INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Variation in Perceptual Sensitivity and Compensation for Coarticulation Across Adult and Child Naturally-produced and TTS Voices

Aleese Block, Michelle Cohn, and Georgia Zellou

Phonetics Lab, University of California, Davis, USA
{asblock, mdcohn, gzellou}@ucdavis.edu

Abstract

The current study explores whether perception of coarticulatory vowel nasalization differs by speaker age (adult vs. child) and type of voice (naturally produced vs. synthetic speech). Listeners completed a 4IAX discrimination task between pairs containing acoustically identical (both nasal or oral) vowels and acoustically distinct (one oral, one nasal) vowels. Vowels occurred in either the same consonant contexts or different contexts across pairs. Listeners completed the experiment with either naturally produced speech or text-to-speech (TTS). For same-context trials, listeners were better at discriminating between oral and nasal vowels for child speech in the synthetic voices but adult speech in the natural voices. Meanwhile, in different-context trials, listeners were less able to discriminate, indicating more perceptual compensation for synthetic voices. There was no difference in different-context discrimination across talker ages, indicating that listeners did not compensate differently if the speaker was a child or adult. Findings are relevant for models of compensation, computer personification theories, and speaker-indexical perception accounts.

Index Terms: speech perception, coarticulatory compensation, speaker age, TTS voices

1. Introduction

People now regularly talk to voice-activated artificially intelligent (voice-AI) assistants (e.g., Amazon's Alexa, Apple's Siri, etc.) [1]. Yet, how people perceive the text-to-speech (TTS) used in these systems, compared to naturally produced speech, is an area with many open questions. Do people perceive acoustic-phonetic detail the same way in TTS as in naturally produced speech? Examining whether perceptual processing for synthesized speech is the same as for natural speech can speak to theories of computer personification. For example, the Computers Are Social Actors (CASA) account proposes that individuals subconsciously apply the same behaviors from human-human interaction to interactions with technology [2], [3]. Modern voice-AI systems have highly human-like features, such as apparent gender [4] and personality traits [5]. People even assign apparent speaker age for TTS voices, e.g., Siri voices are rated as being approximately 40s or 50s [6] and Amazon Polly voices are rated as being in their 30s (female voice) or 50s (male voice) [7]. The present study investigates the extent to which voice age shapes how listeners perceive coarticulation in naturally produced and TTS voices.

1.1. Coarticulation

Coarticulation refers to articulatory overlap of discrete speech segments, resulting in a "blended" acoustic signal where cues to multiple phonemes occur at once. For example, coarticulatory vowel nasalization occurs when velum-lowering for a nasal coda begins during the preceding vowel, yielding context-dependent nasality (e.g., [bɛ̃n] "Ben"). Coarticulation contributes to the lack of invariance problem wherein phonemic segments do not always have a 1:1 correlation with acoustic cues [8]. Yet, it has been shown to be a feature used by listeners to aid comprehension of the speech signal [9]–[13].

There is a line of work aimed at understanding the processing mechanisms involved in perceiving coarticulatory variation. When identifying sounds in context, listeners compensate for, or "factor out", coarticulatory overlap by attributing it to a source. For example, [14] found that while listeners are able to hear vowels as nasalized when they were spliced into oral contexts (i.e., [ɛ̃] into the word "bed"), they have difficulty judging a vowel's nasality when it occurs adjacent to a nasal segment (i.e., less likely to hear [ɛ̃] as nasalized when it is in the word "men"). However, more sensitive perceptual tasks demonstrate that listeners are still able to hear some residual nasalization on vowels in the context of nasal consonants; this is referred to as partial compensation [15]. In fact, when listeners are presented with CVN (consonant-vowel-nasal) tokens containing greater amounts of vowel nasalization, they are able to identify segments faster [9], and they also compensate less by displaying more veridical vowel perception [12]. This indicates that variation in coarticulatory magnitude modulates the extent to which compensation occurs, with listeners displaying less compensation as degree of nasality on the vowel increases.

1.2. Perception of coarticulation in TTS voices

A more limited body of work has investigated how people perceive coarticulation in synthetic speech. For example, listeners are better able to identify synthesized speech segments if the vowels contained the appropriate coarticulatory cues for /ɹ/ or /z/ [13]. Thus, coarticulation was crucial for successful phoneme identification in synthesized speech, similar to how coarticulation improves perception of naturally produced speech [9], [12]. Yet, direct comparisons of TTS and naturally produced speech have revealed different responses to coarticulation. For instance, [16] found that nasal coarticulation (measured via acoustic nasality) present in TTS was more ambiguous than that in human speech. This led to distinct patterns of speech adaptation: listeners were more likely to shift their categorizations of nasalized vowels as oral for TTS than natural voices. In the present study, a prediction is that we will also see differences in how listeners perceive coarticulation in naturally produced vs. TTS voices due to fundamentally different acoustic patterns. On the other hand, the increased naturalness of modern TTS speech might mediate any acoustic differences. For instance, [17] found differences in listeners' perceptions of the human-likeness of different TTS types. Neural TTS, such as that generated using long-short term

http://dx.doi.org/10.21437/Interspeech.2021-228

memory (LSTM) neural networks for Amazon Polly voices (the speech used in the present study), is generated based on the overall speaker patterns as well as local phonetic context and yields highly naturalistic speech [18]. Thus, an alternative prediction for the current study is that perception of coarticulation produced by neural TTS and naturally produced speech will be similar, in line with CASA [2], [3].

1.3. Age as a factor in coarticulation

Prior work has shown that speaker-indexical information mediates how listeners perceive speech [19], [20]. For example, presenting listeners with pictures of different aged speakers can lead to different phoneme categorizations [19]. Coarticulatory patterns have been shown to vary across ages, e.g., in older versus younger adults [21]. Therefore, it is possible that listeners form social indices for different talkers based on experience with speaker age groups whose coarticulatory distributions vary. Yet, a recent study found that age-guises of older vs. younger adult speakers did not affect patterns of perceptual compensation [22].

Yet, speech produced by children and adults differs more markedly. For one, children's speech is less intelligible than adults' in conversational settings [23]; and vowels are easier to identify as children get older [24]. Children have been shown to display greater coarticulation than adults in some studies [25]–[27]. Children also display greater variability in production of coarticulation [28], thought to be driven by differences in motor control. In the current study, we compare perception of coarticulation across adult and child speech and predict that listeners will perceive coarticulation differently based on speaker age. Furthermore, perception patterns might differ if the speech is naturally produced (i.e., containing different degrees of coarticulation in adult vs. child) vs. if it is generated via TTS.

1.4. Current Study

The current study tests whether the perception of coarticulatory vowel nasalization differs for adult or child speakers. Further, we compare talker age-related patterns across naturally-produced and TTS speech. First, in order to confirm that the adult and child voices are perceived as distinct ages, we conducted an age ratings experiment (Experiment 1). The acoustic vowel nasalization patterns in their productions are also analyzed. Then, a paired vowel discrimination paradigm (Experiment 2) [15] tested if speaker age (adult vs. child) and voice type (TTS vs. naturally produced) influence listeners' ability to discriminate coarticulatory vowel nasality in nasal consonant contexts.

2. Experiment 1: Age Ratings Task

2.1. Stimuli

The stimuli consisted of two sets of CVC-CVN minimal pairs (i.e., words in oral and nasal codas), matched for onset and coda place of articulation, using the same vowels as in [15] (/ɛ, oʊ/): bed-Ben, bode-bone. For the TTS voices, the stimuli were generated from two US-English Amazon Polly male voices, Kevin (child) and Matthew (adult), using a neural TTS method (no information on ages of voice actors available). Neural TTS voices were chosen due to their naturalistic productions [18]. For naturally produced speech, two female native speakers of American English (aged 9 and 35 years old) produced the tokens.

2.2. Age Rating Study

Participants (n=17, 12 F) were native English speakers with no reported hearing impairments. Listeners were presented with two recordings of each voice (the words bed and Ben) and were asked to indicate how old they thought the speaker was by typing a number 0-100 in a text box. Ratings were analyzed with two sample t-tests. There was no difference in age ratings across speech type for the adult voices [t(56.35)=1.37, p=0.16] (natural mean = 30.6 years, SD = 12.4; TTS mean = 34.0 years, SD = 7.9) and the child voices [t(51.28)=1.97, p=0.06] (natural mean = 8.1 years, SD = 5.7; TTS mean = 10.3 years, SD = 3.1).

2.3. Acoustic Analysis & Results

Degree of vowel nasalization ('acoustic nasality,' A1-P0) was measured acoustically at vowel midpoint with a Praat script [29]. A1-P0 is a measure of the difference in amplitudes between the first formant spectral peak (A1) and the lowest frequency nasal formant peak (P0) [30], [31]. Higher degree of nasality is represented by lower A1-P0 values.

A1-P0 values for oral and nasalized vowels across ages for naturally produced and TTS voices are provided in Figure 1. A t-test revealed that across all voices, nasal vowels have a lower A1-P0 value than oral vowels, indicating that they are more nasalized [t(228.85)=5.04, p<0.001]. Furthermore, the relative difference in degree of nasalization is smaller within TTS than naturally-produced vowels, indicating that there is less coarticulation in TTS than naturally produced speech. Within natural speech, the adult has a larger difference between vowels than the child [t(116.45)=5.05, p<0.001], indicating a greater degree of coarticulation in the adult voice. Within TTS, there was not a significant difference in degree of nasalization [t(84.22)=-0.06, p=0.95] between the adult and child voices.

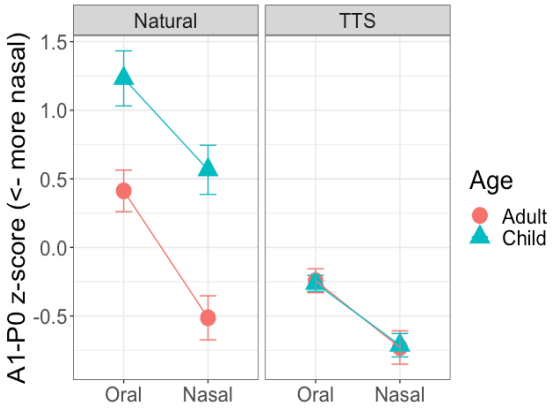

Figure 1: *Mean A1-P0 z-scores for natural (left) and TTS (right) speech across ages; lower values indicate greater nasalization (error bars show standard error).*

3. Experiment 2: Discrimination Task

In Experiment 2, listeners completed a 4-interval forced-choice (4IAX) paired discrimination paradigm [15] to assess their ability to discriminate oral and nasal vowels in CVN and CVC contexts across voice age and type. Overall, we expect an effect of Consonant Context on listeners' discrimination performance. We expect listeners to display higher discrimination for vowels in identical consonant contexts because equal attribution of any acoustic variation to adjacent segments will occur when context is held constant [22], [32]. Differences in discrimination in

different consonant context trials can be used to infer differences in patterns of partial compensation across voices.

3.1. Participants & Stimuli

100 listeners (mean age: 20.8 years, sd = 2.1; 84 female, 11 male, 5 non-binary), native speakers of American English who received course credit for their participation, completed Experiment 1. All reported no history of hearing impairment.

Stimuli consisted of word pairs containing the oral and nasalized vowels extracted from the 4 tokens produced by each of the human and TTS voices used in Experiment 1 (see Section 2.1). First, these vowels were normalized for duration by taking the average of each oral-nasal vowel pair and extracting or adding pulses in Praat [33] with the VocalToolkit [34] until the oral-nasal durations matched within speaker for a given phoneme. Next vowels were normalized for f0 with a linearly falling pitch contour. Then, vowels were spliced back into both b_d and b_n contexts. Tokens were either same-spliced (e.g., an oral vowel into an oral context) or cross-spliced (e.g., an oral vowel into a nasal context). This created two versions of each word for each voice, one with a vowel containing the appropriate coarticulation, and one with different coarticulation. Tokens were normalized for intensity (60 dB). In total, 32 tokens were created (8 words, 4 speakers).

3.2. Procedure

In each trial, two pairs of tokens are presented to a listener: one pair contains acoustically identical vowels (either both nasal or oral); the other pair contains acoustically different vowels (oral v. nasal). Listeners' task is to identify the pair with different vowels.

Participants completed two types of trials (randomly presented), which varied the consonantal context of the pairs: same context and different context. Same context trials contained the same consonant frame across both pairs (all CVC or CVN) and tested listeners' baseline perceptual sensitivity to vowel nasality. Different context trials had varying consonantal frames (CVC v. CVN). Same context example: [bɛ̃n] [bɛ̃n] vs. [bɛn] [bɛ̃n]; different context example: [bɛ̃d] [bɛ̃n] vs. [bɛd] [bɛ̃n] (bolded pair has acoustically distinct vowels).

Differences in performance in different-context trials shows changes in listeners' compensation for vowel nasality across age/TTS conditions.

Order of differing vowels within and across pairs was counterbalanced across trial types. Subjects heard the same type of trials across two blocks varying in speaker age, which were randomly ordered across participants. In total, participants completed 64 trials. Voice type (naturally produced vs. TTS speech) was a between-subjects variable: naturally produced speech condition n = 54 participants; TTS speech condition n = 46 participants.

The experiment was conducted online via Qualtrics. Listeners were instructed to complete the experiment using headphones in a quiet environment. Two practice trials and one listening comprehension question was included within the experiment. A participant's data were retained only if they answered the comprehension question correctly (n=100).

3.3. Statistical Analysis

Listeners' responses were coded binomially: if they selected the pair with acoustically different vowels (=1) or not (=0). We analyzed these responses with a mixed effects logistic regression (lme4 R package; [35]). Main effects included Speaker Age (Adult, Child), Consonant Context (Same, Different), Voice Type (TTS, natural), and all two- and three-way interactions. Random effects included by-Subject random intercepts and by-Subject random slopes for Speaker Age and Consonant Context and their interaction. Contrasts were sum coded. (glmer syntax: Acc ~ Context * Age * VoiceType + (1 + Context * Age | Subject).)

3.4. Results

Figure 2 shows the mean proportion of selecting the acoustically different vowel pair. The model output is provided in Table 1.

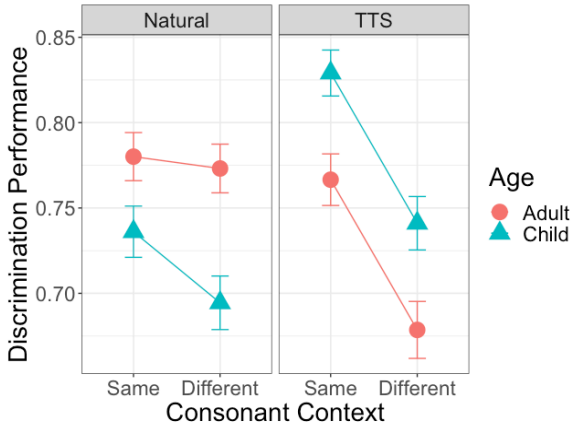

Figure 2: *Mean proportion of acoustically distinct vowels identified (error bars show standard error) for both natural (left) and TTS (right) speech by Speaker Age and Consonant Context. Chance performance is 0.50.*

Table 1: *Model output (see Section 3.3 for model details including structure and glmer syntax).*

	Coef.	SE	*z*	*p*
(Intercept)	1.49	0.18	8.15	<0.001
Context(diff)	-0.69	0.16	-4.25	<0.001
AgeCategory(child)	0.39	0.17	2.34	<0.05
VoiceType(natural)	0.04	0.25	0.17	0.86
Context(diff) * Age(child)	-0.02	0.20	-0.09	0.93
Context(diff) *Voice(natural)	0.46	0.22	2.08	<0.05
Age(child) * Voice(natural)	-0.76	0.21	-3.60	<0.001
Context(diff) * Age(child) * Voice(nat.)	-0.02	0.26	-0.08	0.94

First, we observe a main effect of Consonant Context: as expected, participants showed higher vowel discrimination in trials where the consonant context was the same across all tokens (e.g., [bɛd] [bɛd] vs. [bɛd] [bɛ̃d]), relative to when consonant contexts varied (e.g., [bɛd] [bɛn] vs. [bɛd] [bɛ̃n]).

There was also an effect of Speaker Age on vowel discrimination. Yet, this effect was mediated by an interaction between Speaker Age and Voice Type: listeners displayed higher discrimination of oral and nasalized vowels for the adult's voice in the naturally produced speech, but higher discrimination for the child's voice in the TTS condition.

The model also showed an interaction between Consonant Context and Voice Type: in natural speech, the difference in performance between same-context and different-context trials was smaller than in TTS voices. This indicates that listeners displayed less perceptual compensation for the naturally

produced speech, relative to the TTS. No other effects or interactions were significant.

4. Discussion

The current study examined perceptual sensitivity and compensation for coarticulatory vowel nasalization across talker ages in naturally produced and TTS speech. Across both age and voice types, listeners were more likely to hear acoustically distinct vowels as different when they occurred in identical consonantal contexts compared to when they occurred in contexts where the nasalization could be attributed to coarticulation. This is consistent with perceptual compensation for coarticulation: when the consonant context varies, ability to discriminate differences in coarticulation lower since acoustic variation is attributed to the consonantal source, making phonetic differences harder for listeners to hear [12], [15]. Yet, performance in trials where the consonant context varied was still above chance, consistent with partial perceptual compensation for coarticulation, where some coarticulatory detail remains perceptible [12], [15].

We aimed to investigate whether perceptual sensitivity and partial compensation for nasal coarticulation vary across different types of voices. One prediction was that perceptual sensitivity and patterns of compensation would differ due to the phonetic patterns of coarticulation in natural and TTS speech, suggesting that these processes are acoustically driven. Another possibility was that listeners would apply the same processing mechanisms during the perception of coarticulation in natural and TTS-generated speech. We found differences in overall vowel discrimination based on speaker age, yet critically, this was mediated by the type of voice. In TTS voices, listeners displayed more veridical acoustic perception for the child than the adult voice. Meanwhile, listeners were better able to discriminate the naturally-produced adult voice compared to the naturally produced child voice. This could be due to the overall lower intelligibility of child speech [23] or presence of less coarticulation in the speech signal, as children are known to produce more variable coarticulation [28]. Acoustic analysis of the stimuli showed that the TTS child voice had a smaller difference in acoustic nasality between oral and nasalized vowels, suggesting that there was a smaller degree of coarticulation present in that voice than in the naturally produced child voice. [12] showed that greater degree of nasal coarticulation leads to better vowel discrimination; since the naturally produced child voice had less coarticulation, that could explain why overall vowel discrimination for this voice is lower than the TTS child voice, which contained greater coarticulation. Furthermore, if the TTS child voice was resynthesized from an adult voice, it might contain more adult-like articulations, an additional explanation as to why listeners show better discrimination of vowels in this voice.

Meanwhile, there were no differences in perceptual sensitivity between voice types (listeners were not overall better at vowel discrimination for TTS and naturally produced speech). Hence, overall perceptual sensitivity of nasal coarticulation appears to be comparable for naturally-produced and TTS speech. This is important because one of the key goals in synthesizing speech is to make TTS speech as intelligible as possible [1]; our findings indicate that listeners are able to discriminate between oral and nasalized vowels at similar rates in each type of speech.

Yet, while there were no differences in overall perceptual sensitivity between naturally produced and TTS speech, there were differences in patterns of perceptual compensation: listeners showed more compensation for coarticulation in TTS voices than for naturally-produced voices. Our TTS stimuli contained relatively less nasal coarticulation compared to the naturally-produced voices. The different patterns of compensation can be explained by acoustic differences in the two types of voices: the larger acoustic difference between oral and nasalized vowels in naturally produced speech explains why they were easier to discriminate than the vowels in TTS. Thus, there are distinct differences in patterns of compensation across TTS and natural speech which must be considered in optimizing the voices of digital assistants and other TTS voices that people interact with presently.

However, speaker age did not influence patterns of partial compensation, even when the stimuli contained distinct phonetic coarticulatory patterns. While TTS voices had no difference in relative degree of coarticulation between child and adult voices, the naturally produced speech did; child tokens had less coarticulation than the adult tokens. Hearing an apparent child or adult voice did not trigger top-down differences in expectations of coarticulatory patterns. Furthermore, although the TTS and naturally produced child voices differed in acoustic differences across oral and nasalized vowels from the adult voices, neither one triggered different patterns of compensation for coarticulation. This could be due to the larger variation in coarticulation in child speech [28] meaning listeners may not have a specific expectation for children's speech, so they compensate for these differences.

Overall, the present results extend previous work on phonetic variation in perception by examining how acoustic information present across talkers and in natural and TTS speech can influence vowel discrimination and compensation for coarticulation. This study showed listeners display similar patterns of partial compensation across adult and child speakers; thus, listeners are good at handling variation across speaker ages and adjust so that they compensate to similar extents regardless of age-related acoustic differences. This is similar to [22], which found that when listeners are presented the same stimuli with different age guises, they do not change their patterns of compensation.

Listeners did not show differences in perceptual sensitivity between naturally produced and TTS speech, despite distinct acoustic patterns, consistent with predictions made by CASA [2], [3]. Yet listeners displayed differences in patterns of perceptual compensation, indicated by differences in performance between context conditions [15]. Furthermore, these results support the idea that perceptual sensitivity and compensation for coarticulation is acoustically driven, following patterns of coarticulation present across voice types. From these patterns, we can speculate about the implication of the current findings for further research into human-computer interaction. As methods for TTS evolve, scientists are working to make device voices more naturalistic, reproducing speech and phonetic variation from naturally produced speech more and more accurately. It is important to not only consider what sounds most "human-like" when developing these voices, but also what is perceived most similarly by listeners who interact with digital assistants. Understanding how acoustic differences between the TTS voices and naturally produced speech influence speech perception processes such as vowel discrimination and compensation for coarticulatory vowel nasalization can aid in improving device voices in years to come (cf. comparing perceptual compensation across TTS types in [36]).

5. References

[1] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, Search, and IoT: How People (Really) Use Voice Assistants," *ACM Trans. Comput.-Hum. Interact. TOCHI*, vol. 26, no. 3, pp. 1–28, 2019.

[2] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[3] C. Nass, Y. Moon, J. Morkes, E.-Y. Kim, and B. J. Fogg, "Computers are social actors: A review of current research," *Hum. Values Des. Comput. Technol.*, vol. 72, pp. 137–162, 1997.

[4] F. Habler, V. Schwind, and N. Henze, "Effects of Smart Virtual Assistants' Gender and Language," in *Proceedings of Mensch und Computer 2019*, 2019, pp. 469–473.

[5] I. Lopatovska, "Personality dimensions of intelligent personal assistants," in *Proc. of the 2020 Conference on Human Info. Inter. and Retrieval*, 2020, pp. 333–337.

[6] G. Zellou, M. Cohn, and B. Ferenc Segedin, "Age- and Gender-Related Differences in Speech Alignment Toward Humans and Voice-AI," *Front. Commun.*, vol. 5, pp. 1–11, 2021, doi: 10.3389/fcomm.2020.600361.

[7] M. Cohn, P. Jonell, T. Kim, J. Beskow, and G. Zellou, "Embodiment and gender interact in alignment to TTS voices," in *Proceedings of the Cognitive Science Society*, Toronto, Canada, 2020, pp. 220–226.

[8] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code.," *Psychol. Rev.*, vol. 74, no. 6, p. 431, 1967.

[9] P. S. Beddor, K. B. McGowan, J. E. Boland, A. W. Coetzee, and A. Brasher, "The time course of perception of coarticulation," *J. Acoust. Soc. Am.*, vol. 133, no. 4, pp. 2350–2366, Apr. 2013, doi: 10.1121/1.4794366.

[10] R. Scarborough and G. Zellou, "Clarity in communication:'Clear' speech authenticity and lexical neighborhood density effects in speech production and perception," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 3793–3807, 2013.

[11] G. Zellou and D. Dahan, "Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in English," *J. Phon.*, vol. 76, p. 100910, 2019.

[12] G. Zellou, "Individual differences in the production of nasal coarticulation and perceptual compensation," *J. Phon.*, vol. 61, pp. 13–29, Mar. 2017, doi: 10.1016/j.wocn.2016.12.002.

[13] S. Hawkins and A. Slater, "Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech," 1994.

[14] H. Kawasaki, *Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. In. JJ Ohala, & Jaeger, JJ (Eds.), Experimental phonology (pp. 239-252)*. Orlando, FL: Academic Press, 1986.

[15] P. S. Beddor and R. A. Krakow, "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2868–2887, 1999.

[16] B. Ferenc Segedin, M. Cohn, and G. Zellou, "Perceptual Adaptation to Device and Human Voices: Learning and Generalization of a Phonetic Shift Across Real and Voice-AI Talkers.," in *INTERSPEECH*, 2019, pp. 2310–2314.

[17] M. Cohn and G. Zellou, "Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes," in *Proceedings of Interspeech*, Shanghai, China, Oct. 2020, pp. 1733–1737, doi: http://dx.doi.org/10.21437/Interspeech.2020-1336.

[18] A. Van Den Oord *et al.*, "WaveNet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[19] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *J. Phon.*, vol. 34, no. 4, pp. 458–484, 2006.

[20] N. Niedzielski, "The effect of social information on the perception of sociolinguistic variables," *J. Lang. Soc. Psychol.*, vol. 18, no. 1, pp. 62–85, 1999.

[21] J. Harrington, F. Kleber, and U. Reubold, "Compensation for coarticulation,/u/-fronting, and sound change in standard southern British: An acoustic and perceptual study," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2825–2835, 2008.

[22] G. Zellou, M. Cohn, and A. Block, "Does top-down information about speaker age guise influence perceptual compensation for coarticulatory/u/-fronting?," in *Cognitive Science Society*, Toronto, Canada, 2020, pp. 3483–3489.

[23] P. Flipsen, "Measuring the intelligibility of conversational speech in children," *Clin. Linguist. Phon.*, vol. 20, no. 4, pp. 303–312, 2006.

[24] H.-Y. Sim, C.-H. Choi, and S. H. Choi, "Characteristics of Vowel Formants, Vowel Space, and Speech Intelligibility Produced by Children Aged 3-6 Years," *Audiol. Speech Res.*, vol. 12, no. 4, pp. 260–268, 2016.

[25] N. Zharkova, N. Hewlett, and W. J. Hardcastle, "Coarticulation as an indicator of speech motor control development in children: An ultrasound study," *Motor Control*, vol. 15, no. 1, pp. 118–140, 2011.

[26] N. Zharkova, N. Hewlett, W. J. Hardcastle, and R. J. Lickley, "Spatial and temporal lingual coarticulation and motor control in preadolescents," *J. Speech Lang. Hear. Res.*, vol. 57, no. 2, pp. 374–388, 2014.

[27] A. Noiray, M. Wieling, D. Abakarova, E. Rubertus, and M. Tiede, "Back from the future: Nonlinear anticipation in adults' and children's speech," *J. Speech Lang. Hear. Res.*, vol. 62, no. 8S, pp. 3033–3054, 2019

[28] G. Barbier *et al.*, "What anticipatory coarticulation in children tells us about speech motor control maturity," *Plos One*, vol. 15, no. 4, p. e0231484, 2020.

[29] W. Styler, *Nasality Automeasure Script Package*. GitHub, 2018.

[30] M. Y. Chen, "Acoustic correlates of English and French nasalized vowels," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2360–2370, 1997.

[31] W. Styler, "On the acoustical features of vowel nasality in English and French," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 2469–2482, 2017.

[32] . S. Beddor, J. D. Harnsberger, and S. Lindemann, "Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates," *J. Phon.*, vol. 30, no. 4, pp. 591–627, 2002.

[33] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2018.

[34] R. Corretge, "Praat vocal toolkit: A praat plugin with automated scripts for voice processing," Retrieved 2016-09-05, from http://www. praatvocaltoolkit. com, 2012.

[35] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, Oct. 2015, doi: 10.18637/jss.v067.i01.

[36] G. Zellou, M. Cohn, and A. Block,. "Partial compensation for coarticulatory vowel nasalization across concatenative and neural text-to-speech". *J. Acoust. Soc. Am.,* vol. 149, num. 5, pp. 3424-3436, 2021.