Relative importance of sound segments for the identification of Swedish stops in VC and CV syllables

Öhman, S.

journal: STL-QPSR
volume: 2
number: 3
year: 1961
pages: 006-014

http://www.speech.kth.se/qpsr
A. RELATIVE IMPORTANCE OF SOUND SEGMENTS FOR THE IDENTIFICATION OF SWEDISH STOPS IN VC AND CV SYLLABLES

Low level speech perception

The role of the listener in natural face-to-face conversation is extremely complicated, and it is likely that the decoding of spoken messages involves a simultaneous activation of a hierarchy of parallel analysis and probability mechanisms (1). In order to obtain a correct understanding of that which has been spoken the listener is able to use his knowledge about the social situation in which the conversation takes place, about the probable content of present and forthcoming utterances, about the talker's pronunciation and formulation habits, about the syntactical, grammatical, and phonemic rules for the language in which the conversation is held, and about the acoustic structure of the sounds which actually reach his tympanic membrane.

In the experiment to be described here, we have tried to isolate that part of the speech decoding activity which works exclusively with the acoustic signal. In an identification task like that of our test, the listener is thought of as a phoneme detecting device which matches stored sound patterns with those of the stimuli heard and then responds with a phoneme letter corresponding to the sound pattern giving the best match. When identification of phonemes follows this model we call it "low-level identification".

Summary of experiment

For a more detailed description of stimuli, test, and data processing the reader is referred to QPSR 2/1961 (2), where also some of the results are briefly surveyed. In the following we concentrate on the response matrices for the voiceless and voiced stop consonants.

The low parts of Fig: s II-1 through II-8 show sonagram pictures of the stimulus words. These sonagrams were prepared by means of halving the play-back speed of the tape-recorder and expanding somewhat the frequency scale. The steps of the scale along the time dimension are of 20 msec duration. The ends and beginnings of
these steps correspond to points where the words were cut into two parts by means of an electronic gating system (having relatively steep opening and closing characteristics). Each of these cuts leaves us with two stimuli, one initial VC-like sound, and one final CV-like sound. The VC-like stimuli were randomized and presented to listeners for identification in one test, and so were the CV-like ones in another test, these tests being called "the VC test" and "the CV test" respectively. The results for the stops are shown in the graphs above the sonagrams, responses in "the VC test" being indicated by arrowed curves pointing in the direction with time, and those of "the CV test" in the direction against time.

Model

Let us consider the situation where one of our subjects, $S_1$, is presented with one of the test words, say $w_1 = [zZP]$, in its complete, un-cut version. (Obs. This is not our test situation.) $S_1$ is supposed to tell what consonant he hears, and all consonants are equally probable to occur.

Since the listening conditions are optimal, since $S_1$ has good ears, and since $w_1$ is a well-formed, carefully pronounced Swedish (meaningless) word, the identification of the consonant $c_1 = [p]$ will be correct with probability almost one. Now, $w_1$ contains a number of cues of varying importance for $c_1$, and they are non-uniformly distributed over $w_1$ in time. ($w_1$ begins at $t_0$ and ends at $t_n$.) $S_1$ takes account of these cues and on the basis of them identifies the consonant $c_1$ correctly (i.e. with probability one). But this is like saying the following:

Importance value

Each segment of the word $w_1$, i.e. the acoustic content of each infinitesimal or finite time interval $t (t_0 \leq t \leq t_n)$ of $w_1$, can be thought as having an "importance value", $v(t)$, for the identification of $c_1$ in $w_1$ under the experimental conditions specified above. (Obs. $v$ is dependent on both $w$, $c$, $t$, and the experiment.) These values are summated over time by $S_1$, and when the sum function has
reached a certain critical value or "threshold region", \( S_1 \) responds with the letter for \( c_i \).

In view of this it can be asked: if identification is 100% correct after \( t_n \) of our stimulus words, what is it at times earlier than \( t_n \)? We answer this question by setting up the following hypothesis:

At a time \( t_1 < t_n \) the identification function \( i(t_1) \) is less than or equal to 100%, and in general, \( i(t_1) \) is proportional to the integral of \( v(t) \) from \( t_0 \) to \( t_1 \). Also, if the word is not switched off somewhere before \( t_n \) but rather switched on at a time later than \( t_0 \), say at \( t_j \), so that now only the final part of the word is heard, the identification \( i(t_j) \) of (the now initial consonant) \( c_i \) will be proportional to the integral of \( v(t) \), this time taken from \( t_n \) to \( t_j \).

By moving, in an identification experiment, the cut-point \( t_i \) from an early position and successively forwards, and, in another experiment, \( t_j \) from a late position and successively backwards in time, it should be possible to estimate the above-mentioned integrals and hence parts of the function proportional to \( v(t) \). These integrals are shown in the graphs of Figs 11-1 to 11-8.

If \( v(t) \) is independent of context, the differentiated identification function of the VC test for some consonant should be identical with the differentiated function of the CV test for the same consonant. This ideal is approached in a large class of the test words, but perhaps not in most of those discussed here. It seems that consonants which are, so to speak, acoustically homogeneous and tied up with the preceding and the succeeding vowels in a symmetrical way (such as /a\2j a\2/, /a\2v a\2/, /a\2l a\2/, and so on), that they also have symmetrical VC and CV identification curves. This problem will be discussed in more detail in a later publication.

Perceptual phoneme boundary

Since \( i(t) \) is monotone increasing, \( v(t) \) must be defined to be non-negative for all values of \( t \). The point in time where \( \frac{d}{dt} i(t) \) is maximum is the point which, according to the above reasoning, has
the greatest importance value for the identification of $c_1$. This is 
also the point at which $i(t)$ is steepest. We call this point "the per-
ceptual VC boundary or CV boundary for our stimulus words in our test".

In the preceding QPSR we defined "perceptual boundary" in 
a different way and made a comparison of that sort of boundary with 
boundaries determined from purely acoustic phonetic criteria. Con-
siderable divergence between the placements of the two were than ob-
served. With the definition given above, however, a much better agree-
ment between physical and perceptual point of division is obtained in 
the whole material. As a matter of fact they coincide with pulse minus 
a few msec deviation in much more than 50% of the cases.

Since, in a sense, $v(t)$ is an error function, we have 
tried to fit our data with normal probability curves (ogives). The 
fit is quite good for a number of stimuli (not shown here), notably 
in the liquids and voiced and voiceless fricatives. However, this 
part of the study is not yet completed and is hence only mentioned in 
passing. Obviously, if the normal hypothesis fits, $v(t)$ can be esti-
imated for all values of $t$, and a "boundary indeterminacy" measure can 
be defined as, e.g., the time interval between the points at which 
$i(t) = 50\% \pm \sigma$. (In a transition from a vowel /a2/ to an /l/ this 
indeterminacy interval is no more than 15 msec.)
For voiceless stops generally $s$ was much smaller in responses of the CV test than those of the VC test. It can be expected that $s$ is large when the uncertainty of the subject as to the identity of the consonant is great. And hence, this uncertainty was greater when the vowel was heard before the voiceless stop than when it was heard after the voiceless stop. When the stimulus contains poor acoustical information about the consonant, the uncertainty of the subject is great, and when $s$ is large it is not entirely unambiguous which way is the best way of drawing curves to fit the data. However, $a_2p$, $a_2t$, and $a_2rt$ are probably the most difficult stimuli to judge for the subjects in our whole material since the voice source breaks off before any considerable formant transitions occur in the final part of the initial $a_2$ and, accordingly, a minimum of acoustical information about the consonant is given to the subject. From this we conclude that $s$ is not greater in the responses to the rest of the stimulus words than it is for $a_2p$, $a_2t$, and $a_2rt$, and that the curves drawn in the graphs are sufficiently accurate.

The responses to $a_2p$, $a_2t$, and $a_2rt$ differ from the rest of the stops in the VC test in that the former are not recognized by the subjects until the burst is heard. The relatively large figure of 40% p-responses to $a_2p$ from cut-point 1 to 13 is influenced by the fact that the closing characteristic of the gate was relatively steep. Any vowel cut off by this gate gives the impression of ending in a weak p-sound.

Fig. II-1, voiceless bilabial stop

Out of all the responses to $a_2p$ an average of 18.3% were /f/, /v/, or /b/. As seen in the graph 13.1% were /f/ responses. Most of the missing 43% or so were /h/ responses. These have not been indicated in the graph since they, together with the /p/ responses, were the favorite letters used by subjects when they could not make up their minds.

No /f/ responses occurred in the CV part of the test. Instead a weak /v/ component can be analyzed out of the response matrices. This is an instance of a trend observed all through the voiceless
stops. An intervocalic voiceless stop followed by /a_2/ and with no vowel before it, is heard as voiced. (Cf. ref.(3) where the same phenomenon is found in stops after /s/ in English when the /a/ has been spliced out manually.) The /p/ responses to /p_a_2/ were given by those subjects who had some phonetic experience. The majority of the rest, however, answered with /b/, and lumping /p/ and /b/ responses together the "/b/ or /p/" curve of the graph is obtained.

The reason for the voiced responses to the voiceless stop part of the CV test is probably to be found in the fact that Swedish voiceless stops are much less aspirated intervocally than in initial position, whereas the voiced stops possess about the same degree of aspiration initially as the voiceless ones have intervocally. Moreover, the burst seems to be much more important than the transition of the initial part of the final /a_2/ considering that identification is almost 100 % at cut-point 13 (i.e., when the burst is included in the stimulus) but only about 43 % at cut-point 14 (i.e., when the burst is not included in the stimulus).

We want to put a phoneme boundary where the identification curve is steepest. This occurs somewhere between points 13 and 14, i.e. at the p-spike on the sonagram. According to our model, this point coincides with the point of maximum importance (in this case in passing from a consonant to a vowel). The corresponding point of maximum importance in the VC transition cannot be determined for /p/.

Fig. II-2, voiceless dental stop

The response picture for /a_2ta_2/ has many features in common with that for /a_2pa_2/. In the VC test the /t/ responses never exceed 20 % before the burst is included at the end of the stimulus. At that point, however, there is a rapid rise in the identification curve. 15 = 8.2 % for /t/ in the VC test as compared with 6.5 % in the test of other time direction for "/t/ or /a/". Not indicated in the graph for lack of space. The point of maximum importance coincides with the burst spike, that is between cut-points 13 and 14 of the sonagram. Final vowel transitions seem to be of mediocre importance.
Fig. II-3, voiceless supradental stop

The low /r̥t/ response percentage in the CV test is perhaps explained by 1) the fact that this phoneme never occurs initially in Swedish, and 2) that naive listeners are not conscious of its being an autonomous phoneme since orthography lacks any separate sign for it. According to Elert (4) /r̥t/ and /r̥d/ can on grammatical grounds be resolved as being composed of two phonemes /r/ and /t/ or /d/. As seen in the graph the majority of responses were of stop character. The VC test never included the burst.

The large spreads (10.9 % for /-r̥t/) are to be expected since an additional uncertainty factor is introduced in that the sound was unfamiliar as a phoneme to the subjects.

Fig. II-4, voiceless palatal stop

The response picture for this consonant is more similar to those of the voiced stops than to the voiceless ones. The preconsonantal vowel contributes a lot to the identification. However, the forward pointing curve never exceeds 68 % before the burst is included. The steepest part of it seems to lie between cut-points 3 and 4, which is about the moment when the positive F2-transition of the initial vowel begins.

Also here the tendency to respond "voiced" in the CV part of the test recurs. The "/-k/ or /-g/" curve is steepest where the aspirative part of the burst ends. A clear /j/ response component is observed when the burst is not included in the CV test but vanishes during the burst. The transitions of the final vowel alone seem to be sufficient for a correct identification of the stop.

Fig. II-5, voiced bilabial stop

The /b/ response curve for the VC test is steeper than that for the CV test indicating a broader indeterminacy region in the transition from stop to final vowel than from vowel to stop. The /-b/ curve seems to be steepest at cut-point 4 or somewhat later, i.e. coinciding with the beginning of the occlusive interval. In the region
before this point, /-p/ responses were quite common.

The /b-/ curve of the CV test changes maximally between points 12 and 13, in which interval the b-burst occurs. Around points 10 and 11 /v-/ responses were not uncommon, and the /b-/ curve does not reach its maximum but at point 7, suggesting that a "good" initial /b/ requires a certain minimum duration of the voice bar.

Fig. II-6, voiced dental stop

The /da2/ stimuli were easier to identify than the /a2d/ stimuli which is seen from the fact that the /-d/ curve does not exceed 75%. In standard Swedish pronunciation it is not uncommon that a final /d/ is released with a short neutral vowel which probably helps identifying the sound.

Perceptual boundaries are about points 3 to the left and between 11 and 12 to the right.

Fig. II-7, voiced supradental stop

There is the same problem with the initial /rd/ stimuli as with the corresponding initial voiceless stop, /rt/. The phoneme is not possible in this position in correct Swedish, and naive listeners do not regard it as an autonomous "sound". This might explain why the /rd/- curve ends at 65%. Listeners tended to use the letter "d" as response to the stimulus and treating /d/ responses and /rd/ responses as the same we obtain the broken "/d-" or /rd-" curve. This is steepest between points 12 and 13 where also the burst of the stop is situated.

The /-rd/ curve slope has its maximum between points 2 and 3. During this interval there is a very sudden negative transition of the fourth formant, typical of the moderate retroflexion which precedes this stop. During the final part of the initial vowel /rt/ responses are common. These are replaced by /rd/ responses when the subjects are given more and more of the vowel. From this it can perhaps be concluded that in this experiment first the place of articulation is established by the listener (on the basis of the F4 transition),
and then the voiced/voicedless distinction is determined (on the basis of the duration cue).

Fig. II-8, voiced palatal stop

Perceptual boundaries are situated between cut-points 6 and 7 to the left, and 14 and 15 to the right. Initially /\textit{t/} responses are observed and finally /\textit{j/} responses are very common, as seen in the graph. Here as in the other stimulus words described, the stop was more readily identified when it occurred initially than finally.

Conclusion

The models and explanations given above are not intended as definitive, but they rather serve the purpose of illustrating some possible lines of thought which we hope can make justice to the outcomes of the tests. A more comprehensive report on the complete material is in preparation.

S. Öhman

Fig. II-1 Bottom: Stimulus word /a2pa2/. Half play-back speed from tape recorder. The scale along the time axis has 20 msec steps. At each step the word is cut into two pieces, one initial and one final.

Top: Response percentages corresponding to each of the cut points. Only systematic response trends are shown.
Fig. II-3  Stimulus word /a₂rt a₂/. See text under Fig. II-1.
Fig. II-4 Stimulus word /a₂k a₂/. See text under Fig. II-1.
Stimulus word /a₂b₂/. See text under Fig. II-1.
Fig. II-6 Stimulus word /$a_2d_2a_2$/. See text under Fig. II-1.
Fig. II-8 Stimulus word /a₂g a₂/. See text under Fig. II-1.