Evaluating Low-Level Speech Features Against Human Perceptual Data

Caitlin Richter
Dept. of Linguistics
University of Pennsylvania
ricca@sas.upenn.edu

Naomi H. Feldman
Dept. of Linguistics and UMIACS
University of Maryland
nhf@umd.edu

Harini Salgado
Dept. of Computer Science
Pomona College
harini.salgado@gmail.com

Aren Jansen
HLTCOE
Johns Hopkins University
arenjansen1@gmail.com

Abstract

We introduce a method for measuring the correspondence between low-level speech features and human perception, using a cognitive model of speech perception implemented directly on speech recordings. We evaluate two speaker normalization techniques using this method and find that in both cases, speech features that are normalized across speakers predict human data better than unnormalized speech features, consistent with previous research. Results further reveal differences across normalization methods in how well each predicts human data. This work provides a new framework for evaluating low-level representations of speech on their match to human perception, and lays the groundwork for creating more ecologically valid models of speech perception.

Index Terms: speech perception, automatic speech recognition, speaker normalization, cognitive models

1 Introduction

Understanding the features that listeners extract from the speech signal is a critical part of understanding phonetic learning and perception. Different feature spaces imply different statistical distributions of speech sounds in listeners' input (Figure 1), and these statistical distributions of speech sounds influence speech perception in both adults and infants (Clayards et al., 2008; Maye et al., 2002).

The performance of automatic speech recognition (ASR) systems is also affected by the way in which these systems represent the speech signal. For example, changing the signal processing methods that are used to extract features from the speech waveform (Hermansky, 1990; Hermansky and Morgan, 1994) and applying speaker normalization techniques to these features (Wegmann et al., 1996; Povey and Saon, 2006) can improve the performance of a recognizer. Recently there has been considerable interest in representation learning, in which new features that are created through exposure to data from the language to be recognized, or through exposure to other languages, lead to better system performance (Grézl et al., 2014; Heigold et al., 2013; Kamper et al., 2015; Sercu et al., 2016; Thomas et al., 2012; Wang et al., 2015). It is potentially useful to know how closely the feature representations in ASR resemble those of human listeners, particularly for low-resource settings, where systems rely heavily on these features to guide generalization across speakers and dialects.

In this paper we introduce a method for measuring the correspondence between low-level speech feature representations and human speech perception. This allows us to compare different feature spaces in terms of how well the locations of sounds in the space can predict listeners' perception of acoustic similarity. Our method has potential relevance both in ASR, for understanding how well the feature representations in ASR systems capture the similarity structure that guides human perception, and in cognitive science, for evaluating hypotheses regarding the feature spaces that human listeners use.

We evaluate the ability of a feature space to capture listeners' *same-different* responses in AX discrimination tasks, in which listeners hear two sounds and decide whether they are acoustically identical. Rather than assuming that listeners' ability to discriminate

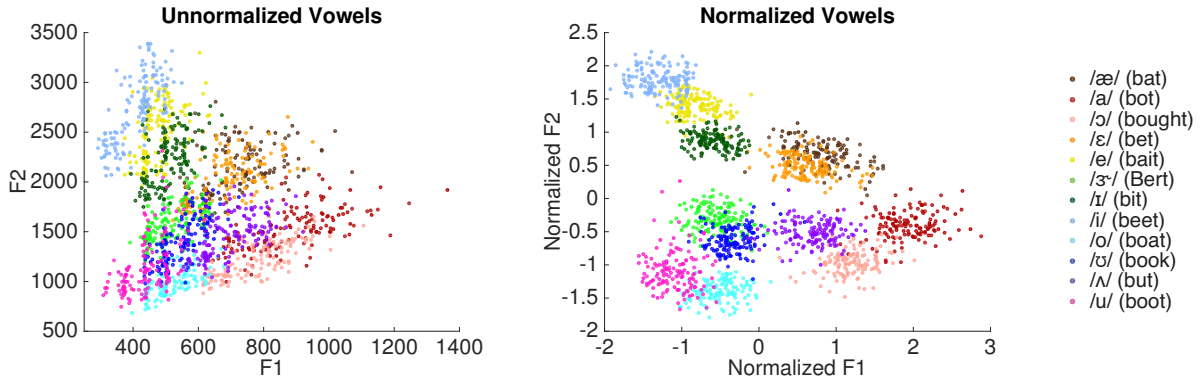

Figure 1: Acoustic characteristics of vowels produced in hVd contexts by men, women, and children from Hillenbrand et al. (1995), plotted as raw formant frequencies (left) and z-scored formant frequencies (right). These feature spaces yield different distributions of sounds in acoustic space. If listeners' perception is biased toward peaks in these distributions, the feature spaces make different predictions for listeners' behavior in perceptual discrimination tasks.

sounds is directly related to those sounds' distance in a feature space, we instead adopt a cognitive model of speech perception which predicts that listeners' perception of sounds is biased toward peaks in the distribution of sounds in their input. This leads listeners to perceive some sounds as closer together in the feature space than they actually are, and others as farther apart (Figure 2). The model has previously been shown to accurately predict listeners' *same-different* discrimination judgments when listening to pairs of sounds (Feldman et al., 2009; Kronrod et al., 2016).

Our innovation in this work is to estimate the distribution of sounds in the input from a corpus, using different feature spaces. Under the model, listeners are expected to bias their perception toward peaks in the distribution of sounds in the corpus. Those peaks occur in different locations for different feature spaces. Thus, different feature representations yield different predictions about listeners' discrimination. Features that yield a better match with listeners' discrimination are assumed to more closely reflect the way in which listeners generalize their previous experience when perceiving speech.

In addition to providing a way to evaluate feature representations, adapting a cognitive model to operate directly over speech recordings lays the groundwork for building more ecologically valid models of speech perception, by enabling cognitive scientists to make use of the same rich corpus data that is often used by researchers in ASR. These speech corpora will allow models to be trained and tested on data that more faithfully simulate the speech environment that listeners encounter, rather than on artificially simplified data.

As an initial case study, we use the perceptual model to evaluate features derived from two speaker normalization algorithms, which aim to reduce variability in the speech signal stemming from physical characteristics of the vocal tract. We compare these normalized features to a baseline set of unnormalized features. Speaker normalization has previously been found to improve phonetic categorization (Cole et al., 2010; Nearey, 1978), increase a phonetic categorization model's match with human behavior (Apfelbaum and McMurray, 2015; McMurray and Jongman, 2011), and improve the performance of speech recognizers (Povey and Saon, 2006; Wegmann et al., 1996). However, the degree to which specific normalization techniques from ASR match human perception is not yet known. Experiments in this paper replicate the general benefit of speaker normalization using our perceptual model, while also providing new data on the degree to which different normalization algorithms from ASR capture information that is similar to what humans use in perceptual tasks.

The paper is organized as follows. We begin by characterizing the speech features tested in this paper. The following section describes the method we use for evaluating these features against human discrimination data. Experiments are presented testing how well each representation predicts human data. We compare our method to previous work, and conclude

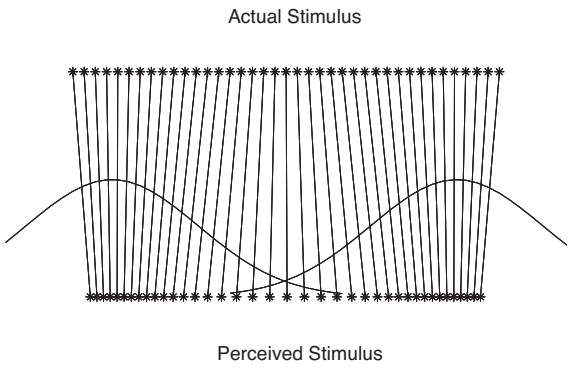

Figure 2: Illustration from Feldman et al. (2009) showing listeners' perceptual bias toward peaks in their prior distribution over sounds in a feature space. This bias leads listeners to perceive some sounds as closer together, and other sounds as farther apart, than they actually are. Reprinted from "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference" by N. H. Feldman, T. L. Griffiths, & J. L. Morgan, 2009, *Psychological Review*, 116, p. 757. Copyright 2009 by the American Psychological Association. Reprinted with permission.

by discussing implications and future directions.

2 Speech features

Speech recognition systems typically consist of a front-end, which transforms short time windows of the speech signal into numeric feature vectors, and a back-end, which uses the output of the front-end system together with phonetic models, language models, dictionaries, and so on to infer what was said. Our work analyzes the front-end feature vectors directly, examining the effects of speaker normalization algorithms on these feature vectors.

Speech contains commingled effects of linguistic, paralinguistic, and purely physical sources of variation. It has often been proposed that human listeners normalize the speech signal to generalize phonetic content across speakers. Listeners could normalize for vocal tract length, for example, either by attending to cues that are relatively constant across speakers (Miller, 1989; Monahan and Idsardi, 2010; Peterson, 1961) or by relying on relative, rather than absolute, encoding of cues that are affected by vocal tract length (Barreda, 2012; Cole et al., 2010). Alternatives to normalization include theories of adaptation in which listeners create representations that are specific to individual speakers or groups of speakers (Dahan et al., 2008; Kleinschmidt and Jaeger, 2015), and theories which argue that no normalization or adaptation mechanisms are necessary at all (Johnson, 1997; Johnson, 2006). The empirical literature on speaker normalization and adaptation includes results that support each of these approaches.

Our focus here is on replicating an advantage of normalized features found in cognitive models of categorization. McMurray and Jongman (2011) and Apfelbaum and McMurray (2015) conducted a direct comparison between normalized and unnormalized encodings of cues to fricative categorization. They found that using normalized features yields a better match with human categorization performance. We aim to replicate the advantage of normalized over unnormalized features using our discrimination model, in order to validate our new evaluation method against an existing cognitive modeling approach.

We test two methods for speaker normalization: vocal tract length normalization (VTLN), which is widely used in ASR systems (Wegmann et al., 1996); and z-scoring, which has been applied in linguistic analyses as well as in ASR systems (Lobanov, 1971; Viikki and Laurila, 1998). Neither VTLN nor z-scoring is a cognitive model of speaker normalization, but both have the potential to capture information similar to what listeners might obtain if they generalize across utterances from speakers with different vocal tract lengths.

We apply these normalization methods to vowels that are represented by mel frequency cepstral coefficients (MFCCs), which have been widely employed as an input representation in ASR systems (Davis and Mermelstein, 1980). MFCCs are a 12-dimensional vector that describe a timepoint of speech by capturing information about the spectral envelope, reflecting vocal tract shape. They are computed by taking the discrete time Fourier transform of the speech signal, mapping it onto a mel frequency scale with log power (to reflect how people process frequency and loudness), and applying the discrete cosine transform to result in a cepstrum (inverse spectrum) whose real values are the MFCC representation. This representation of the signal is indirectly related to formant frequencies, which are peaks in the spectral envelope that are known to correlate with vowel identity and are often used to characterize vowels in studies of hu-

man speech perception. However, MFCCs have the advantage that they are computed deterministically from the speech signal, and thus are not subject to the error inherent in automatic formant tracking.

This section gives an overview of both normalization methods and characterizes their effects; we defer details of their technical implementation to Section 4.

2.1 Vocal tract length normalization

The first normalization method we test, vocal tract length normalization (VTLN), is a technique developed for automatic speech recognition (Cohen et al., 1995; 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 compensates for speaker differences in vocal tract length, which affects the resonant frequencies of the vocal tract, by applying a piecewise linear transformation to expand or compress the frequency axis of the productions of each speaker. The particular adjustment to apply for each speaker can be chosen by attempting to maximise the similarity of their adjusted speech to other speakers. We use /i/ tokens as the basis of this comparison across speakers. VTLN is widely successful in ASR systems, where it substantially decreases the word error rate (Giuliani et al., 2006).

2.2 Z-scoring

The second normalization algorithm, within-speaker z-scoring (Lobanov, 1971), has been suggested for descriptive sociolinguistic research (Adank et al., 2004), as it highlights learned linguistic content while removing speaker-body confounds, such as vocal tract length, from vowel formants. Z-scoring is also used in ASR (Viikki and Laurila, 1998), where it is often referred to as cepstral mean and variance normalization (CMVN). Z-scoring has been shown to reduce word error rate in speech recognizers, to improve generalization to noisy input, and to improve generalization across speakers (Haeb-Umbach, 1999; Molau et al., 2003; Tsakalidis and Byrne, 2005).

2.3 Effects of normalization

The method that we use in Section 4 for testing hypothesized representations of speech against human perception relies on the idea that different feature representations imply different distributions of sounds in listeners' input. To examine these distributions of sounds, we computed MFCCs, MFCCs with VTLN, and z-scored MFCCs from vowel recordings in the Vowels subcorpus of the Nationwide Speech Project (NSP) (Clopper and 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. For our analyses, these vowel tokens are represented by the speech features of their midpoint.[1] Because the temporal midpoint of a word may not be the midpoint of its vowel, we detected the vocalic portion of each NSP token as determined by relatively high total energy (more than one standard deviation above the mean) and low zero-crossing rate (below 30%). We used the temporal midpoint of this span.

We characterize the effects of each normalization method on the distribution of vowels in the NSP corpus by comparing vowel distributions across genders and dialect regions (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. To compare these distributions, we computed symmetrized Kullback-Leibler divergence, a measure of difference between two probability distributions, using a nearest-neighbor algorithm (Wang et al., 2006). Lower K-L divergence indicates greater similarity between two distributions.

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 ei-

[1] We believe that dynamic representations of sounds are critical in human perception. However, this simplification is unlikely to affect model performance for perceiving the sounds in our experiments (Section 4), which have constant formant values across the duration of each vowel token.

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

	MFCC	VTLN	z-scored
Gender KLDiv	7.84	6.14	4.58
Dialect KLDiv	4.41	4.19	2.09

ther of these normalizations than they are in unnormalized MFCCs. Z-scoring using all 10 NSP vowels also sharply reduces K-L divergence between dialect pairs. In contrast, VTLN that matches across speakers on the basis of /i/, which differs little across US English dialects (Clopper and Pisoni, 2006), has minimal effects on dialect K-L divergence: cross-dialect vowel pronunciation differences remain distinct after VTLN. Overall, K-L divergence is lowest when z-scoring by speaker. VTLN removes somewhat less of the gender and dialect variation.

In summary, MFCCs, MFCCs with VTLN, and z-scored MFCCs 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, allowing us to ask which of these feature representations best corresponds with human perception.

3 Model of speech perception

Our goal is to evaluate MFCCs, MFCCs with VTLN, and z-scored MFCCs on their match to human perception. We perform this evaluation by integrating each type of speech feature into a cognitive model of speech perception. The model we adopt has previously been shown to accurately predict listeners' perception of both vowels and consonants (Feldman et al., 2009; Kronrod et al., 2016), but has not yet been implemented directly on speech recordings.

3.1 Model overview

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. This differs from the problem of inferring a phoneme label, which is typically the goal of ASR systems, but is consistent with a large body of empirical evidence showing that human listeners recover acoustic detail from the speech signal in addition to category information (Andruski et al., 1994; Blumstein et al., 2005; Joanisse et al., 2007; Pisoni and Tash, 1974; Toscano et al., 2010).

To correct for noise in the speech signal, listeners bias their perception toward acoustic values that have high probability under their prior distribution. This creates a dependency between the distribution of sounds in the input, which determines listeners' prior distribution, and listeners' perception of sounds.

Formally, speakers and listeners share a prior distribution over possible acoustic values that can be produced in a language, $p(T)$. The acoustic values are assumed to lie in a continuous feature space, such as MFCCs. Prototypical sounds in the language have highest probability under this distribution, but the distribution is non-zero over a wide range of acoustic values, corresponding to all the ways in which speech sounds might be realized. Speakers sample a target production T from this distribution. The target production carries 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,

$$p(S|T) = \mathcal{N}(T, \Sigma_S) \tag{1}$$

Both T and S are in $\mathbb{R}^d$, where d is the dimensionality of the feature space. In our experiments, this feature space is defined by either MFCCs, MFCCs with VTLN, or z-scored MFCCs.

Listeners hear S and reconstruct T by drawing a sample from the posterior distribution,

$$p(T|S) \propto p(S|T)p(T) \tag{2}$$

We refer to a listener's sample from the posterior distribution as a *percept*. The top row of Figure 2 corresponds to S, and the bottom row corresponds to listeners' reconstruction of T.

3.2 Implementation on speech corpora

Previous work using this model has inferred listeners' prior distribution by measuring their perceptual cat-

egorization, and estimating the parameters of Gaussian phonetic categories that listeners appear to be using to make these categorization judgments. In the experiments below, we instead use vowel productions from the NSP corpus to estimate listeners' prior distributions. We avoid making parametric assumptions about the form of this prior distribution (such as expecting sounds from the corpus to fall into Gaussian categories) by using importance sampling, as proposed by Shi et al. (2010).

We assume that vowels in the NSP corpus constitute a set of samples $\{T^{(i)}\}$ drawn from listeners' prior distribution $p(T)$. We apply likelihood weighting, weighting each sound $T^{(i)}$ by the probability that it generated the stimulus being perceived, $p(S|T^{(i)})$. We then sample a sound from the corpus according to its weight. A sound from the corpus that is sampled in this way behaves as though it were drawn from the posterior distribution, $p(T|S)$.

This way of approximating the model through importance sampling allows us to sample a percept from the model's posterior distribution without knowing the analytical form of the prior distribution. We require only samples drawn from the prior, and can reweight these in order to obtain a sample from the posterior. In addition, this implementation of the model does not require sounds from the corpus to have phoneme labels, as the weights $p(S|T^{(i)})$ are defined by the model's Gaussian likelihood function, corresponding to speech signal noise.

3.3 Modeling discrimination data

The model can be used to predict listeners' discrimination behavior. In AX discrimination tasks, listeners hear two stimuli and decide whether they are acoustically identical. We assume 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, follows a binomial distribution whose parameter is the probability that the percepts for the two sounds are within a distance ϵ of each other,

$$p(same) = p(|T_1 - T_2| < \epsilon | S_1, S_2) \qquad (3)$$

We approximate the probability from (3) as

$$p(same) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{|T_1^{(i)} - T_2^{(i)}| < \epsilon} \qquad (4)$$

where $T_1^{(i)}$ and $T_2^{(i)}$ are samples drawn through importance sampling from posterior distributions $p(T_1|S_1)$ and $p(T_2|S_2)$. We draw 100 pairs of target productions from the posterior for each experimental trial and approximate listeners' probability of responding *same* by estimating the proportion of these pairs that yielded a *same* response, using add-one smoothing.[2] We use this estimate to compute model likelihoods – the probability of the human responses, given the model predictions – based on listeners' actual *same-different* responses.

The noise covariance Σ_S and the response threshold ϵ are free parameters of the model, which we fit to maximize model likelihoods.[3] 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.

4 Experiments

Experiments 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 correspond to a similarity space more similar to the dimensions that human listeners use in speech perception.

[2] The distance computation in (4) uses the Mahalanobis distance between $T_1^{(i)}$ and $T_2^{(i)}$, where the covariance matrix is the noise variance Σ_S. If listeners' sensitivity is related to their perceptual noise, then Σ_S is the appropriate scaling factor for this distance. In support of this, Feldman et al. (2009) found in their one-dimensional model that increasing Σ_S by adding white noise increases listeners' decision threshold to the same degree.

[3] Although our use of Mahalanobis distance allows a tradeoff between Σ_S and ϵ in the distance computation, Σ_S also affects the weighting of exemplars through the perceptual model's likelihood function. Thus, these parameters are identifiable.

4.1 Human perceptual data

We use vowel discrimination data from an AX discrimination task conducted by Feldman et al. (2009) in quiet listening conditions (Figure 3). Twenty participants heard stimuli consisting of 13 isolated vowels ranging from /i/ (as in 'beet') to /e/ (as in 'bait'), which were synthesized to simulate a male speaker. To control the placement of stimuli along the vowel continuum, first and second formant frequencies were equally spaced between /i/ and /e/ values along the mel frequency scale. Participants heard all ordered pairs of stimuli and judged whether each pair was acoustically identical, responding *different* if they could hear any difference between the stimuli. MFCCs, MFCCs with VTLN, and z-scored MFCCs computed from these 13 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.

4.2 Speech representations

The samples $\{T^{(i)}\}$ that serve as the model's prior distribution are vowels from the Vowels subcorpus of the NSP. Vowel midpoints were represented in the model either as MFCCs, MFCCs with VTLN, or z-scored MFCCs, computed using a 25 ms window and a 10 ms step size.[4] Features were scaled to have zero mean and a variance of one across all vowels.[5]

VTLN was implemented using a procedure from Wegmann et al. (1996) adapted for an unsupervised setting, which compensates for vocal tract length variation by finding a frequency adjustment for each speaker that maximizes the similarity of their tokens to all the other speakers'. This VTLN implementation first applies and then selects among several frequency rescaling factors, ranging from -20% to +20% of the original vocal tract length in steps of 5%. Together with the original speech (scale factor of 0%) there are therefore 9 possible variants. The rescaling factor for each speaker is chosen using expectation maximization (Dempster et al., 1977). The E-step trains a Gaussian mixture model with 9 components and spherical covariance from all /i/ vowel frames in the NSP corpus. The M-step finds a rescaling factor for each speaker that maximizes the likelihood under the current Gaussian mixture model. After EM converges, the maximum likelihood rescaling factor for the stimuli is selected under the final NSP-based mixture model. Normalizing the stimuli is straightforward, due to the reliance of our VTLN procedure on only /i/ tokens. Vowel midpoints of corpus tokens and stimuli were excerpted after normalization.

Z-scoring was applied to MFCCs at vowel midpoints. The NSP is an ideally balanced case for the z-scoring normalization, because each speaker says the same vowels the same number of times. MFCCs for the stimulus 'speaker' were only available for the /i/-/e/ vowel continuum, so it was not possible to apply z-scoring directly to the stimuli in the same way as it was applied to each NSP speaker. However, the stimuli were synthesized according to average formant values for male speakers (Peterson and Barney, 1952), and we verified that normalizing MFCCs for these stimuli according to the average z-scoring factors of the 30 male NSP speakers projected the stimuli into an appropriate region of the acoustic space.

Finally, although speech recognition systems typically use 12 MFCC dimensions, we additionally include experiments for all feature types that omit subsets of the higher dimensions, as formant information that is useful for perceiving vowels is primarily reflected in the lower dimensions.

4.3 Fitting parameters

The NSP corpus was divided into equally sized sets to fit parameters and test model likelihoods in 2-fold cross-validation. Two methods were used for dividing the corpus. In one case, the two halves contained different tokens from the same speakers. In the other case, the two halves contained tokens from different speakers (balanced for gender and dialect region). Each division of the corpus was created once, and used across experiments with all speech feature types.

The fitting procedure consisted of selecting the

[4]The zeroth cepstral dimension was excluded from the representation in experiments; it mainly describes the total amount of energy at each time in the speech signal, which is useful in detecting whether or not the speech is vocalic but contributes little to distinguishing different vowel sounds. Additionally, although ASR systems frequently append deltas and double deltas to the cepstral coefficients, corresponding to the first and second derivatives of the MFCCs, we did not use these as they add free parameters to the model.

[5]This was different from the z-scoring method for speaker normalization, which scaled the vowels from each individual speaker to have zero mean and a variance of one.

response threshold ϵ and the noise variance Σ_S, constrained to be diagonal, to best fit the perceptual data. This was done using Metropolis-Hastings sampling (Hastings, 1970) with parallel tempering, which flexibly interleaves broad exploration of the search space with fine-tuning of the best parameter values. The sampler used Gaussian proposal distributions for each parameter, with proposals Σ_S^* and ϵ^* drawn based on current estimates Σ_S and ϵ as $\Sigma_S^* \sim \mathcal{N}(\Sigma_S, \sigma_1^2\mathbf{I})$ and $\epsilon^* \sim \mathcal{N}(\epsilon, \sigma_2^2)$. Given this symmetric proposal distribution, the acceptance ratio is

$$A = \frac{p(\mathbf{k}|\Sigma_S^*, \epsilon^*)}{p(\mathbf{k}|\Sigma_S, \epsilon)} \quad (5)$$

where $\mathbf{k}$ is a vector of counts corresponding to *same-different* responses from human participants in the experiment. The likelihood $p(\mathbf{k}|\Sigma_S, \epsilon)$ is a product of the binomial likelihood functions for each speech sound contrast, whose parameters are approximated according to (4). We run several chains, indexed $1..C$, at temperatures $\tau_1..\tau_C$, by raising each likelihood $p(\mathbf{k}|\Sigma_S, \epsilon)$ to a power of $\frac{1}{\tau}$. At each iteration there is a proposal to swap the parameters between neighboring chains. The acceptance ratio for swapping parameters between chains c_1 and c_2 is

$$A = \frac{p(\mathbf{k}|\Sigma_S^{(c_2)}, \epsilon^{(c_2)})^{\frac{1}{\tau_{c_1}}} p(\mathbf{k}|\Sigma_S^{(c_1)}, \epsilon^{(c_1)})^{\frac{1}{\tau_{c_2}}}}{p(\mathbf{k}|\Sigma_S^{(c_1)}, \epsilon^{(c_1)})^{\frac{1}{\tau_{c_1}}} p(\mathbf{k}|\Sigma_S^{(c_2)}, \epsilon^{(c_2)})^{\frac{1}{\tau_{c_2}}}} \quad (6)$$

where τ_{c_1} and τ_{c_2} are the temperatures associated with chains c_1 and c_2, respectively. Proposals are accepted with probability $min(1, A)$. Model parameters were fit using this procedure, with 50,000 iterations through eleven chains at different temperatures ranging from 0.01 to 1,000. Our analyses use the last sample in the lowest temperature chain, with a temperature of 0.01.

In each run of the model, one half of the corpus was used to fit model parameters, while the other half was used to compute model likelihoods at test. The roles of the two sets were then reversed, resulting in 2-fold cross-validation. Each half of the corpus served as a test set for 10 runs of the model, so points and error bars in Figures 4 and 5 represent means and standard error calculated across all 20 replications; the relatively small error bars indicate that results were consistent across replications.[6]

4.4 Results

Stimuli in the behavioral discrimination task were synthesized to mimic a single speaker, and thus VTLN and z-scoring are not expected to directly impact relative distances between stimuli S in the feature space. Instead, differences in model behavior across the three feature types should arise primarily because VTLN and z-scoring impact the distribution of vowels from the NSP corpus. This affects the model's prior distribution and thus its perceptual biases, leading to different reconstructions of T.

The predicted discrimination patterns that result from different speech representations are illustrated qualitatively in Figure 3, together with human data and a model benchmark which is described at the end of this section. Patches of darker shading in the confusion matrices of Figure 3 indicate higher proportions of *same* responses, corresponding to reduced discrimination between stimuli in these regions. The models vary in how well they reflect humans' perceptual biases. One-dimensional models are extremely poor regardless of feature type. Unnormalized MFCC models, as shown most clearly in 4 and 6 dimensions, often differ from human listeners by excessively reducing discrimination around the initial (/i/) end of the stimulus continuum. Z-scoring models, particularly in the low to mid range of dimensionality, reflect human judgments better than MFCC models but still tend to deviate from human performance in visible areas of reduced discrimination centered around stimuli 6 and 9. Models using VTLN representations of moderately low dimensionality show the most faithful fit with human patterns of perceptual bias.

These differences in model performance across feature types can be quantified by computing the log likelihood of a model, given the human data. This measure treats the human *same* and *different* responses as ground truth, and assesses model performance on the binary decision task against that ground truth; higher log likelihoods indicate a closer match of a model to human perceptual data. Log likelihoods shown in Figure 4 quantify the performance of models across feature types; this complements the

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

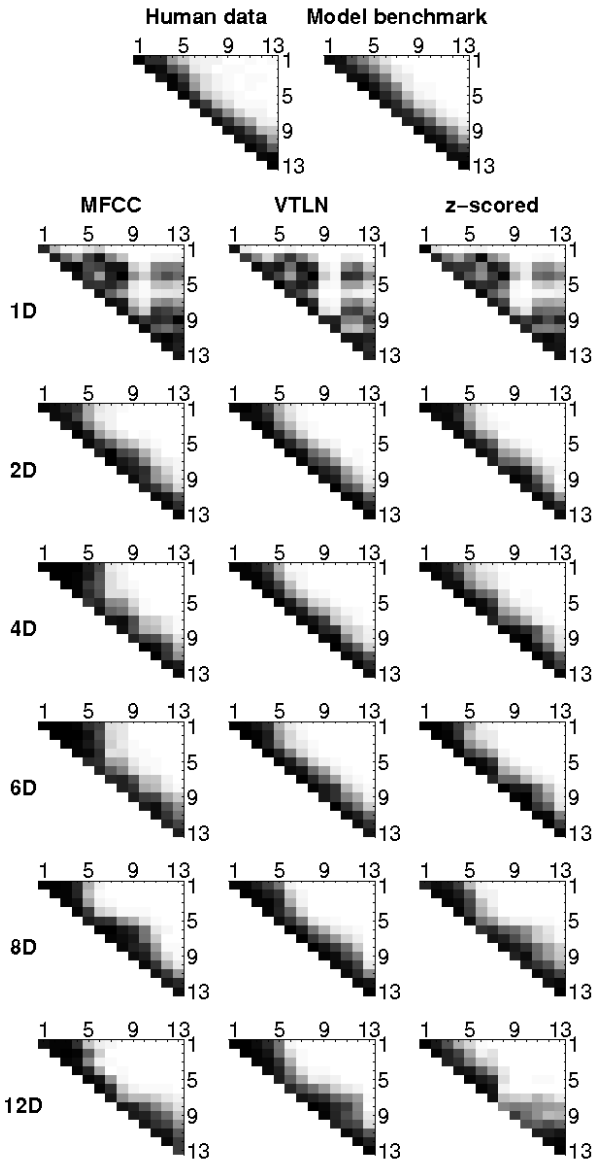

Figure 3: Confusion matrices for how often each pair of stimuli is judged to be the same (black) vs. different (white) by humans, by a benchmark exemplar model, and by representative models testing the different speech features on familiar speakers. Axis labels of 1-13 denote stimulus numbers from an AX trial.

detailed views of model behavior in Figure 3. Recall that our primary goal is to replicate previous findings with categorization models showing that normalized representations are more consistent with listeners' perception than unnormalized representations. We do replicate this result with our discrimination model: in almost all cases, unnormalized MFCCs have the lowest likelihoods among the three representations tested (Figure 4).

We also find that MFCCs normalized by VTLN outperform z-scored MFCCs. This occurs in spite of the fact that z-scoring within speakers puts male and female speech into the most directly comparable space and in general neutralizes the most differences between speakers as measured by K-L divergence (Table 1). This indicates that more effective speaker normalization algorithms (in the sense of minimizing differences between speakers in a feature space) do not always translate into better matches with human perception. Furthermore, it provides evidence that the better performance of the VTLN models is not merely an artifact of high acoustic similarity among vowels in the corpus.

As suggested by Figure 3, models using speech representations of one dimension are always extremely poor, with log likelihoods of -1300 to -1600; the first cepstral coefficient on its own does not provide enough information for the perceptual task. In some experiments, such as for z-scored features, test likelihood decreases with higher numbers of dimensions, likely due to overfitting of parameters to particular sounds from the corpus.[7] The lower MFCC dimensions, particularly the second dimension, appear to contain information relevant to listeners' discrimination of an /i/-/e/ continuum.

As a representative example of the best fitting parameters, one trained model using four dimensions of VTLN features found parameters[8] of $\Sigma_S = 0.1617$, 0.0092, 2.2400, 1.8194 and $\epsilon = 3.4388$.[9] Dimensions where noise is low indicate those that the model found informative for this perceptual task: low noise causes even numerically small differences between tokens to be attended to as differences in perceived vowel quality, whereas high noise variance generates a strong perceptual bias toward the center of vowel

[7] The corresponding training likelihoods for z-scored dimensions remained stable or even increased at higher numbers of dimensions; there was not failure to converge in the training procedure. Similarly, the low likelihood observed at six dimensions for MFCCs was due to several runs that achieved high likelihood on the training set and low likelihood on the test set.

[8] The covariance matrix Σ_S was constrained to be diagonal; here we list only the diagonal elements.

[9] Across the 10 training replications, the standard deviation in these trained parameters was respectively 0.0823, 0.0006, 0.8684, 1.1518 for Σ_S and 0.1645 for ϵ.

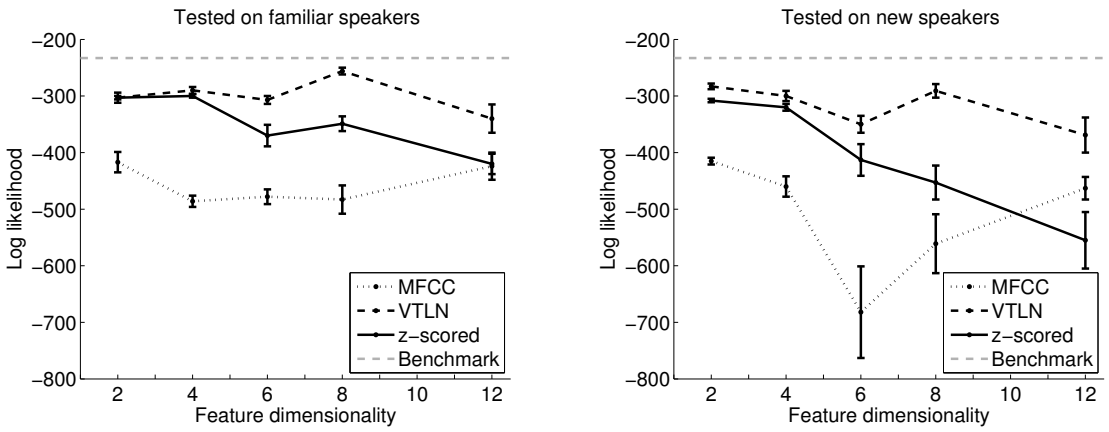

Figure 4: Model fit to human data, when generalizing to familiar speakers (left) and new speakers (right).

space that collapses the space between vowel percepts along that dimension. Dimensions with high noise parameter values therefore indicate that the model benefited from perceiving all samples from the prior distribution as being similar along that dimension, with variability along those dimensions treated as noise relative to listeners' perceptual judgments about this vowel contrast. Across all feature types, successful models had low noise in dimension 2 and low variance in this noise parameter across training replications. This consistency, together with the large difference in the model's ability to match human performance between feature dimensionalities of 1 and 2, indicates that the second cepstral coefficient is important for capturing listeners' discrimination of sounds along the /i/-/e/ vowel continuum.

Our next analyses examine the speech tokens sampled as percepts by the model when perceiving the experimental stimuli. The model did not use the NSP's speaker and phoneme labels for selecting these percepts, but we take advantage of these labels for our analyses. Across all models, the 100 percepts drawn from the posterior distribution for each stimulus contained on average 30 different tokens, indicating that a number of speech productions (from different speakers) are treated as linguistically similar to each other. As a measure of model quality and interpretability, we examine two aspects of the tokens sampled by the model during each perceptual judgment (Figure 5). 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 tokens 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 tokens, 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 tokens; this confirms that the z-scored representation is highly effective at neutralizing differences between speakers of different genders (Table 1), although as noted above, it is not the representation that gives the best match to human discrimination performance (Figure 4).

While models with 2 through 6 dimensions consistently treated the experimental stimuli as being most similar to high front vowels in the corpus, models with higher orders of cepstral coefficients did not (Figure 5), reinforcing the importance of the lower MFCC dimensions in capturing listeners' perception of these stimuli. We suspect that failure to perceive stimuli as high front vowels in models with higher numbers of dimensions emerged due to the artificial synthesis of the experimental stimuli, which resulted in these stimuli being most similar to low back vowels from the corpus in two of the higher MFCC dimensions. The model can perceive stimuli as low

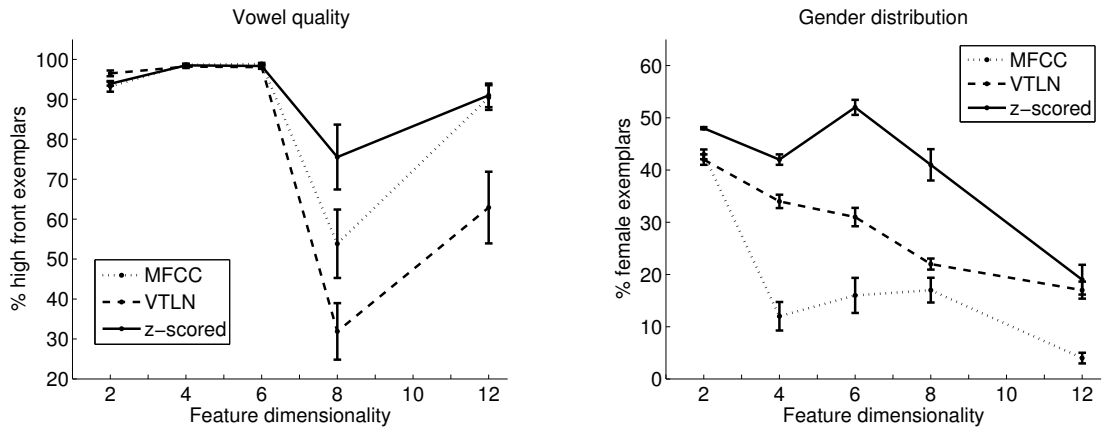

Figure 5: Secondary evaluations, showing how often the stimuli were correctly perceived as high front vowels (left) and how often the model based perceptions on female utterances (right). Data are averaged across the familiar and new speaker test cases, which were very similar on these measures.

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. Nevertheless, even if we omit results from models that were affected by this mismatch between corpus tokens and experimental stimuli, the relative ordering of the three feature types was qualitatively consistent across a range of dimensionalities.

Finally, we can compare the log likelihoods from Figure 4 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 production data in speech recordings. This method of estimating listeners' prior distributions sidesteps the problem of mapping between perception and production in that it estimates prior distributions directly from perceptual data. Using their estimate, the model yields an average log likelihood of -233, somewhat higher than those obtained here. Our models should approach this value as the distribution of sounds in the corpus approaches the prior distribution that listeners use in perceptual tasks.

5 Relation to previous work

To our knowledge, this is the first time a cognitive model has been used to evaluate ASR representations against human behavioral data. Front-end feature representations in ASR are typically evaluated on their ability to improve performance in an ASR system (Junqua et al., 1993; Welling et al., 1997). Measuring performance in an ASR system provides a direct evaluation relative to a particular task, but provides only indirect evidence about the correspondence of feature representations with human perception. Our work measures the extent to which these feature representations match human judgments, parallel to analyses that have been conducted in other domains of language to compare representations used in language technology to human data (Chang et al., 2009; Gershman and Tenenbaum, 2015). As a feature evaluation metric, our approach also has the advantage that it examines front-end feature representations directly, and is not subject to arbitrary decisions about back-end systems in an ASR pipeline.

There have also been proposals that use direct measures of feature variability (Haeb-Umbach, 1999) or minimal pair ABX discrimination tasks (Carlin et al., 2011; Schatz et al., 2013; Schatz et al., 2014) to evaluate feature representations, and thus do not depend on particular back-end ASR systems. These approaches differ from ours in that they evaluate whether discrimination performance is correct, rather than whether discrimination performance is similar to that of human listeners. In addition, these approaches assume a direct mapping between distance in the feature space and perceived similarity between

two stimuli, whereas ours takes into account human listeners' perceptual bias toward sounds that occur frequently within the feature space.

Research in cognitive science has previously examined the dimensions that guide listeners' perception, but has relied on data from categorization tasks, in which listeners decide which category a sound belongs to (Apfelbaum and McMurray, 2015; Idemaru et al., 2012; McMurray and Jongman, 2011; Nittrouer and Miller, 1997; Repp, 1982). Our approach instead focuses on acoustic discrimination tasks. 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 labeled training data, building a discrimination model does not, as explained in Section 3.

6 Discussion

In this paper a cognitive model of speech perception was implemented directly on speech recordings and used to evaluate the low-level feature representations corresponding to two speaker normalization methods: VTLN and z-scoring. Both normalization methods improved the model's fit to human perceptual data. Between the two normalization methods, VTLN outperformed z-scoring, despite being less effective at collapsing gender and dialect differences.

The ability of VTLN and z-scoring to improve the match of a feature space with human perception may be a relevant consideration in supporting the use of these normalization methods in ASR systems. Furthermore, although we do not mean to imply that VTLN or z-scoring are cognitive models of speaker normalization, the better performance of VTLN over z-scoring does suggest that VTLN facilitates generalization across speakers in a way that is more similar to human perceivers. For example, human listeners may collapse across genders while retaining differences across dialects. Knowing how to model the way in which listeners generalize across speakers could have practical implications for assisting listeners with hearing impairments (Liu et al., 2008), and our work complements previous experimental and computational evidence related to this question (Barreda, 2012; Halberstam and Raphael, 2004; McMurray and Jongman, 2011; Nearey, 1989).

Speaker normalization and speaker adaptation are often treated as competing theories of human speech perception (Dahan et al., 2008), whereas speaker normalization algorithms are applied in conjunction with speaker adaptation algorithms in the ASR literature (Gales and Young, 2007). Our experiments focused only on the effects of normalization, without considering whether there might be a role of adaptation in explaining listeners' discrimination behavior. It is not obvious how we would apply speaker adaptation in our discrimination model, which does not incorporate explicit category representations.

Despite the advantage observed for normalized over unnormalized features, none of the representations tested here match human perception exactly. A prior distribution estimated from perceptual categorization still outperforms prior distributions measured from a speech corpus. Our primary goal has been to introduce a new method for assessing speech feature representations against human data, and to validate that method against previous findings that showed a benefit for normalization.

Our method is compatible with a wide range of feature spaces. 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 experimental stimuli and sounds from a corpus, and for which a distance metric between sounds in the corpus can be compared to a threshold ϵ. The likelihood function does not need to be Gaussian, and the feature space does not need to have fixed or even finite dimensionality. Even when assuming a Gaussian likelihood function, not all feature spaces have uncorrelated dimensions, and the model may need to be extended to use a full covariance matrix for Σ_S. This extension is straightforward to derive mathematically, but would require additional human perceptual data to constrain the model's free parameters.

The speech corpus used in these experiments consisted of /hVd/ utterances that were balanced across vowel, gender, and dialect. Conducting initial experiments with a corpus of vowels in neutral contexts allowed us to investigate algorithms for speaker normalization while sidestepping issues of how listeners generalize across phonological contexts. However,

differences between this controlled corpus and listeners' experience may account in part for the lower performance of the corpus-based models compared to the model whose prior distribution was estimated from perceptual categorization data. Future work that estimates listeners' prior distributions from speech recordings can do so using more realistic corpora of conversational speech. The model's prior distribution can in principle be estimated from any speech corpus, and does not require phoneme labels.

Using larger corpora of conversational speech would be particularly beneficial because it could provide sufficient data to derive and evaluate features from representation learning algorithms, which typically involve training models with many parameters. Recent work in ASR has proposed ways of optimizing feature representations based on linguistic data (Grézl et al., 2014; Heigold et al., 2013; Kamper et al., 2015; Sercu et al., 2016; Thomas et al., 2012; Wang et al., 2015). These proposals differ from each other in their objective functions, and our method can help determine which of these objective functions yield representations that capture human listeners' generalizations. Testing the outcome of these learning algorithms against human data in high resource languages could help select representation learning algorithms for low resource settings. Such investigations would not necessarily be limited to speaker normalization algorithms, but could apply to any representation learning algorithm designed to optimize perception for a particular language.

Perhaps the most exciting future application of this work is in investigating the way in which human listeners' feature spaces become tuned to their native language. Research in cognitive science has begun asking how language-specific feature spaces might be learned (Holt and Lotto, 2006; Idemaru and Holt, 2011; Idemaru and Holt, 2014; Liu and Holt, 2015; Nittrouer, 1992). The method proposed here provides a way of testing the predictions of these learning theories against human data, and of testing whether representation learning algorithms from the ASR literature have analogues in human perceptual development. Unlabeled speech corpora are available in multiple languages, and cross-linguistic perceptual differences are typically measured through discrimination tasks; our discrimination model can take advantage of these data. Thus, using the method proposed here, predictions of human representation learning theories can be evaluated with respect to adults' cross-linguistic differences in sensitivity to perceptual dimensions. If the cognitive model were extended to infant discrimination tasks, a similar strategy could be used to evaluate theories of human representation learning with respect to children's discrimination abilities throughout the learning process. In this way, implementing a cognitive model directly on corpora of natural speech can lead to a richer understanding of the way in which listeners' perception is shaped by their environment.

Acknowledgments

We thank Josh Falk for help in piloting the model, Phani Nidadavolu for help computing speech features, and Eric Fosler-Lussier, Hynek Hermansky, Bill Idsardi, Feipeng Li, Vijay Peddinti, Amy Weinberg, the UMD probabilistic modeling reading group, and three anonymous reviewers for helpful comments and discussion. Previous versions of this work were presented at the 8th Northeast Computational Phonology Workshop and the 38th Annual Conference of the Cognitive Science Society. The work was supported by NSF grants BCS-1320410 and DGE-1321851.

References

Patti Adank, Roel Smits, and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5):3099–3107.

Jean E. Andruski, Sheila E. Blumstein, and Martha Burton. 1994. The effect of subphonemic differences on lexical access. *Cognition*, 52:163–187.

Keith S. Apfelbaum and Bob McMurray. 2015. Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin and Review*, 22(4):916–943.

Santiago Barreda. 2012. Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *Journal of the Acoustical Society of America*, 132(5):3453–3464.

Sheila E. Blumstein, Emily B. Myers, and Jesse Rissman. 2005. The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, 17(9):1353–1366.

Michael A. Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. 2011. Rapid evaluation of speech

representations for spoken term discovery. *Proceedings of Interspeech*.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22*.

Meghan Clayards, Michael K. Tanenhaus, Richard N. Aslin, and Robert A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

Cynthia G. Clopper and David B. Pisoni. 2006. The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, 48(6):633–644.

Jordan Cohen, Terri Kamm, and Andreas G. Andreou. 1995. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *Journal of the Acoustical Society of America*, 97(5):3246–3247.

Jennifer Cole, Gary Linebaugh, Cheyenne M. Munson, and Bob McMurray. 2010. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38:167–184.

Delphine Dahan, Sarah J. Drucker, and Rebecca A. Scarborough. 2008. Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, 108:710–718.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of the IEEE*, pages 357–366.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

Naomi H. Feldman, Thomas L. Griffiths, and James L. Morgan. 2009. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782.

Mark Gales and Steve Young. 2007. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

Samuel J. Gershman and Joshua B. Tenenbaum. 2015. Phrase similarity in humans and machines. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Diego Giuliani, Matteo Gerosa, and Fabio Brugnara. 2006. Improved automatic speech recognition through speaker normalization. *Computer Speech & Language*, 20(1):107–123.

František Grézl, Ekaterina Egorova, and Martin Karafiát. 2014. Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure. *IEEE Spoken Language Technology Workshop*, pages 48–53.

Reinhold Haeb-Umbach. 1999. Investigations on inter-speaker variability in the feature space. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 397–400.

Benjamin Halberstam and Lawrence J. Raphael. 2004. Vowel normalization: the role of fundamental frequency and upper formants. *Journal of Phonetics*, 32:423–434.

W. Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Georg Heigold, Vincent Vanhoucke, Andrew Senior, Phuongrang Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jesse Dean. 2013. Multilingual acoustic models using distributed deep neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8619–8623.

Hynek Hermansky and Nelson Morgan. 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.

Lori L. Holt and Andrew J. Lotto. 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, 119(5):3059–3071.

Kaori Idemaru and Lori L. Holt. 2011. Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, 37(6):1939–1956.

Kaori Idemaru and Lori L. Holt. 2014. Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 40(3):1009–1021.

Kaori Idemaru, Lori L. Holt, and Howard Seltman. 2012. Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *Journal of the Acoustical Society of America*, 132(6):3950–3964.

Marc F. Joanisse, Erin K. Robertson, and Randy Lynn Newman. 2007. Mismatch negativity reflects sensory and phonetic speech processing. *NeuroReport*, 18(9):901–905.

Keith Johnson, 1997. *Speech perception without speaker normalization: An exemplar model*, pages 145–165. Academic Press, New York.

Keith Johnson. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499.

Jean-Claude Junqua, Hisashi Wakita, and Hynek Hermansky. 1993. Evaluation and optimization of perceptually-based ASR front-end. *IEEE Transactions on Speech and Audio Processing*, 1(1):39–48.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Dave F. Kleinschmidt and T. Florian Jaeger. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148–203.

Yakov Kronrod, Emily Coppess, and Naomi H. Feldman. 2016. A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin and Review*, 23(6):1681–1712.

Ran Liu and Lori L. Holt. 2015. Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6):1783–1798.

Chuping Liu, John Galvin III, Qian-Jie Fu, and Shrikanth S. Narayanan. 2008. Effect of spectral normalization on different talker speech recognition by cochlear implant users. *Journal of the Acoustical Society of America*, 123(5):2836–2847.

B. M. Lobanov. 1971. Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49(2):606–608.

Jessica Maye, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111.

Bob McMurray and Allard Jongman. 2011. What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246.

James D. Miller. 1989. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85(5):2114–2134.

Sirko Molau, Florian Hilger, and Hermann Ney. 2003. Feature space normalization in adverse acoustic conditions. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 656–659.

Philip J. Monahan and William J. Idsardi. 2010. Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and Cognitive Processes*, 25(6):808–839.

Terrance Michael Nearey. 1978. *Phonetic feature systems for vowels*, volume 77. Indiana University Linguistics Club.

Terrance M. Nearey. 1989. Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85(5):2088–2113.

Susan Nittrouer and Marnie E. Miller. 1997. Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America*, 101(4):2253–2266.

Susan Nittrouer. 1992. Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, 20:351–382.

Gordon E. Peterson and Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Gordon E. Peterson. 1961. Parameters of vowel quality. *Journal of Speech and Hearing Research*, 4(1):10–29.

David B. Pisoni and Jeffrey Tash. 1974. Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, 15(2):285–290.

Daniel Povey and George Saon. 2006. Feature and model space speaker adaptation with full covariance Gaussians. *Proceedings of Interspeech*.

Bruno H. Repp. 1982. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1):81–110.

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proceedings of Interspeech*, pages 1781–1785.

Thomas Schatz, Vijayaditya Peddinti, Xuan-Nga Cao, Francis Bach, Hynek Hermansky, and Emmanuel Dupoux. 2014. Evaluating speech features with the minimal-pair ABX task (II): Resistance to noise. *Proceedings of Interspeech*.

Tom Sercu, Christia Puhrsch, Brian Kingsbury, and Yann LeCun. 2016. Very deep multilingual convolutional neural networks for LVCSR. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Lei Shi, Thomas L. Griffiths, Naomi H. Feldman, and Adam N. Sanborn. 2010. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, 17(4):443–464.

Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. 2012. Multilingual MLP features for low-resource LVCSR systems. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4269–4272.

Joseph C. Toscano, Bob McMurray, Joel Dennhardt, and Steven J. Luck. 2010. Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, 21(10):1532–1540.

Stavros Tsakalidis and William Byrne. 2005. Acoustic training from heterogeneous data sources: Experiments in Mandarin conversational telephone speech transcription. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 461–464.

Olli Viikki and Kari Laurila. 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. 2006. A nearest-neighbor approach to estimating divergence between continuous random vectors. *IEEE International Symposium on Information Theory*, pages 242–246.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. 2015. Unsupervised learning of acoustic features via deep canonical correlation analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Steven Wegmann, Don McAllaster, Jeremy Orloff, and Barbara Peskin. 1996. Speaker normalization on conversational telephone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 339–341.

L. Welling, N. Haberland, and H. Ney. 1997. Acoustic front-end optimization for large vocabulary speech recognition. *Proceedings of Eurospeech*, pages 2099–2102.