Latent speech representations learned through self-supervised learning predict listeners' generalization of adaptation across talkers

Zhengyang Jin (Zhengyang.Jin@uga.edu)
Department of Computer Science, University of Georgia, GA, USA

Yuhao Zhu (yzhu@rochester.edu)
Department of Computer Sciences, University of Rochester, Rochester, NY, USA

T. Florian Jaeger (fjaeger@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

Abstract

Unfamiliar accents can pose a challenge to speech recognition. However, listeners often adapt quickly to novel accents, and even generalize this adaptation across talkers with the same accent. We investigate how such cross-talker generalization—critical to effective speech perception—is achieved. We take advantage of advances in automatic speech recognition to test whether comparatively simple similarity-based inferences can explain cross-talker generalization in human listeners. We use the latent perceptual space learned by the HuBERT model—shaped by the statistics of the speech signal and the objective to recognize speech—to meaningfully measure the similarity between talkers' pronunciation. We find that word-level similarity in this latent space predict listeners' ability to successfully generalize across talkers. We discuss consequences for theories of adaptive speech perception. In particular, our results explain why cross-talker variability is *not* a prerequisite for cross-talker generalization (contrary to influential accounts).

Keywords: speech perception; cross-talker generalization; similarity-based inference; automatic speech recognition

Introduction

Talkers' speech is shaped by their physiology, social identity, and language background. These factors result in pronunciation variability at both the level of individual talkers (idiolects) and groups of talkers (second language/L2 accents, regional dialects, and sociolects). In theory, this variability in the mapping from acoustics to sound categories, words, and meaning might pose a formidable computational challenge for speech recognition. However, humans are remarkably skilled at adapting to cross-talker variability, including the generalization of unfamiliar speech patterns across talkers (for review, Bent & Baese-Berk, 2021).

Such cross-talker generalization is critical for effective speech recognition, especially when faced with unfamiliar accents. Yet, the mechanisms that afford cross-talker generalization remain largely unknown. An influential early account focused on the role of *variability* (Bradlow & Bent, 2008): exposure to cross-talker variability was hypothesized to be necessary for generalization, allowing listeners to distinguish accent- from talker-level variability. Support for this hypothesis came from studies that found successful generalization only after exposure to multiple talkers (ibid, Baese-Berk, Bradlow, & Wright, 2013). The idea that variability during exposure/training is *inherently* beneficial has since influenced research on speech perception and beyond, including fields ranging from pedagogy to rehabilitation therapy.

More recent studies, however, suggest a need to revisit this idea (Bradlow, Bassard, & Paller, 2023; Xie & Myers, 2017; Xie, Liu, & Jaeger, 2021). These studies find cross-talker generalization even after exposure to a single talker, and even when that exposure is rather short—as few as 16 sentence recordings. Based on these findings, Xie et al. (2021) proposed an alternative, similarity-based, account: if listeners (1) maintain information about recently experienced speech input and (2) use this information to categorize subsequent speech input (as hypothesized in most modern theories of speech perception), successful cross-talker generalization should depend on the cross-talker *similarity* in the perceptual space(s) relevant to speech representations (phonemes, syllables, or words). Under this alternative account, variability does *not necessarily* facilitate generalization. Instead, potential benefits of variability during exposure are *mediated*, resulting from an increased probability that exposure carries relevant information about test.

A number of studies have now provided evidence for the type of similarity-based generalization proposed in Xie et al. (2021). For instance, generalization from one talker with an unfamiliar accent to another talker with the same accent seems to depend on the similarity of those talkers' phonetic distributions (Alexander & Nygaard, 2019; Xie & Myers, 2017; see also Kraljic & Samuel, 2007; Reinisch & Holt, 2014). However, existing tests of the similarity-based hypothesis have relied on largely qualitative comparisons. And the few studies that have carefully evaluated phonetic characteristics have focused on a small number of phonetic cues and contrasts (e.g., three cues to final stop voicing in Xie & Myers, 2017; F1-F2 and duration for some vowel contrasts in Alexander & Nygaard, 2019). None of these studies quantified similarity, and assessed whether the similarity between the speech during exposure and test can predict listeners' ability to generalize from exposure to test. These studies thus also leave open whether similarity-based inferences can explain a non-trivial share of cross-talker generalization.

To the best of our knowledge, only two previous studies have attempted to quantify how similarity between exposure and test recordings affects generalization. Perhaps surprisingly, neither of these studies found convincing evidence for similarity-based generalization. Key to understanding the reasons for the mixed results of previous work is, we submit, that neither study actually measured whether the exposure

and test recordings were similar *in the relevant way*—i.e., in the mapping from acoustics onto speech categories. For instance, Xie et al. (2021) found that the subjective similarity of exposure and test talkers—estimated through a separate norming study—was weakly predictive of listeners' ability to generalize exposure benefits to test. However, as discussed by Xie et al, subjective similarity ratings can be strongly affected by holistic similarities in talkers' voice quality (e.g., speech rate, pitch, vocal fry), which are not necessarily informative about how talkers' accents affect the mapping from acoustics to speech categories in their study (we describe this study in more detail below). Bradlow et al. (2023) instead used five acoustic features (speech rate, pitch, vowel dispersion, etc.) to estimate talker-to-talker similarity in the acoustic signal. Estimated this way, similarity was not predictive of generalization. Critically, this approach, too, does not capture *how acoustic features map onto speech categories*.

Here, we begin to address this issue. We take advantage of advances in automatic speech recognition (ASR) to estimate how much helpful information a set of exposure recordings contains about subsequent test recordings. We apply this approach to the data from Xie et al. (2021) and test whether word-level similarity in the relevant perceptual space can explain cross-talker generalization of adaptation to natural L2 accents. The latent representations learned by the ASR model we use are obtained through self-supervised learning and fine-tuned to the objective of recognizing words. Such ASR models might thus learn latent perceptual spaces, as well as category representations in those spaces, that resemble those learned by human listeners.

The pipeline we develop is inspired by recent applications of ASR models to quantify how the relative 'non-nativeness' of second language (L2) speech affects its perception by native (L1) listeners (Chernyak, Bradlow, Keshet, & Goldrick, 2024; Kim, Chernyak, Keshet, Goldrick, & Bradlow, 2025). Chernyak et al. compared sentence recordings of L2 talkers to recordings of the same sentences by L1 talkers. As we describe in more detail below, this makes it possible to calculate the distance of two recording in the latent space learned by an ASR model. Chernyak and colleagues found that L1 listeners' accuracy in transcribing an L2 talker's speech decreases, the more distant an L2 talkers' speech is, on average, from L1 speech (see also Kim et al., 2025).

We use the approach developed by Chernyak, Kim, and colleagues to test whether similarity-based inference can explain cross-talker generalization after exposure to L2 speech. Compared to the overall intelligibility of L2 talkers by L1 talkers—the question addressed by Chernyak, Kim, and colleagues—the effects of recent exposure on subsequent generalization that we seek to understand are substantially more subtle (smaller effects). We thus extend the approach pioneered by Chernyak, Kim, and colleagues in a few ways. First, we do not assume the relevant perceptual distances are necessarily Euclidean. Second, we use perceptual similarity—an exponential function of distance—rather than distance to predict human behavior. We also improve the statistical analyses in Chernyak et al., 2024 in two ways. First, we use word-level, rather than talker-level, similarity to model word-level recognition accuracy, rather than talker-level intelligibility. Second, we use mixed-effect logistic, rather than Beta, regression to account for the amount of information available from participants.

Study 1 sets all parameters of our exemplar model to reasonable defaults established in previous work (Apfelbaum & McMurray, 2015), though we note that these previous uses of exemplar models (i) employed a small number of handpicked phonetic features—unlike the high-dimensional latent space learned by modern ASR models—and (ii) tested exemplar models against categorization tasks over isolated segments, syllables, or nonce-words—rather than the full complexity of word recognition within a sentence context. In Study 2, we instead optimize the exemplar model's parameters against the data from Xie and colleagues, and begin to assess whether different layers of the ASR model's DNN architecture differ in how predictive they are about human behavior.

Methods

Next, we describe the data we used to test whether similarity-based inferences can explain cross-talker generalization. Then we describe our ASR-based approach, and how we used it to estimate the similarity of the exposure and test recordings. This measure can be seen as a (coarse-grained) approximation of the amount of information that exposure provides about the speech patterns encountered during test. Finally, we describe how we tested whether the similarity between the exposure and test recordings predicts listeners' ability to generalize from exposure to test.

Data

We used Experiment 1a from Xie et al. (2021, henceforth X21). X21 is a large-scale replication (N=320 participants) of a classic study on the perception of L2-accented English by L1-English listeners (Bradlow & Bent, 2008). During test, all groups of participants transcribed 16 short sentence recordings from a single Mandarin-accented talker. Transcription accuracy was assessed for 3-4 keywords per sentence, yielding 51-52 scored keywords per participant (e.g., "boy", "fell", "window" for the sentence "the boy fell from the window").

During a preceding exposure phase, participants transcribed 80 different, but similarly short, sentence recordings. Depending on the exposure group that the participant was randomly assigned to, exposure recordings consisted of the same 16 sentences each from five different L1 talkers (*control* exposure), the same sentences from five L2 talkers different than the test talker (*multi-talker*), five repetitions of the same recordings from one L2 talker different from the test talker (*single-talker*), or five repetitions of the same recordings from the same talker as during test (*talker-specific*). Participants' transcription accuracy during test thus measured how well they were able to generalize from exposure to test.

The X21 data are particularly suitable for the present purpose because they contained a comparatively large number of exposure-test combinations. Specifically, Xie et al repeated the design described above for four different L2-accented test talkers. This resulted in 32 unique combinations of exposure and test talkers: four combinations of exposure- and test-talker for the control, multi-talker, and talker-specific conditions, and 20 variants of the single-talker conditions. Additionally, the design counterbalanced which of two sets of 16 sentences were used during during test, resulting in 64 combinations of exposure and test recordings. With 51-52 keywords per test, the data contain a total of 3296 unique combinations of (i) exposure talkers, (ii) test talker, and (iii) test keyword. For each combination, X21 includes the number of times participants transcribed the keyword correctly/incorrectly.

Estimating exposure-to-test perceptual similarity

Figure 1 describes how we estimated the perceptual similarity between exposure and test talkers for each of the 3296 combinations. We decided to estimate *word-level* similarities for the present study. In contrast to sentence-level similarities, this takes advantage of the fact that X21 contains keyword-level human transcription accuracies, while avoiding the challenges associated with segment-level similarities (e.g., the need for segment-aligned transcripts, not included in X21). Specifically, we estimated the similarity of each keyword recording during test to the recordings of the same word by the exposure talkers. This introduces an assumption since listeners actually never heard the same keywords or sentences during exposure and test (see above): we assume that exposure recordings were sufficiently informative about exposure talkers' pronunciation to let listeners estimate how the exposure talker(s) *would* have pronounced the test keyword.[1]

Defining a Latent Perceptual Space To quantify word-level similarities between exposure and test talkers, we need to project the sentence recordings from those talkers into a latent perceptual space. This space needs to capture the acoustic dimensions relevant to speech recognition while still maintaining fine-grained acoustic differences between talkers. We used a self-supervised learning (SSL) ASR model to achieve this goal, specifically HuBERT-Large with fine-tuning for word recognition (Hsu et al., 2021). HuBERT is trained on English input (containing mostly L1 English).

Like similar mainstream SSL-ASR models, HuBERT-Large consists of two network blocks: an encoder network and a context network. The encoder network, also known as the feature extractor, is composed of a seven-layer convolutional neural network (CNN), while the context network consists of 24 bidirectional Transformer layers. For Study 1, we adopted the 512-dimensional output of the encoder network (the 7th layer of the CNN) as the perceptual space representation. In Study 2, we consider alternative approaches.

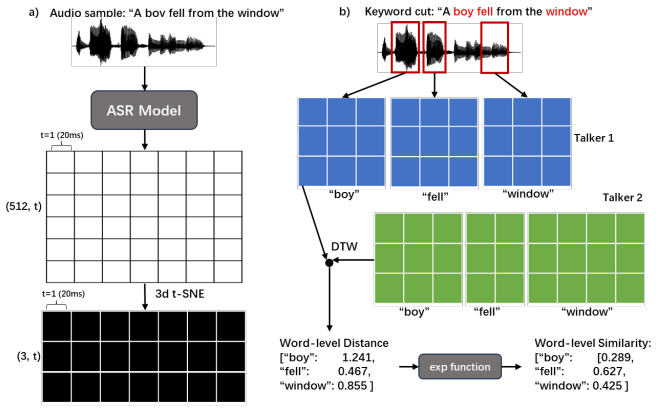

Figure 1: Approach. **a)** Projecting speech recordings into latent perceptual space. **b)** Calculating word-level similarities.

HuBERT is initially trained to learn audio features (latent clusters in the acoustic signal) from MFCCs inputs (which mimic the human auditory system's increased sensitivity to differences between low acoustic frequencies, compared to differences between high frequencies). Using masked-prediction, HuBERT then is trained in a self-supervised fashion to learn latent structure in the speech signal. After training, the HuBERT-Large that we used was fine-tuned against the objective of speech recognition (Hsu et al., 2021).

Word-level perceptual similarities Our approach closely followed Chernyak et al. (2024), but at the word-, rater than sentence-level. We used t-distributed stochastic neighbor embedding (t-SNE, van der Maaten & Hinton, 2008) to reduce the dimensionality of HuBERT's perceptual space, and thus the complexity for subsequent computations (e.g., from 512 to 3 latent dimensions for each time window t of 20ms length; see left side of Figure 1). Specifically, we applied t-SNE to the combined 352 sentence recordings of the 5 L1-accented and 6 L2-accented talkers in X21. We then used manually annotated word boundaries (contained in X21) to extract the trajectories through the t-SNE space for each test keyword. Due to differences in speechrate and pronunciation, the length of this trajectory—and the mapping of each of its 20ms time windows onto the word's phonological segments—can differ between recordings. We thus used dynamic time warping (DTW) to align recordings of the same word by two different talkers, yielding two aligned trajectories (matrices with three rows and n columns; see Figure 2).

Finally, we calculated the perceptual similarity for each pair of aligned trajectories of the same word by two talkers (right side of Figure 1). We follow Apfelbaum and McMurray (2015), and define the distance between two feature vectors in perceptual space as:

[1] While the use of segment-level similarities would have ameliorated the need for this assumption, it would not have removed it: even many of the segments experienced during test did not occur during exposure (especially, if *n*phones are considered, which is necessary to capture the often substantial effects of surrounding phonological context on segment realization). Ultimately, a feature-based approach might be most promising, which, however, comes with its own challenges that we hope to address in future work.

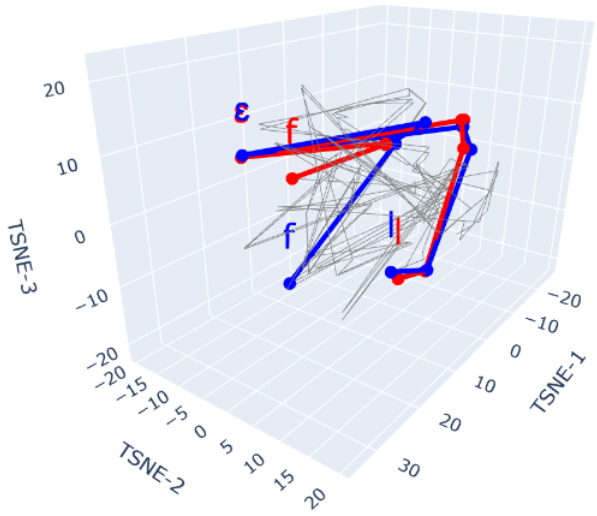

Figure 2: Time-aligned trajectory of two talkers' pronunciation of the sentence "A boy fell from the window" in the 3D t-SNE-derived projection of the 512-dimensional latent space. Highlighted is the word "fell", for which the L1- (blue) and L2-talker (red) differ strongly in their realization of /f/. Each point represents a 20ms time step.

$$dist(i,j) = \sqrt[\tau]{\sum_m w_m |v_{m,i} - v_{m,j}|^\tau} \quad (1)$$

where $v_{m,i}$ is the value of feature vector i in dimension m. For Study 1, we set all feature weights $w_m = 1$ and $\tau = 2$ to obtain Euclidean distances. Using this distance metric, we define the distance between two word recordings $\mathbf{w}_x$ and $\mathbf{w}_y$ as the minimal distance between their trajectories in the t-SNE space that can be found by DTW:

$$D(\mathbf{w}_x, \mathbf{w}_y) = \min_{\pi \in \mathcal{P}} \sum_{(i,j)\in\pi} dist(S(f(\mathbf{w}_x))_i, S(f(\mathbf{w}_y))_j) \quad (2)$$

where π is the alignment path, and $\mathcal{P}$ is the set of all possible alignments between the trajectories, f extracts the representation in latent layer of the ASR model, and S applies t-SNE. This yields the normalized word-level similarity between the two recordings (again following Apfelbaum & McMurray, 2015), ranging from 0 to 1:

$$\text{similarity}_{\mathbf{w}_x,\mathbf{w}_y} = \exp\left(\frac{-D(\mathbf{w}_x,\mathbf{w}_y)^k}{|\pi_{min}|}\right) \quad (3)$$

where $|\pi_{min}|$ is the length of the best path resulting from DTW, and k determines how much quickly similarity decreases with distance in the latent perceptual space. For Study 1, we set and $k = 1$. In Study 2, we explore alternatives.

Figure 3 summarizes the median word-level similarities between all pairs of talkers in our data, using the approach described for Study 1. This shows that word-level similarities were, on average, highest between pairs of L1 talkers (top-left red square) and lowest between pairs of L1 and L2 talkers (pairs not contained in either red square). This *qualitatively* aligns with the results in Xie et al. (2021), where transcription accuracy during test was highest in the talker-specific condition, followed by the multi- and single-talker conditions, and finally the control condition with native exposure.

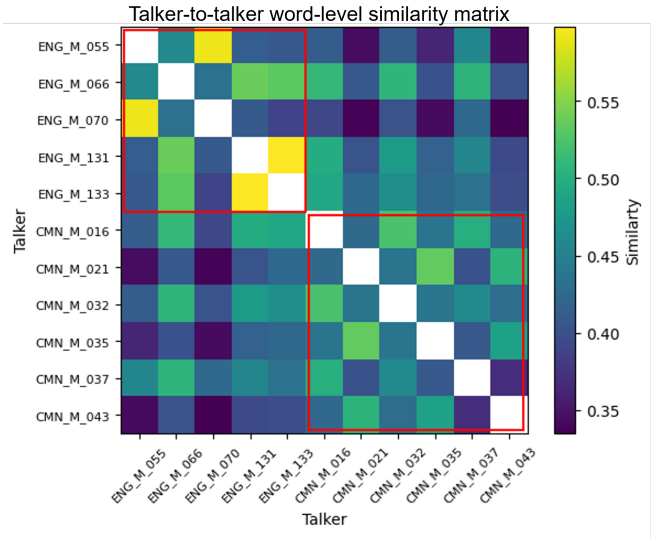

Figure 3: Median word-level similarities between the 11 talkers in X21. Shown for Study 1, using perceptual representations from the final CNN layer of HuBERT-Large, with all $w_m = 1$, $\tau = 2$, $k = 1$. Top red square indicates five L1 talkers; bottom square indicates six L2 talkers.

Exposure-to-test word-level similarity One further aggregation step is necessary totest whether the derived word-level similarities are predictive of listeners' ability to generalize from exposure to test. Both the control and the multi-talker exposure contained five talkers. For Study 1, we calculated the similarity between exposure and test for each keyword as the *maximum* word-level similarity across the five exposure talkers. In Study 2, we explore alternatives.

Figure 4 illustrates the resulting distribution of word-level similarities in Study 1 for one of the four test talkers. Unsurprisingly, there is substantial variability between keywords (crossing lines). This highlights that one can*not* safely conclude from the ordering of a exposure-test talker combination's mean similarity whether the derived word-level similarities can qualitatively predict human perception in X21. This motivates our analysis approach, presented next.

Predicting human perception from similarity

To test whether the 3,296 word-level similarity estimates derived from our ASR-based approach are predictive of human perception, we used mixed-effects logistic regression (`glmer` in R package `lme4`). Specifically, we regressed how often human participants transcribed a keyword correctly or incor-

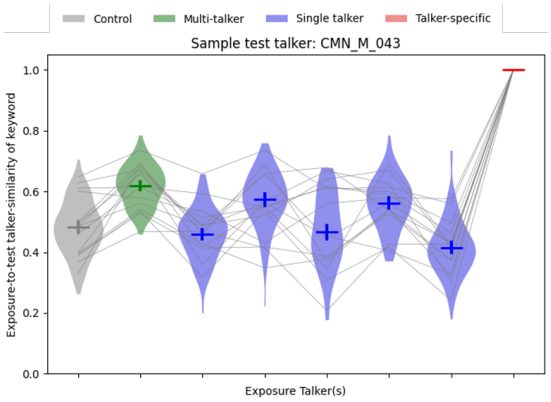

Figure 4: Distribution of word-level similarities in Study 1 between exposure talkers and one of the four test talkers. Point ranges shows medians and their 95% bootstrapped CI over all keywords, violins show density of those word-levle similarities. Thin lines illustrate grouped structure of the data for 15 randomly selected keywords. Per our approach, similarity = 1 in the talker-specific condition(see *Discussion*).

rectly during test against the keyword's exposure-to-test similarity. To avoid Type I error inflation, we included random intercepts by keyword (nested under sentence) and by test talker. This approach accounts for the amount of information available about each keyword's average transcription accuracy, while also accounting for the data's repeated-measures structure. Additional control analyses are described below.

Study 1

When all 3,296 word-level similarity values were included in the model, similarity was a highly significant positive predictor of human transcription accuracy ($\hat{\beta} = 0.8$, $z = 10.06$, $p < .0001$). This result held when talker-specific observations—for which word-level similarities were always 1—were removed from the data, though the effect of similarity was much reduced ($\hat{\beta} = 1.08$, $z = 4.3$, $p < .0001$).

Critically, similarity continued to have significant positive effect in a second regression analysis (w/ talker-specific data: $\hat{\beta} = 1.13$, $z = 4.0$, $p < .0001$; w/o: $\hat{\beta} = 1.35$, $z = 4.6$, $p < .0001$), when we added exposure condition to the regression—the only predictor used in previous studies (Xie et al., 2021). This shows that variation in exposure-to-test similarity explains variation in listeners' behavior beyond that accounted for by exposure condition. The fact that the effect of similarity was reduced when condition is included in the analysis (smaller z-value) suggests that differences in the average exposure-to-test similarity *between conditions* contribute to the overall effect of similarity. Finally, similarity did not account for *all* of the effect of condition (adding condition improved the fit of the model (w/ talker-specific data: $\chi^2(3) = 65.4$, $p < .0001$; w/o: $\chi^2(2) = 65.9$, $p < .0001$).

Study 2

To assess the robustness of our findings, Study 2 repeated the analysis from Study 1, while varying degrees of freedom in the computational architecture. Specifically, we considered:

- The **ASR layer** used to calculate word-level perceptual similarity: the final *CNN* layer (as in Study 1) or the final *Transformer* layer
- The **aggregation function for word-level similarities** in conditions with multiple exposure talkers (control, multi-talker): the *maximum* (as in Study 1) or *mean* similarity of any exposure talker for that keyword and test talker.
- The **distance metric**: $\tau \in (.5, 1, 2, 4, 8)$. $\tau = 1$ is most appropriate for clearly separable feature dimensions, whereas $\tau = 2$ (assumed in Chernyak et al., 2024; Kim et al., 2025) is effective when this is not the case.

For each of these combinations, we used BFGS optimization to find the scaling parameter k that best fit listeners' responses (using the first GLMM described in Study 1, with similarity as the only fixed effect, plus random effects).

Table 1 summarizes the results. In all cases, similarity had a highly significant positive effect on listeners' transcription accuracy during test. This effect was strongest (largest z-value; lowest model BIC), when the latent space of the final Transformer layer was used to obtain perceptual representations, with little effect of the aggregation function (the same held when the talker-specific data is excluded; not shown).

Network layer	Sim. aggregation	τ	best k	Sim. z-value	BIC
CNN	mean	.5	0.10	10.27	6930
CNN	max	.5	0.20	10.08	6934
Transformer	mean	4	1.91	**13.33**	**6859**
Transformer	max	4	3.39	13.33	6862

Table 1: Results of Study 2. Final two columns show how predictive similarity estimates were of listeners' accuracy. Since τ had little effect on fit (z-values differences $< .1$), we show results for only the best-fitting τ for each layer and aggregation function.

For the best-fitting combination of layer (transformer layer 24), aggregate function (mean), τ (4) and k (1.91), we also fit a mixed-effects logistic regression with similarity and exposure condition. This analysis replicated the finding of Study 1 that similarity explains variability in listeners' transcription accuracy both *within* and *between* exposure conditions, but does not explain all variability between conditions (comparison against regression with just condition: $\chi^2(1) = 117.4$, $p < .001$); against regression with just similarity: $\chi^2(3) = 94.4$, $p < .001$). This is visualized in Figure 5: participants in the control condition had reliably lower accuracy during

test than expected based on the ASR-derived word-level similarities (gray line below all other lines). One potential reason for this—to be tested in future research—is that control participants experience the most striking change in speech styles from L1 exposure to L2 test, which can create additional processing difficulty (for review, Magnuson, Nusbaum, Akahane-Yamada, & Saltzman, 2021).

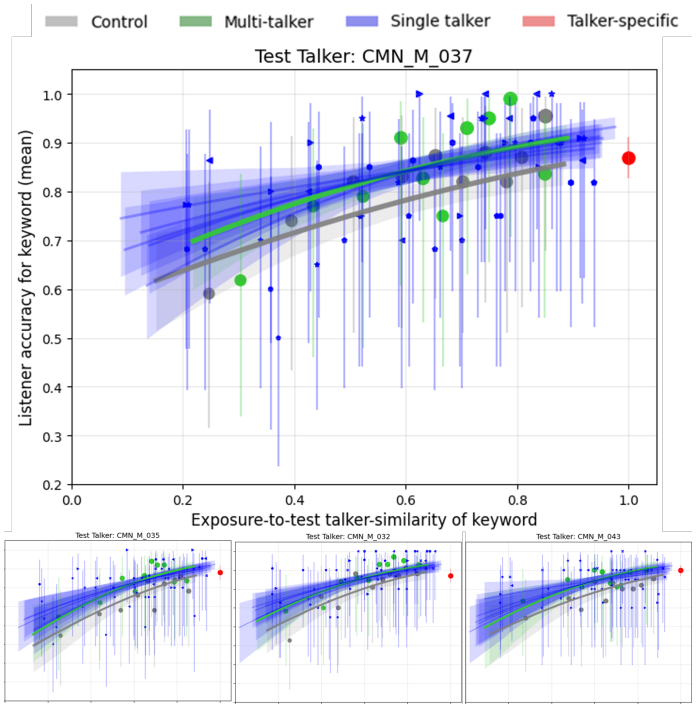

Figure 5: Word-level similarity between exposure and test is a significant predictor of listeners' transcription accuracy during test. Panels show the four test talkers. Lines show fit of ordinary logistic regression fit separately to each combination of exposure condition and test talker. Points show listeners' accuracy for individual words (size indicates number of participant responses for that word).

Discussion

The perceptual similarity of speech inputs to previously experienced speech has long been hypothesized to be critical to speech perception (Goldinger, 1998; Johnson, 1997; Kleinschmidt & Jaeger, 2015). Here we set out to test whether similarity between talkers' realization of speech categories is predictive of cross-talker generalization after exposure to L2-accented speech, as hypothesized in Xie et al. (2021). Our results support this hypothesis. This explains why multi-talker exposure is *not* necessary for cross-talker generalization. It also suggests that it is time to rethink *why* variability during exposure can be helpful for generalization: not because it is inherently helpful, but primarily because it increases the probability that exposure inputs are similar in the relevant ways to those that listeners are later tested on.

To be able to estimate word-level similarity in talkers' pronunciation of speech categories, we drew on advances in ASR. We build on recent work that has used ASR-derived perceptual spaces to investigate speech perception (Chernyak et al., 2024; Kim et al., 2025). These pioneering works found that an L2 talker's similarity to L1 (native) pronunciations in an ASR-derived perceptual space predicts how *a priori* intelligible the L2 talker's speech is for L1 listeners. Here, we extended the approach developed in those works approach to ask whether exposure-driven changes in perception—including cross-talker generalization—can be explained by similarity-based inferences.

Even under simplifying assumptions that arguably bias against the hypothesis (see below), we found that similarity-based inferences predict a substantial amount of variability in cross-talker generalization. To the best of our knowledge, this is the first time this has been demonstrated for an unconstrained task like transcription that begins to resemble the demands and affordances of everyday speech perception. To the best of our knowledge, this constitutes the first direct demonstration that similarity-based inferences can predict a substantial amount of variability in cross-talker generalization.

Finally, our findings have potential implications for ASR system design. By leveraging insights into human perceptual generalization, future ASR systems could incorporate training regimens that mimic diverse exposure conditions, enhancing their robustness to speaker variability.

Methodological Considerations and Limitations

Our approach relies heavily on ASR-derived features. This might introduce biases based on the limitations of the underlying ASR model. For example, HuBERT ensures preservation of acoustic characteristics, but it does not fully account for higher-level contextual dependencies that might influence human perception. Future work could explore the integration acoustic features with contextual embeddings to better capture the range of human speech processing capabilities.

Similarly, our current architecture does not model how listeners' representations might change dynamically during exposure, as listener *integrate* the exposure exemplars into representations derived from previous speech input. This integration process is expected to depend, for instance, on whether a listener actually was able to correctly recognize the speech input during exposure (and thus 'label' the exposure exemplar). Future work might address this limitation by modeling which phones (or context-sensitive variants, such as diphones) listeners recognized during exposure. This will also address another limitation of the present approach, which relies on word-level representations. Words never were repeated between exposure and test in X21. The use of word-level similarities between talkers thus assumes that listeners somehow extract the relevant phonetic properties from the exposure speech that are necessary to generalize to the words encountered during test. While this assumption strikes us as plausible, future work will benefit from an approach that models listeners' generalization process more directly.

References

Alexander, J. E., & Nygaard, L. C. (2019). Specificity and generalization in perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *145*(6), 3382–3398.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *PBR*, *22*, 916–943.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the acoustical society of America*, *133*(3), EL174–EL180.

Bent, T., & Baese-Berk, M. (2021). Perceptual learning of accented speech. *The handbook of speech perception*, 428–464.

Bradlow, A. R., Bassard, A. M., & Paller, K. A. (2023). Generalized perceptual adaptation to second-language speech: Variability, similarity, and intelligibility. *The Journal of the Acoustical Society of America*, *154*(3), 1601–1613.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Chernyak, B. R., Bradlow, A. R., Keshet, J., & Goldrick, M. (2024, 06). A perceptual similarity space for speech based on self-supervised speech representations. *JASA*, *155*(6), 3915-3929.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251.

Hsu, W., Bolte, B., Tsai, Y., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM*.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (p. 145-146). San Diego: Academic Press.

Kim, S.-E., Chernyak, B. R., Keshet, J., Goldrick, M., & Bradlow, A. R. (2025). Predicting relative intelligibility from inter-talker distances in a perceptual similarity space for speech. *Psychonomic Bulletin & Review*, 1–12.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *APP*, *83*, 1842–1860.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *JMLR*.

Xie, X., Liu, L., & Jaeger, T. (2021). Cross-talker generalization in the perception of non-native speech: a large-scale replication. *JEP:General*.

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, *97*, 30–46.