Introduction

There are two broad traditions in the study of speech perception, ~~and these two traditions have~~ very different approaches to *variability* in speech. On the one hand, for the cognitive/psycholinguistic tradition, variabiltiy is a *problem* that listeners must cope with. In this view, variability is such a severe problem that it is traditionally referred to only indirectly, as the "lack of invariance" (Liberman et al. 1967). The sociolinguistic tradition, on the other hand, views variation as a rich source of social information. The particular variety of language that you speak says a lot about who you are as a person.

~~However,~~ these two approaches have recently begun to converge in how they approach variability. In particular psycholinguistic theories of speech perception have ~~started to take variation seriously~~~~, instead of wishing it would go away.~~ In part this realization comes from computational-level analyses of speech perception (Clayards et al. 2008; Feldman, Griffiths, and Morgan 2009; Feldman et al. 2013; Norris and McQueen 2008; Kleinschmidt and Jaeger 2015). These approaches start from the hypothesis that the speech perception system is organized in order to be good at speech perception in the world that it has to operate in. In the spirit of ideal observer approaches to other domains (like visual perception, Marr 1982; or memory, Anderson 1990; Anderson 1991) these approaches focus on spelling out how the nature of the task, the available information, and the structure of the world constrain, in principle, how well a listener can do.

Applied to speech perception, this approach ~~reveals two important things~~. First, it suggests that speech perception can be thought of as a process of *inference under uncertainty* (Clayards et al. 2008; Norris and McQueen 2008): because of noise in production processes, each linguistic unit is realized—even by a single talker—as a *distribution* of acoustic cues (cf. Lisker and Abramson 1964; Peterson and Barney 1952; Hillenbrand et al. 1995; Allen, Miller, and DeSteno 2003; Newman, Clouse, and Burnham 2001). As such, the best a listener can do is to *infer* how likely each possible linguistic unit is as an explanation of the cues they observe, based on their knowledge of these cue distributions.

Second, this approach provides a new perspective on talker variability, which is formalized in the "ideal adapter" framework (Kleinschmidt and Jaeger 2015). One consequence of talker variability is that the distribution of cues for each linguistic unit *changes* from one situation to the next, depending on who's talking (Cynthia G. Clopper, Pisoni, and Jong 2005; Newman, Clouse, and Burnham 2001; Allen, Miller, and DeSteno 2003). Given that effective speech perception—in this perspective—depends on good knowledge of the underlying cue distributions, this means that listeners must *also* constantly be inferring the current talker's linguistic generative model (the probabilistic distributions of cues they produce for each underlying linguistic structure).

This second insight leads to the following prediction. An ~~"~~ideal adapter~~"~~ will take advantage of any additional structure in the world that is informative about how cue distributions vary from one situation to the next. This structure may be as simple as the fact that individual talkers tend to be consistent in the cue distributions they produce (Heald and Nusbaum 2015), meaning that prior experience with a familiar talker is informative about the cue distributions they will produce in the future. But structure can occur at other levels, too, and here is one place that psycholinguistic theories begin to parallel sociolinguistics. To the extent that variables like gender, class, regional origin, etc. are sociolinguistically relevant, they are reliably informative about how linguistic variables are realized, and hence helpful for speech perception.

This means that a listener can potentially learn a lot about a talker's cue distributions just by knowing *who* a talker is. Conversely, listeners can learn a lot about who a talker is based on the distributions of cues that they produce. ~~These two long-standing insights from sociolinguistics~~ follow straightforwardly from the ideal adapter framework. However, the extent to which ~~either of these is plausible~~ *in practice* depends on exactly how much a particular socio-indexical variable—a particular sense of *who* a talker is—actually influences the cue distributions that talkers produce.

Background

What do we already know about the relationship between speech perception and socio-linguistic variables?[1]

First, we know that the *amount* and *structure* of talker variability differs between cues and phonetic categories. Vowels and fricatives have a lot of talker variability, much of it conditioned on sex (Peterson and Barney 1952; Hillenbrand et al. 1995; Jongman, Wayland, and Wong 2000; McMurray and Jongman 2011; Newman, Clouse, and Burnham 2001). For vowels this can be largely attributed to differences in vocal tract length, but not entirely (Bladon, Henton, and Pickering 1984; Johnson 2005; Johnson 2006). There appears to be less talker variabiltiy for stop voicing (e.g., /b/ vs. /p/), and little systematic effect of sex (Allen, Miller, and DeSteno 2003; Lisker and Abramson 1964; Chodroff et al. 2015). Despite differences in the overall degree of talker variabiltiy, there is stylistic variation in both. For instance, regional dialects of American English differ in their pronunciation of vowel categories (Cynthia G. Clopper, Pisoni, and Jong 2005; Labov, Ash, and Boberg 2005). Use of voice onset time (VOT) as a cue to stop voicing varies based on language background (e.g., monolingual English speakers vs. French bilinguals; Caramazza et al. 1973; Flege 1987; Pineda and Sumner 2010; Sumner 2011), as well as regionally in the UK (Docherty et al. 2011).

We also know that listeners use socio-indexical group information to guide speech recognition. Niedzielski (1999) found that if listeners believe that a talker is Canadian, they hear more Canadian raising than if they believe the talker is American. Hay and Drager (2010) found a similar sensitivity to dialect group using an even subtler manipulation, manipulating listeners perceptions based on a stuffed animal that cued New Zealand or Australia. Perception of vowels and fricatives is affected by the perceived gender of a talker, which can be cued by voice quality, visual presentation of a male or female face, or even explicit instruction (Strand 1999; Johnson, Strand, and D'Imperio 1999; Strand and Johnson 1996) Indirect evidence that listeners are sensitive to socio-indexical grouping variables comes from recalibration (also known as perceptual learning) studies that show different patterns of generalization from male to female talkers (and vice-versa) for different cues/contrasts (Eisner and McQueen 2005; Kraljic and Samuel 2005; Kraljic and Samuel 2007; Reinisch and Holt 2014)

Finally, listeners can *infer* socio-indexical variables based on speech, but it's not clear what linguistic variables they use (cf. Thomas 2002). Of particular interest, listeners can infer a talker's regional dialect based on short excerpts (Cynthia G Clopper and Pisoni 2006,Cynthia G. Clopper and Pisoni (2007)). Listeners can also infer talker *identity* from sine-wave speech (Remez, Fellowes, and Rubin 1997), which removes non-phonetic voice quality cues to identity but preserves most phonetic information.

Our goals

In the current study, we aim to use the theoretical and computational tools of the ideal adapter framework to quantify the relationship between a number of socio-indexical and linguistic variables. Most (but not all, e.g., Cynthia G. Clopper, Pisoni, and Jong 2005; McMurray and Jongman 2011) of the work on this relationship between has been descriptive, aimed at establishing that differences between particular groups of talkers *exist* in the first place, and that listeners are sensitive to these differences at all. But the mere existence of differences does not establish exactly how *informative* or *useful* such grouping variables are for speech perception. In the ideal adapter framework, whether or not a listener is predicted to track cue distributions conditional on a particular grouping variable (like sex or dialect) critically depends on the informativity and utility of that grouping for speech perception.

We thus have two main goals in the current paper. First, quantifying this relationship between socio-indexical and linguistic variables would go a long way towards refining the predictions of the ideal adapter framework, and in turn understanding how listeners efficiently adapt to many different talkers. Second, we aim to show that the conceptual and computational tools of the ideal adapter framework offer a unifying perspective on linguistic and socio-linguistic perception. In this view, both can be seen as processes of *inference*, which relies on the same basic knowledge about the cue distributions that correspond to different underlying lingiustic and social variables.

[1]For the current study, we restrict ourselves to English, and focus primarily (but not exclusively) on *American* English.

Specifically, we aim to answer these three questions:

1. How *informative* are socio-indexical variables about the distribution of different acoustic-phonetic cues?
2. How *useful* is socio-indexical grouping for correct speech recognition?
3. How well can you *infer* socio-indexical variables based on acoustic-phonetic cue distributions alone?

We address these goals at different levels of socio-indexical grouping, and for two different sets of phonetic cues/contrasts. Next, we describe the datasets we analyze, the techniques we use, our findings, and finally the conclusions we can draw.

Methods

Datasets

We analyze speech from two corpora, one focusing on vowels and the other on stop consonant voicing.

Vowels

For vowels, we used data from the Nationwide Speech Project (Cynthia G. Clopper, Pisoni, and Jong 2005). Specifically, we analyzed first and second formant frequencies (F1xF2, measured in Hertz) recorded at vowel midpoints in isolated, read "hVd" words. This corpus contains 48 talkers, 4 male and female from each of 6 regional varieties of English: North, New England, Midland, Mid-Atlantic, South, and West (see map in Cynthia G. Clopper, Pisoni, and Jong 2005; regions based on Labov, Ash, and Boberg 2005). Each talker provided approximately 5 repetitions of each of 11 English monophthong vowels (plus "ey"), for a total of 2659 observations.

Because much of the variability in talkers is due to overall differences in formant frequencies, we also used Lobanov-normalized formant frequencies as input, in addition to the un-normalized formant frequencies in Hertz. Lobanov normalization z-scores F1 and F2 separately for each talker (Lobanov 1971).

Stop voicing

We analyzed data on word-initial stop consonant voicing in conversational speech from the Buckeye corpus (Pitt et al. 2007, extracted by Wedel, *in prep*). Voice onset time (VOT) was automatically extracted for 5984 word initial stops, 2264 voiced and 3720 voiceless, for labial, coronal, and dorsal places of articulation. Data came from 24 talkers, who were balanced male and female and younger/older than 40 years. On average, each talker produced 42 tokens for each phoneme (range of 5 – 156).

The major strength of this dataset is that it contains observations of both voiced and voiceless stops, which allows us to assess the utility of socio-indexical grouping factors for recognizing voiced vs. voiceless stops. However, it does not not contain data from talkers who vary on socio-indexical variables that are known to correlate with differences in VOT distributions, like native language background. Preliminary analyses of a collection of VOTs for only voiceless stops from French-English bilinguals (Lev-Ari and Peperkamp 2013) suggests that, even though these groups are known to produce different VOT distributions, the size of this effect is much smaller than even talker-level variability *within* the monolingual talkers in the Buckeye corpus (which, to foreshadow our results, is substantially smaller than for vowels).

Modeling

Each phonetic category was modeled as a normal distribution: stop voicing as univariate distributions of VOT, and vowels as bivariate distributions of F1 and F2. We used the maximum likelihood estimators for the model parameters, which are the sample mean and covariance matrix.

For each socio-indexical grouping level, we trained separate models for each phonetic category based on all tokens from that category and grouping level (holding out test data when necessary, see below). The grouping levels we considered were

- Marginal (all tokens)
- Sex (male/female)
- Age (younger/older than 40, VOT only)
- Dialect (six regions, vowels only)
- Diaelct+Sex (12 levels, vowels only)
- Talker

Comparing cue distributions

In order to evalute the *informativity* of socio-indexical variables with respect to cue distributions themselves, we use Kullback-Leibler (KL) divergence to measure how much the group-specific cue distributions differ from the overall (marginal) cue distributions. This measures the expected cost (in bits of extra message length) of encoding data from each group using a code that's optimized for the marginal distribution. We do this separately for each linguistic category and group, and then average the results, calculating bootstrapped confidence intervals over groups.

The KL divergence of Q from P is $DL(Q||P) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ (with density functions q and p respectively). In our case, $P = \mathcal{N}_G$ is a multivariate[2] normal cue distribution for the group, with mean μ_G and covariance Σ_G, while $Q = \mathcal{N}_M$ is the marginal multivariate normal cue distribution with mean μ_M and covariance Σ_M. With some simplification[3], the KL divergence of the marginal from the group distribution works out to be

$$DL(\mathcal{N}_M||\mathcal{N}_G) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_M^{-1}\Sigma_G) + (\mu_M - \mu_G)\Sigma_M^{-1}(\mu_M - \mu_G) - d + \log\frac{|\Sigma_M|}{|\Sigma_G|}\right)$$

where d is the dimensionality of the distribution, and logs are natural logarithms (we report KL divergence in bits, which is the above quantity divided by $log(2)$).

Speech recognition

Next, to address the *utility* of socio-indexical grouping for speech recognition, we calculate, for each level of socio-indexical grouping, the probability of correct recognition of phonetic categories. We do this using an "ideal listener" model (Clayards et al. 2008; Kleinschmidt and Jaeger 2015) that compute the posterior probability of a category given an observed cue value based on the likelihood of that cue being generated by each category's cue distribution. By doing this using, for instance, the cue distributions of each category produced by female talkers provides an estimate of how well a listener would be able to recognize speech from an unfamiliar female talker if all they knew was the talker's sex.

We want to determine the phonetic category v_i[4] of each of the cues x_i produced by a talker. If we assume that the listener knows that this talker belongs to group $g = j$, this inference is a straightforward application

[2]The math is the same for the univariate special case, as with VOT.

[3]See, for instance, http://stanford.edu/~jduchi/projects/general_notes.pdf, p. 13.

[4]x_i refers to a single observed cue value (possibly multidimensional, in the case of vowel formants), and x (without subscript) refers to a *vector* of multiple observations (from a single talker, unless otherwise specified). v_i refers to observation i's category, and g to a talker's group. $\sum_j$ refer to a sum over all possible values of j.

of Bayes Rule:

$$p(v_i|x_i, g=j) \propto p(x_i|v_i, g=j)p(v_i)$$

If, on the other hand, the listener does not know which group the talker belongs to, they have to marginalize out group. This ammounts to taking a weighted average of the posterior probabilities under each group, weighted by the probability that the talker belongs to that group, $p(g|x)$ (which we compute below):

$$p(v_i|x_i) = \sum_j p(v_i|x_i, g=j)p(g=j|x)$$

(where x refers to all the tokens produced by this talker).

For vowels, we classified vowel categories directly. For voicing, the only cue available is VOT, which does not (reliably) distinguish place of articulation. Thus, we classified voicing separately for each place of articulation, and then average the resulting accuracy.

Indexical group recognition

Finally, to address much cue distributions tell listeners about socio-indexical variables themselves, we classify each talker's socio-indexical group (at each level). This provides a measure of how well a listener would be able to determine, for instance, whether a talker was male or female based only on the distributions of cues they produce (even without knowing the intended category of each production).

As for speech recognition, we use an ideal observer model. That is, we compute the posterior probability of each socio-indexical group $g=j$, given all of the talker's observed cue values x:

$$p(g|x) \propto p(x|g)p(g) = \left(\prod_i p(x_i|g)\right) p(g)$$

The only complication is that, without knowing the the phonetic category of each observation a priori, each observation may have been generated by any of the phonetic categories. Thus, to determine the *overall* likelihood of observing a cue value x_i under group g, we first have to marginalize over categories v_i:

$$p(x_i|g) = \sum_k p(x_i|v_i=k, g)p(v_i=k|g)$$

For all of these classifications, we assume a flat prior on categories/groups. We perform this analysis separately for each level of socio-indexical grouping. For instance, we compute both $p(\text{sex}|x)$ and $p(\text{dialect}|x)$ for each talker.

Controls and assessing significance

We use a bootstrapping procedure to assess group differences and calculate confidence intervals for our estimates of utility, informativity, and inferability. In general, we resample talkers, in order to estimate the population-level variability from the limited sample of talkers in our datasets. However, for informativity, we have a single observations from each *group* (not talker), and so we resample groups instead.[5] For testing group differences, we use the proportion of resampled datasets with same-sign difference as the bootstrapped p value.

[5] While it's possible to resample talkers *before* calculating the KL divergence of group-level distributions from marginal, this systematically *increases* the KL divergence. The reason for this is that in bootstrapping, talkers are resampled with replacement, which means that the variance of the resulting resampled group-level distributions goes down, increasing the KL divergence from the (already higher variance) marginal distributions.

Cross-validation For classification, if test data is included in the training set, this artificially inflates accuracy at test (James et al. 2013, Section 5.1). Cross-validation controls for this by splitting data into training and test sets. For group-level models (sex, age, dialect, and dialect+sex), we use leave-one-talker-out cross-validation: train each group's models with test talker's observations held out. For the talker-specific models, we use 6-fold cross-validation (or leave-one-out when there were fewer than 6 tokens in a category for a talker), where each phonetic category is randomly split into 6 approximately equal subsets. Then, one subset of each category is selected for test, the models are trained on the remaining five, and the test data is classified as above.

Cross-validation is not only important because it provides an unbiased measure of classifier accuracy. It is also essential for testing the hypothesis that *group-level* cue distributions are useful to listeners. If the test talker is included in the training dataset, then the utility of that talker's own productions is confounded with any utility of the group itself.

Different group sizes For the vowel data, the different levels of grouping have very different group sizes, and this requires some caution. The broadest (sex) has 24 talkers per group (23 after holdout), while the most specific (dialect+sex) has only 4 (3 after holdout). This introduces a systematic bias in favor of broader groupings, because small sample sizes lead to noisier estimates of the underlying model, and hence lower accuracy (on average) at test (James et al. 2013, Section 2.2.2). To correct for this, after holding out the test talker, we randomly subsampled talkers (without replacement) within each group in the training set to be the same as the smallest group (3 talkers, based on Dialect+Sex[6]).

We use 20 different random subsamples for each talker, averaging accuracy over each resampled training set. A different subsampling is used for every talker, and thus any additional variance introduced by this procedure is accounted for by bootstrapping talkers. The estimates obtained in this way allow us to compare accuracy across groupings with different group sizes, but at the cost of underestimating the true group-level accuracy across the board. As such, they must be considered a useful lower bound on the utility of socio-indexical groupings.

Results

Informativity of socio-indexical groupings about cue distributions

We first analyze how informative each socio-indexical grouping variable is about the cue distributions of each phonetic category. As described in the methods, we measure informativity by the average KL divergence between the group-conditional cue distributions and the unconditioned (marginal) cue distributions.

Figure 1 plots the KL divergence of cue distributions at different levels of grouping from marginal distributions, across contrasts (vowels and stop voicing) and cues (vot, raw/Lobanov-normalized F1xF2). There are two clear patterns.

First, socio-indexical variables are in general more informative for vowels than for stop voicing, even using normalized formants as input. Strikingly, the *most* informative variable for VOT—talker identity—is roughly as informative as the *least* informative variable for Lobanov-normalized F1xF2 (Sex). As Figure 2 shows, this means that, on average, individual talkers' VOT distributions diverge from the marginal distribution (B) at roughly the same level that the male and female distributions of normalized F1xF2 diverge from the marginal normalized F1xF2 distributions (A).

Second, there are substantial differences in the informativity of the socio-indexical grouping variables we considered. The overall pattern is that more specific grouping factors are more informative than broader groupings. The noteable exception to this pattern is the Sex is the most informative variable for un-normalized

[6] We also ran the analyses resampling each group to 7 talkers, which corresponds to the Dialect-level group size after holdout (excluding the Dialect+Sex grouping, since there are only 4 talkers per group before holdout). Besides a small increase in overall accuracy (because of the reduced variance of the distribution estimates), this did not substantially change the results.

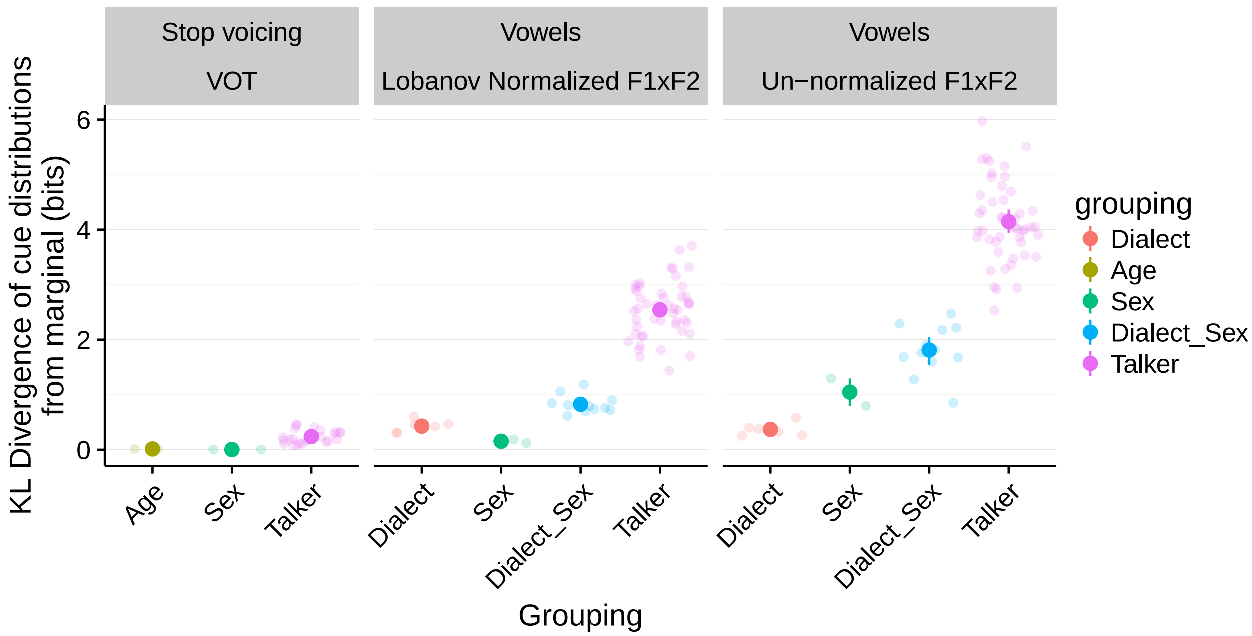

Figure 1: Socio-indexical variables are more informative about cue distributions for vowel (formants) than for stop voicing (vot). On top of this, more specific groupings (like Talker and Dialect+Sex) are more informative than broader groupings (Sex). This is indicated by higher KL divergence of each grouping level from marginal (showing mean and 95% bootstrapped CIs over groups).

F1xF2 distributions, which reflects the fact that overall sex differences explain much (but not all) of the talker variation in F1xF2 (Hillenbrand et al. 1995; Johnson 2006).

As Figure 3, while the relative ordering of grouping variables' informativity is consistent across vowels, their actual degree of informativity varies quite a bit. Dialect (and Dialect+Sex) is particularly informative for `aa`, `ae`, `eh`, and `uw`, vowels with distinctive variants in at least one of the dialect regions from the NSP. `aa` is undergoing a merger with `ao` in some regions, `ae` and `eh` participate in the northern cities chain shift, and `uw` is fronted in some regions (and in others, but only by female talkers; Cynthia G. Clopper, Pisoni, and Jong 2005).

Finally, individual dialects also vary in how informative they are about vowel formant distributions. Figure 4 shows that talkers from the North dialect region produce vowels—`ae` and `aa` in particular—with formant distributions that deviate markedly more from the marginal distributions than any of the other dialects. Other dialects have, on average, similar deviations from marginal. The high deviation of `uw` by New England talkers is the result of these talkers producing a very low-variance, back (low F2) distribution. Similarly, Mid-Atlantic talkers produce a low-variance `ey` distribution that is higher and fronter than average. Finally, the Mid-Atlantic `aa` is, like the Northern `aa`, non-merged with `ao` (Cynthia G. Clopper, Pisoni, and Jong 2005) and hence deviates from the marginal `aa` substantially.

Utility of socio-indexical groupings for speech recognition

Next, we evaluate the *utility* of each grouping variable for speech recognition. The utility of a grouping variable—like dialect—can be quantified as the probability of correct recognition given the cue distributions conditioned on each group—e.g., each dialect—using an ideal listener model (Clayards et al. 2008; Kleinschmidt and Jaeger 2015).

Figure 5 shows the probability of correct recognition for stop voicing/vowel, based on the cue distrbutions at each level of grouping. As with informativity about the distributions themselves, there's an asymmetry

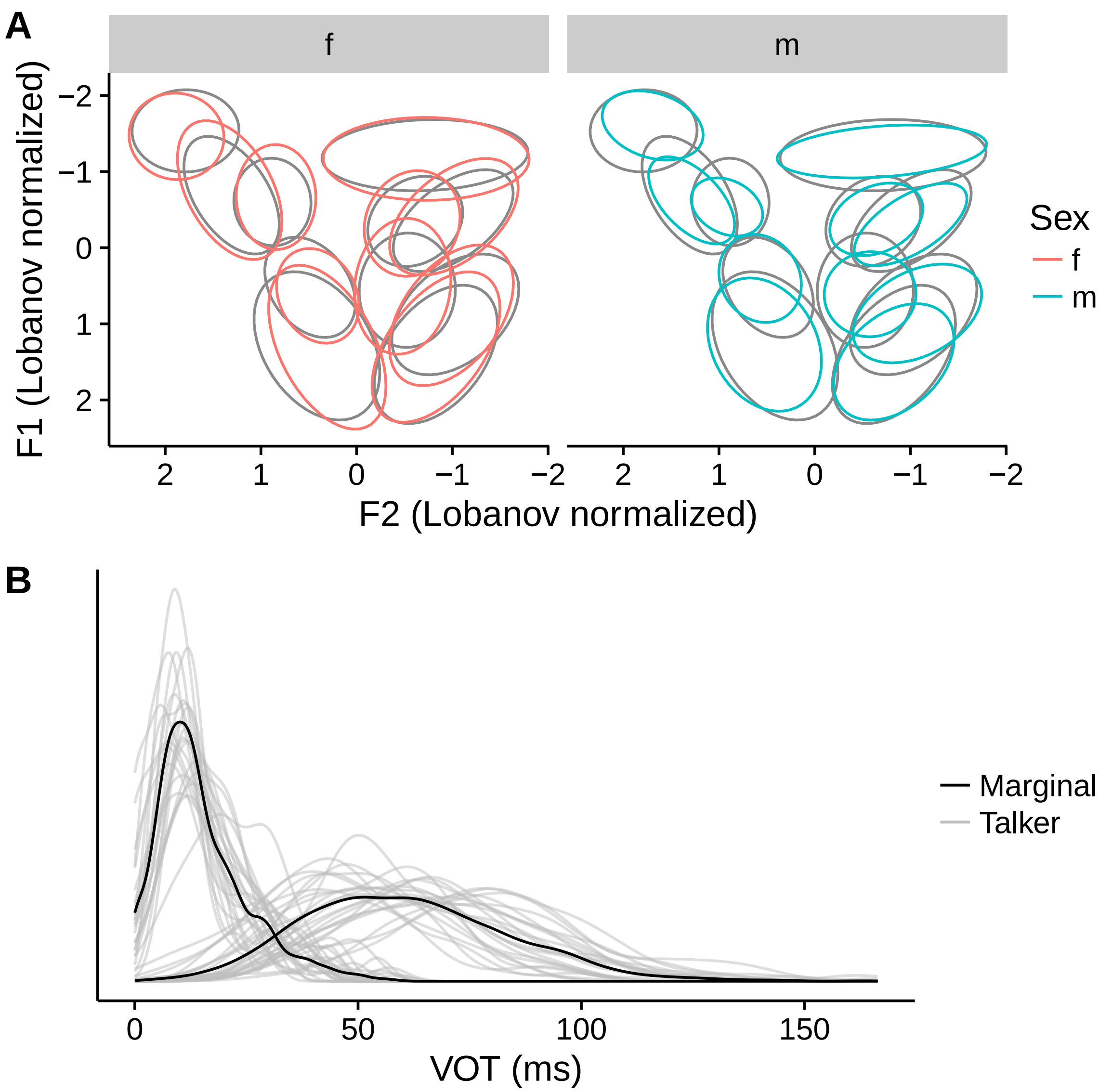

Figure 2: Male and female distributions of Normalized F1xF2 diverge from the marginal distributions (A) only slightly less than talker-specific VOT distributions diverge from marginal (B).

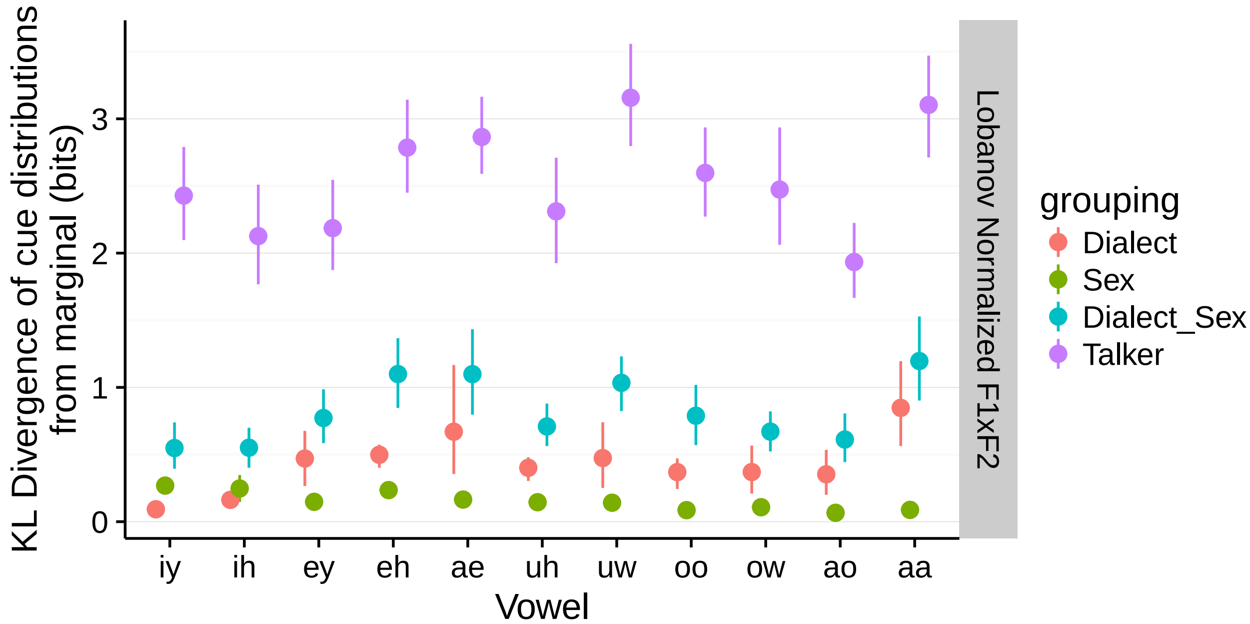

Figure 3: Individual vowels vary substantially in the informativity of grouping variables about their cue distributions. Only normalized F1xF2 is shown to emphasize dialect effects.

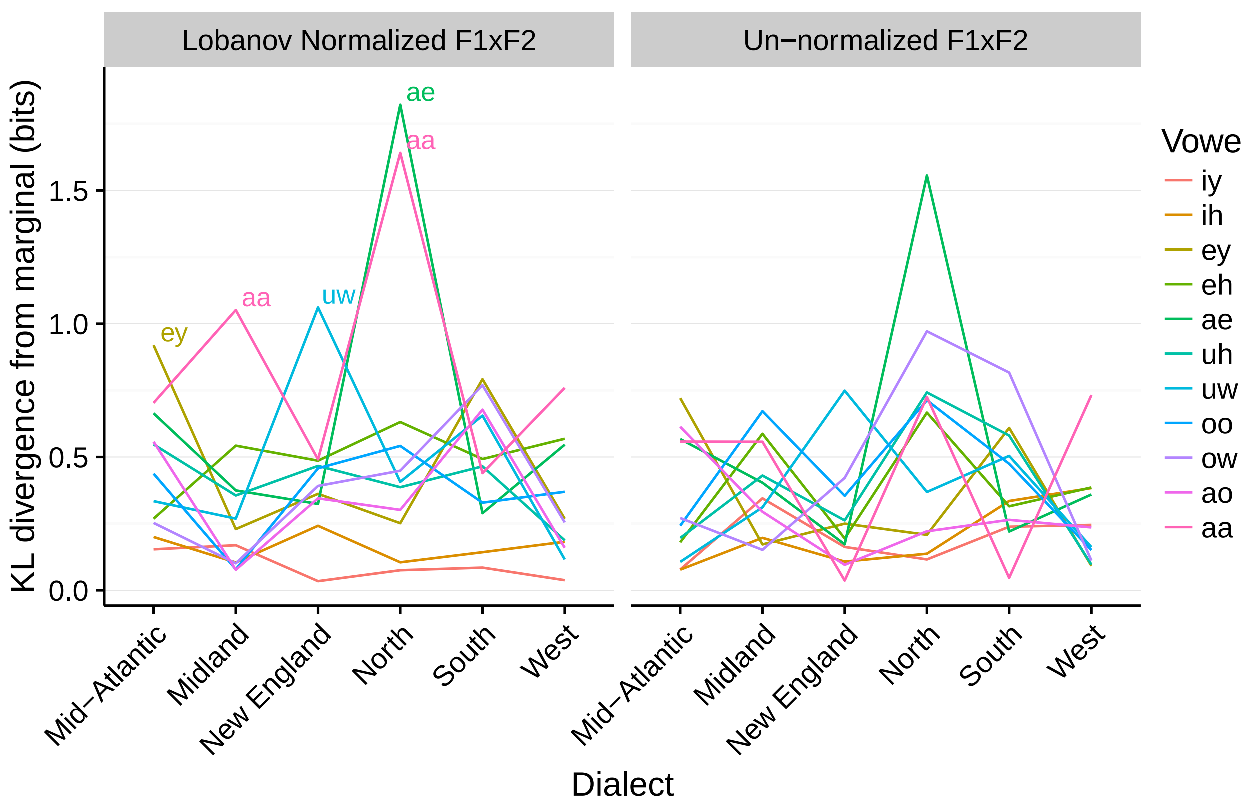

Figure 4: A small number of dialect/vowel combinations account for most of the divergence of dialect-specific vowel formant distributions. In particular, the distribution of `ae` and `aa` produced by Northern talkers diverge markedly more than any other vowel/dialect combination.

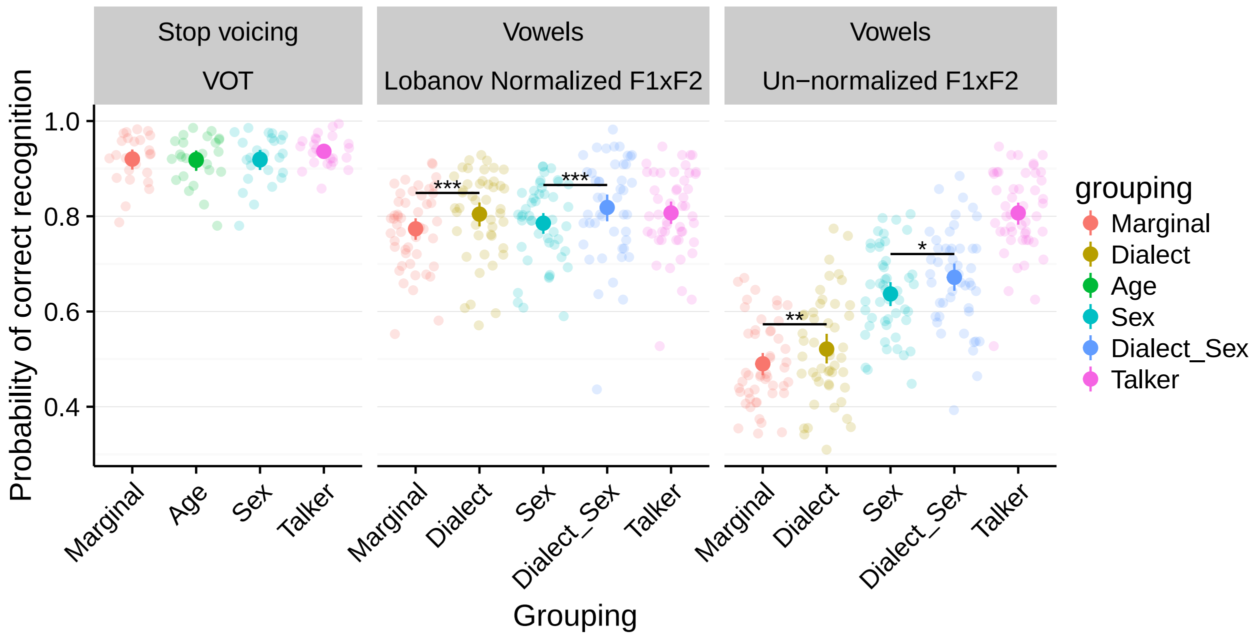

Figure 5: Speech recognition accuracy using for marginal, group-level, and talker-specific cue distributions. Small points show individual talkers, and large points and lines show mean and bootstrapped 95% CIs over talkers. Marginal and group-level accuracy is based on leave-one-talker out cross-validation, and talker-specific on 6-fold cross-validation (or leave-one-token-per-category out if there are fewer than 6 tokens per category). Bars and stars show signficant increases in accuracy when conditioning on dialect, alone or in addition to sex. Here and elsewhere: `*` $p < 0.05$, `**` $p < 0.01$, and `***` $p < 0.001$.

between vowels and stop voicing in the overall utility of socio-indexical variables for speech recognition. Probability of correct recognition is overall higher for stop voicing than vowels. Voicing recognition is also less sensitive to the particular grouping variable, which is consistent with the finding above that VOT distributions do not differ substantially at different levels of socio-indexical grouping. There is, however, a slight advantage for using talker-specific VOT distributions for recognition, over marginal, age-, and sex-conditional distributions (on the order of 2% increase in accuracy, all three p < 0.01).[7]

Normalized input results in higher vowel recognition accuracy across the board, again paralleling the findings about the cue distributions themselves. The one exception is at the level of talker-specific distributions, where recognition accuracy is unchanged (since Lobanov normalization is a linear transformation of the input, which leaves the structure of the categories within each talker unchanged).

Also paralleling the cue distributions themselves, classifying according to sex-specific distributions provides the biggest boost in recognition for un-normalized formant accuracy. For normalized input, none of the socio-indexical grouping factors provide much of an advantage over the marginal distributions. In both cases, dialect provides a small but consistent advantage for recognition, both alone and in combination with sex (increasing accuracy by 3% on average, all $p < 0.05$).

The utility of socio-indexical grouping for recognizing individual vowels largely mirrors the overall pattern (Figure 6). The exception is that the utility of conditioning on dialect varies substantially from one vowel to the next. For most vowels, conditioning on dialect makes little difference to correct recognition. But for a few—particularly `ae` and `eh`—conditioning on dialect increases accuracy by nearly 10%. In the case of normalized formant input, accuracy using dialect-conditioned distributions actually equals or surpasses the accuracy with talker-specific distributions.

The overall utility of dialect also varies substantially based on the talker's actual dialect (Figure 7). This

[7]These comparisons are also significant when using log-odds of correct recognition, rather than raw probabilities.

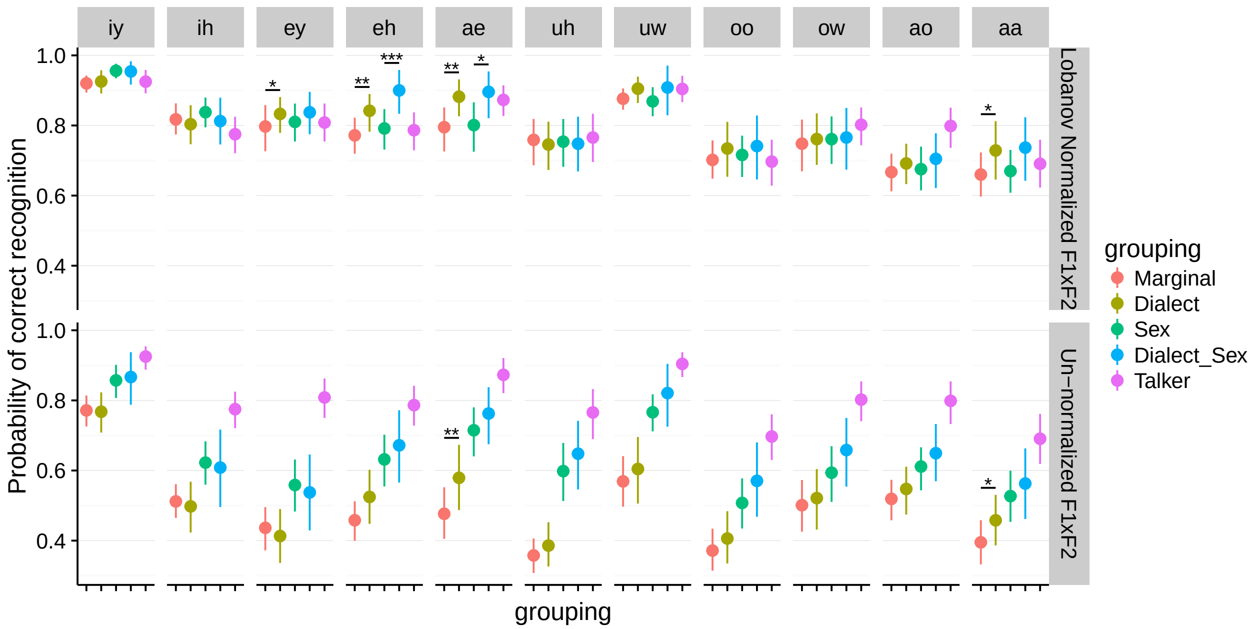

Figure 6: Probability of correct recognition varies across vowels, overall and according to the socio-indexical grouping variable. Bars and stars show significant improvement from conditioning on dialect, above marginal or in addition to sex alone.

is consistent with the fact that dialects themselves vary in how much they deviate from both the norms of Standard American English (Cynthia G. Clopper, Pisoni, and Jong 2005) and from the marginal cue distributions in this dataset (Figure 4). Most notably, conditioning on dialect provides a consistent advantage for talkers from the North dialect region, on the order of 10%.

Inferring socio-indexical variables from cue distributions

Finally, we assess how well listeners could infer socio-indexical variables from unlabeled acoustic-phonetic cues, based on the group-conditional cue distributions. We measure this by the accuracy with which an "ideal listener" can categorize a talker's group membership for each socio-indexical grouping variable.

For most groupings, it is possible to infer each talker's group with above chance accuracy, given only that talker's unlabeled observed cues (Figure 8; all $p < 0.01$, except for inferring a talker's sex based on their VOT distributions, and dialect from un-normalized F1xF2 distributions, both $p > 0.15$).

In some respects, these results mirror the patterns of informativity about the cue distributions themselves (Figure 1). Vowel formant distributions vary more according to group than do stop VOT distributions, and likewise socio-indexical group can, on the whole, be inferred with higher accuracy based on vowel formants than on VOT. For vowels specifically, much of the variability across talkers is driven by sex differences, and this is the grouping variable that's easiest to infer of the three we tested.

However, in other respects, these results do *not* simply mirror informativity. For instance, un-normalized F1xF2 distributions diverge from marginal more for dialect+sex than they do for sex alone, but accuracy at inferring a talker's sex is nearly at ceiling, while accuracy is barely above chance for inferring a talker's dialect+sex. Likewise, normalized F1xF2 distributions given dialect and dialect+sex diverge from marginal less than non-normalized, but accuracy at inferring these two grouping variables is higher for normalized than non-normalized.

Why the discrepancy? The informativity measure we used was the average divergence of *each category's* cue distribution. For inferring indexical groups, we assumed that listeners do not know the intended category

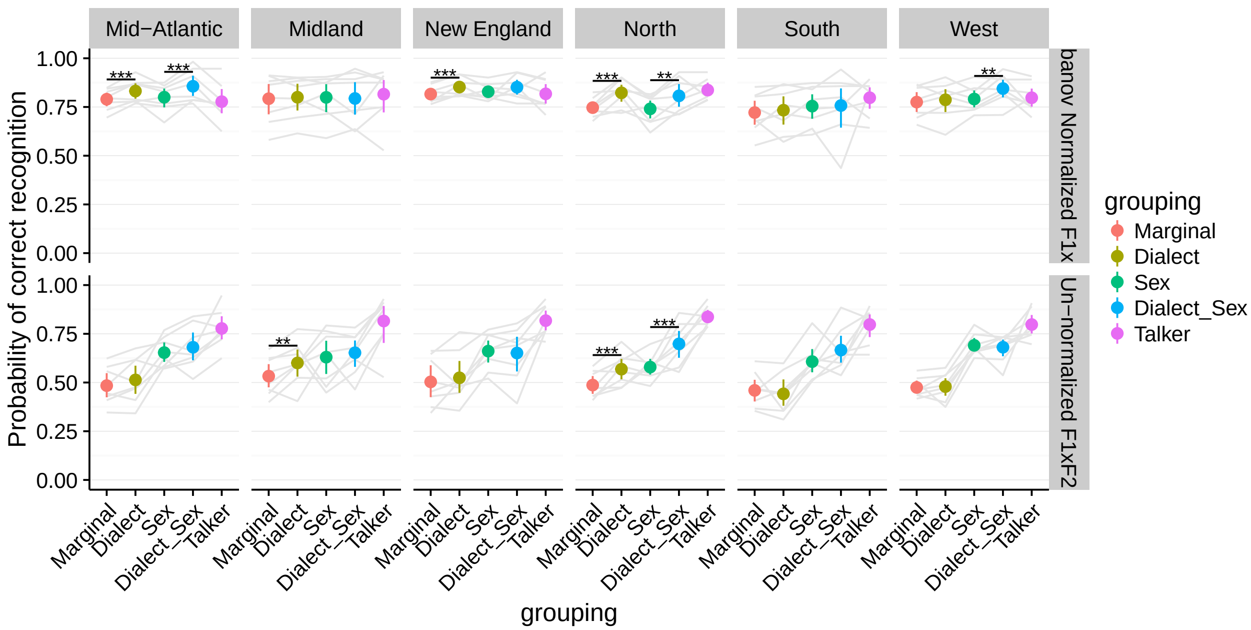

Figure 7: The utility of socio-indexical variables varies across dialects. Dialect itself is particularly informative only for talkers from the Mid-Atlantic and North regions. Each line shows a single talker, to emphasize within-talker changes in accuracy with grouping level, and large points and confidence intervals show mean accuracy and bootstrapped 95% CIs over talkers.

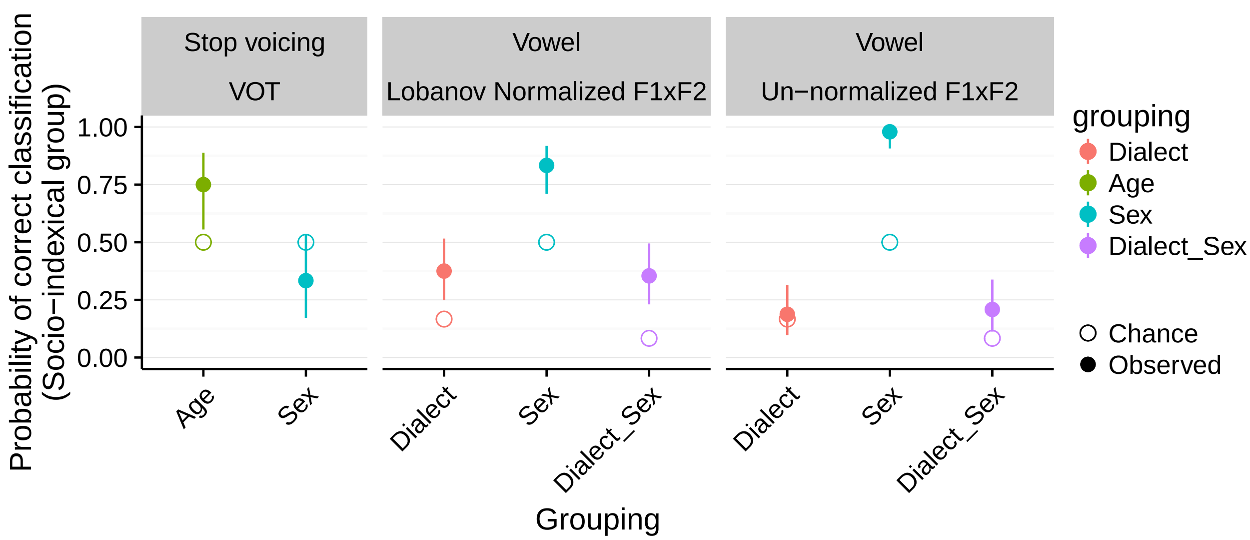

Figure 8: Probability of correctly classifying a talker's socio-indexical group varies with the grouping variable, contrast, and cues. A talker is correctly classified if the overall posterior probability of their actual group given their unlabeled productions is the highest of all groups.

of each observation, and so the relevant likelihoods are each based on a *mixture* of the category-specific distributions. Even if there are some categories whose individual distributions diverge across groups, the overall mixture distribution of all categories may still be too similar to allow for the group to be reliably inferred.

General Discussion

Our results show that, on the whole, socio-indexical grouping variables are *informative* about phonetic cue distributions, *useful* for improving speech recognition, and can be *inferred* from phonetic cues themselves. However, the extent to which these are true depends on the particular grouping variable and particular phonetic categories/cues involved. Socio-indexical variables are more useful, more informative, and more easily inferred for vowels than for stop consonant voicing. Some variables are broadly useful (sex, talker identity) while others are useful only in certain, specific contexts (dialect for certain vowels/dialect combinations).

Our results also speak to the relationship between informativity, utility, and inferrability themselves. In general, informativity and utility mirror each other: conditioning on a socio-indexical variable is more useful for speech recognition when the corresponding conditional cue distributions diverge more from the overall or marginal distributions. But being useful for speech recognition does not always mean that a socio-indexical variable can be easily inferred from phonetic cues alone, or vice-versa.

Here we discuss the implications of these results. First, the ideal adapter generally predicts that listeners should track conditional distributions for groups that are informative and useful for speech recognition. By directly quantifying the utility and informativity of a number of grouping variables, our results are an important step towards making more specific predictions about what group-level representations listeners should maintain if, as assumed by the ideal adapter, they are taking advantage of the structure that is actually present in cross-talker variability. Second, our results shed light on discrepancies between phonetic contrasts in listeners' willingness to generalize recalbration/perceptual learning from one talker to another. Third, this paper provides a proof of concept for the idea that, like phonetic judgements, socio-linguistic judgements can be productively viewed as a sort of inference under uncertainty. This suggests the potential for a tighter integration of sociolinguistic and psycholinguistic perspectives on speech perception.

What to track?

Treating speech perception as a problem of inference under uncertainty—as the ideal adapter does—highlights the importance of listeners' knowledge about the distributions of cues that are produced for each linguistic unit. A major question that this perspective raises is *what* linguistic, socio-indexical, and acoustic/phonetic variables listeners are learning distributions for. The ideal adapter does not directly answer this question, but provides a set of conceptual and quantitative tools for addressing it. The studies reported here take these tools and apply them to data on how many different talkers produce two different sets of phonetic categories. We hope that by doing so we provide a proof of concept for the broad usefulness of these tools. One purpose they might be put to is to formulate hypotheses about what distributions listeners should track.

At the highest level, the ideal adapter predicts that listeners should not track *everything*. Rather, listeners need only track the joint distributions of variables that are informative. At the level of phonetic categories themselves, this means that (for instance) there's no reason for listeners to track each vowel's distribution of preceding VOT[8] (or more absurdly, completely unrelated physical quantities like temperature or barometric pressure). This also applies at the level of socio-indexical grouping variables: listeners get no benefit for tracking separate distributions for different groups of talkers for a cue that does not systematically vary between those groups.

In fact, it can actually *hurt* a listener to track cue distributions at a level that's not informative. The reason for this is related to the idea of bias-variance tradeoff from machine learning (James et al. 2013, Section 2.2.2).

[8]Barring, of course, the possibility that VOT is systematically affected by neighboring vowels, and hence informative about them.

Given the same amount of data, tracking multiple, specific distributions will result in noisier, less accurate estimates than lumping together all the observations in a single distribution. This price may be worth paying for a listener when there are large enough differences between groups that treating all observations as coming from the same distribution *biases* the estimates of the underlying distribution (and hence the inferences that listeners make based on those distributinos) far enough away from the true structure of the data. To take a concrete example, modeling each vowel as a single distribution of (un-normalized) formants across all talkers results in high-variance, overlapping distributions which have low recognition accuracy. But modeling them as two distributions—one for males, and one for females—provides much more specific estimates and higher classification accuracy, as shown by Figures 1 and 5.

Thus, the ideal adapter predicts that listeners should learn separate cue distributions for levels of a socio-indexical grouping variable when that variable has high *informativity* about some categories' cue distributions and/or high *utility* for speech recognition. However, However, the notions of informativity and utility apply beyond better *speech* recognition per se. Listeners extract a lot of non-phonetic/linguistic information from speech signal. To properly define the informativity or utility of a particular grouping variable, we need to consider the *goals* of speech perception, which go beyond just recognizing phonetic categories. Sociolinguistics recognizes that, in many cases, the communication of social information is just as if not more important than the communication of linguistic information. Groupings that are *socially meaningful* can thus be informative and justify being tracked, even if ignoring them has a negligible effect on speech recognition, as long as the corresponding cue distributions carry some information about relevant social variables. In our results, dialect is a good example: on the whole, ignoring dialect doesn't have huge consequences on recognition accuracy. But it can be inferred (at least above chance) based on vowel F1 and F2, and listeners are plausibly interested in determining a talker's regional origins for a variety of reasons.

An additional consideration is that listeners are not simply told which variables are informative and which are not. They must actually *infer* what distributions are actually worth tracking. Moreover, every listener's experience with talker variability will be different, and so a variable that is informative in one listener's experience may be irrelevant in another's. While the analyses we present here go a long way toward focusing the predictions of the ideal adapter framework, they must be combined with knowledge of each listener's own personal history—either assumed, or somehow measured, even approximately—in order to make specific predictions for a particular subject or population of subjects. This same logic applies to which socio-indexical variables are of direct interest to a listener: social categories that are highly important in one person's social world may be completely meaningless in another's. An important aspect of the research program laid out by the ideal adapter framework is to probe listeners' prior beliefs *directly* (which the previous chapter is a first step towards)

Finally, our results suggest that the input representation—the cue space over which categories are distributions—can affect which variables are informative or not. For vowels, using Lobanov-normalized formants as input substantially reduces the informativity and utility of sex as a grouping factor, but *increases* the utility of dialect in many cases. From a listener's perspective, dialect would appear to be relatively uninformative without normalization. This points to a complex interaction between normalization and adaptation/perceptual learning as strategies for coping with talker variability. Both strategies are, in fact, used by listeners, but the interaction between them is poorly understood (Weatherholtz and Jaeger 2016).

Consequences for adapting to unfamiliar talkers

The results of this study also tell us something about how listeners might adapt to an unfamiliar talker. The ideal adapter links informativity to adaptation, and the results here allow us to make more specific predictions based on the ideal adapter, in two ways.

First, the informativity of talker identity is a measure of the variability across talkers. When talker identity is highly informative, there's more variabiltiy across talkers, and the ideal adapter predicts that prior experience with other talkers will be less relevant, resulting in faster and more complete adaptation to an unfamiliar talker. We found here that talker identity is less informative about VOT distributions than it is for vowel formant distributions. Hence, the ideal adapter predicts that listeners will adapt to talker-specific VOT distributions more slowly, and be more constrained by prior experience with other talkers. The first prediction is borne out by Kraljic and Samuel (2007), who compared recalibration of a voicing contrast with a fricative place contrast. It's also borne out indirectly by the modeling work in Chapters NNN and NNN, which found that the effective prior sample size for /b/-/d/ (which, like vowels, is primarily cued by formant requencies) is much lower than for /b/-/p/ (cued by VOT).

Second, the informativity of higher-level grouping variables is linked to *generalization* across talkers: if two talkers are from groups that tend to differ, listeners should treat them separately and not generalize from experience with one talker to the other. Likewise, if two talkers are from the same group, listeners *should* generalize. We found that talker sex is informative for vowel formant distributions, but not for VOT, which means that listeners *should* generalize from a male to a female talker (and vice-versa) for a voicing contrast, but *not* for a vowel contrast. Listeners do, in fact, tend to generalize voicing recalibration across talkers of different sexes (Kraljic and Samuel 2006; Kraljic and Samuel 2007). While there's (to our knowledge) no data on cross-talker generalization for vowel recalibration, listeners tend not to generalize across talkers for fricative recalibration (Eisner and McQueen 2005; Kraljic and Samuel 2007), which (like vowels) are cued by spectral cues that vary across talkers and by gender (Newman, Clouse, and Burnham 2001; Jongman, Wayland, and Wong 2000; McMurray and Jongman 2011).

There is also evidence for the prediction of generalization *within* informative groups. In the absence of evidence that two talkers from the same group (e.g. two males) produce a contrast differently, experience with one provides an informative starting point for comprehending (and adapting to) the other. Along these lines, Zande, Jesse, and Cutler (2014) found that listeners generalize from experience with one male talker's pronunciation of a /b/-/d/ contrast to another, unfamiliar male.

Finally, it's important to point out that these predictions are best thought of as *biases* that can be overcome with enough of the right kind of evidence (Kleinschmidt and Jaeger 2015). For instance, listeners can overcome their bias to generalize experience with VOT and learn talker-specific VOT distributions, but it requires hundreds of observations from talkers who produce very different VOT distributions (Munson 2011). Likewise, listeners will generalize recalibration of a fricative contrast from a female to a male talker given the right kind of test stimuli (Reinisch and Holt 2014).

Sociolinguistic inference

Our findings suggest that socio-linguistic judgements can—like linguistic judgements—be viewed as probabilistic inference. In this view, both social and linguistic judgements rely on knowledge of how different underlying categories—social and linguistic—are probabilistically realized as distributions of observable cues. Just like each vowel (for instance) is realized as a distribution of F1 and F2 values, each dialect is *also* realized as an F1xF2 distribution (along with many other cues). When a listener hears a talker produce particular cue values, they can use knowledge of these distributions to compare how well each possible underlying social variable *explains* the speech they've observed. We find that this kind of model can classify a talker's dialect at roughly the same accuracy (10-40%) as human listeners in a forced-choice task based on sentences spoken by the same talkers (Cynthia G Clopper and Pisoni 2006).

The idea of socio-linguistic judgements of inference fits naturally within the ideal adapter framework, which holds that listeners are simultaneously making at least three kinds of inferences in the normal course of speech perception:

1. *What* a talker is saying
2. *How* that talker says things
3. *Who* that talker is, in relation to other talkers

The third level of inference is essential for talker-invariant speech perception: knowing *who* a talker is allows listeners to take advantage of their prior experience with other, similar talkers (Kleinschmidt and Jaeger 2015). Of course, listeners likely also want to know who a talker is for reasons that have nothing to do with accurate speech recognition per se. To the extent that a talker's way of realizing linguistic variables says anything about who they are their speech is informative about their identity, at the same time as their identity is informative about their speech. Thus both sociolingusitic and psycholinguistic considerations lead to the idea that social inferences may well be inextricable from linguistic inferences.

Realizing that socio-linguistic judgements can be treated as a kind of inference is a potentially powerful idea, but it is impotant to realize that it is not, per se, a complete *model* of socio-linguistic judgements. Rather, it is a framework for developing such a model. In this view, the particular inferences that a listener would draw based on particular linguistic input depends not only on the distributions of cues in the world but just as much on the listener's own, internal model of how social variables relate to each other. Or, as it's more commonly put, a listener's stereotypes or ideologies about language use and social identity.

Careful sociolinguistic work is required to tease these factors out. One example comes from Levon (2014). He finds that when UK listeners hear a male talker with high /s/ spectral center of gravity (COG), they infer that the talker is a gay man. But when they hear a male talker with high /s/ spectral COG *and* TH-fronting (i.e., /f/ for /TH/), they judge the talker to be a working class straight man. That is, the inference that the talker is working class *blocks* the inference that he is gay. These sorts of effects are perfectly compatible with an inference-based perspective, but they depend on the specific contents of the listeners internal model of how social variables are related to each other and to observable cues (for examples in other domains, see R. A. Jacobs and Kruschke 2010). Such internal models are not directly derivable from production data like we analyze here, but rather require probing a listner's subjective, implicit beliefs (as in the previous chapter).

A lower bound

Finally, it is important to note that our results here constitute a *lower bound* on the informativity or utility of different levels of socio-indexical grouping.[9] We model cue distributions for a particular group as a *single* normal distribution over observed cue values. In reality, a hierarchical model is more appropriate, since different levels of grouping can nest within each other. For instance, each dialect group is likely better modeled as a *mixture* of talker-specific distributions, which each exhibit dialect features to a varying degree. This is especially important for *adaptation* to an unfamiliar talker, since a group-level distribution conflates *within* and *between* talker variation, both of which have separate roles to play in belief updating.

The approach to group-level modeling that we take here is roughly equivalent to the *posterior predictive* distribution of a fully hierarchical model, which integrates over lower levels of grouping to provide a single distribution of cues given the group (and phonetic category). This corresponds to the best guess a listener would have *before* hearing anything from an unfamiliar talker, if the only information they had about that talker was their group membership. As the listener hears more cue values from the talker, the hierarchical nature of grouping structure becomes more important and can provide (in principle) a significant advantage over what we measured here. But modeling this process is quite a bit more complicated and we leave it for future work. Nevertheless, modeling each category as a single, "flat" distribution per group may well prove a useful approximation, or even a boundedly-rational model of how listeners take advantage of different levels of grouping structure (and similar approaches have been used in, e.g., motor control Körding et al. 2007).

Conclusion

Socio-linguistic variables like age, sex, and regional origin have been identified by sociolinguistics as factors that systematically affect the realization of linguistic categories. Using an ideal observer framework, we quantified the extent to which a range of these variables are *informative* about the distributions of acoustic cues corresponding to linguistic categories, *useful* for recognizing those categories, and can themselves be *inferred* from unlabeled cues. Our results show that the utility and informativity of a particular socio-indexical variable are closely related but not identical, while inferrability is distinct. Moreover, we demonstrate how

[9] Even above and beyond the limitations imposed by unqueal numbers of talkers in each group, which necessitates subsampling talkers in the larger groups in order to meaningfully compare accuracy.

this method for quantifying these factors allows them to be compared across phonetic categories as well as cues/contrasts (VOT vs. F1xF2).

Together, these results show that the idea of inference under uncertainty, when applied to speech perception, provides a unifying perspective on both linguistic and socio-linguistic perception. This perspective leads to conceptual and computational tools for addressings questions that are of interest to psycholinguistics and sociolinguistics, as well as developing new bridges between the two.

Allen, J. Sean, Joanne L Miller, and David DeSteno. 2003. "Individual talker differences in voice-onset-time." *The Journal of the Acoustical Society of America* 113 (1): 544. doi:10.1121/1.1528172.

Anderson, John R. 1990. *The adaptive character of thought*. Studies in Cognition. Hillsdale, NJ: Lawrence Erlbaum Associates.

———. 1991. "The adaptive nature of human categorization." *Psychological Review* 98 (3): 409–29. doi:10.1037//0033-295X.98.3.409.

Bladon, R. A W, C. G. Henton, and J. B. Pickering. 1984. "Towards an auditory theory of speaker normalization." *Language and Communication* 4 (1): 59–69. doi:10.1016/0271-5309(84)90019-3.

Caramazza, a, G H Yeni-Komshian, E B Zurif, and E Carbone. 1973. "The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals." *The Journal of the Acoustical Society of America* 54: 421–28. doi:10.1121/1.1913594.

Chodroff, Eleanor, John Godfrey, Sanjeev Khudanpur, and Colin Wilson. 2015. "Structured Variability in Acoustic Realization : A Corpus Study of Voice Onset Time in American English Stops." In *Proceedings of the 18th International Congress of Phonetic Sciences*, edited by The Scottish Consortium for ICPhS. Glasgow, UK: the University of Glasgow.

Clayards, Meghan A, Michael K Tanenhaus, Richard N Aslin, and Robert a Jacobs. 2008. "Perception of speech reflects optimal use of probabilistic speech cues." *Cognition* 108 (3): 804–9. doi:10.1016/j.cognition.2008.04.004.

Clopper, Cynthia G, and David B Pisoni. 2006. "Effects of region of origin and geographic mobility on perceptual dialect categorization." *Language Variation and Change* 18 (2): 193–221. doi:10.1017/S0954394506060091.

Clopper, Cynthia G., and David B. Pisoni. 2007. "Free classification of regional dialects of American English." *Journal of Phonetics* 35 (3): 421–38. doi:10.1016/j.wocn.2006.06.001.

Clopper, Cynthia G., David B. Pisoni, and Kenneth de Jong. 2005. "Acoustic characteristics of the vowel systems of six regional varieties of American English." *The Journal of the Acoustical Society of America* 118 (3): 1661. doi:10.1121/1.2000774.

Docherty, Gerard, Dominic Watt, Carmen Llamas, Damien Hall, and Jennifer Nycz. 2011. "Variation in Voice Onset Time Along the Scottish-English Border." In *Proceedings of ICPhS XVII*, edited by Wai-Sum Lee and Eric Zee, 591–94. August. International Phonetic Association.

Eisner, Frank, and James M McQueen. 2005. "The specificity of perceptual learning in speech processing." *Perception & Psychophysics* 67 (2): 224–38.

Feldman, Naomi H, Thomas L Griffiths, and James L Morgan. 2009. "Learning phonetic categories by learning a lexicon." *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–13.

Feldman, Naomi H, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. 2013. "A role for the developing lexicon in phonetic category acquisition." *Psychological Review* 120 (4): 751–78. doi:10.1037/a0034245.

Flege, James Emil. 1987. "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification." *Journal of Phonetics* 15 (1): 47–65.

Hay, Jennifer, and Katie Drager. 2010. "Stuffed toys and speech perception." *Linguistics* 48 (4): 865–92. doi:10.1515/ling.2010.027.

Heald, Shannon L M, and Howard C. Nusbaum. 2015. "Variability in vowel production within and between days." *PLoS ONE* 10 (9). doi:10.1371/journal.pone.0136791.

Hillenbrand, J, L A Getty, M J Clark, and K Wheeler. 1995. "Acoustic characteristics of American English vowels." *The Journal of the Acoustical Society of America* 97 (5.1): 3099–3111.

Jacobs, Robert A, and John K Kruschke. 2010. "Bayesian learning theory applied to human cognition." *Wiley Interdisciplinary Reviews: Cognitive Science*, n/a–/a. doi:10.1002/wcs.80.

James, Gareth, Daniela Witten, Robert Tibshirani, and Trevor Hastie. 2013. "An Introduction to Statistical Learning with Applications in R." *Book*, 431. doi:10.1007/978-1-4614-7138-7.

Johnson, Keith. 2005. "Speaker normalization in speech perception." In *The Handbook of Speech Perception*, edited by David B Pisoni and Robert E Remez, 363–89. Figure 1. Oxford: Blackwell Publishers.

———. 2006. "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology." *Journal of Phonetics* 34 (4): 485–99. doi:10.1016/j.wocn.2005.08.004.

Johnson, Keith, Elizabeth A Strand, and Mariapaola D'Imperio. 1999. "Auditory-visual integration of talker gender in vowel perception." *Journal of Phonetics* 27 (4): 359–84. doi:10.1006/jpho.1999.0100.

Jongman, Allard, Ratree Wayland, and Serena Wong. 2000. "Acoustic characteristics of English fricatives." *The Journal of the Acoustical Society of America* 108 (3): 1252. doi:10.1121/1.1288413.

Kleinschmidt, Dave F, 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. doi:10.1037/a0038695.

Körding, Konrad P, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. 2007. "Causal inference in multisensory perception." *PLoS ONE* 2 (9): e943. doi:10.1371/journal.pone.0000943.

Kraljic, Tanya, and Arthur G Samuel. 2005. "Perceptual learning for speech: Is there a return to normal?" *Cognitive Psychology* 51 (2): 141–78. doi:10.1016/j.cogpsych.2005.05.001.

———. 2006. "Generalization in perceptual learning for speech." *Psychonomic Bulletin & Review* 13 (2): 262–8.

———. 2007. "Perceptual adjustments to multiple speakers." *Journal of Memory and Language* 56 (1): 1–15. doi:10.1016/j.jml.2006.07.010.

Labov, William, Sharon Ash, and Charles Boberg. 2005. *The Atlas of North American English*. doi:10.1515/9783110206838.

Lev-Ari, Shiri, and Sharon Peperkamp. 2013. "Low inhibitory skill leads to non-native perception and production in bilinguals' native language." *Journal of Phonetics* 41 (5). Elsevier: 320–31. doi:10.1016/j.wocn.2013.06.002.

Levon, Erez. 2014. "Categories, stereotypes, and the linguistic perception of sexuality." *Language in Society* 43 (5): 539–66. doi:10.1017/S0047404514000554.

Liberman, Alvin M, F S Cooper, D P Shankweiler, and M Studdert-Kennedy. 1967. "Perception of the speech code." *Psychological Review* 74 (6): 431–61.

Lisker, L., and A.S. Abramson. 1964. "A cross-language sudy of voicing in initial stops: Acoustical measurements." *Word* 20 (3): 384–422.

Lobanov, B. M. 1971. "Classification of Russian Vowels Spoken by Different Speakers." *The Journal of the Acoustical Society of America* 49 (2B): 606. doi:10.1121/1.1912396.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY, USA: Henry Holt; Co., Inc.

McMurray, Bob, 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): 219–46. doi:10.1037/a0022325.

Munson, Cheyenne M. 2011. "Perceptual learning in speech reveals pathways of processing." PhD thesis, University of Iowa.

Newman, Rochelle S., Sheryl a. Clouse, and Jessica L. Burnham. 2001. "The perceptual consequences of within-talker variability in fricative production." *The Journal of the Acoustical Society of America* 109 (3): 1181–96. doi:10.1121/1.1348009.

Niedzielski, N. 1999. "The Effect of Social Information on the Perception of Sociolinguistic Variables." *Journal of Language and Social Psychology* 18 (1): 62–85. doi:10.1177/0261927X99018001005.

Norris, Dennis, and James M McQueen. 2008. "Shortlist B: A Bayesian model of continuous speech recognition." *Psychological Review* 115 (2): 357–95. doi:10.1037/0033-295X.115.2.357.

Peterson, Gordon E, and Harold L Barney. 1952. "Control Methods Used in a Study of the Vowels." *The Journal of the Acoustical Society of America* 24 (2): 175–84. doi:10.1121/1.1906875.

Pineda, Marisa, and Meghan Sumner. 2010. "Phonetic adaptation in non-native speech: Insights from a distributional analysis of long-lag voice onset time." *The Journal of the Acoustical Society of America* 127 (3): 1853. doi:10.1121/1.3384388.

Pitt, Mark A, Laura C Dilley, Keith Johnson, S Kiesling, W Raymond, E Hume, and E Fosler-Lussier. 2007. "Buckeye Corpus of Conversational Speech (2nd release)." Columbus, OH: Department of Psychlogy, Ohio State University.

Reinisch, Eva, and Lori L Holt. 2014. "Lexically guided phonetic retuning of foreign-accented speech and its generalization." *Journal of Experimental Psychology. Human Perception and Performance* 40 (2): 539–55. doi:10.1037/a0034409.

Remez, Robert E, Jennifer M Fellowes, and Philip E Rubin. 1997. "Talker identification based on phonetic information." *Journal of Experimental Psychology: Human Perception and Performance* 23 (3): 651–66. doi:10.1037/0096-1523.23.3.651.

Strand, Elizabeth A. 1999. "Uncovering the Role of Gender Stereotypes in Speech Perception." *Journal of Language and Social Psychology* 18 (1): 86–100. doi:10.1177/0261927X99018001006.

Strand, Elizabeth A, and Keith Johnson. 1996. "Gradient and Visual Speaker Normalization in the Perception of Fricatives." In *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielfelt*, edited by D Gibbons, 14–26. Berlin: Mouton de Gruyter.

Sumner, Meghan. 2011. "The role of variation in the perception of accented speech." *Cognition* 119 (1). Elsevier B.V.: 131–6. doi:10.1016/j.cognition.2010.10.018.

Thomas, Erik R. 2002. "Sociophonetic Applications of Speech Perception Experiments." *American Speech* 77 (2): 115. doi:10.1215/00031283-77-2-115.

Weatherholtz, Kodi, and T Florian Jaeger. 2016. "Speech perception and generalization across talkers and accents." In *Oxford Research Encyclopedia of Linguistics*.

Zande, Patrick van der, Alexandra Jesse, and Anne Cutler. 2014. "Cross-speaker generalisation in two phoneme-level perceptual adaptation processes." *Journal of Phonetics* 43. Elsevier: 38–46. doi:10.1016/j.wocn.2014.01.003.