AUGUST 27 2025

Language-specific phonetic realisation of stop voicing contrasts in English and Japanese synthesised speech

James Tanner ; Yasuaki Shinohara ; Faith Chiu

JASA Express Lett. 5, 085202 (2025)
https://doi.org/10.1121/10.0039066

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Language-specific phonetic realisation of stop voicing contrasts in English and Japanese synthesised speech

James Tanner,[1,a] Yasuaki Shinohara,[2] and Faith Chiu[1]
[1]English Language and Linguistics, University of Glasgow, Glasgow, G12 8QQ, United Kingdom
[2]Faculty of Commerce, Waseda University, Tokyo, 169-8050, Japan
james.tanner@glasgow.ac.uk, y.shinohara@waseda.jp, faith.chiu@glasgow.ac.uk

Abstract: Speech synthesis has improved dramatically over recent years, enabled by large datasets and advances in neural network architectures. Little is known, however, about how synthesised speech patterns are realized from a phonetic perspective. By synthesising speech in two languages with differing implementations of stop voicing, we observe that synthesised speech broadly follows expected patterns for each language, though partially diverges for specific segments. Synthesising speakers into the opposing language also results in stops similar to target language distributions. These findings demonstrate the capability of speech synthesis models to encode phonetic information and further motivate questions regarding the phonetics of synthesised speech.

[Editor: Charles C. Church] https://doi.org/10.1121/10.0039066

Received: 16 June 2025 **Accepted:** 4 August 2025 **Published Online:** 27 August 2025

1. Introduction

Recent years have seen a substantial increase in both the performance and widespread application of machine learning models, including large language models (LLMs), across many spheres of life. In the domain of speech technology, since the release of the Tacotron speech synthesis model (Wang *et al.*, 2017), a vast array of speech synthesis models have been released with the capacity to generate highly fluent and naturalistic speech, including the "cloning" or "zero-shot synthesis" of a target speaker's voice with small amounts (1 ∼ 5 min) of input speech (e.g., Betker, 2023; Wang *et al.*, 2023; Du *et al.*, 2024; Liao *et al.*, 2024; Casanova *et al.*, 2024). In contrast to the previous "state-of-the-art" approaches to speech synthesis (see Tokuda *et al.*, 2013, for a review), these models are trained to predict a spectrogram given an input text, in many cases without utilising any intermediate phonological representations (i.e., phoneme sequences). Instead, these models utilise the LLM framework of using high-dimensional "tokens" to represent the complex relationship between text and spectral information (Guo *et al.*, 2025).

Despite the highly fluent and expressive speech generated from these speech synthesis models, there remains little understanding as to how these models store and represent linguistic information (see, e.g., tom Dieck *et al.*, 2022; Liu *et al.*, 2023, for analysis of similar models), as well as the extent to which this synthesised speech exhibits similar patterns of linguistic variability to speech produced by humans. The quality of speech synthesis models is typically evaluated impressionistically (see Kirkland *et al.*, 2023, for a review) or through applying speech recognition to their outputs (Taylor and Richmond, 2021), leaving unclear how linguistic and phonetic information is reproduced in the synthesised speech signal. Indeed, very little work has investigated the production of speech by modern speech synthesis systems from an acoustic-phonetic perspective, leaving open many questions about how speech synthesis models "behave" phonetically, such as how linguistic contrasts are phonetically reproduced in synthesised speech, and how speaker-specific and linguistic information is disentangled in order to synthesise a particular speaker (e.g., Gwizdzinski *et al.*, 2023; Song *et al.*, 2025). We explore these questions in this study, focusing specifically on the voicing contrast in stop consonants and how synthetically-generated speech represents these contrasts acoustically. Put differently, *in what ways do synthesised stops pattern similarly or differently from human-produced stops?* To explore this issue, we compare the acoustic-phonetic patterning of synthesised stops to that expected from the previous literature based on speech produced by humans. Specifically, we examine the marking of the phonological voicing contrast in both English and Japanese: two languages that share the same phonological voicing contrast but differ greatly in how it is implemented phonetically. While both languages utilise a two-way voiced/voiceless distinction in singleton stops ({/p/, /t/, /k/} vs {/b/, /d/, /g/}), English is characterised as an "aspirating" language with the phonological voicing in stops predominantly distinguished by (positive) voice onset time (VOT); this contrasts with the characterisation of Japanese as a "voicing" language, utilising greater use of voicing during stop

[a]Author to whom correspondence should be addressed.

closures to mark the phonological contrast (Iverson and Salmons, 1995; Nasukawa, 2005). Our two research questions, which explore the extent to which modern speech synthesis models learn and reproduce patterns of linguistic and phonetic variation, are as follows:

- RQ1: To what extent does the acoustic realisation of stop voicing in synthesised speech follow the expected phonetic patterns from human speech (e.g., in both English and Japanese respectively)?
- RQ2: What is the acoustic-phonetic profile of stops when synthesised in the opposite language of the input speaker (e.g., a Japanese speaker → English)?

Specifically, we investigate these questions concerning three acoustic cues to the stop voicing contrast in utterance-medial stops: VOT, closure duration (CD), and the degree of closure voicing (CV). English has been found to have medial positive VOT values of 7–20 milliseconds (ms) for voiced stops, and 40-60ms for voiceless stops (Edwards, 1981; Byrd, 1993; Sonderegger *et al.*, 2020). Japanese medial VOT is typically realised with approximately 12 ms for voiced stops (Tanner *et al.*, 2025a; Shimizu, 1996) and between 25 and 60 ms for voiceless stops, depending on place of articulation (Tanner *et al.*, 2025a; Riney *et al.*, 2007). English CD has also been found to be between 40–50 ms for voiceless stops, and 50-80ms for voiced stops (Edwards, 1981; Sonderegger *et al.*, 2020; Byrd, 1993), compared with 35 ms for voiced and 65 ms for voiceless stops for Japanese (Homma, 1981; Tanner *et al.*, 2025a). Keating (1984) reports that CV is common in English intervocalic positions, with voicing "bleeding" from the preceding segment (Docherty, 1992; Davidson, 2016). While CV itself has been analysed with multiple different approaches across studies, it has been found that English voiceless stops are likely to contain approximately 20%–30% voicing (Docherty, 1992; Edwards, 1981), with 70%–80% voicing for voiced stops (Jacewicz *et al.*, 2009; Edwards, 1981). In contrast, CV is less frequent in Japanese voiceless stops (∼10%), but near-obligatory (>80%) voicing for voiced stops (Gao and Arai, 2019; Tanner *et al.*, 2020).

2. Methods

2.1 Data

The data used for this study comes from the *University of Tsukuba Multi-Language Corpus* (Itahashi, 2006), each containing reading passages (*North Wind & Sun*) and single-word productions (days of the week, numerals, etc) from each speaker for 11 languages. For this study, we use the English and Japanese recordings from eight (four female) English and 13 (six female) Japanese speakers.[1] The speech synthesis model used to generate speech samples for this study was the X-TTS speech synthesis model (Casanova *et al.*, 2024), chosen because it supports both zero-shot speaker synthesis for both languages of interest, in order to avoid the potential confound of comparing different models with different architectures. The X-TTS model consists of a generative pretrained transformer (GPT) architecture trained to predict a spectrogram given a sentence of written text and language code (in the form "`[lang]text`") and converted to a 25 kHz waveform. Speaker information is encoded within a 1028-dimensional embedding space, which is used to condition the vocoding process during upsampling of the waveform.

For Japanese, 100 sentences were selected from the phoneme-balanced subset of the *Voice Actress Corpus* (Sonobe *et al.*, 2017). For English, 582 sentences (due to the relative rarity of intervocalic stops) were selected from *The Rainbow Passage*, *Please Call Stella*, *Comma Gets A Cure*, *Arthur The Rat*, and the TIMIT corpus (Garofolo *et al.*, 1993). As the X-TTS model does not natively support multi-speaking finetuning, all sentences in both languages were synthesised in a zero-shot fashion (i.e., without model finetuning), resulting in speech generated in both the same language as the target speaker (i.e., English → English; Japanese → Japanese) as well as the opposite language (English → Japanese; Japanese → English). All synthesised sentences were then aligned using the Montreal Forced Aligner (McAuliffe *et al.*, 2017a), using the `english_mfa` and `japanese_mfa` phonesets and acoustic models, respectively. Data for all utterance-medial stops, including segmental information (e.g., place of articulation, duration, speech rate), and speaker metadata (e.g., synthesised language, speaker original language) were extracted from the aligned corpora extracted using *PolyglotDB* package (McAuliffe *et al.*, 2017b). VOT was estimated using the *AutoVOT* package (Keshet *et al.*, 2014), using the default pretrained acoustic model, with approximately 25% tokens of each dataset spot-checked. CD was calculated as the duration between the onset of the force-aligned stop boundary and the start of VOT. CV was calculated using the *VoiceReport* function in the *Parselmouth* (Jadoul *et al.*, 2018) Python wrapper for Praat (Boersma and Weenink, 2023), based on speaker-level estimated Pitch tracks. The *fraction of unvoiced frames* within the stop closure was then subtracted from 100 (to reflect a fraction of voiced frames) and then divided by 100 to reflect a range of closure voicing values between 0 and 1 [e.g., (100–70)/100 = 0.3, Tanner *et al.*, 2025a)]. Japanese stops either followed or preceded by devoiced vowels were excluded. In total, 16 632 English and Japanese utterance-medial stops were included in the analysis, which is summarised in Table 1.

2.2 Models

To address our two research questions, we take an approach to linear regression modelling that best suits the properties and distribution of the three acoustic properties of interest (VOT, CD, CV), made possible by the flexibility of model specification in the *brms* (Bürkner, 2021) Bayesian regression interface to the Stan programming language (Carpenter *et al.*, 2017).

We model VOT and CD—for which the logarithm of their values is normally distributed—using the `Lognormal` model family. As CV is measured as a bounded [0,1] value, we model closure duration using the `Zero-One Inflated Beta` (ZOIB) regression family. ZOIB consist of two distributional response parameters, corresponding to the mean (μ) and spread (ϕ) of the Beta distribution, as well as two additional logistic/binomial responses capturing the probability of values at the extreme 0 and 1 values (Ospina and Ferrari, 2012).[2]

The structure of these regression models was designed to specifically address our research questions while accounting for other sources of variability in stop realisation, including speech rate, place of articulation, and surrounding vowel height. Specifically, our model structure contained a binary term for the phonological voicing category of the stop ({`voiced`,`voiceless`}) as well as a 4-level factor reflecting the four combinations of speaker original language and synthesised language ({`English-English`, `English-Japanese`, `Japanese-English`, `Japanese-Japanese`}), as well as an interaction term between phonological voicing and this 4-level factor, allowing the size of the voicing effect to differ for each combination. We also included terms for (log-transformed) speech rate, place of articulation of the stop, and the height of the preceding and following vowel. To further account for sources of variability for each synthesised-language/speaker-language combination, we also included interaction terms between speech rate and synthesised-language, place of articulation and synthesised-language, speech rate and phonological voicing category, and place of articulation and voicing category. All models included by-speaker random effects for voicing category (correlated with intercepts), place of articulation, speech rate, and previous and following vowel height. The models were fit with "weakly-informative regularising" priors, which discourage extreme effect sizes without biasing parameter estimation towards a particular size or direction (Vasishth *et al.*, 2018). The VOT and CD models were specified with a $t(3, 2.9, 2.5)$ prior for the intercepts, while the CV model was specified with $t(3, 0, 2.5)$ priors for the μ and ϕ intercepts and $logistic(0, 1)$ priors for the zero-one inflation intercepts. Fixed effect terms for all models were specified with a $Normal(0, 2)$ prior, and random effects were specified with a $t(3, 0, 2.5)$ prior. The speaker-level correlation terms were specified with a $lkj(1.5)$ prior to discourage extreme $(-1,1)$ correlation estimations. Model posteriors were sampled for 4000 iterations (2000 warmup) across four chains, resulting in 8000 posterior samples.

3. Results

We address our research questions by considering the relative differences in both overall effects (e.g., average VOT) and the size of the voicing contrast (e.g., the average voiced-voiceless difference in VOT) first (RQ1) in speech synthesised in the same language as the input speaker (i.e., English–English vs Japanese–Japanese), and second (RQ2) in speech synthesised in both the speaker's original language and opposite language (e.g., English–English vs English–Japanese). We report these differences as model-estimated marginal mean effects between conditions calculated with the *emmeans* package (Lenth, 2023), and report the median estimated difference ($\hat{\Delta}$) and the 100% Credible Interval (CI).[3] To assess the degree of evidence for a particular effect (e.g., a language-specific difference in VOT), we evaluate the degree to which the parameter's credible interval falls within the *region of practical equivalence* (ROPE)—range of values with negligible effect size (Kruschke, 2010)—which we treat as the range $[-0.1, 0.1]$ (Kruschke and Liddell, 2018). As we are considering the full range of the posterior distribution (100% confidence interval, CI) with respect to the ROPE, we consider there to be evidence for an effect if the percentage of the distribution within ROPE (%*ROPE*) is less than 2.5 (Makowski *et al.*, 2019).[4]

With respect to RQ1, while there is evidence for the languages differing in the size of the VOT voicing contrast ($\hat{\Delta} = 0.41, CI = [0.09, 0.68], \%ROPE = 0$), Fig. 1(A) illustrates that these estimated VOT values differ from those expected from previous studies of human speech. While English voiced stops fall within the expected VOT range (10–19 ms), we observe that the synthesised English voiceless stops exhibit a shorter-than-expected VOT range of 17–36 ms, compared with the range of 40–60 ms observed for human-produced stops (e.g., Byrd, 1993; Sonderegger *et al.*, 2020). Similarly, for Japanese, synthesised voiceless stops are realised with much shorter VOT in synthesised speech (10–18 ms) than that expected for human speech (25–60 ms, Riney *et al.*, 2007; Tanner *et al.*, 2025a). The voicing contrast for CD similarly differs between synthesised language ($\hat{\Delta} = -0.34, CI = [-0.54, -0.11], \%ROPE = 0$), as illustrated in Fig. 1(B). Specifically, we find little evidence for a CD-based voicing contrast in English ($\hat{\Delta} = 0.08, CI = [-0.05, 0.23], \%ROPE = 72.26$), with voiceless closures shorter (36-56ms) than the expected 50-80ms (e.g., Byrd, 1993). This contrasts with the estimated CDs for Japanese stops of 35–50 ms and 52–75 ms for voiced and voiceless stops, respectively, which closely follow the patterns for human-produced Japanese stops (Homma, 1981; Tanner *et al.*, 2025a). We find that languages differ in the degree to which CV differs between voiced and voiceless stops in synthesised speech ($\hat{\Delta} = 1.02, CI = [0.55, 1.48], \%ROPE = 0$),

Table 1. Counts of synthesised tokens used in the analysis, grouped by original language of the speaker.

	Speakers		Tokens	
Speaker language	Male	Female	English	Japanese
English	8	4	3672	2277
Japanese	13	6	6772	3911

though this mainly reflects differences in CV for voiced stops specifically [Fig. 1(C)]. For English, voiced CD has an estimated range of 42%–56%, substantially lower than that expected for English human speech (70%–90%, Jacewicz *et al.*, 2009). Japanese synthesised voiced stops are similarly estimated to exhibit lower CD (46%–57%) than the >%80 expected from previous studies (Gao and Arai, 2019).

When comparing the acoustic properties of synthesised stops between the two speaker languages (RQ2), we find that English stops synthesised from both English and Japanese speakers differ in the size of the VOT voicing contrast ($\hat{\Delta} = -0.22, CI = [-0.50, 0.03], \%ROPE = 0.03$) [see Fig. 1(A)]. This is not the case for stops synthesised in Japanese, however, which do not differ in the VOT voicing contrast size ($\hat{\Delta} = 0.06, CI = [-0.26, 0.29], \%ROPE = 73.67$). We also observe negligible evidence for differences between input speaker language, for either English or Japanese stops, for either CD [Fig. 1(B)] or CV [Fig. 1(C)].

4. Discussion

Despite the large advances in the quality and naturalness of speech generated from modern speech synthesis models, little is known about how these models represent and reproduce patterns of linguistic variability, including the extent to which these synthesised speech patterns are similar to human speech from an acoustic-phonetic perspective. The goal of this study is to provide an initial examination into the linguistic behaviour of modern neural-network-based speech synthesis models, focusing specifically on how acoustic cues to the phonological voicing contrast in stops are acoustically realised in synthesised speech. By comparing synthesised speech in two languages that differ in the phonetic implementation of stop voicing (English and Japanese), we find that the acoustic properties of synthesised stops follow the expected patterns for each language to an extent, though crucially exhibit some differences from human-produced stops in VOT, CD, and CV. This finding suggests that modern speech synthesis models can learn and reproduce the acoustic-phonetic properties of linguistic contrasts in a language-specific way, but also partially diverge from the expectations for specific cues. When comparing the synthesis of a speaker into the opposite language (e.g., English → Japanese), we find that these stops pattern similarly to the language being *synthesised* instead of the original language of the speaker. This suggests that the model is successfully able to disentangle the language-specific and speaker-specific information from the speech signal, and favours the language-specific implementation of the stop over that of the implementation it receives as input for conditioning the synthesised output.

Together, these findings suggest that modern speech synthesis models are capable of encoding and reproducing patterns of phonetic variation, though the extent to which synthesised speech patterns are similar to human speech varies between acoustic cues and between languages. While the fact that models can capture broad patterns of phonetic variability might not be immediately surprising (given the high fluency and naturalness ratings given to the outputs of these models), the observations reported here raise a number of questions necessary for developing a more comprehensive understanding of the linguistic behaviour of modern speech synthesis models. For example, why do some synthesised acoustic dimensions pattern more closely to human speech than others? While this study explicitly explored acoustic patterns of synthesised stops, it remains unclear the extent to which non-contrastive linguistic properties may be distangled from the speech signal. For example, to what extent are other speaker-specific dimensions—such as pitch, voice quality, and intonation—maintained (or disregarded) when synthesising a speaker into a different language? To address these questions more directly, a clear direction for future work would be to compare synthetic outputs for a given speaker directly with that speaker's genuine (human) productions. The synthesised speech in this study was created without finetuning: how would

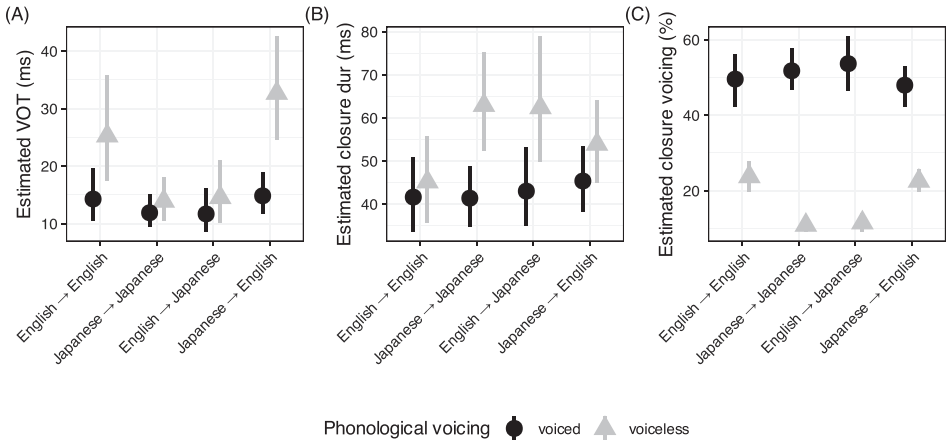

Fig. 1. Model-estimated VOT (A), CD (B), and CV (C) for voiced (black circle) and voiceless (gray triangle) synthesised stops. The *X*-axis represents the language of the target speaker and the synthesised language (e.g., English → Japanese = English speaker synthesised in Japanese). Points indicate the median of the posterior distribution, with lines indicating the 95% high-density interval.

the acoustic realisation of stops differ following finetuning? As the model used in this study explicitly takes the language of a text as part of the input, to what extent are the findings observed here dependent on the specific architecture of this model? To what extent can these models capture *within-language* (i.e., cross-dialectal) variation in the realisation of linguistic contrasts (Gwizdzinski *et al.*, 2023)? How are speaker-specific acoustic properties retained for speakers that fall outside of the typical speech datasets used for training speech synthesis models, such as child speech or speakers from non-standard dialects? Given recent work demonstrating patterns of linguistic bias in modern neural speech models (e.g., Chang *et al.*, 2024), it is essential for further research to focus on understanding how speech synthesis models learn, represent, and reproduce linguistic information, and the extent to which speech generated by modern speech synthesis systems may or may not pattern similarly to speech from human speakers.

Supplementary Material

See the supplementary material for full regression model tables.

Acknowledgements

This research was supported by a British Academy Postdoctoral Fellowship awarded to J.T.

Author Declarations

Conflict of Interest

The authors have no conflicts of interest to declare.

Data Availability

The code and data that support the findings of this study are available in Tanner *et al.* (2025b) at https://doi.org/10.17605/OSF.IO/3GDVT.

References

[1]While all speakers are native speakers of their respective languages, demographic information about speakers, including age and dialect background, is not provided in the corpus metadata.
[2]See Tanner *et al.* (2025a) for the application of the ZOIB response family to the modelling of stop closure voicing patterns.
[3]See Torres and Steffman (2023) for a similar analysis of stop voicing using estimated marginal means.
[4]See Soo and Babel (2025) for an example of ROPE analysis applied to phonetic data.

Betker, J. (**2023**). "Better speech synthesis through scaling" arXiv:2305.07243.
Boersma, P., and Weenink, D. (**2023**). "Praat: Doing phonetics by computer [computer program] (version 6.3.09)" http://www.praat.org/ (Last viewed 22 May 2025).
Bürkner, P.-C. (**2021**). "Bayesian item response modeling in R with brms and Stan," J. Stat. Softw. **100**(5), 1–54.
Byrd, D. (**1993**). "54,000 American stops," UCLA Work. Papers Linguistics **83**, 97–116.
Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (**2017**). "Stan: A probabilistic programming language," J. Stat. Softw. **76**(1), 1–32.
Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. (**2024**). "XTTS: A massively multilingual zero-shot text-to-speech model," in *Proceedings of Interspeech 2024*, pp. 4978–4982.
Chang, K., Chou, Y.-H., Shi, J., Chen, H.-M., Holliday, N., Scharenborg, O., and Mortensen, D. R. (**2024**). "Self-supervised speech representations still struggle with African American Vernacular English," in *Proceedings of Interspeech 2024*, pp. 4643–4647.
Davidson, L. (**2016**). "Variability in the implementation of voicing in American English obstruents," J. Phon. **54**, 35–60.
Docherty, G. (**1992**). *The Timing of Voicing in British English Obstruents* (Foris, New York).
Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., Gao, Z., and Yan, Z. (**2024**). "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," arXiv:2407.05407.
Edwards, T. J. (**1981**). "Multiple features analysis of intervocalic English plosives," J. Acoust. Soc. Am. **69**, 535–547.
Gao, J., and Arai, T. (**2019**). "Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery," J. Phon. **77**, 100932–100933.
Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM," NIST Report No. 4930, National Institute of Standards and Technology, Gaithersburg, MD.
Guo, Y., Li, Z., Wang, H., Li, B., Shao, C., Zhang, H., Du, C., Chen, X., Liu, S., and Yu, K. (**2025**). "Recent advances in discrete speech tokens: A review," arXiv:2502.06490.
Gwizdzinski, J., Barreda, S., Carignan, C., and Zellou, G. (**2023**). "Perceptual identification of oral and nasalized vowels across American English and British English listeners and TTS voices," Front. Commun. **8**, 1307547.
Homma, Y. (**1981**). "Durational relationship between Japanese stops and vowels," J. Phon. **9**, 273–281.
Itahashi, S. (**2006**). "University of Tsukuba Multilingual Speech Corpus (UT-ML)," DSC Reference Portal, Dataset. https://doi.org/10.32130/src.UT-ML
Iverson, G., and Salmons, J. (**1995**). "Aspiration and laryngeal representation in Germanic," Phonology **12**, 369–396.
Jacewicz, E., Fox, R. A., and Lyle, S. (**2009**). "Variation in stop consonant voicing in two regional varieties of American English," JIPA **39**, 313–334.
Jadoul, Y., Thompson, B., and de Boer, B. (**2018**). "Introducing parselmouth: A Python interface to Praat," J. Phon. **71**, 1–15.

Keating, P. A. (**1984**). "Phonetic and phonological representation of stop consonant voicing," Language **60**, 286–319.
Keshet, J., Sonderegger, M., and Knowles, T. (**2014**). "AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction," https://github.com/mlml/autovot/ (Last viewed 28 May 2025).
Kirkland, A., Mehta, S., Lameris, H., Henter, G. E., Szekely, E., and Gustafson, J. (**2023**). "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," in *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW2023)*, pp. 41–47.
Kruschke, J. K. (**2010**). "What to believe: Bayesian methods for data analysis," Trends Cogn. Sci. **14**, 293–300.
Kruschke, J. K., and Liddell, T. M. (**2018**). "The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective," Psychon. Bull. Rev. **25**, 178–735.
Lenth, R. V. (**2023**). "*emmeans: Estimated Marginal Means, aka Least-Squares Means*, R package version 1.9.0," https://CRAN.R-project.org/package=emmeans (Last viewed 6 June 2025).
Liao, S., Wang, Y., Li, T., Cheng, Y., Zhang, R., Zhou, R., and Xing, Y. (**2024**). "Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis," arXiv:2411.01156.
Liu, O., Tang, H., and Goldwater, S. (**2023**). "Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces," arXiv:2305.12464.
Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (**2019**). "bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework," JOSS **4**(40), 1541.
McAuliffe, M., Scolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017a**). "Montreal forced aligner [computer program]," https://montrealcorpustools.github.io/Montreal-Forced-Aligner/ (Last viewed 16 June 2025).
McAuliffe, M., Stengel-Eskin, E., Socolof, M., and Sonderegger, M. (**2017b**). "Polyglot and Speech Corpus Tools: A system for representing, integrating, and querying speech corpora," in *Proceedings of Interspeech 2017*.
Nasukawa, K. (**2005**). "The representation of laryngeal-source contrasts in Japanese," in *Voicing in Japanese*, edited by J. van de Weijer, K. Nanjo, and T. Nishihara (De Gruyter Mouton, Berlin), pp. 71–87.
Ospina, R., and Ferrari, S. L. (**2012**). "A general class of zero-or-one inflated beta regression models," Comput. Stat. Data Anal. **56**, 1609–1623.
Riney, T. J., Takagi, N., Ota, K., and Uchida, Y. (**2007**). "The intermediate degree of VOT in Japanese initial stops," J. Phon. **35**, 439–443.
Shimizu, K. (**1996**). *A Cross-Language Study of the Voicing Contrasts of Stop Consonants in Asian Languages* (Seibido, Tokyo).
Sonderegger, M., Stuart-Smith, J., Knowles, T., MacDonald, R., and Rathcke, T. (**2020**). "Structured heterogeneity in Scottish stops over the twentieth century," Language **96**, 94–125.
Song, J. Y., Rojas, C., and Pycha, A. (**2025**). "Factors modulating perception and production of speech by AI tools: A test case of Amazon Alexa and Polly," Front. Psychol. **16**, 1520111.
Sonobe, R., Takamichi, S., and Saruwatari, H. (**2017**). "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv:1711.00354.
Soo, R., and Babel, M. (**2025**). "Processing pronunciation variation with independently mappable allophones," J. Phon. **110**, 101402.
Tanner, J., Igarashi, Y., and Maekawa, K. (**2025a**). "Speech rate effects on the realisation of multiple acoustic cues to the Japanese stop voicing contrast," J. Acoust. Soc. Am. **157**, 2624–2635.
Tanner, J., Shinohara, Y., and Chiu, F. (**2025b**). "Language-specific phonetic realisation of stop voicing contrasts in English and Japanese synthesised speech," https://osf.io/3gdvt/ (Last viewed 16 June 2025).
Tanner, J., Sonderegger, M., and Stuart-Smith, J. (**2020**). "Structured speaker variability in Japanese stops: Relationships within versus across cues to stop voicing," J. Acoust. Soc. Am. **148**, 793–804.
Taylor, J., and Richmond, K. (**2021**). "Confidence intervals for ASR-based TTS evaluation," in *Proceedings of Interspeech 2021*, pp. 2791–2795.
Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (**2013**). "Speech synthesis based on hidden Markov models," Proc. IEEE **101**(5), 1234–1252.
tom Dieck, T., Pérez-Toro, P. A., Arias, T., Noeth, E., and Klumpp, P. (**2022**). "Wav2vec behind the scenes: How end2end models learn phonetics," in *Proceedings of Interspeech 2022*, pp. 5130–5134.
Torres, C., and Steffman, J. (**2023**). "Stop voicing in drehu: Effects of place of articulation, speaker sex, and language attitudes," in *Proceedings of the 20th ICPhS*, Prague, Czech Republic, pp. 2996–3000.
Vasishth, S., Nicenboim, B., Beckman, M., Li, F., and Kong, E. J. (**2018**). "Bayesian data analysis in the phonetic sciences: A tutorial introduction," J. Phon. **71**, 147–161.
Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. (**2023**). "Neural codec language models are zero-shot text to speech synthesizers," arXiv:2301.02111.
Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (**2017**). "Tacotron: Towards end-to-end speech synthesis," arXiv:1703.10135.