Phonological Primes and McGurk Fusion

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

ABSTRACT

We illustrate the power of phonological primes in representing coarticulation processes by modeling nasal assimilation in English and Arabic, then devoicing in German and Arabic. The primes of Government Phonology 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. Unlike coarticulation, which is modeled by assimilation and loss of primes, fusion requires cancellation of those resonance primes that carry conflicting evidence in audio and visual channels. The same cancellative mechanism works equally well with incongruence in the place of articulation of onsets, codas. The mechanism also works with incongruence of vowel quality in nuclei.

1. MCGURK FUSION IN AUDIO-VISUAL SPEECH PRESENTATIONS

The seminal work of MacDonald and McGurk on fusion [1] in audio-visual speech perception was presented in psychophysical terms. Incongruent data (a labial sound 'baa' aligned with an image of a speaker saying palatal 'daa') elicited a fusion response in many subjects (the velar sound 'gaa'). This response to multi-modal incongruity contrasts starkly with more familiar unimodal fusion phenomena in perception. Unimodal incongruities between left and right channels elicit, for example, depth perceptions in subjects responding to stereovisual or to stereophonic data. The early work on multimodal incongruity avoided cognitive effects of linguistic context by working with nonsense syllables. However, McGurk fusion phenomena are now known to survive embedding in the many natural languages, at least for subjects using English [2], French [3], Dutch [4], Finnish [5], German & Spanish bilingually [6], Chinese [7] and Japanese [8]. Subjects with very different language models governing their perceptions have reported fusion responses to word stimuli having incongruent audio and visual channels. 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 Mueller [9] 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. In this paper we attempt to put our first findings in an appropriate theoretical framework, using subsegmental representations that allow coarticulation phenomena to be modeled in a principled way.

The experimental results on vowel incongruities [10] and consonant incongruities [11] are published elsewhere. They are concerned with variation of fusion response with the linguistic site of an incongruity. In sum, they show that fusion rates (as measured by the percentage of subjects reporting a fusion response to an incongruent stimulus) differ significantly between long and short vowels and between onset and coda consonants. Our first estimates of rates from double-blind experiments are, in decreasing order of fusion rate, short V (67%), coda C (60%), onset C (48%), long V (16%). The rates vary with accuracy of alignment of the incongruent stimuli, but the ordering of rates is invariable and statistically significant. The difference between coda and onset fusion rates is great enough that the response rates could be used to test empirically hypotheses about constituent structure. Although on sonority gradient grounds it is not controversial to syllabify in English using codas (e.g. [12]), there are rival coda and no-coda views on syllabification in, for example, Arabic [13]. Fusion experiments could be designed to settle the controversy by studying variation of fusion rates in Arabic consonant clusters. But before entering into such controversies, we plan first to check the behaviour of fusion rates in broader contexts such as branching constituents phrase context.

Besides such scientific probing of the cognitive models of speech, fusion phenomena have technological spin-offs, too. Multi-media presentation of speech in video-telephony and video-conferencing has become commonplace, and equipment lacking the bandwidth to handle such data accurately in real time communication has been widely marketed. Also, the programmable talking avatar is increasingly used as a representation of a software agent. These constructs seem set to become a vital part of the human-computer interface; yet, like overloaded telephony channels, they can create accidental incongruities when acoustic and video channels suffer synchronisation errors. Our technical objective, therefore, is directed towards understanding the errors that can arise when using these types of multi-media representation of speech. We have started to map the most likely locations for errors (for example in codas), but in this paper we look closely at the

errors of vowel quality and of consonant place of articulation that McGurk fusion generates. A historical example of improving communication by understanding perception errors occurs in German sports stadia where the poor acoustics makes it hard for listeners to distinguish *zwei*=2 from *drei*=3. The vowels, being more sonorous, are more easily heard than consonants, so announcers are taught to say *zwo* instead of *zwei*.

2. PHONOLOGICAL PRIMES

In this paper we approach fusion through the subsegmental representations of general phonology. The phonological framework that best suits our investigation is not the popular articulatory feature framework of Chomsky and Halle, but a phonology based on unary primes. The Government Phonology (GP) proposed by Kaye, Lowenstamm and Vergnaud [14] proposes a small set of such subsegmental primes - elements of cognitive representation equally engaged in articulation and audition. They are thought to function as elements of the mental models conditioning articulation and also speech perception. And there is evidence that they are phonetically detectable both in combination and isolation.

The set of elements has changed as GP has evolved, but the most widely used elements are as follows. Contrasts in voicing state are represented by presence of **H** in voiceless segments, absence in voiced segments. Manner of articulation is represented by noise element **h** (present in fricatives and the burst phase of plosives) and occlusion element **ʔ** (present in fricatives, affricates and the closure phase of plosives). Contrasts in place of articulation of consonants are represented using velar element **A**, palatal element **I**, and labial element **U**. For example, the Arabic velar/uvular plosive *qaaf* has a sound **q** represented by a combination of **ʔ** + **h** with **A**, while the palatal/velar *kaaf* with a sound **k** is represented by **ʔ** + **h** with **I** and **A**. Some GP theorists [14] add a fourth element **R** to the inventory of place elements but the examples used in this paper do not make use of it. The place elements are also known as resonance elements because they are employed to represent vowel contrasts and vowel sounds are resonances of the vocal tract. In a vowel segment the element **A** alone describes the vowel sound **aː**; the elements **I** and **U** representing the vowel sounds **ɪː** and **ʊː** respectively. In the vowel cardinal diagram they are the vertex vowels. Central vowel **ə** is represented by a combination of all three resonance elements. The case for detectability of GP elements is based on cluster analysis of the spectra of speech segments. A detailed account of how to acquire acoustic signatures of elements from clusters is given in Ingleby and Brockhaus [15]. More recently, Harris and Lindsey [16] have proposed, at least for the resonance elements, visual signatures made up of facial gestures from a video channel.

The elements themselves provide, *inter alia*, an elegant representation of phonological processes. Here are some examples:

(1) ENGLISH /increase/

ɪ n k r ɪː s —velarise→ **ɪ ŋ k r ɪː s**

(2) ARABIC / أَنْفُس =souls /

ʔ æ n f ʊː s —labialise→**ʔ æ m f ʊː s**

(3) GERMAN / *Band*= ribbon /

b a n d ø —devoice→**b a n t ø**

(4) ARABIC / إِجْحَاف =injustice /

ʔ i ʒ ø ħ aː f —devoice→ **ʔ i ʃ ø ħ aː f**.

The first is typical behaviour of nasals, modelled by assimilation by the palatal segment **n** of velar element **A** from **k** to form the velar nasal **ŋ**. The second example is also a rather standard assimilation of element **U** from labio-dental **f** to convert **n** to **m**. Example (3) is more surprising: it depends on the way of representing the empty timing slot **ø** (needed at the end of words to handle epenthetic material). Because **ø** has no phonation, one can regard it as the most isolated manifestation of GP element **H**. Taking this view, the **H** is assimilated by word-final **d** which becomes **t**. The word-final devoicing, widely attested in German and Dutch, has thus a simple GP representation. Example (4) shows a case of word-medial devoicing in Arabic, not attested in Germanic languages (except in compounds like *mundtot* = struck dumb). As the Arabic orthography shows, there is an empty slot, the phonological expression **ø** of the vowel diacritic *sukuun* (◌ْ)between the after the letter *jiim* (ج ,which has the voiced fricative sound **ʒ**). The **H** in the GP representation of **ø** can thus be assimilated converting **ʒ** to **ʃ**. A more detailed account of the phonological manifestations of *sukuun* in Arabic has been presented by Ingleby [15]. Indeed Baothman [13] has developed GP representations of a full inventory of the phonological processes of Arabic. In the next section, we attempt to examine the power of GP to represent multi-modal fusion. The attempt succeeds by adding one further subsegmental event to GP inventory of assimilation and loss.

3. REPRESENTING FUSION USING GP ELEMENTS

The experimental results on fusion in simple words are summarised below in cardinal diagram form. Subjects were asked to report their perceptions by choosing words arranged on a vowel quality triangle. In the first graph, Figure 1 (left), two positions corresponding to vowel quality in audio and visual channels are shown, together with the area where subjects' fusion responses cluster. Because the GP resonance elements also represent place contrasts for consonants, the place of articulation of codas and onsets can be represented on the same type of graphical display, Figure 1 (second graph-right). Generally, we found that the fusion response areas remained unchanged when we switched the contents of audio and visual channels, and

that the same fusion area locations occurred with onset-, coda- and nucleus-incongruities. Before we analysed results, it seemed reasonable to expect, as in coarticulation, that there would be assimilation of elements across the audio and visual channels. Cross-channel assimilation leads one to expect perceptions to fall into an area straddling the line segment joining the points representing audio and visual signals. This expectation was only born out when subjects opted for the most salient channel, ignoring incongruity. In every case of a fusion response, the reported perceptions were outside the expected shaded area. We therefore abandon the assimilative model of fusion.

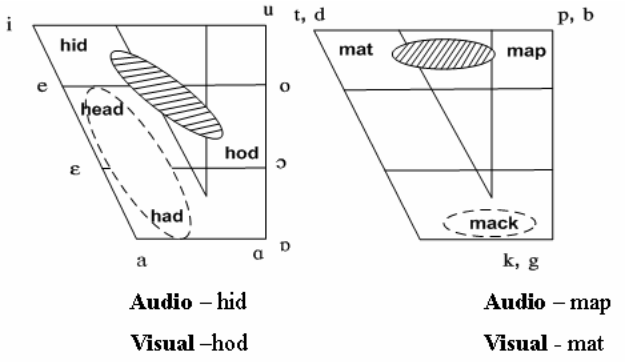

Figure 1: Fusion in consonants and vowels

The graphs show that the perceived vowel quality (or consonant POA) is determined by cancellation. With audio /hid/ and video /hod/, fusion responses varying over the choices /head/ and /had/ were obtained. The dominant **I** in the audio segment conflicts with the dominant **U** in the video segment. The subjects first hesitate over the conflict of evidence (we have measured their decision times [11]). Then they resolve the conflict of evidence by deciding that the element over which there is no conflict is dominant: elements in cross-mode conflict are cancelled. The same pattern shows in the audio /mat/, video /map/ leading to perception /mack/. It also shows in stimuli with onset incongruity.

In unimodal coarticulation, element cancellation is not needed to model processes. In the case of lenition processes, there is element loss when, for example, a plosive **t** is glottalised to **ʔ**. The plosive is represented by manner elements **ʔ** + **h** while the glottal stop has no **h** in its subsegmental representation. The process is simply loss of a single element of subsegmental material [12]. We thus believe that element cancellation is the key mechanism of conflict resolution when a mental model processes conflicting evidence. This conjecture needs further investigation outside the linguistic domain. We suggest that experimentation on hue perception using incongruent spectra at left and right eye would be helpful. Similar experimentation on response to incongruities of musical pitch at left and right ear could be used to test our conjecture.

4. CONCLUSIONS

The main conclusion from the related experimental papers [10] and [11] is that the qualitative patterns of response to vowel quality and POA incongruity are the same in the three linguistic contexts examined (onset, nucleus and coda). In this paper we have represented this common pattern by a new type of subsegmental event. The output of the fusion process is represented by canceling (from a default combination of all resonance/place primes) any conflicting primes. Primes are deemed to be in conflict if they are present on one channel and not on the other. The restriction to resonance primes in the cancellation rule has its origin in the fact that only these have visual signatures on the speaker's lips. Harris and Lindsey [17] portray these signatures as rounded lips for **U** , a wide mouth for **A** and a rectangular, tight lip-shape for **I**.

The cancellation event is not, in general, representable by the assimilation and loss events that suffice for representation of co-articulation processes. When, for example, we get velar output coda of /tick/ from channel inputs /tit/ (palatal coda) and /tip/ (labial coda), the velar element in the output cannot have been assimilated from either input channel. Not only is it absent from both input codas, it is also absent from both the onset and the nucleus of the channel words! If, however, one adds to the basic events of assimilation and loss in the subsegmental tier, our element cancellation event (expressed as the rule of striking out conflict elements from a default), then one can predict what sort of fusion response will be elicited by a given incongruity. The new event adds significantly to the predictive power of the phonological theory.

It would, in principle, be possible to describe the cancellation rule in other phonological frameworks than GP. Each cognitive element of GP corresponds to a bundle of binary features that may be considered to flip in harmony from + to – states when an element moves. Simple rules for nasal assimilation in GP can be re-cast as more complex rules in a binary feature framework. Element motion associated with cancellation events is not different from motion associated with the more usual events of assimilation and loss. We do not, therefore, feel that the patterns and rules that we pick out to model fusion are tied to the GP framework alone. The rule syntax for cancellation events would just be more complex in a feature or feature-bundle framework than it is in a theory like GP with privative primes.

REFERENCES

[1] J. MacDonald, J. and H. McGurk.Hearing "Lips and Seeing Voices". *Nature* 264, pp. 746-748, 1976.

[2] D.W. Dekle, C.A.Fowler and M.G.Funnell. "Audiovisual integration in perception of real words". *Perception and Psychophysics,* 51 (4), pp. 355-362, 1992.

[3] C. Colin, M.Radeau and P. Deltenre. "Intermodal Interactions in Speech: A French Study", in *Audio Visual Speech Perception*, Terrigal, (New South Wales, Australia), December 1998. http://www.cmis.csiro.au/avsp98/papers/

[4] B. DeGelder, P.Bertelson, P.Vroomen and H.C.Chen. "Interlanguage Differences in the McGurk Effect for Dutch and Cantonese Listeners". Proc. *4th European Conference on Speech Communication and Technology*, Madrid, pp 1699-1702, 1995.

[5] M. Sams, P.Manninen, V.Surakka, P.Helin and R.Katto. "McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context". *Speech Communication*, 26 (1-2) pp. 75-87,1998.

[6] A. Fuster-Duran, "Perception of Conflicting Audio-Visual Speech: An Examination across Spanish and German", in *Speechreading by Humans and Machines,* D.G. Stork & M. E. Hennecke (eds.), Springer-Verlag, 1995.

[7] D. Burnham, and S.Lau. The effect of tonal information on auditory reliance in the McGurk effect. In *Audio Visual Speech Perception*, Terrigal, (New South Wales, Australia), December 1998. http://www.cmis.csiro.au/avsp98/papers/

[8] Y. Hayashi, and K.Sekiyama. "Native-Foreign Language Effect in the McGurk Effect: A Test With Chinese and Japanese". In *Audio Visual Speech Perception*, Terrigal, (New South Wales, Australia), December 1998. http://www.cmis.csiro.au/avsp98/papers/

[9] K. Müller, "Probabilistic Context-free Grammars for Phonology". Proc *6th Workshop of ACL SIG Computational Phonology*, Philadelphia, July 2002.

[10] A.N. Ali, and M. Ingleby, "Perception Difficulties and Errors in Multimodal Speech: The Case of Vowels". Proc. 9th Australian International Conference on Speech Science & Technology, Melbourne, 2002, pp. 438-443.

[11] A.N. Ali, "Perception Difficulties and Errors in Multimodal Speech: The Case of Consonants". Proc. *15th International Congress of Phonetic Sciences*, Barcelona, 2003.

[12] J. Harris, *English Sound Structure*, Blackwell, Oxford, 1994.

[13] F. Baothman, and M. Ingleby, "Pharyngeal Spreading versus Nasal Assimilation Processes" in Arabic Speech". Proc. *Sixteenth Annual Symposium on Arabic Linguistics*, Arabic Linguistics Society and Cambridge University, 2002.

[14] J. Kaye, Löwenstamm, J and Vergnaud, J. "The Internal Structure of Phonological Elements: a Theory of Charm and Government". *Phonology Yearbook* 2, pp. 305-328, 1985.

[15] M. Ingleby, and F. Baothman. "Empty Nuclei in Arabic Speech Patterns and the Diacritic sukuun". Proc. *Sixteenth Annual Symposium on Arabic Linguistics*, Arabic Linguistics Society and Cambridge University, 2002.

[16] M. Ingleby, M. and Brockhaus, W.G. "Acoustic Signatures of Unary Primes", in *Phonology: from phonetics to cognition,* Durand, J. et al (eds.),. Oxford University Press, 2002.

[17] J. Harris, and G. Lindsey, "Vowel Patterns in Mind and Sound", in *Phonological Knowledge: Its Nature and Status,* Burton-Roberts, N., et al. (eds.), , Oxford University Press, 2000, pp.185-205.