Formants in text-to-speech systems - comparing TTS voices of Blizzard Challenge 2013

Ayushi Pandey[1], Sébastien Le Maguer[1], Julie Carson-Berndsen[2], Naomi Harte[1]
[1] Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin
[2] ADAPT Centre, School of Computer Science, University College Dublin, Ireland
pandeya@tcd.ie, lemagues@tcd.ie, julie.berndsen@ucd.ie, nharte@tcd.ie

Abstract

Modern trends in synthesis evaluation attempt to capture finer aspects of the human experience of synthetic speech. However, a feature-based exploration of the synthetic speech signal, especially in comparison with human speech signal is still missing from the discussion.

Using the Blizzard Challenge 2013 speech database, we propose an analysis of synthetic speech systems using two types of acoustic-phonetic features: F1-F2 vowel space characteristics; and formant transitions in stop-vowel (CV) sequences. We observe that a subset of vowel space characteristics (within-category dispersion among vowels, in particular) can predict an above-chance correlation with MOS scores of Similarity to the natural voice (66%) and perceived Naturalness (64%).

F2 transitions were not found to be as clearly correlated with perceived MOS scores. However, system-specific characteristics were still revealed, when HMM-based systems showed a statistically significant raising of the onset of the F2. Based on these results, we believe that a clearer concept of acoustic-phonetic correlates of naturalness may inform future approaches to the analysis of synthetic speech, and help relate perceptual responses to phonetic attributes in the signal.

Introduction

Speech synthesizers have long been analyzed as models of speech production. Some of the earliest known works included the creation of pattern-playback synthesis, which began to be used as a model of perceptual listening (Delattre et al., 1955). This feedback between perception and synthesis can be seen even in recent times. Synthetic speech is used for studying perception of targeted linguistic phenomena (Story & Bunton, 2010), and on the other hand, articulatory data has been used to supplement data-driven models for increased naturalness (Csapó et al., 2021).

The *evaluation* of present day data-driven TTS systems, (unit-selection, HMM and Hybrid systems) is not viewed as a potential case for perceptual or production modelling of speech. Currently, TTS evaluation has been dominated by subjective listening tests, which collect mean opinion scores (MOS) on a set of perceivable attributes of synthesized speech. At the same time, since the MOS scores are obtained by averaging multiple listeners' responses over several utterances, its diagnostic capabilities (Wagner & Betz, 2017) , that is, locations of distortions are quite limited.

As an alternative, some improved evaluation designs (Gutierrez et al., 2021) explicitly request their listening participants to mark where the distortion in the signal can be observed. Complementary to this, behavioural metrics such as pupillometry (Govender et al., 2019), EEG based studies (Antons et al., 2013; Parmonangan et al., 2019) have tapped into the subconscious decision making of human participants, and correlate these responses (e.g, pupil dilation, neuronal activity) with MOS scores. To overcome the reliance on subjective methods, objective methods of MOS score prediction have achieved good correlations with human responses (Lo et al., 2019; Mittag & Möller, 2020).

The automatic prediction of MOS enables feedback during development stages of TTS, while redesigning the subjective evaluation adds variety to the sources through which users' responses can be obtained. While each of these designs hold enormous potential, they do not discuss the inherent properties of the speech signal itself, especially in comparison with natural speech. Acoustic-phonetic characteristics of speech are easily extractable, and can offer a wide array of features that can be used to offer feedback during development stages of TTS. At the same time, they can make the evaluation more diagnostic, by identifying those features which strongly differ from natural.

In our previous work (Pandey et al., 2021), we demonstrated that contrastive properties of obstruent consonants can be used to compare quality of systems of Blizzard Challenge 2013. In this paper, we compare the properties of vowels in synthetic speech, using the steady-state and transitional cues from their formants. Speech generated by HMM, Hybrid and Unit-selection techniques in the Blizzard Challenge 2013 have been compared against the natural voice using a set of features derived from vowel formants.

Feature values obtained per system, have been correlated against the subjective MOS score for that system. The analysis has been supplemented by linear mixed effects models, to investigate further the source of deviation from natural voice. Through this analysis, we aim not only to describe the properties of synthetic speech, but also to correlate these results with the obtained MOS scores, and situate this approach within the speech synthesis evaluation paradigm.

Formants and formant transitions

Vowel formants, or the peaks in the acoustic spectrum corresponding to the resonances in the vocal tract, provide important distinctive features for the perception of vowel quality. The vowel space describes the location of a vowel in an X-Y plane, where the vertical axis represents the first formant (F1) and the horizontal axis represents the second formant (F2).

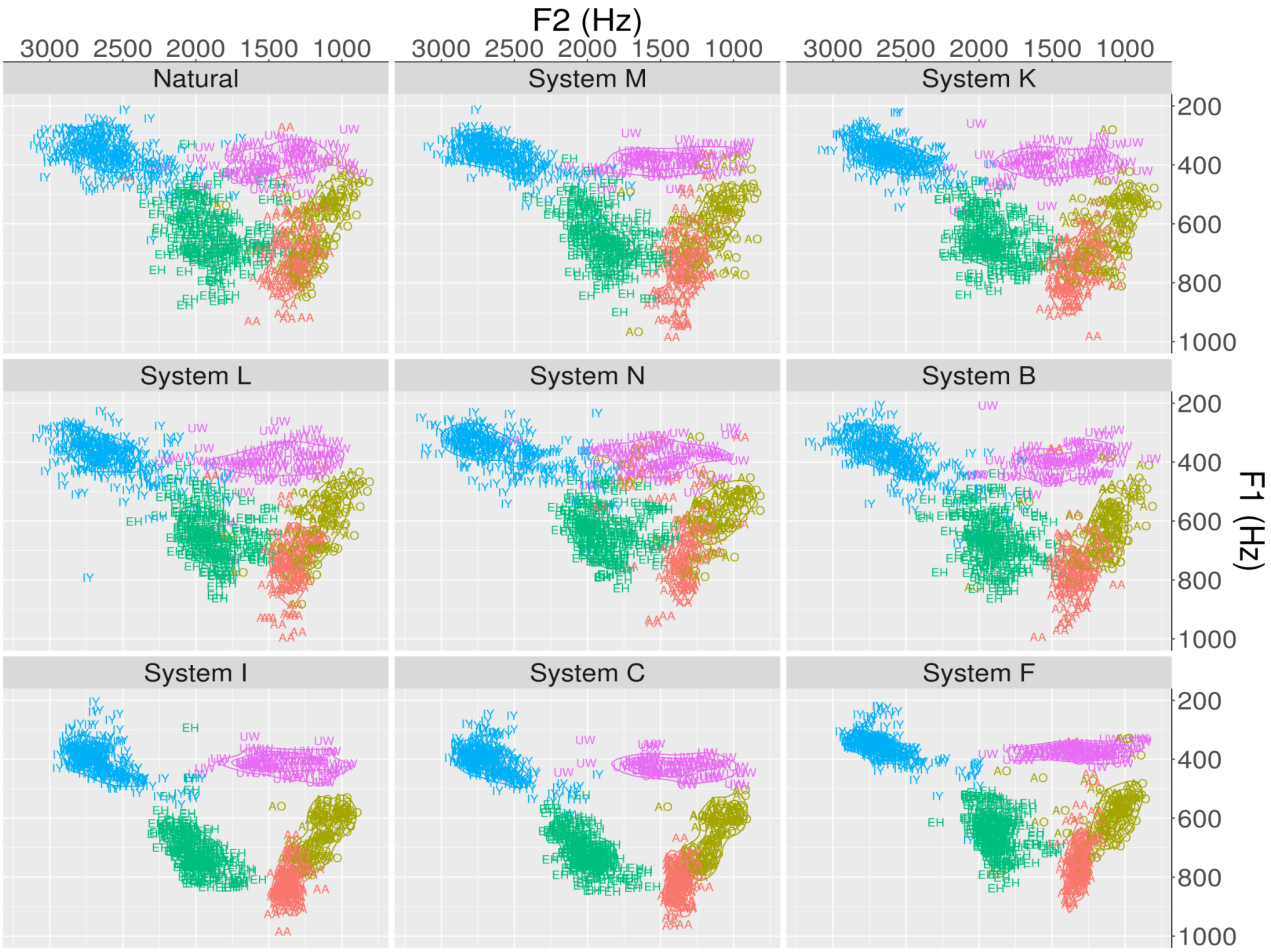

Figure 1: Vowel formant plots arranged in rows by system-family, and in columns by quality. The first row has Natural, and 2 Hybrid systems (M, K). The second row has 3 Unit-selection systems (L, N, B). The last row has 3 HMM systems (I, C, F). HMM systems can be clearly seen to cluster instances of vowels around respective vowel means.

Vowel spaces have been traditionally examined for applications such as increased vowel space area and F2-F1 measurements were found to be close correlates of intelligible speech. More specific to our case, vowel spaces have also been used to evaluate intelligibility in foreign directed speech (Bradlow et al., 1996), pronunciation assessment (Minematsu et al., 2006) and training (Wik & Escribano, 2009), developmental stages of vowel production (Vorperian & Kent, 2007). Particularly relevant here, Chen et al. (2010) report a set of vowel space metrics for non-native speech that correlate with perceived intelligibility.

While vowel-space characteristics provide measurements for the steady-state of the vowel, the dynamic movement of the formant within the vowel can inform the concatenative strength of the synthesized speech. Formant transitions are considered important predictors of consonantal place of articulation (Nearey & Shammass, 1987; Sussman et al., 1991; McCarthy, 2019;). A linear regression fitted between formant values at the midpoint (F2 midpoint) and the initial portion (F2 onset) of the vowel, displays a strongly linear relationship (Sussman et al., 1991). The slope, which determines the extent of co-articulation between the stop consonant and the vowel, was shown to be steepest for bilabial consonants, intermediate for velars and flattest for alveolar stops. However, vocalic context also strongly influences this relationship.

In this paper, we correlate the MOS scores obtained in the Blizzard Challenge, with a) a set of vowel space characteristics and, b) stop-vowel formant transition patterns. Next, we analyze the behaviour of systems with respect to each of the features for which the correlation was found significant.

Experimental Setup

Description of the dataset

The Blizzard Challenge (BC) is an international task designed to independently compare state of the art corpus-based speech synthesis systems[1]. Using a common training dataset, participants must all submit identical sentences as output by their synthesis system. These are then evaluated with extensive listener tests. The expressive speech corpus proposed for the 2013 edition was extracted from audiobooks dictated by an American English female speaker. Of the 10 entries, 5 used parametric HMM-based techniques (systems C,F,H,I,P), 3 used Unit-Selection (systems B,L,N), and 2 used Hybrid method (systems K,M) for synthesis.

The dataset used for analysis was a parallel dataset, composed of 100 sentences submitted by each system. Based on subjective evaluation tests on a subset of this dataset, system M was rated as the most natural and most similar to the original speaker. Systems K, I, L, and C were the next most highly ranked. P was the lowest ranked. The diversity of its TTS generation

[1] https://www.synsig.org/index.php/Blizzard_Challenge

techniques, single-speaker dataset, especially in English, and accompanying MOS scores, made BC 2013 an ideal corpus to compare characteristics of synthetic speech systems.

Formant extraction

First, phoneme-level segmentation was conducted for each systems using the Montreal Forced Aligner (McAuliffe et al., 2017). A hand analysis, supplemented by a global analysis of vowel durations across the systems validated the accuracy of this segmentation.

Using Praat (Boersma & Weenink, 2018), formant values (F1-F2) were extracted for 13 American English vowels : /ɪ, i, æ, e, ɛ, a, ə, ɚ, ʌ, ɔ, o, ʊ, u/ at 20% (onset) and 50% (midpoint) of the duration of the vowels. The optimal ceiling value for each vowel was determined by the Escudero optimization procedure (Escudero et al., 2009), where the appropriate ceiling frequency minimized the within-vowel variance in the dataset. The window size was set to 25ms with default values used for all other parameters.

Results

Analysis of the vowel space

Chen et al. (2010) identified seven characteristics of vowel space which were closely related to intelligibility of non-native speech. These characteristics are: *vowel space area*; *overall dispersion* of each instance; *within-category dispersion* of each vowel type; *F1 range*; *F2 range* as well as the *F1-F2 distance* for /i/ and /a/.

In this paper, we explore how these factors might relate to synthetic speech. To achieve this, we computed the Spearman's correlation between each of these characteristics and the BC MOS values for naturalness and similarity. The results are presented in Table 1.

The results show that, at a p-value of 0.05, only within-category (W-C) vowel dispersion and F1 range were significant.

As exemplified in Figure 1, F1-F2 vowel space for the 5 vowels /i, æ, e, a, ɔ, u/ showed that the HMM synthesis consistently lead to a more contracted vowel space. Therefore, we conducted a statistical analysis of vowel dispersion per type of synthesis to further investigate the previous observation. In order to perform statistical evaluation on both F1 and F2 individually, we used a modified euclidean distance used in Chen et al., (2010). This was done to observe system-specific behaviour on vowel-dispersion along each of the F1 and F2 dimensions. Using the distance for each specfic vowel, a linear mixed effects model (Kuznetsova et al., 2017) was applied to analyze the dispersion. The system type (*HMM*, *Hybrid* and *Unit-Selection*) as well as the vowel type *Front* or *Back* are considered as fixed effects. The utterance index is considered as a random effect. These results are presented in Table 2.

These results show that the natural speech is producing more dispersed F1 values than any type of synthetic speech, as exemplified in Figure 1.

Table 1. Spearman's correlation of each feature with MOS ratings for Similarity and Naturalness. Values in bold denote significant correlations at p-val = 0.05

Feature	Similarity	Naturalness
Area	0.05	0.05
Dispersion	0.25	0.24
W-C dispersion	**0.66**	**0.64**
F1 range	**0.62**	**0.73**
F2 range	0.6	0.55
F2 - F1 (/a/)	-0.08	-0.07
F2 - F1 (/i/)	-0.04	0.07

In addition, the dispersion of F1 values for HMM synthesis is significantly less than the other types of synthesis. HMM synthesis produces even less dispersed F2 values. This is a consequence of the well-known over-smoothing effect, due to the statistical nature of parametrical synthesis (Zen et al., 2009) , which leads to a reduction of the variability of the generated speech.

From these results, we can conclude that HMM synthesis produces more distinct vowels than any type of synthesis, but it fails to generate more extreme or more nuanced vowel instances which may be fundamental to naturalness. Additionally, we observed a significant effect of vowel-type on dispersion, where front vowels were seen to lower dispersion (β_{front}= -5.933, 95% CI [-9.71, -2.16], p<0.001). The relationship of this particular finding to perceived naturalness requires more analysis.

Analysis of locus equations

The locus equations are an effective representation of the extent of coarticulation between the preceding consonant and a vowel (McCarthy, 2019; Sussman et al., 1991). These equations compute the trajectory slope of the second formant (F2) between the onset and the mid-point of the vowel.

We focus on three type of consonants: bilabial (/p, b/), alveolars (/t, d/) and velars (/k, g/). Based on the results of previous studies using these equations (Sussman et al., 1991; McCarthy, 2019), we expect to find these patterns in natural speech: velars followed by front vowels have the steepest slope; velars followed by back vowels bilababial produces a moderately steep slope; the slope is the flattest when the consonant is a bilabial.

In the BC-2013 corpus, the overall pattern of coarticulation is not canonical (labial > velar > alveolar) as reported by Sussman et al. (1991). For natural voice, the fitted regression lines of alveolars and labials are nearly parallel, whereas the velar shows a sharper slope across all vowel contexts.

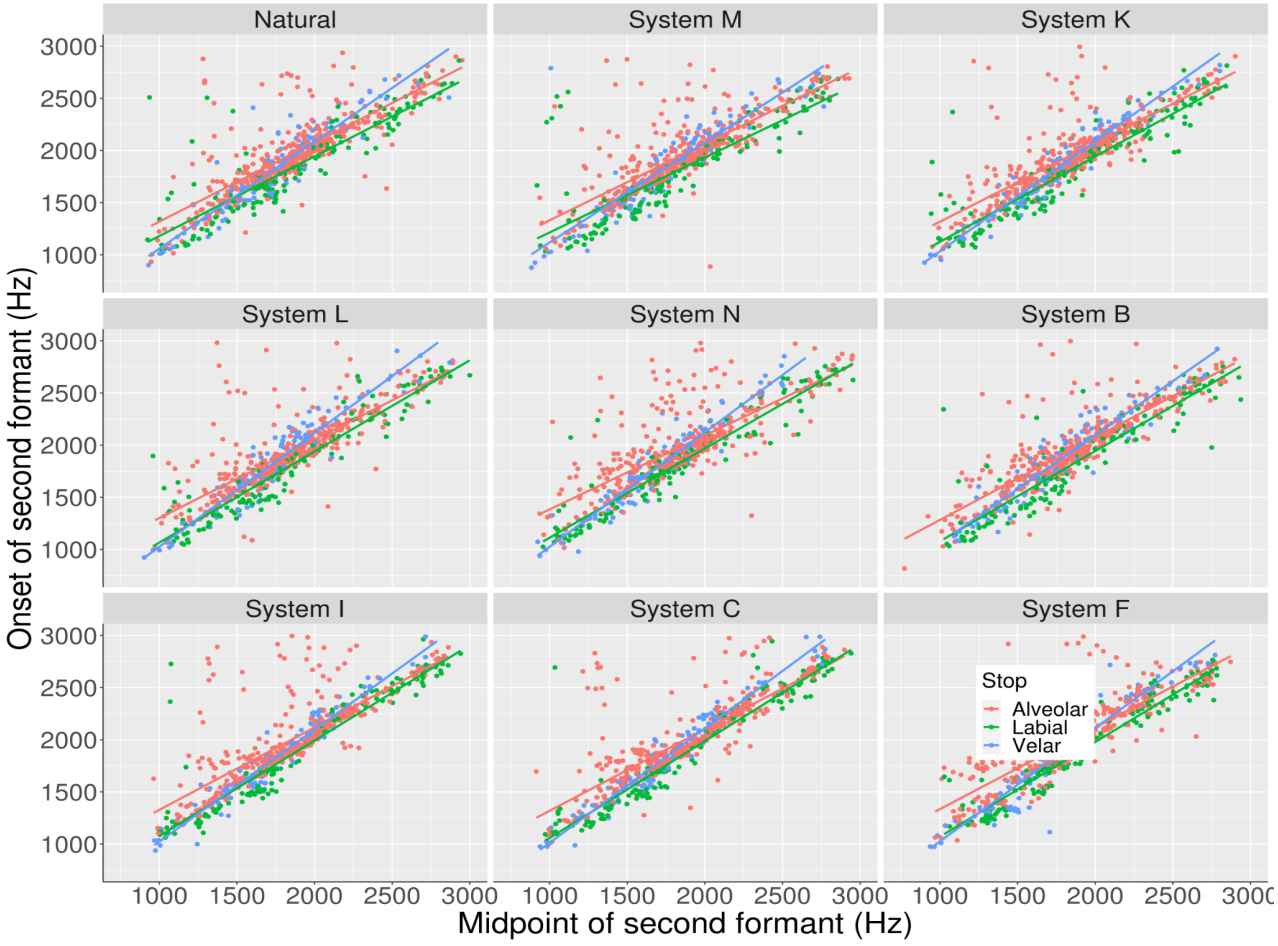

Figure 2: Relationship between the onset and midpoint of vowels in different consonantal contexts displayed for each system. Systems are arranged in rows by system-family, and in columns by quality. The best-performing Systems M and K can be seen to resemble the transition patterns of natural voice.

The consistent sharp slope observed in the velar contexts can be attributed to the higher number of front-vowels in the velar context (front vowels:58, back vowels: 31).

Table 2: Linear mixed-effects model fits on F1 and F2 dispersion. HMM systems show dispersion to be greatly reduced in both F1 and F2.

	System	β **coefficient**	**95% CI**
F1	Unit-Sel	-7.38	[-11.44, -3.31]
	Hybrid	-11.58	[-15.90, -7.26]
	HMM	-29.12	[-33.06, -25.18]
F2	Unit-Sel	-7.49	[-13.80, -1.17]
	Hybrid	-9.80	[-16.50, -3.10]
	HMM	-40.18	[-46.29, -34.06]

To identify if corpus characteristics were responsible for this, we repeated the analysis on the female speakers of the TIMIT corpus (Garofolo et al., 1993). This data shares characteristics with BC-2013, as it is also read speech. Since the replication of our results on a hand-annotated corpus validated our methods, a deeper discussion on these trends is not the scope of this paper.

The most important observation is that Systems M and K, the highest scoring Hybrid systems, closely resembles the co-articulation patterns of natural voice. This can be seen in Figure 2.

Then, we conducted a linear mixed effects modeling to analyze the influence of system-type on the relationship between F2-onset and F2 midpoint. To do so, we analyze the onset value with the mid point value as a main effect. The system family (HMM, Unit Selection or Hybrid) is set as fixed effect. The only statistically significant results found is that HMM systems raise the onset value (= 36.78, 95% CI [9.39, 64.17], $p < 0.001$) when the consonant is an alveolar. As described in (McCarthy, 2019), back vowels for alveolars raise the onset of the second formant because of the reduced length back cavity.

However, in the case of HMM contexts, we found the central vowel /ə/ to increase the F2 onset (mean increase: 79 Hz). While all vowels show statistical averaging and clustering, the /ə/ assumes a particularly backed position. One possible reason can be that although the /ə/ is an unstressed medial vowel, its stressed counterpart /ʌ/, (the vowel in “but", and “gut")

assumes a further back position. Since the specification of linguistic stress is made in the lexicon of HMMs based modeling (Mametani et al., 2019; Watts et al., 2010), it is possible that this distinction between stressed/unstressed was normalized.

Conclusion

In this paper, we propose the use of formant-driven features in the analysis of synthetic speech. The central finding of this paper was that HMM systems cluster the vowel space more densely around their within-vowel means. Secondly, within-category vowel dispersion patterns were correlated at an above-chance level with the Blizzard Challenge 2013 MOS scores, both for naturalness and similarity. We also found that on the basis of F2 transitions, the correlation-based ranking did not display clear significance.

We plan to conduct a full-scale exploration of acoustic-phonetic features to include a variety of phonological classes, and enhance the dataset to include multi-speaker and multilingual TTS voices. We also plan to extend this analysis onto modern neural voices such as WaveNET and WaveGAN vocoders.

References

Antons, J.-N., Schleicher, R., Arndt, S., Moller, S., Porbadnigk, A. K., & Curio, G. (2012). Analyzing speech quality perception using electroencephalography. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 721–731.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3), 255.

Chen, L., Evanini, K., & Sun, X. (2010). Assessment of non-native speech using vowel space characteristics. *IEEE Spoken Language Technology Workshop*, 139–144.

Csapó, T. G., Tóth, L., Gosztolya, G., & Markó, A. (2021). Speech Synthesis from Text and Ultrasound Tongue Image-based Articulatory Input. *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 31–36. https://doi.org/10.21437/SSW.2021-6

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773.

Delattre, P., Cooper, F. S., Liberman, A. M., & Gerstman, L. (1954). Acoustic Loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *26*(1), 137–137.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393.

Govender, A., Valentini-Botinhao, C., & King, S. (2019). Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis. *Speech Synthesis Workshop (SSW)*.

Gutierrez, E., Oplustil-Gallegos, P., & Lai, C. (2021). Location, Location: Enhancing the Evaluation of Text-to-Speech Synthesis Using the Rapid Prosody Transcription Paradigm. *ArXiv Preprint ArXiv:2107.02527*.

King, S., & Karaiskos, V. (2013). The Blizzard Challenge 2013. *The Blizzard Challenge Workshop*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019). MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. *International Conference on Speech Communication and Technology (Interspeech)*, 1541–1545. https://doi.org/10.21437/Interspeech.2019-2003

McCarthy, D. T. P. D. (2019). *The acoustics of place of articulation in English plosives* [PhD Thesis]. Newcastle University.

Minematsu, N., Asakawa, S., & Hirose, K. (2006). Structural representation of the pronunciation and its use for CALL. *IEEE Spoken Language Technology Workshop (SLTW)*, 126–129.

Mittag, G., & Möller, S. (2020). Deep Learning Based Assessment of Synthetic Speech Naturalness. *International Conference on Speech Communication and Technology (Interspeech)*, 1748–1752. https://doi.org/10.21437/Interspeech.2020-2382

Nearey, T. M., & Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, *15*(4), 17–24.

Pandey, A., Le Maguer, S., Carson-Berndsen, J., & Harte, N. (2021). Mind your p's and k's–Comparing obstruents across TTS voices of the Blizzard Challenge 2013. *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 166–171.

Parmonangan, I. H., Tanaka, H., Sakti, S., Takamichi, S., & Nakamura, S. (2019). Speech Quality Evaluation of Synthesized Japanese Speech using EEG. *International Conference on Speech Communication and Technology (Interspeech)*, 1228–1232.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *Journal of the Acoustical Society of America*, *121*(5), 3044.

Story, B. H., & Bunton, K. (2010). *Relation of vocal tract shape, formant transitions, and stop consonant identification*.

Sussman, H. M., Fruchter, D., & Cable, A. (1995). Locus equations derived from compensatory articulation. *The Journal of the Acoustical Society of America*, *97*(5), 3112–3124.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*.

Wik, P., & Escribano, D. L. (2009). Say 'Aaaaa' interactive vowel practice for second language learning. *Workshop on Speech and Language Technology in Education (SLaTE)*.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064.