NATURAL CHOICE: COMPARING PLACE CLASSIFICATION BETWEEN NATURAL AND TACOTRON FRICATIVES

Ayushi Pandey[1], Sébastien Le Maguer[1], Jens Edlund[2], Naomi Harte[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

[2]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
pandeya@tcd.ie, lemagues@tcd.ie, edlund@speech.kth.se, nharte@tcd.ie

ABSTRACT

As commercial prevalence of high-quality TTS voices continues to grow, little is known about their suitability for speech sciences. In this paper, we compare feature importance in obstruent place classification between natural voice, and two TTS systems: Tacotron WaveNet and Tacotron WaveGAN. Support Vector Machines (SVMs) are used for classification, using 20 acoustic-phonetic features. First, individual feature strength was evaluated by passing each feature individually to the SVMs. Then, classification was made using the full 20 featureset, to test the most high performing feature combination.

Top-ranked predictors in natural voice were compared with each of the Tacotron voices. We find that Tacotron voices greatly resemble natural in contrasts involving sibilant fricatives w.r.t similarity with high-performing features. But in non-sibilant fricatives, we find greater dissimilarity between natural and Tacotron voices.

Keywords: obstruents, Tacotron, WaveNet, WaveGAN, support-vector-machines

1. INTRODUCTION

In days of formant synthesis, speech science and synthesis technology enjoyed a reciprocal relationship. However, synthetic speech thus produced was more unintelligible than human speech, and its findings often did not extend to human language [1]. Although intelligibility improved with statistical parametric synthesis, it brought a roboticity or “unnaturalness" to the speech output [2]. But today, with neural Text-To-Speech (TTS), high-quality, natural-sounding synthetic speech has become quite accessible. Malisz et al. [3] show that data-driven TTS is now more realistic, highly intelligible, and perceptually closer to human speech. They argue that conclusions drawn on speech produced by modern TTS may generalize better to the human voice. It may thus be time, once again, for a closer exchange between speech science and synthesis technology.

Using synthetic speech as a research tool has attracted many phoneticians. Vector space embeddings in WaveNet synthesizers have been used to analyze prosodic patterns in Lombard speech, proposing new methodologies for prosodic research [4]. Controllable architectures like WaveBender [5] have reinstated the scope of feature-based phonetics within modern TTS. Additionally, using conversational synthetic speech has provided greater flexibility over controlled variation in determining paralinguistic characteristics [6]. Speech science has also increasingly contributed *to* TTS. Forensic investigation of speech spoofed through TTS provided important insights into human-likeness [7]. Segmental analysis of obstruents using acoustic-phonetic features in [8], have revealed system-specific weaknesses in WaveNet TTS. Furthermore, evaluation paradigms have borrowed actively from phonetics [9].

However, the majority of these papers involve prosodic explorations of synthetic speech. Phonemic contrast, which took centre stage in early formant synthesizers, has been overlooked in neural TTS. If contrastive trends in synthetic speech can generalize well to human speech, dependence on data collection can be immensely reduced. Phonemic contrasts could be synthesized in various positional, vocalic and cluster contexts, including speaking styles. Additionally, the success of multi-speaker, and accented TTS [10] can ensure the necessary diversity required to understand acoustic invariance.

To investigate whether phonemic contrast is maintained in neural TTS in the same way as in a human voice, this paper provides **place classification** of English fricatives as a targeted test case. Recently, voiceless fricatives have been shown to deviate from a natural voice in neural TTS [8]. However, despite deviation, if contrast

is encoded in the same parameters as in the natural voice, then TTS voices may be suitable tools for speech science research. Conversely, if phonemic contrast is indexed by divergent trends in TTS voices, then generalization may become more difficult. Additionally, unexpected acoustic detail may enforce a cognitive load condition and increase reliance on lexical cues [11]. This can cause greater problems for non-native listeners [12].

We compare **feature importance** for place classification between a natural voice, and two TTS systems: Tacotron WaveGAN and Tacotron WaveNet. The corpus for our analysis comes from the recently extended Blizzard Challenge 2013 (BC-2013) [13]. Features are analyzed both individually and as an ensemble, using Support Vector Machines (SVM)s for place classification between contrastive pairs (e.g, /s-ʃ/). Top-ranked features in each contrastive pair are compared. Through this analysis, we explore whether contrastive trends in Tacotron voices are similar to the natural voice in terms of top-ranked features. We find that sibilant fricatives in both the voices show very similar trends to the natural voice, while non-sibilant fricatives need further exploration.

2. EXPERIMENTAL SETUP

2.1. Dataset and features

2.1.1. Description of the dataset

The Blizzard Challenge 2013 [14] provides single-speaker read-speech data, collected from a female speaker of standard American English. Recently, the original 2013 challenge was extended [13] to include modernTTS voices. For the present analysis, we selected Tacotron WaveGAN (system R) , Tacotron WaveNet (system Z), and the natural voice. While both systems use Tacotron [15] as the acoustic model, waveform generation is handled by WaveGAN [16] and WaveNet [17]. With this dataset of a 100 sentences produced by each of the three voices (300 sentences total), we analyzed place classification in the following contrastive pairs of fricatives (*voiceless:*/f-θ/, /θ-s/, /s-ʃ/, *voiced:*/f-ð/, /ð-z/). Obstruents occuring in unstressed and consonant cluster positions were removed, hence the voiced sibilants /z-ʒ/ were excluded.

2.1.2. Feature extraction

A set of acoustic-phonetic features were extracted from these obstruent consonants in natural and each of the Tacotron voices. In English fricatives, place is determined largely by static spectral characteristics such as spectral tilt, shape, center frequency [18, 19]. Formant transitions have also frequently been discussed [20], although with inconsistent findings.

From the **vocalic** portion of the CV syllable, we extracted: *vowel duration*; *vocalic RMS amplitude*; *Formant values (F1-F3) at onset and midpoint*; *relative amplitude of F3*; *vocalic spectral tilt* and *Delta F1-F3*.

From the **consonantal** portion of the syllable, we extracted: *consonant duration*; *noise duration*; *RMS amplitude*; *spectral tilt*; *spectral shape*; *peak frequency*; *peak amplitude*; and *dynamic amplitude*. All these features have been compared across all English obstruents in [21] for various contrastive strength. They have also been used in analysing obstruents of WaveNet vocoders in [8].

2.2. Feature importance: individual and ensemble

Feature importance provides important links between speech stimuli and phonemic perception. Feature importance was analyzed in R using a two-pronged SVM based classification, and complemented with statistical analysis. This approach is greatly inspired by Styler's [22] work on nasality feature analysis. All features were pre-processed by median-normalization and scaling, and imbalances are removed by undersampling the abundant phonemic class. SVMs use a 5-fold cross validation, and a radial-basis kernel throughout.

2.2.1. Single-feature model (SF model)

In this step, place classification is performed using only 1 feature at a time. This reveals the independent predictive capacity of the predictor, before it is cast into a high-dimensional space.

Feature importance is calculated simply by checking the accuracies as returned by each feature. Next, in addition to accuracy, top-ranked features identified above were also checked for their statistical signficance in the dataset, using the Kruskal-Wallis rank sum test. If complementary information is provided by both the dataset and SVM, then it is crucial for the feature to be properly reproduced in Tacotron voices.

Similarity between natural and Tacotron voices is manually compared, using predictor accuracies, statistical significance scores, and median differences between phonemes.

2.2.2. Full-feature ensemble model (FF model)

In this step, we provide all 20 features to the SVM as an ensemble. This resembles real-world scenarios of speech perception, where all the contrastive information is available to the listener. This step investigates whether the same *feature combination* is shared between natural and Tacotron voices.

First, we calculate the accuracies and F1 score for each contrastive pair, for natural and both Tacotron voices. Then, ***feature importance*** is obtained using the DALEX [23] library, which ranks features based on the 1-AUC score. Thus, we obtain the top 7 (1/3rd of 20) most performant features for every contrastive pair in natural and Tacotron voices. Next, we compute a ***similarity score*** based on the Sørensen's-Dice coefficient. This calculates the magnitude of overlap among the top-7 features between natural voice and each Tacotron system. A high-score indicates greater overlap between high-performing feature combinations.

3. RESULTS

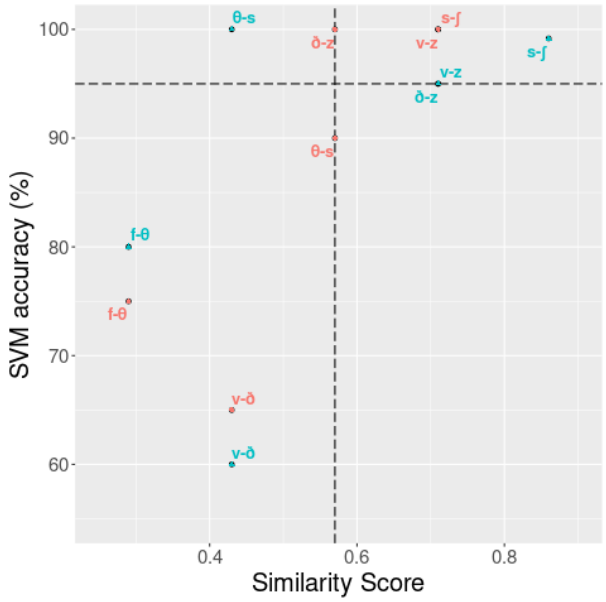

Figure 1: Relationship between accuracy and feature similarity with natural voice. Labels for contrastive pair presented for Tacotron WaveGAN **(R)** and Tacotron WaveNet **(Z)**. Dashed lines show medians on every axis.

In this section, we present the results of place classification in sibilants, between sibilant-non-sibilants, and between non-sibilants. First, we report single-feature accuracies and statistical significance of top-ranked features in each voice using the SF model. These values are presented comparatively between natural and each Tacotron voice. Then, using the FF model, we report the accuracies of place classification, and similarity between top-7 features of natural and each Tacotron voice. Figure 1 displays the relationship similarity and accuracy obtained using the FF model.

3.1. Within-sibilants

***Voiceless* /s-ʃ/**: In sibilant place identification, both systems Z and R show remarkable similarity to natural voice. Through the single-feature analysis, we find **spectral shape** to be the most powerful predictor of sibilant place, with its standalone accuracy reaching 100% in natural voice, as well as in both the synthetic voices . In natural voice, the spectral shape of /ʃ/ is higher by +14.12 dB compared to /s/, [$\chi^2(1) = 86.26$, p-val < 0.0001]. The distinction of higher spectral shape in /ʃ/ is consistently maintained in system Z (median difference +13.29 dB), and in R by +12.83 dB respectively. This means that spectral shape is a robust indicator of within-sibilant place in natural voice, and gains a similar prominence in synthetic voices as well.

Using the FF model, we find near-perfect accuracy for place classification between sibilant fricatives. Figure 1 shows their clusters in the top-right quadrant, displaying high similarity with natural, as well as above-median accuracy. System Z scores 0.99, and R 0.71, on the Sørensen's distance metric. This indicates a very strong match in the features selected by the natural voice, and those by both the Tacotron systems.

Therefore, in addition to spectral shape, other high-performing features are also consistent in natural as well as synthetic voices for place identification in sibilants.

3.2. Sibilant vs non-sibilant

***Voiceless* /θ-s/**: Here, both Tacotron voices show high accuracy of classification, but share only a few features with natural voice. Through a single-feature analysis, we find a source of divergence between systems on the basis of **spectral tilt**. In Tacotron voices consonantal spectral tilt shows high standalone accuracy [Accuracy = 100%, $\chi^2(1) = 8.75$, p-val < 0.0001]. In natural voice, although spectral tilt for /s/ is significantly higher (p-val $<$ 0.001), it is a weak predictor for SVM, with only 55% individual accuracy. Using the FF model, system Z shows a 100% accuracy, with Sørensen's score of only 0.43. This means that the feature combination is not similar to natural, but a divergent

model gives comparable accuracy of classification.

Therefore, features in /θ-s/ place identification are more divergent from the natural voice, compared to within-sibilant place identification. However, significantly higher spectral tilts in /s/ means that the synthetic data, at least, preserves contrastive trends like the natural voice.

***Voiced* /ð-z/**: For the voiced sibilant-non-sibilant contrast /ð-z/, we see that Tacotron voices show greater similarity with natural in terms of features. Through a single feature analysis, we find that consonantal **spectral tilt**, **dynamic amplitude** and **peak amplitude** are robust across natural voice, as well as both the Tacotron voices. Here, dynamic amplitude between /ð/ and /z/ show statistically significant differences, [$\chi^2(1) = 12.79$, p-val < 0.001], with median dynamic amplitude of /z/ being +30.5 dB higher than θ. Comparable trends can be seen in Tacotron voices as well. Next, a high Sørensen's score of 0.71 and 0.57 in systems Z and R, indicate a majority of features in the full-model SVM are shared with natural.

Therefore, we can see that place classification between voiced sibilant-non-sibilant pair /ð-z/ shares great similarity with natural.

3.3. Within non-sibilants

***Voiceless* /f-θ/**: For this contrast, we can see very divergent trends. Figure 1 displays them in the bottom-left quadrant, indicating a low similarity with natural, as well as below-median accuracy. Next, we find that the contrast is very weak and confusable in natural voice, with only a few features showing modest statistical differences (p-val < 0.01). An SF model reveals consonantal RMS amplitude as the strongest predictor in natural voice [Accuracy 75%, p-val = ns]. On the other hand, both the Tacotron voices show high standalone accuracies for noise duration (System Z: [Accuracy 75%, $\chi^2(1) = 10.30$, p-val < 0.01], System R: [Accuracy 80%, $\chi^2(1) = 9.08$, p-val < 0.01]), which is conspicuously absent as a high-ranking feature in natural voice.

Using the FF model, we find that accuracy in natural voice is only 55%. This means that accuracy drops for natural when more features are added, and feature selection must be much more careful. Conversely, in Tacotron voices, accuracies are higher than chance in the FF model, even though it is lower than median.This means that features selected by SVM for /f-θ/ are different from natural.

***Voiced* /v-ð/**: Similarly in the voiced condition, both Tacotron voices show a low Sørensen's score, indicating only a minimal overlap with natural voice in the full-model SVM. As observed in the voiceless condition, noise duration is an important predictor of place in Tacotron voices [Accuracy 70%, $\chi^2(1) = 3.28$, p-val > 0.05] in system Z, and [Accuracy 70%, $\chi^2(1) = 3.28$, ns] in system R.

Therefore, in place identification between non-sibilants, i.e, /f-θ/, and /v-ð/, we see strongly divergent trends from natural voice in terms of feature importance, and classification results. So, a much deeper exploration is needed before their use can be recommended for speech sciences.

4. DISCUSSION & CONCLUSION

In this paper, we compared feature importance in place classification of obstruents between a natural voice and two Tacotron voices. Using a single-feature SVM model, we compared individual predictors that were most informative in each voice. Then, using a full-feature model in SVM, we compared feature importance between the voices using a similarity metric. We found that sibilant fricatives produced by both Tacotron vioces followed similar contrastive trends to the natural voice. However, those involving non-sibilant fricatives differ from the natural voice. A clear separation can be seen in Figure 1. Through these findings we can recommend that studies conducted on synthetically produced sibilant contrasts may be generalized to human speech.

Our findings are consistent with previous work on sibilants, w.r.t. the resilience of their features in impaired listening conditions [24]. Similarly, in non-sibilants, inadequate feature distinctions [25], and place confusions between /f-θ/ have been reported [26] for natural speech. Both Tacotron voices appear to overcome this inadequacy, basing classification primarily on longer noise durations in the labiodental /f/ and /v/. However, previous evidence suggests that spectral features are more informative of place in non-sibilants, instead of durations [19]. This indicates that insufficient contrastive information in the natural voice may cause cue trading in Tacotron voices to maintain contrast. This introduces an acoustic detail which may not be expected by listeners. Whether this impacts cognitive load, needs to be substantiated with carefully controlled perception experiments. Similarly, while this study shows featural importance through SVM and statistical tests, the perceptual relevance of these features should be established through subjective category perception tests, and supported further through cross-linguistic explorations.

5. ACKNOWLEDGEMENTS

This work has the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, the ADAPT Centre (Grant 13/RC/2106), and a Google Faculty Award.

6. REFERENCES

[1] S. A. Duffy and D. B. Pisoni, "Comprehension of synthetic speech produced by rule: A review and theoretical interpretation," *Language and Speech*, vol. 35, no. 4, pp. 351–389, 1992.

[2] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[3] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *International Congress of Phonetic Sciencesnces ICPhS 2019 5-9 August 2019, Melbourne, Australia Melbourne Convention and Exhibition Centre*, 2019.

[4] J. Šimko, M. Vainio, A. Suni *et al.*, "Analysis of speech prosody using wavenet embeddings: The lombard effect," in *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[5] G. T. D. Beck, U. Wennberg, Z. Malisz, and G. E. Henter, "Wavebender gan: An architecture for phonetically meaningful speech manipulation," *arXiv preprint arXiv:2202.10973*, 2022.

[6] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[7] C. Kirchhuebel and G. Brown, "Spoofed speech from the perspective of a forensic phonetician," *Interspeech 2022*, 2022.

[8] A. Pandey, S. L. Maguer, J. Carson-Berndsen, and N. Harte, "Production characteristics of obstruents in wavenet and older tts systems," in *INTERSPEECH*, 2022.

[9] E. Gutierrez, P. O. Gallegos, and C. Lai, "Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm," *ArXiv*, vol. abs/2107.02527, 2021.

[10] H. B. Moss, V. Aggarwal, N. Prateek, J. González, and R. Barra-Chicote, "Boffin tts: Few-shot speaker adaptation by bayesian optimization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7639–7643.

[11] S. L. Mattys and L. Wiget, "Effects of cognitive load on speech recognition," *Journal of memory and Language*, vol. 65, no. 2, pp. 145–160, 2011.

[12] S. L. Mattys, L. M. Carroll, C. K. Li, and S. L. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers," *Speech communication*, vol. 52, no. 11-12, pp. 887–899, 2010.

[13] S. L. Maguer, S. King, and N. Harte, "Back to the future: Extending the blizzard challenge 2013," in *Interspeech*, 2022.

[14] S. King and V. Karaiskos, "The blizzard challenge 2013," in *The Blizzard Challenge Workshop*, 2013, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.

[15] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *ArXiv*, vol. abs/1703.10135, 2017.

[16] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[17] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, 2016.

[18] A. Jongman, "Duration of frication noise required for identification of english fricatives," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1718–1725, 1989.

[19] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[20] A. Wagner, M. Ernestus, and A. Cutler, "Formant transitions in fricative identification: The role of native fricative inventory," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 2267–2277, 2006.

[21] C. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," *PhD dissertation, University of Kansas, Lawrence*, 2020.

[22] W. Styler, "On the acoustical and perceptual features of vowel nasality," Ph.D. dissertation, University of Colorado at Boulder, 2015.

[23] P. Biecek, "Dalex: explainers for complex predictive models in r," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3245–3249, 2018.

[24] J. Meyer, L. Dentel, and F. Meunier, "Speech recognition in natural background noise," *PloS one*, vol. 8, no. 11, p. e79279, 2013.

[25] S. J. Behrens and S. E. Blumstein, "Acoustic characteristics of english voiceless fricatives: A descriptive analysis," *Journal of Phonetics*, vol. 16, no. 3, pp. 295–298, 1988.

[26] K. Maniwa, A. Jongman, and T. Wade, "Acoustic characteristics of clearly spoken english fricatives," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3962–3973, 2009.