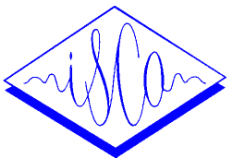

ISCA ITRW on
Spoken Word Access Processes (SWAP)
Nijmegen, The Netherlands
May 29-31, 2000

LIPREADING AND THE COMPENSATION FOR COARTICULATION MECHANISM

Jean Vroomen & Beatrice de Gelder
Tilburg University, Tilburg, The Netherlands

ABSTRACT

Listeners compensate for coarticulatory influences of one speech sound on another. We examined whether lipread information penetrates this perceptual compensation mechanism. Experiment 1 replicated that when an /as/ or /aʃ / sound preceded a /ta/-/ka/ continuum, more velar stops were perceived in the context of /as/ ([1]). Experiments 2 and 3 investigated whether the same phoneme boundary shift would be obtained when the context was lipread instead of heard. An ambiguous sound between /as/ and /aʃ/ was dubbed on the video of a speaker articulating /as/ or /aʃ /. Subjects relied on the lipread information when identifying the ambiguous fricative sound, but there was no boundary shift in the following /ta/-/ka/ continuum. These results indicate that biasing of the fricative and compensation for coarticulation can be dissociated.

1. INTRODUCTION

In psychological research, it has been a tacit assumption that tasks involving the identification and/or discrimination of phonemes (or other sub-lexical elements) tap much of the same processes involved in the processing of natural speech and recognizing words. This idea, though, may be wrong. In our study, we present an example of a dissociation between phoneme identification and the so-called compensation for coarticulation mechanism. It has been shown that ambiguous stops between /t/ and /k/ tend to be heard as /k/ after /s/ and as /t/ after /ʃ/ ([1]). The explanation given for this finding is that the perceptual system compensates for coarticulation in production. During production of a fricative-stop pair, the place of articulation of the stop is supposed to shift towards that of the fricative. A /k/ in the context of /s/ is thus supposed to be more anterior than in neutral context. The perceptual system compensates for this by adjusting the category boundary such that an 'anterior' /k/ will nevetheless be heard as a velar /k/ when preceded by /s/. This effect has been replicated with an ambiguous fricative (henceforth /?/) midway between /s/ and /ʃ/ that was embedded in a lexical context biassing is towards /s/ or /ʃ/ ([2]). More /k/ responses were obtained after *christma?* than after *fooli?*, as if listeners had heard /s/ in *christma?* and /ʃ/ in *fooli?*.

Subsequent research, though, complicated the picture. Thus, ([3]) argued that ([2]) had confounded lexical context with transitional phoneme probablities (TPs). An /s/ is more likely to follow /ə/ (as in *christmas*), and /ʃ/ is more likely to follow /I/ (as in *foolish*). When words were controlled for TPs (e.g., as in *juice* and *bush*), ([3]) *only* found biases in the labeling of the fricative (listeners were lexically biased reporting *bush* when hearing *bu?* and *juice* when hearing *jui?*), but no subsequent shift in a /t/-/k/ continuum that immediately followed these words. In contrast, nonwords with TPs that biased the fricative either toward /s/ (as in *mi?*) or /ʃ/ (as in *nai?*) had an effect on labeling of *both* fricative and stop. ([3]) argued that models like TRACE ([4]) have problems accounting for such a dissociation because TRACE has a single phoneme level that serves two functions, namely: 1) as outlet for phoneme decisions, and 2) as an intermediate processing stage between features and words. Biases in fricative identification and compensation for coarticulation should in such architecture affect the same level and the effects should therefore come and go hand-in-hand, without a dissociation.

However, the interpretation of ([3]) may also be confounded because it is known that compensation for coarticulation may at least partly be due to an auditory contrast effect since it happens with non-speech precursors, and quails apparently show the original effect ([5)]. Since the vowels in ([3]) were not matched (short vowel in *bu?* and long vowel in *jui?*), the dissociation may be based on different auditory contrast effect. In the present study, we avoided this difficulty by using lipread information to bias the identification of the fricative.

2. EXPERIMENT 1

To ensure that there were no stimulus confounds, we wanted to replicate that when an /as/ or /aʃ / sound precedes a /ta/-/ka/ continuum, more velar stops are perceived in the context of /as/.

3. METHOD

Subjects. Ten native speakers of Dutch took part in the experiment.

Stimuli. A nine-step /ta/-/ka/ continuum was created that was preceded by /as/or /aʃ/. Natural tokens of /ta/, /ka/, /as/, /aʃ/, /asta/, /aska/, /aʃta/ and /aʃka/ were produced by a male native speaker of Dutch and recorded in a sound-damped booth. The speaker was also recorded on Sony U-Matic video for the audio-visual experiments. The auditory stimuli were low-pass filtered at 9.8 kHz, and then digitised at 20 kHz with a 12 bit analog-to-digital converter. The nine-step /ta/-/ka/ continuum was created by adding in proportion the amplitudes of the waveforms of the first 55 ms of a /t/ and /k/. The proportion of the /k/ increased from .2 to 1.0 in nine steps of .1. such that the /t/ endpoint contained .2 /k/ and .8 /t/, while the /k/ endpoint

contained 1.0 /k/ and 0. /t/. The vocalic portion of the /ka/ sound was appended at the zero crossing of the so created stimuli. Pilot tests showed that the phoneme boundary was about in the middle of the continuum.

From naturally produced /as/ and /aʃ/ tokens, the /s/ and /ʃ/ were cut out with the cut being made at a zero crossing. Then, the /ʃ/ was made of equal length with the /s/ by cutting out the final 64 msec, making each stimulus last 175 msec. The /s/ and /ʃ/ were appended at a zero crossing to the vocalic part of /as/ to make /as/ and /aʃ/ tokens. Finally, the /as/, and /aʃ/ tokens were prepended to the /ta/-/ka/ continuum with a 150 msec silence interval between them. The stimuli were natural sounding with no discontinuities in the waveform that were audible in the form of clicks.

Procedure. The stimuli were presented in two blocks of 90 stimuli each. Within each block, the context sound of the /ta/-/ka/ continuum was either /as/ or /aʃ/. The presentation order of the blocks was counterbalanced across subjects. Each block was made up of 10 lists of random sequences of the nine possible /ta/-/ka/ stimuli. A three-second interval separated the individual stimuli, with longer pauses between the sequences. Each block was preceded by a warming-up session in which each stimulus was presented once in random order.

Each subject participated in two sessions of 8 minutes each. The stimuli were low-pass filtered at 9.8 kHz and then stored for presentation on a digital audiotape via a Philips 850 DAT recorder. Subjects were asked to respond orally whether they heard /ta/ or /ka/. The responses were written down by the experimenter.

4. RESULTS

Figure 1 shows the identification functions of the /asta/-/aska/ and /aʃta/-/aʃka/ continuum. The effect was exactly as predicted: an /as/ context increased the percentage of /ka/ responses if compared with /aʃ/ because the phoneme boundary of the continuum was shifted toward the /ta/-end.

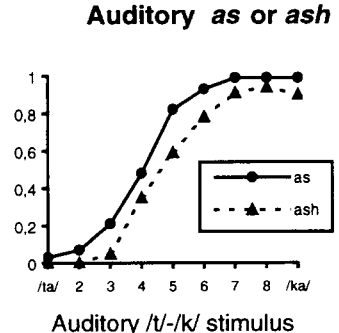

Figure 1. The effect of an auditory /s/ or /ʃ/ sound on the perceived place of articulation of a following stop consonant.

The proportion of /ka/ responses was submitted to a 2 (context) x 9 (auditory levels) analysis of variance (ANOVA). As intended, the identification function raised from left to right as the target stimuli varied along the /ta/-/ka/ continuum, $F(8,72) = 106.9$, $p < .001$. The effect of context was significant because more /ka/ responses were given when /as/ was the context, $F(1,9) = 10.80$, $p < .01$. The interaction between context and the auditory information was significant indicating that the context effect was largest for the ambiguous /ta/-/ka/ tokens, $F(8,72) = 2.12$, $p < .05$. Experiment 1 thus replicated the basic compensation for articulation effect and thus shows that there were no stimulus confounds.

5. EXPERIMENT 2

Experiment 2 tested whether lipread information about the fricative can influence labelling of the fricative and whether it does, in addition, affect labelling of the stop.

6. METHOD

Subjects. Ten new subjects were paid for their participation.

Stimuli. The nine stimuli from the /ta/-/ka/ continuum were now preceded by an /a?/. This sound track was then dubbed on the video of the speaker articulating either /aska/ or /aʃka/. The ambiguous /?/ was made in a similar way as the /ta/-/ka/ continuum, namely by adding up the waveforms of a /s/ and /ʃ/. The tokens were spliced out from a naturally produced /as/ and /aʃ/ and they were made of equal length. By taking a proportion of .8 /s/ and .2 /ʃ/ a /?/ sound was made which, in a pilot test, was chosen as /s/ about 50% of the time. The rest of /a?ta/-/a?ka/ tokens was exactly the same as in Experiment 1. The tokens were dubbed on the video recording of the speaker saying /aska/ or /aʃka/. During the pronunciation of /s/ in /aska/, the speaker's lips were retracted, and during the pronunciation of /aʃ/ in /aʃka/, the lips were rounded. The pronunciation of the /k/ phoneme in the video recording was visually indistinguishable from /t/ since both phonemes belong to the same viseme cluster. Lipreading thus distinguished between /s/ or /ʃ/, but not between /t/ or /k/. The audio was synchronised with the video at the onset of the initial vocalic portion of the stimuli. This resulted in natural looking stimulus events in which no desynchronisation of the audio and video could be detected.

Procedure. The stimuli were presented in 10 lists of random sequences of the 18 possible stimuli (nine auditory levels and two visual articulations). There was a 4-sec pause between the successive stimulus events and a 10-sec pause between two sequences. Prior to testing, subjects were given 10 practice trials. Subjects were tested individually in a sound-damped booth. They viewed a 63-cm television monitor that presented both the auditory and visual dimensions of the speech stimuli. Subjects were seated at a distance of about two meter from the monitor. The audio was set at a comfortable listening level with the peak of the amplitude at approximately 61 dB-A. Testing lasted about 20 min. Subjects were asked to respond orally whether they

heard /asta/, /aska/, /aʃ ta/ or /aʃ ka/. Responses were written down by the experimenter.

7. RESULTS

Figure 2 displays the proportion of /ka/ responses, (i.e. the proportion of /aska/ plus /aʃ ka/ responses) as a function of the auditory levels, separately for the visual /as/ or /aʃ / context. An ANOVA performed on the proportion of /ka/ responses indicated that the identification functions raised from left to right as the target stimuli varied from /ta/ to /ka/, $F(8,72) = 169.05$, $p < .001$. The visual pronunciation of /as/ or /aʃ / had a marginally significant effect on the identification of the /ta/-/ka/ stimuli, $F(1,9) = 4.45$, $p = .064$. The interaction among the visual and auditory dimensions of the stimuli was not significant, $F(8,72) < 1$. However, although there was a small effect of the lipread context, it was in the opposite direction of what was expected. Instead of /as/, it was /aʃ / that increased the proportion of /ka/ responses from .463 for the /as/ context to .483 for /aʃ/.

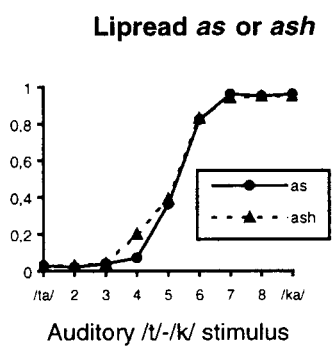

Figure 2. The effect of a lipread /s/ or /ʃ/ on the perceived place of articulation of a following stop consonant.

Performing a logit transformation on the data showed that the mean phoneme boundary in the /as/ context was 5.31 stimulus units while it was 5.15 in the /aʃ/ context, $F(1,9) = 4.09$, $p = .074$. There was thus no phoneme-boundary shift towards the /ta/-end of the continuum when a visual /as/ pronunciation was seen. Individual analysis confirmed that only two out of ten subjects had a shift in the /ta/ direction.

An explanation for the failure of a visual /as/ to increase the proportion of /ka/ responses could have been that the lipread information did not affect the identification of the fricative /s/ or /ʃ/. Since we asked our subjects to report both context and target stimulus, we were able examine this. When a /s/ was seen, the proportion of /s/ responses (i.e. /asta/ plus /aska/) was .510 while the proportion of /s/ responses was .292 when a /ʃ/ was seen, $F(1,9) = 8.14$, $p < .02$. Lipreading did thus contribute to the identification of the fricative, but not the stop. One might, however, argue that the effect of lipreading was too small for an effect to be observed on the following /ta/-/ka/ continuum, since only 61 percent of the context was correctly identified. We therefore tried to magnify the influence of lipreading in a second analysis by discarding all responses that were not in agreement with the provided lipread information. In the case a visual /s/ was seen, all responses with /ʃ/ were rejected (i.e. /aʃta/ and /aʃka/) and similarly, when a /ʃ/ was seen, all responses with /s/ (i.e. /asta/ and /aska/) were rejected. Thus, in this analysis the lipread context was always correctly identified. An ANOVA performed on these corrected measures indicated that the auditory variable was significant, $F(8,56) = 99.28$, $p < .001$, but there was again neither an effect of lipreading, $F(1,7) < 1$, $p = NS$, nor was the interaction significant, $F(8,56) = < 1$, NS. (Two subjects had to be discarded because the appropriate proportions could not be computed unless a division was made by zero). The mean phoneme boundary was 5.31 stimulus units in the /as/ context, while it was 4.95 in the /aʃ/ context, $F(1,7) = 1.53$, $p = .256$. Thus, the perceived fricative did not exert any influence on the phoneme boundary of the following /ta/-/ka/ continuum, even though in this analysis all subjects reported consistently /as/ or /aʃ / depending on the context.

We were, however, still not satisfied with these results because in the previous analysis about 39 percent of the data had to be discarded. Moreover, it might be the case that coarticulatory effects would have been found if the /ta/-/ka/ tokens were more ambiguous since compensation effects are usually largest for the most ambiguous stimuli. In the next experiment, we therefore decreased the loudness. It was hoped that this would boost the contribution of vision in the identification of the context and that the auditory information of the /ta/-/ka/ tokens would become more ambiguous.

8. EXPERIMENT 3

The same stimuli were used as in the previous experiment, except that the loudness was lowered so as to increase the relative contribution of the lipread information.

9. METHOD

Subjects. Seven new subjects were paid for their participation.

Stimuli and Procedure. The same as in Experiment 2, except that the loudness of the audio-visual stimuli was set at a level in which it became hard to distinguish /t/ from the /k/ (the peak of the amplitude was at 51 dB-A).

10. RESULTS

We first investigated whether subjects relied on vision in the identification of the context sound. The proportion of /s/ responses when /s/ was seen was .84, whereas it was .21 when a /ʃ/ was seen, $F(1,6) = 39.23$, $p < .001$. Thus, lipreading did contribute to the identification of the

fricative (81 percent correct) and the visual influence was boosted if compared with Experiment 2 (which was 61 percent correct). Moreover, as can be seen in Figure 3, identification of the /ta/-/ka/ tokens was much more difficult since the usual steep categorical identification functions were not obtained. However, there was again no effect of the visual information on the /ta/-/ka/ continuum. Lipreading an /s/ did thus not elevate the proportion of /ka/ responses.

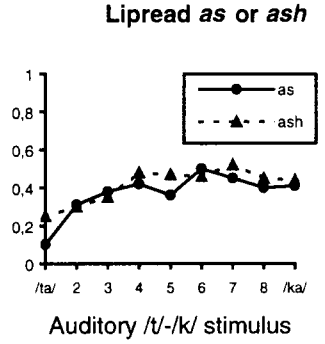

Figure 3. The effect of a lipread /s/ or /ʃ/ on the perceived place of articulation of a following stop consonant.

An ANOVA performed on the proportion of /ka/ responses confirmed that there was an auditory effect, $F(8,48) = 3.16$, $p < .006$, and a small nonsignificant effect of vision, $F(1,6) = 4.12$, $p = .09$, but again the latter effect was in the opposite direction of that of heard stimuli. When an /s/ was lipread, the proportion of /k/ responses was .39 whereas it was .44 when a /ʃ/ was lipread. (We only report responses in which the context was correctly identified. Including responses in which the context was incorrectly lipread did not change the results). Thus, although we succeeded in increasing the contribution of vision and decreasing the quality of the /t/-/k/ tokens, there was still no effect of coarticulation.

11. DISCUSSION

A tacit assumption in psychological research is that tasks involving the identification and/or discrimination of phonemes (or other sub-lexical elements) tap much of the same processes involved in the processing of natural speech and recognizing words. This idea is most likely wrong. The present results show that lipread information can have an effect on identification of a fricative, but not on compensation for coarticulation. This finding is in line with ([3]) who observed a similar effect with lexical information. The assumption made by TRACE and many others that phoneme identification is a direct reflection of one of the intermediate stages of speech processing therefore needs to be reconsidered.

Of course, this is not new, and many studies have pointed out a similar contrast. From the neuropsychological literature it is known that the ability to identify CV syllables is a poor predictor of auditory comprehension deficits. On the basis of this, ([6]) speculate that there is, akin to vision, a ventral cortical pathway for word recognition and a dorsal pathway for sublexical discrimination and identification. In ([7]) there is also a strict distinction between phoneme units and decision units (for comments, see ([8]). Moreover, it is well-known illiterates, Chinese, or dyslexics have problems with tasks requiring manipulation of speech segments at a subsyllabic level, while at the same time there is no obvious deficit in spoken word recognition ([9,10]). Apparently, then, spoken word recognition is thus less transparent than psycholinguistic tasks may suggest.

12. REFERENCES

[1] Mann, V.A. & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America,* **69**, 548-558.

[2] Elman, J. L. & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language,* **27**, 143-165.

[3] Pitt, M. A. & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language,* **39**, 347-370.

[4] McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology,* **18**, 1-86.

[5] Lotto, A. J. & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics,* **60**, 602-619.

[6] Hickok, G. & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*, 131-138.

[7] Norris, D., McQueen, J. M. & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Brain and Behavioral Sciences,* **23**.

[8] Vroomen, J. & de Gelder, B. (in press). Why not model spoken word recognition instead of phoneme monitoring? *Brain and Behavioral Sciences.*

[9] Bertelson, P., de Gelder, B., Tfouni, L. & Morais, J. (1989). The metaphonological abilities of adult illiterates: new evidence of heterogeneity. *The European Journal of Cognitive Psychology,* **1**, 239-250.

[10] de Gelder, B., Vroomen, J. & Bertelson, P. (1993). Effects of alphabetic reading competence on language representation in bilingual Chinese subjects. *Psychological Research,* **55**, 315-321.