INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Perceptual adaptation to device and human voices: learning and generalization of a phonetic shift across real and voice-AI talkers

Bruno Ferenc Segedin, Michelle Cohn, and Georgia Zellou

Department of Linguistics, University of California, Davis, USA
bferencsegedin@ucdavis.edu

Abstract

Voice-activated artificially-intelligent digital devices are a new type of interlocutor. Like for human talkers, they have idiosyncratic speech patterns that require listeners to perceptually adapt to during language comprehension. One question is how perceptual adaptation to a novel accent in speech produced by a digital device voice compares to adaptation to human voices. Furthermore, adaptation to one talker can *generalize* to novel voices. Hence, we also tested whether perceptual adaptation to accented device voices generalizes to novel human voices, and vice versa. In this study, listeners were first exposed to words with a shifted phoneme realization in either a device or human voice. Later, participants were tested on whether they shifted their identification of words in the shifted talker. Additionally, we tested whether listeners applied the shift to novel device and human voices not heard in exposure. Results reveal talker-specific learning for both device and human voices. Yet, the size of the shift was larger for the device voices. Furthermore, listeners exposed to the shift in device voices showed generalization to novel human voices, and vice versa. These patterns of adaptation and generalization for device and human talkers have implications for models of speech perception models and human-computer interaction.

Index Terms: speech recognition, perceptual adaptation, phonetic variation, human-computer interaction.

1. Introduction

Perceptual adaptation, or recalibrating the speech-to-phoneme mapping, reflects the flexibility of the human perceptual system during language comprehension [1]. The ability to quickly adjust to talkers who produce shifted speech patterns, i.e. a novel phonetic realization of existing phoneme categories, allows for successful communication across a variety of contexts. Recently, a new type of interlocutor has entered speech communities: voice-activated artificially-intelligent (AI) devices, such as Amazon's Alexa. The speech produced by voice-AI systems is increasingly naturalistic, but just like for any individual they have idiosyncratic productions that require perceptual fine-tuning during comprehension. We ask how listeners adapt to the speech patterns of voice-AI systems, relative to human speech patterns. Furthermore, prior work has observed that when people adapt to a shifted phonetic pattern in one voice, they can generalize that learning to novel voices. We ask whether listeners generalize patterns learned from a human or device voice to a novel voice, and whether they apply these patterns only to the voices of the same interlocutor type (device or human).

1.1. Perceptual adaptation to TTS voices

Many humans regularly communicate with non-human artificially intelligent entities [2], such as voice-AI systems (e.g., Amazon's Alexa), where the voices are created via text-to-speech (TTS) synthesis [3]. One question is to what extent we apply the same perceptual adaptation mechanisms observed for human interlocutors to understanding TTS speech. On the one hand, we might predict less adaptation for TTS voices, given their more limited phonetic repertoire compared to real human voices, whose productions vary on the basis of context, e.g., [4], interlocutor, e.g., child-directed, non-native speaker directed speech [5], and natural articulatory and cognitive constraints, e.g. [6]. TTS voices are considerably less phonetically variable than human voices, cf. [7]: for example, they produce more constrained types of prosodic inflections and do not adapt to their interlocutor's communicative needs. Furthermore, there is evidence suggesting that the way we interact with TTS voices is distinct from human interactions: people produce different patterns of phonetic variation when talking to a computer avatar, relative to a human, e.g., longer segments, greater vowel space expansion [8]. This supports the possibility that we have distinct linguistic-phonetic expectations during interactions with devices, similar to how we adapt our speech patterns to L2 speakers, e.g., [6]. Given the artificiality of device voices and listeners' distinct experiences with phonetic variation in synthetic voices, one prediction is that a shifted phonetic pattern in a device voice will be processed differently than a similar shift in a human voice. Such a prediction is supported by prior work finding that people process speech from natural and computer voices differently in variety of psycho-acoustic tasks [9], [10].

On the other hand, computer personification theories, i.e., "Computers are social actors" framework, or CASA [11], propose that when people sense cues of apparent humanness in a system, they apply conventions of human-human interaction. Indeed, one prior study did observe perceptual learning for a phonetic shift in a TTS voice: listeners exposed to lowered front vowels, i.e., "wicked" to "wecked", in a Macintosh 'Bruce' voice were more likely to later accept those pronunciations as real words in a lexical decision task, but note that there was not a human voice comparison [12]. More recently, advances in TTS synthesis have led to even more human-like pronunciation; voice-AI technology, such as Amazon's Alexa is hyper-realistic, exhibiting greater cues of humanness and social characteristics, such as having names, apparent genders, and personas. Thus, another prediction is that there will be no difference in perceptual adaptation to humans and these TTS voices, given robust cues of "humanness".

http://dx.doi.org/10.21437/Interspeech.2019-1433

1.2. Generalization during perceptual adaptation

Once a pattern has been learned (via adaptation), the next question is whether it is abstracted and applied to other talkers. A shift heard in one talker's voice might be considered an idiosyncratic aspect of that individual's speech, with no generalization to different talkers; this was observed by [13], [14]. Other studies observe generalization of perceptual learning to new talkers during testing who had not been heard in exposure [15]. These findings are mixed, but the ideal adapter framework has been one approach to unifying the empirical work [16]-[18]. The ideal adapter proposal is that listeners form representations about the *structured variation* in the acoustic signal, both for a given phoneme contrast and for a given talker. The prediction is that talker generalization occurs based on factors that group novel speakers into similar or different perceived categories as the shifted talker, e.g., L1, dialect region, gender, etc. Supporting this comes from adaptation studies where accented talkers share particular features or characteristics, either in their phonetic patterns or social properties. For example, native English listeners showed talker-independent perceptual adaptation to Chinese-accented L2 English, after exposure to multiple speakers with that accent, suggesting that generalization of learning is facilitated when talkers share phonetic characteristics [19].

The ideal adapter framework provides a way of setting up predictions about generalization of perceptual learning within talker categories: device vs. human. We ask whether generalization is *mediated* by category membership; that is, whether a pattern learned in a device voice goes on to influence lexical comprehension only in novel device voices. Similarly, we might predict that patterns learned for a human voice might generalize only to novel human voices. These constraints on generalization might be based on differences in the way people think exist between synthetic and natural human speech, potentially based on folk knowledge of voice-AI. On the other hand, we might predict that the mechanisms of generalization for device and human voices may be similar, given the increasing human-likeness of the voice-AI system. That is, if Alexa sounds more like a person, perhaps listeners are more inclined to generalize Alexa's patterns to a human voice (and vice versa), a prediction that stems from theories of computer personification.

1.3. Current study

The current study was designed to test two questions. We explore how perceptual learning of a shifted phonetic pattern in voice-AI talkers compares to learning of a similar shift in human talkers. We also compare *generalization* of adaptation in either a device or human voice to novel voice-AI or human voices.

2. Methods

2.1. Stimuli

The words used to generate the stimuli consisted of 4 CVC-CVN-NVN minimal triplets (*bed-ben-men; bud-bun-mun; dead-den-nen; dud-done-none*). Each triplet had matching onset and coda place of articulation and contained an /ɛ/ or /ʌ/.

Recordings of these words were made by 4 distinct talkers: 2 human and 2 Alexa voices, with a female and male voice in each category. The two humans (one male, one female, both native English speakers) produced each word twice, randomly presented, in a sound booth. The recordings were digitized at a 44kHz sampling rate, using Shure WH20 XLR head-mounted microphone. Recordings of the words were also generated in two Alexa voices ("Joanna" and "Matthew"), using Speech Synthesis Markup Language (SSML) in the Alexa Skills Kit. All recordings of the words were amplitude normalized to 60 dB.

Stimuli for the exposure phase (Section 2.2.1) were created with "shifted" and "unshifted" nasality patterns in CVC and CVN words in each voice. The "unshifted" pattern consisted of C_C and C_N frames same-spliced with a vowel taken from different production of that word. For the "shifted" pattern, C_C and C_N frames were cross-spliced with vowels from CVN and NVN contexts, respectively. Thus, vowels in the "shifted" words contained a novel phonetic pattern where the oral vs. nasal lexical contrast surfaced as nasalized (e.g., [dɛ̃d] 'dead') vs. hyper-nasalized vowel realization (e.g., [dɛ̃̃n] 'den'). I.e., each of the 8 words in the shifted pattern contained vowels with more nasalization than typical.

Stimuli for the test phase (Section 2.2.2) consisted of CV syllables cross-spliced with nasal vowels (from the corresponding CVN word) and oral vowels (from the correspondence CVC token). Syllables were gated into wide-band noise, 5dB less than the vowel's peak intensity (following methods in [20]) to reduce perceptual biases toward a final stop coda.

2.2. Participants and Procedure

120 native English-speaking undergraduates recruited from the UC Davis subjects' pool (107 female, 1 nonbinary; mean age= 19.9 years, sd = 1.7 years) completed the experiment in a sound attenuated booth, facing a computer screen.

First, all subjects were shown an introductory slide, where they were told they would be repeating words produced by four talkers: "Matthew" and "Joanna" (Alexa devices) and "Melissa" and "Carl" (humans). This introductory slide included four pictures of the talkers: the images for Joanna and Matthew were two separate Echo devices while the images for the human talkers, consisted of two stock images of smiling adult humans of corresponding genders.

2.2.1. Exposure Phases

102 subjects completed a version of the experiment consisting of two **exposure phases**, followed by a test phase. The purpose of the exposure phase was to provide listeners experience with lexically-guided talker-specific phonetic patterns [1]. Participants heard 2 talkers in the exposure phase. One of the talkers was "shifted" (e.g., [dɛ̃d] labeled 'dead' and [dɛ̃̃n] labeled 'den'); the second talker was "unshifted" (e.g., [dɛd] labeled 'dead'; [dɛ̃n] labeled 'den'). Since there were four different talkers, we generated 4 versions of the exposure phases where each of those talkers was assigned to be the shifted talker, and a talker was assigned to be the unshifted talker, always the opposite gender (counterbalanced across subjects, *n*=23-26 in each condition).

In the first exposure phase, listeners heard each of target words produced by the shifted and unshifted talkers, as they saw the word on the screen (randomly presented). Subjects' task was to press a button after they heard each word (80 trials: 8 words x 2 talkers x 5 repetitions). In the second exposure phase, subjects were again presented the target words and were asked to identify the gender of the talker (male or female) using a labeled button box (64 trials: 8 words x 2 talkers x 4 repetitions).

2.2.2. Test Phase

Listeners then completed the **test phase**, where they performed a word completion task. The test phase was identical across participants, where every participant was tested on all four talkers: the “shifted” and “unshifted” talker, and two novel talkers, relative to their exposure phase condition.

In critical test trials, participants heard a CV syllable containing a nasal vowel (e.g., [dɛ̃]) and identified whether the fragment they heard was taken from a CVC or a CVN word (binary forced-choice option). Participants heard an equal number of control trials, syllables with an oral vowel (e.g., [dɛ]), so as not to bias their responses toward nasal lexical choices. There was a total of 96 critical test trials and 96 filler trials, randomly presented (24 syllables produced in each of the 4 voices).

2.2.3. Control Condition (No Exposure)

In order to interpret whether learning and generalization has occurred, we ran a **control** condition, collecting a set of test responses from listeners with no prior exposure to a shifted or unshifted phonetic patterns. 18 participants completed only the test phase (Section 2.2.2). The performance of these participants in identifying nasal vowels in syllables without prior exposure to the shifted pattern, can be taken as the response pattern to these talkers’ nasal vowels with no prior exposure to the novel accent.

3. Analysis and Results

We collected 11,520 lexical choice responses to critical trials, with nasal vowels, for subjects who completed the Exposure phrase (n=102) and those in the control condition (n=18) (120 participants x 4 talkers x 24 nasal vowel trials). Responses were coded as binomial data: CVC option (=1), or a CVN option (=0). Figure 1 displays the aggregate responses, as a function of Test Talker and Exposure Condition (details about these variables are provided in figure legend and caption).

We modeled responses to nasal vowel syllables selected as CVCs with a mixed effects logistic regression model using the *glmer()* function in the *lme4* package [21]. The model included two fixed effects predictors: Test Talker (2 levels: Human and Device) and Exposure Condition corresponding to the prior exposure a given participant had with the **test voice** (5 levels: Shifted, Unshifted, Same Humanness, Different Humanness, No Exposure), (see Figure 1 and Table 1). If listeners perceptually adapt to the shifted talker from exposure they should be more likely to classify a CV syllable with a nasal vowel as a CVC, relative to categorizations of that talker’s nasal vowels in the control condition. Meanwhile, if generalization occurs across talkers, the likelihood of CVC categorization for nasal vowels for novel talkers should also increase, relative to the control.

In addition, to test whether adaptation and generalization was *different* across human and device voices, we included a two-way interaction between Test Talker and Condition in the model. The model also included by-subject random intercepts and by-subject random slopes for Test Talker. Table 1 presents the model output.

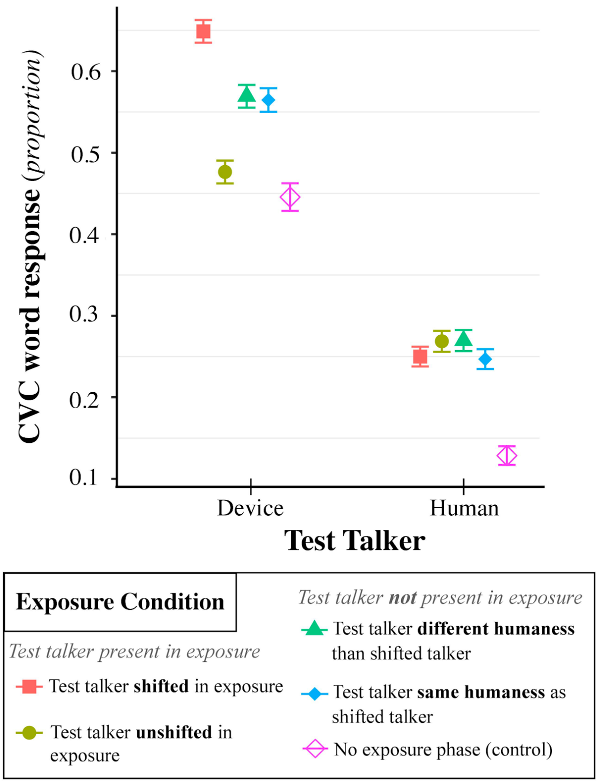

Figure 1: *Proportion of nasal vowel syllables selected as CVCs for Test Talker categories based on Exposure Condition. In two conditions, the Test Talker was present in exposure (either as shifted or unshifted voice). In the other three conditions, the Test Talker was not present in exposure: participants were either exposed to other talkers with the shift (varying in having the same or different humanness with the Test Talker) or participants experienced no exposure phase (control condition).*

Table 1: *Logistic regression model output.*

Effect	*Est*	*SE*	*z*	*p*
Test Talker: Device [ref=Human]	1.81	0.19	9.4	<0.001
Exposure Condition: Shifted [ref=No Exposure]	0.86	0.22	3.9	<0.001
Exposure Condition: Unshifted [ref=No Exposure]	0.96	0.22	4.2	<0.001
Exposure Condition: different humanness [ref=No Exposure]	0.96	0.22	4.3	<0.001
Exposure Condition: same humanness [ref=No Exposure]	0.84	0.22	3.8	<0.001
Test Talker: Device * Exposure Condition: Shifted	0.04	0.25	0.2	<0.001
Test Talker: Device * Exposure Condition: Unshifted	-0.81	0.25	-3.3	<0.001
Test Talker: Device * Exposure Condition: Diff humanness	-0.42	0.25	-1.7	0.09
Test Talker: Device * Exposure Condition: Same humanness	-0.32	0.25	-1.3	0.2

Overall, Test Talker was a significant predictor of CVC responses to nasal vowels. As seen in Figure 1, device voices' nasal vowels are more ambiguous than human voices' nasal vowels, reflected in higher proportion of CVC responses for the control (no exposure phase) condition. There is also an overall effect of Exposure Condition: participants in the Shifted Condition were more likely to respond that a nasal vowel signaled a CVC word than participants in the No Exposure Condition, indicating that perceptual learning of the pattern has indeed occurred across both device and human voices. In fact, participants were more likely to indicate that nasal vowels were CVC items in all conditions, relative to the No Exposure Condition, indicating that any exposure to a shifted voice leads to learning of the shift.

Critically, significant interactions between Test Talker and Exposure Condition reveal that the patterns of veridical adaptation and generalization of learning are *different* for device and human voices. First, we observed an interaction between Test Talker and Shifted Exposure Condition; as seen in Figure 1, the difference in the proportion of CVC categorizations for the control condition relative to the shifted condition is *larger* for device voices than for human voices. In other words, people show a larger perceptual shift for devices that produce a shift than for humans with a comparable shift.

We additionally observed a significant interaction between Test Talker and Unshifted Exposure: we see a *larger* difference of the proportion of CVC responses for unshifted human voices (relative to the human control condition), than for the unshifted device voices (relative to the device control condition). That is, listeners generalized a phonetic shift to a human voice even when they heard that voice in exposure as *unshifted*, while they were less likely to do this for an unshifted device voice.

Finally, interactions between Test Talker and Different Humanness and between Test Talker and Same Humanness were not significant. Since Different Humanness and Same Humanness were significant main effects of Exposure, meaning CVC categorizations were more likely overall in those conditions relative to control, this suggests that listeners generalized to both novel device and human talkers similarly.

4. Discussion

The current study was designed to investigate two open questions about perceptual learning for non-human, artificially-intelligent voices. First, we tested whether perceptual learning to a shifted accent in digital device voices is similar for human voices. Second, we explored generalization of learning across device and human voices. Cross-talker generalization might be constrained by the categories of the talkers, a prediction in line with the *ideal adapter framework* [16]-[18]; on the other hand, generalization from device to human voices (and vice versa) might be similar across device/human categories of talkers, a hypothesis stemming from *computer personification* theories, i.e., CASA [11]. Addressing whether there are asymmetries in the learning and generalization of a phonetic shift across device and human voices can speak to our understanding of the role of voice-AI systems in human speech communities. For example, one question is whether interactions with voice-AI subsequently impact our comprehension of other AI systems and humans. Furthermore, this work can inform the processes of generalizing learning across voices.

We find that both exposure to a shifted pattern in a device voice and in a human voice leads listeners to learn those patterns in those voices. Talker-specific learning for device voices has been observed in prior work, e.g., [12]. However, we observe that the *magnitude* of the learning is different across device and human talkers: listeners show more robust learning for the speech patterns in device voices than for human voices. This was true following exposure to the shifted device voice, as well as for the lack of a shift in the unshifted device voice (i.e., talker-specific learning of the non-shift only occurred for device voices, not for human voices).

One explanation for these differences comes from the fact that the device voices were more acoustically ambiguous with respect to the feature in question (nasality). Indeed, an acoustic analysis of the vowel nasalization patterns in the stimuli confirms this: when we take an acoustic measure of vowel nasalization (A1-P0 dB, [22]), at the midpoint in the oral and the nasal vowels from these talkers, the difference in acoustic nasality between oral and nasal vowels was greatest for the human voices (Melissa: 3.81, Carl: 3.82) and smallest for the device voices (Joanna: 0.72, Matthew: 2.1), supporting the observation that the oral-nasal distinction was inherently more ambiguous for the voice-AI productions. Taken together, this might mean that listeners are more accepting of learning a shifted pattern in a voice that is more ambiguous with respect to nasality.

Another possibility is that greater learning for a phonetic shift in device voices is due to a top-down effect of listeners' knowledge that the device voices were artificially generated. This explanation is supported by results in [9], which found that differences in psychoacoustic processing of computer and human voices were not the result of differences in acoustic features, but were related to more holistic, representational factors. In this case, listeners could be more accepting of atypical pronunciations from synthetic talkers, due to top-down expectations that TTS voices can sound unnatural, reflected in stronger learning of a phonetic shift.

In terms of generalization to novel talkers, we observe no difference in generalization of learning from a shifted voice to novel device or novel human voices. In other words, these talker categories, i.e., the distinction between humans and devices, do not constrain generalization of learning of a phonetic shift. This has relevance both for theories of speech perception and for device-human interaction research. For one, that we see generalization *across* the human and device categories can inform our understanding of talker groups and adaptation strategies, cf. ideal adapter framework. Furthermore, equivalent generalization of learning across device and human voices aligns with the CASA framework of computer personification that humans engage with computers *like people* - here, we extend CASA to apply to generalization across device and human talkers of prior perceptual learning.

Here, we observe that vocal interaction with an idiosyncratic device voice influences how we subsequently process human speech. The finding of transfer of learning from device to human ultimately raises questions about the impact of voice-AI on human speech communities. As human interaction with devices using speech becomes more common, the effect these systems have on our speech and language is an area that warrants further study.

5. Acknowledgements

This work was partially funded by an Amazon Faculty Research Award to GZ.

6. References

[1] D. Norris, J.M. McQueen, and A. Cutler, “Perceptual learning in speech,” *Cognitive Psychology*, vol. 47, no. 2, pp. 204-238, 2003.

[2] K. Olmstead, “Nearly half of Americans use digital voice assistants, mostly on their smartphones,” Pew Research Center, 2017.

[3] D.C. Plummer, M. Reynolds, C.S. Golvin, A. Young, P.J. Sullivan, A. Velosa, and M. Bhat, “Top strategic predictions for 2017 and beyond: Surviving the storm winds of digital disruption,” *Gartner report G00315910*. Gartner. Inc., 2016.

[4] E. Lombard, “Le signe de l'élévation de la voix” *Annales des Maladies de L'oreille et du Larynx, vol. XXXVII, No. 2*, pp. 101-109, 1911.

[5] M. Uther, M.A. Knoll, and D. Burnham, “Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech.” *Speech Communication*, vol. 49, no. 1, pp. 2-7, 2007.

[6] D. Recasens, M.D. Pallares, and J. Fontdevila, “A model of lingual coarticulation based on articulatory constraints.” *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 544-561, 1997.

[7] G. Németh, M. Fék, and T.G. Csapó, “Increasing prosodic variability of text-to-speech synthesizers.” Eighth Annual conference of the International Speech *Communication Association*, 2007.

[8] D. Burnham, J. Sebastian, and L. Rice, “Computer-and human-directed speech before and after correction.” *space*, no. 6, vol. 7, 2007.

[9] S. Lattner, G. Maess, Y. Wang, M. Schauer, K. Alter, and A.D. Friederici, “Dissociation of human and computer voices in the brain: Evidence for a preattentive gestalt-like perception.” *Human Brain Mapping*, vol. 20, no. 1, pp. 13-21, 2003.

[10] M. White, R. Rajakrishnan, I. Kiwako, and S.R. Speer, “Eye tracking for the online evaluation of prosody in speech synthesis: Not so fast!,” *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[11] C. Nass, and Y. Moon, “Machines and mindlessness: Social responses to computers,” *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000.

[12] J. Maye, R. N. Aslin, and M.K. Tanenhaus, “The weckud wetch of the wast: Lexical adaptation to a novel accent.” *Cognitive Science*, vol. 32, no. 3, pp. 543-562, 2008.

[13] T. Kraljic and A. G. Samuel, “Perceptual adjustments to multiple speakers.” *Journal of Memory and Language*, vol. 56, no. 1, pp. 1-15, 2007.

[14] D. Dahan, S. J. Drucker, and R. A. Scarborough, “Talker adaptation in speech perception: Adjusting the signal of the representations?” *Cognition*, vol. 108, no. 3, pp. 710-718, 2008.

[15] T. Kraljic, and A. G. Samuel, “Generalization in perceptual learning for speech,” *Psychonomic bulletin & review*, vol. 13, no. 2, pp. 262-268, 2006.

[16] D. F. Kleinschmidt, K. Weatherholtz and F. T. Jaeger, “Sociolinguistic perception as inference under uncertainty,” *Topics in Cognitive Science*, vol. 10, no. 4, pp. 818-834, 2018.

[17] D.F. Kleinschmidt, & T.F. Jaeger. “What do you expect from an unfamiliar talker?” *Proceedings of the 38th annual meeting of the cognitive science society. Austin, TX: Cognitive Science Society*, 2016.

[18] D. F. Kleinschmidt, “Structure in talker variability: How much is there and how much can it help?”, *Language Cognition and Neuroscience*, vol. 34, no. 1, pp. 43-68. 2019.

[19] A.R. Bradlow, and T. Bent, “Perceptual adaptation to non-native speech.” *Cognition,* vol. 106, no. 2, pp. 707-729, 2008.

[20] J. J. Ohala, and M. Ohala, “Speech perception and lexical representation: the role of vowel nasalization in Hindi and English,” in *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, pp. 41-60, 1995.

[21] D. Bates, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using lme4,” *arXiv preprint arXiv:1406.5823*, 2015.

[22] M. Y. Chen, “Acoustic correlates of English and French nasalized vowels,” *The Journal of the Acoustical Society of America, vol.* 102, no. 4, pp. 2360-2370. 1997.