THE McGURK EFFECT AND !Xóõ CLICKS

Anthony Traill
University of the Witwatersrand, Johannesburg

ABSTRACT

Visual information can influence the perception of speech sounds, particularly concerning place of articulation (POA). However, the perception of manner features relies on the auditory modality [3]. This study explores these factors with the five basic click consonants of !Xóõ, a San language. The clicks utilize labial or corono-dorsal POA and thus allow for the replication of POA-by-visual-effect-manipulation used in studies employing conventional speech sounds. Since click contrasts rely on a number of salient manner distinctions involving spectral emphasis, intensity, and noisiness, it was hypothesized that the auditory stimulus would over-ride conflicting visual cues, particularly where labial POA was not cross classified for the abrupt click types. Results provide weak confirmation for the hypothesis when auditorily salient features over-ride conflicting visual cues. In other cases visual cues strongly influence perception.

1. INTRODUCTION

It is a well-known fact that in speech perception, the auditory modality can be effected by the visual modality. Studies have demonstrated how visual information conveyed chiefly by the shape of the mouth can radically change the perception of sounds, particularly regarding place of articulation (POA). Features such as manner and voicing, however, rely on the auditory modality [3]. Although these studies rely on matching auditory and visual information in ways that do not occur naturally, their results are nevertheless of interest for a theory of speech perception which tries to integrate multiple sources of information [1,3]. The conclusions have been based on the investigation of natural and synthetic stimuli drawn from sound types produced on the pulmonic airstream mechanism. In this paper, the role of the auditory and visual modalities is assessed in the perception of the five natural clicks in !Xóõ.

This may be of interest for a number of reasons. Firstly, since the clicks involve either labial-dorsal ([ʘ]) or corono-dorsal ([ǀ ǁ ǃ ǂ]) double articulations, they allow the experimenter to replicate and manipulate the visual factors associated with labial and coronal POA used in other studies. Secondly, clicks have very distinctive auditory properties: certain clicks are unusually intense sounds and all of them cross-classify according to two other manner features, namely, whether they are impulse-like (ǃ ǂ) or not (ʘ ǀ ǁ) and whether they have a high frequency spectral emphasis (ǀ ǂ) or not (ǃ ǁ) ([ʘ] has spectral feature of both classes). It was hypothesized that the auditory modality exclusively would determine the perception of these manner features. Since there is no high intensity impulse-like click with labial POA, it was further hypothesized that when these clicks (i.e. [ǃ ǂ]) were paired with labial visual information, the auditory stimulus would resolve the anomalous situation.

2. EXPERIMENTAL METHOD

The experimental technique used to investigate these hypotheses was that of the classical McGurk and Macdonald study [2], in which a videotape is produced of all the permutations of visual and audio information for the class of consonants of interest. This is played to subjects who judge which sound they heard. In this application, simultaneous video and DAT audio recordings were made of a native speaker of !Xóõ producing the syllables [gʘa gǀa gǁa gǃa gǂa ba da ga]. A videotape of all the visual and audio permutations of these utterances was prepared and was played three times to 10 native listeners of !Xóõ. They were asked to judge what they heard by repeating the stimulus syllable. The non-click syllables were included as a control. The perception of the audio stimuli was tested independently as well (10 subjects, six tokens for each stimulus).

3. RESULTS

3.1 Audio Test. Testing the audio stimuli independently proved to be critically important because the results revealed unexpected areas of difficulty in the perception of certain sounds. This had an unfortunate effect on the general results of the experiment because these stimuli were used for the audio track of the videotape. Nevertheless there were tendencies which provided some interesting results.

Perceived as (%) N=60									
Stimulus	b	d	g	gʘ	gǀ	gǁ	gǃ	gǂ	ŋg
b	**93**		7						
d		**20**	67		5	7			1
g			**95**						1
gʘ				**53**	47				
gǀ					**100**				
gǁ						**100**			
gǃ	2		5				**90**		3
gǂ			**37**				7	**55**	1

Table 1. Results of the Audio Test

While the poor perception of the labial click is to be expected because of its ambiguous acoustic properties, the poor perception of [d] and its misperception mainly as [g] is puzzling, as is the misperception of [gǂ] as a non-click. Both [d] and [gǂ] are otherwise extremely salient sounds, and there is nothing atypical in the spectra for these stimuli to explain the responses. The responses to [d] and [gǂ] in the experiment will, therefore, at beast be difficult to interpret. Moreover, the misperception of [d] as [g] confounds any attempt to replicate certain details of the original McGurk and MacDonald experiment.

3.2 Results from the experimental videotape. The main results of the perception test with both visual and audio cues may be summarized as follows.

3.2.1 General comments. Generally, the segments perceived poorly in the Audio Test remain poorly perceived even when visual information is available. Thus [d] is only perceived 23% correctly (cf. 20% in the Audio Test) and [gǂ] is only perceived 20% correctly (cf. 55% in the Audio Test). This confirms that the

audio stimuli for these sounds were ambiguous. The exceptional case is the labial click which improves its score from 53% correct on the Audi Test to 97% correct (the single misperception as [ǀ] represents one token from a listener who perceived the click correctly in the two other trials). This result dramatically demonstrates the role of visual information in the perception of this click.

3.2.2. **Perception of the manner feature "click".** There is no general visual cue to the manner distinction "click" vs. "non-click". This may seem to be a superfluous remark, since, as noted above, manner features rely on acoustic cues. However, clicks have some unique articulatory properties and it is possible that these may be accompanied by visual cues to which native speakers are attuned. If we examine the number of category misperceptions, under all visual conditions, it is clear that clicks are overwhelmingly heard as clicks and non-clicks as non-clicks. The exceptions cluster round [d] and [g‡], the poorly perceived segments in the Audio Test. For example, when [g‡] is paired with visual [b] it is perceived as the labial click [gʘ] in only 5/30 responses, but when it is paired with visual [gʘ] it is perceived as the labial click in 26/30 responses. It is possible that the slight pursing of the lips which accompanies the suction for the labial click used in the experimental token is the visual cue which determines this response.

The !Xóõ labials [b] and [gʘ] offer a good opportunity to assess whether the ±click manner distinction may have visual cues. Table 2 shows a weak effect of the labial click articulation on the perception of the non-click consonant.

Audio	Perceived as	%	Visual
b	b	100	b
	b	87	gʘ
	gʘ	13	gʘ
gʘ	gʘ	97	gʘ
	gʘ	93	b
	b	0	b

Table 2. Visual labial effects on [b] and [gʘ]

These are not the only cases where visual differences between [b] and the labial click appear to influence perception (see **3.3.4**).

3.3.3 The perception of [gʘ] and [gǀ]. The perception of these low intensity, noisy clicks is readily influenced by visual information. Although [gǀ] is a salient sound in the Audio Test (100% correct), it is frequently perceived as the labial click [gʘ] with visual labial (83% of the time with [gʘ] labial, 60% with [b] labial); [gʘ] is perceived as [gǀ] with [d] visual (73%), [g] visual (93%) and oral click visual (i.e. [gǀ gǂ g! g‡]) (76%). Examination of the acoustic spectra for the two stimuli shows similarities, but [gʘ] has a flatter spectrum and its noiseburst is weaker. These differences are evidently not very salient under these experimental conditions.

3.3.4 The role of salient acoustic features. The hypothesis that salient acoustic features of clicks may be expected to over-rule contradictory visual information may be tested with the results for pairs like acoustically intense [!] and [ǀ] paired with visual labial (as pointed out above, [‡] cannot be used for this because of its ambiguous auditory status). The results are mixed: pairing [g!] with visual labial clearly caused perceptual confusion, seen firstly in the number of different response types (7 types with [gʘ] visual, 5 types with [b] visual) and secondly in some atypical, mixed, non-native sequences such as [b! gb bg]. Although the majority of responses is [g!] (33% with visual [b], 53% with visual [gʘ]), in which the audio stimulus resolves the bizarre audio-visual pairing, 16% is mixed, with 4 tokens of [b!] and one each of [gb bg]. Counter to the hypothesis, the labial visual cue over-rides the [g!] audio stimulus in the [gʘ] responses (27% with visual [b], 30% with visual [gʘ]). These labial click responses represent a perceptual compromise because [gʘ] lacks any of the acoustic properties of the stimulus (except perhaps the click feature). However, in six tokens (four with visual labial [b] and two with visual labial [gʘ]) the visual stimulus completely over-rides the auditory stimulus as seen in the [b] responses.

The response types to audio [gǀ] paired with visual labial, are much more constrained. With visual [b] the majority of [gǀ] responses (57%) ignores the visual cue; the only other response is [gʘ] (43%) in which the visual cue over-rides the salient auditory one. The saliency of [gǀ] lies in its intensity compared to [gʘ] (approximately +10dB) and this probably determines the majority response. One would therefore expect a labial click with this intensity to be auditorily anomalous; nevertheless the visual cue exploits the phonological feature of non-abruptness which cross-classifies the two clicks, resulting in the [gʘ] responses. With visual labial click articulation, the same pair of responses occurs, but in this case the visual cue is stronger: [gǀ] is perceived 40% of the time, and [gʘ] 60% of the time. We attribute the higher number of [gʘ] responses to a distinctive visual cue associate with the labial click, most likely slight lip pursing, which accompanies the noisy click.

Although these results for [ǀ] and [!] may be seen as providing weak support for the hypothesis concerning the role of auditory salience, the results for the lateral, labial and dental clicks show that phonological categorization is also a factor. The fact that both labial and oral POA for clicks are cross-classified for the noisy release, but the labial POA is not cross-classified for the feature of abrupt release (i.e. the clicks [! ‡]), means that labial visual cues can have a stronger influence on the perception of [ʘ ǀ].

3.3.5 The perception of non-clicks. The findings in (4) must be compared with the behaviour of non-clicks under comparable conditions so as to establish whether the role of visual cues in the perception of non-clicks differs.

The problem with the ambiguity of the [d] audio stimulus necessarily focuses the discussion on the perception of [b g].

As has been reported for other studies, the perception of [b] is strongly influenced by visual cues [3]. With the appropriate labial cue it is perceived perfectly and it is perceived very well with the visual cue for the labial click (87%). However, with the visual cues for oral articulations (coronal or dorsal), it is perceived correctly only once. In these cases it is perceived most often as [g] (100% with dorsal visual, and 82% with oral click visual). Notice that the pairing [b] with dorsal visual does not yield the expected middle articulation [d] as it did in the McGurk and MacDonald studies [1, 2]. However, the click version of this pairing (i.e. audio [gʘ] paired with visual dorsal articulation) does, giving the response with a dental click. The basis of this is straightforward because these two clicks do indeed cross-classify with intensity and noisiness and the spectra for the tokens used in the experiment are similar. We return to this below.

[g] is clearly a very salient sound in this experiment because it has the lowest error rate under all visual conditions. The asymmetry noted in other experiments [3] between the effects of a dorsal articulation vs.labial articulation on perception is confirmed. Whereas the perception of [b] has high error rates of 100% incorrect with [g] dorsal visual, and 82% incorrect with

coronal (click) visual, [g] has much lower error rates of only 33% incorrect with [b]-labial visual and of 37% incorrect with [ʘ] labial visual. Labial articulations thus have less effect on the perception of [g] than intra-oral articulations have on the perception of [b]. We return here to the exception to this, noted above, involving the labial click [ʘ] and the dental click [ǀ]. In these two cases the effects are almost symmetrical: intra-oral visual cues (coronal and dorsal) induce an average error rate of 87% incorrect in the perception of [gʘ] and labial visual cues (for [b] and [gʘ]) induce an average error rate of 72% incorrect in the perception of [gǀ]. The most obvious explanation for this is that the two clicks are acoustically similar despite their prominent articulatory differences, and they are therefore likely to behave in similar ways under these experimental conditions.

4. CONCLUSION

The hypothesis that clicks should largely resist the effects of conflicting visual information because the distinctions between them rely on salient manner of articulation contrasts, is partly confirmed. Despite the limitations introduced by the ambiguity of two of the audio stimuli used in the experiment, some interesting tendencies can be seen in the data. While a truly anomalous pairing of auditory and visual information such as occurs when the impulse-like click [!] is paired with labial articulation may often be resolved auditorily, the prominent visual cue leads to resolutions with non-native sequences or to a lame compromise with the labial click [gʘ]. This distinctly unnatural result should be attributed to the artificial experimental conditions rather than to any natural phonetic linguistic processes. In most other cases, resolutions are determined by the extent to which acoustic features and POA cross-classify with one another. Where they do, visual information strongly influences perception. Finally, there is some evidence that subjects sometimes used the subtle visual difference between the articulation of a labial stop and a labial click to influence their judgements.

REFERENCES

[1] MacDonald, J. and H. McGurk. 1978. Visual influences on speech perception processes. *Perception and Psychophysics*, 24, 253-257.

[2] McGurk, H and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 164, 746-748.

[3] Massaro, D.W. 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale NJ, L. Erlbaum.