Cross-Modal Fusion: Context Effects in Lexical Words*

Azra Ali

University of Huddersfield

The study focuses on the response of participants to audiovisual presentations of talking heads, and examines the effect of noise and temporal misalignment of channels in English monosyllabic words. The results show that McGurk fusion of phonetic segments is sensitive to the linguistic context of a segment: coda consonants elicit fusion more frequently than onset consonants and short vowels elicit fusion more often than long vowels. The second part of the research focuses on representing fusion using the subsegmental elements of the Government Phonology (GP) framework. The framework models fusion as a place-of-articulation (POA) phenomenon and results show that POA elements on different channels tend to be cancelled in the perception of participants responding to incongruent sensory data.

1 CROSS-MODAL SPEECH

People use both acoustic and visual modalities to understand speech, although many are not aware of the visual component. Strong evidence for the visual component is found amongst the many people with a hearing impairment who can understand speech by lip-reading (Sumby and Pollack 1954). Studies have shown that visual information provided by the movements of a speaker's mouth and face strongly influences what an observer perceives, even when the auditory signal is clear and the observer's hearing is good (Massaro 1998, Campbell 1998). The way in which the channels reinforce varies from segment to segment. On the one hand, perceiving a place-of-articulation contrast, such as that between labial /b/ and palatal /d/, is difficult via sound but relatively easy via sight. On the other hand, a contrast of voicing state, such as that between voiceless /s/ and voiced /z/ is easy to hear but scarcely visible. 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. The complementarity makes accurate speech perception more resistant to channel noise (Robert-Ribes et al 1998). Complementarity is of particualar importance to communication with people suffering from hearing impairment short of total hearing loss (Beskow et al 1997).

Evidence for strong psychophysical interaction between audio and visual speech channels in human speech perception is found in the well-known McGurk effect (MacDonald and McGurk 1976). If humans are presented with temporally aligned but conflicting audio and visual stimuli – known as 'incongruent stimuli' – the perceived sound may differ from that present in either channel. McGurk and MacDonald asked their recording technician to create a videotape with the audio syllable [ba] dubbed onto a visual [ga], most normal adults reported hearing [da] or [tha]. But when the participants were presented with only one modality (visual or audio, not too noisy) they reported the syllables correctly. The McGurk effect highlights the nature of human speech perception, which is clearly bimodal. The human observer of a 'talking head' perceives misaligned video data as if it were aligned. Such issues are of considerable importance for the effective design and use of new multimedia applications involving audio-visual speech synthesis, such as in video telephones, video conferencing, and information retrieved via talking head interfaces. Animated cartoons with talking heads are increasingly used to represent software agents. The design of such animated

* I would like to thank Dr Michael Ingleby for suggesting a GP representation of McGurk fusion and encouraging me to look at coda-onset differences. His help and support throughout my research and in producing this paper is very much appreciated.

Camling 2003: 53-59.

talking agents demands linguistic knowledge in some phonetic detail (Massaro 1998, Beskow et al 1997). Generally speaking, a multimedia presentation of a spoken message is more accurately perceived than voice-only presentation: the audio and visual data reinforce in accessing the same language primitives in the listeners cognitive lexicon and hence elicit responses that are more resistant to noise than single-channel presentations. But a necessary condition for reinforcement is good synchrony.

Previous studies have concentrated only on participants' responses to congruent and incongruent multimodal speech stimuli in nonsense syllables. We have begun an extended study of the effect of linguistic context on fusion, working first on the simplest English monosyllabic words (long or short vowel nucleus between a non-branching onset and coda). Studies by Müller (2002) using the BNC and a standard pronouncing dictionary for syllabification shows that simple monosyllables of this nature make up 65% of the total for British English.

2 GOVERNMENT PHONOLOGY

An important secondary aim of the research was to investigate how successfully the Government Phonology (GP) framework can model the McGurk fusion experienced by a group of participants. People's perception of audio-visually mismatched stimuli can change depending on which place of articulation is presented auditorally and which is presented visually. This GP framework was originally developed for modelling single-channel speech cognition, and we are taking it into multi-media linguistics. The primes of GP are used because they have acoustic (Ingleby and Brockhaus 2002) and in some cases visual signatures (Harris and Lindsey 2000), making them detectable in combination and isolation. We use them to model the cross-channel processes that produce McGurk fusion in simple English monosyllables. Using the GP framework enabled us to pinpoint the fusion, in terms of place and manner of articulation, for consonants and for vowels using the cardinal diagram. The results from the experiments and theoretical analysis provide a deeper understanding of audiovisual speech perception but also provide an experimental framework for testing phonological representation.

The Government Phonology framework is based on cognitive primes (Kaye, Löwenstamm and Vergnaud 1985, Harris 1994) and is known for its power in modelling coarticulation phenomena as assimilation of speech primes. It also expresses the distinction between onset and coda in syllabic structure. GP characterises a speech segment by a phonological expression consisting of a subset from a small fixed set of elements, fewer in number than the features of SPE (Chomsky and Halle 1968). GP expressions can be used to model the phonologically significant processes of the world's languages more easily (Ingleby and Brockhaus 1998). Phonological elements being E={A, I, U, H, L, ?, h}, the sounds are thus represented in isolation or in combination (fusion of one or more elements). The elements can be broadly categorised into three segments: two manner elements, **ʔ** (occlusion) and **h** (noise); three place elements **U** (labial), **I** (palatal) and **A** (back) and finally two source (voicing state) elements **H** (halt phonation) and **L** (low tone).

The elements at the head of a list represent the most salient manner of articulation of the phone, while the other elements represent less salient properties - place of articulation and vowel quality contrasts, voicing state etc. Therefore, the GP elements differ from binary features in that 'only their presence needs to be specified within a segmental representation' (Chalfont 1997:41). For instance e.g., a /b/ sound is a voiced stop and in GP terms is expressed as **ʔ+h+U** and a /p/ is a voiceless stop expressed as **ʔ+h+U+H**. Generally, examples, all fricatives contain **h** as head and, less saliently, **ʔ**. The voiceless fricatives contain **H**, too, but voiced ones do not. The place of articulation of consonants is specified by combinations of place elements – **U** alone for /f/ and /v/, **A** alone in /k/ and /g/. The place elements are also used in combination to represent vowel qualities. The elements **A**, **I** and **U**

(as heads) represent the extremes of the vowel triangle: for example in long vowels /ɑː/, /ɪː/ and /ʊː/ respectively. The vowel /**e**/ has a quality represented by combination **A+I**, while the central schwa vowel /**ə**/ is represented by combination **A+I+U**.

As mentioned earlier, GP is known for its power in modelling coarticulation phenomena for example assimilation is the movement of an element. This can be illustrated using a simple nasal assimilation, for example, the words 'input' and 'income' are ideally pronounced /ɪ n p ʊ t/ and /ɪ n k ʌ m/ but the phone /**n**/ in rapid speech changes its place of articulation by assimilating material from the following consonant, thus / ɪ **n** p ʊ t / becomes / ɪ **m** p ʊ t / and / ɪ **n** k ʌ m / becomes / ɪ **ŋ** k ʌ m /. In GP, these processes are modelled respectively by moving labial element **U** or a back element **A** from the segment following the nasal. Element **U** present in /p/ spreads to /n/ which is perceived as labial nasal /**m**/, while element **A** present in /k/ spreads to /n/ which is perceived as velar /**ŋ**/.

The movement and assimilation of GP elements can explain all the usual phonological processes. In this study we investigate the explainability of the McGurk effect on similar lines. We regard the elements as having a visible signature in the facial gestures of speakers as well as the usual acoustic signature. We seek to model the perceptions of participants presented with re-aligned speech in terms of GP elements present either in the audio channel (represented by an acoustic signature) or in the visual channel (*via* a visible signature).

3 EXPERIMENTS: CROSS-MODAL FUSION IN LEXICAL WORDS

Experiments were designed around artificially re-aligned vowel and consonant audiovisual stimuli, essentially video recordings of speakers articulating common English words. The re-aligned stimuli were interspersed amongst control stimuli: natural recordings without realignment. Vowel stimuli were split into two groups, short and long vowels. Each stimulus was made up of a vowel nucleus embedded between the same onset and coda. For the creation of re-aligned stimuli, word pairs were selected that were well separated in the vowel cardinal diagram, choices being as in Table 1. With consonant stimuli (Table 2), word pairs used had congruent nuclei but incongruencies in either onset alone and coda alone – so that any fusion elicited by consonant incongruity could be compared in different linguistic contexts. All the words were selected from Longman English Pronunciation Dictionary (Wells 2000).

SHORT / LONG	s1	s2	l1	l2	l3	l4
Audio	hod	hod	hoard	hoed	hard	hard
Visual	had	hid	hoed	who'd	who'd	heed

Table 1: Incongruent vowel stimuli

ONSET / CODA	o1	o2	o3	o4	o5	c1	c2	c3	c4	c5
Audio	tail	seal	pat	date	fill	map	tap	bus	lot	ram
Visual	fail	teal	tat	bait	sill	mat	tat	but	loss	ran

Table 2: Incongruent consonant stimuli

Our experiments also probed the hesitations of participants *via* two measures. One was the total time taken by a participant to select a response to a stimulus from an open list of possible responses. The other was the number of replays used by the participant before reaching a decision about the stimulus. Both measures were logged automatically by our experimental software.

3.1 Creating the stimuli and procedure

Video recordings were made of a male (aged 23 years) and a female (aged 22 years) native speaker 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 realigned stimuli, words were grouped into contrasting pairs, as detailed in Tables 1 and 2 with few natural controls. To some re-aligned clips, 'cocktail party' noise was added acoustically. All clips were then saved as *.avi files with a frame rate of 30 per second and frame size of 320mm x 270mm.

The participants were provided with report forms on which to record 'what they thought the speaker was saying' when receiving an experimental token. The report forms included text-words corresponding to the audio channel of the token, the video channel of the token, a number of possible results of channel fusion and some random words, presented either on a vowel cardinal diagram (for vowels) or a similar geometry for consonant place of articulation. The experiments were carried out in a Usability Lab with minimal background noise. Participants sat about half a metre from the monitor screen and used headphones connected to the computer to listen to the audio.

3.2 Participants

Fifty participants took part in the 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. Both the vowel and consonant stimuli were randomised so that the participants did not know which stimulus they were viewing.

3.3 Discussion of results

The comparison 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 observed rates in the experiments reported below were a lowest (16%) for long vowels, 48% for onset consonants, 60% for coda consonants and a highest 67% for short vowels, as shown in Figure 1.

In Figure 2, the graph clearly shows that in incongruent channels (with and without noise), participants took longer to decide what the talking head was saying, this is echoed also in the number of replays of the stimuli. Full details of the experimental findings, fusion rate and decision time can be found in Ali (2003) and Ali and Ingleby (2002).

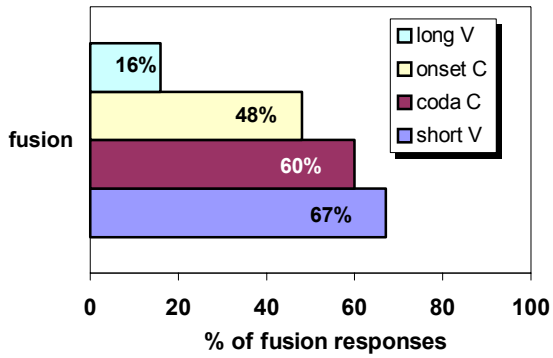

Figure 1: Fusion rate

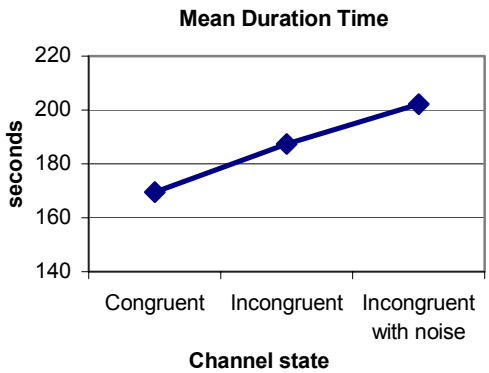

Figure 2: Decision time

4 FUSION IN THE GOVERNMENT PHONOLOGY FRAMEWORK

The primes of GP are used because they have acoustic and in some cases visual signatures – making them detectable in combination and isolation. We use them to model the cross-channel processes that produce McGurk fusion in simple English monosyllables. Using the GP framework enabled us to pinpoint the fusion, in terms of place and manner of articulation, for consonants and for vowels using the cardinal diagram.

4.1 Modelling vowel fusion

A vowel sound is represented by a single GP element or by combination (fusion of two or more elements). In the first example Figure 3(a), where the audio (A) channel is '**i**d' is represented by GP element **I** and the visual channel (V) '**o**d' as a fusion of two GP elements **U+A** (the ordering is not important), thus resulting in a McGurk fusion (F) as '**ea**d' which can be expressed as **A+I**. In the second example Figure 3(b), the audio channel is '**u**d' → GP = **U** and the visual channel '**a**d' → GP = A resulting in a McGurk fusion as '**o**d' → GP = **U+A**.

Figure 3: Vowel fusion

4.2 Modelling consonant fusion

Consonant fusion can also be modelled in a similar way by adapting the cardinal diagram. This can be achieved by applying the 'one mouth' principle first noted by Jakobson (Anderson 1985:121) who suggested that a single system could be used to describe both the vowel and consonant sounds. The principle makes /p/, /t/, /k/ contrasts similar to those distinguishing to /u/, /i/, /a/: both can be presented in a triangle which is equivalent to the vowel cardinal diagram. In GP the same elements are used to describe both of these vowel and consonant contrasts, using elements **U, I A,** as shown in Figure 4.

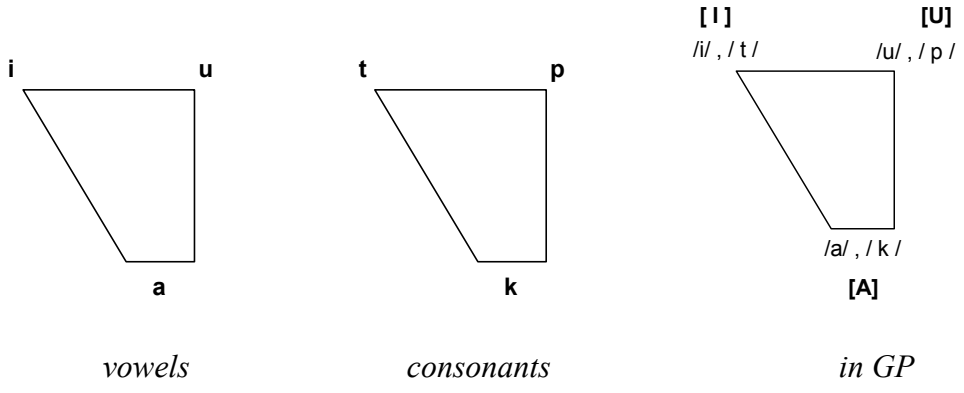

Figure 4: 'one mouth' principle

In the report forms used by our participants, cf. Figure 5, which shows voiced and unvoiced obstruents, the various labelling of vertices that have been used in the literature is also shown, e.g. Jakobson (Anderson 1985:121) and Shane (1973:23). Figure 6 shows the regions of interest in the cardinal diagram for an investigation of onset consonant fusion. In this case, some of the participants reported a perception '**f**ate' when presented with visual '**d**ate' and audio '**b**ait'. This pattern of perception is <u>not</u> typical of participant responses.

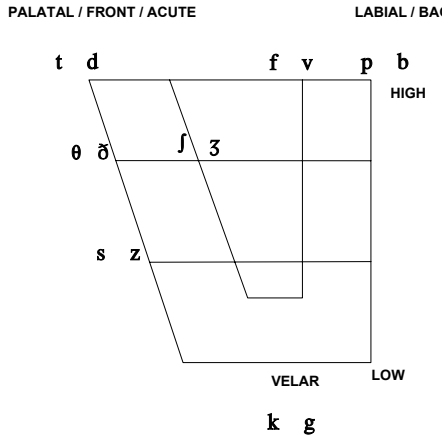

Figure 5: Consonant cardinal diagram

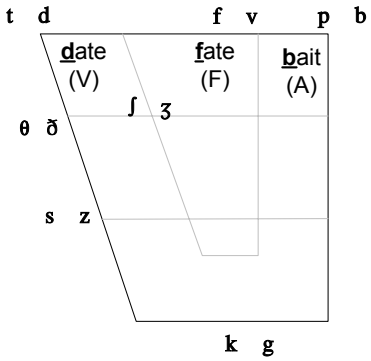

Figure 6: onset consonant fusion

The most interesting finding is rather than the fusion being <u>agglomerative</u>, similar to the assimilation phenomena of coarticulation effects, McGurk fusion is revealed to be <u>cancellative</u>. If the fusion were agglomerative, a 'hid' in the acoustic channel could assimilate material from 'h**a**d' **A** in the visual channel to be perceived as 'h**ea**d' **A+I**. In fact, participants tended to report 'h**oo**d' **U**. Conflict between channels cancels the **I** and **A** leaving behind a more salient **U**. Similarly, for consonants, a conflict between different elements in audio and visual channels results in neither showing up in the perceptual channel, this can be explained in terms of GP primes. If for example, palatal /t a **t**/ ("tat") in the visual channel conflicts with labial /t a **p**/ ("tap") in the audio channel, both **I** (/t/) and **U** (/p/) elements are cancelled from the coda consonant leaving behind a more salient **A**, corresponding to the perception /t a **k**/ ("tack"). A detailed account of this cancellative effect explained in terms of GP primes can be found in Ingleby and Ali (2003).

5 CONCLUSION

Taking a wider view of these results, the decreasing vulnerabilities, to fusion amongst the syllabic constituents, short-vowel, coda, onset, long-vowel, which we know to be statistically significant (Ali 2003) and (Ali and Ingleby 2002), offers a hope of probing mental models of syllabic phenomena in multi-media linguistics. This adds to the growing arsenal of experimental probes into cognitive models of language – extending the phrase boundary probe based on bilingual code switching, and the morphological domain-boundary probe based on epenthetic expletives.

REFERENCES

Ali, Azra (2003) 'Perception difficulties and errors in multimodal speech: The case of consonants.' In proceedings of the *15th International congress of phonetic sciences*, Barcelona (to appear).

Ali, Azra and Michael Ingleby (2002) 'Perception difficulties and errors in multimodal speech: The case of vowels.' In proceedings of the *9th Australian international conference on speech science and technolog,* Melbourne, Australia: Australian Speech Science and Technology Association (Inc), pp. 438-443.

Anderson, Stephen (1985) *Phonology in the Twentieth Century*. Chicago: University of Chicago Press.

Beskow, Jonas., Martin Dahlquist, Björn Granström, Magnus Lundeberg, Kark-Erik Spens, and Tobias Öhman (1997) 'The Teleface project – multimodal speech communication for the hearing impaired.' In the proceedings of the oc. 5th European conference on speech communication and technology: Eurospeech, Greece: Typoffset, pp. 2003-2010.

Campbell, Ruth., Barbara Dodd, and Denis Burnhan (1998) *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*. East Sussex: Psychology Press.

Chalfont, Carl (1996) *Automatic Speech Recognition - A Government Phonology Perspective on the Extraction of Subsegmental Primes from Speech Data*. PhD Thesis, University of Huddersfield.

Chomsky, Norm. and Morris Halle (1968) *The Sound Pattern of English*. New York: Harper and Row.

Harris, John (1994) *English Sound Structure*. Oxford: Blackwell.

Harris, John and Geoff Lindsey (2000) 'Vowel patterns in mind and sound.' In N. Burton-Roberts et al (eds.) *Phonological Knowledge: Its Nature and Status.* Oxford: University Press, pp.185-205.

Ingleby, Michael and Azra Ali (2003) 'Phonological primes and McGurk fusion.' In the proceeding of the *15th International congress of phonetic sciences*, Barcelona, (to appear).

Ingleby, Michael and Wiebke Brockhaus (2002) 'Acoustic signatures of unary primes.' In J. Durand et al (eds.) *Phonology: from phonetics to cognition.* Oxford: University Press, pp. 131-150.

Ingleby, Michael and Wiebke Brockhaus (1998) 'A concurrent approach to the automatic extraction of subsegmental primes and phonological constituents from speech.' In the proceedings of the *COLING-ACL Conference*, Montreal. France: Institut de la Communication Parlee, pp.578-582.

Müller, Karin (2002) 'Probabilistic context-free grammars for phonology.' presented at the *Workshop on Morphological and Phonological Learning Computational Phonology 2002*, Philadelphia. http://www.coli.uni-sb.de/~kmueller/publication.html

Kaye, Johnathan., Jean Löwenstamm, and Jean-Roger Vergnaud (1985) 'The internal structure of phonological elements: a theory of charm and government.' *Phonology Yearbook* 2:305-328.

MacDonald, John and Harry McGurk (1976) 'Hearing lips and seeing voices.' *Nature* 264, December 23/30:746-748.

Massaro, Dominic (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle*. Cambridge, MA: MIT Press.

Robert-Ribes, Jordi., Jean-Luc Schwartz, Mohamed-Tahar Lallouache, and Pierre Escudier (1998) 'Complementarity and synergy in bimodal speech.' *Journal of the Acoustic Society of America* 103(6):3677-3689.

Shane, Sanford. (1973) *Generative Phonology*. Englewood Cliffs London: Prentice Hall.

Sumby, W. and Irwin Pollack (1954) 'Visual contribution to speech intelligibility in noise.' *Journal of the Acoustical Society of America* 26:212-215.

Wells, John (2000) *Longman Pronunciation Dictionary*. Second Edition, Harlow: Longman.

Azra N. Ali

School of Computing and Engineering
Canalside West, Firth Street
University of Huddersfield
Huddersfield
HD1 2LN
United Kingdom

a.n.ali@hud.ac.uk
http://scom.hud.ac.uk/scomana