Visual Speech

Audiovisual Processing CMP-6026A

Dr. David Greenwood

david.greenwood@uea.ac.uk

SCI 2.16a University of East Anglia

October 28, 2021

Content

- Speech Production
- Visual Speech
- Visemes
- Coarticulation

Speech Production

in a visual context

Speech Production

Speech can be regarded as a *filtering* process.

- Air is expelled from the lungs.
  - the excitation signal
- This air is forced through the vocal tract.
  - the filter
- The air exits via the nose and mouth.
  - the filtered signal

Speech Production

The filter *response* is determined by the vocal tract **shape**, which is dependent on the position of the speech **articulators**.

- The filter is non-stationary since the response changes over time.
- Speech is time-varying in nature.

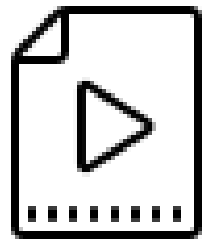
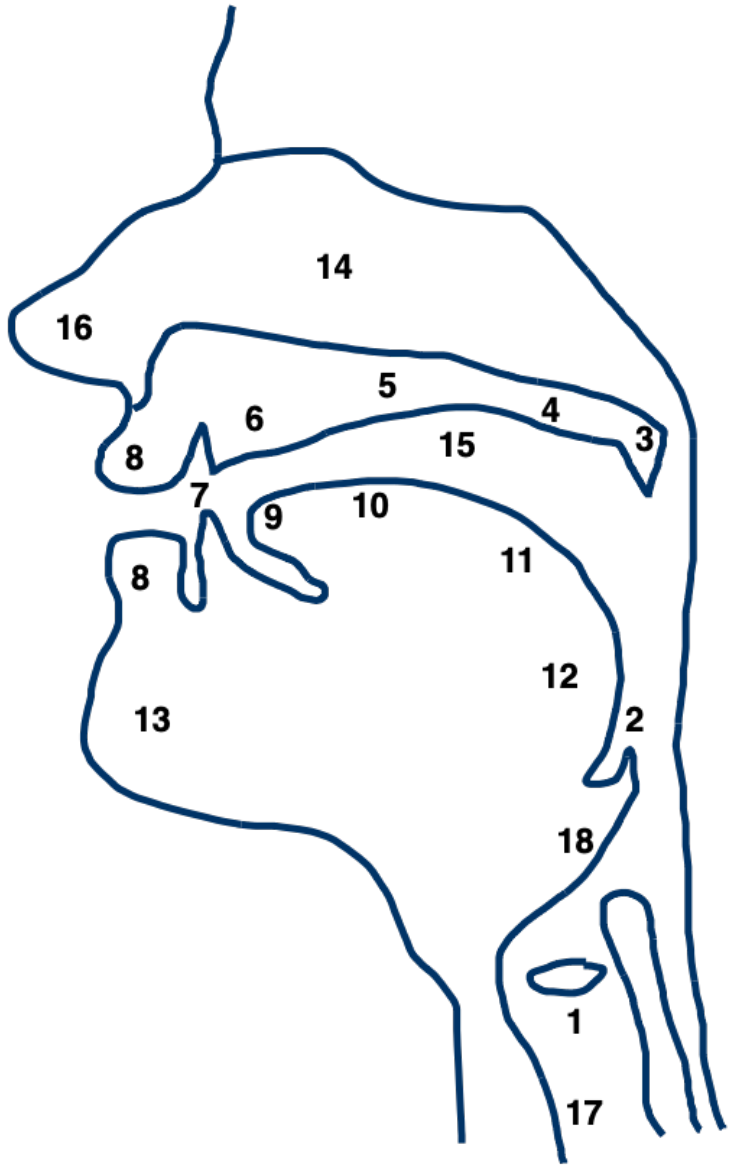


Figure 1: An MRI of the vocal tract.

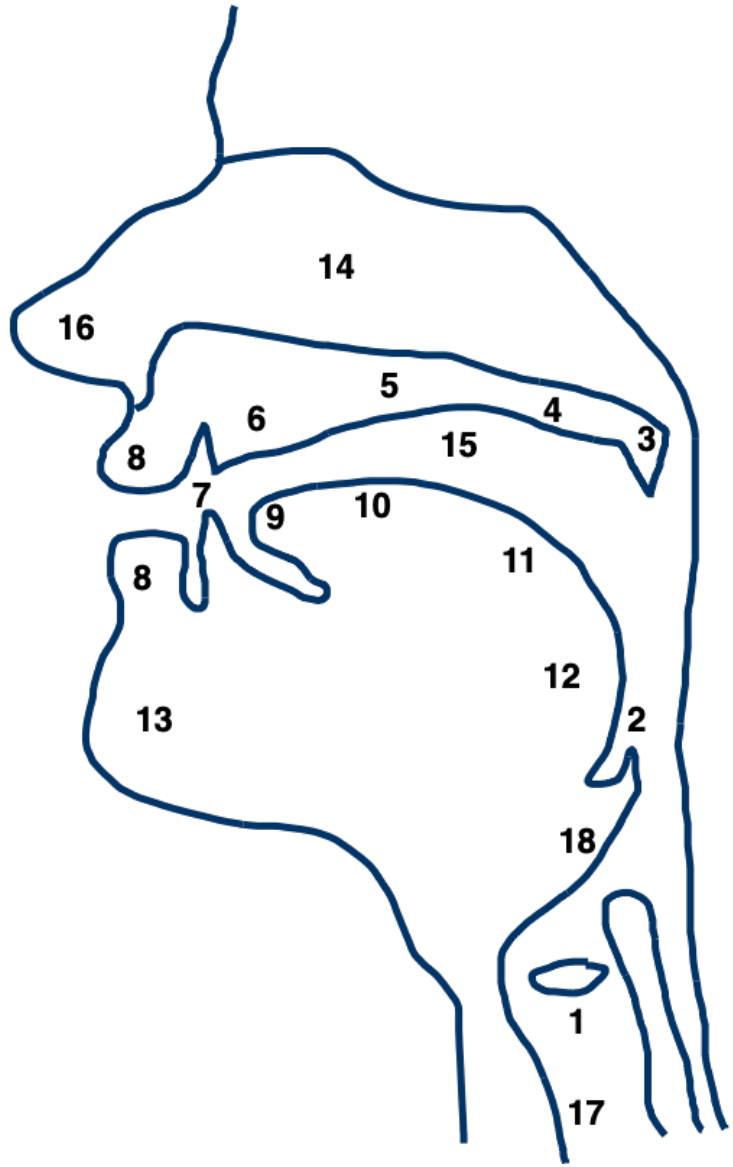
Speech Articulators

1. vocal choords
2. pharynx
3. velum
4. soft palate
5. hard palate
6. alveolar ridge
7. teeth
8. lips
9. tongue tip
10. blade
11. dorsum
12. root
13. mandible
14. nasal cavity
15. oral cavity
16. nostrils
17. trachea
18. epiglottis



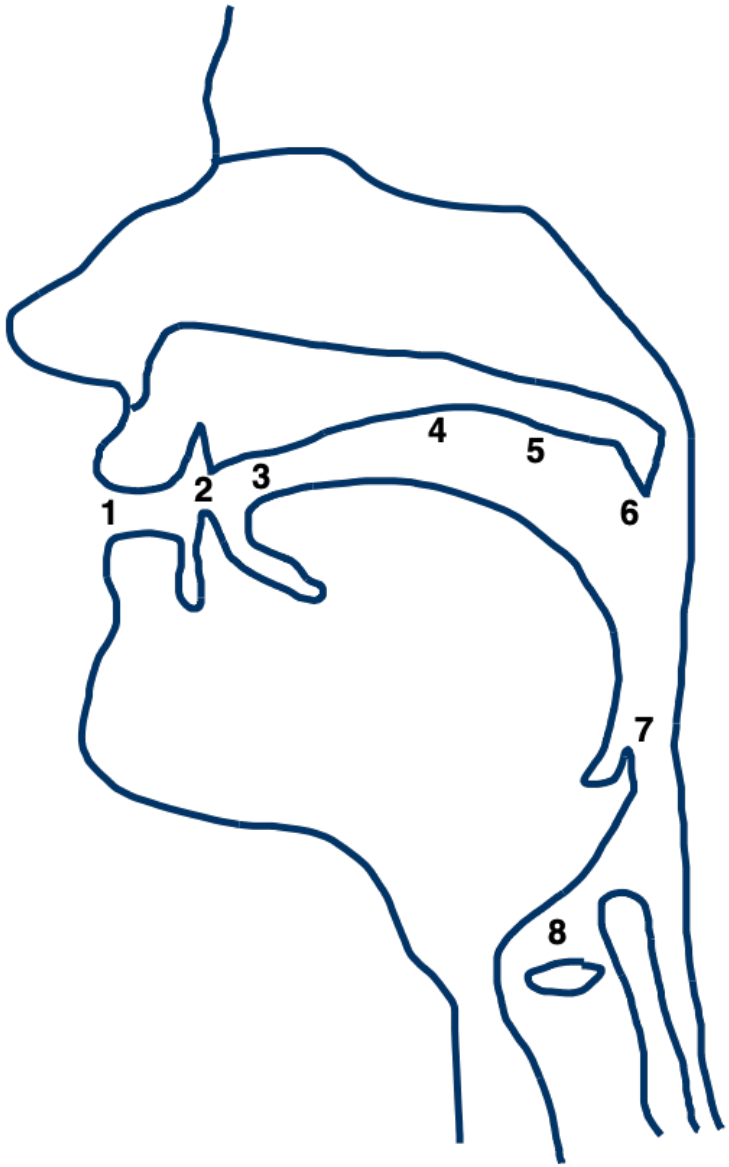
Speech Articulators

1. vocal choords
2. pharynx
3. velum
4. soft palate
5. hard palate
6. alveolar ridge
7. teeth
8. lips
9. tongue tip
10. blade
11. dorsum
12. root
13. mandible
14. nasal cavity
15. oral cavity
16. nostrils
17. trachea
18. epiglottis



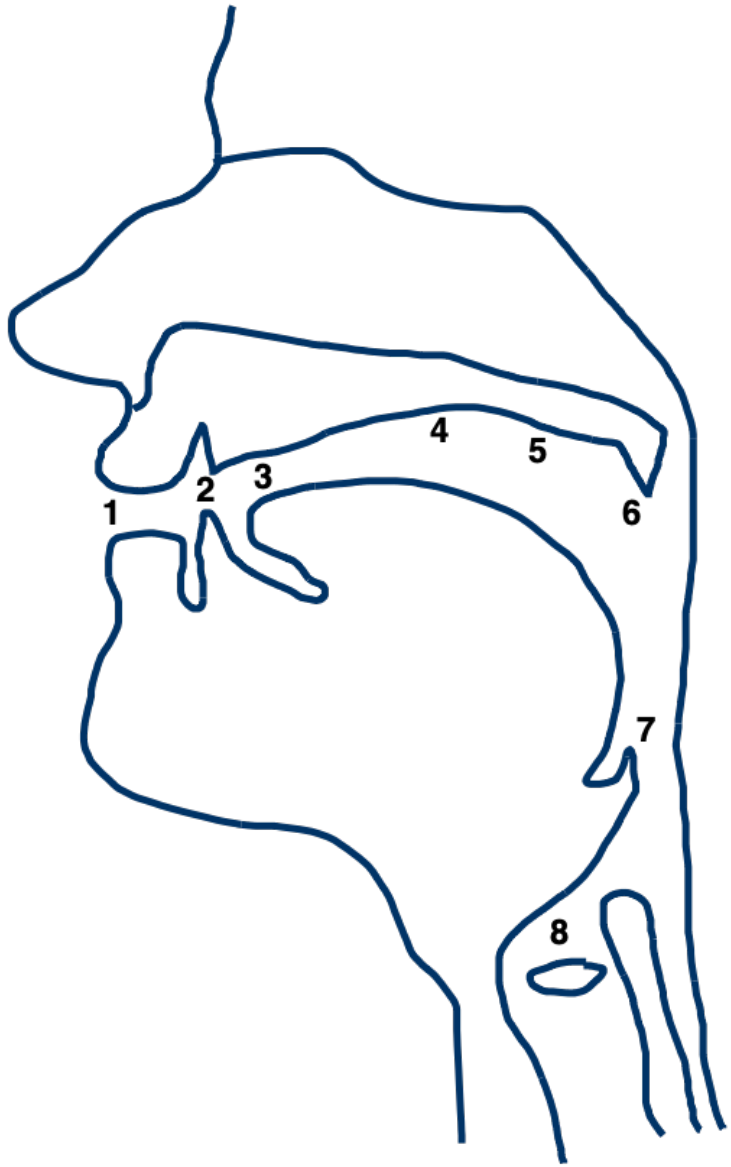
Places of Articulation

1. labial
2. dental
3. alveolar
4. palatal
5. velar
6. uvular
7. pharyngea
8. glottal



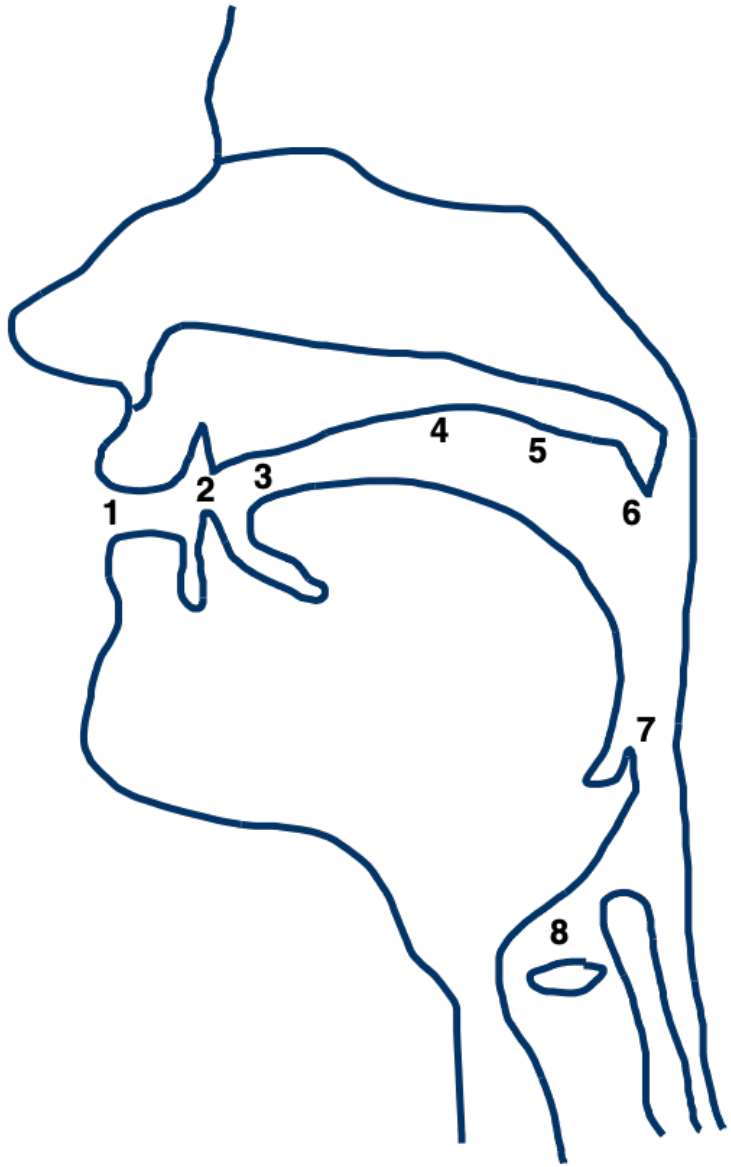
Places of Articulation

1. labial /b
2. dental /t/
3. alveolar /l/
4. palatal /y
5. velar /k
6. uvular
7. pharyngea
8. glottal



Places of Articulation

1. labial /b
2. dental /t
3. alveolar /l
4. palatal /y
5. velar /k
6. uvular
7. pharyngea
8. glottal



Articulation

The **place** of articulation describes *where* a speech sound is formed.

Articulation

The **manner** of articulation describes *how* a speech sound is formed.

Articulation

- Stop
  - a complete blockage is formed along the vocal-tract.
- Nasal
  - airflow can exit through the nose (velum is lowered).
- Fricative
  - a partial blockage is formed causing a turbulent airflow.

Articulation

- Approximant
  - a partial blockage, but insufficient enough to cause a fricative.
- Lateral
  - airflow is blocked along the centre of the vocal-tract.

Note: these manners of articulation are not mutually exclusive.

Consonants

Consonants are characterised by the place and manner of articulation.

Consonants

- /*p*/ is a voiceless bilabial stop (plosive).
- /*m*/ is a voiced bilabial nasal.
- /*f*/ is a voiceless labiodental fricative.
- /*k*/ is a voiceless velar stop.
- /*j*/ is voiced palatal lateral approximant.

Vowels

For vowels the airflow is relatively unobstructed.

Vowels

Vowels **cannot** be characterised by the place or manner of articulation.

Vowels

Vowels **are** characterised by:

- The degree of lip-rounding.
- The front to back position of the high-point of the tongue.

Vowels

Diphthongs are the *concatenation* of two vowels.

Visual Speech

Speech is about more than just sounds.

Visual Speech

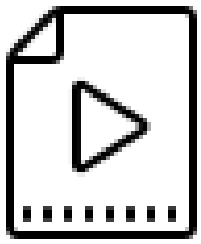
- The formation of *some* speech can be **seen**.
- We all use visual speech to help disambiguate similar sounds.
- In a noisy environment you tend to watch the person you are speaking with more closely.

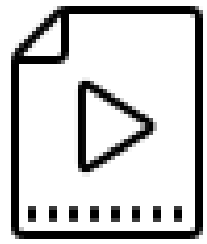
Visual Speech

Speech formation can be felt.

Some deaf-blind people use the **Tadoma** method of communication.

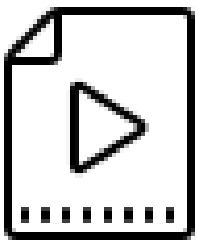
Visual Speech

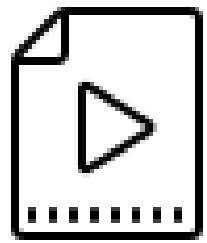




Can you discriminate between “dog” and “bog”, in noisy audio?

Visual Speech





Can you discriminate between “dog” and “bog”, when the articulators are visible?

Visual Speech

Audiovisual speech is *complementary* in nature.

- Sounds that **sound** similar often look different

  eg. /b/, /d/, /m/, /n/, /f/, /s/
- The formation of sounds that **look** the same sound different

  eg. /f/, /v/, /s/, /t/, /b/, /p/

Visual Speech

Visual information provides an effective improvement of $\approx 11dB$ in signal-to-noise ratio.

Visual Speech

Vision can improve understanding of hard-to-understand utterances.

Visual Speech

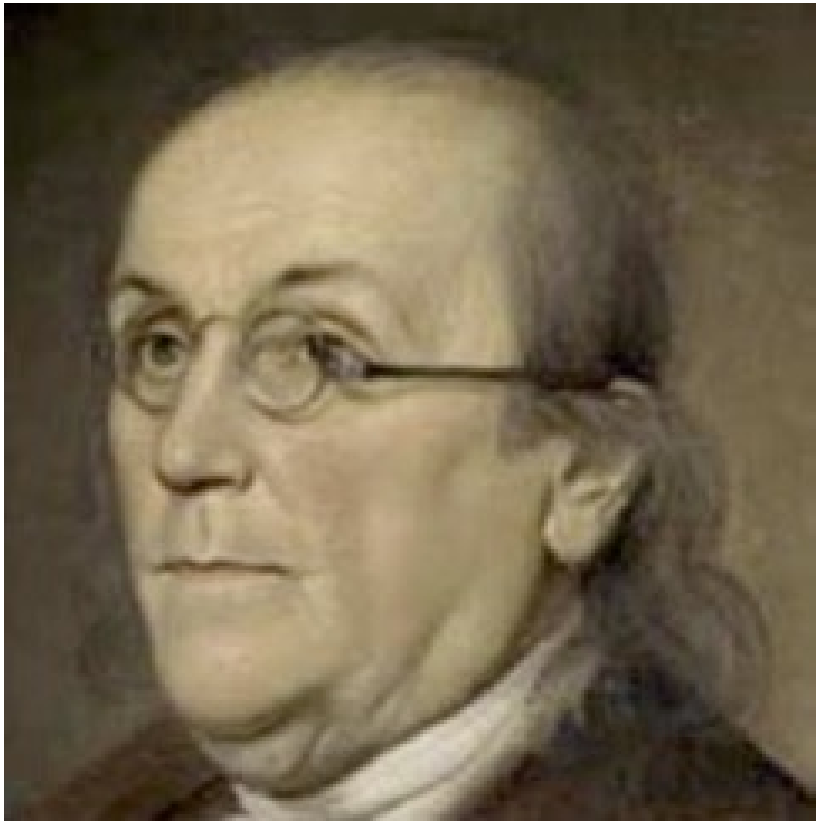


Figure 2: Benjamin Franklin

Benjamin Franklin invented bi-focal spectacles to help better understand French!

"... since my being in France, the glasses that serve me best at table to see what I eat, not being the best to see the faces of those on the other side of the table who speak to me;
... and when one's ears are not well accustomed to the sounds of a language, a sight of the movements in the features of him that speaks helps to explain...
so that I understand French better by the help of my spectacles."
– Benjamin Franklin, in 1785

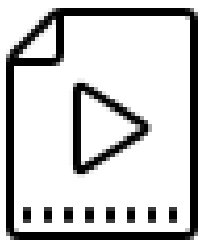
McGurk Effect

Visual speech can **alter** our perception of a sound.

This is illustrated by the **McGurk** effect.

McGurk & MacDonald, Hearing lips and seeing voices. 1976

McGurk Effect



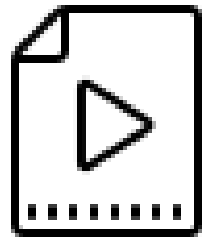
- you hear “baa” . . .
- you see “gaa” . . .
- you perceive “daa” . . .

McGurk Effect

Auditory “baa” with visual “gaa” is often perceived as “daa”.

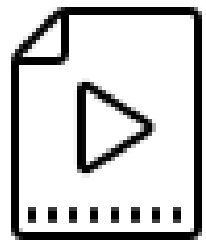
- What is perceived is neither seen nor heard!
- happens even when the viewer is aware of the effect
- The effect persists across age, gender and language.

McGurk Effect



“baa” or “faa”?

McGurk Effect



“Bill”, “pail”, “mayo”?

McGurk Effect

Also on YouTube:

- https://youtu.be/KiuO_Z2_AD4
- https://youtu.be/xlXaNJR-1Oo
- https://youtu.be/G-IN8vWm3m0

Visemes

- The basic building block of auditory speech is the **phoneme**.
- The closest visual equivalent is the **viseme** (visual phoneme).

Visemes

- The mapping from phonemes to visemes is **many-to-one**.
- Many phonemes map to the same viseme.

Visemes

- Visemes are usually derived using *subjective* experiments.
- Viewers are asked to identify the consonant in isolated nonsense words.

Visemes

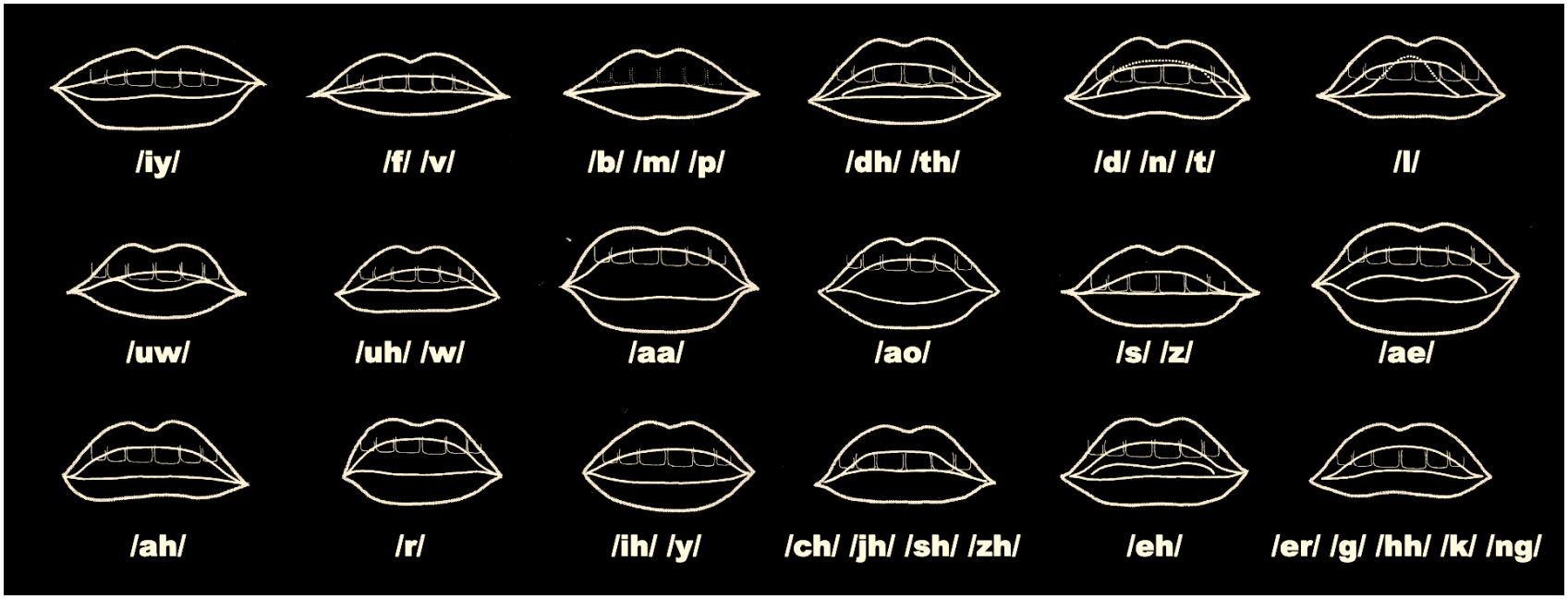


Figure 3: F.Parke and K.Waters, Computer Facial Animation, A K Peters, 1996.

Coarticulation

Coarticulation

Phonemes are abstract representations of sound.

Coarticulation

- We could think of speech as being a string of phonemes.
- Each has an idealised articulator configuration
- Speech is produced by smoothly varying from one vocal tract configuration to the next.

Coarticulation

WRONG!!

Coarticulation

The articulator positions **do not** depend only on the current sound.

- Neighbouring sounds influence each other.

Coarticulation

The articulators never reach their *ideal* target.

- They only move close enough to *approximate* the required sound.
- What you see is a by-product of this.

Coarticulation

This is known as **coarticulation**.

Coarticulation

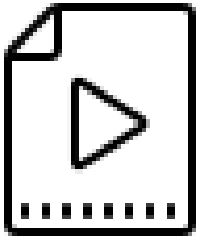
There are two forms of coarticulation:

- anticipatory coarticulation
- perseverative coarticulation

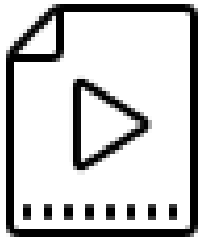
Coarticulation

The same phoneme in different contexts both sounds and **looks** different.

/k/



/t/



Models of Coarticulation

There is no definitive model of coarticulation.

Look Ahead Model

One is the “Look ahead” model.

Speech gestures begin as early as possible provided there are no constraints on the articulators.

Look Ahead Model

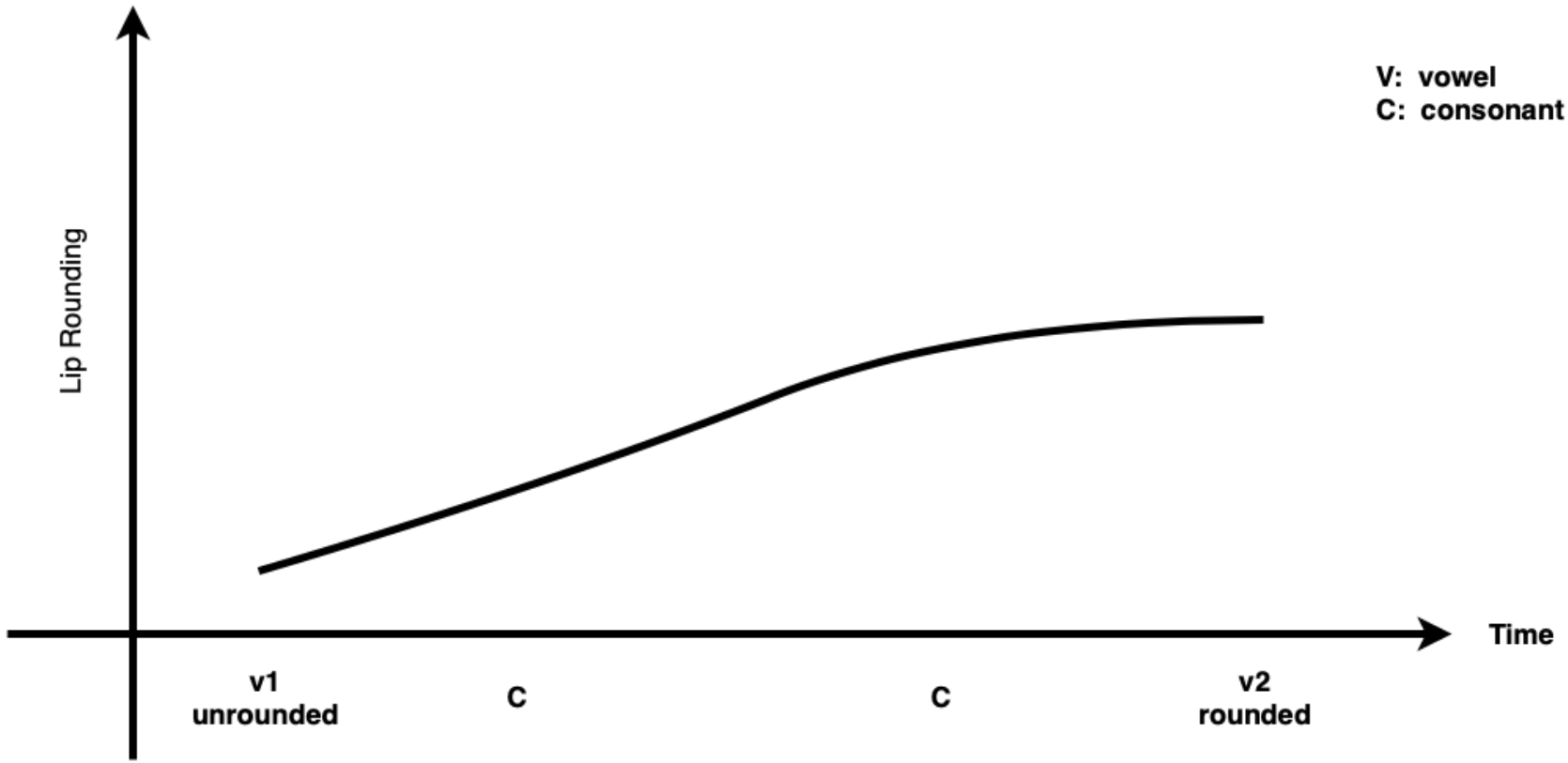


Figure 4: Look Ahead Model

Look Ahead Model

The look ahead model assumes lazy speech production and allows gradual transitions between speech targets.

Temporal Model

An alternate model is the temporal model.

The temporal model assumes that speech gestures begin at a fixed time prior to the onset of a sound.

Temporal Model

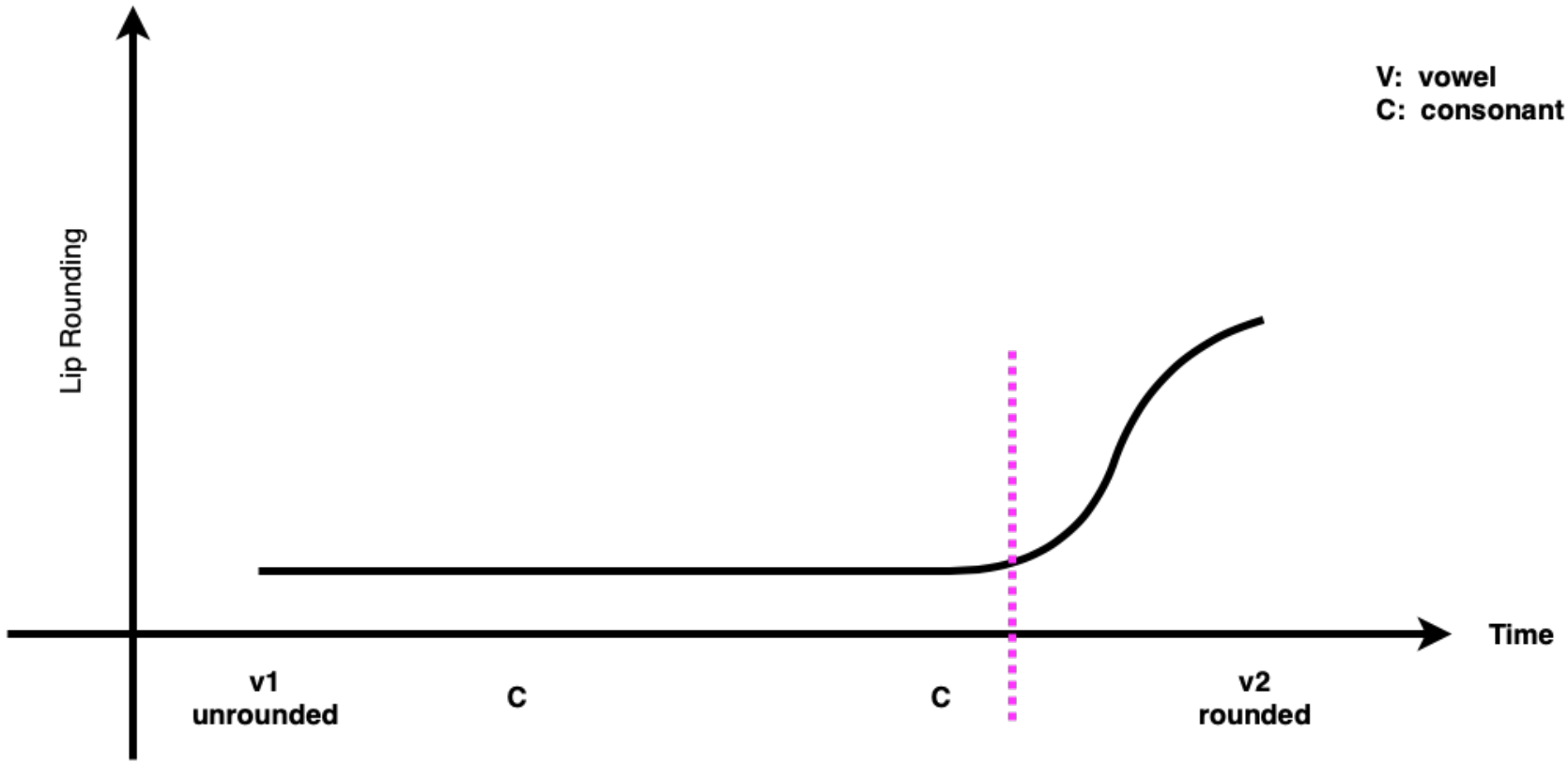


Figure 5: Temporal Model

Temporal Model

The temporal model assumes that speech gestures are largely independent and that speech is the superposition of the gestures.

Hybrid Model

There are also hybrid models:

- Combine both the look ahead and temporal models.
- Initial movement is gradual and starts early.
- Later movement is more rapid, at a fixed time in advance of the pose.

Gestural Model

- A phoneme is represented by a set of **dominance** functions for each articulator.
- The function specifies how dominant an articulator is at different points in time during the articulation of a sound.
- The dominance increases to a peak and then decreases over time.

Gestural Model

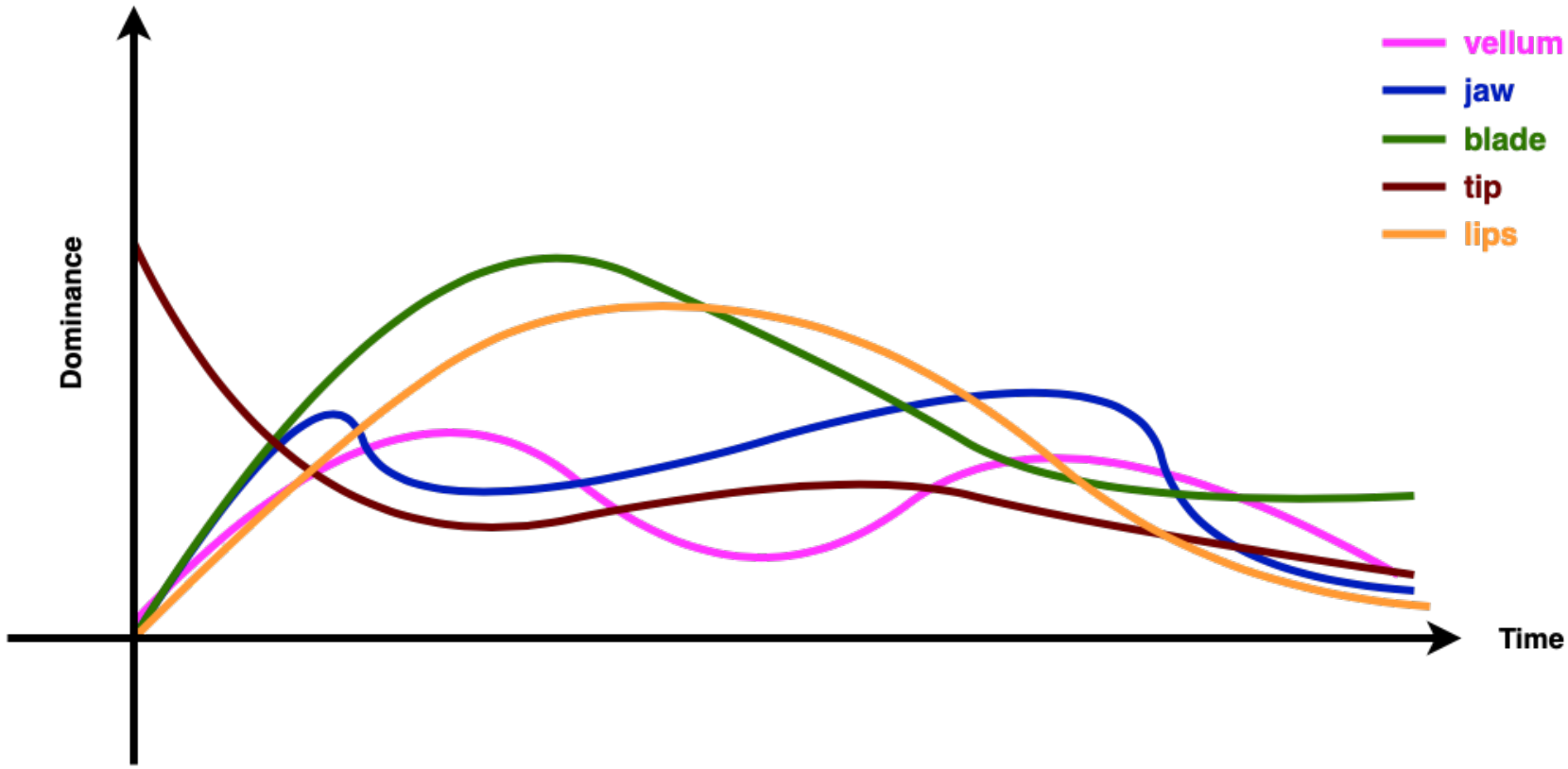


Figure 6: Gestural Model

Models of Coarticulation

- Different coarticulation models exist because different studies use different experimental conditions and linguistic factors.
- Each model might fit the particular conditions for a given experiment.
- The lack of a formal definition of a viseme and a definitive model of coarticulation make recognition (and synthesis) of visual speech difficult!

Summary

Speech is multi-modal in nature!

Summary

A view of the articulation is useful for disambiguating similar sounds.

To a limited extent we all *lip-read* regardless of our awareness.

Summary

Visual speech is poorly defined compared with acoustic speech.

- A viseme is assumed to be the visual analogue of the phoneme.
- Coarticulation means that visemes as lip shapes are not a good unit.
- The same sound has many different visual appearances.