CONTEXT-DEPENDENT RELEVANCE OF BURST AND TRANSITIONS FOR PERCEIVED PLACE IN STOPS: IT'S IN PRODUCTION, NOT PERCEPTION

Roel Smits

Dept. Phonetics and Linguistics
University College London
United Kingdom

ABSTRACT

Several studies on place perception of prevocalic stop consonants have shown that the apparent perceptual weight of release burst and formant transitions depends on the vowel context: bursts carry higher perceptual weight in high front vowel contexts like /i/ than in low non-front vowel contexts like /a/, while the reverse holds for formant transitions. This finding is generally interpreted as reflecting a context-dependent "reweighting" of burst and transition cues by the perceptual system. In this paper it is shown that the observed effect can be entirely accounted for by contextual variation of the distributions of the relevant cues themselves: Naturally produced stop bursts appear to be acoustically more distinctive in high front vowel contexts than in low non-front vowel contexts, while the reverse is true for formant transitions. The apparently context-dependent perceptual weight of burst and transitions can be reproduced with such cue distributions, even if the classification model itself contains fixed, context-independent linear boundaries. This claim is supported with acoustical and perceptual data of a burst-splicing experiment involving burst-spliced stop-vowel utterances containing the stops /p, t, k/ and vowels /a, i, y, u/.

1. INTRODUCTION

Fischer-Jorgenson (1972) performed a perception experiment in which the relevance of release burst and formant transitions for place perception in stop consonants was tested. From naturally produced stop-vowel syllables in vowel contexts /a, i, u/ the release bursts were excised and exchanged between utterances with conflicting place of articulation using a tape-splicing technique. A similar procedure was carried out for the formant transitions. The cross-splicing was only carried out within vowel contexts. The resulting stimuli were presented to listeners for stop classification. The results showed that the relative perceptual relevance of release burst and formant transitions highly depends on the vowel context. Generally speaking, the release burst appeared to determine place of articulation in vowel context /i/, while the formant transitions were dominant in vowel context /a/. This phenomenon of a vowel dependency of the perceptual relevance of burst and transitions has subsequently been corroborated by several studies in which bursts were either deleted from utterances or presented in isolation (e.g. Dorman et al., 1977; Pols, 1979).

Visual inspection of spectrograms of the stimuli led Fischer-Jorgenson to hypothesise that "... transitions, aspirations, and explosions of a given consonant which differ ... from those of other consonants before the same vowel are more efficient than those which differ little from one consonant to the other before the same vowel" (Fischer-Jorgenson, 1972, p. 158, 159). In particular, the formant transition patterns appeared to be very different for the three places of articulation in /a/ context while the bursts were rather similar. The reverse was true in the /i/ context. It was suggested by Fischer-Jorgenson that the perceptual system weighs the more effective cues more heavily in the stop classification than the less effective cues.

How can this, rather qualitative, hypothesis be implemented in a more formal model of the classification process? Assuming that the unit of recognition is the segment or smaller, Fischer-Jorgenson's hypothesis seems to suggest that the following more or less discrete steps are made in the classification process: (1) vowel recognition; (2) "fine-tuning" of stop-recognition model in terms of cue weighting; (3) stop recognition. Obviously, a mechanism like this is much more complex than a recognition model in which the consonant is recognised directly without dependency of the vowel.

In this paper it is shown that the latter direct recognition model can indeed explain the observed seemingly vowel-dependent behaviour, and explicit mechanisms which fine-tune the stop-recognition to the vowel identity are not required.

2. PLACE-PERCEPTION IN STOPS: STUDY BY SMITS (1995)

Recently, Smits (1995) (see also Smits et al., submitted a, b) carried out a study of place perception in stops. The purpose of the study was to investigate which of the large number of cues for stop place that have been proposed over the years are perceptually most important, and to provide a quantitative description of how the cues are used by the perceptual system in the classification task. The research consisted of two parts: (1) a perception experiment with burst-spliced stop-vowel utterances; (2) a formal simulation of listeners' classification behaviour.

2.1. Perception Experiment

Two male Dutch speaker produced stop-vowel syllables, with stops from the set /p, t, k/ and vowels from the set /a, i, y, u/. Note that Dutch unvoiced stops are unaspirated. From these syllables the burst was electronically removed and spliced onto the burst-less parts of syllables with the same vowel but a different stop, thus mimicking Fischer-Jorgenson's procedure.

The task of the subjects was to label each stimulus as either P, T, or K.

The results show general trends which are very similar to the ones described by Fischer-Jorgenson. In the vowel context /a/ the transitions dominate the perceived place of articulation and the burst is hardly relevant. The reverse holds in the /i/ and /y/ contexts. The listeners' classifications of the burst-spliced utterances of one speaker are summarised in Table 1.

	burst	**trans**	**other**
/a/	14.9	81.4	3.7
/i/	69.1	23.9	7.0
/y/	82.0	13.8	4.2
/u/	52.7	27.7	19.6
total	54.7	36.7	8.6

Table 1: Percentages of listeners' classifications of the burst-spliced utterances in accordance with the burst (**burst**), the transitions (**trans**), or the remaining class (**other**) broken down for vowel context.

An analysis of variance on the dependent variable *percentage responses in accordance with burst* showed that the vowel identity was a highly significant factor. The same was found when *percentage responses in accordance with transitions* was used as dependent factor.

2.2. Simulation study

The second part of the investigation was a simulation of the classification behaviour of the listeners. It was assumed that the listeners' classification of a stimulus essentially consists of two steps: (1) a mapping of the speech signal onto a vector in a multidimensional cue space, (2) classification of the cue vector as P, T or K.

First, a large number of acoustic cues for stop place of articulation that have been proposed over the years were measured on all stimuli. Both "detailed" cues, such as formant and burst frequencies, and "gross" cues, such as global spectral tilt and compactness, were measured using semi-automatic and manual procedures.

Next, all possible subsets of 1 to 7 cues out of the total set of 17 cues were mapped onto the observed responses using a formal model of human classification behaviour. The classification model was based on the *single-layer perceptron* (SLP). The properties of this model will be discussed later. Suffice it to say at this point that the model simply maps an incoming cue vector (representing a stimulus) onto an output vector containing the probabilities of choosing each of the possible responses (P, T, or K), without having knowledge of the vowel identity. In the model-fitting procedure a formal cross-validation technique known as "leaving one out" was applied to avoid over-fitting (Smits and Ten Bosch, submitted). Separate model fits were made for the data associated with the two different speakers. Eventually, for each of the two speakers, the combination of cue set and model which produced the highest level of goodness-of-fit was selected for further study. Here we will only discuss the best cue-set-plus-model for one of the two speakers.

The cue set associated with the best-fitting model consisted of the following 5 cues: burst length l_b, frequency of F2 at voicing onset $F2_O$, frequency of F3 at voicing onset $F3_o$, frequency of a broad mid-frequency peak at burst onset F_o, and the relative level of this peak L_o. The responses predicted by the model on the basis of this cue set was, like the listeners' responses, analysed in terms of the percentage responses in accordance with the burst or transitions. The results of this analysis are listed in Table 2. To facilitate a comparison with the listeners' response pattern listed earlier in Table 1, the results for both the listeners' responses as well as the model responses are shown graphically in Figure 1.

	burst	**trans**	**other**
/a/	12.1	78.4	9.5
/I/	58.3	19.9	21.8
/y/	76.5	14.1	9.4
/u/	41.6	33.8	24.6
total	47.1	36.5	16.4

Table 2: Percentages of the model's classifications of the burst-spliced utterances in accordance with the burst (**burst**), the transitions (**trans**), or the remaining class (**other**) broken down for vowel context.

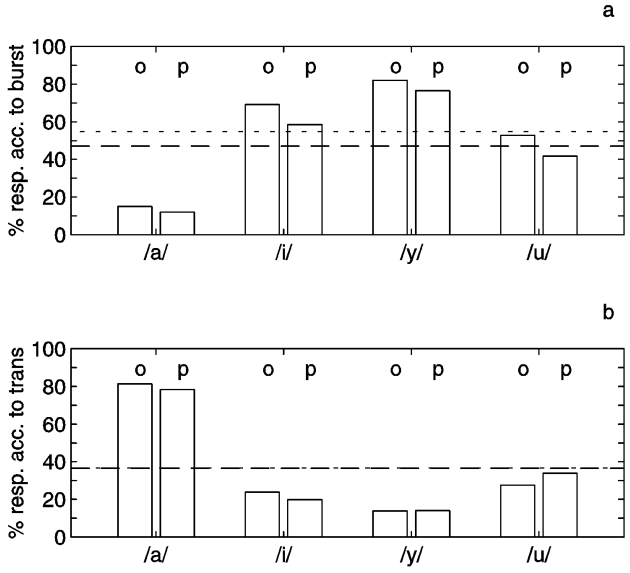

Figure 1: Percentages of the observed ("o") listeners' classifications and predicted ("p") model's classifications of the burst-spliced utterances in accordance with the burst (1a), or the transitions (1b), broken down for vowel context. The dashed lines indicate average levels, with the short and long dashes referring to the observed and predicted classifications, respectively.

The model predictions closely mimic the response pattern found for the listeners' classifications of the burst-spliced utterances. Interestingly, the model classifies the burst-spliced stimuli mostly in accordance with the formant transitions in vowel context /a/, while in contexts /i/ and /y/ the burst largely determines classification, just like we found for the listeners. This is somewhat surprising because the input information to the model does not contain any explicit information on the vowel identity. The model is therefore not capable of making a vowel-dependent adjustment in the weighting of burst cues and formant cues. Hence, the three step mechanism discussed earlier, in which first the vowel is recognised, next the stop-recognition model is fine-tuned, and finally the stop is recognised, is ruled out.

3. CUE DISTRIBUTIONS AND CLASSIFICATION PROPERTIES

In order to understand the somewhat paradoxical finding in the previous section we will look more closely at the properties of our classification model. First, let us define a *response region* of a class C as the subspace of the cue space where class C is the most likely response. Furthermore, we define a *class boundary* as the subspace of the cue space which separates different response regions.

The SLP-based model used in the simulation supports linear class boundaries, which means that the class boundaries contain only zero- and first-order terms of the acoustic cues. The class boundaries implemented by our best-fitting model are given in Eqs. (1)-(3).

$$B_{pt} = 0.2l_b - 2.0F2_o + 0.05F3_o - 3.6F_o + 1.0L_o + 1.9 = 0 \quad (1)$$

$$B_{pk} = -3.8l_b - 4.9F2_o + 0.4F3_o + 4.8F_o + 1.6L_o - 0.4 = 0 \quad (2)$$

$$B_{tk} = -4.1l_b - 2.9F2_o + 0.3F3_o + 8.4F_o + 0.6L_o - 2.3 = 0 \quad (3)$$

where Bpt, Bpk and Btk indicate the class boundaries between p and t, p and k, and t and k, respectively. The acoustic cues appearing in the boundary equations are actually normalised to unit variance by subtracting their mean followed by a division by their variance.

Now we define an acoustic cue's *weight* to the perception of a given phonetic distinction as the absolute value of its coefficient in the relevant boundary equation. For example, the weight of $F2_O$ in the labial-coronal distinction (Eq. 1) is 2.0. We define a cue's *average weight* as the average of the weights for the three distinctions. For example, the average weight of F_o is (3.6+4.8+8.4)/3 = 5.6. It is necessary to define a cue's weight relative to its natural variance because otherwise the coefficient's magnitude would have arbitrary size, depending on the relevant unit. For example, if we would change the unit of the cue l_b from [ms] to [s], without doing the normalisation, its weight would increase a thousandfold). Obviously, other definitions of the weight of a cue are possible but the one chosen here is particularly suited to our purpose.

Graphically, a cue's weight is related to the slant of the hyperplane representing the class boundary relative to the cue axis. If the boundary is almost perpendicular to the cue axis the cue's weight is large, whereas as the boundary is almost parallel to the cue axis the cue's weight is close to zero. If the weight of a cue changes the class boundary rotates.

The point of all this is the following. If the cue distributions differ per vowel context, the cue weights effectively differ as well due to different normalisations. If we recalculate the cue distributions separately per vowel, we indeed find that they differ greatly per vowel context. The means and standard deviations of the five cues in the four vowel contexts are listed in Table 3.

		l_b (ms)	$F2_o$ (ERB)	$F3_o$ (ERB)	F_o (ERB)	L_o (dB)
/a/	μ	11.1	18.4	21.5	19.7	3.68
	σ	4.3	1.5	0.8	1.2	1.8
/i/	μ	23.7	20.8	23.1	21.8	5.50
	σ	17.8	0.3	1.1	3.7	2.52
/y/	μ	25.3	19.9	21.4	18.9	4.89
	σ	3.7	0.3	0.2	4.8	1.6
/u/	μ	19.0	15.9	20.4	19.3	5.74
	σ	8.8	2.7	1.5	4.8	2.1
all	μ	19.8	18.7	21.6	19.9	4.95
	σ	11.4	2.4	1.4	3.9	2.1

Table 3: Means μ and standard deviations σ of the five cues used in the best-fitting model calculated separately per vowel context as well as across all contexts. The frequency cues are expressed in the psychoacoustically plausible unit ERB (Equivalent Rectangular Bandwidth) rather than Hz.

Let us concentrate on the most important burst cue F_o and the most important transition cue $F2_o$. We can calculate the effective average weight of each of these cues in the four vowel contexts by dividing the average weight by the overall standard deviation and multiplying the result by the standard deviation in the respective vowel context. The resulting "effective" average weights of F_o and $F2_o$ in the four vowel contexts are listed in Table 4.

	$F2_o$	F_o
/a/	2.0	1.7
/i/	0.4	5.3
/y/	0.4	6.8
/u/	3.7	6.8
all	3.3	5.6

Table 4. Effective average weights of $F2_O$ and F_o calculated separately per vowel context as well as across all vowels. This is an item in a bulleted list.

We find that in context /a/ $F2_O$ is weighted slightly more heavily than F_o. In context /i/, on the other hand, the weight of F_o is more than 10 times as large as that of $F2_O$. The situation is even more extreme for /y/. Finally, in the /u/ context F_o is weighted slightly more heavily than $F2_O$. This reflects exactly the

behaviour found for the listeners as well as the model: The burst is most important in the contexts /y/ and /i/ and least important in the /a/ context.

Figure 2 displays a 2-dimensional cross-section of the 5-dimensional cue space along the F_o and $F2_o$ axes. The solid lines represent the model's linear class boundaries separating the response regions for P, T, and K. The two ellipses are equi-probability contours of two 2-dimensional Gaussian probability density functions (pdfs) fitted to the acoustic cues for the vowel context /a/ (short dashes) and /y/ (long dashes). The figure clearly shows that in the /a/ context the transition cue $F2_o$ largely determines the classification, while in context /y/ burst cue F_o dominates the classification**.**

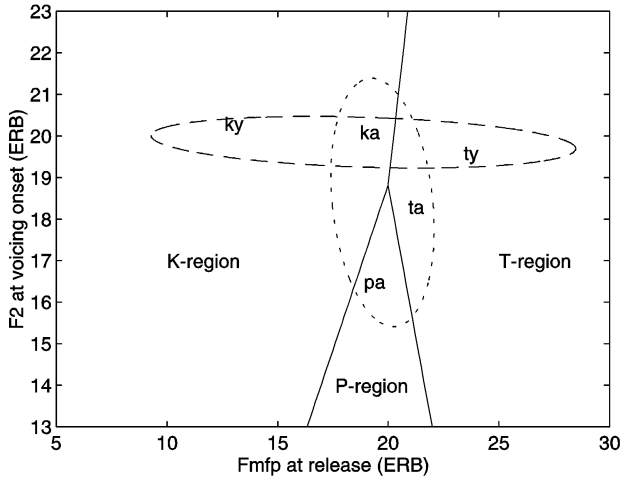

Figure 2: 2-dimensional cross-section of the 5-dimensional cue space along F_o (horizontal axis) and $F2_o$ (vertical axis). The solid lines are the model's linear class boundaries. The two ellipses are the 2σ equi-probability contours of two 2-dimensional Gaussian probability density functions fitted to the acoustic cues for the vowel contexts /a/ (short dashes) and /y/ (long dashes). The syllables printed within the ellipses represent typical tokens.

4. DISCUSSION

We have demonstrated that the acoustic distributions of the cues affect the effective weights of cues in each vowel context, even though the classification model is not "actively" adjusted per vowel context. Context-dependent cue-weighting mechanisms, such as the one suggested in the introductory section, are therefore not necessary for an adequate modelling of the classification behaviour. Obviously, these vowel dependent differences in the cue distributions originate in articulatory processes. We therefore conclude that the context-dependence of the perceptual relevance of burst and transition cues is not caused by any perceptual reweighting processes, but by differences in acoustic cue distributions generated in production, hence the title of this paper.

More generally speaking, it may be the case that mechanisms such as the one described in this paper are partially responsible for the apparent "tuning-in" on different signal portions when listening to different speakers or in different listening conditions. For example, when speaker 1 produces stop release bursts which provide relatively unambiguous information on place of articulation, while speaker 2 does so for the formant transitions, a fixed perceptual system will, as it were, automatically weigh the more effective cues more heavily.

Finally, at a level of general scientific strategy, we would like to argue that knowledge about classification models is important for speech perception research. The study presented in this paper clearly demonstrates that a formal simulation of the listeners' phonetic classification behaviour followed by a close scrutiny of the classification process itself may generate fundamental insights into the relevant psychological processes. Moreover, assumptions which unnecessarily increase the complexity of a perception model may be avoided.

5. REFERENCES

1. Dorman, M.F., Studdert-kennedy, M., and Raphael, L.J. “Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues,” *Perception & Psychophysics 22*: 109-122, 1977.

2. Fischer-Jorgensen, E. “Tape-cutting experiments with Danish stop consonants in initial position,” *Annu. Rep. Inst. Phon., Univ. Copenhagen 6*: 104-168, 1972

3. Pols, L.C.W. “Coarticulation and the identification of initial and final plosives,” in: J. Wolff and D. Klatt (eds.), *ASA 50 Speech Communication Papers*, Acoust. Soc. Am., New York, 459-562, 1979.

4. Smits, R. *Detailed versus gross spectro-temporal cues for the perception of stop consonants*, Doctoral dissertation, Inst. Percep. Res. (IPO), Eindhoven, Netherlands, 1995.

5. Smits, R. and Ten Bosch, L. “The perceptron as a model of human categorization behavior,” submitted to J. Math. Psychology.

6. Smits, R., Ten Bosch, L., and Collier, R. “Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants: I. Perception experiment,” submitted (a) to J. Acoust. Soc. Am.

7. Smits, R., Ten Bosch, L., and Collier, R. “Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants: II. Modeling and evaluation,” submitted (b) to J. Acoust. Soc. Am.