PERCEPTUAL TARGET OF PHONETIC ACCOMMODATION: A PATTERN WITHIN A SPEAKER'S PHONETIC SYSTEM OR THE RAW ACOUSTIC SIGNAL?

Kuniko Nielsen[1] & Rebecca Scarborough[2]

[1]Oakland University, [2]University of Colorado Boulder
nielsen@oakland.edu, rebecca.scarborough@colorado.edu

ABSTRACT

Previous studies have shown that speakers implicitly imitate phonetic details of recently heard speech. Although this generally results in acoustic convergence, it is unclear whether speakers are responding to linguistic or acoustic aspects of the speech signal. To address this, we compared convergence on linguistic vs. acoustic patterns using an AXB perceptual similarity test. Targets (X tokens) were produced by a speaker with acoustically low A1-P0 (correlated with *high* nasality) but were modified to exhibit *reduced* vowel nasality for that speaker. Comparison stimuli (A & B tokens) were produced by talkers of two types: those who decreased their nasality (imitating within-speaker linguistic/phonetic patterns, but diverging acoustically), and those who increased their nasality (converging acoustically, but diverging linguistically). Listeners judged linguistic imitation tokens as more similar to the target than acoustic imitation tokens, revealing within-speaker phonetic pattern as the basis for listeners' similarity judgments and the likely target of phonetic accommodation.

Keywords: phonetic accommodation, phonetic imitation, speech perception, speech production

1. INTRODUCTION

It is well attested that speakers adapt to various aspects of an ambient linguistic environment over time, including the fine phonetic details [8, 15]. This plasticity of the speech system has been demonstrated in laboratory settings as well, showing that participants' speech becomes more similar to a model talker as the result of brief exposure [7, 5]. Although this generally results in acoustic convergence between interlocutors (or a speaker and a model talker) [1, 2, 11, 12], it is not yet understood what attributes in the speech signal speakers employ when they converge phonetically. [13] addressed this question and examined the relationship between perceived phonetic convergence and various acoustic measures such as duration, F0, F1, and F2. Their results showed that a combination of acoustic measures predicted the perceived phonetic convergence better than any of the individual acoustic attributes alone, revealing the complex nature of phonetic accommodation. Their results also suggest that accommodative behaviors are likely guided by some sort of holistic abstraction instead of memory of particular acoustic features.

The degree of phonetic accommodation is often assessed by using either acoustic measures such as DID (Difference in Distance, e.g., [1, 13, 17]) or an AXB perceptual similarity test [2, 7, 12]. Neither of these measures, however, considers within-speaker linguistic/phonetic patterns as a target of phonetic accommodation. It is well known that speaker normalization takes place in speech perception, enabling us to perceive speech signals with little conscious effort despite the wide inter-talker variability on various phonetic dimensions, e.g., [9]. [9] argues that listeners perceive speech signals relative to an internal representation of the talker; in other words, the percept of speech is within-speaker linguistic/phonetic patterns. If phonetic accommodation is a reflection of speech perception, it is reasonable to assume that speakers refer to within-speaker phonetic patterns instead of raw acoustic targets when they converge phonetically.

[19] presents a unique case that explores this possibility. In [19], target tokens for imitation were produced by a model speaker with low values for A1-P0 (an acoustic correlate of *high* nasality, [3]) but which were modified to exhibit *reduced* vowel nasality (i.e., reduced nasal coarticulation) for that speaker.[1] Even after manipulation to raise A1-P0 in the vowels, i.e., to reduce nasality, his A1-P0 in the vowels was still lower than all participants' A1-P0. In other words, his vowels were acoustically more nasal than all participants', even when the degree of coarticulatory vowel nasality was reduced. The

[1] A1-P0 is a spectral measure that refers to the difference between the amplitudes of the first formant harmonic peak (A1) and the lowest frequency nasal peak (P0). As vowels become more nasalized, P0 increases and A1 decreases, yielding lower A1-P0.

results showed that the majority of participants who listened to the model talker's speech with reduced coarticulatory vowel nasality decreased their vowel nasality in post-exposure productions. [19] presented two interpretations of the results: 1) speakers diverged from the acoustic target (i.e., the generally low A1-P0), or alternatively, 2) speakers converged toward the modeled linguistic target (i.e., toward within-speaker decreased coarticulation patterns), as their productions reflected changes toward the modeled pattern of change (reduced nasality).

The production data in [19] do not allow us to determine which type of target the participants' perceived that subsequently triggered the change in their speech production. However, perceptual judgment of imitation such as an AXB perceptual similarity test could provide a way to distinguish linguistic convergence and acoustic divergence. If the post-exposure tokens produced by those who decreased their nasality (acoustically diverging / imitating within-speaker linguistic/phonetic patterns) are perceived as more *similar* to the target tokens than the baseline tokens, it will suggest that the target of phonetic accommodation is within-speaker linguistic/phonetic patterns of nasality. On the other hand, if the post-exposure tokens produced by those who increased their nasality (imitating acoustic features) are perceived as more similar to the target tokens than the baseline tokens, it will suggest that the target of phonetic accommodation is the acoustic realization of nasality.

The aim of the current study is to explore the target of phonetic accommodation, specifically, whether speakers are responding to linguistic patterns or to raw acoustic aspects of the speech signal. To this end, we compare convergence on what we will refer to as the linguistic vs. acoustic patterns observed in [19] by assessing perceptual similarity using an AXB perceptual similarity test. Further, we examine the role of phonological neighborhood density (ND) in these perceptual judgments, as it has been shown to modulate both patterns of nasality [16] and phonetic accommodation [19]. (Since high ND words with a nasal coda are produced with greater nasality in general [16], these tokens might be judged as less similar to the model talker's tokens than low ND tokens if the listeners' basis of similarity judgement is linguistic pattern of nasality. On the other hand, high ND tokens might be judged as more similar to the model talker's tokens if the listeners are tuning into raw acoustic of nasality.)

2. METHODS

2.1. Participants

Nineteen native speakers of English (12 female) participated in the AXB perceptual similarity test. All listeners reported normal hearing and speech, and received course credit for their participation.

2.2. Stimuli

Stimuli included 32 monosyllabic English words with vowel-nasal sequences, providing a coarticulatory context for vowel nasalization. Half of the words had a high neighborhood density (ND), half had low ND. In the AXB design, the X tokens were the model talker stimuli from [19] with the low A1-P0 (-5.73 dB A1-P0 for these tokens). The speaker was a male native speaker of English, and the tokens were acoustically modified to be less nasal than the speaker's natural baseline (+3 dB A1-P0). The modification was achieved through spectral mixing of the target (the naturally-produced token with a nasal coda and a nasalized vowel) and a phonetically-matched oral vowel minimal pair. (E.g., for *ban*, the stimulus is generated by mixing *ban* and *bad*.) A range of nasal-oral proportions were generated, and the token with the targeted degree of nasality (measured as A1-P0) was selected. This methodology results in an increase in A1-P0, but also in modified realizations of other possible cues for coarticulatory nasality.

The A and B comparison stimuli (of the AXB design) were created from the post-exposure recordings of twelve participant talkers (7 female) from [19]. Talkers were selected on the basis of fitting one of the following 4 imitation types (3 talkers in each type): 1) Ling_Max: talkers who imitated the linguistic change in the target speech to the greatest extent by reducing their vowel nasality in the post-exposure block (average change in A1-P0 = 4.83dB), 2) Ling_Less: talkers who imitated the linguistic change in the target by reducing nasality in the post-exposure block, but to a lesser degree (average change in A1-P0 = 1.83dB), 3) No_Change: talkers who didn't change their nasality after exposure (average change in A1-P0 = 0.07dB), and 4) Acoustic: talkers who imitated the acoustic aspect of the target speech and became more nasal in the post-exposure block (average change in A1-P0 = -1.97dB).

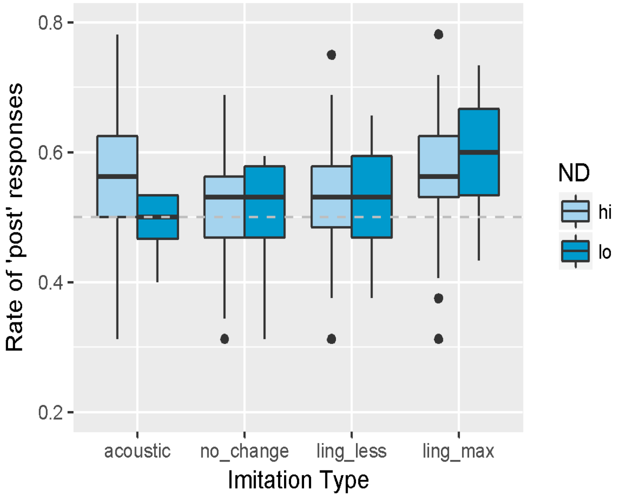
2.3. Procedure

On each trial, three versions of the same word were presented, with the model talker's token as X, and the participants' pre-exposure (baseline) and post-exposure tokens as A and B (one repetition per order, i.e., AXB and BXA). Listeners were instructed to decide as quickly as possible whether the first or the last item (A or B) sounded more like the middle item (X). Trials are blocked by talker, and the order of the twelve talkers was randomized for each listener.

3. RESULTS

Figure 1 summarizes the results of the AXB task, displaying perceived phonetic convergence (= rate of "post" response) across the four imitation types, presented separately by the neighborhood density of the stimuli. As seen, Ling_Max shows the greatest perceived phonetic convergence among four imitation types. In addition, it shows that the effect of neighborhood density on phonetic convergence varies across imitation types.

Figure 1: Perceived phonetic convergence by Imitation Type and neighborhood density.



Responses (re-coded as "pre" or "post") were analyzed with a generalized linear mixed model with Imitation Type (Ling_Max, Ling_Less, No_Change, Acoustic), Presentation_Order (post-exposure token as A or B), ND (High or Low), and Trial umber as fixed factors and random intercepts by-participant and by-item. Results showed a significant overall preference for post-exposure items (54%, z=3.81), indicating that listeners judged the post-exposure items to be more similar to the target stimuli (X) than the items produced before exposure. This corroborates previous studies in phonetic accommodation, indicating that productions become more similar to a heard target. Further, imitation type Ling_Max showed a significantly greater post-exposure preference compared with the Acoustic type (z=2.170). In other words, it is the tokens that are maximally decreased in nasality (following the linguistic pattern of decreased nasality in the target stimuli) that most trigger a "post" response. The model fit is significantly improved by including Imitation Type, compared to the model without the predictor (X^2=29.876, p>0.01). Similarly, ND and Trial number significantly improved the model fit (X^2=15.891, X^2=17.612, respectively). The main effect of ND (z=-2.575) as well as the interaction between Ling Max and ND (z=2.754) were significant, showing that Low ND tokens had higher "post" response in Ling_Max, while High ND tokens had higher "post" response in Acoustic. The effect of Trial number was not significant (z>1, p>0.05).

When the AXB analysis was replicated with the more granular measure of talker-average DID in A1-P0 as a factor instead of Imitation Type (to classify talkers), there was no significant effect on the rate of post response (z<1, p>0.1). Thus, although the within-speaker linguistic/phonetic pattern has a stronger influence on listeners' similarity judgments than raw acoustic similarity, its predicting power for similarity judgments is not completely robust, suggesting that there are other factors involved when listeners make perceptual similarity judgments.

4. DISCUSSION AND CONCLUSION

The current study investigated the target of phonetic accommodation by comparing two types of phonetic accommodation in a perceptual similarity assessment (AXB): linguistic imitation (imitating within-speaker linguistic/phonetic patterns of reduced nasality) and acoustic imitation (imitating acoustically high nasality). We observed a greater perceived phonetic convergence for tokens with decreased nasality (reflecting linguistic imitation) than for those with increased nasality (reflecting acoustic imitation). The perceived phonetic convergence was comparable between tokens that showed acoustic imitation and tokens with no change at all. In other words, it was the tokens that were maximally decreased in nasality (following the linguistic pattern of decreased nasality in the target stimuli) that most triggered a "post" response and not the tokens that were increased in nasality (following the acoustic pattern of greater nasality in the target stimuli), strongly suggesting that within-speaker linguistic/phonetic similarity, rather than raw acoustic similarity, was the basis for listeners' perceptual similarity judgments. Given this, the phonetic accommodation of reduced nasality observed in [19] should be interpreted as convergence toward the modeled linguistic target, even though speakers might have diverged acoustically.

In [2], male speakers yielded stronger F0 accommodation as measured by DID than female speakers, but a group of listeners judged the female speakers as exhibiting greater convergence. These findings may point to a similar distinction between linguistic and acoustic targets. Although it is possible that the difference between the acoustic measure and the perceptual measure is due to listeners tuning into cues other than F0, it is also possible that the difference results from perceptual judgments based on within-speaker linguistic patterns, rather than absolute acoustics.

[13] argued that a holistic measure of phonetic convergence should take into account a combination of various acoustic measures. Our results suggest that in addition to acoustic measures, phonetic accommodation should be interpreted with respect to linguistic patterns. When degree of phonetic accommodation is assessed by measures like DID that consider only acoustic distance along a given dimension, linguistically-predicted direction of change should be considered as well. In the F0 study [2] described above, for instance, it is possible that female speakers' acoustic accommodation would have appeared greater if it had been calculated relative to within-speaker range of F0.

In an account of speaker normalization, [9] argues that listeners perceive speech signals in general relative to an internal representation of the talker. In other words, the very percept of speech is within-speaker linguistic/phonetic patterns. (See also [6].) Insofar as phonetic accommodation is a reflection of speech perception, it is reasonable to assume that speakers employ within-speaker phonetic patterns rather than a raw acoustic target when they converge phonetically.

In fact, notably, the listeners in the current study demonstrated their reliance on these linguistically-relative representations in their perception of both the model talker and the comparison talkers. In order to recognize the Ling_Max post-exposure productions (as opposed to baseline productions) as more similar to the model talker, they had to extract the linguistic pattern of reduced vowel nasality in both the model talker's and potential imitators' (AB talkers') speech. The fact there was no Trial effect in the current study indicates that the extraction of these linguistic patterns can happen very quickly without much exposure to the talker or to the patterns. Presumably, the AB talkers who imitated in the earlier study employed similar normalization in their perception of the model talker. And the fact that those talkers imitated even in a post-test following exposure (not in an immediate shadowing task) indicates that the influence must be durable [19].

Previous work suggests that listeners are indeed sensitive to degree of coarticulatory nasality, as it is modified systematically in various perceptually sensitive lexical [16, 17] and communicative [17] contexts. Further, we know that listeners perceive words with an increased degree of appropriate coarticulatory nasality better than words with less [17]. In the stimuli of the current study, we speculate that the relative modification (reduction) of degree of nasality could be perceived and interpreted by listeners either through comparison of the nasality of the vowel (the only modified portion of each stimulus word) with the nasality of the unmodified adjacent oral and nasal consonants or through sensitivity to the trajectory of A1-P0 change. Although specific A1-P0 values vary across speakers (e.g., [18]), the nasality trajectory across a pre-nasal vowel is consistently a cline from nearly oral to nasal. Our nasality modification methodology involved mixing the vowel portion of a word with nasal coda with that of non-nasal coda, yielding a shallower trajectory of change in nasality across the vowel in our target stimuli than in naturally-produced tokens.

In the current study, the observed influence of ND indicates that durable representation is playing a role in the perception task as well. Thus, [19]'s results, along with the results of the current study suggest that in processing speech, listeners evaluate representations that are dynamically alterable and linguistically sensitive in a speaker-specific way. (See also [4].)

The idea of dynamically alterable phonological representation is consistent with various existing representational accounts of phonetic accommodation (e.g., exemplar models [7, 10] and the interactive alignment model [14]), which assume that accommodations occur because representations are aligned or updated with details from heard utterances. Our results indicate further that these representations must be sensitive to within-speaker phonetic patterns, since these within-speaker phonetic patterns seem to be the target of phonetic accommodation.

5. REFERENCES

[1] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. J. Phonetics 40(1), 177–189.

[2] Babel, M., Bulatov, D. 2012. The role of fundamental frequency in phonetic accommodation. Language and speech, 55(2), 231-248.

[3] Chen, M. 1997. Acoustic correlates of English and French nasalized vowels, J. Acoust. Soc. Am. 102(4), 2360–2370.

[4] Dahan, D., Drucker, S. J., Scarborough, R. A. 2008. Talker adaptation in speech perception: Adjusting the signal or the representations?. Cognition, 108(3), 710-718.

[5] Delvaux, V., Soquet, A. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. Phonetica, 64(2-3), 145-173.

[6] Fowler, C. A. 1994. Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. Percept. Psychophys. 55(6), 597–610.

[7] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105(2), 251–279.

[8] Harrington, J., Palethorpe, S., Watson, C. 2000. Does the Queen speak the Queen's English? Elizabeth II's traditional pronunciation has been influenced by modern trends. Nature 408, 927–928.

[9] Johnson, K. 2005. Speaker Normalization in Speech Perception. In Remez, R. and Pisoni, D.B. (Eds.) The Handbook of Speech Perception. Oxford: Blackwell, 363-389.

[10] Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. J. Phonetics, 34(4), 485-499.

[11] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. J. Phonetics, 39, 132–142.

[12] Pardo, J. S., Jay, I. C., Krauss, R. M. 2010. Conversational role influences speech imitation. Atten. Percept. Psychophys. 72, 2254–2264.

[13] Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E. 2013. Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. J. Mem. Lang., 69(3), 183-195.

[14] Pickering, M. J., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190.

[15] Sancier, M., Fowler, C. A. 1997. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. J. Phonetics 25, 421–436.

[16] Scarborough, R. 2013. Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. J. Phonetics 41(6), 491-508.

[17] Scarborough, R., Zellou, G. 2013. Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. J. Acoust. Soc. Am. 134, 3793-3807.

[18] Styler, W. 2017. On the Acoustical Features of Vowel Nasality in English and French. J. Acoust. Soc. Am.. 142(4), 2469-2482.

[19] Zellou, G., Scarborough, R., Nielsen, K. 2016. Phonetic imitation of coarticulatory vowel nasalization. J. Acoust. Soc. Am.., 140(5), 3560-3575.