

Michael Ingleby & Azra N.Ali

*School of Computing and Engineering
University of Huddersfield*

Government Phonology and Patterns of McGurk Fusion

audio-visual perception

(b a: | d a:) → (ga:)_F

statistical

(A 25%, V 15%, F 60%)

incongruous data

(the first segments in audio and visual channels are inconsistent)

site of the incongruity

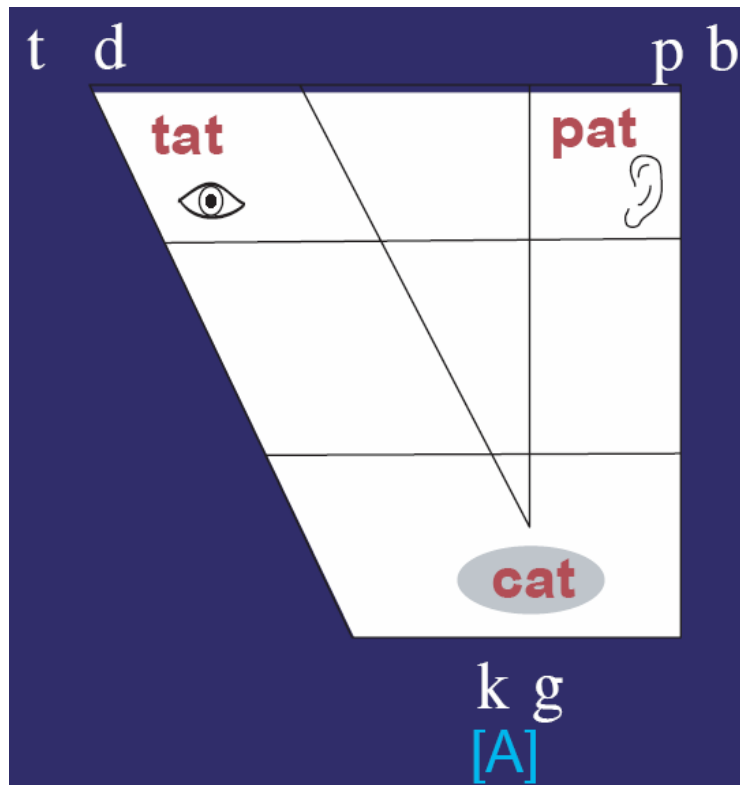
The Background Research Project:

- An interdisciplinary study of human response to software artefacts in a ‘usability lab’
- Focused on screen displays of talking heads
- Azra is a doctoral student investigating transfer of information using talking head displays
- Michael is supervising, and works on statistical pattern recognition, including automatic recognition of phonological primes (GP elements)
- Assisted by Dr Phil Marsden, a cognitive psychologist specialising mental models and reliability aspects of the human-computer interface

Scope of the talk:

- McGurk fusion in words – incongruous vowel segments and consonant segments embedded in words
- Double-blind experiments – to eliminate transfer of expectation from experimenter to participating subjects
- We describe types of fusion – assimilative *versus* cancellative –and GP models thereof
- We note change of fusion rate with site of incongruity and pick out sites most favourable to fusion
- The experiments include words with branching constituents
- Survival of fusion in polysyllabic words and trite phrases and sentences is described briefly

Results on a simple onset site/slot:



the figure is derived from the cardinal diagram for vowel qualities, which we adapt to consonant POA by the 'one mouth' principle

its vertices correspond to the GP resonance elements I, U and A

Reference

Ali (2003c)

$(p \text{ æ } t \mid t \text{ æ } t) \longrightarrow (k \text{ æ } t)_F$

- typical statistical pattern reported by participants -

A	25-30%
V	25-30%
F	~ 45%
- the statistical pattern is different for vowel and coda sites, with fusion rates F -

coda	> onset
short vowel	> long vowel

Statistical patterns for coda sites:

$(m \text{ æ } p \mid m \text{ æ } t) \rightarrow (m \text{ æ } k)_F$

 highest F (~ 60%)

$(m \text{ æ } p \mid m \text{ æ } k) \rightarrow (m \text{ æ } t)_F$

$(m \text{ æ } t \mid m \text{ æ } p) \rightarrow (m \text{ æ } k)_F$

 highest V (~ 60%)

$(m \text{ æ } k \mid m \text{ æ } p) \rightarrow (m \text{ æ } t)_F$

- The labial feature is very visible as lip-rounding in the visual channel
- If there is labiality in the visual channel, it tends to override other perceptions and make participants report a response with labial features
- The same visibility effect can be found amongst voiced plosives (bɪb bɪd bɪg), for example, whether in coda or onset position
- Such patterns suggest that fusion phenomena are related to the relative visibility and audibility of subsegmental features (hence tractable, using a subsegmental framework)

Why use the GP framework ?

1. Its subsegmental primes (elements) have acoustic signatures that make them audible in isolation or in combination
 - e.g. *Ingleby and Brockhaus 2002 for a case statement*
2. At least some of its elements have visual signatures, lip/mouth shapes that make them visible
 - e.g. *Harris and Lindsey 2000 give shape signatures for elements A, I and U*
3. It has been argued that GP elements can represent the mental lexicon that mediates audition and articulation of speech
 - e.g. *Jonathan Kaye's 'Phonology, a cognitive view' 1989... but the argument applies to any framework that can model coarticulation phenomena*
4. GP models coarticulation effects rather elegantly, as assimilation or loss of subsegmental material in the form of elements

GP models of assimilation:

ENGLISH increase

ɪ n k r ɪː s —velarise→ ɪ ŋ k r ɪː s

(*velar element A assimilated from k segment by n to form m*)

ARABIC أنْفُس = souls

ʔ æ n f uː s —labialise→ ʔ æ m f uː s

(*labial element U assimilated from f segment by n to form m*)

GERMAN Band = ribbon

b a n d ø —devoice→ b a n t ø

(*halt-phonation element H assimilated from pause ø by d to form voiceless t*)

INCONGRUENT STIMULUS

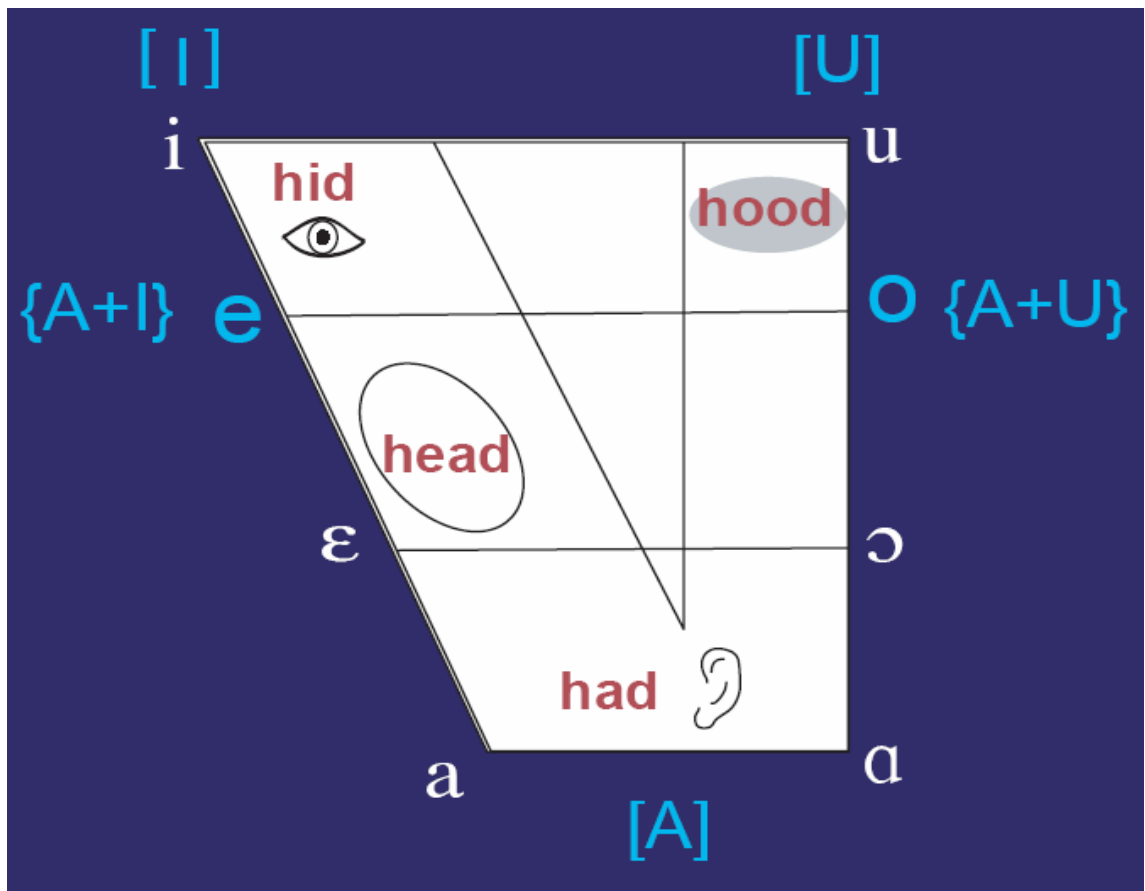
(s ɪː l | t ɪː l) → (θ ɪː l)_F

(b ʌ s | b ʌ t) → (b ʌ f)_F F ~25%

(*noise element h assimilated from the audio channel to join elements I + U from the visual channel to form fricative segment θ in the perception channel*)

- **BUT very few of Azra's experimental results on incongruent speech stimuli fit this assimilation model**

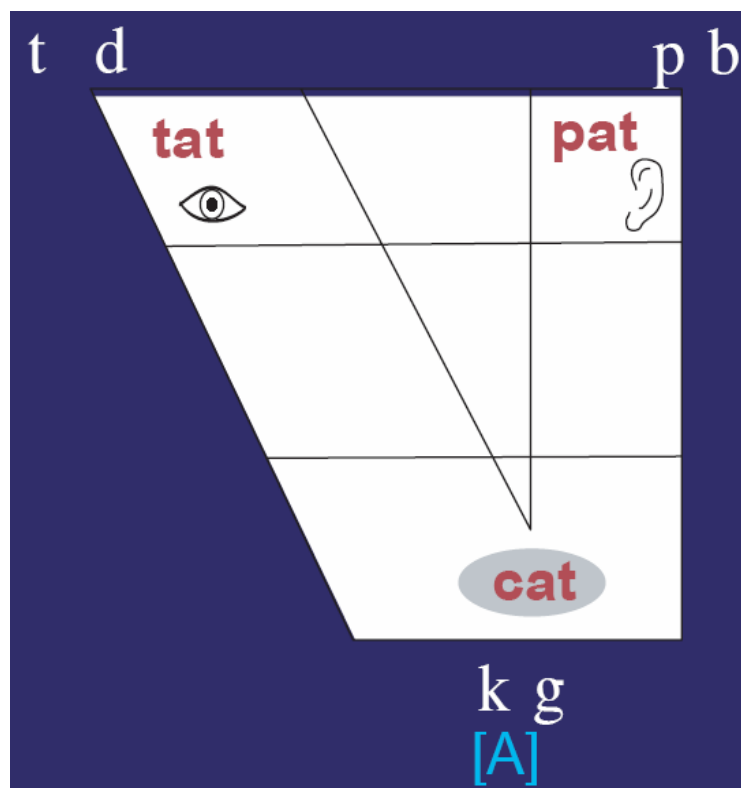
A cancellative type of fusion :



vowel : (h æ d | h I d) → (h U d)_F
 * (h æ d | h I d) → (h e d)_F

- If assimilation occurs, the A from the audio channel and the I from the visual channel would be assimilated in the perception channel to give an e segment
- The fusion that actually occurs involves cancellation of conflicting elements to leave as default the place element U and hence an u in the perception channel

Cancellative fusion at an onset site:



- The labial element U in the audio channel and the palatal element I in the visual channel cancel to leave as default a velar A in the perception channel
 - The default velarity does not appear to come from the vowel nucleus, nor from the other consonant
- stronger evidence from coda fusion

$(t \text{ I } p \mid t \text{ I } t) \rightarrow (t \text{ I } k)_F$

$(p \text{ A } p \mid p \text{ A } t) \rightarrow (p \text{ A } k)_F$ with fusion rate ~60%

(in these cases, there is no velar element in onset or coda or vowel in either audio or visual channel, yet one appears in the perception channel !)

Extreme cases of cancellative fusion:

- In a few cases, incongruence in a segment produces a complete cancellation of material in a segment
- This only happened at onset consonant sites and in sentence context, as in the following

(p ɒ d | t ɒ d) → (ɒ d)_F
(p ɪ l | k ɪ l) → (ɪ l)_F
(b əʊ l d | g əʊ l d) → (əʊ l d)_F

- We are not sure that this is perceptual – it may be an experimental artefact caused by alignment of incongruent segments
- Alignment and relative quality of the audio and visual channels does have an effect on fusion responses of participants
- Therefore, it would be a good idea to place a ‘standard’ sample of incongruent stimuli at a laboratory phonology website so that researchers could replicate, confirm and extend experiments with new participants
- If the original McGurk and MacDonald paper were to be submitted now to Nature, a condition of publication would be the availability of stimuli at the journal website for such replication studies

Design of experiments in this kind of laboratory phonology:

- Participants usually want to please experimenters, so they must not be given clues about what we would like them to perceive
- They are told simply that the video-clips are of variable quality and they should look and listen carefully before reporting what the recorded speaker is saying
- They report their perceptions of a vowel cardinal diagram with many possible response words marked at appropriate positions, and space for alternatives
- Some video clips have congruent audio and visual channels and added noise – of white noise or ‘cocktail party’ types
- The recordings are relabelled by an administrator so that neither experimenter nor participant can know the content until after a session ends
- This kind of double-blind experimental design is mandatory in testing human response to drugs and is considered vital by cognitive psychologists seeking to probe mental models in ‘objective’ ways

Analysis of statistical patterns in results:


- When testing theories in the empirical sciences, it is generally accepted that null hypotheses should be set up and tested using appropriate statistical measures of confirmation
- for example, a null hypothesis on fusion rates for responses to incongruent stimuli might be H_0 - *that the incongruent stimuli rates are the same, on average, as those arising from perception errors for noisy congruent stimuli*

evidence for rejecting this

Means and standard deviations of fusion responses to noisy congruent and incongruent coda experiments with 50 subjects can be put into a t-test

Azra has found, typically, that the observed differences of means could have happened by chance, given H_0 , with probability less than 0.01%

a null hypothesis is rejected if the conditional probability is below 5%



Findings from double-blind experiments and statistical hypothesis tests:

- 1. Average fusion rates in non-branching codas of monosyllabic English words are significantly greater than those for non-branching onsets**

Reference

Ali & Ingleby (2002), Ali (2003b)

- 2. With branching constituents, average fusion rates in both branches of an onset are not significantly different**

Examples

(pru:d | tru:d) → (kru:d)_F with fusion rate ~ 11%
(spɪl | stɪl) → (skɪl)_F with fusion rate ~ 10%

- 3. Average fusion rates in both branches of a coda are not significantly different**

Examples

(bɪbz | bɪdz) → (bɪgz)_F with fusion rate ~ 19%
(wɪsp | wɪst) → (wɪsk)_F with fusion rate ~ 22%

Reference

Ali (2003a)

Emerging results on polysyllabic words:

- work checking the patterns of fusion in polysyllabic word contexts considers

(s l I: p I ŋ | s l I: t I ŋ) → (s l I: k I ŋ)_F

(b ɒ b i | b ɒ d i) → (b ɒ g i)_F etc

In such cases, the fusion rates remain similar to those for a coda without a morphological suffix and significantly different from the rate for an onset.

Note also that the fusion response is a much rarer word than that in audio and visual channels, so the fusion is robust against salience effects.

Reference

Ali & Ingleby (2003)

- In the case of different suffixes,
(s l I p ə | s l I k ə) → (s l I t ə)_F
(f l I p ə | s l I t I ŋ) → (s l I: k I ŋ)_F
(h ɒ p i | h ɒ k i) → (h ɒ t i)_F

the hypothesis that the coda-like fusion rate is suffix independent is not rejected. Affixes like *y* and *-ing* elicit more fusion in the foregoing stem than the agent affix *-er*. It seems that certain affixes are more prone to induce stem perception errors and ‘that morphological information is represented in the mental lexicon in a quite detailed way’

References

Janssen, D. and Humphreys, K. (2002)

McQueen, J.M. and Cutler, A. (1998). (see bibliographic hand-out)

Emerging results on words in sentences:

- Fusion still occurs even when the incongruent data is embedded in a sentence context.
- We begin to see extreme cases of cancellative fusion

Example

‘all that glisters is not (b əʊ l d | g əʊ l d) → (əʊ l d)_F’

‘two peas in a (p ɒ d | t ɒ d) → (ɒ d)_F’

- In sentence context, onsets are more prone to fusion than codas.
- We are able to test the strength of semantic cueing using probabilistic grammars.

Example:

‘No pain, no (b eɪ n | g eɪ n) → (d eɪ n)_F’

- The cueing from a sentence does not block fusion, even when the cue favours the audio or the visual channel.

Knowledge Representations *versus* theories:

- The present use of GP and other subsegmental concerns the representation of knowledge
(...about the phonological processes attested amongst classes of speaker, about distributional constraints associated with licencing, etc.)
- We have extended this representation to model attested responses to incongruent multi-media data, but a knowledge representation falls short of being a scientific theory
- A ***bona fide*** theory must have the power to predict beyond the empirical data used to confirm it
- A possible theory based on our GP models of fusion can be formulated as a claim to language universality of fusion
(already there is data on fusion in words from other languages than English)
- there are, however, other ways of claiming ‘theoryhood’

A cognitive theory of cancellation:

- Audiovisual incongruity is not the only kind that perception psychologists are interested in
- Using the stereo channels of a sound lab, one can create incongruity between left-ear and right-ear audio stimuli, and there is a body of experimentation on such diotic incongruity
- The diotic stimulus work has been concerned with the way people adduce sound source location from diotic incongruence, and the way location helps to pick out a speech signals from noise created at a different location
- We predict that a McGurk cancellative fusion can occur in response to such diotic stimuli
- We are attempting to experiment with a diotic, purely acoustic McGurk effect, but our acoustic equipment has still not been made reliable enough to discuss results

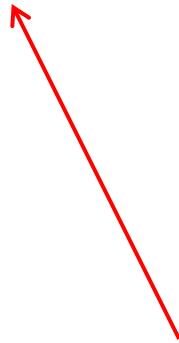
reference M.Ingleby, 2003 *Human Response to Incongruent Sensory Data*

Coda-onset differences and CV languages:

- An abiding pattern in all the contexts that we have investigated is that cancellative fusion rates remain significantly different for codas and onsets
- If a theory of language universality holds, this difference would not be observed in pure CV languages
- In the case of languages for which syllabification is contested, one could use incongruent stimuli to seek for the segments that show too much fusion to be onsets, thereby doing empirical syllabification
- **Arabic** is a case in point
 - The Arabic tradition of Sybawaih, on which the (phonetic) alphabet is founded, uses CV units symbolised orthographically by a consonant with a vowel diacritic
 - The Western tradition of classical scholars, treating Arabic like Latin and Greek, postulate that there are CVC, CVVC, CVCC syllables
 - We have programmed some experiments for June and July, with Arabic phonologist Ali Idrissi of Montreal, to test this using subjects from Saudi Arabia, Egypt and Jordan

Do these native Arabic speakers have codas in their mental models of Arabic speech ?

Ali, Azra, Fatmah and I will let you know !



Fatmah Baothman and I developed a CV model of Arabic speech patterns on GP lines. It represents all the known coarticulation processes of Arabic and leads to a stress-prediction algorithm that is much simpler than those based on the syllabification of the Western classical tradition