Phonetic convergence in shadowed speech: A comparison of perceptual and acoustic measures.

Jennifer S. Pardo
Department of Psychology, Montclair State University, New Jersey, United States
paroj@mail.montclair.edu

Abstract
Phonetic convergence is highly variable across studies, measures, and analyses. The current paper describes a study that examined multiple acoustic measures in concert with a perceptual measure of phonetic convergence. The study employed a shadowing task in which multiple talkers shadowed words from a set of models. Across different scales of analysis, the acoustic measures were highly variable, yielding inconsistent results. Perceptual assessment of phonetic convergence provided a measure that was more stable, reliable, and valid than any single acoustic attribute. Mixed-effects regression modeling assessed the relative contributions of each acoustic attribute to perceived phonetic convergence on a word-by-word basis. This study demonstrates the utility of an approach that combines acoustic and perceptual measures of phonetic convergence.

Index Terms: phonetic convergence, speech perception, speech production

1. Introduction
Speech production is highly variable, both between and within talkers. Some of this variability is related to physiological differences between talkers, but other factors also play a role. These factors include talker dialect and idiolect, as well as relatively transient aspects of a conversational setting. Recent research on phonetic convergence has demonstrated that talkers become more similar in acoustic-phonetic form to an interlocutor or to an auditory model on multiple acoustic-phonetic dimensions. Across numerous studies, investigators have assessed convergence in many acoustic attributes, with differing degrees of success and consistency. The current paper proposes a paradigm that calibrates the relative contribution of individual acoustic-phonetic parameters by including them as predictors of perceptual measures of phonetic convergence in mixed-effects regression modeling.

Previous research on phonetic convergence has employed both perceptual and acoustic measures. In studies that use perceptual measures, phonetic convergence is assessed in an AXB perceptual similarity test with matched lexical items. This paradigm was first introduced by Goldinger [1] for measuring imitation in a shadowing task, then adapted by Pardo [2] for use with items taken from a conversational task. In the AXB task, a separate set of listeners decide which of the two flanking items (A or B), both produced by one talker, sound more like the middle item (X), produced by another talker. The two items being compared (A & B) comprise baseline and repeated utterances. Phonetic convergence is quantified as the proportion of repeated items (either in a shadowing or conversational task) that sounded more similar to a model talker’s items than baseline items. Many studies have used AXB perceptual assessment of phonetic convergence, finding influences of word frequency and repetition, talker sex, and conversational role [3-11].

Perceptual measures of phonetic convergence provide a holistic assessment of change in multiple acoustic-phonetic parameters simultaneously. A talker who converges to (or diverges from) a model/partner does so on multiple attributes within a single item, and the relative convergence of any single attribute varies across words. There is no one acoustic attribute that serves as a metric of convergence across words and talkers. Therefore, perceptual assessment is crucial to any investigation of phonetic convergence because listeners effectively collapse across the multidimensional space when making similarity judgments.

Across studies of phonetic convergence that have focused exclusively on acoustic measures, the landscape of potential attributes is very large. However, there is currently no compelling rationale or standard for choosing one acoustic attribute over another. Some of the factors reported in the literature include intensity, speaking rate, duration, voice onset time, F0/pitch, vowel formants, and voice quality. Few studies have examined more than one attribute at a time, and even fewer have compared acoustic and perceptual measures in the same study. Unfortunately, this approach yields data that is relatively inconsistent and chaotic.

For example, a few studies have examined convergence in vowel formants. In one study, multiple male and female talkers shadowed words produced by 2 male talkers [12]. The items varied in target vowel (5 vowels: /i/, /æ/, /ɒ/, /o/, and /u/), allowing for comparison of convergence across the different vowels. Overall, all vowels were found to converge, but /æ/ and /ɒ/ showed the greatest levels convergence. However, this pattern of vowel convergence differs from that reported in a study of convergence during conversational interaction [7]. In that study, talkers converged on /i/ and /u/ and diverged on /æ/ and /ɒ/. In a study of convergence in college roommates across the academic year, there were no consistent patterns of convergence in vowel formants across pairs of talkers [9]. Instead, some talkers converged in average vowel formants over the course of the academic year while others diverged. At the same time, however, perceptual measures of convergence indicated modest phonetic convergence overall that was unrelated to convergence in vowel formants. Finally, Vallabha and Tuller [13] reported that talkers were unable to imitate their own vowels faithfully, exhibiting biased productions that were most likely related to dialectal differences among the talkers.

In a comprehensive study of convergence on multiple acoustic parameters, Levitan & Hirschberg [14] assessed conversational convergence in intensity, pitch, speaking rate, jitter, shimmer, and noise-to-harmonic ratio. They reported four kinds of convergence in all measures: session level proximity of conversational partners, changes in proximity from the first to the second half of the first game in the interaction and for the entire multi-game session, and turn-by-turn synchrony. Overall, intensity was strongest in both
proximity and convergence. This finding is not surprising given the well-known Lombard sign [15].

At the session level, the talkers were more similar to their partners than to others in intensity (mean and max), pitch (max only) and speaking rate. During the first game, talkers converged on intensity (mean only), shimmer, and noise-to-harmonics ratio. Pitch mean and jitter converged over the halves of the entire session, but were not significant in proximity in the second half. Inter-talker correlations in turn-by-turn synchrony, proximity, and convergence for all acoustic measures were significant, but small, at 0.50 or less.

This thorough approach yields interesting but chaotic data, and it is not known which acoustic attributes are perceptible to listeners, and which play a relatively minor role. On the one hand, a unit change in intensity is not perceived in the same manner as a unit change in F0/pitch. On the other hand, convergence in one acoustic attribute might offset divergence in another. Perceptual assessment of phonetic convergence provides a common metric for evaluating the relative contribution of individual acoustic attributes. Moreover, any acoustic change that is not perceptible is unlikely to play a social role in interaction [16].

The current paper proposes a more comprehensive paradigm for exploring phonetic convergence that integrates acoustic and perceptual measures in the same study. Previously, measures of articulation rates and vowel formants have been found to differ from perceived phonetic convergence [7-10]. However, those studies either used ordinary linear regression or correlation analyses, along with traditional analyses of variance. The current approach adopts mixed-effects regression modeling to assess the relative contribution of distinct acoustic-phonetic parameters to perceived phonetic convergence [17-19].

2. Method

The current study assessed phonetic convergence in 80 words produced by a large set of model talkers and shadowers. The words were used in an AXB perceptual test of phonetic convergence. Three acoustic attributes representing some of the most commonly reported attributes in the literature were compared to the perceptual assessment of phonetic convergence: Duration, F0, and Vowel Formants. A Praat script measured Duration, F0, and Vowel Formants (F1 & F2) at the midpoint of the words. The acoustic measures were converted to measures of Euclidean distance to the model words than the baseline measures, Euclidean distances in (normalized) F1 by F2 space derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

2.1. Recordings

Twenty talkers (10 female) provided model utterances of 80 monosyllabic words sampling across 8 vowels of American English: /ɪ/, /ɪ:/, /e/, /æ/, /a/, /o/, /u/, and /ʊ/ [taken from Munson & Solomon, 20]. Twenty shadowers (10 female) provided baseline and shadowed versions of the words. The shadowers were instructed to say each word as quickly and clearly as possible, whether prompted visually or over the headphones. Each shadower heard words from one model talker in same-sex pairings.

The recordings were collected in a soundproof booth using Sennheiser HMD 280 headsets, and the items were prompted by a Mac PowerBook Pro running SuperLab 4.5.

2.2. Perceptual Assessment

The words were normalized to 80% intensity in SoundStudio and used in multiple AXB perceptual similarity tests. Each model-shadower pair’s items were used to compose a separate test that was completed by 5 listeners, for a total of 100 listeners across the 20 pairs of talkers. In each test, the shadower’s baseline and shadowed utterances (A/B) were compared to the model talker’s utterances (X), counterbalancing for order of presentation of the baseline and shadowed items across trials. Each test comprised 2 blocks of trials in which each word was heard once with the baseline item first and once with the shadowed item first. Responses were scored as correct if a listener selected the shadowed version as more similar to the model utterance. If shadowers converged to the model talkers, then listeners should select the shadowed item as more similar with a probability greater than 0.50.

The perceptual data were collected in quiet testing rooms using Sennheiser Pro headphones, and the trials were prompted by PCs running SuperLab 4.5. Listeners indicated their response on each trial by pressing a key on the keyboard.

2.3. Acoustic Measures

Three acoustic attributes representing some of the most commonly reported attributes in the literature were assessed: Duration, F0, and Vowel Formants. A Praat script measured duration, F0, and vowel formants (F1 & F2) at the midpoint of each vowel. These measures were checked for inaccuracies and outliers and either hand-corrected or removed from the analyses. To permit between-talker comparisons, the Vowel Formant measures were normalized using the Labov technique in the vowels package for R [21, 22]. This extrinsic scaling method derives speaker-specific scaling factors that are then used to scale the original Hertz values during normalization. Because the comparisons were within sex across the same monosyllabic words, the F0 and duration measures were not normalized.

The acoustic measures were converted to measures of convergence by first calculating the distance (absolute value) between the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

3. Results

3.1. Convergence

On average, the shadowers appeared to converge toward the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

The perceptual data were collected in quiet testing rooms using Sennheiser Pro headphones, and the trials were prompted by PCs running SuperLab 4.5. Listeners indicated their response on each trial by pressing a key on the keyboard.

2.3. Acoustic Measures

Three acoustic attributes representing some of the most commonly reported attributes in the literature were assessed: Duration, F0, and Vowel Formants. A Praat script measured duration, F0, and vowel formants (F1 & F2) at the midpoint of each vowel. These measures were checked for inaccuracies and outliers and either hand-corrected or removed from the analyses. To permit between-talker comparisons, the Vowel Formant measures were normalized using the Labov technique in the vowels package for R [21, 22]. This extrinsic scaling method derives speaker-specific scaling factors that are then used to scale the original Hertz values during normalization. Because the comparisons were within sex across the same monosyllabic words, the F0 and duration measures were not normalized.

The acoustic measures were converted to measures of convergence by first calculating the distance (absolute value) between the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

3. Results

3.1. Convergence

On average, the shadowers appeared to converge toward the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

The perceptual data were collected in quiet testing rooms using Sennheiser Pro headphones, and the trials were prompted by PCs running SuperLab 4.5. Listeners indicated their response on each trial by pressing a key on the keyboard.

2.3. Acoustic Measures

Three acoustic attributes representing some of the most commonly reported attributes in the literature were assessed: Duration, F0, and Vowel Formants. A Praat script measured duration, F0, and vowel formants (F1 & F2) at the midpoint of each vowel. These measures were checked for inaccuracies and outliers and either hand-corrected or removed from the analyses. To permit between-talker comparisons, the Vowel Formant measures were normalized using the Labov technique in the vowels package for R [21, 22]. This extrinsic scaling method derives speaker-specific scaling factors that are then used to scale the original Hertz values during normalization. Because the comparisons were within sex across the same monosyllabic words, the F0 and duration measures were not normalized.

The acoustic measures were converted to measures of convergence by first calculating the distance (absolute value) between the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.

3. Results

3.1. Convergence

On average, the shadowers appeared to converge toward the model talker’s words and the shadower’s baseline words and then the corresponding distance between the model talker’s words and the shadowed words. Then, the shadowed distances were subtracted from the model distances, for a derived difference-in-distance measure (DID). For the vowel measures, Euclidean distances in (normalized) F1 by F2 space between talkers were used. If the shadowed words were closer in acoustic distance to the model words than the baseline words, the DID values should be greater than 0.
DID: \( t(79) = 10.79, \ p < 0.0001 \); F0 DID: \( t(79) = 9.19, \ p < 0.0001 \); Vowel DID: \( t(79) = 2.51, \ p = 0.01 \). This reflects the fact that the acoustic measures were highly variable and sensitive to the scale of the analysis.

Each measure, each panel displays AXB perceived phonetic together, split by vowel. To highlight the distinct patterns for /g37/ /g31/ /g39/ /g40/ /g1/ /g38/ /g32/ /g43/ /g38/ /g32/ /g43/ /g38/ /g32/

Figure 1: Phonetic convergence measures split by vowel. Each panel displays the perceptual measure of convergence with the acoustic DID estimates of convergence in Duration (top), F0 (middle), and Vowel Formants (bottom) superimposed.

Figure 1 displays the convergence data for all measures together, split by vowel. To highlight the distinct patterns for each measure, each panel displays AXB perceived phonetic convergence across vowels, with the DID measures for Duration (top), F0 (middle), and Vowel Formants (bottom) superimposed. In all measures, convergence varied as a function of vowel, however, the effect of vowel was not significant in analyses of variance for any measure except for Vowel DID (by word, but not by shadower; \( F(7,72) = 6.23, \ p < 0.0001 \)). Therefore, although it appears that convergence was strongest for /i/ and /u/, the pattern was only significant in the acoustic measure, not in the perceptual AXB measure.

As apparent in the figure, only the measure of Vowel DID was correlated with the perceptual measure of phonetic convergence (bottom panel, \( r(6) = 0.84, \ p < 0.01 \)). However, this correlation weakened when collapsing the data by word (\( r(78) = 0.38, \ p < 0.001 \)), and did not carry through when collapsing the data by shadower (\( r(18) = 0.05, \) ns). In contrast, Duration DID was only correlated with the perceptual measure when collapsing by shadower (\( r(18) = 0.59, \ p < 0.01 \)), and F0 DID was not correlated with perceived convergence in any data treatment. None of the acoustic measures were correlated with each other at any level of analysis.

### 3.1.1. Mixed-Effects Regression Modeling

The previous analyses are highly variable and do not capture the full range of variation in this dataset. The next set of analyses treated the acoustic DID measures as predictors of the perceptual AXB measure in a mixed-effects logistic regression model using R [17-19, 22, 23]. The binomial data from every trial in the AXB tests were modeled as a function of the acoustic DIDs in Duration, F0, and Vowel Formants. To permit comparisons across fixed effects, the acoustic data were normalized by conversion to z-scores before entering the model.

First, a control model was created that included the three random effects (shadower, word, and listener) and a fixed factor for order. This factor was necessary because listeners were biased to select the first item in the AXB test trials, and this bias differed across listeners. To control for the difference in the bias across listeners, random slopes for order were included with the random effect of listener. In addition, random slopes for all acoustic factors were included with the relevant sources (word, shadower), as recommended by Barr et al. [23]. Finally, inclusion of each fixed effect was confirmed through model testing.

Table 1. Model parameters.

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>( \beta )</th>
<th>( SE(\beta) )</th>
<th>( \chi^2 )</th>
<th>( p(\chi^2) )</th>
<th>( \chi^2 )</th>
<th>( p(\chi^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.065</td>
<td>0.023</td>
<td>3.076</td>
<td>0.045*</td>
<td>25.05</td>
<td>0.0009*</td>
</tr>
<tr>
<td>Order</td>
<td>-0.216</td>
<td>0.077</td>
<td>-2.122</td>
<td>0.035</td>
<td>124.69</td>
<td>2.1e-16*</td>
</tr>
<tr>
<td>zDur DID</td>
<td>0.065</td>
<td>0.023</td>
<td>3.076</td>
<td>0.045*</td>
<td>25.05</td>
<td>0.0009*</td>
</tr>
<tr>
<td>zF0 DID</td>
<td>0.065</td>
<td>0.031</td>
<td>2.055</td>
<td>0.045*</td>
<td>70.70</td>
<td>2.2e-16*</td>
</tr>
<tr>
<td>zVowel DID</td>
<td>-0.057</td>
<td>0.025</td>
<td>2.321</td>
<td>0.025*</td>
<td>46.07</td>
<td>5.84e-7*</td>
</tr>
<tr>
<td>zDur*F0</td>
<td>-0.076</td>
<td>0.038</td>
<td>1.913</td>
<td>0.045*</td>
<td>10.19</td>
<td>0.0016*</td>
</tr>
</tbody>
</table>

The acoustic DID measures were added to the control model, and the model parameters appear in Table 1. According to the model, each acoustic DID measure was a significant predictor of the perceptual AXB data, and including all three acoustic DID measures together, plus an interaction between Duration and F0, provided the best fitting model. The significant intercept in the model confirms that the overall phonetic convergence detected in the AXB tests (0.58) was significantly different from chance (0.50). The positive coefficients for the acoustic DID parameters indicate that greater convergence in the acoustic measures was related to...
greater perceived convergence (i.e., greater likelihood of selecting a shadowed item as more similar to a model item than a baseline item). From the magnitude of the parameter weights, it appears that listeners’ judgments were most influenced by DIDs in Duration and F0, followed by Vowel.

4. Discussion
In this study, phonetic convergence during a shadowing task was measured in 4 ways, using a perceptual AXB task and 3 acoustic measures. These data demonstrate the inherently variable nature of convergence across multiple dimensions. While the perceptual AXB data yielded results that were consistently significant at all scales of analysis, the acoustic DIDs measures were highly variable across words and shadowers, yielding inconsistent results. In particular, the patterns of convergence in Duration and F0 across vowels were markedly different from each other and from perceived convergence, as shown in Figure 1. Despite the fact that Vowel convergence was correlated with perceived AXB convergence when collapsing by vowel, this effect diminished or disappeared when collapsing by words and shadowers, and yielded the weakest parameter weight in the regression model.

Perceptual measures are necessary for assessing phonetic convergence for two main reasons. First, it is not possible to characterize the phenomenon in a comprehensive manner without including multiple dimensions. Talkers can converge on one dimension at the same time that they diverge or produce random variation in other dimensions. Moreover, this flexibility extends across different items. For example, convergence in F0 on one item does not imply that the talker will always or only converge on F0. On another item, the talker might converge on vowel formants or duration instead. Indeed, this is the kind of pattern that occurred in this dataset. When listeners perform an AXB perceptual similarity task, they are effectively collapsing across these multi-dimensional acoustic patterns, providing a more reliable measure of phonetic convergence than any single acoustic-phonetic attribute.

A second reason to prefer perceptual measures relates to the potential social function of phonetic convergence. According to Giles’ Communication Accommodation Theory, talkers converge or diverge in order to manage social distance, or to display affiliation [16]. Although a particular acoustic measure might show reliable convergence, it is not necessarily the case that listeners would detect convergence in that attribute. There might be other attributes that diverged, and that are more salient to a listener, or the attribute of interest might be a correlate of another attribute that was more salient to both the listener and the talker who produced the item. Therefore, perceptual assessment of phonetic convergence provides a more valid measure for generalizing to social settings.

In order to examine phonetic convergence in acoustic and perceptual measures in concert, it is necessary to employ mixed-effects regression modeling. As shown in the current dataset, phonetic convergence measures are highly variable across multiple scales of analysis. Mixed-effects modeling can accommodate this variability because it treats all of the data on an item-by-item basis—all sources of variability are entered into the model without collapsing. Finally, these models can assess the relative contributions of individual acoustic-phonetic attributes to perceived phonetic convergence.

In the current study, only three acoustic attributes contributed to the model. The fit of the model depended on including all three acoustic measures together. Models that did not include one of the measures did not maximize fit, as confirmed by the significant chi-square tests for each acoustic parameter. However, the resulting model leaves much to be explained. The Somers index of concordance for the final model was 0.6, which is barely above the recommended cutoff of 0.5. A more comprehensive set of acoustic measures would likely improve the model. Future investigations would benefit from a more comprehensive approach that assesses the relative contributions of multiple dimensions to phonetic convergence in speech production.

5. Conclusions
Studies of phonetic convergence that only focus on a single acoustic-phonetic attribute provide only a glimpse of the phenomenon. Because of the inherently variability in speech production and perception, such glimpses are likely to yield an incomplete (at best) or inaccurate (at worst) characterization of phonetic convergence. Mixed-effects modeling of perception can be used to calibrate the relative prominence of multiple dimensions in speech production. The current paper demonstrates the utility of incorporating multiple acoustic measures in concert with perceptual assessment of phonetic convergence.

6. Acknowledgements
The author is grateful for the assistance of the members of the Speech Communication Laboratory: Kelly Jordan, Rolliene Mallari, Caitlin Scanlon, Eva Lewandowski, Diana Cabezas, Hannah Gash, and Catherine Ode. Completion of this paper was supported in part by startup funds from the Dean’s office of the College of Humanities & Social Sciences at Montclair State University and by NSF PAC program grant #BCS-1229033.

7. References