occurrences of a vowel. As shown in the Formant Extraction block in Figure 1, for dynamic vowel space, F1 and F2 values are captured at multiple time points, with averages computed per time point across all occurrences of a vowel. Averaging formants across multiple occurrences of a vowel mitigates variability in automatic segmentation and formant extraction (Peterson and Barney, 1952; Kiefte and Nearey, 2017; Harrington et al., 1997).

Vowel Space Plots: *TTSVowelViz* illustrates the extracted formants in an interactive GUI with static and dynamic vowel space plots (Figure 2). The static vowel space plots average (F1, F2) pair per



(a) Static vowel space. (b) Dynamic vowel space.

Figure 2: Static and dynamic vowel space plots generated by *TTSVowelViz*.

vowel category, while the dynamic vowel space traces F1-F2 trajectories over time. Both representations use the Bark scale, where distances between formants reflect perceptually equal differences, making F1 and F2 values comparable (Ross, 2024, p. 87). The plots include the following to facilitate interpretability:

- *Interactive Components:* Each plot contains multiple frames, navigable via a slider (Figure 2a). A frame represents a vowel space at a particular step. Generated using Plotly[1] library, the plots allow users to hover over points for formants, zoom in for details, and switch frames to observe vowel space evolution.
- *Ground Truth:* The first frame shows the vowel space derived from the dataset, reflecting the language being learned. The static plot connects point vowels to define the vowel space shape (Figure 2a), while the dynamic plot connects formants across time points to form F1-F2 trajectories (Renwick and Stanley, 2020) (Figure 2b).
- *Synthesised Vowel Space:* The static plot tracks formant shifts over training steps, with lines connecting points to indicate progression (Figure 2a). The dynamic plot shows evolving formant trajectories smoothed with splines (An et al., 2024) and includes (F1, F2) vowel points from the static plot as a reference (Figure 2b). The dynamic vowel space captures how adjacent sounds influence vowel production, illustrating coarticulation effects and consonant-vowel interactions (Stanley et al., 2021; Jansen, 2019).

Visualising vowel spaces at every training step can be time-consuming, so users can select a subset of steps or all if needed. Our tool allows customisation through various user inputs, including training steps, target language/accent with point vowels, vowels for static and dynamic plots, time points for formant extraction, text inputs for synthesis, synthesis logic, and ground-truth dataset.

3. Experiments

To assess *TTSVowelViz*, we fine-tuned a speech model from General American English (GAE) to New Zealand English (NZE), explored the changes in vowel spaces generated by *TTSVowelViz* during training, and validated these changes through a perception test.

3.1. Model Training and Fine-tuning

A Tacotron 2 (Shen et al., 2018) model, pre-trained for 120000 steps[2] on the LJSpeech dataset (Ito and Johnson, 2017) (13100 audio clips, ≈24 hours, GAE female), was fine-tuned on the Mansfield corpus (Watson and Marchi, 2014) (831 audio clips, ≈1 hour, NZE female) for 28000 steps. Fine-tuning was done on an NVIDIA Tesla K80 GPU with a learning rate of 0.001 for ≈36 hours and stopped at 28000 steps based on the loss curve.

[1] https://plot.ly

[2] https://github.com/TensorSpeech/TensorFlowTTS

Monophthongs			Diphthongs		
IPA	**Wells**	**Word**	**IPA**	**Wells**	**Word**
e	DRESS	head	ɔɪ	CHOICE	lloyd
iː	FLEECE	heed	æɪ	FACE	aid
ʊ	FOOT	foot	əʉ	GOAT	oat
ʉː	GOOSE	goose	æɔ	MOUTH	how'd
ɪ	KIT	hid	eː	NEAR/SQUARE	hear/hare
ɔ	LOT/CLOTH	hod	ɑɐ	PRICE	hide
ɜː	NURSE	heard			
ɐː	START/BATH	hard			
ɐ	STRUT	hud			
oː	THOUGHT	thought			
æ	TRAP	had			

Table 1: NZE vowels (Wells, 1982b) represented in International Phonetic Alphabet (IPA) symbols and Wells' keywords, with the words used in the synthesised word lists.

Parameter		Value
Training steps		0, 1000, 3000, 7000, 10000, 16000, 20000, 28000
Target accent		New Zealand English (NZE)
Point vowels		FLEECE, THOUGHT, START
Static vowel space	Vowels	NZE monophthongs (Table 1)
	Time points	50% (midpoint)
Dynamic vowel space	Vowels	NZE diphthongs (Table 1)
	Time points	20%, 50%, 80% (An et al., 2024; Watson et al., 1998)
Inference		Tacotron 2 (Shen et al., 2018)
Ground truth		Mansfield corpus (Watson and Marchi, 2014)

Table 2: *TTSVowelViz* configuration.

3.2. Vowel Space Visualisation

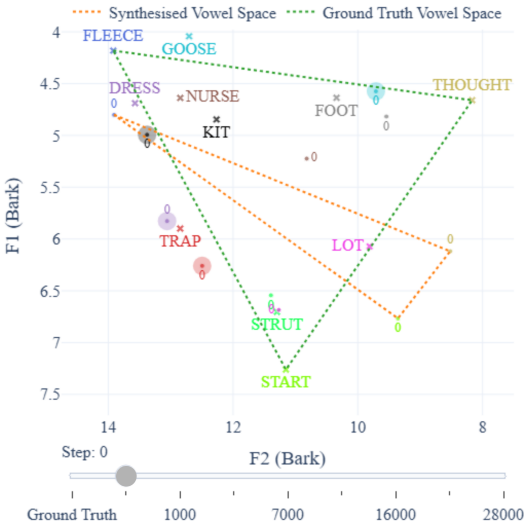
Table 1 presents the NZE vowels, with the corresponding words used in the synthesised word lists to reduce coarticulation effects (Peterson and Barney, 1952; Harrington et al., 1997). In this paper, the vowels are represented using Wells' Lexical Set (Wells, 1982a). *TTSVowelViz* was configured using the inputs listed in Table 2. Twenty-five word lists were created separately for monophthongs and diphthongs, with words listed in Table 1 in different orders to balance synthetic speech variations due to neighbouring words and sentence positioning (Abeysinghe et al., 2022).

At the selected training steps, *TTSVowelViz* synthesised speech, segmented it, extracted the formants, and generated static and dynamic vowel spaces (Figure 3). The crosses and dots in Figure 3 represent the ground-truth and synthesised vowels, respectively. The ground-truth vowel space with dark green dotted lines in Figure 3 follows the expected NZE triangular shape with three point vowels (Ross, 2024, p. 5). In Figures 3a, 3b, and 3c, each colour represents a different NZE monophthong. Synthesised vowels at each selected step are labelled 0 to 7. Figure 3a shows the static vowel space before fine-tuning, reflecting the synthetic voice of the pre-trained GAE model. Figures 3b and 3c illustrate vowel formant progressions (as explained in Section 2) with solid lines. For instance, in Figure 3b, lines connecting points labelled 0 and 1 of the same colour indicate static vowel space changes from 0 to 1000 steps. Figures 3d, 3e, and 3f illustrate the F1-F2 trajectories of diphthongs, with marker sizes increasing from the first time point to the last. They also include the monophthongs at the corresponding steps.

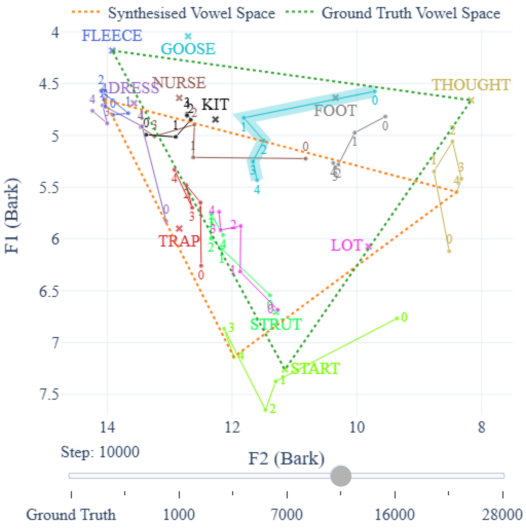
3.3. Perception Test

We conducted a perception test[3] to assess whether formant changes for each vowel in the static vowel spaces produced by *TTSVowelViz* align with human perception. The test targeted nine NZE monophthongs, including those differing most between GAE and NZE (DRESS, FLEECE, GOOSE, KIT, LOT, NURSE, START, and TRAP (Abeysinghe et al., 2022; Ross et al., 2023)) and NZE point vowels (FLEECE, START, and THOUGHT). Each vowel appeared in the carrier sentence, "Say the word ... again," with the ellipsis replaced by a **word** con-

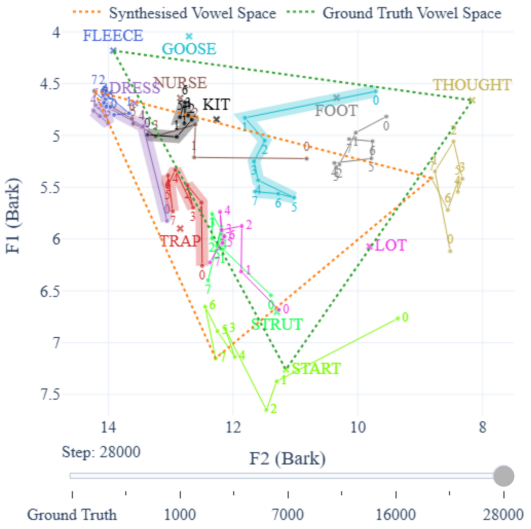
[3]Approved by the University of Auckland Human Participants Ethics Committee (Ref. UAHPEC22681) on 01/10/2021 for 3 years.



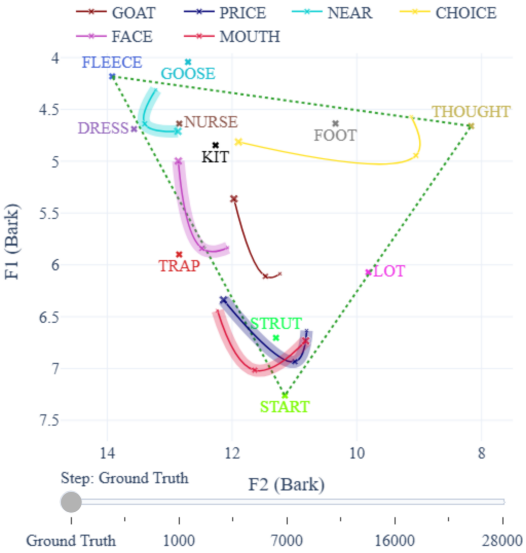
(a) Static vowel space at step 0.



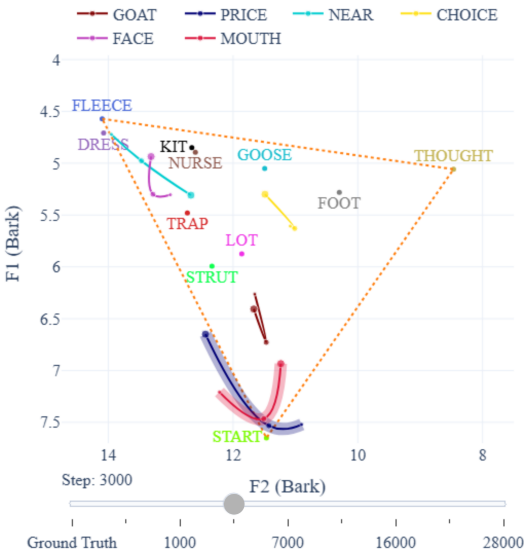
(b) Static vowel space at step 10000.



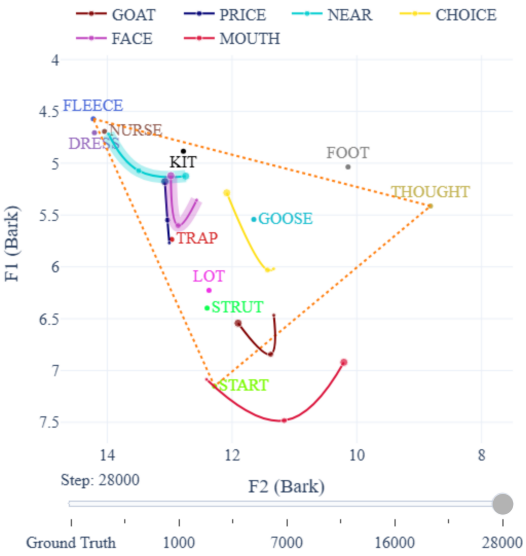
(c) Static vowel space at step 28000.



(d) Dynamic vowel space of ground truth.



(e) Dynamic vowel space at step 3000.



(f) Dynamic vowel space at step 28000.

Figure 3: Static and dynamic vowel spaces at selected training steps. The highlighted vowels correspond to those focused on in Section 4. Numeric labels in the plots correspond to fine-tuning steps: 0=0, 1=1000, 2=3000, 3=7000, 4=10000, 5=16000, 6=20000, 7=28000.

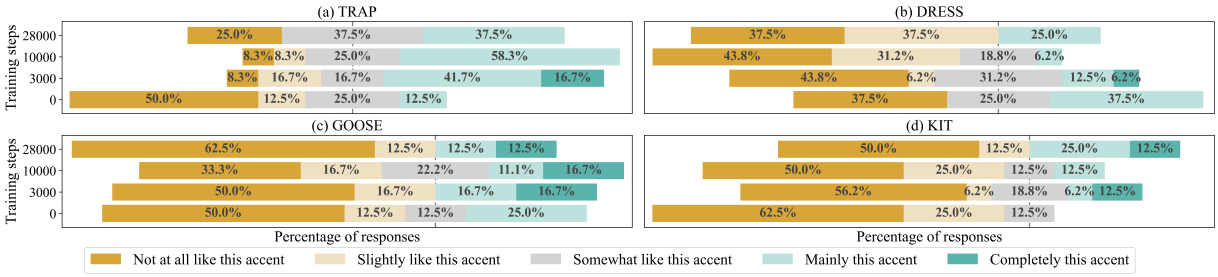


Figure 4: Distribution of Likert-scale responses for the NZE accent in the perception test, visualised using stacked bar graphs.

taining the vowel (Table 1). For each vowel, two sentences were synthesised per step, resulting in $18(9 \times 2)$ sentences per step and $144(18 \times 8)$ total across eight steps. In this online test, 23 participants (aged 22–57, $mean = 33, SD = 12$) rated synthetic speech on a Likert scale from *Not at all like this accent* to *Completely this accent*, adjusting sliders for NZE and GAE. Among them, 13 were first-language English speakers. All had lived in New Zealand for over a year, were proficient in English, and reported no hearing impairments.

4. Results and Discussion

This section presents the vowel space and perceptual analysis results, examining how formant movements visualised with *TTSVowelViz* correspond to the perception of accent change during training.

4.1. Monophthong Analysis

Plots generated by *TTSVowelViz* (Figures 3a, 3b, and 3c) show how the model learns each NZE monophthong *during* training. Comparing the vowel space shape before fine-tuning (Figure 3a) with the one at 28000 steps (Figure 3c) shows a shift towards the triangular vowel space shape of ground-truth NZE speech (Figure 3c), indicating an accent shift from GAE to NZE. Differences in vowel height between the ground-truth and synthesised vowel spaces (Figure 3c) likely stem from differences in speaking style, as the ground-truth speech was carefully articulated, while synthetic speech has continuous patterns with minimal pauses.

Through a vowel-based analysis, we determined whether the formant shifts in the generated vowel spaces align with human perception. “Increase in perception" refers to participants perceiving the synthesised sentences as closer to NZE. Figure 4 presents the Likert-scale responses at selected training steps, with stacked bars representing the distribution of responses. A larger green section indicates more participants perceiving the synthetic speech as *mainly/completely NZE*-accented. Our observations fall into four categories:

1) The vowel space **moved towards the expected** NZE region, and perception **increased** (FLEECE, NURSE, START, THOUGHT, TRAP)

TRAP is a mid-front vowel in NZE (Ross, 2024, p. 44), while in GAE, it is a low-front vowel. Before fine-tuning, TRAP was in the low-front region (highlighted red 0 in Figure 3a). It shifted to the mid-front region in NZE during training (highlighted red path in Figure 3c). This aligns with perception (Figure 4a), where *mainly/completely NZE* responses increased from 12.5% at step 0 to 37.5% at step 28000. FLEECE, NURSE, START, and THOUGHT followed a similar trend where the accent perception aligned with the vowel space movement.

2) The vowel space **did not move towards the expected** NZE region, and perception **decreased** (GOOSE)

GOOSE is a high-front vowel in NZE (Ross, 2024, p. 14). Initially positioned in the high-back region before fine-tuning (highlighted cyan 0 in Figure 3a), it shifted towards the high-central region by step 3000 (point marked as 2 in the highlighted cyan path in Figure 3b) and moved to the mid-central region by step 28000 (Figure 3c). This aligns with perception results (Figure 4c), where *mainly/completely NZE* responses increased from 25% at step 0 to $33.4\%(16.7\% + 16.7\%)$ at step 3000 as GOOSE got closer to the expected NZE region, before dropping back to $25\%(12.5\% + 12.5\%)$ at step 28000.

From the above observations, it can be seen that all the long vowels (FLEECE, NURSE, START, THOUGHT, GOOSE) influenced accent perception. When GOOSE was not in the expected region in the vowel space, participants struggled to identify the accent of the synthesised speech. This indicates that future NZE datasets should reconsider the distribution of GOOSE. Also, more focus needs to be placed on GOOSE to ensure that the speech model learns it accurately, thereby facilitating accurate accent adaptation.

3) The vowel space **moved towards the expected** NZE region, but perception **decreased** (DRESS)

Initially, the highlighted purple 0 of DRESS in Figure 3a appeared in the mid-front region. By 28000 steps, it shifted towards the high-front region (highlighted purple path in Figure 3c). Despite DRESS being known as a high-front vowel in NZE, *mainly/completely NZE* responses declined from 37.5% at 0 steps to 25% at 28000 steps (Figure 4b). This could be because the ground-truth dataset, recorded in 2014 (Watson and Marchi, 2014), features a raised DRESS, a well-known NZE marker. However, Ross (2024) found that DRESS has been lowering in NZE, which is likely what the perception test participants expected.

4) The vowel space **did not move towards the expected** NZE region, but perception **increased** (KIT, LOT)

KIT is centralised in NZE (Ross, 2024, p. 7). Though it stayed in the high-front region throughout training (see highlighted black path in Figure 3c), *mainly/completely NZE* responses increased from 0% at 0 steps to $37.5\%(25\%+12.5\%)$ at 28000 steps (Figure 4d). LOT also showed a similar behaviour.

The observations on KIT and LOT contradict our expectations but are likely due to the quality of the synthesised voice or due to both vowels being short, giving participants less time to perceive accent differences. However, the results also indicate that more emphasis needs to be given to the distribution of these vowels in the training dataset, and model training could be modified to focus more on them.

The accent shift from GAE to NZE is evident as most vowels in the synthesised vowel space moved towards expected NZE regions (Ross, 2024, p. 44). Perception test results reinforce this, with rising *mainly/completely NZE* responses aligning with declining *not at all/slightly GAE* responses. This confirms that *TTSVowelViz* effectively visualises the perceived accent change.

We observed notable and rapid formant shifts within the first 10000 steps. To analyse their statistical significance, we used linear mixed models with training steps as the fixed effect and vowel occurrences as the random effect. The results indicate that all monophthongs, except FOOT, exhibit significant formant shifts $(p < .001)$ within 10000 steps. This suggests that model training could potentially stop at this point. However, further investigation is needed to verify this claim.

4.2. Diphthong Analysis

Formant trajectories of diphthongs in the dynamic vowel spaces generated by *TTSVowelViz* (Figures 3d, 3e, and 3f) visualise F1 and F2 changes over a vowel's duration and their evolution during training. At step 28000 (Figure 3f), FACE and NEAR vowels closely resemble the ground truth (Figure 3d). Interestingly, MOUTH and PRICE match the ground truth better at 3000 steps (Figure 3e), indicating they were learned earlier, but accuracy was reduced with further fine-tuning. A future perception test could assess whether human perception aligns with these trajectory shifts. As with the static vowel spaces, by observing the dynamic vowel spaces, we determined that the most notable formant shifts occur within the first 10000 steps, which cannot be inferred solely from Figures 3d, 3e, and 3f.

TTSVowelViz provides a linguistically grounded and interpretable method for evaluating speech models, and allows researchers to visualise the evolution of vowel spaces in speech models. Future work could use vowel space changes to determine optimal early stopping points for training, complementing loss curves. Additionally, quantifying the differences between ground-truth and synthesised vowel spaces could refine loss functions or attention mechanisms, thereby improving the naturalness of speech synthesis.

5. Conclusion

This study introduces *TTSVowelViz*, a tool for visualising static and dynamic vowel spaces *during* speech model training. It offers an intuitive way to evaluate how well a model aligns with expected language characteristics, providing an interpretable approach to improve TTS systems. *TTSVowelViz* paves the way for model training methods that can adapt to learn vowel spaces accurately. We aim to encourage speech researchers to consider vowel space accuracy more closely in model development, contributing to speech synthesis informed by linguistic knowledge.

6. Ethical Considerations

TTSVowelViz does not collect new data. But its evaluation involves speech corpora and human perception testing, which raises ethical considerations.

- *Data Use and Consent:* All speech data were obtained from publicly available or appropriately licensed datasets. For perception tests, participants provided informed consent, and no personally identifiable information was collected. Participation was anonymous and voluntary, with the option to withdraw at any time.
- *Potential Misuse: TTSVowelViz* is not susceptible to misuse within TTS, as it functions solely as an interpretability tool. However, TTS models analysed using the tool could potentially be misused for impersonation, voice cloning, or the generation of deceptive content. Therefore, the responsible deployment of TTS systems

should incorporate safeguards such as watermarking and usage policies.

- *Broader Impact:* By introducing vowel space visualisation into TTS model training, this work promotes linguistically informed interpretability. This can support the identification of modelling biases and enable more informed speech evaluation across languages and accents.

7. Limitations

As one of the first efforts to introduce linguistically grounded visualisation of TTS training, *TTSVowelViz* focuses on vowels and does not capture other phonetic and prosodic features such as consonants, rhythm, or intonation. Future research may extend the approach to additional speech features.

The visualisations rely on automatic segmentation and formant extraction from synthetic speech, which may introduce variability. To mitigate this, we average formant values across multiple occurrences of a vowel.

The empirical evaluation in this study is limited to a single English accent. However, *TTSVowelViz* is designed to be both language- and accent-agnostic. Validation across additional languages, accents, and TTS architectures would further demonstrate its general applicability.

TTSVowelViz is intended as an interpretability tool. It complements, rather than replaces, existing objective and perceptual evaluation methods in TTS research.

8. Acknowledgements

This work was supported by the Marsden Fund Council from government funding, managed by Royal Society Te Apārangi. The authors thank the participants who took part in the perception test and the University of Auckland Faculty of Engineering and Design for providing participation incentives.

9. Bibliographical References

Binu Abeysinghe, Jesin James, Catherine I. Watson, and Felix Marattukalam. 2022. Visualising Model Training via Vowel Space for Text-To-Speech Systems. In *Interspeech*, pages 511–515.

Sven Albrecht, Rewa Tamboli, Stefan Taubert, Maximilian Eibl, Günter D. Rey, and Josef Schmied. 2022. Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems: Measuring Monophthong Naturalness. In *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*.

Henry An, Jesin James, Catherine Watson, and Binu Abeysinghe. 2024. Visualising Vowel Dynamics for the Training of Text to Speech Systems. In *Australasian International Conference on Speech Science and Technology*, pages 197–201.

David Brett. 2004. Computer generated feedback on vowel production by learners of English as a second language. *ReCALL*, 16(1):103–113.

Michael Carey. 2004. CALL Visual Feedback for Pronunciation of Vowels. *CALICO Journal*, 21(3):571–601.

Xinyuan Chao, Charbel El-Khaissi, Nicholas Kuo, Priscilla K. John, and Hanna Suominen. 2020. An Automatic Vowel Space Generator for Language Learner Pronunciation Acquisition and Correction. In *Annual Workshop of the Australasian Language Technology Association*, pages 54–64.

Felicity Cox, Joshua Penney, and Sallyanne Palethorpe. 2024. Australian English Monophthong Change across 50 Years: Static versus Dynamic Measures. *Languages*, 9(3).

Anushiya R. Gladston and K. V. Pradeep. 2023. Exploring Solutions for Text-to-Speech Synthesis of Low-Resource Languages. In *International Conference on Signal Processing and Communication*, pages 168–172.

Jonathan Harrington, Felicity Cox, and Zoe Evans. 1997. An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics*, 17(2):155–184.

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111.

Ewa Jacewicz, Robert A. Fox, and Joseph Salmons. 2007. Vowel space areas across dialects and gender. In *International Congress of Phonetic Sciences*, pages 1465–1468.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71.

Sandra Jansen. 2019. Change and stability in GOOSE, GOAT and FOOT: back vowel dynamics in Carlisle English. *English Language and Linguistics*, 23(1).

Allard Jongman, Marios Fourakis, and Joan A. Sereno. 1989. The Acoustic Vowel Space of Modern Greek and German. *Language and Speech*, 32(3):221–248.

Michael Kiefte and Terrance M. Nearey. 2017. Modeling consonant-context effects in a large database of spontaneous speech recordings. *The Journal of the Acoustical Society of America*, 142(1):434–443.

Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. 2023. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications*, 82:15171–15197.

Philipos C. Loizou. 2011. Speech Quality Assessment. In *Multimedia Analysis, Processing and Communications*, volume 346, pages 623–654. Springer Berlin Heidelberg.

Gordon E. Peterson and Harold L. Barney. 1952. Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Margaret Renwick and Joseph Stanley. 2020. Modeling dynamic trajectories of front vowels in the American South. *The Journal of the Acoustical Society of America*, 147:579–595.

Brooke Ross, Elaine Ballard, and Catherine Watson. 2023. Front vowel lowering in New Zealand English: investigating a regional sound change. In *International Congress of Phonetic Sciences*, pages 3666–3670.

Brooke C. Ross. 2024. *A new look at sound change in New Zealand English*. Ph.D. thesis, The University of Auckland.

Steven Sandoval, Visar Berisha, Rene L. Utianski, Julie M. Liss, and Andreas Spanias. 2013. Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, 134(5):EL477–EL483.

Florian Schiel. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. In *International Congress of Phonetic Sciences*, pages 607–610.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783.

Joseph A. Stanley, Margaret E. L. Renwick, Katherine I. Kuiper, and Rachel M. Olsen. 2021. Back Vowel Dynamics and Distinctions in Southern American English. *Journal of English Linguistics*, 49(4):389–418.

Catherine Watson and Jonathan Harrington. 1999. Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America*, 106(1):458–468.

Catherine Watson, Brooke Ross, Elaine Ballard, Helen Charters, Richard Arnold, and Miriam Meyerhoff. 2018. Preliminary Investigations into Sound Change in Auckland. In *Australasian International Conference on Speech Science and Technology*, pages 17–20.

Catherine I. Watson, Jonathan Harrington, and Zoe Evans. 1998. An acoustic comparison between New Zealand and Australian English vowels. *Australian Journal of Linguistics*, 18(2):185–207.

Catherine I. Watson, Peter J. Keegan, Margaret A. Maclagan, Ray Harlow, and Jeanette King. 2017. The Motivation and Development of MPAi, a Māori Pronunciation Aid. In *Interspeech*, pages 2063–2067.

Catherine I. Watson, Margaret A. Maclagan, Jeanette King, Ray Harlow, and Peter J. Keegan. 2016. Sound change in Māori and the influence of New Zealand English. *Journal of the International Phonetic Association*, 46(2):185–218.

John C. Wells. 1982a. *Accents of English 1: An Introduction*, volume 1. Cambridge University Press.

John C. Wells. 1982b. *Accents of English 3: Beyond the British Isles*, volume 3. Cambridge University Press.

10. Language Resource References

Keith Ito and Linda Johnson. 2017. *The LJ Speech Dataset*. PID https://keithito.com/LJ-Speech-Dataset/.

Catherine Watson and Amélie Marchi. 2014. *Resources Created For Building New Zealand English Voices*. In *Australasian International Conference on Speech Science and Technology*, pages 92–95.