# MODELING THE ACQUISITION OF COVERT CONTRAST

*James P. Kirby*

University of Edinburgh, UK
j.kirby@ed.ac.uk

## ABSTRACT

This paper explores the learnability of covert contrasts (impressionistically homophonous categories that can be reliably distinguished at the phonetic level) through a series of model-based clustering simulations using human production data. Allowing the models to learn both the number and parameters of those categories provides a way to explore the potential stability of category structures. The results indicate that while a statistical learner can be quite effective at inducing covert contrasts, success depends crucially on the number and distributional characteristics of the relevant cue dimensions.

**Keywords:** near merger, Dutch, incomplete neutralization, mixture models, unsupervised learning

## 1. INTRODUCTION

Phonological contrasts are often argued to be neutralized in certain contexts, but a growing body of research suggests that many cases of apparent neutralization may in fact be reliably distinguished at the phonetic level [9, 11, 12]. Together with studies demonstrating variability in the perception and production of near-merger at the population level [5], the question arises of whether such COVERT CONTRASTS represent an essentially transitory stage in language evolution, or whether even subtly cued contrasts could persist indefinitely within and/or 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. Previous computational studies of phonetic category acquisition [1, 2, 10] have all focused on vowel contrasts, which tend to be well-separated in a low-dimensional acoustic space. Covert contrasts might be expected to present greater difficulty for statistical learning mechanisms on account of the high degree of overlap along multiple cue dimensions, suggesting that may be more likely to neutralize in the course of acquisition or transmission.

Reports of contrast reduction often remark on its uneven distribution in a population, with some members of a speech community distinguishing a contrast in production and/or perception, while others do not [4, 5]. Through a series of statistical learning simulations, the present study explores the role of this individual variation in cue perception and production on the acquisition of covert contrast.

## 2. CATEGORY RESTRUCTURING AS MODEL SELECTION

The task of phonetic category induction may be intuitively recast as a (model-based) clustering problem: determining the intrinsic structure of a set of data without prior knowledge of that structure. Here, clustering was performed using a Gaussian mixture model (GMM) [7]. In a GMM, a *D*-dimensional data point x = ($x_1$, . . . , $x_D$ ) is assumed to be generated by a *K* -component mixture model with density

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

where $\theta = ((\pi_1 , \mu_1, \Sigma_1), . . . , (\pi_K , \mu_K , \Sigma_K ))$ is a *K* (*D* + 2)-parameter structure containing the component weights $\pi_k$, mean vectors $\mu_k$, and covariance matrices $\Sigma_k$ of the *D*-dimensional Gaussian densities *N*. The component weights $\pi$ must obey the constraint that $\sum_1^K \pi_k = 1$.

Fitting a *K* -component GMM involves finding $\theta$, usually via the method of maximum likelihood estimation: given an observation vector X = $x_1$, $x_2$ , . . ., $x_N$ , find $\theta$ that maximizes the log-likelihood $L$ = ln P (X|$\theta_{\max}$ ). However, there remains the problem of determining an optimal value of *K*. One strategy 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 in a number of ways; here, we employ the BAYESIAN INFORMATION CRITERION (BIC), a metric which penalizes models based on the number of free parameters they contain [8]. BIC-based model selection proceeds as follows. Given a series of *N* i.i.d. *D*-dimensional observations **X**, let *L* be the maximized log-likelihood of a GMM with *K* *D*-dimensional components characterized by

parameters $\theta$, and let $Q$ stand for the number of independent parameters in that model:

(2) $Q = K\,(D + D(D + 1)/2) + K - 1$

The BIC is then defined as

(3) $BIC = -2L + ln(N)Q$

The first term, $-2L$, measures the model's accuracy (fit to the data), while the second term, $ln(N)Q$, represents the model's complexity. Given two models fit on the same data, the model with the smaller BIC value is considered superior in terms of the fit-complexity trade-off. Here, neutralization is predicted when the BIC-OPTIMAL model fit to data generated from a model with $K$ components has fewer than K components.

## 3. COVERT CONTRAST IN DUTCH

The voicing contrast between word-final obstruents in Dutch is a case that has been argued to be both complete [6] and incomplete neutralization [11]. Compelling evidence in favor of incomplete neutralization is provided by Warner, et al. [11], who show that a distinction between word-final /t/ and /d/ is not only supported by small but statistically significant differences in the production of cues such as the duration of the stop burst, but that listeners are also able to distinguish forms such as those in Table 1 on the basis of cues which appear to overlap significantly in production, such as the degree of voicing during closure.

**Table 1:** Dutch minimal pairs differing in under- lying voicing of the final obstruent, from [11].

| voiceless | | voiced | |
|---|---|---|---|
| baat /bat/ | 'benefit' | baad /bad/ | 'bathe-1sg' |
| boot /bot/ | 'nut' | bood /bod/ | 'offered-1sg' |
| eet /et/ | 'eat-sg.' | eed /ed/ | 'oath' |
| meet /met/ | 'measure-sg' | meed /med/ | 'avoided- |
| 1sg' noot /not/ | 'nut' | nood /nod/ | 'necessity' |
| smeet /smet/ | 'threw-sg' | smeed /smed/ | 'forge-1sg' |
| zweet /zwet/ | 'nut' | Zweed /zwed/ | 'Swede' |

The original production data set gathered by Warner et al. contains 2,160 tokens gathered from 15 native speakers: two repetitions each of 72 Dutch lexical items, forming 36 minimal pairs, containing either phonologically long or short vowels. The simulations reported here consider just the seven minimal pairs shown in Table 1 (420 tokens), the subset of the data containing phonetically short vowels.

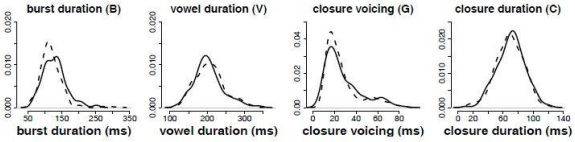
### 3.1. Series 1: Pooled data

Warner et al. [11] measured the duration of four potential cues to underlying final stop voicing in Dutch: the duration of the release burst (B), the duration of the preceding vowel (V), the duration of the closure (C), and the duration of the voiced period of the closure (G). The means and standard deviations for these cues based on the productions of all 15 speakers are given in Table 2, and plotted in Fig. 1.

**Table 2:** Mean (s.d.) for cues to Dutch final /t/ and /d/ for all speakers in [11].

| Category | B | V | G | C |
|---|---|---|---|---|
| voiced | 118 (36) | 207 (39) | 27 (16) | 69 (18) |
| voiceless | 130 (34) | 208 (40) | 28 (17) | 71 (19) |

**Figure 1:** Density plots of cues to underlying final /t/ (solid) and /d/ (dashed) for all speakers in [11].




Two general types of model-fitting simulations were performed using these data, to assess (a) whether neutralization would be predicted (based on the BIC-optimal number of mixture components) and (b) the robustness of the solution to random variation in the training data. First, to explore individual variation in attention to cue, a series of 75 GMMs were fit to a set of $N$ =500 observation vectors generated from a 2-component, 4-dimensional Gaussian mixture with the parameters given in Table 2, representing all non-empty subsets of the set of 4 cues (e.g. {B, V, C, G, BV, BC, VC...}) crossed with 1 to 5 components. A single-component model was BIC-optimal in all cases, despite marginal improvements in classification accuracy for models with 3 or more components.

In order to insure that the simulation results did not simply reflect an idiosyncratic statistical property of the particular observation data to which the models were fit, each simulation was repeated 1,000 times with different sets of input data generated from the parameters in Table 2. The results of these typicality experiments confirm the initial findings: the optimal solution was a single-component model in all cases except for the cue combination BGC, where $K = 1$ models were selected 90% of the time and $K = 2$ models just 10%.

### 3.2. Series 2: Individual data

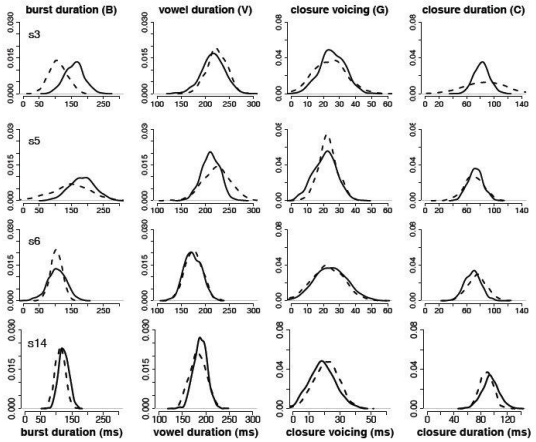
The results of the simulations reported above predict complete neutralization of the Dutch voicing contrast in final position, contra Warner et al., who found that listeners were able to distinguish between voicing categories with greater-than-chance accuracy based on differences between cues which did not covary with voicing in

production. However, the authors also noted a certain amount of between-speaker variability in their data; by pooling the production data for all speakers, this individual variation is obscured. To explore the effect of individual variability in production on the recoverability of this contrast, simulations were repeated using data from four speakers representative of the range of variation seen in the Warner et al. data set (subjects 3, 5, 6, and 14), shown in Table 3 and Fig. 2.

**Table 3:** Mean (s.d.) for cues to Dutch final /t/ and /d/ for individual speakers in [11].

| | | $s_3$ | $s_5$ | $s_6$ | $s_{14}$ |
|---|---|---|---|---|---|
| B | vcd | 159 (30) | 183 (39) | 105 (29) | 124 (16) |
| | vcls | 109 (29) | 149 (60) | 104 (18) | 113 (16) |
| V | vcd | 216 (24) | 211 (19) | 174 (13) | 118 (14) |
| | vcls | 224 (21) | 223 (28) | 176 (18) | 183 (18) |
| G | vcd | 26 (8) | 21 (7) | 25 (10) | 19 (8) |
| | vcls | 24 (10) | 23 (6) | 23 (10) | 21 (8) |
| C | vcd | 82 (11) | 73 (11) | 69 (12) | 94 (13) |
| | vcls | 86 (30) | 70 (14) | 75 (14) | 88 (10) |

**Figure 2:** Density plots of cues to underlying final /t/ (solid) and /d/ (dashed) for individual speakers.




First, BIC scores for all non-empty subsets of cue dimensions were computed for 4 sets of 75 models, each set trained on data generated from one of the subjects in Table 3. Fig. 3 shows the optimal number of cue dimensions (1=light grey, 2=dark grey) computed for these 300 GMMs (models with more than 2 components were never BIC-optimal). In some cases, learners were able to recover the underlying contrasts, but the outcome varied both with the source of training data as well as with the dimensionality of the input. For instance, while a 2-component GMM was BIC-optimal when fit to data from $s_6$ containing just vowel duration (V) information, adding additional cue dimensions resulted in 1-component models emerging as optimal, while the reverse was generally true for models fit to data from subject $s_3$.

As in Series 1, each of the clustering simulations based on individual speaker data were repeated 1,000 times. The typicality of the results shown in Fig. 3 is summarized in Fig. 4, which shows the proportion (out of 1,000) of 1, 2, and (in a very few cases) 3-category solutions learned. These simulations highlight the potentially stochastic nature of the category restructuring process, demonstrating that even small fluctuations in the input may cause the number of categories posited by one learner to differ from the number posited by another learner, even when exposed to nominally the same data. This illustrates how the fate of a covert contrast may depend heavily on the particulars of the input to which learners are exposed, in addition to individual differences in the saliency or integrity of acoustic cues to a contrast.

**Figure 3:** Number of BIC-optimal categories (1=light grey, 2=dark grey) for GMMs fit to individual subject data by input dimensionality.

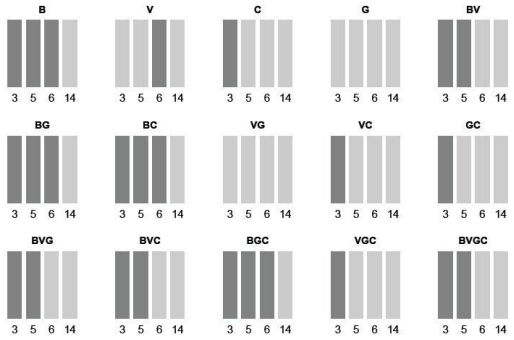
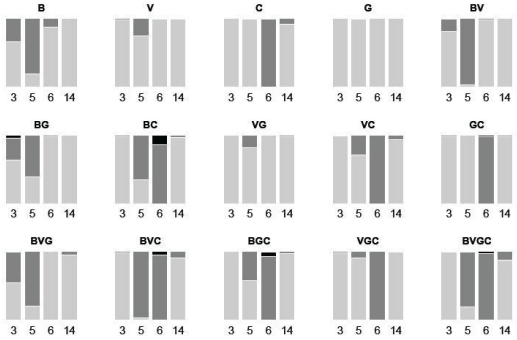



**Figure 4:** Proportion of 1 (light grey), 2 (dark grey), and 3 (black) BIC-optimal category solutions for input data from individual Dutch speakers by input dimensionality.




Furthermore, the typicality simulations illustrate the danger in assessing the potential separability of a contrast based on a single acoustic dimension, even if highly salient. While 1-component solutions were generally preferred for observation data which included only B and V cues, 2-component solutions were more likely to be optimal for other types of 2-dimensional observation data, such as that

containing B or G cues. When burst (B) information (which covaried most robustly with the underlying voicing specification for all 4 speakers) was surpressed, 1- component models were nearly always optimal, but the mere presence of the burst cue alone was not al- ways sufficient for a learner to recover the contrast.

## 4. DISCUSSION

One key finding of this study is that access to a cue dimension that, when considered on its own, allows a learner to posit a *K* -category solution as optimal does not necessarily imply that a *K* -category solution will be optimal when this dimension is considered simultaneously with other dimensions. This is consistent with [11], who found a significant effect of continuum step when examining listener sensitivity to closure duration (C), a cue which did not significantly covary with underlying voicing in production. One interpretation of these results is that even though an individual acoustic dimension may not vary systematically with an underlying category distinction in production, human listeners may nonetheless possess some awareness of the role it plays in distinguishing between categories.

The finding that GMMs were sometimes able to recover the contrast when fit to subsets of the data predicts that the likelihood of recovering a covert contrast may vary stochastically in a population. That is, whether a learner posits e.g. 2 categories rather than 1 cannot be predicted solely on the basis of considering pooled data from a population, since learners are exposed to different subsets of the population and/or may attend to or integrate the range of potential cue dimensions in different ways. This is consistent with the findings that some members of a speech community show covert contrast/near mergers in production, perception, or both, while others neither produce nor perceive such contrasts [5].

## 5. CONCLUSIONS

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. This suggests that covert contrasts may represent potential stable states, rather than just temporary phases in the evolution of a contrast. However, the results also demonstrate that an underlying contrast cannot necessarily be inferred from separability along individual acoustic-phonetic dimensions, especially when considerable individual variation exists in the input. These results underscore the importance of considering individual-level variation in the production and perception of cues when investigating the acquisition and evolution of sound patterns, whether computationally or experimentally.

## 6. ACKNOWLEDGEMENTS

All data used in the simulations reported here were collected by Natasha Warner, Allard Jongman, Joan Sereno, and Rachèl Kemps, originally discussed and analyzed in [11]. Their willingness to share these data is gratefully acknowledged; however, any errors in the presentation or analysis should be attributed solely to the present author.

## 7. REFERENCES

[1] de Boer, B., Kuhl, P. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters On-line* 4(4), 129-134.

[2] Feldman, N.H., Griffiths, T.L., Morgan, J.L. 2009. Learning phonetic categories by learning a lexicon. *Proc. CogSci* 31.

[3] Hewlett, N. 1988. Acoustic properties of /k/ and /t/ in normal and phonologically disorderd speech. *Clinical Linguistics and Phonetics* 2, 29-45.

[4] Labov, W. 1994. *Principles of Linguistic Change Vol. 1: Internal Factors.* Oxford: OUP.

[5] Labov, W., Karen, M., Miller, C. 1991. Nearmergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 33-74.

[6] Lahiri, A., Schriefers, H., Kuijpers, C. 1987. Contextual neutralization of vowel length: Evidence from Dutch. *Phonetica* 44, 91-102.

[7] McLachlan, G.J., Peel, D. 2000. *Finite Mixture Models*. New York: Wiley.

[8] Schwarz, G.E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6(2), 461-464.

[9] Scobbie, J.M., Gibbon, F., Hardcastle, W.J., Fletcher, P. 2000. Covert contrast as a stage in the acquisition of phonetics and phonology. In Broe, M., Pierrehumbert, J. (eds.), *Papers in Laboratory Phonology V.* Cambridge: CUP, 194-207.

[10] Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F., Amano, S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *PNAS* 104(33), 13273-13278.

[11] Warner, N., Jongman, A., Sereno, J., Kemps, R. 2004. Incomplete neutralization and other subphonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32, 251-276.

[12] Yu, A.C.L. 2007. Understanding near mergers: the case of morphological tone in Cantonese. *Phonology* 24, 187-214.