Perception Difficulties and Errors in Multimodal Speech: The Case of Consonants

Azra N. Ali
School of Computing and Engineering, University of Huddersfield,
Huddersfield, England
E-mail: a.n.ali@hud.ac.uk

+

ABSTRACT

This study focuses on the perception difficulties and errors of audiovisual presentations of talking heads, and examines the effect of noise and temporal misalignment of channels for two groups of consonants; onsets and codas in monosyllabic words. The study confirms that incongruence of audio and visual channels cases McGurk fusion even in the linguistic context of monosyllabic words. The results show that, in coda position, incongruent consonant stimuli elicit fusion more readily than those in the onset position. In all cases with incongruence, subjects are found to take more time to reach decisions than with congruent stimuli. It is argued that decision times – which we measure in two different ways – are indicators of perceptual difficulty experienced by subjects when interpreting a stimulus.

1. INTRODUCTION

Human communication is multimodal; information is conveyed through face and other gestures and speech sounds in combination. Although speech perception is usually considered an auditory process, studies have shown that visual information provided by the movements of a talker's mouth and face strongly influences what an observer perceives, even when the auditory signal is clear and unambiguous [1],[2],[3]. Therefore, people use both the acoustic and the visual modality to understand speech. Many people with a hearing impairment can understand speech by lip-reading, which shows that linguistic information, can be conveyed by vision [4]. When the audio channel is noisy, information from the visual channel significantly improves the accuracy of speech perception, as was demonstrated quantitatively by Sumby and Pollack [5]. Thus, congruent audio and video speech channels not only provide two independent sources of information, but do so complementarily: each is strong when the other is weak. Furthermore, the complementarity makes accurate speech perception more resistant to channel noise [6].

Evidence for strong interaction between audio and visual speech channels in human speech perception is found in the well-known McGurk effect [7]. If humans are presented with temporally aligned but conflicting audio and visual stimuli – now known as 'incongruent stimuli' - the perceived sound may differ from that present in either channel.

2. BIMODAL PERCEPTION OF CONSONANTS

Earlier researchers showed the extent to which McGurk effect can occur in syllables of the CV, for e.g. /ba/ and /ga/ or VCV types, for e.g. /aba/ and /aga/. This tells us very little in terms of which speech sounds are vulnerable to McGurk fusion. The study therefore focuses on English monosyllabic words with two contextual types of consonant slot: onset and coda. The aim of the study was primarily to provide a better understanding of the McGurk Effect phenomenon in monosyllabic words. Such issues are of considerable importance for the effective design and use of new multimedia applications involving audiovisual speech synthesis. Animated cartoons with talking heads are increasing used to represent software agents. The design of such animated talking agents demands linguistic knowledge in some phonetic detail [1, 4].

Two groups of contrasting word pairs were targeted; those with contrasting onset, and those with coda contrasts. The design of the experiment targeted pairs of words contrasting independently in POA, manner-class and voicing-state. Major consideration was given to selecting word pairs with the same contrasts in both coda and onset groups. The objective was to look for possible differences of fusion phenomena in coda and onset positions. The word pairs used had the same vowel nucleus in both members of the pair.

For comparison purposes, the consonant incongruent mononuclear word stimuli prepared from the above pairs were mixed with a few stimuli with incongruities in nuclear position. These stimuli included cases of both short and long incongruent vowels embedded between congruent stimuli: a non-branching onset /h/ and a simple coda /d/ [8]. The results showed, in fact, that though long vowels fuse much less readily than consonants, the short vowels are more prone to fusion than consonants. The study measured fusion rate as the proportion of incongruent stimuli eliciting a fusion response instead of a channel response. The

 ISBN 1-876346-48-5

observed rates in the experiments reported below were a lowest (15%) for long vowels, 47% for onset consonants, 59% for coda consonants and a highest 67% for short vowels.

2.1 MEASURING DECISION TIMES

The experiment also probed the hesitations of subjects *via* two measures of decision time. One was the total time taken by a subject to select a response to a stimulus from an open list of possible responses. The other was the number of replays used by the subject before reaching a decision about the stimulus. Both determined the perceived level of task complexity, and were logged automatically by the experimental control software. To verify that there are differences between onset and coda consonants, some incongruent stimuli by adding distracting noise to the audio channel: in fact, 'cocktail party' background chatter. Its effect was to increase decision-times, whichever measure was used.

3. METHOD

3.1 CREATING THE STIMULI

Video recordings were made of a male (aged 23 years) and a female (aged 22 years), both native speakers of British English articulating common English monosyllabic words. The video recordings were done inside a quiet, controlled Usability Lab using a standard 8 mm digital Sony Camcorder with built-in microphone for audio. To prepare the realigned stimuli words were grouped into contrasting pairs of consonants, see Table 1. The re-aligned clips, some with 'cocktail party' noise added acoustically and a few natural controls were then saved as *.avi files with a frame rate of 30 per second and frame size of 320mm x 270mm. The tokens were then incorporated into purpose built software with a built in functionality to log decision times and participant's personal details, such as name, age and gender.

Consonant tokens were split into two groups onsets and codas. All the words were selected from Longman English Pronunciation Dictionary (Wells, 2000). For control purposes, the word pairs used had congruent nuclei but incongruencies in either onset alone or coda alone, Table 1, so that any fusion elicited by consonant incongruity could be compared in different linguistic contexts.

Onset Contrasts			*Coda Contrasts*		
	Audio	**Visual**		**Audio**	**Visual**
O1	tail	fail	C1	map	mat
O2	seal	teal	C2	tap	tat
O3	pat	tat	C3	bus	but
O4	date	bait	C4	lot	loss
O5	fill	sill	C5	ram	ran

Table 1: Pairs of Consonant Fusion

As a control on the quality of recordings and the ambient environment, the experiment also included some congruent data in the same format as the stimuli with incongruence.

3.2 PROCEDURE

Fifty subjects took part in the vowel experiment, a mixture of both females and males, with an age range between 21 to 54 years. None had hearing problems and all either had normal vision or wore prescribed corrective lenses. The subjects were provided with a report form on which to record 'what they thought the speaker was saying' when receiving an experimental stimulus, replaying the clip as many times as they needed to reach a decision. The form included the following text-words: *one* corresponding to the audio channel of the stimulus, *one* for the video channel, one each for possible results of channel fusion, some random words and a space to write in a word not included on the form. There were no time limits set to complete the experiment, and no feedback about the experiment was given to the subjects and the experiment was conducted double-blind to eliminate experimental effects.

4. DISCUSSION OF RESULTS

For congruent channels, over 90% of the participants accurately reported what the talking heads were saying and for incongruent channels, there are distinctly different vulnerabilities to fusion in coda and onset consonant, detailed below.

4.1 FUSION RATES

Exploratory analysis revealed that short codas are more likely to undergo McGurk fusion than onset consonants, as shown in the bar chart Figure 1(a) and 1(b). Not all participants experience the fusion, and the bar-chart simply shows that more subjects do so with coda consonants than onsets. The clip which is most vulnerable to McGurk fusion was C1 in (a) and C1, C2 in (b), involving the fusion of **/p/** with **/t/** and **/t/** with **/s/**.

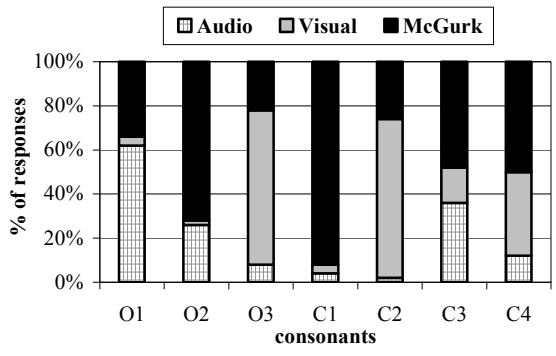

Figure 1(a): Percentage responses for incongruent onset and coda consonants with no noise

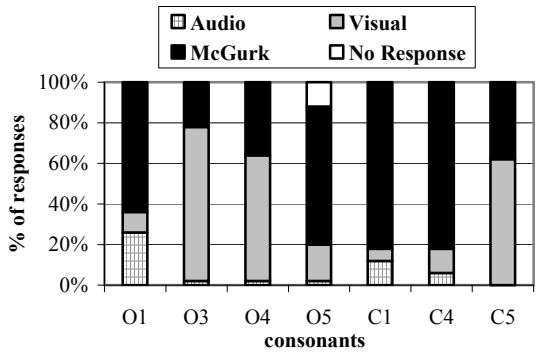

Figure 1(b): Percentage responses for incongruent onset and coda consonants with added 'cocktail party' noise

A more detailed statistical analysis using chi-squared tests revealed the fusion rate differences between incongruent onset and coda consonants are highly significant. The tests of null hypothesis that fusion rates are the same for onset and coda consonants were rejected: with no noise at $p<0.01$, with noise at $p<0.01$, and with noisy and noiseless cases together at $p<0.001$. Differences of fusion rates brought about by noise were tested: the hypothesis of null difference was rejected at $p<0.000$. The effect of background noise was most masked in stimuli with coda incongruities. The faint background speech noises that we used increased the fusion rate.

4.2 DECISION TIMES

With all the stimuli together (congruent and incongruent) the exploratory analysis was focused on the mean decision times per stimulus over all subjects, decision times being measured both by time to respond (duration time) and the number of replays of the stimulus. The mean decision times between onset and coda consonant are summarised in Table 2. The table clearly shows that are differences between consonant in the onset position and consonant in the coda position: subjects take longer to decide onset than coda.

Mean decision times per stimulus	Onset	Coda
Time to respond (secs)	16.68	14.55
Number of replays	0.92	0.68

Table 1: Summary of decision times

These averages conceal, however, significant differences between decision times for congruent and incongruent stimuli.

(a) Time to respond

The decision times were examined for variation with the state of the two channels (congruent, incongruent without noise and incongruent with noise) as shown graphically in Figure 2. The graph shows that in most cases the time to respond for coda is shorter than for onset consonant.

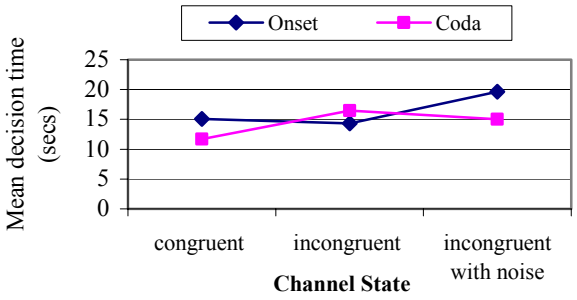

Figure 2: Duration times for various channel states

Detailed statistical analysis revealed by using ANOVA, with all the stimuli, the null hypothesis that time to respond is independent of channel state was rejected ($F=15.04$, $p<0.001$). With incongruent stimuli only, the null hypothesis that times to respond to onset and coda consonants were the same was rejected ($F=5.80$, $p<0.05$ with no noise; $F=17.35$, $p<0.000$ with noise). Note that the effect of noise was to blur the differences of response times between onset and coda consonants.

Time to respond was also analysed by partitioning the type of response into; decision for audio, decision for visual, and decision for fused. The results indicated that the decision time in noisy condition was greater than in noiseless condition for all types of responses. Detailed findings indicated that the mean decision time per stimulus in noiseless cases was higher for codas, but lower in noisy cases, as shown in Figure 3a and Figure 3b.

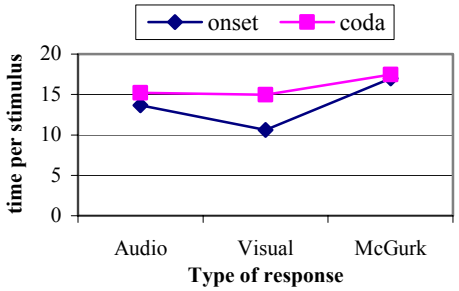

Figure 3a: Mean decision times in noiseless condition

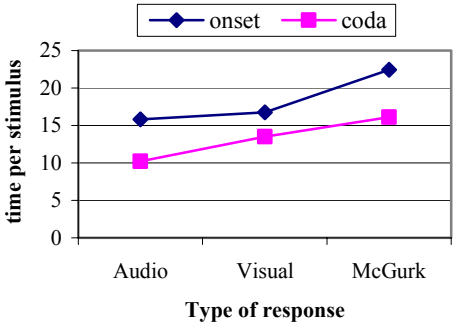

Figure 3b: Mean decision times in noisy condition

ISBN 1-876346-48-5

A more detailed statistical analysis revealed that, with all incongruent stimuli partitioned by types of response (decision for audio, decision for visual, decision for fused), null hypotheses that decision times are the same for onset and coda consonants were rejected for fusion responses (F=10.92, p <0.001). With only noisy incongruent stimuli partitioned by type of response, the null hypothesis of equal decision times was rejected for audio, visual and fusion type responses (F=5.869, p<0.05; F=4.83, p<0.05 and F=15.822, p<0.001) respectively. With noiseless incongruent stimuli partitioned by response, the equal decision times hypothesis was rejected for visual type responses (F=14.61, p<0.001).

(b) Number of replays of stimuli

The number of replays was examined for variation with the state of the two channels (congruent, incongruent without noise and incongruent with noise) as shown graphically in Figure 4. The graph shows that in most cases the number of replays of the stimulus is considerably shorter in consonant in the coda position than for the onset position.

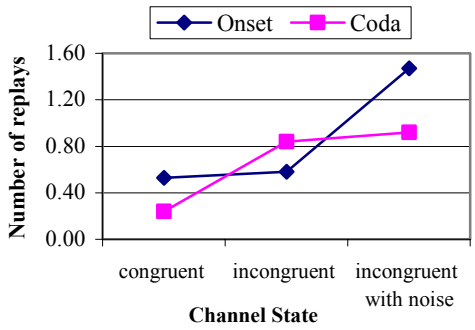

Figure 4: Mean number of replays for various channel states

Using ANOVA, with all the stimuli, the null hypothesis that number of replays is independent of channel state was rejected (F=12.43, p<0.001). With incongruent stimuli only, the null hypothesis that number of replays for onset and coda consonants was the same was also rejected (F=6.84, p<0.01 with no noise; F=15.66, p<0.001 with noise). Differences partitioned by types of response (decision for audio, decision for visual, decision for fused) brought about by noise were tested, the null hypothesis that replays are the same was rejected, F=9.058, p<0.01 for audio; F=12.716, p<0.000 for visual and F=18.013, p<0.000 for fused responses.

The emerging pattern is that the distinctly different vulnerabilities to fusion. For the most vulnerable coda consonants, it appears that fusion events bring about a speeding up of perceptions, especially in noisy conditions.

5. CONCLUSION

The study revealed that there are different vulnerabilities to 'perceptual sports' such as McGurk fusion in English monosyllabic words. The experimental findings can therefore be summarised to reflect that under re-alignment, a consonant in the coda position is more likely to cause fusion than a similar consonant in the onset position. For the most vulnerable coda consonants, it appears that fusion events bring about a speeding up of perceptions. These are the segments over which the greatest care must be taken in designing outputs of 'talking head/agents'. These findings confirm the impression of phonologists that a coda is less stable entity than an onset.

REFERENCES

[1] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle*. Cambridge, Mass: MIT Press, 1998.

[2] R. Campbell, B. Dodd, and D. Burnham, *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, Psychology Press, 1998.

[3] K. Tiippana, M. Sam, and T. Andersen, "Visual attention influences audio visual speech perception", in *Proceedings Audio Visual Speech Processing*, pp. 167-171, 2001.

[4] J. Beskow, M. Dahlquist, B. Granström,, M. Lundeberg, K. Spens, T. Öhman, T. The Teleface project - Multimodal speech communication for the hearing impaired, in *Proceedings of Eurospeech '97*, Rhodes, Greece, 1997.

[5] W. Sumby, and I. Pollack, "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, 26, pp. 212-215, 1954.

[6] K. Robert-Ribes, J. Schwartz, T. Lallouache, T. & P. Escudier, "Complementarity and Synergy in Bimodal Speech", *Journal of Acoustic Society of America* ,103(6), pp. 3677-3689, 1998.

[7] J. MacDonald, J. and H. McGurk, "Hearing lips and seeing Voices", *Nature* 264, December 23/30: pp. 746-74, 1974.

[8] Ali, A. N. and Ingleby, M (2002) "Perception Difficulties and Errors in Multimodal Speech: The Case of Vowels", in *Proceedings of the 9th Australian International Conference on Speech Science & Technology,* Melbourne, Australia, December 2 to 5, pp. 438 – 443, 2002.