PREDICTING DEPENDENCIES IN PHONETIC CATEGORIZATION FROM DISTRIBUTIONS OF ACOUSTICAL CUES IN NATURAL UTTERANCES

Roel Smits
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

ABSTRACT
This study addressed several basic issues in the perception of coarticulated phonemes. First, the potential consequences of coarticulation on the distributions of acoustical cues were discussed. It was argued that a pattern classifier would deal with the problem by introducing dependencies between the recognition of successive phonemes. Next, the theoretical argument was applied to the perception and production of Dutch syllables /si sy fi fy/. Based on acoustical cue distributions measured on natural utterances of these syllables, listeners’ perceptual strategies were predicted. Predictions were compared with listeners’ categorizations of a synthetic two-dimensional fricative-vowel continuum. Despite some quantitative discrepancies, good correspondence between predicted and observed categorization dependencies were found.

1. INTRODUCTION
One of the central problems in speech perception research is how listeners decode coarticulated phonemes. Massaro [1] and Nearey [2] have investigated this problem from a pattern classification perspective. Massaro assumes that listeners use the syllable as a recognition unit. Nearey has shown, however, that listeners’ categorizations of coarticulated phonemes can usually be accounted for by a model that adopts independent phoneme-based categorizations, augmented by a diphone bias term. Smits [3] recently introduced the HICAT model of hierarchical categorization. Like Nearey’s logistic regression model, HICAT assumes phoneme-based categorization, but also explicitly allows for hierarchical, i.e., unidirectional, dependencies between categorizations of successive phonemes. The HICAT model distinguishes 3 types of dependencies: dependency of the location, steepness, and orientation of the boundary of one categorization on the other categorization. Among other things, this paper shows how these distinct types of dependencies naturally arise from a pattern-classification-based analysis of the acoustical consequences of coarticulation.

2. CUE DISTRIBUTIONS AND CATEGORIZATION STRATEGIES
2.1. Acoustical consequences of coarticulation
Consider the Dutch fricative-vowel (FV) syllables /si sy fi fy/. In Dutch, /fi/ is unrounded. Acoustically, the /si-/fi/ distinction is well-captured by the position of the low-frequency edge, henceforth indicated as F3. If /s/, we would record a large number of such FV syllables and measure F3, and F3, we would find acoustical distributions much like those given in figure 1a. In reality, FV syllables are produced with regressive rounding assimilation, i.e., /si/ and /fi/ are unrounded when followed by /i/ and rounded when followed by /y/. Such assimilation may have two distinct acoustical effects. First, rounding may shift the fricative resonance down in frequency by an equal amount for /si/ and /fi/, as displayed in figure 1b. Second, the rounding may cause the fricative resonance of /si/ and /fi/ to converge, as shown in 1c. In practice, a combination of these two effects will usually occur, but in the context of the present analysis it is useful to highlight the distinction.

2.2. Pattern classification strategies
How would a pattern classifier deal with the problems shown in figure 1? First, we restrict the discussion to fuzzy classifiers, i.e., classifiers that give probabilities of each response as output, rather than a single response. Next, we define a phoneme-based classifier’s performance as the similarity of its output to a syllable-based classifier (given enough training, a syllable-based classification is optimal for the current 4-choice problem). For the problem of figure 1a, a phoneme-based classifier employing independent phoneme categorizations, with category boundaries indicated by the solid lines, would give optimal performance, i.e., its output would be indistinguishable from that of a syllable-based classifier. The solid lines in 1b indicate the optimal independent categorization strategy for the “shifted” geometry. Here, performance can be improved by either making the position of the /si-/fi/ boundary dependent on the vowel (dotted lines) or making the position of the /fi/-fy/ boundary dependent on the fricative (dash-dotted lines). The solid lines in 1c indicate the optimal independent categorization strategy for the “converged” geometry. Performance can be improved either by making the steepness of the /si-/fi/ boundary dependent on the vowel (steeper for /i/ than for /y/, indicated by the dotted lines and sigmoid shapes), or making the orientation of the /fi/-fy/ boundary dependent on the fricative (dash-dotted lines).

In summary, likely acoustic consequences of coarticulation are a shift and/or a convergence of acoustical distributions associated with one phonological distinction, depending on the phonological context. From a pattern-classification perspective, these patterns cause a decrease in performance of a classifier adopting independent phoneme categorizations. This can be repaired, however, by making the position, steepness, and/or orientation of the category boundary for one distinction dependent on the categorization of the other.
The present study investigated whether listeners base their categorization strategies on acoustical cue distributions in the fashion discussed above. A difficulty with this methodology is that we do not know the cue distributions (i.e., the “training material”) for separate listeners. Therefore, the additional assumption is made that each listener’s training material is well approximated by his or her own productions.

The study consisted of 5 steps:
1. Acoustic analysis of naturally uttered syllables /si sy /śi śy/;
2. Predicting categorization strategies from acoustical cue distributions;
3. Testing listeners’ categorization behavior for a synthetic /si sy śi śy/-continuum;
4. Analyzing perception data for dependent categorization strategies using the HICAT model;
5. Comparing predicted and observed categorization strategies.

3. PRODUCTION EXPERIMENT

3.1. Subjects.
Two male and two female native speakers of Dutch participated in the experiment. None had any history of speech or hearing disorders.

3.2. Procedure
3.2.1. Recording procedure. Subjects were seated in front of a computer screen in a soundproof booth. At each trial an orthographic representation of one of the four test syllables appeared on the screen, after which the subject spoke the syllable into a microphone. The visually presented test words were taken from a randomized list of 50 repetitions of each of the syllables /si sy śi śy/.

3.2.2. Measurement procedure. On each of the recorded syllables two measurements were made: the frequencies of a fricative resonance (Ffr) and of F3 in the vowel. The fricative noise spectra usually had low energy in the low-frequency region and a wide plateau of high energy in the mid to high frequencies.

Ffr was defined as the frequency of the left-hand (low-frequency) edge of this plateau. This definition of Ffr is compatible with the frequency of the lower of the two fricative formants used in the synthesis procedure that will be discussed later. Ffr and F3 were measured in relatively stationary portions of the signals, i.e., not in the transitional regions between fricative and vowel.

3.3. Results.
The results of the production experiment are shown separately for the four subjects in Figure 2. Several general observations about figures 2a-d are worth making. First of all, all subjects have coarticulation patterns that are combinations of the basic shift and convergence patterns shown in figures 1b and c. Second, different subjects have different mixes of the two basic patterns: subject one (2a) has a relatively strong convergence component, while subject four (2d) has a relatively weak convergence component. While the patterns for subjects one, three and four...
mainly convey “one-way” coarticulation, subject two (2b) shows considerable “two-way” coarticulation. That is, not only do we see a shift and convergence of the /sy/ and /fy/ probability-density functions (pdfs) along the F2 axis compared to /si/ and /fy/, there is also a convergence of the /si/ and /sy/ pdfs compared to /si/ and /fy/ along the F3 axis. Finally, the pdfs of the male speakers (2b and d) are located at lower frequencies than those for the female speakers (2a and c), which was expected based on differences between male and female vocal tract sizes.

4. PREDICTED CATEGORIZATION STRATEGIES

Based on the measured acoustic parameters, predictions were made for likely categorization strategies separately per subject. The method is illustrated for subject two in figure 3. For figure 3a, an independent, phoneme-based (IP) strategy was assumed, i.e., it was assumed that listeners use the phoneme as the unit for categorization, and that the fricative and vowel categorizations are independent. Twodimensional Gaussian pdfs were estimated for the vowels and fricatives. In these estimations it was assumed that the covariance matrices for the /i/ and /y/ pdfs are equal, and that the covariance matrices for the /s/ and /f/ pdfs are equal. The two dashed ellipses in 3a that are oriented more or less horizontally represent equi-probability contours of the pdfs of /i/ and /y/. The fricative pdfs are represented by the more vertically oriented dashed ellipses. The assumptions of independence and equal covariance matrices cause the category boundaries for the fricative and vowel distinctions to be straight rather than curved. These boundaries are represented by the solid lines in 3a. Based on the phoneme pdfs, “underlying” syllable pdfs can be calculated that would lead to the same categorization functions as the phoneme-based pdfs. The underlying syllable pdfs are represented by the dotted ellipses in 3a.

For figures 3b, c, and d, a dependent phoneme-based (DP) strategy was assumed, with the fricative categorization depending on the vowel categorization. Like for figure 3a, the two vowel pdfs are assumed to be equal, as are the “underlying” syllable pdfs. Figures b, c, and d allow the position, orientation, and steepness of the fricative boundary, respectively, to depend on the vowel. The steepness of the /s/-/f/ boundary in 3d is higher for /i/ than for /y/ (the boundary steepness cannot be represented in a territorial plot).

For figure 3e, it was assumed that listeners use the syllable as the recognition unit (S strategy). Again the syllable pdfs are assumed to be equal for /sy/ and /fy/.

Finally, figure 3f represents subject two’s categorization behavior as fitted by the HICAT model. Figure 3f will be discussed in a later section.

How can we use the calculations represented by figures 3a to e to predict what categorization strategy the subject will use? Recall that the performance of a phoneme-based classifier was defined as the similarity of its output to the output of the syllable-based (optimal) classifier. The similarity of the classifier’s output to the S classifier was approximated by the similarity of the underlying syllable pdfs for the IP and DP classifiers to the actual syllable pdfs for the S classifier, defined as the inverse of the average Bhattacharyya distance [4] between the respective syllable pdfs. Using this method, it was found that, for all four subjects, a dependency of the fricative categorization on the vowel led to higher performance than the reverse dependency. Of the 3 types of dependency, the steepness dependency led to highest performance. Thus, on the basis of the production experiment, it

Figure 3. Predicted categorization boundaries based on subject 2’s utterances assuming IP (a), DP (b, c, d), and S (e) strategies. Dashed ellipses are equi-probability contours of bivariate Gaussian pdfs for phonemes, dotted ellipses for syllables. In panels b, c, and d, the position, orientation, and steepness of the fricative boundary depend on the vowel. Panel f gives subject 2’s actual categorization boundaries inferred from his perception data. The small rectangle indicates the range of the stimulus continuum.
was predicted that all subjects would have a perceptual strategy in which the steepness of the /s/-/ʃ/ boundary depends on the vowel.

5. PERCEPTION EXPERIMENT

5.1. Method
A 64-member, twodimensional synthetic FV continuum was created by varying F3 in 8 steps from 2450 Hz to 2625 Hz, and F3 in 8 steps from 2890 Hz to 3310 Hz. 60 repetitions of each of these stimuli were randomized and presented through headphones to listeners for categorization as /si/, /sy/, /ʃi/, or /ʃy/. The same four subjects who participated in the production experiment took part in the listening experiment.

5.2. HICAT model fits
Separate HICAT model fits were made for the perception data of each subject, testing for dependency direction and type. The results are given in table 1. Figure 3f shows the category boundaries for the best-fitting HICAT model for subject two’s perception data.

<table>
<thead>
<tr>
<th>subject</th>
<th>dependency direction</th>
<th>dependency type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fric. depends on vowel</td>
<td>steepness</td>
</tr>
<tr>
<td>2</td>
<td>Fric. depends on vowel</td>
<td>orientation</td>
</tr>
<tr>
<td>3</td>
<td>Vowel depends on fric.</td>
<td>position</td>
</tr>
<tr>
<td>4</td>
<td>Fric. depends on vowel</td>
<td>steepness</td>
</tr>
</tbody>
</table>

Table 1. Dependency direction and type for the best-fitting HICAT model for the perception data of each of the four subjects.

It is obvious from table 1 that subjects are very different in their categorization strategies. However, for 3 out of 4 listeners the fricative categorization depends on the following vowel.

6. COMPARISON OF PREDICTED AND OBSERVED STRATEGIES

6.1. Comparison within subjects
Overall, prediction of categorization strategies within subjects showed mixed success. Qualitatively, the individual predictions for subjects 1 and 4 were in agreement with their perception data, while those for subjects 2 and 3 were not. Several other quantitative aspects of the predicted strategies for all subjects were in disagreement with the observed strategies. For example, predicted steepness of category boundaries was much too high compared to observed steepness in the perception data. A possible reason for these discrepancies is that the variability in the acoustic data within syllables was unnaturally low because syllables were uttered in highly controlled conditions and in isolation. If the syllables had been taken from whole words, the within–syllable variability may be much higher.

6.2. Comparison between subjects
A comparison between subjects of the effects of increasing the complexity of the classifier provided much more encouraging results. For all four subjects, the relative reduction in badness-of-fit was calculated that was achieved by changing the categorization strategy from IP to DP, and from DP to S. These calculations were made both for the prediction stage, based on the production data, and for the HICAT model fits on the perception data. The orderings for the subjects for decreasing reduction in badness-of-fit are given in the table 2.

<table>
<thead>
<tr>
<th></th>
<th>IP to DP</th>
<th>DP to S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production</td>
<td>3-1-4-2</td>
<td>2-4-1-3</td>
</tr>
<tr>
<td>Perception</td>
<td>3-1-2-4</td>
<td>2-4-1-3</td>
</tr>
</tbody>
</table>

Table 2. Ordering of subjects for decreasing reduction in badness-of-fit as predicted from the production data and inferred from the perception data.

For example, of all four subjects, the recognition performance on subject 3’s utterances improved most by making the fricative categorization dependent on the vowel. Also, of all subjects, the HICAT model fit to subject 3’s perception data benefited most from allowing the fricative categorization to depend on the vowel. In conclusion, comparison of predicted and observed strategies between subjects shows excellent agreement.

7. DISCUSSION AND CONCLUSIONS

This paper addressed several basic issues in the perception of coarticulated phonemes. First the potential consequences of coarticulation on the distribution of acoustical cues was discussed. It was argued that a pattern classifier could boost recognition performance by making the position, orientation, or steepness of the category boundary for one phoneme dependent on the adjacent one. Next, the theoretical argument was applied to the perception and production of Dutch syllables /si sy ji jy/, which is a case of regressive rounding assimilation. Based on acoustical cue distributions measured on natural utterances of these syllables, listeners’ categorization strategies were predicted. Predictions were compared with HICAT model analyses of listeners’ categorizations of a synthetic two-dimensional fricative-vowel continuum. It was found that

1. Individual listeners differ in their categorization strategies, but for 3 out of 4 listeners the fricative categorization depends on the following vowel.
2. Quantitative discrepancies were found between predicted categorization strategies for individual subjects based on their own productions and observed strategies inferred from their perception data. The estimation of the listeners’ “training data” needs to be improved.
3. Predicted ordering of subjects according to their likelihood of using independent phoneme-based, dependent phoneme-based, or syllable-based strategies correlated well with observed ordering.

REFERENCES