Exemplar models of speech perception do not prevent the need for normalization

**Abhishek Sharma** (ashar58@ur.rochester.edu)  
Goergen Institute for Data Science, 1209 Wegmans Hall, Rochester, NY 14627, USA   
  
**Anna Persson** (anna.persson@su.se)   
Department of Swedish and Multilingualism, Stockholm University, SE-106 91 Stockholm, Sweden

**T. Florian Jaeger** (fjaeger@ur.rochester.edu)  
Brain and Cognitive Sciences and Goergen Institute for Science, Meliora Hall,

Rochester, NY 14627, USA

Abstract

We compare two competing models of speech perception—exemplar theory with and without auditory normalization—against human behavior in an 8-way forced-choice categorization task. We trained exemplar categories on a phonetic database, optimized free parameters against listeners responses, and evaluated the two types of models against held-out responses from listeners. Our results suggest that exemplar storage alone is *in*sufficient to explain human speech perception, replicating and extending findings obtained under the assumption of parametric (Gaussian) category representations. We also determine the reason for the insufficiency of exemplars: without normalization, exemplars do not sufficiently cluster in the acoustic-phonetic space, reducing exemplar effects to the single closest neighbor—thus exaggerating the curse of dimensionality.

**Keywords:** speech perception; exemplar models; normalization

# Introduction

Understanding how humans perceive and interpret spoken language remains a complex challenge, particularly given the variability introduced by different speakers, accents, and contextual factors. This project seeks to unravel the cognitive processes underlying speech perception by comparing exemplar theory and auditory normalization models.

…

En bild som visar text, diagram

AI-genererat innehåll kan vara felaktigt.

Figure : Illustrating cross-talker variability in formant values for US English vowels (data: Xie & Jaeger, 2020)

By leveraging phonetic and perceptual datasets, the study evaluates which model—or combination thereof—best mirrors human speech perception behavior, particularly focusing on whether exemplar-based representations are sufficient to explain listeners’ perception.

…

# Exemplar models

The core idea behind exemplar models of speech perception is that listeners store previously experienced speech inputs along with both their (inferred) phonetic category label and representations of the context the speech input occurred in. These exemplars are then used to categorize subsequent speech input. Specifically, the posterior probability that a speech input *x* is an instance of category *c*, p(*c* | *x*) is assumed to be a function of the similarity between *x* and previously experienced exemplars stored in listeners’ memory. Here we follow the formulation provided in Apfelbaum & McMurray (2015), with some notational differences:

where p(*c*) is prior probability of the category *c* in the current context, and is the similarity between the input *x* and the stored exemplar *j* of category *c*. This similarity is the function of the distance *D* between *x* and the exemplar in the relevant perceptual space, scaled by a factor *s*:

The scaling factor determines how quickly similarity decreases as a function of distance. For , all exemplars affect the categorization of the input *x* equally, regardless of their distance from *x*. The larger *s*, exemplars that are further away will increasingly affect categorization less.

Various distance functions *D* have been proposed in the literature. Here we follow previous work (Apfelbaum & McMurray, 2015, ONE-MORE), and assume that distances are calculated as a weighted sum of distances along *m* = 1, …, *M* separate perceptual feature dimensions . This independence assumption avoids the curse of dimensionality that would otherwise result from calculating distances in the type of high-dimensional feature space that is relevant to speech perception:

where determines the type of distance, and the weights are constrained to sum to 1, . In total, this affords the exemplar model *M+1* degrees of freedom for *M* features (the scaling factor *s*, the distance metric , and *M*-1 feature weights).

We evaluate this model in terms of its ability to predict listeners’ perceptual responses in two different experiments on vowel perception, depending on whether the perceptual features are talker-normalized prior to exemplar-based categorization.

# Study 1 – Exploring the exemplar space

We begin by comparing various instantiations of the exemplar model without talker-normalization. The best of these models will provide the most conservative baseline for our research question—addressed in Study 2—whether talker normalization provides a better account of listeners’ speech categorization even when listeners are assumed to store rich exemplars. Specifically, we manipulate:

1. *What features are included in the model:* the **F1-F2 model** includes only the two primary formant cues to vowel identity in US English (F1 and F2); the **F1-F3 model** adds F3, which has been argued to capture information about lip rounding (REF); the **spectro-temporal model** adds the fundamental frequency f0 and vowel duration.

For each of these three models, we consider two variants: **static** **models** use formant measures (F1-F3) that estimate the stable formants in the center of the vowel (as remains the standard in the field); **dynamic models** include formant measures from the start, center, and end of the vowel. The latter model can thus capture information formant dynamics, which are known to affect perception (REF).

1. *Distance metric:* **We consider .** A is expected to be most effective for separable feature dimensions, whereas is expected to more effective for integral dimensions (Nosofsky, 1986).
2. *Similarity scaling:* **We consider** , the first two of which have been considered in previous work (Apfelbaum & McMurray, 2015).

This resulted in 3 (features included) x 2 (static vs. dynamic) x 2 () models x 3 (*s*) = 36 models. We determined the best-fitting feature weights through optimization, as described below.

## Methodology

Next, we describe the cross-validation used to optimize the fit of each model, while simulating individual differences between listeners’ experience (Figure 1).

En bild som visar text, skärmbild, Teckensnitt, linje

AI-genererat innehåll kan vara felaktigt.

Figure : Cross-validation approach

**Exemplar data.**To approximate the types of vowel exemplars a typical L1 listener of US English might have stored in their memory, we use a database of 1,240 US English vowel recordings (Xie & Jaeger, 2020). The data comprise recordings of eight monophthong vowels ([ɑ], [æ], [ʌ], [ɛ], [ɪ], [i], [ʊ], [u]) in the context of *hVd* words (e.g., “heed”, “hid”, etc.) from 14 L1 talkers from the Northeast of the US (12 male, 5 female). We excluded one talker that was used in the experiment we report below. The data include the vowel durations, mean f0 during the vowels, and formant measurements (F1, F2, F3) 35%, 50%, and 65% into the vowel segment.

As part of the cross-validation, we randomly split talkers into five bins of 2-3 talkers each, and created five unique folds by combining bin 1-4, bin 2-5, bin 1,3-6 etc. Each fold thus consists of XXX-XXX exemplars from XXX-XXX talkers. Together, the five folds provide a (very crude) estimate of the type of variability in exemplar knowledge that might be expected across listeners of the dialect.

**Perception data.**To evaluate the predictions of all models, we use publicly available data from two perception experiments (Experiment 1a and 1b, Persson et al., 2024). In each experiment, listeners listened to one talker produce the same eight monophthong vowels in the same *hVd* word context as in the database described in the previous section. On each trial listeners, answered which word (and thus vowel) they heard, using an 8-way response grid. Experiment 1a consisted of 72 natural recordings (9 instances of each vowel) from the talker excluded above. Experiment 1b consisted of XXX resynthesized recordings that spanned that entire vowel space (for details, see Persson et al., 2024). The data contain the same acoustic information, obtained under the same annotation procedure, as the data in the previous section.

Twenty-eight and 31 participants completed Experiments 1a and 1b, respectively, each hearing two instances of each stimulus, for a total of 3983 and 8970 categorization responses, respectively. For cross-validation, we randomly split the participants in each experiment into two bins, creating a *training fold* for optimize the free parameters () against listeners’ responses, and a *test fold* to assess the fit on held-out data.

**Data preparation.** We log-transformed all duration, f0 and formant values since this is known to both better approximate the neural representations of frequency information in the human brain (REF), and to better fit human behavior (REF).

To increase interpretability of feature weights, and to aid model fitting, we followed previous work and z-score the features . For this, we calculated the mean and SD of each feature over the combined exemplar and perception data. As z-scoring is a linear transform, it does not affect the results.

**Model fitting.**Sequential least squares programming was used to determine the feature weights that maximize the log-likelihood of listeners’ responses in the training data, separately for each of the five exemplar folds. Distributed over XXX cores on a BlueHive cluster, the 180 optimizations (=36 models x 5 folds) took XXX hours to complete.

## Results

Log-likelihoods on the test data did not significantly differ from log-likelihoods on the test data, suggesting that we did not overfit to the data (mixed-effect logistic regression with random intercepts by fold, *p*>XXX).

As shown in Figure 3, …

XXX

Figure : Mean per-response log-likelihood of listeners’ vowel categorizations in held-out test data

[Core findings: Is dynamic better than static? Does number of features matter? Is one tau and/or s reliably better?]

# Study 2: Does normalization help?

Based on the results of Study 1, we decided to estimate the effects of normalization for XXX models: … [explain choice: F1-F2 only because it’s often used]

We took the exact same approach as in Study 1, except that we included the scaling parameter *s* in the optimization.

Normalization to consider: C-CuRE over log-transformed features (same as Nearey log-mean) and/or Nearey uniform scaling (single log-mean).

# Old Results

**Parameter Optimization**

The optimization process revealed interesting patterns in the optimal parameter values across models. The feature weights showed systematic differences between normalized and unnormalized models, suggesting different relative importance of F1 and F2 in these contexts. The similarity scaling parameter demonstrated sensitivity to speech type (natural vs. synthetic).

**Optimization Results:**

Table 1: Optimized models parameters

| Model | w1 | w2 | Similarity Scaling |
| --- | --- | --- | --- |
| No norm (1a) | 0.5158 | 0.4815 | 4.4158 |
| No norm (1b) | 0.5426 | 0.4574 | 4.1211 |
| Norm (1a) | 0.4907 | 0.5093 | 4.4165 |
| Norm (1b) | 0.5151 | 0.4829 | 3.4938 |

**Model Performance**

**Mean Tuning Log Likelihoods:**

All the models converged to an optimal mean log likelihood over the tuning data fold from the perceptual dataset. The following table summarizes these values.

Table 2: Log-Likelihood on tuning fold

| Model | Tuning Log Likelihood | |  |
| --- | --- | --- | --- |
| No norm (1a) | | -2947.9452 | |
| No norm (1b) | | -6614.4842 | |
| Norm (1a) | | -2990.8505 | |
| Norm (1b) | | -6335.1150 | |

Table 3: Log-Likelihood on held-out fold

| Model | Tuning Log Likelihood | |  |
| --- | --- | --- | --- |
| No norm (1a) | | -3003.4248 | |
| No norm (1b) | | -7752.9370 | |
| Norm (1a) | | -2959.1136 | |
| Norm (1b) | | -7594.9935 | |

**No normalization (Exp 1a)**

**A screenshot of a computer screen

Description automatically generated**

Figure 5: No normalization (Exp 1a)

**A screenshot of a computer screen

Description automatically generated**

Figure 6: No normalization (Exp 1b)

A screenshot of a computer screen

Description automatically generated

Figure 7: Normalization (Exp 1a)

**A screenshot of a computer screen

Description automatically generated**

Figure 8: Normalization (Exp 1a)

# Discussion

The implementation successfully demonstrates the relative merits of exemplar-based and normalization approaches to speech perception. The results suggest that while both approaches capture important aspects of perception, if we combine them (as in MODEL-2) then we do not get a clear improvement in model performance.

The technical implementation highlights several key insights:

* The importance of proper parameter optimization
* The role of cross-validation in ensuring robust results
* The value of comprehensive visualization in understanding model behavior

# Conclusion

This implementation provides a rigorous framework for comparing speech perception models. The results suggest that incorporating CCure- normalization into exemplar-based models does not offer advantages for vowel dataset. Future work could explore additional normalization techniques and extend the model to other phonetic contrasts.

# References

Apfelbaum, K. & McMurray, B. (2015). XXX

Nosofsky

Persson, S., et al. (2024). XXX