Acquisition of covert contrasts: an unsupervised learning approach[*]

James P. Kirby
University of Chicago

1 Introduction

Phonological contrasts are often argued to be neutralized in certain contexts. In many dialects of American English, for example, it has been claimed that the contrast between /t/ and /d/ is neutralized to [ɾ] when followed by an unstressed vowel, leading to homophony between pairs such as *metal–medal* and *cuttle–cuddle* (Giegerich 1992). In Dutch, word-final obstruent devoicing results in homophony between word pairs such as those in Table 1 (Lahiri *et al.* 1987).

voiceless			voiced		
baat	/bat/	'benefit'	*baad*	/bad/	'bathe-1sg'
noot	/not/	'nut'	*nood*	/nod/	'necessity'
voet	/vut/	'foot'	*voed*	/vud/	'feed-1sg'

Table 1: Dutch minimal pairs differing in underlying voicing of final obstruent.

In recent years, a growing body of research has suggested that many contrasts which were thought to be neutralized may in fact be distinguished by subtle yet statistically significant differences in production and perception. In the Dutch case, for instance, Warner *et al.* (2004) have shown that the contrast between word-final /t/ and /d/ is not only distinguished by small differences in the distributions of acoustic cues such as the duration of the stop burst, but that listeners are able to distinguish forms such as those in Table 1 on the basis of other cues which do not differ significantly in production, such as vowel duration and the degree of voicing during closure (see Figure 1).

These types of situations, where impressionistically homophonous categories can be reliably distinguished at the phonetic level, have been referred to collectively as SUSPENDED or COVERT CONTRASTS (Hewlett 1988; Labov *et al.* 1991; Scobbie *et al.* 2000; Yu 2007). Covert contrasts are particularly interesting from the standpoint of language change because they provide a window into sound changes in progress. In fact, it has been suggested that many instances of supposed historical neutralization are in fact suspended contrasts or 'near-mergers' (Labov *et al.* 1991).

[*] I would like to thank the audience at CLS 46 for useful questions and comments as well as Morgan Sonderegger and Karen Livescu for evaluating an earlier version of this work, which led to considerable improvement of both the content and the presentation. Naturally, I alone bear responsibility for any remaining shortcomings or errors.

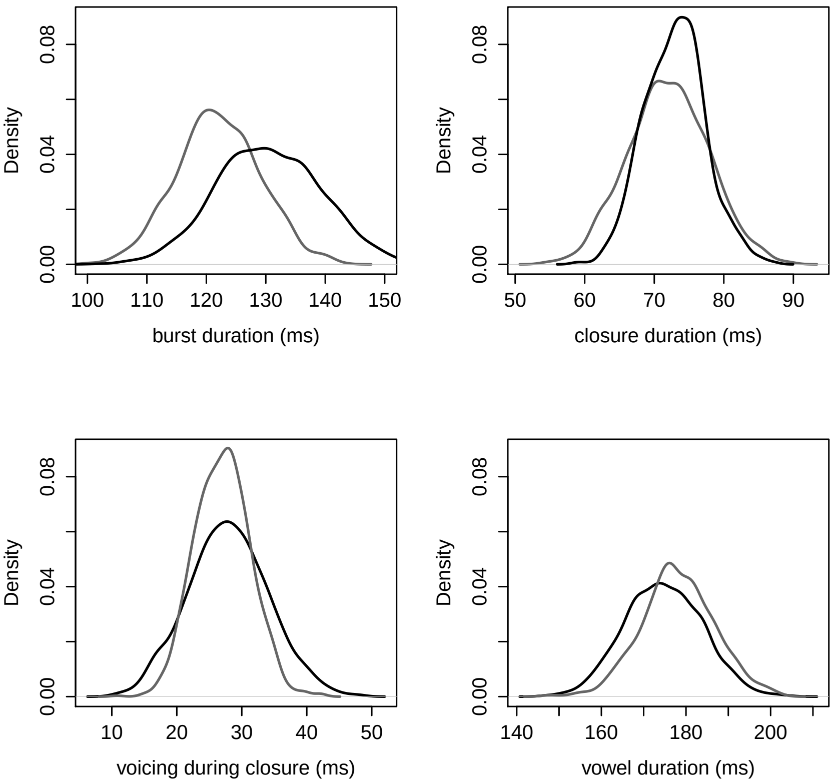

Figure 1: Distribution of 4 acoustic cues to Dutch underlying /t/, /d/ in final position. Black lines give distribution of underlyingly voiceless stops, gray lines underlyingly voiced stops. Parameters estimated based on data in Warner *et al.* (2004).

Covert contrast has also been used to describe a stage of phonological acquisition during which a child's articulations of a contrast, although perceptually inaccessible to adult members of the speech community, are nonetheless instrumentally detectable (Macken and Barton 1980; Hewlett 1988; Scobbie *et al.* 2000). This raises the question as to whether covert contrasts are in fact sound changes in the latter stages of completion, or whether a contrast, however subtly cued, can persist indefinitely within a speech community and across generations of speakers.

One way to address this question is to ask if the categories in a particular instance of covert contrast are *in principle* separable by a statistical learning algorithm. Research indicating that distributional statistics on cues to phonetic category membership are accessible to both adults (Clayards *et al.* 2008) and infants (Maye *et al.* 2002, 2007) has been supported by computational models of phonetic category acquisition, which demonstrate that sound patterns can in fact be induced based on the distributions of acoustic cues (de Boer and Kuhl 2003; Vallabha *et al.* 2007; Feldman *et al.* 2009). However, previous computational studies have only considered contrasts which are well-separated in a low-dimensional acoustic space, such as subsets of the vowel system. Covert contrasts, such as the case of final devoicing in Dutch described above, might be expected to present greater difficulty for statistical learning mechanisms (including those potentially used by humans) on account of the high degree of overlap along multiple cue dimensions. As a result, covert contrasts may be more likely to neutralize in the course of acquisition.

This paper explores the learnability of covert contrasts through a series of statistical learning simulations based on the Dutch contrast discussed above. The results indicate that while a statistical learner can be quite effective at inducing extremely subtle distinctions, providing an existence proof that covert contrasts can persist in the acquisition of phonetic categories, the learner's success depends a great deal on the number and distributional characteristics of the relevant cue dimensions.

2 Mixture models and model-based clustering

The issue of how to best account for the mapping from a continuous, highly variable acoustic speech signal to a discrete set of phonologically equivalent linguistic categories has led an increasing number of researchers to explore approaches to speech perception grounded in pattern recognition (Oden and Massaro 1978; Nearey 1997; Smits *et al.* 2006; Clayards *et al.* 2008; Holt and Lotto 2010; Toscano and McMurray 2010). The problem of determining the number of categories which generated a set of acoustic observations can be intuitively recast as a clustering problem: determining the intrinsic structure of a set of data without prior knowledge of that structure. Clustering is a type of unsupervised learning, since neither the number of clusters (or components) which generated the observations nor the parameters of those components are known in advance.

Model-based clustering is a powerful unsupervised learning technique which allows for information or assumptions about the underlying distribution of the observation data to be modeled directly, usually in the form of a FINITE MIXTURE MODEL (McLachland and Peel 2000). In a finite mixture model, each cluster is assumed to have its own weight and probability distribution. Since acoustic-phonetic cues are often roughly normally distributed, a GAUSSIAN MIXTURE MODEL (GMM) may be appropriate for modeling speech categories. Formally, a GMM is a weighted sum of K component densities

$$f(\mathbf{x};\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

where $\mathbf{x}$ is a D-dimensional feature vector, π_k is the k^{th} component weight, and $\theta = (\theta_1, \ldots, \theta_K) = ((\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K))$ is a $K(D+2)$-parameter structure containing the component weights π_k, mean vectors $\boldsymbol{\mu}_k$, and covariance matrices $\boldsymbol{\Sigma}_k$ of the D-dimensional Gaussian densities

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \tag{2}$$

The component weights $\pi_1, \ldots, \pi_K$ must sum to 1.

Fitting a K-component GMM involves finding θ, usually via the method of maximum likelihood (ML) estimation. Given a series of N observation vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$, ML finds θ that maximizes the log likelihood

$$\log p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

Since this cannot be solved in closed form, iterative techniques such as the EM (expectation maximization) algorithm (Dempster *et al.* 1977) are often employed.

However, there remains an additional problem: how to determine the best value of K when it is not known in advance. ML estimation is of little use here, as maximum likelihood is ultimately achieved by associating each observation with its own Gaussian. One means of avoiding this kind of overfitting is to pick the simplest model consistent with the data, where 'simplest' is defined with respect to the number of parameters in the model. This trade-off between model fit and model complexity can be measured by the BAYESIAN INFORMATION CRITERION (BIC), an approximation of the Bayes factor which penalizes models based on the number of free parameters they contain (Schwarz 1978; Fraley and Raftery 2007). The larger the value of the BIC, the stronger the evidence for the model.

BIC-based model selection proceeds as follows. Given a series of N independent, identically distributed D-dimensional observations $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$, first find ML parameter estimates θ for a series of GMMs with K in the range $1, \ldots, K_{max}$ by maximizing Equation (3). Now let $L(D, K)$ be the maximum log-likelihood of a GMM with K D-dimensional components and $Q(D, K)$ independent parameters:

$$\begin{aligned} L(D,K) &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (4) \\ Q(D,K) &= K(D + D(D+1)/2) + K - 1 && (5) \end{aligned}$$

The BIC-optimal model is the one which maximizes the quantity

$$BIC(D, K) \equiv 2 \cdot L(D, K) - Q(D, K) \log N \tag{6}$$

A larger BIC implies fewer model parameters, better fit, or both.

3 Case study: Dutch final devoicing

Warner *et al.* (2004) compared the durations of preceding vowel, stop closure, voicing during closure, and burst of Dutch stops in neutralizing and non-neutralizing environments. They found that vowel duration was a significant predictor of voicing in both environments, while burst duration was significant in the neutralizing environment only in cases where the preceding vowel was long. Neither duration of voicing during the closure nor of the closure itself emerged as significant predictors of underlying voicing; however, listeners were able to use these cues to discriminate between categories in a two-alternative forced choice perception experiment when all other predictors were held constant. These results suggest that while cues which covary strongly with an underlying contrast in production will be important for a learner attempting to perceptually recover that contrast, other cues may also play a role. To explore this possibility, I conducted two sets of simulation experiments to determine if (1) a computational learner could recover an underlying contrast when given observation data containing only those cues which covary in production with the underlying category contrast (here, voicing) and (2) if having access to additional cue dimensions would help or hinder acquisition of the contrast.

3.1 Series 1

The first set of simulations varied the number and type of cue dimensions made available to the learner to determine the predictions of model-based clustering using the BIC about an individual learner's success at acquiring a covert contrast.

3.1.1 Methods

Since model-based clustering can only usefully compare models fit to a single set of observations, the experiments in Series 1 fit a series of GMMs to a set of $N = 500$ 4-dimensional data vectors generated from a Gaussian mixture with equally weighted components using the parameters given in Table 2 (the long vowel neutralization environment data of Warner *et al.* 2004).[1] Models were fit using the EM-based estimator implemented in the R package `mclust` (Fraley and Raftery 2006). The first set of experiments compared two-dimensional models using only the burst and vowel duration information (the two cues that Warner *et al.* found to reliably differentiate phonological voicing in both production and perception). The second set of experiments added a third dimension (duration of closure voicing or duration of the closure itself), while the third set of experiments included all four dimensions. For each value of D, the BIC score of models with up to 5 components were compared. The model-fitting procedure (described in §2) was identical for all experiments; only the number of cue vector dimensions available to the learner changed.

[1] Since Warner *et al.* did not report the standard deviations of the cues they studied, standard deviations were estimated based on distributions of the same cues to word-final stops in the author's own corpus of American English production data as well as German data from Jessen (1998).

Measurement	Underlying voicing	Mean	s.d.
Vowel duration (VDUR)	Voiceless	175	9
	Voiced	178	9
Closure duration (CDUR)	Voiceless	73	4
	Voiced	72	6
Burst duration (BURST)	Voiceless	131	10
	Voiced	122	7
Closure voicing duration (VGCL)	Voiceless	28	6
	Voiced	27	4

Table 2: Initial parameter settings used to generate observation data.

3.1.2 Results

An overview of the results for a single set of simulation runs is given in Table 3. Where models with larger values of K failed to result in a reduction in BIC and/or error rate, results have been omitted for clarity. For comparison, the Bayes error rate – the theoretical minimum error rate of any classifer – was 0.2. For observation data in 2 dimensions, the BIC-optimal model contained just a single component; for data in 3 and 4 dimensions, the BIC-optimal models contained 2 components.

	BURST, VDUR		+VGCL		+CDUR		+VGCL, CDUR	
K	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*	*BIC*	*err*
1	**-14695**	**0.5**	-20784	0.5	-20858	0.5	-26782	0.5
2	-14702	0.45	**-20739**	**0.45**	**-20813**	**0.4**	**-26762**	**0.27**
3			-20779	0.36	-20854	0.33	-26799	0.47
4							-26804	0.32

Table 3: BIC scores and error rates for models of 2, 3, and 4 dimensions. K = number of categories (components); columns show the cue dimensions made available in the observation data. Bold items indicate the BIC-optimal solutions.

Figures 2–5 show the results of model fitting for several configurations of interest. The *solution* plots (left columns) show the observation data, with underlyingly voiced stops shown as solid gray and underlyingly voiceless stops as open black circles, along with the 90% confidence ellipses for the estimated Gaussians, which provide a rough visual indication of the fitted clusters as well as the correlation between the cue dimensions. The *accuracy* plots (right columns) show the correctly (black $+$) versus incorrectly (gray $\times$) categorized data points based on the fitted model. For models with more than two categories (i.e. where $K > 2$), the error rates reported are the best possible interpretations of the model predictions.

Figure 2 shows both $K = 1$ (panels A and C) and $K = 2$ (panels B and D) so-

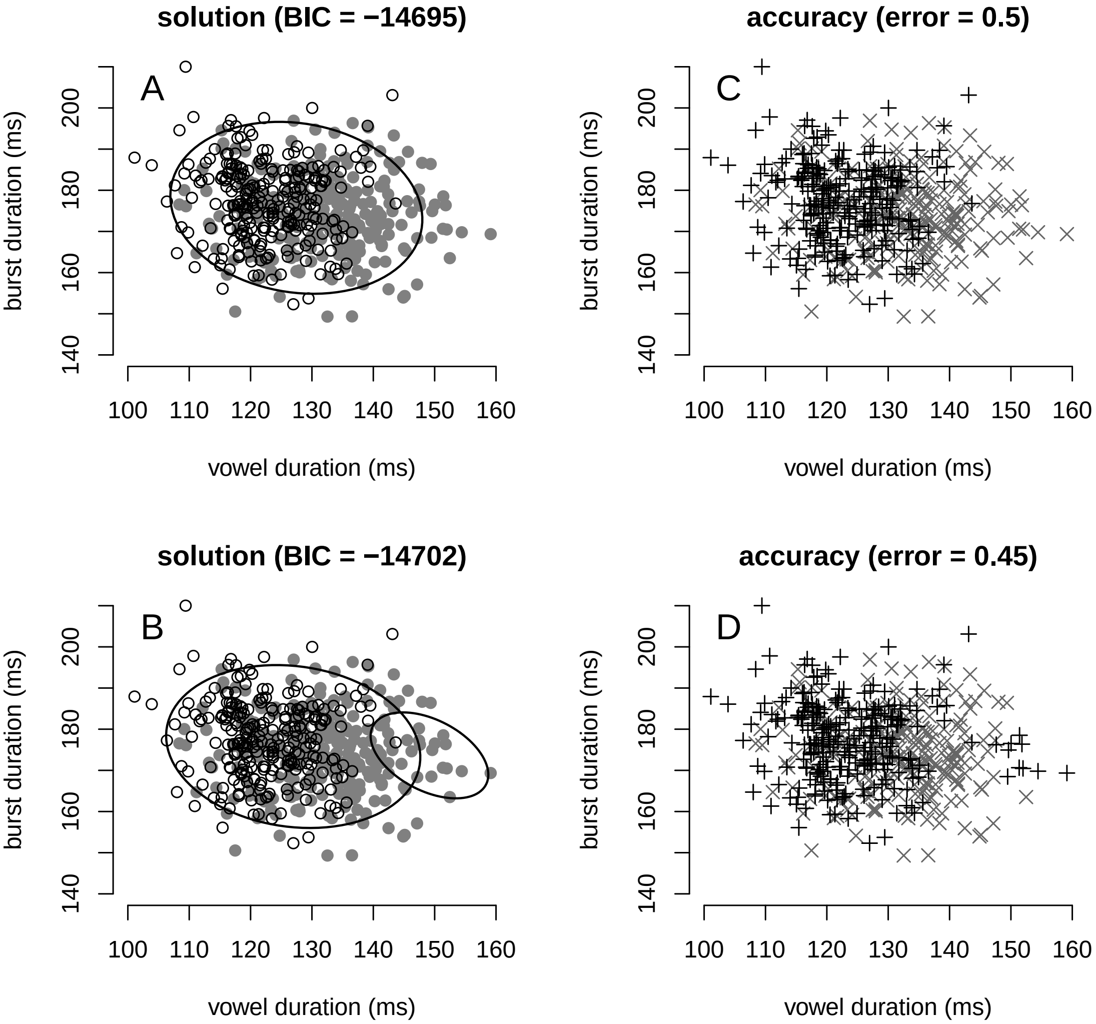

Figure 2: A–B: Distributions of voiceless (black) and voiced (gray) tokens for training data in 2 dimensions, with ellipses corresponding to the 90% confidence intervals of the estimated Gaussians for the $K = 1$ (A) and $K = 2$ (B) solutions. C–D: Classification accuracy based on the $K = 1$ (C) and $K = 2$ (D) solutions.

lutions for the $D = 2$ space. Although classification based on the $K = 1$ solution is at chance, this solution is preferred to the $K = 2$ solution, whose slightly increased accuracy fails to be justified by the increased complexity on the BIC metric. Increasing K failed to reduce error below 0.45.

Figures 3 and 4 show 2-dimensional projections of the $K = 2$ and $K = 3$ solutions, respectively, fit to observation data in 3 dimensions (columns 2 and 3 of Table 3). Here, although the addition of a third component resulted in a more significant

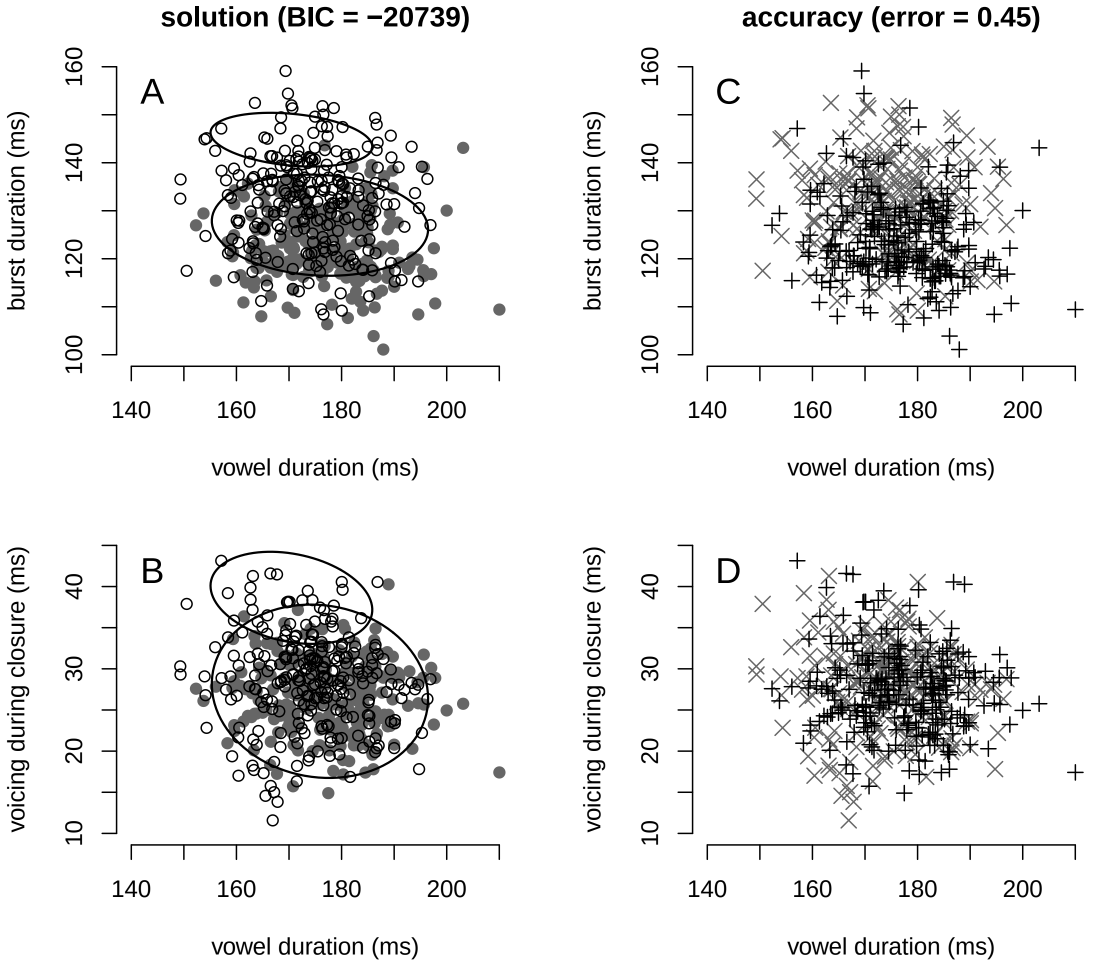

Figure 3: A-B: Underlying distributions of voiceless (black) and voiced (gray) tokens plus 90% confidence intervals for observation data in 3 dimensions. C-D: Classification accuracy based on the $K = 2$ solution.

increase in accuracy compared to the $D = 2$ case, it again failed to justify the increase in model complexity. The results were comparable regardless of whether the third cue was CDUR or VGCL. Note that neither solution proposes categories which appear to be motivated by the underlying voiced/voiceless structure of the data, with the $K = 2$ solution performing at barely above chance levels.

Figure 5 shows the 2-dimensional projections for a $K = 2$ solution fit to observation data in 4 dimensions (column 4 of Table 3). The $K = 2$ solution achieves near-optimal accuracy and is also selected as the optimal model on the basis of the BIC; models with higher K had reduced accuracy in addition to increased complex-

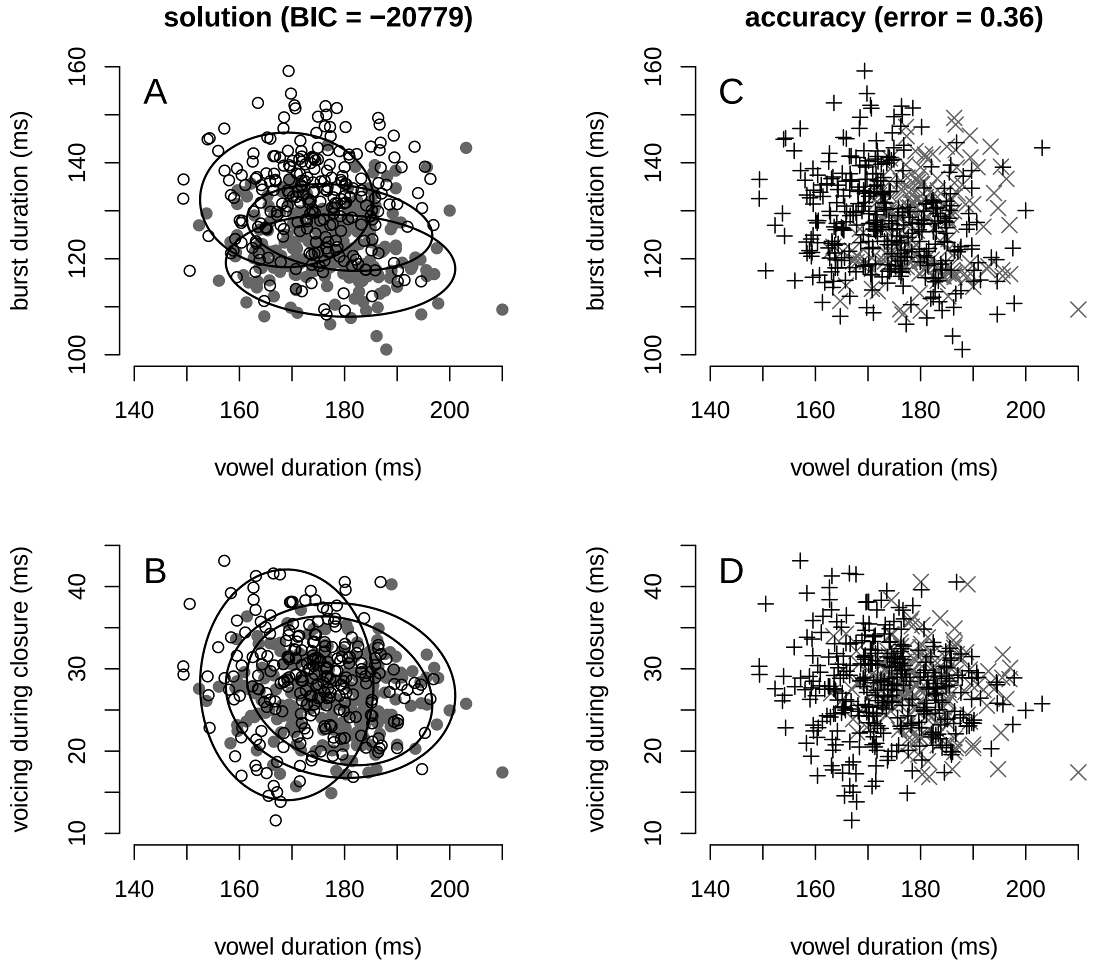

Figure 4: A-B: Underlying distributions of voiceless (black) and voiced (gray) tokens plus 90% confidence intervals for observation data in 3 dimensions. C-D: Classification accuracy based on the $K = 3$ solution.

ity (these plots are omitted for reasons of space). Although a learner with access to 4 cue dimensions arrives at the same number of optimal components as the learner with access to only 3, the error rate of the resulting classifier is far lower and the resulting categories are much better aligned with the underlying category structure.

3.2 Series 2

The experiments in Series 1 fit a variety of models to a single set of observations in order to assess the general utility of the BIC in modeling the acquisition of a covert contrast. While the results suggest that the BIC may be a useful metric in assessing

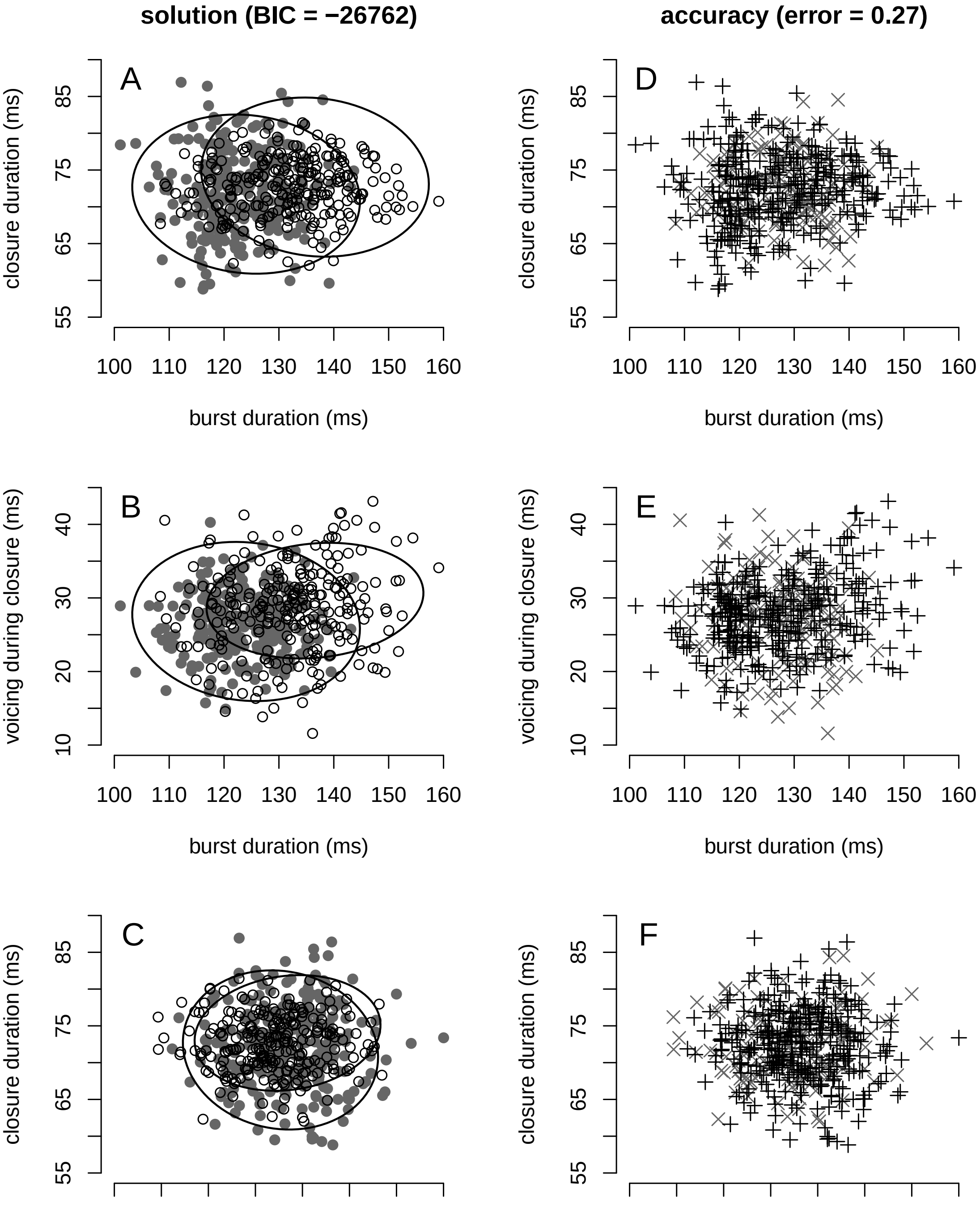

Figure 5: A-C: Underlying distributions of voiceless (black) and voiced (gray) tokens plus 90% confidence intervals for observation data in 4 dimensions. D-F: Classification accuracy based on the $K = 2$ solution.

the learnability of covert contrast, models fit to any single set of observations may not be representative of the larger range of potential outcomes. Therefore, a second set of experiments were conducted in which models were also fit to a range of observations generated from a mixture model with fixed parameters, in order to gain a better sense of the typicality of the Series 1 solutions.

3.2.1 Methods

The general model fitting procedure was the same as in Series 1, except that each model parameterization was fit to 1,000 different series of observation vectors drawn from a GMM with parameters shown in Table 2. The BIC-optimal number of model components was recorded for each fit.

3.2.2 Results

The results of the Series 2 experiments are given in Table 4, which shows the largest proportion of BIC-optimal model components for observation data of a given dimensionality in bold. The d_x columns indicate the cue dimensions included in the observed data. For most (but not all) 2-dimensional observation data, the largest proportion of optimal models included just a single component; for 3- and 4-dimensional observation data, $K = 1$ and $K = 2$ solutions were optimal with roughly equal frequency.

d_1	d_2	d_3	d_4	$K = 1$	$K = 2$	$K = 3$
BURST	VDUR			**0.61**	0.39	
BURST	VGCL			0.39	**0.61**	
BURST	CDUR			0.41	**0.58**	0.01
VDUR	VGCL			**0.96**	0.04	0.01
VDUR	CDUR			**0.99**	0.01	
VGCL	CDUR			**0.95**	0.04	0.01
BURST	VDUR	VGCL		**0.51**	0.49	
BURST	VDUR	CDUR		0.46	**0.53**	0.01
BURST	VGCL	CDUR		0.31	**0.68**	0.01
VDUR	VGCL	CDUR		**0.58**	0.41	0.01
BURST	VDUR	VGCL	CDUR	0.42	**0.58**	

Table 4: Proportion of BIC-optimal category solutions for the Dutch experiments in terms of percentage of 1,000 fits. d_x columns indicate the cue dimensions included in the observed data. Bold entries show the greatest proportion of possible solutions for models with given dimensionality.

4 Discussion

The results of the experiments in Series 1 indicate that while a statistical learner can successfully acquire a covert contrast (in this case, between two underlying voicing categories), success depends not only on access to cue dimensions which reliably covary with an underlying category specification, but also those that do not. For example, despite marginally better accuracy, the $K = 2$ solution was dispreferred on the BIC metric when only the BURST and VDUR cue dimensions were available; the improved accuracy did not justify the increase in model complexity over a model with a single category. As more cue dimensions were made available to the learner, a convergence between models with the highest accuracy and models with the BIC-optimal number of components was observed.

However, as illustrated by the replication experiments in Series 2, small differences in the observation data to which a statistical learner is exposed can have a considerable impact on the number of the components in the optimal solution. While the $K = 1$ solution was generally preferred for observation data which included only BURST and VDUR information, the $K = 2$ solution was more likely to be optimal for other types of 2-dimensional observation data, such as that containing only BURST and VGCL information. When BURST information (the dimension which covaried most robustly with the underlying voicing specification) was unavailable, a model with a single component was nearly always optimal.

As the number of cue dimensions made available to the learner increased, interpretation of the results became even less straightforward. The factors affecting the relative likelihood of selecting a model with one or two components for 3-dimensional data are not immediately obvious; the relative likelihood of the optimal solution did not appear to vary with the presence or absence of any single cue dimension. In experiments with 4-dimensional observation data, a model with two components was selected as optimal just 58% of the time. Thus, while access to multiple cues *may* assist learners in recovering a covert contrast, there is no guarantee that this will be the case: the higher accuracy afforded models with access to more cue dimensions is only justified if the increase in model complexity is not too great, given the increase in likelihood. Empirically speaking, this is an extremely desirable property of the model, as it reflects the considerable variation in the data to which human listeners are exposed in the course of language acquisition, as well as variation in whether, and to what extent, individual members of a speech community show near-merger or sensitivity to covert contrast (Labov *et al.* 1991).

The results presented here may also be considered in light of Warner *et al.*'s findings that burst duration and vowel duration were the only two cues to reliably covary with underlying voicing contrast in production. The present results suggest the possibility that the ability of a learner to recover the contrast may depend on attending to a potentially larger set of cues, beyond those which reliably covary in isolation with an underlying contrast. Indeed, there is some experimental evidence that this is the case. When Warner *et al.* examined listener sensitivity to closure duration in a forced-choice identification task, they found a significant effect of continuum step. This suggests that even when a cue does not vary systematically with an underlying category in production, it may nonetheless play a role in distinguishing between categories.

From the statistical learning standpoint, this raises the question of how cues which, on their own, appear statistically inseparable could possibly improve on the performance of a classifier built from more robustly separable data vectors. While the specifics of cue integration in human speech sound categorization are a matter of ongoing investigation (Clayards 2008; Toscano and McMurray 2010), statistically speaking, even dimensions along which categories are at best slightly separable can improve classification performance when considered together, because categories which are not well-separated when projected down to any single dimension may still be well-separated in a higher-dimensional space.[2]

The increased likelihood of an optimal $K = 2$ solution when using certain subsets of the full set of cue dimensions suggests the possibility that some experimental findings showing that human listeners cannot discriminate between supposedly neutralized categories at above chance levels may be misleading, in that their design may not allow for a positive outcome. If categories are only recoverable when learners have access to a wide range of acoustic cues, then failure of participants to discriminate categories in a traditional two-alternative forced choice paradigm, where one acoustic dimension is varied while all others are held constant, should not be taken as evidence that the categories are in principle indiscriminable, or even that the acoustic dimension tested plays no role in category discrimination. Accordingly, future laboratory exploration of covert contrast may need to consider alternative experimental approaches to more accurately assess the role of complexity in human categorization and category learning (Plauché 2001; Pothos and Chater 2002; Goudbeek *et al.* 2008; Pothos and Close 2008).

A related issue concerns the cognitive status of the BIC and other likelihood-based clustering methods. One objection to these methods (such as those based on the Minimum Description Length principle) is that such techniques operate in 'batch mode', with computations referencing the sum total of experienced tokens each time an observation is assigned a category label. While potentially unrealistic from an neural-implementational standpoint, this class of models can nonetheless shed light on the general functional-computational issues of category learning and inductive inference (Marr 1982).

[2] Note that predictions made about the category structure based on metrics like the BIC will depend not only on the number of cue dimensions made available, but also on the degree to which they reliably separate underlying categories; the present results could vary significantly if other distributional parameters, such as the cue variance or distance between the means, were systematically manipulated (Kirby 2010).

5 Conclusion

The results of model-based clustering indicate that an unsupervised statistical learner is in principle capable of recovering covert contrasts, with a success rate dependent on the type and number of cues provided. Statistical learners with access only to cues which covaried reliably with an underlying contrast in production were unable to learn the correct distribution, while access to additional cues facilitated category learning. This suggests both that (i) covert contrast could be successfully transmitted and acquired as such by human learners and (ii) covert contrast may be a stable state unto itself, rather than just a temporary phase in the loss or acquisition of a contrast. These results demonstrate how taking a pattern recognition approach to phonetic categorization allows for a more nuanced understanding of the factors which contribute to acquisition and transmission of phonetic categories as well as the conditions under which the number of functional categories may change.

References

de Boer, Bart, and Patricia Kuhl. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters On-line* 4.129–134.

Clayards, Meghan, 2008. *The ideal listener: making optimal use of acoustic-phonetic cues for word recognition*. University of Rochester dissertation.

——, Michael K. Tanenhaus, Richard Aslin, and Robert A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108.804–809.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39.1–38.

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*.

Fraley, Chris, and Adrian E. Raftery. 2006. MCLUST Version 3 for R: Normal mixture modeling and model-based clustering. Technical report 504, Department of Statistics, University of Washington.

——, and ——. 2007. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* 24.155–181.

Giegerich, Heinz J. 1992. *English Phonology*. Cambridge: Cambridge University Press.

Goudbeek, M., A. Cutler, and R. Smits. 2008. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication* 50.109–125.

Hewlett, Nigel. 1988. Acoustic properties of /k/ and /t/ in normal and phonologically disorderd speech. *Clinical Linguistics and Phonetics* 2.29–45.

Holt, Lori L., and Andrew J. Lotto. 2010. Speech perception as categorization. *Attention, Perception, and Psychophysics* 72.1218–1227.

Jessen, Michael. 1998. *The phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins.

Kirby, James, 2010. *Cue selection and category restructuring in sound change*. University of Chicago dissertation.

Labov, William, Mark Karen, and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3.33–74.

Lahiri, Aditi, Herbert Schriefers, and Cecile Kuijpers. 1987. Contextual neutralization of vowel length: Evidence from Dutch. *Phonetica* 44.91–102.

Macken, Marlys A., and David Barton. 1980. The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language* 7.41–74.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

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

——, Daniel J. Weiss, and Richard N. Aslin. 2007. Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science* 11.122–134.

McLachland, Geoffry J., and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Nearey, Terrence. 1997. Speech perception as pattern recognition. *Journal of the Acoustical Society of America* 101.3241–3254.

Oden, Gregg C., and Dominic W. Massaro. 1978. Integration of featural information in speech perception. *Psychological Review* 85.172–191.

Plauché, Madelaine, 2001. *Acoustic cues in the directionality of stop consonant confusions*. University of California, Berkeley dissertation.

Pothos, Emmanuel M., and Nick Chater. 2002. A simplicity principle in unsupervised human categorization. *Cognitive Science* 26.393–343.

——, and James Close. 2008. One or two dimensions in spontaneous classification: A simplicity approach. *Cognition* 107.581–602.

Schwarz, Gideon E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6.461–464.

Scobbie, James M., Fiona Gibbon, William J. Hardcastle, and Paul Fletcher. 2000. Covert contrast as a stage in the acquisition of phonetics and phonology. In *Papers in Laboratory Phonology V: Language Acquisition and the Lexicon*, ed. by Michael Broe and Janet Pierrehumbert, 194–207. Cambridge: Cambridge University Press.

Smits, Roel, Joan Sereno, and Allard Jongman. 2006. Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance* 32.733–754.

Toscano, Joseph C., and Bob McMurray. 2010. Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science* 34.434–464.

Vallabha, Gautam K., James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104.13273–13278.

Warner, Natasha, Allard Jongman, Joan Sereno, and Rachèl Kemps. 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32.251–276.

Yu, Alan C. L. 2007. Understanding near mergers: the case of morphological tone in Cantonese. *Phonology* 24.187–214.