A Framework for Evaluating Speech Representations

Caitlin Richter[1], **Naomi H. Feldman**[2*], **Harini Salgado**[3], **Aren Jansen**[4]
[1]Department of Linguistics, University of Pennsylvania, Philadelphia, PA 19104
[2]Department of Linguistics and UMIACS, University of Maryland, College Park, MD 20742
[3]Pomona College, Claremont, CA, 91711
[4]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD 21211
*Address for correspondence: nhf@umd.edu

Abstract

Listeners track distributions of speech sounds along perceptual dimensions. We introduce a method for evaluating hypotheses about what those dimensions are, using a cognitive model whose prior distribution is estimated directly from speech recordings. We use this method to evaluate two speaker normalization algorithms against human data. Simulations show that representations that are normalized across speakers predict human discrimination data better than unnormalized representations, consistent with previous research. Results further reveal differences across normalization methods in how well each predicts human data. This work provides a framework for evaluating hypothesized representations of speech and lays the groundwork for testing models of speech perception on natural speech recordings from ecologically valid settings.

Keywords: speech perception, speaker normalization, Bayesian modeling, approximate inference

Listeners track statistical distributions of sounds in their language. Adults are sensitive to these distributions when perceiving speech (Clayards, Tanenhaus, Aslin, & Jacobs, 2008), and infants' discrimination is influenced by these distributions (Maye, Werker, & Gerken, 2002). Statistical properties of the input can differ depending on the dimensions that listeners extract from the speech signal. For example, acoustic characteristics of vowels are highly variable when represented by their formant frequencies, but the variability is greatly reduced when they are represented by the z-score of their formant frequencies relative to other vowels by the same speaker (Figure 1; see also Cole, Linebaugh, Munson, & McMurray, 2010). Because the distributional characteristics of speech depend so heavily on the dimensions used, understanding the dimensions that listeners extract from the speech signal is a critical part of understanding phonetic learning and perception.

In this paper we introduce a novel approach to evaluating hypotheses about the dimensions that support listeners' perception. We adopt a cognitive model of speech perception from Feldman, Griffiths, and Morgan (2009), which predicts that listeners' perception is biased toward peaks in the acoustic distribution of sounds in their input. This model provides a formal link between the distribution of sounds in the input and listeners' discrimination abilities. We measure the input from a speech corpus and use the model to predict listeners' discrimination behavior. Different ways of representing the speech signal imply different distributions of sounds in the corpus, yielding different predictions about listeners' discrimination. We are interested in learning which representations of speech best predict listeners' actual discrimination.

We model AX discrimination tasks, in which listeners decide whether two sounds are acoustically identical. Previous

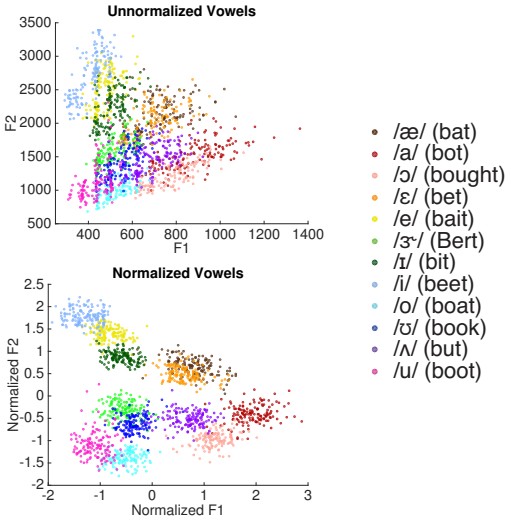

Figure 1: Acoustic characteristics of vowels produced in hVd contexts by men, women, and children from Hillenbrand et al. (1995), plotted as raw formant frequencies (top) and z-scored formant frequencies (bottom). If listeners' perception is biased toward peaks in these distributions, these feature spaces make different predictions about listeners' performance in perceptual discrimination tasks.

approaches to evaluating speech dimensions have instead focused on categorization tasks, in which listeners decide which category a sound belongs to (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011). Discrimination provides several advantages over categorization: it is a more fine-grained measure than categorization, and can be reliably measured in both adults and infants, even when listeners do not have well-formed categories for a given set of sounds. In addition, whereas building a categorization model requires speech recordings to have been labeled with phoneme identities, building a discrimination model does not, as we explain below.

As an initial case study, we examine speech dimensions defined by two speaker normalization techniques: z-scoring, which has been proposed in the cognitive science literature (e.g., Lobanov, 1971), and vocal tract length normalization (VTLN), which is widely used in automatic speech recognition (ASR) systems (e.g., Wegmann, McAllaster, Orloff, & Peskin, 1996). We find that both normalization methods yield a robust improvement over unnormalized representations in predicting listeners' discrimination, consistent with previous research. We also find that VTLN predicts human data better than z-

scoring, despite being less effective at eliminating speaker variability. These results illustrate our method for evaluating these dimensions against discrimination data and provide clues to the dimensions that guide listeners' perception. Adapting a cognitive model to operate over speech recordings also lays the groundwork for testing models of speech perception in more ecologically valid settings, by enabling cognitive scientists to make use of the same rich corpus data that is often used by researchers working in automatic speech recognition.

Speaker normalization

We begin by characterizing the speech representations tested in this paper. Speech contains commingled effects of linguistic, paralinguistic, and purely physical sources of variation. The goal of speaker normalization is to find representations that diminish some of the variability in the speech signal, like that from the speaker's body, while retaining task-appropriate information such as the variability that represents different phonemes. While some models have questioned whether listeners normalize across speakers (Johnson, 1997), most evidence suggests some degree of normalization. Normalized representations have been found to improve phonetic categorization (Cole et al., 2010; Nearey, 1978), increase a categorization model's match with human behavior (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011), and improve the performance of speech recognizers (Wegmann et al., 1996; Povey & Saon, 2006). We test the effects of two specific methods for speaker normalization, z-scoring and vocal tract length normalization, in a cognitive model of vowel discrimination.

Within-speaker z-scoring (Lobanov, 1971) has been suggested for descriptive sociolinguistic research (Adank, Smits, & Hout, 2004), as it highlights learned linguistic content while removing speaker-body confounds from vowel formants. A related manipulation, linear regression, has also been shown to improve cognitive models of fricative perception (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011).

Vocal tract length normalization (VTLN) is a technique developed for automatic speech recognition. VTLN compensates for speaker differences in vocal tract length by stretching or compressing the frequency axis of the productions of each speaker (Wegmann et al., 1996). The aim of VTLN is to adjust the corpus so that it is as if all the speakers had identical vocal tract lengths. VTLN is widely successful in ASR systems, where it substantially decreases the word error rate (e.g., Giuliani, Gerosa, & Brugnara, 2006). In performing this normalization, we use a procedure from Wegmann et al. (1996) adapted for an unsupervised setting, which selects frequency adjustments for speakers on the basis of their /i/ productions by maximising the similarity of all /i/ tokens across speakers.

We apply both normalization methods to vowels that are represented by mel frequency cepstral coefficients (MFCCs; Davis & Mermelstein, 1980). MFCCs are widely employed as an input representation in ASR systems (although we do not implement an ASR system here). MFCCs are a 12-dimensional vector that describe a timepoint of speech by capturing information about the spectral envelope, reflecting vocal tract shape. Thus, they capture information similar to formant frequencies, but have the advantage that they can be computed automatically from the speech signal, without being subject to the error inherent in automatic formant tracking.[1]

Table 1: Effect of normalization on K-L divergence.

	MFCCS (unnormalized)	Z-score normalized	VTLN normalized
Gender KLDiv	7.84	4.58	6.14
Dialect KLDiv	4.41	2.09	4.19

Effects of normalization Our method for testing hypothesized representations of speech against human perception relies on the idea that different representations of speech yield different distributions of sounds in the input. To examine the distribution of sounds in the input, we computed MFCCs, z-scored MFCCs, and MFCCs with VTLN from vowel recordings in the Vowels subcorpus of the Nationwide Speech Project (NSP; Clopper & Pisoni, 2006). This corpus contains ten different vowels pronounced in the context hVd (hid, had, etc.) by 5 female and 5 male speakers from each of 6 dialect regions of the United States. Each of these 60 speakers repeats each of the 10 hVd words 5 times, for a total of 3000 hVd tokens balanced across vowel, gender, and dialect.[2]

We characterize the effects of each normalization method on the distribution of vowels in the NSP corpus by computing symmetrized Kullback-Leibler divergence, a measure of difference between two probability distributions (Wang, Kulkarni, & Verdú, 2006). Lower K-L divergence indicates greater similarity between two distributions. We estimated K-L divergence across gender and across dialect (averaged over 15 pairwise comparisons of 6 dialect regions). Male and female speakers differ in their vocal tract lengths, and thus normalization algorithms would be expected to increase similarity between their vowel productions. Dialects also differ in their pronunciations of different vowels; although this variation is not related to vocal tract length, it may nevertheless be impacted by normalization algorithms that seek to neutralize speaker differences.

K-L divergence between genders and between dialect pairs is highest in MFCCs with no speaker normalization, reflecting the effects these factors have on the original acoustic signal (Table 1). Both VTLN and z-scoring reduce K-L divergence between genders, as predicted, so that male and female speakers saying the same vowel appear more similar after either of these normalizations than they are in unnormalized MFCCs. Z-scoring using all 10 NSP vowels also reduces K-L divergence between dialect pairs. In contrast, VTLN matching

[1] Z-scoring has previously been applied to formant frequencies, but we show in the next section that it is also effective at normalizing across speakers when applied to MFCCs. Vowel-intrinsic normalization methods such as formant ratios were not tested here because they are not straightforward to apply to MFCCs.

[2] While this corpus does not correspond exactly to listeners' experience with language, conducting initial simulations with a corpus of vowels in neutral contexts allows us to investigate algorithms for speaker normalization while sidestepping issues of how listeners generalize across phonological contexts.

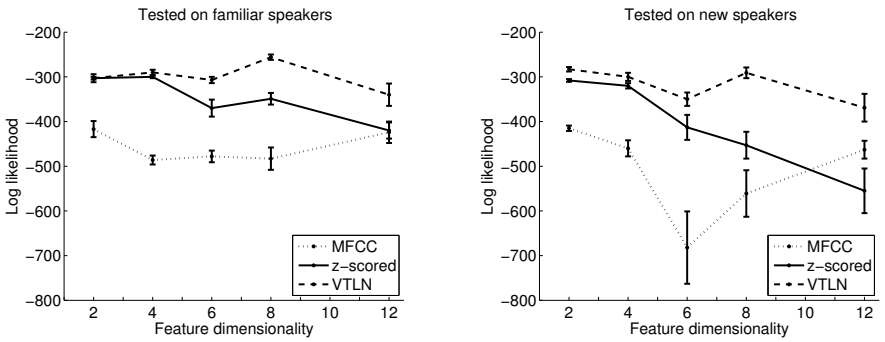

Figure 2: Model fit to human data, when generalizing to familiar speakers (left) and new speakers (right). Results for 1-dimensional features are not shown on this scale as they are extremely poor, with log likelihoods of -1300 to -1600.

across speakers on the basis of /i/, which differs little across the dialects (Clopper & Pisoni, 2006), has minimal effects on dialect K-L divergence: cross-dialect differences remain nearly as distinct after VTLN. Overall, when z-scoring by speaker K-L divergence is lowest (though nonzero; gender and dialect information remains in the representation); VTLN removes somewhat less of the gender and dialect variation.

In summary, MFCCs, z-scored MFCCs, and MFCCs with VTLN each correspond to different distributions of vowels in the input, with z-scoring being the most effective at increasing the overlap between the distributions of vowels spoken by different speakers. The next section describes a cognitive model that uses these input distributions to quantitatively predict listeners' vowel discrimination in the laboratory.

Cognitive model

The model of discrimination we adopt has been shown to accurately predict listeners' discrimination of both vowels and consonants (Feldman et al., 2009; Kronrod, Coppess, & Feldman, 2012), but has not yet been implemented directly on speech recordings. The model formalizes speech perception as an inference problem. Listeners perceive sounds by inferring the acoustic detail of a speaker's target production through a noisy speech signal. Because listeners need to correct for uncertainty in the speech signal, their perception is biased toward acoustic values that have high probability under their prior distribution. This creates a dependency between the listeners' prior distribution and their perception of sounds.

Formally, speakers and listeners share a prior distribution over possible acoustic values that can be produced in the language, $p(T)$. Prototypical sounds in the language have highest probability under this distribution, but the distribution is nonzero over a wide range of acoustic values, corresponding to all the ways in which speech sounds might be realized. When producing a sound, speakers are assumed to sample a target production T from this distribution. The target production can carry meaningful information aside from category identity, such as dialect information or coarticulatory information about upcoming sounds, making its acoustic value something that listeners wish to recover. The stimulus S heard by listeners is similar to the target production T, but is assumed to be corrupted by a noise process defined by a Gaussian likelihood function, $p(S|T) = \mathcal{N}(T, \Sigma_S)$. Both T and S are d-dimensional vectors, where d is the dimensionality of the feature space. In our simulations, the feature space is defined by either MFCCs, z-scored MFCCs, or MFCCs with VTLN.

Listeners hear S and reconstruct T by drawing a sample from the posterior distribution, $p(T|S) \propto p(S|T)p(T)$. We refer to a listener's sample from the posterior distribution as a percept. The percept is a continuous acoustic value, rather than a category label; this is consistent with a large body of evidence showing that listeners recover fine-grained acoustic detail from the speech signal (e.g., Pisoni & Tash, 1974).

The model can be used to predict listeners' discrimination behavior in the laboratory. In AX discrimination tasks, listeners hear two stimuli and decide whether they are acoustically identical. The model assumes that for each stimulus, listeners sample a percept from their posterior distribution, $p(T|S)$. They then compute the distance between their percepts of the two stimuli and compare it to a threshold ε. If the percepts are separated by a distance less than ε, listeners respond *same*; otherwise they respond *different*. Given these assumptions, the proportion of *same* responses for two stimuli, S_1 and S_2, is predicted to follow a binomial distribution whose parameter is the probability that the percepts for the two sounds are within a distance ε of each other, $p(|T_1 - T_2| < \varepsilon | S_1, S_2)$. The noise covariance Σ_S and the response threshold ε are free parameters which are optimized to best predict discrimination data.

Whereas previous work with this model has estimated listeners' prior distribution from perceptual categorization data, here we estimate listeners' prior distribution directly from production data in the NSP corpus; we make the simplifying assumption that the prior distribution directly mirrors the distribution of sounds in the input. Previous work has also assumed that listeners' prior distribution is a mixture of Gaussians, with one Gaussian distribution corresponding to each phonetic category. We avoid making this assumption by using an exemplar-based implementation of the model.

Shi, Griffiths, Feldman, and Sanborn (2010) showed that exemplar models provide a way of approximating Bayesian inference. Specifically, exemplar models implement a form of approximate inference known as importance sampling. To

use importance sampling for our simulations, we need a set of exemplars $\{T^{(i)}\}$ that are sampled from listeners' prior distribution $p(T)$. We assume the vowels in the NSP constitute this set of exemplars. We then weight each exemplar by its likelihood with respect to a stimulus S, $p(S|T^{(i)})$, and select an exemplar according to its weight. An exemplar from the corpus sampled in this way behaves as though it were drawn from the posterior distribution $p(T|S)$. This method does not require us to know a parametric form for the prior distribution $p(T)$, because the prior distribution is represented only through exemplars. In addition, it does not require the exemplars from the corpus to have category labels, as the weights $p(S|T^{(i)})$ are defined by the model's Gaussian likelihood function, corresponding to the speech signal noise.

We estimate the model's probability of responding *same* on each trial by using importance sampling to obtain 100 pairs of percepts corresponding to the pair of stimuli presented in that trial. The proportion of these pairs of percepts that are within distance ε of each other provides an estimate[3] of the probability of responding *same* on that trial. We use these probabilities to predict listeners' actual *same-different* responses in an experiment. We implement the model several times with different speech representations: MFCCs, z-scored MFCCs, or MFCCs with VTLN. Comparing model likelihoods across the three speech representations allows us to ask which representations best predict listeners' discrimination.

Simulations

Simulations implemented the perceptual model with normalized and unnormalized representations, comparing model predictions to human discrimination data. To the extent that different representations of speech yield different distributions of sounds in a corpus, they should make different predictions about the biases that listeners will exhibit in a speech perception experiment. Representations that yield more accurate predictions can be assumed to contain information more similar to the dimensions that listeners use in speech perception.

Human perceptual data We use vowel discrimination data from an AX discrimination task conducted by Feldman et al. (2009) in quiet listening conditions. Twenty participants heard a continuum of 13 isolated vowels that were synthesized to simulate a male speaker. First and second formants of these stimuli ranged linearly in mels from /i/ (as in 'beet') to /e/ (as in 'bait'). Participants heard all ordered pairs of stimuli and judged whether each pair was acoustically identical. MFCCs, z-scored MFCCs, and MFCCs with VTLN computed from these thirteen stimuli serve as input to the model, as the stimulus S. Model predictions are then compared with listeners' *same-different* responses to each pair of stimuli.

Speech representations The exemplars that serve as the model's prior distribution are vowels from the Vowels subcorpus of the NSP. Vowels were represented either as MFCCs, z-scored MFCCs, or MFCCs with VTLN, computed at their midpoint. Although speech recognition systems typically use 12 MFCC dimensions, we additionally include simulations that omit subsets of the higher dimensions, as the lower dimensions are better able to capture information from formants traditionally used to describe vowel quality.

The NSP is an ideal case for the z-scoring normalization, because each speaker says the same tokens the same number of times. However, MFCCs for the stimulus 'speaker' were only available for the /i/-/e/ vowel continuum. Because the stimuli were originally synthesized according to average formant values for male speakers, we handled this missing data by normalizing the stimuli according to average z-scoring factors of the 30 male NSP speakers. Due to the reliance of our VTLN procedure on only /i/ tokens, this normalization was straightforward to apply to the stimuli.

Fitting parameters The NSP corpus was divided into two balanced, equally sized sets of exemplars. Two methods were used for dividing the corpus. In one case, the two halves contained different exemplars from the same speakers, while in the other case the two halves contained exemplars from different speakers (balanced for gender and dialect region). Each division of the corpus was created once, and used for simulations with all speech feature types.

In each simulation, one set of exemplars was used to fit parameters: the response threshold ε and the noise variance Σ_S (constrained to be diagonal) were selected using Markov chain Monte Carlo to maximize model likelihoods given perceptual data.[4] The other set of exemplars was used to compute model likelihoods at test. The roles of the two sets of exemplars were then reversed, resulting in 2-fold cross-validation. Each set of exemplars served as a test set for 10 simulations. Points and error bars in Figures 2 and 4 represent means and standard error calculated across all 20 simulations; the relatively small error bars indicate that results were consistent across replications.[5]

Results Model performance is assessed by computing the log likelihood of the model, given the human data. Higher log likelihoods indicate a closer match to perceptual data.

Previous work with categorization models suggests that normalized representations are more consistent with listeners' perception than unnormalized representations (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011) . We replicate this result with our method: in almost all cases, unnormalized MFCCs have the lowest likelihoods among the three representations tested (Figure 2). We also find that MFCCs normalized by VTLN outperform z-scored MFCCs, although z-scoring within speakers neutralizes more inter-speaker variation as measured by K-L divergence (Table 1). Thus, the better performance of the VTLN models is not merely an artifact of acoustic similarity among vowels in the corpus; this also may imply that human representations underlying vowel discrimi-

[3] We use add-one smoothing to compute this estimate.

[4] Z-scoring and VTLN were applied prior to parameter fitting; therefore neither adds free parameters to the model, and no optimization to fit human data is involved in their application.

[5] Numerical values fit for model parameters were also consistent across the 20 replications for each speech feature type.

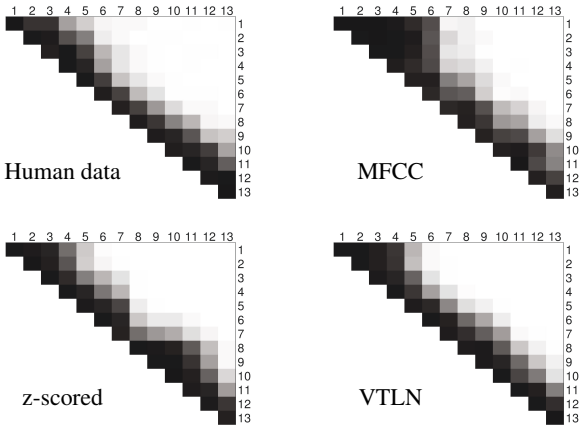

Figure 3: Confusion matrices for how often each pair of stimuli is judged to be the same (black) vs. different (white) by humans and by 4-dimensional models tested on familiar speakers. Axis labels denote stimulus numbers from an AX trial.

nation do not completely normalize across speakers. Examples of *same-different* responses from models using each type of feature are shown in Figure 3, together with human data.

Models using speech representations of one dimension are always extremely poor; the first cepstral coefficient on its own does not provide enough information for the perceptual task. In some simulations, such as for z-scored features, test likelihood decreases with higher numbers of dimensions beyond a certain point, likely due to overfitting of parameters to particular exemplars.[6] It appears that the lower MFCC dimensions, particularly the second dimension, contain information relevant to listeners' discrimination of an /i/-/e/ continuum.

Although the NSP's speaker and phoneme labels are not used by our cognitive model, we take advantage of this information in analyses for insight on the types of exemplars sampled as percepts by the model. Across all models, the 100 percepts drawn from the posterior distribution for each stimulus contained on average 30 different exemplars, indicating that a number of exemplars (from different speakers) are treated as linguistically similar to each other. As a measure of model quality and interpretability, we examine the identity of the exemplars sampled by the model during each perceptual judgment (Figure 4). The percentage of samples that belong to the classes of vowels along the /i/-/e/ continuum (NSP 'heed', 'hayed', and 'hid' tokens; henceforth referred to as high front vowels) gives information on model quality, because all the stimuli are perceived by US English speakers as falling along this continuum. The proportion of times a model samples female exemplars to recover the linguistic target for this experiment's male-speaker stimuli gives an indication of the model's ability to generalize linguistic content across genders.

Models using unnormalized MFCCs tend to make the least use of female exemplars, indicating that this representation does not recognize very much similarity between male and female speakers saying the same vowel. Models with z-scored features are closest to sampling 50% female exemplars; this confirms that the z-scored representation is highly effective at neutralizing difference between speakers of different genders (Table 1), although it is not the representation that gives the best match to human perceptual performance (Figure 2).

While simulations with 2 through 6 dimensions consistently treated the experimental stimuli as being most similar to high front vowels in the corpus, simulations with higher orders of cepstral coefficients did not (Figure 4), reinforcing the importance of the lower MFCC dimensions in capturing listeners' perception of these stimuli. We suspect that this behavior emerged due to the artificial synthesis of the experimental stimuli, which resulted in these high front vowel stimuli being most similar to low back vowels from the corpus in two of the higher MFCC dimensions. This can cause the model to perceive stimuli as low back vowels in cases where it generalizes along those dimensions. This underscores the difficulty of bringing together ecologically valid speech corpora with the type of controlled stimuli typically used in experimental settings, and illuminates areas in which future research may provide insight by addressing these challenges.

Finally, we can compare the log likelihoods from Figure 2 to a benchmark showing ideal model performance. Feldman et al. (2009) estimated listeners' prior distributions for the /i/ and /e/ categories from perceptual categorization data, rather than from speech recordings. Using their estimate, the model yields average log likelihood of -233, somewhat higher than those obtained here. A corpus-based model should approach this value as the distribution of sounds in the corpus approaches the prior distribution that listeners use in perceptual tasks.

Discussion

In this paper a novel method was introduced for evaluating hypotheses regarding the dimensions that guide listeners' speech perception. A cognitive model of AX discrimination was implemented directly on speech recordings and used to evaluate two speaker normalization methods. Both normalization methods improved the model's fit to perceptual data, consistent with previous research. Between the two normalization methods, VTLN outperformed z-scoring, despite being less effective at collapsing gender and dialect differences.

The advantage of VTLN over z-scoring in modeling human perceptual data suggests that VTLN allows the model to generalize across speakers in a way that is more similar to human perceivers. For example, listeners may track statistical distributions of speech in ways that allow them to collapse across gender while retaining differences across dialects. Nevertheless, a prior distribution estimated from perceptual categorization data still outperforms the prior distributions measured from a corpus; none of the three representations tested here match human perception exactly. Instead, these simulations provide initial clues to the dimensions that guide listeners' perception

[6] The corresponding training likelihoods for z-scored dimensions remained stable or even increased at higher numbers of dimensions. Similarly, the low likelihood observed at six dimensions for MFCCs was due to several runs that achieved high likelihood on the training exemplars and low likelihood on the test exemplars.

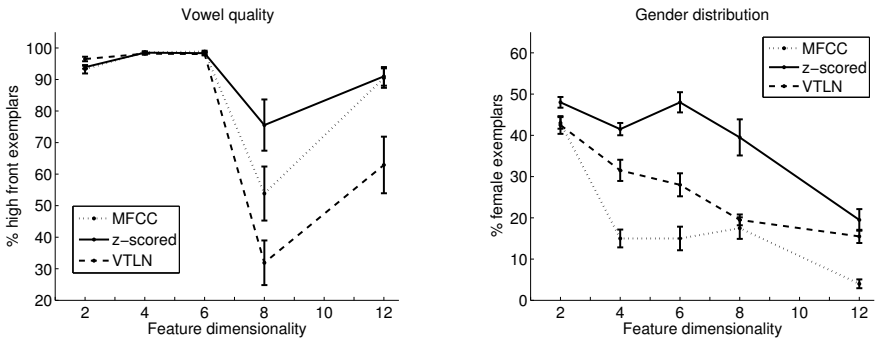

Figure 4: Secondary evaluations, showing how often the stimuli were correctly perceived as high front vowels (left) and how often the model based perceptions on female exemplars (right). Data are averaged across the familiar and new speaker test cases, which were very similar on these measures.

while also allowing us to validate a novel method for assessing speech representations against human data.

Our method provides a general tool for investigating proposals regarding the dimensions that guide listeners' perception. The model's prior distribution can in principle be estimated from any speech corpus, and does not require phoneme labels. This ability to make use of unlabeled corpora provides an advantage over evaluation methods that rely on categorization. In addition, the importance sampling approximation used here can be implemented on any speech representation for which a likelihood function $p(S|T)$ can be computed between stimuli and exemplars and for which a distance metric between exemplars in the corpus can be compared to a threshold ε, and thus can be used even for representations that lack a fixed or finite set of dimensions. This flexibility makes it a promising tool for exploring cross-linguistic differences in listeners' sensitivity to perceptual dimensions, as well as for evaluating theories of dimension learning against children's discrimination data.

To our knowledge, this is also the first time the model from Feldman et al. (2009) has been implemented on speech recordings. Modifying models to operate over corpora of natural speech allows them to make use of ecologically valid datasets, and can thus facilitate a richer understanding of the way in which listeners' perception is shaped by their environment.

Acknowledgments We thank Josh Falk for help piloting the model, Phani Nidadavolu for help computing speech features, and Hynek Hermansky, Bill Idsardi, Feipeng Li, Vijay Peddinti, Amy Weinberg, and the UMD probabilistic modeling reading group for helpful comments. This work was supported by NSF grant BCS-1320410.

References

Adank, P., Smits, R., & Hout, R. v. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, *116*, 3099-3107.

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

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804-809.

Clopper, C. G., & Pisoni, D. B. (2006). The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, *48*, 633-644.

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167-184.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of the IEEE*, 357-366.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752-782.

Giuliani, D., Gerosa, M., & Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech & Language*, *20*(1), 107–123.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099-3111.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (p. 145-165). New York: Academic Press.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A unified model of categorical effects in consonant and vowel perception. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, *49*, 606-608.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219-246.

Nearey, T. M. (1978). *Phonetic feature systems for vowels* (Vol. 77). Indiana University Linguistics Club.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, *15*, 285-290.

Povey, D., & Saon, G. (2006). Feature and model space speaker adaptation with full covariance Gaussians. *Proceedings of Interspeech*.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*, 443-464.

Wang, Q., Kulkarni, S. R., & Verdú, S. (2006). A nearest-neighbor approach to estimating divergence between continuous random vectors. *Convergence*, *1000*(1), 11.

Wegmann, S., McAllaster, D., Orloff, J., & Peskin, B. (1996). Speaker normalization on conversational telephone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 339-341.