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

Anonymous CogSci submission

Abstract

Unfamiliar accents can constitute a challenge to speech recognition. Yet, listeners often overcome this challenge quickly, sometimes within sentences. Remarkably, listeners can further *generalize* this adaptation to unfamiliar talkers with similar accents. How such generalization—critical to effective speech perception across individuals—is achieved remains unknown. Here we investigate the extent to which similarity-based inferences, hypothesized by exemplar theory, can explain cross-talker generalization. We take advantage of advances in automatic speech recognition to obtain a latent perceptual space—shaped by the statistics of the speech signal and the objective to recognize phonological segments and words—in which we can meaningfully measure the similarity between talkers' pronunciation. We test whether word-level similarity in this latent space can predict listeners' ability to successfully generalize across talkers. Despite strong simplifying assumptions for this initial test, we find credible effects of similarity-based inferences that capture previously unexplained variance.

Keywords: speech perception; adaptation; cross-talker generalization; automatic speech recognition

Introduction

Speakers' voice is shaped by their vocal tract physiology and social or cultural identity. These differences result in variability at both the level of individual talkers (idiolects) and group of talkers (dialects, sociolects). Humans are remarkably skilled at adapting to this variability, often after only short exposure to unfamiliar talkers or accents (for review, Bent & Baese-Berk, 2021). This ability is critical to human communication, especially in noisy environments.

However, it is still unclear *how* listeners adapt and generalize their speech perception so quickly. One influential hypothesis is that these abilities follow without further assumptions from exemplar theory of speech perception or similar theories (Johnson et al., 1997; for review, Apfelbaum & McMurray, 2015). Exemplar theory holds that listeners continuously store rich perceptual traces (exemplars) of previously recognized speech inputs, and use this information to recognize subsequent speech input. Research over the last few decades has provided substantial evidence from a wide range of speech phenomena that are qualitatively compatible with this idea. Some studies further implemented exemplar *models*, and tested them quantitatively against so-called 'small world' experiments on speech perception. In these experiments, listeners usually hear dozens or hundreds of instances of isolated syllables or words that form minimal pairs (e.g., "heed", "hood", "had", ...). Listeners' task in such experiments tends to be highly constrained (e.g., forced-choice categorization), and stimuli are carefully manipulated to differ along a small number of hand-selected phonetic features (e.g., the first and second formants for vowel perception; for a notable towards a more data-drive approach with 20 phonetic features, see Apfelbaum & McMurray, 2015).

These type of studies have provided important proof of concepts, though it is worth noting that none of them has modeled adaptation to accented speech, or generalization across talkers—the question we aim to contribute here. Critically, the 'small world' approach leaves open whether the mechanisms that might explain listeners' ability in highly constrained tasks over highly simplified stimuli also apply to everyday speech perception. Natural accents differ along large numbers of segmental and supra-segmental features, including both linguistically and socio-indexically relevant features. In real life, listeners tend to encounter these accents in connected speech of high phonetic, lexical, and structural heterogeneity, with variable speech rates, loudness, etc. Both the perceptual and cognitive demands, and the contextual affordances of such speech differ starkly from the type of speech experimenters present in small world experiments. Yet, listeners are known to successfully adapt and generalize even when presented with stimuli and tasks that much more closely approximate everyday speech (e.g., passive listening or transcriptions over non-repeated, non-minimal pair sentence recordings Bradlow & Bent, 2008; Xie, Liu, & Jaeger, 2021; for review, Bent & Baese-Berk, 2021).

Here, we test whether similarity-based inferences over exemplars (or similarly rich representations) might be able to explain listeners' ability to generalize exposure to subsequently experienced speech in these experiments. In order to do so, we address an issue that is of relevance to speech research more generally: the use of phonetic features that make strong theoretical assumptions, rather than being learned directly from the large amounts of speech input humans receive during language development. Beyond the assumptions that come with these features, their use tends to require time-consuming and expensive expert annotation that can only be partly automated. This in turn tends to be one of the main reasons why experiments on speech perception tend to manipulate one or two features at a time, which comes with a host of issues (e.g., listeners likely have expectations about

the covariation between phonetic features that are likely to be violated when experimenters only manipulate isolated properties of the stimulus).

To sidestep the strong assumptions (and costs) introduced by this traditional approach, we take advantage of advances in automatic speech recognition (ASR). The latent representations learned by these ASR models are obtained through fully self-supervised learning, not unlike that assumed to underlie human language acquisition. And, recent findings suggest that distances in these latent perceptual space capture something about human perception: the more different non-native, second language (L2) speech is from native (L1) speech when projected into the ASR-derived latent space, the less intelligible it tends to be *a priori* to L1 listeners Chernyak, Bradlow, Keshet, and Goldrick (2024). We extend the framework developed by Chernyak and colleagues to test, for the first time, whether the type of similarity-based inference hypothesized by exemplar models can explain listeners' adaptation to, and generalization of, naturally accented speech. Specifically, we test an ASR-based exemplar model against a comparatively large experiments that assessed listeners' ability to generalize exposure to non-native, second language (L2) accented speech (Xie et al., 2021). We present two studies. Study 1 sets all parameters of our exemplar model to defaults established in previous work (Apfelbaum & McMurray, 2015). In Study 2, we begin to fit (optimize) some of these parameters against the data from Xie and colleagues. We start by describing our general approach.

Methods

Next, we describe the Xie et al. (2021) experiment. Then we describe how we operationalized the latent perceptual space using an ASR-based approach, and estimated the similarity of the exposure and test recordings. This measure can be seen as a (coarse-grained) approximation of amount of information that exposure provides about the speech patterns listeners encountered during test. Finally, we describe how we tested whether the derived similarity between the exposure and test recordings can predict listeners' ability to generalize from exposure to test.

Data

We use Experiment 1a from Xie et al. (2021). The experiment 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). The experiment investigated the effects of exposure to L1- or L2-accented talkers on the subsequent perception of L2-accented speech during test. A summary of the fully crossed and counterbalanced design is shown in Figure 1.

During test, participants always transcribed 16 short sentence recordings from one Mandarin-accented talker. Transcription accuracy was assessed for the 3-4 non-function word keywords contained in each sentence (50 keywords in total per test). During exposure, participants transcribed 80 similarly short sentence recordings that never were the same

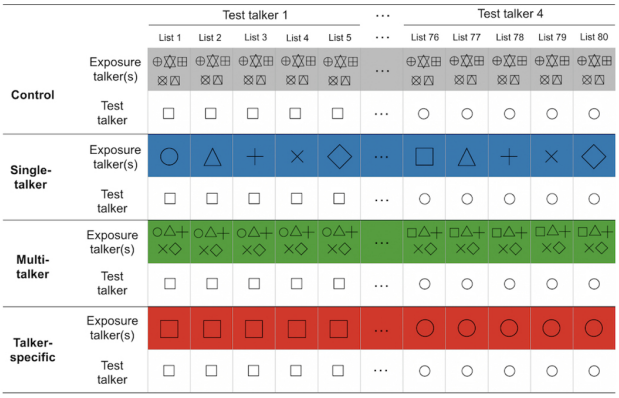

Figure 1: Experiment design assessing human listeners' ability to generalize across talkers (from Xie et al., 2021).

as the test sentences. Depending on the exposure condition, 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*). Transcription accuracy during test thus measures how well listeners were able to generalize from exposure to test.

What makes the data from Xie et al. (2021) suitable for the present purpose is that they contained a comparatively large number of exposure-test combinations. Specifically, Xie and colleagues repeated the design described above for four different L2-accented test talkers while counterbalancing exposure talkers across the different conditions. 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 16 sentences were used during exposure and which during test, yielding 64 unique combinations of exposure and test recordings. Each of these combinations contained 51-52 keywords during test. In total, the data therefore contain 3296 keywords during test that are unique in terms of the combination of exposure talker(s) and test talker that participants had experienced. For each of these keywords, the data contain participants' average transcription accuracy.

Estimating perceptual (dis)similarity between exposure and test talkers

Next, we describe how we estimated the perceptual similarity between exposure and test talkers for each of the 3296 combinations of exposure talker(s), test talker, and keyword. This similarity serves as a (coarse-grained) approximation of the information that the speech recordings experienced during exposure provided about the keyword participants had to transcribe during test.

Figure 2 describes our overall approach. One complication deserves mention upfront: since test sentences never were the same as the exposure sentences (see above), listeners rarely experienced the same keyword during exposure and test.[1] This has consequences since we use the similarity between test talkers' and exposure talkers' pronunciation of the keyword as a stand-in for the information that exposure provided about the keyword during test. The approach we present here therefore assumes that exposure sentence recordings contain enough information about the exposure talkers' pronunciation to let listeners estimate how the exposure talker(s) *would* have pronounced the test keyword. This strikes us as a reasonable initial assumption since exposure always contained 16 sentence recordings from each exposure talker, and these recordings covered a decent amount of the phonetic space of English. We discuss alternative approaches we plan to pursue, and the challenges they face, in the general discussion.

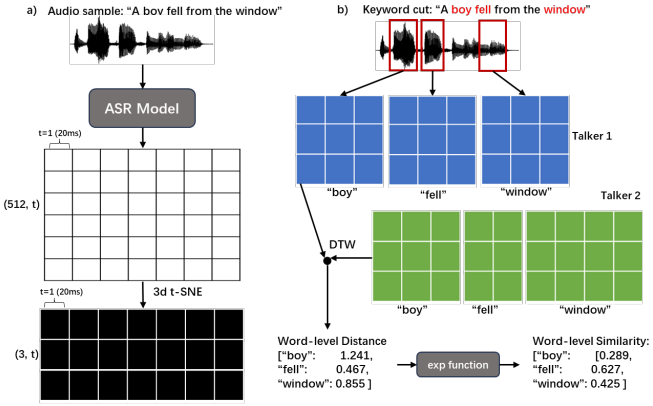

Figure 2: Overview of ASR-based approach. **a)** Projecting speech recordings into latent perceptual space. **b)** Calculating word-level perceptual similarities.

Defining a Latent Perceptual Space To quantify the similarity between exposure and test talkers in Xie et al.'s experiment, 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 (W. Hsu, 2021). HuBERT is trained solely on English input, of which the clear majority is likely to be L1-accented.

Like similar mainstream SSL-ASR models, HuBERT 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 12 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 compare this decision with other alternatives.

We chose HuBERT over other SSL-ASR models, such as Wav2vec2, Conformer, or Whisper, because it has recently been reported to capture qualitative differences in the relative intelligibility of L2-accented talkers (Chernyak et al., 2024). HuBERT has been argued to employ a more advanced representation learning approach, allowing it to capture finer-grained audio features. HuBERT uses a pre-training method to learn audio features, initially extracting them using MFCCs, which closely mimicking the human auditory system's increased sensitivity to differences between low acoustic frequencies, compared to differences between high frequencies. HuBERT then applies the k-means clustering method to group the data and creates hidden units for feature mapping.

Word-level perceptual similarities 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 (from 512 to 3 latent dimensions for each time window t of 20ms length; see left side of Figure 2). Specifically, we applied t-SNE to the combined 352 sentence recordings of the 5 L1-accented and 6 L2-accented talkers used in Xie et al. (2021). We then used manually annotated word boundaries (provided by Xie and colleagues) to extract the trajectories through the t-SNE space for each of the 3-4 keywords from each sentence recording (Figure . 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. To address this, we 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). Figure 3 shows the trajectories of two different recordings of the same sentence before and after DTW.

Finally, we calculated the perceptual similarity for each pair of aligned trajectories of the same word by two talkers (right side of Figure 2). We follow Apfelbaum and McMurray (2015), and define the distance between two feature vectors in perceptual space as:

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

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 pos-

[1] Only one keyword ($< 1\%$) occurred in both sets of 16 sentences.

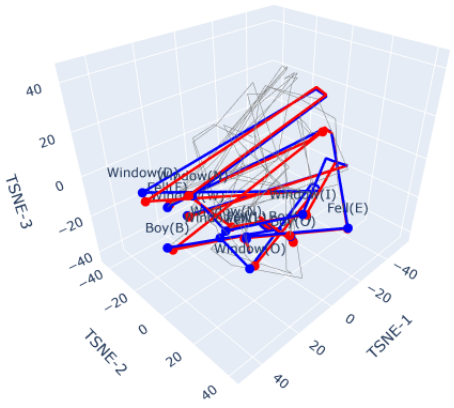

Figure 3: Trajectory of single L1-talker (red) vs single L2-talker (blue) for the example sentence "A boy fell from the window" through the 3d t-SNE-derived projection of the 512-dimensional latent space in which we assess similarity. Each point represents a phonological segment. Bold lines represent keywords' part.

sible 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):

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

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 4 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 of Xie et al. (2021), who found that 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.

Exposure-to-test word-level similarity One further aggregation step is necessary before we can test whether the derived word-level similarities are predictive of listeners' ability to generalize from exposure to test. Both the control (5 L1 talkers) and the multi-talker condition (5 L2 talkers) contained multiple talkers. For those conditions, we calculated the similarity between exposure and test for each keyword as the *maximum* word-level similarity across the five talkers.

Figure 5 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 mean similarity (points) whether the derived word-level similarities can qualitatively predict human perception in Xie et al. (2021). This motivates our analysis approach, presented next.

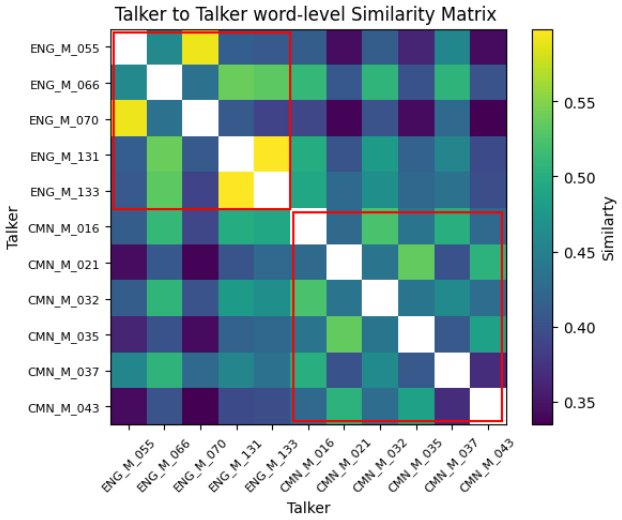

Figure 4: Median word-level similarities between the 11 talkers in the Xie et al. (2021) data. Shown for Study 1, using perceptual representations from the final CNN layer of HuBERT, all $w_m = 1$, $\tau = 2$, $k = 1$. Top-left red square indicates five L1 talkers; bottom-right square indicates six L2 talkers.

Predicting human perception from similarity

To test whether the 3,296 word-level perceptual similarity values derived from our ASR-based approach are predictive of human perception, we used mixed-effects logistic regression (`glmer` in R package `lme4`). Specifically, we predict how often human participants transcribed a keyword correctly or incorrectly during test as a function of the keyword's exposure-to-test similarity. To avoid inflated Type I errors, we included random intercepts by keyword (nested under sentence) and by test talker. This approach adequately accounts for the amount of information available about each keyword's average transcription accuracy, while also accounting for the data's grouped/repeated-measures structure. Additional control analyses are described in each study.

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 during test ($\hat{\beta} = 1.15$, $z = 10.0$, $p < .0001$). This result held when talker-specific observations—for which word-level similarities were always 1 (cf. Figure 5)—were removed from the data, though the effect of similarity was much reduced ($\hat{\beta} = 1.08$, $z = 4.0$, $p < .0001$).

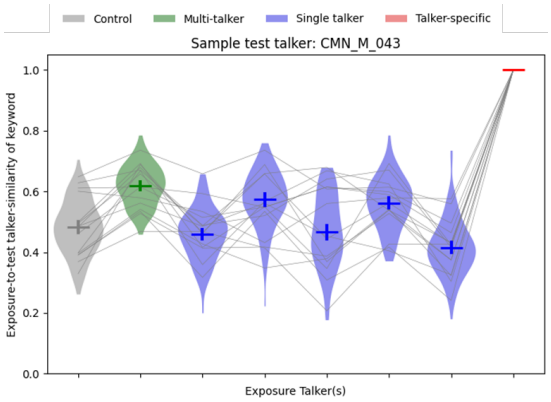

Figure 5: Distribution of word-level similarities in Study 1 between exposure talkers and one of the test talkers. Point ranges shows medians and their 95% bootstrapped CI over all keywords, violins show density of those keyword similarities. Thin lines show 15 randomly selected keywords to illustrate grouped structure of the data. Note that the high similarity in talker-specific condition (always 1) is a trivial consequence of our approach (we return to this limitation in the *Discussion*).

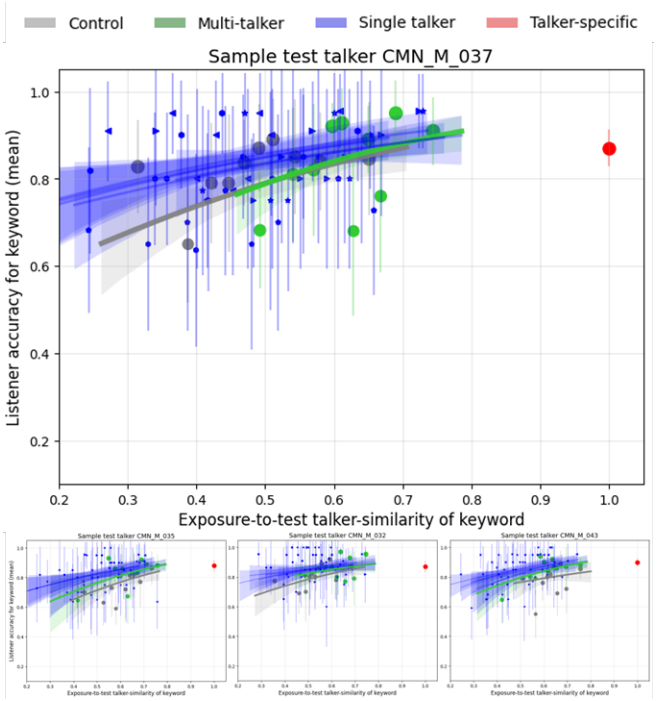

Figure 6: Keyword-level similarity between exposure and test is a significant predictor of listeners' transcription accuracy during test. Panels show the four test talkers. Lines show best fit of mixed-effects logistic regression (incl. the random effects; see text). Points show participants' accuracy for individual keywords (size indicates number of participants that transcribed keyword).

Critically, similarity continued to have significant positive effect in a second regression analysis (w/ talker-specific data: $\hat{\beta} = 1.18$, $z = 3.9$, $p < .0001$; w/o: $\hat{\beta} = 1.39$, $z = 4.4$, $p < .0001$), when we added exposure condition—the only predictor used in previous studies (Xie et al., 2021)—to the model. 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: $\chi^2(3) = 66.23$, $p < .0001$; w/o: $\chi^2(2) = 66.46$, $p < .0001$).

Figure 6 visualizes the fit of the regression with both similarity and condition as predictors against participants' transcription accuracy. This illustrates the effects of similarity, and reveals one reason why similarity does not capture *all* of the effects of exposure condition: 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).

Study 2

To assess the robustness of our finding, Study 2 repeated the process outlined in 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**: $\tau \in (1,2)$. $\tau = 1$ is taken to be most appropriate for clearly separable feature dimensions, whereas $\tau = 2$ is taken to more effective when this is not the case.

For each of these combinations, we found the similarity 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). We used Broyden-Fletcher-Goldfarb-Shanno optimization, as implemented in R's `optim` function.

Table 1 summarizes the results. In all cases, similarity had a highly significant positive effect on listeners' transcription

accuracy during test. This effect was numerically strongest (largest z-value; lowest model BIC), when the latent space of the final Transformer layer was used to obtain perceptual representations, and the maximum similarity to any exposure talker was used for the control and multi-talker conditions (this result replicates when the talker-specific data is excluded from the optimization).

Network layer	Sim. aggregation	τ	best k	Sim. z-value	BIC
CNN	mean	1	2.59	10.43	6928
CNN	mean	2	1.37	10.25	6931
CNN	max	1	1.14	10.09	6935
CNN	max	2	1.24	10.14	6934
Transformer	mean	1	0.56	10.77	6917
Transformer	mean	2	0.77	11.22	6907
Transformer	max	1	0.82	11.83	6893
Transformer	max	2	1.09	**12.52**	**6878**

Table 1: Results of Study 2 (sim. = similarity)

Discussion

Recent work has shown that ASR-derived perceptual representations can to some extent predict how *a priori* intelligible L1 listeners experience L2-accented speech to be (Chernyak et al., 2024). Here, we have extended this approach to ask whether exposure-driven changes in intelligibility can be explained by similarity-based inferences over the same latent space. Our findings suggest that latent perceptual spaces learned through self-supervised learning can be used to investigate speech adaptation and generalization in human listeners. We find that the same computational logic previously shown to explain adaptation in 'small world' experiments can explain a substantial amount of variation in listeners' adaptation and generalization to natural accents in much less constrained tasks—including effects both within and between exposure conditions. To the best of our knowledge, this constitutes the first direct demonstration that storage of exemplars provides a plausible explanation for adaptation and generalization across talkers during everyday speech perception.

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 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. Xie et al. (2021), never repeated words between exposure and test. 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 and an finer grain.

Broader Implications for ASR Design

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. The observed relationship between perceptual similarity and generalization further suggests that embedding spaces optimized for similarity could improve next-generation ASR models.

References

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.

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

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.

Johnson, K., et al. (1997). Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing*, 145–165.

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

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

W. Hsu, Y. T. K. L. R. S. A. M., B. Bolte. (2021). Hubert: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM*.

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