Quantitative measurement of the influence of acoustic cues on the perception of voiced plosives

W. Datschewit

Lehrstuhl für Datenverarbeitung, Technical University of Munich
Franz-Joseph-Str. 38, D-8000 Munich 40, West Germany

ABSTRACT

A series of experiments was conducted to investigate the role of formant transitions on the perception of the voiced plosives /b/, /d/ and /g/. In order to take into account the contextual influence of the surrounding vowels, vowel-consonant-vowel-units with the vowels /a:/, /u:/ and /i:/ were used as speech stimuli. Based on naturally spoken samples the stimuli were systematically manipulated with regard to different acoustic cues. The loci of the second and third formants (according to Delattre’s locus-theory) were varied. Subjects had to estimate their perception of the different plosives. The results show that formant transitions affect the perception very much, but F2-locus alone does not discriminate the different voiced plosives. A combination of F2- and F3-loci may be efficient cues to distinguish between /b/, /d/ and /g/.

1. INTRODUCTION

The purpose of the reported experiments was to obtain a functional relationship between the perception of a plosive and the physical dimension of an investigated acoustic cue. This should be expressed with a direct and quantitative measure for the perception of a sound. It will be referred to below as the strength of the perception of a sound. This way of scaling the perception is different to the many traditional forced-choice-experiments, where subjects had to perform an identification task between different categories. In contrast, in the experiments reported here they have to scale the degree of perception - the strength - of one sound in a range between 0 and 100. The obtained functions between strength and acoustic cue may be subsequently used in a rule-based automatic speech recognition system.

In an earlier study about the perception of voiced plosives [2], the results showed that the burst seems to have less influence on the perception of voiced plosives than it had with their unvoiced counterparts, whereas the duration of the closure and formant transitions affect the perception very much. Formant transitions might be a possible cue for identification of the place of articulation. These findings are emphasized by other experimental results as reported by Pols [6]. There he states, that 'vocalic transition seems to be a very relevant property of speech, since the vocalic transition is probably more resistant to sloppy pronunciation and masking than the burst itself'. In another study Crystal and House [1] reported that less than 50 % of the stops they analyzed in connected speech, contained releases. Lindholm et al. [5] report: '...formant information was the dominant factor in determining the identification responses of normal-hearing listeners.'

These reasons made it seem appropriate to concentrate on the formant transitions. In the experiments presented here, research was focused on the loci of the second and third formant, as it was defined by Delattre in [3]. According to Delattre the term locus means a 'point on the frequency scale at which a transition begins or to which it may be assumed to point'. This point should be characteristic for each consonant and should therefore have a fixed context-independent frequency position. Delattre postulated a locus at 700 Hz for /b/ and at 1800 Hz for /d/. For /g/ he found a locus at 3000 Hz, but for following back-rounded vowels such as /u:/ at about 700 Hz.

2. EXPERIMENTS

All experiments were based on VCV-units with the three voiced plosives /b/, /d/ and /g/ in each of the vowel-contexts /a:/, /u:/ and /i:. These three vowels were selected because they represent the corners of the vowel-triangle in German speech, thereby containing all extreme cases. Original speech-sounds were spoken by one male speaker, lowpass-filtered at 6.5 kHz and recorded with a 12-bit-ADC on a µ-VAX-II-computer; the sampling rate was 16 kHz.

In each experiment one of the original vowel-plosive-vowel-sounds was selected as basic material, LPC-analyzed and the VCV-unit systematically manipulated with respect to one acoustic cue. A continuum of stimuli along the possible physical dimension of the cue was designed, where each sound contained all properties of the original one, except the manipulated cue. For this reason all functions represent the influence of a cue on condition that all other cues are contained in the VCV-unit. An experiment consisted of a series of pairs of such vowel-consonant-vowel-stimuli. The first sound of each pair, the anchor, represented the original (unmanipulated), but also LPC-synthesized signal, the second stimulus was the manipulated version. All manipulations were done with a speech-editor based on a pitch-synchronous LPC-analysis-synthesis-system on a µ-VAX-II-computer. Pitch analysis was performed by evaluating the laryngograph signal.

Sixteen male and female subjects, all of them pupils or students, performed the different tests twice. During each experiment they had to judge the stimuli, which were presented in randomized order with 8 repetitions via earphones. Their task was to scale the strength of the second (manipulated) stimulus in percent compared to the original plosive, which had a relative value of 100. If the strength of the second stimulus is less than that of the anchor, subjects scaled it with a value less than 100. A scaling value of 0 % indicates no perception of the original plosive at all.
3. RESULTS

The results of the experiments are presented as curves with the strength of the stimulus shown on the y-axis and the locus of the stimulus on the x-axis. In the figures the median (denoted with a circle "o") and the interquartile range (denoted by a dash-dotted line) for each stimulus from all subjects and all repetitions are plotted.

3.1 Locus of second formant

3.1.1 vowel-context /a:/

Fig. 1 shows the results of the experiments with the F2-locus in /a:/-context. For the plosive /b/ one can see that for F2-loci below 1.4 kHz subjects estimated a strength of 100, i.e. full recognition, and decreased their answers with increasing F2-locus down to nearly 0 % /b/-strength. On the other hand /g/-strength increases with higher F2-locus values and above 1.4 kHz there is full 100%-perception of the sound. The perception of /d/ is best for F2-Locus between 1.1 and 1.9 kHz with a 100%-peak at about 1.7 kHz. Since this lies within the region of 100 % /g/-strength, F2-locus cannot act as single cue to divide /d/ and /g/.

3.1.2 vowel-context /u:/

In the context /u/ F2-Loci were manipulated in a band between 0.3 and 2.1 kHz. There seems to be a kind of break in all curves of fig. 2 for values of F2-loci between 0.9 and 1.1 kHz. For values higher than 0.9 kHz the strength of /d/ increases, whereas the /b/- and /g/-strength decrease rapidly. It is obvious that /b/ and /g/ cannot be separated by their F2-loci, as both curves show 100%-strength around the same frequency.
3.1.3 vowel-context /i:/

Figure 3 shows the results for perception of the three voiced plosives in /i:/-context. Examining the ranges having a strength of 100%, one obtains three different regions for the three different plosives: for /b/ the F2-Locus is in a range between 0.5 and 1.3 kHz, afterwards subjects tend to decrease their answers. For /d/ this area is between 1.3 and 2.1 kHz, whereas for /g/ 100%-strength ranges from 2.1 kHz upwards.

![Graph showing /b/-, /d/- and /g/-strength as function of F2-locus in /i:/-context](image)

3.2. Locus of third formant

As mentioned above F2-locus is no sufficient cue to distinguish between /d/ and /g/ in /a:/-context. In another experiment F3-loci of the stimuli /a:da:/ and /a:ga:/ were investigated. The results are presented in fig. 4. The more F3-locus is moved away from the original /a:da:/-locus at 2.9 kHz, the more /d/-strength decreases till it reaches 0% at 2.1 kHz. At 2.1 kHz lies the original value of the F3-locus of /a:ga:/; increasing this value to 2.9 kHz results in a decreasing perception of /g/, but the effect is very small.

![Graph showing /d/- and /g/-strength as function of F3-locus in /a:/-context](image)

4. CONCLUSION

Table 1 lists the frequency bands for F2-loci in which 100%-perception was reached for all nine conducted experiments. In addition the values of the second formant of the following vowel are listed. The data show that only in the vowel-context /i:/ might F2-locus act as a single cue for distinction between place of articulation, but not with the vowels /a:/ and /u:/.

A closer look at the results for the velar plosive /g/ in different vowel-contexts shows that the perception of /g/ seems to depend very strongly on the vocalic context. Best recognition of /g/ in /a:/-context is reached with an F2-locus between 0.5 and 0.9 kHz, whereas F2-loci have to be above 1.4 kHz for /a:/-context and 2.1 kHz for /i:/-context. These values are quite similar to those of the second formants of the neighbouring vowels, as listed in table 1. This indicates that there are no big transitional effects for the velar plosive /g/ and therefore no context-independent F2-locus can be found. This opinion is supported by the results of Gay [4]. In a study on articulatory movements in VCV-units he has demonstrated that velar stops are most sensitive to coarticulation effects because the tongue is involved as the primary articulator in the production of velars. Also the fin-
The experiments on the F3-locus in /a:i/ and /a:ga:/ showed that F3-transitions can influence perception too. A combination of F2- and F3-loci might make it possible to distinguish between the different voiced plosives. Further work on this subject is currently in process.

### Table 1: ranges of F2-loci with 100% strength for /b/, /d/ and /g/

<table>
<thead>
<tr>
<th>vowel context</th>
<th>formant 2 [kHz]</th>
<th>range of F2-loci with 100% strength [kHz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>/u:/</td>
<td>0.7</td>
<td>0.5-0.9 / 0.8-1.7</td>
</tr>
<tr>
<td>/a:/</td>
<td>1.3</td>
<td>1.0-1.5 / 1.3-2.1</td>
</tr>
<tr>
<td>/i:/</td>
<td>2.4</td>
<td>&gt;1.1 / &gt;1.5</td>
</tr>
</tbody>
</table>

These findings lead to the opinion that both articulation and perception of velar plosives is strongly influenced by coarticulation. In the case of the voiced stop /g/ nearly no transition is detectable and therefore there cannot exist an explicit consonant-independent locus for the transition of the second formant. This finding is in contrast to Delattre’s locus theory.

For the plosive /b/ perception is best for values of the F2-locus lower than those of /d/ and /g/. This is the fact for all three investigated vowel surroundings. The upper breakpoint for 100% perception rises with increasing F2-value of the following vowel. For /u:/ this value is 0.9 kHz and increases to 1.5 kHz for /i:/ This finding may also be affected by coarticulation. Delattre’s context-independent F2-locus at 700 Hz is within the region from table 1 but could not be found in such an explicit manner.

The consonant /d/ seems to be recognized best with a F2-locus in midfrequency-regions. This finding is not contrary to Delattre, who found a F2-locus at about 1600 Hz but again it is not possible to find such an explicit frequency point. In the context /a:/ the frequency band with 100% /d/-strength is very small because the gap between /b/ and /g/-strength is also small. This gap increases for /i:/ likewise the region for /d/-perception gets larger. In /u:/ context /b/- and /g/-strength is maximum between 0.5 and 0.9 kHz. Because of that /d/-recognition is 100 % for all loci above 1.1 kHz.

### REFERENCES


