Spectrogram reading and Coarticulation

Yves Laprie

Spectrogram reading

- The steps:
  - Display 2 seconds of speech and choose an appropriate frequency range (0-8000Hz). Do not change these parameters too often in order to keep the same scales in mind.
  - Segment the signal in broad classes (vowels, stops, fricatives) and then sonorants and semi-vowels
  - Then start from regions which can be easily interpreted and extend phone identification.
- Process vowels by comparing them with /i,a,u/:
  - Search vowels with maximal F1 over the sentence (likely to be close to /a/), minimal F2 (probably close to /u,o,i/), maximal F3 (probably close to /i/)
  - Search places where F1-F2 get close /u,o/, F2-F3 /y/ or else F3-F4 /i/
  - Consider formant transitions.

Vowels

front

back

F2

close

/i/

/y/

/u/

open

F1

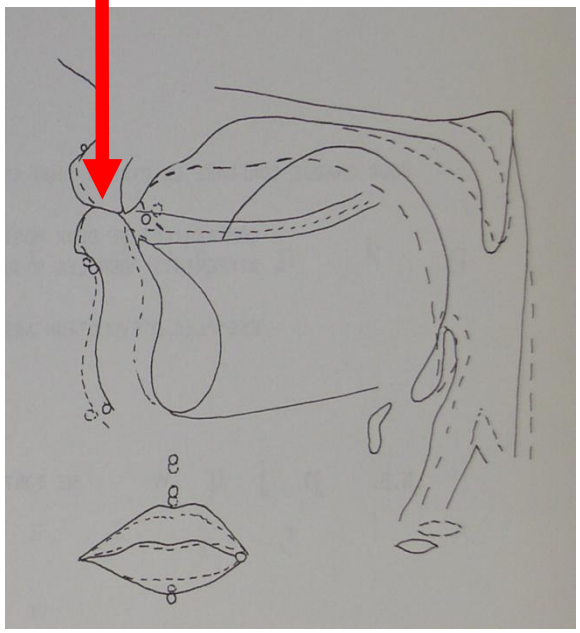
/a/

Effect of occlusives on vowels (CV)

- Labialization lengthens the vocal tract and as a consequence formant frequencies lower near the consonant.
- In general bursts of /p/ are shorter that those of /t/ and /k/.
- There is aspiration in many languages (not in French) which substantially increases the burst duration.
- F2 and F3 of central vowels get closer in the neighbourhood of the consonants /k,g/
- For back vowels there is often a peak in the burst in front of F2 for /k,g/.
- /t,d/ present a locus (called dental locus) for F2 between 1500 and 2000 Hz. There is thus a strong transition for F2 when the following vowel is a back vowel like /u,o/.

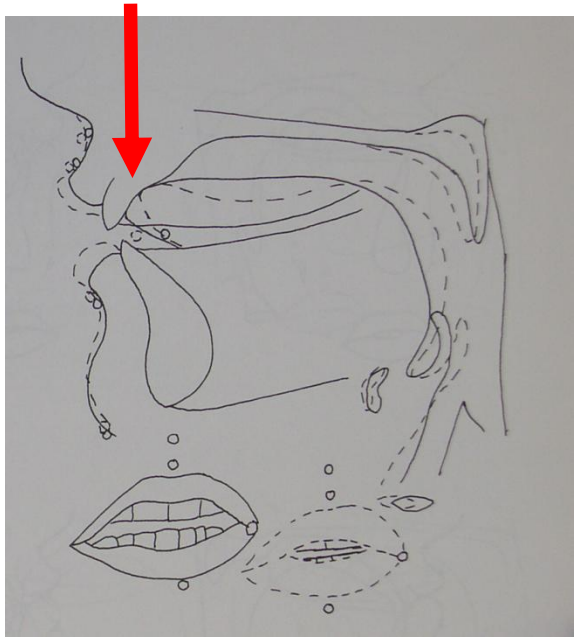
Placetof articulation of French stops

/p/



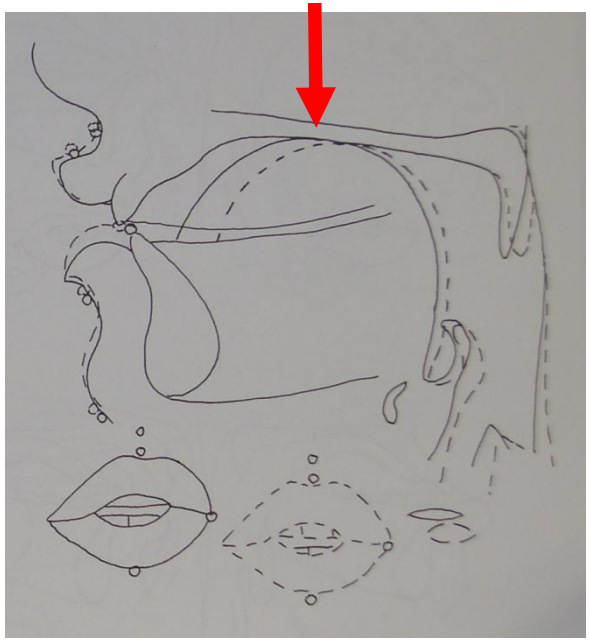
/pət,pRɛ/
dotted, solid

/t/



/pat,tab/
dotted, solid

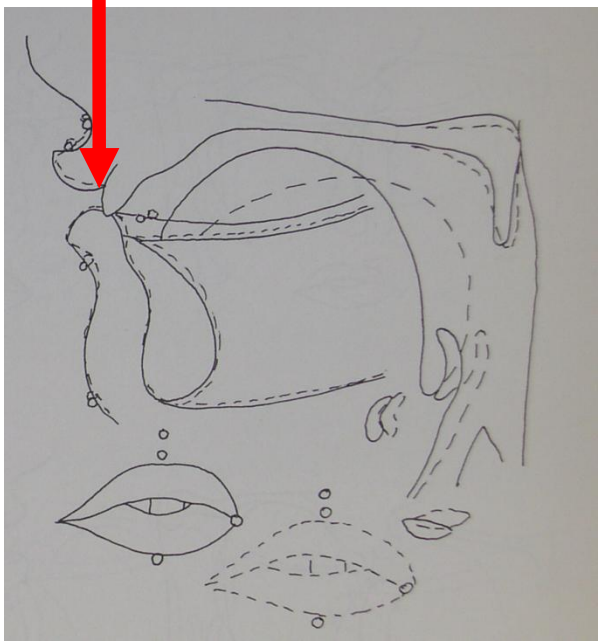
/k/



/ky,ku/

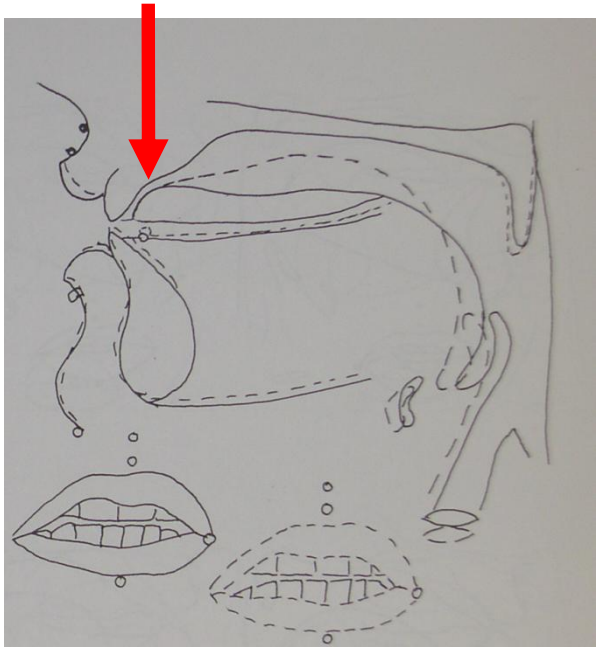
Place of articulation of French fricatives

/f/



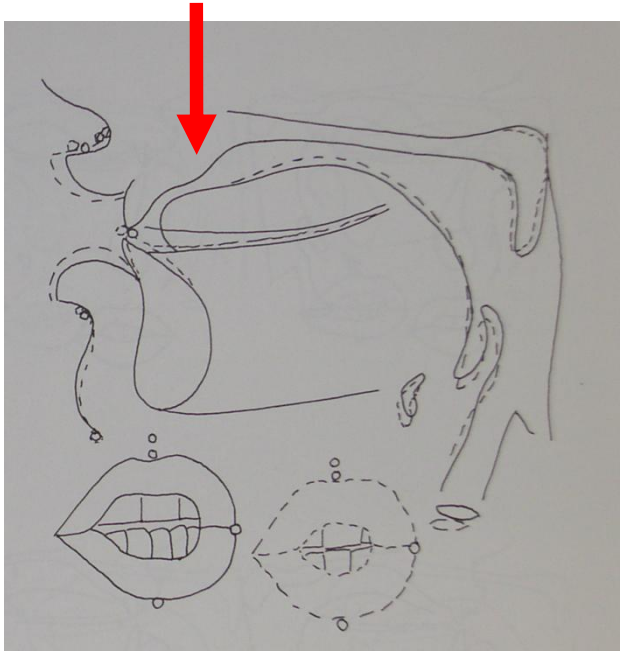
/fym,fɔr/

/s/



/si,sa/

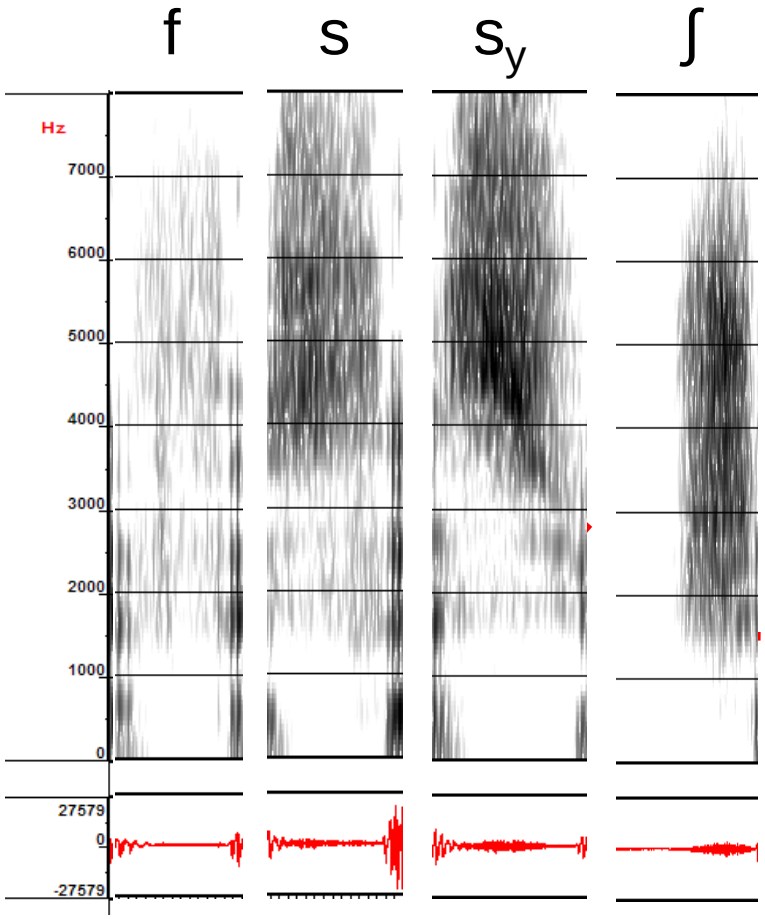
/ʃ/



/ʃø, ʃu/

French fricatives

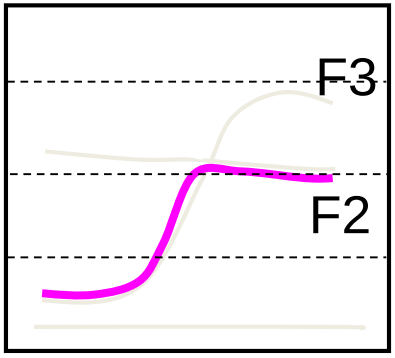
The lower frequency boundary, the intensity and the shape of the frication noise have to be taken into account.



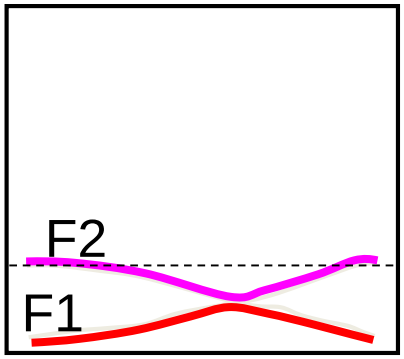
French sonorants et semi-vowels

- Sonorants
  - /R,l/ are highly variable, and especially /R/
  - /m,n/ nasal, closed mouth at lips for /m/, at the alveo-dental level for /n/
- Semi-vowels
  - /w,j,ɥ/ voiced sounds close to /u/, /i/ et /y/, also known short /u/, /i/ and /y/.

Some other acoustic cues



This pattern is characteristic of a back-front transition (/ui, waj/ …) and the flipped version of (/iu, ju/ ...)

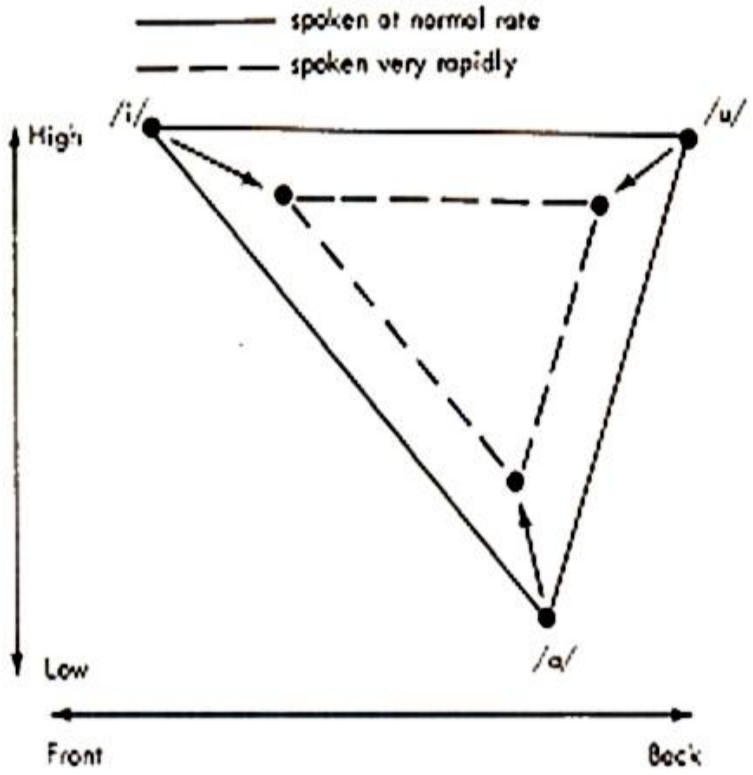


This pattern is characteristic of one /R/ between two vowels.
Be careful, /R/ is the phoneme which gives rise to the highest articulatory/acoustic variability.

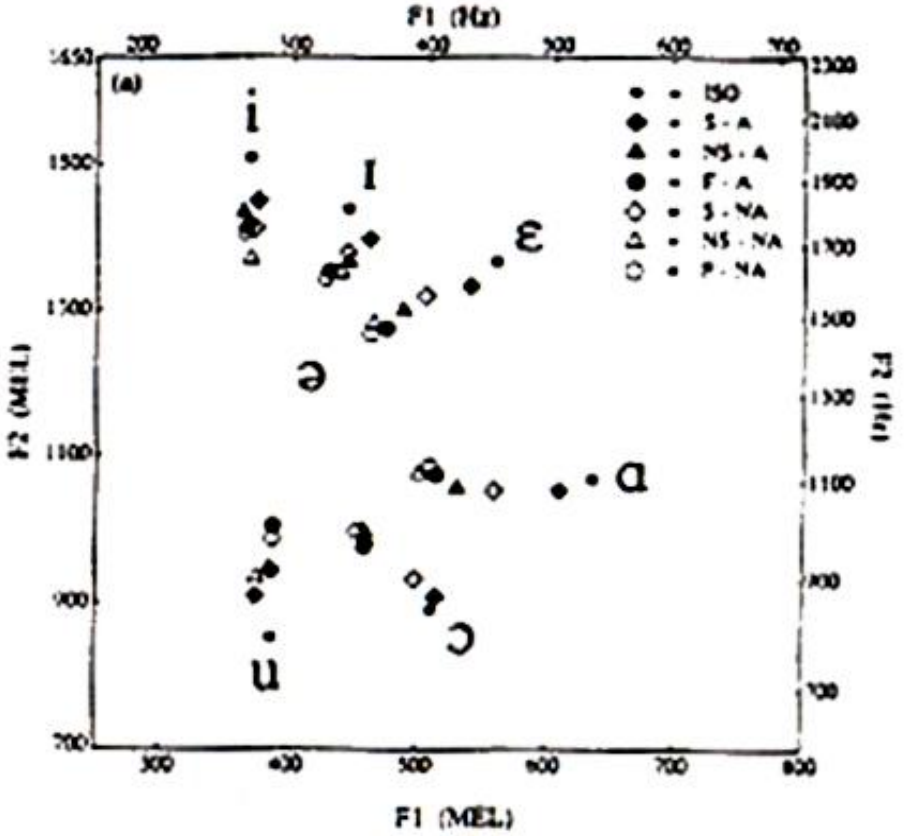
- An abnormaly long burst may be due to the presence of a consonant cluster /kr,pr,pl/…
- There are many other acoustic cues...

Vocalic reduction and vowel neutralization

When the speech rate is fast or when speech is not well articulated formants get close to that oe the neutral vowel (schwa).



From Daniloff

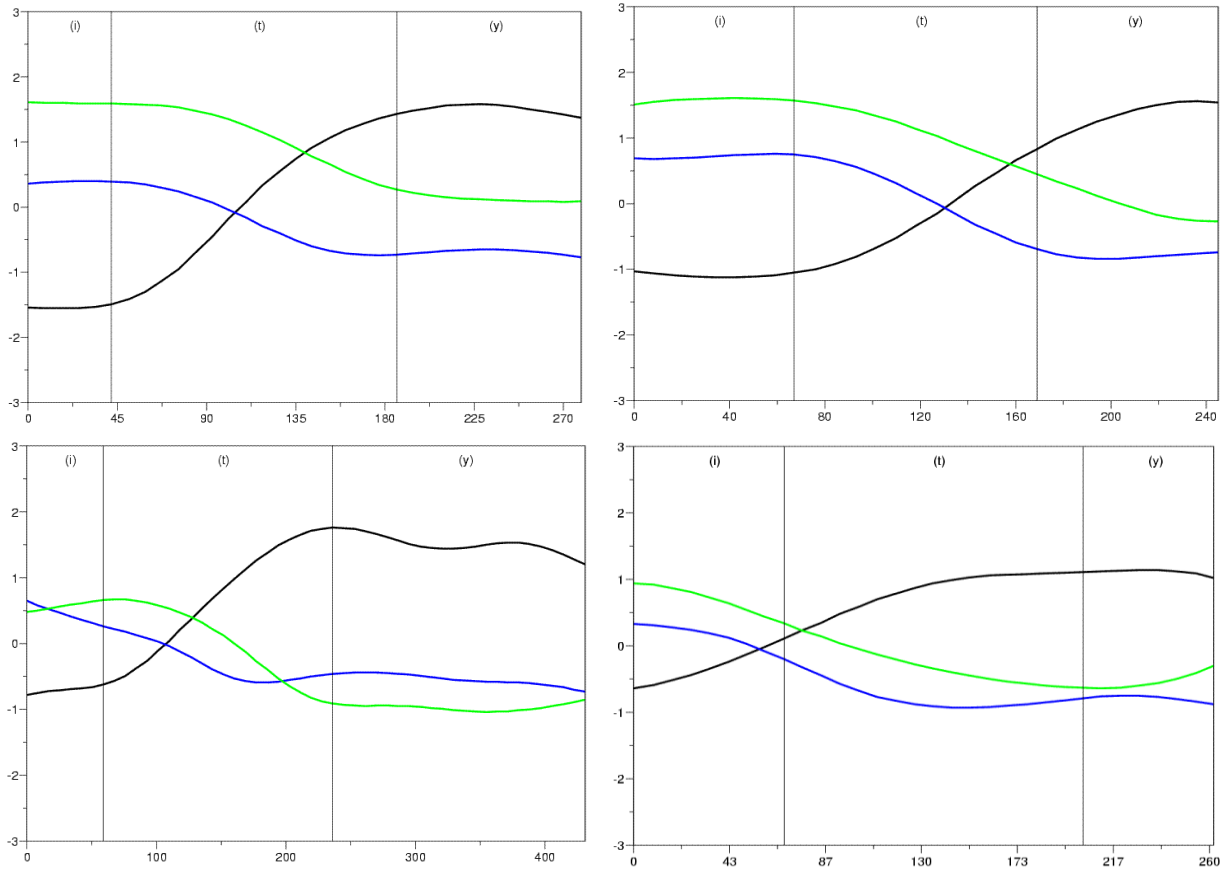


From van Bergem 1999

Coarticulation

- The production of one sound is influenced by that of neighbour sounds (after and before as well):
  - This is due to the articulator characteristics (inertia and dynamics linked to muscles involved) and to the planification of production tasks aiming at minimizing the speaker's articulatory effort.
  - Coarticulation makes automatic speech recognition and synthesis difficult.
- Coarticulation may be:
  - Anticipatory (influence of the next sounds to articulate)
  - Carry-over (influence of the sounds already articulated)
- Three approaches:
  - « Look-ahead » proposed by Henke (1966) and used for French by Benguerel & Cowan (1974). Anticipation starts as soon as possible.
  - « Time-locked » proposed Bell-Berti and colleagues (Bell-Berti & Harris; 1981 ; Boyce *et al*, 1990). The duration of anticipation is determined by the response time of articulators.
  - « Hybrid » propoed by Perkell and Chiang (1986) with variants for instance the expansion model propoed by Abry & Lallouache (1991 ; 1995).

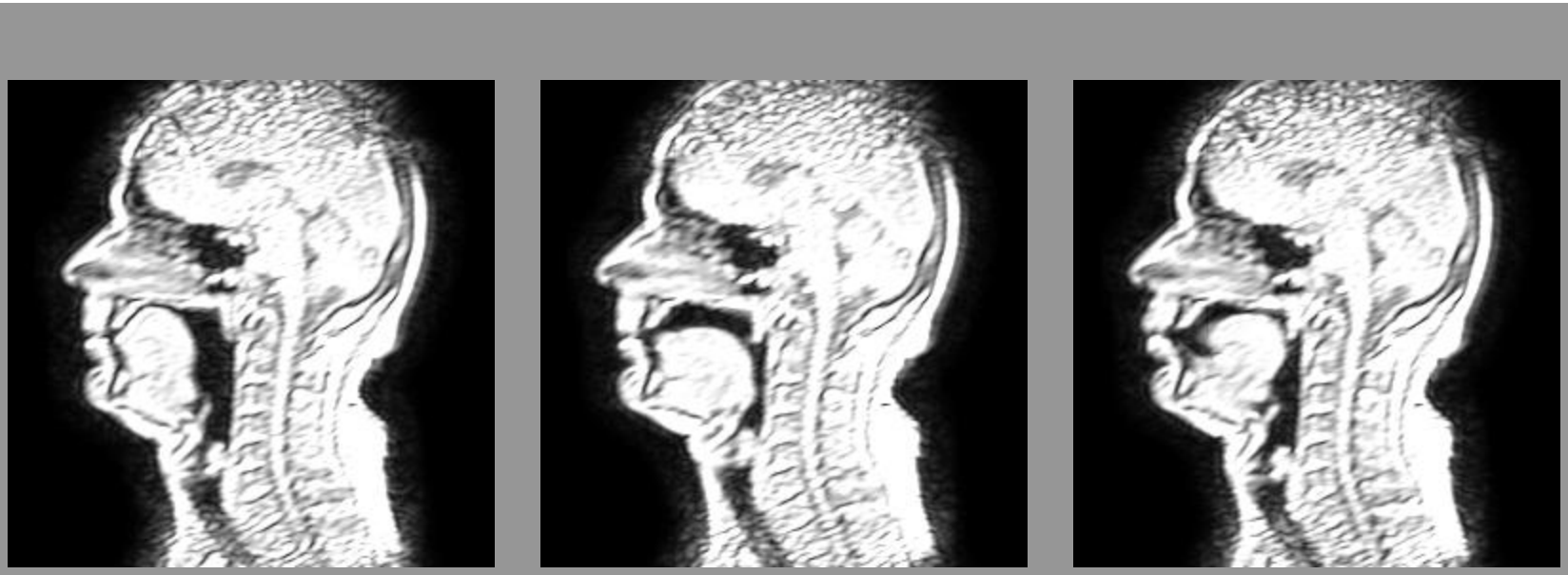
One example: interspeaker variability of the labial anticipation for /ity/



- Anticipation more or less marked:
  - Variable onset
  - Variable duration
- The maximum of protrusion either just before pr during /y/
- Invariant :
  - Anticipation
  - Protrusion of /y/

Tree examples (from a presentation of B. Kröger)

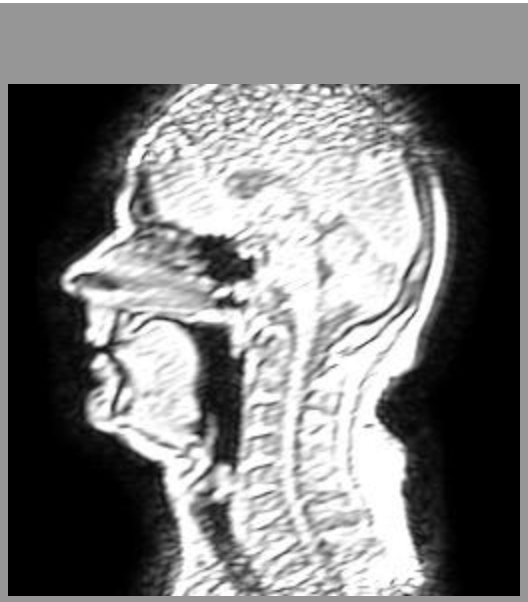
Dynamic MRI-data: [b] (from Kröger)



[ibi] **[aba]** **[ubu]**

Strong coarticulatory positioning of the tongue

Dynamic MRI-data: [d] (from Kröger)



[idi]

[ada]

[udu]

-> gives an idea: what is primary consonantal articulation; what is (vocalic) coarticulation

Dynamic MRI-data: [g] (from Kröger)

[igi]

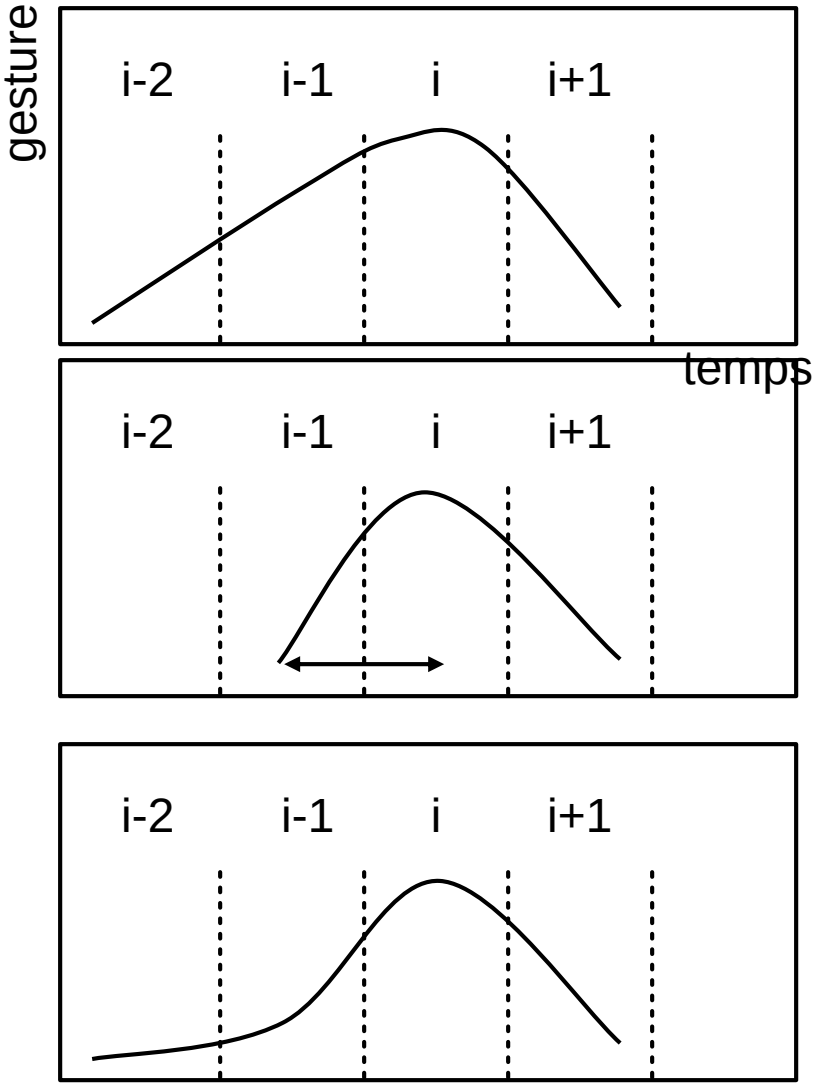
[aga]

[ugu]

Some elements for reflection

- There is always a marked anticipation.
- The maximal value of the articulatory parameter is often reached at the sound onset, or even slightly before.
- Involve very different articulators:
  - Tongue tip also called tongue apex (fast because light)
  - Velum (ight)
  - Lips (fast and light)
  - Tongue body and mandible (more massive)
- It is assumed that articulators ere independent from each other (lips and tongue tip for instance). This is true in a fist approach.
- Even if there are constant antipication movements there exists a great speaker variability... which was often masked by the small number of subjects studied.
- Speakers are consistent for their own coarticulation strategy.
- … data cannot be acquired that simply.

Three coarticulation models used for labial coarticulation



1.Look-ahead

2.Time-locked

3.Hybrid

Look-ahead model

	S1	S2	S3
F	0	0	+
G	+	0	-
H	0	-	+

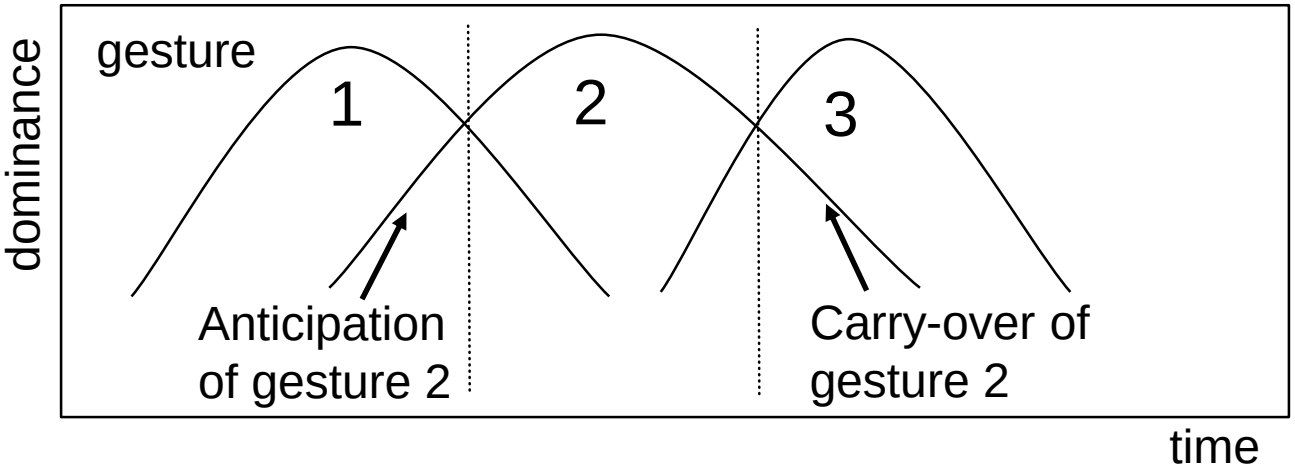
Phonlogical representation

S1	S2	S3
+	+	+
+	-	-
-	-	+

Phonetic representation

- Model inspired by phonology. Features propagate from right to left unless there are not blocked.
- But :
  - coarticulation often starts during the sound which stops the feature to anticipate.
  - vowels are often influenced by transcosonantal vowels in VCV.
  - anticipation has temporal limit even if it can start very early.
  - Segment not specified for one feature are not completely free w.r.t. this feature.

Modèle « Time-locked »



- A model inspired by physics
- The duration of a given articulatory gesture is relatively constane (for a given speaker) because of dynamic constraints on speech articulators.
- Anticipation is thus relatively constant w.r.t. the acoustic onset of the sound to realize.
- Articulatory gestures of the sound which comes before and after overlap with the current gesture (coproduction).

Hybrid models

- Two stages during (labial) anticipation:
  - As soon as possible (≈ look-ahead)
  - A second stages linked to the vowel to reach (≈ time-locked)
- Partially contradicted by the velum coarticulation (Bell-Berti et Krakow). The two part might correspond to two targets.

The expansion model Abry-Lallouache (1995)

- An incompressible duration for the anticipation duration given by the VV transition.
- The anticipation duration increases with the time available (time given by the consonants between the two vowels).
- Slopes (articulatory/time) are speaker dependent.
- This model is more pragmatic than its predecessors because is allows interspeaker variability.

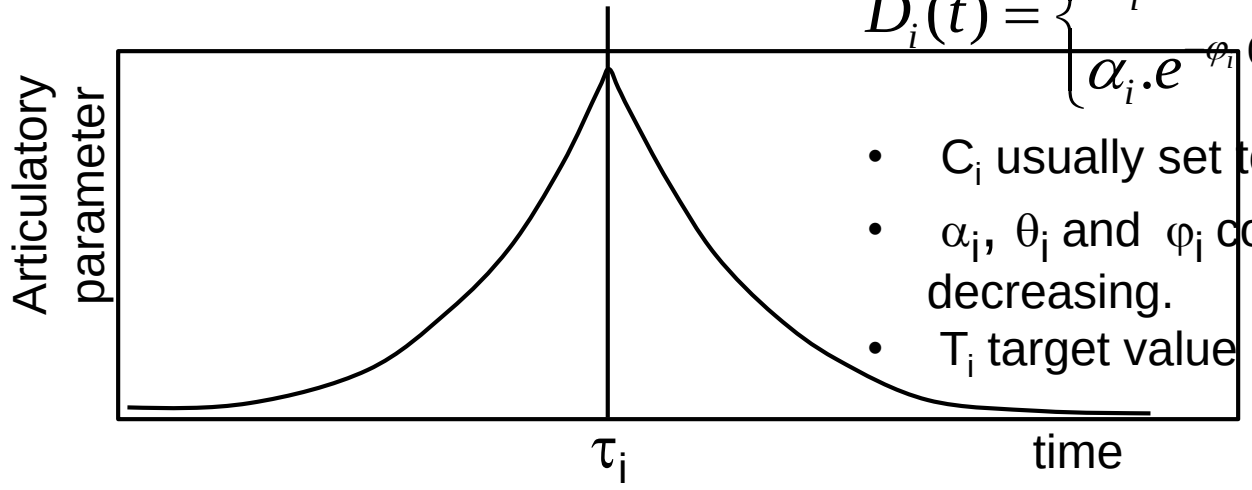
Two models frequently implemented

- Cohen and Massaro:
  - Utilizes dominance functions
  - Widely used to pilot labial coarticulation.
- Öhman → accorde un rôle spécifique aux consonnes et aux voyelles

Cohen & Massaro model (1/3)

- Inspired by the gestural model of production proposed by Löfqvist (overlapping of articulatory gestures)
- Utilization of dominance functions: one for each segment and each articulatory parameter.
- Parameters of the dominance function:

$$D_i(t) = \begin{cases} \alpha_i . e^{-\theta_i (\tau_i - t)^{c_i}} & t < \tau_i \\ \alpha_i . e^{-\varphi_i (t - \tau_i)^{c_i}} & t \geq \tau_i \end{cases}$$



- C_i usually set to 1
- α_i, θ_i and φ_i control the dominance decreasing.
- T_i target value

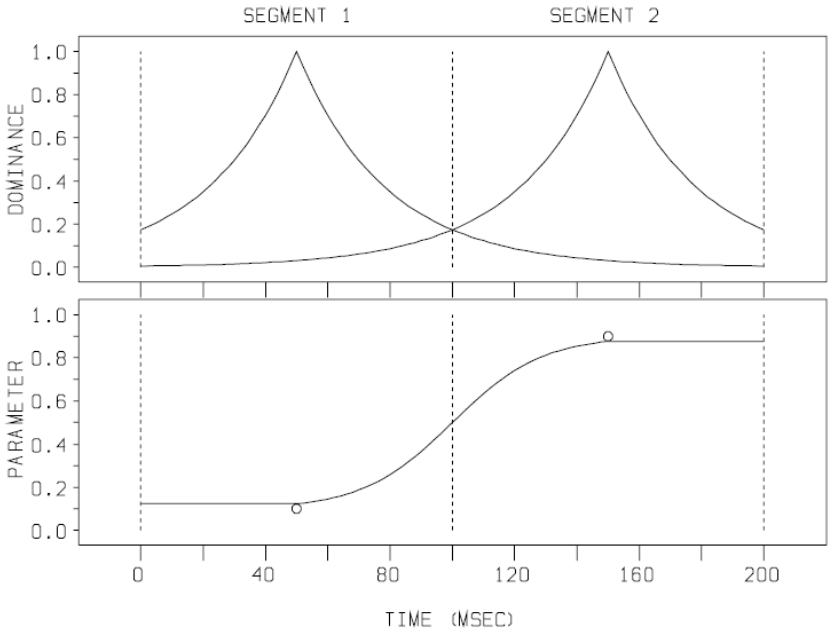
Cohen & Massaro model (2/3)

From dominance functions to articulatory parameters:

Dominance function

Articulatory parameter (T_i is the target value and D_i the dominance of segment i).

$$z(t) = \frac{\sum_{i=1}^{N} T_i D_i(t)}{\sum_{i=1}^{N} D_i(t)}$$



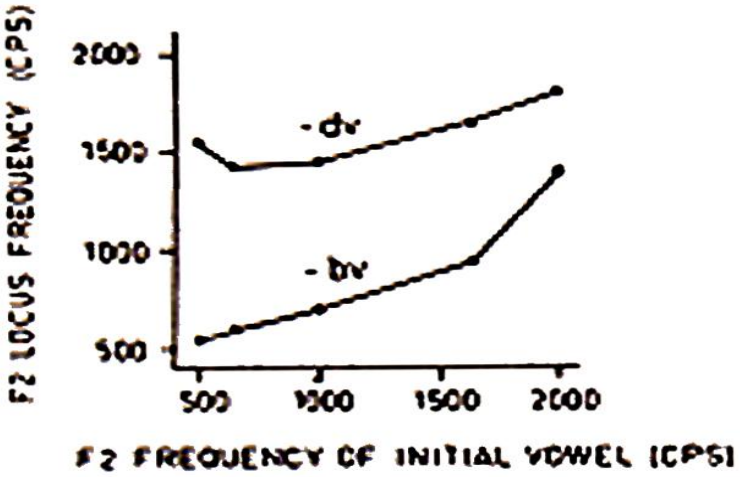
From Cohen & Massaro

Cohen & Massaro model (1993) (3/3)

- Implementation:
  - Requires 5 parameters to learn for each segment and each articulatory feature (viseme or phoneme).
  - Widely used for labial coarticulation (Beskow 2004 for instance)
  - Requires an important training corpus
  - Training by optimizing parameters (either directly or by using the gradient that can be computed easily).
- Weaknesses :
  - There exist corpora for lip coarticulation but very few for the vocal tract (and not with a sufficient size).
  - No contextual effect taken into account (for instance : the target for /k/ is the same for /u/ and /i/).
  - Influence of the training corpus (speaker plus content).

Öhman model (1965)

- Öhman (1965) has shown that the loci of consonants are not independent of the vocalic context.



- And that more that locus is required for /g/.

- *Conclusion :*
- *The consonant gestures are superimposed on those of the contextual vowels.*

Öhman model (1966)

- Idea: to a certain extent (only) articulators involved in the consonantal gesture can be decoupled from those used for vowels.
- Speech is a sequence of continuous vocalic gestures on which fast constriction gestures of consonants superimpose.

$$s(x,t) = v(x) + k(t)\left[c(x) - v(x)\right]w_c(x)$$

where *s(x,t)* is the vocal tract shape at point x and time t, *v(x)* vocal tract shape at point x and time t for a given vowel, *c(x)* the vocal tract shape for the consonant, *k(t)* is an interpolation value (de 0 à 1), and $w_c(x)$ describes the resistance degree of the consonant.

- *v(x)* describes the vocal tract shape which can be the comibation of two vowels V1 et V2 (v(x) thus varies from V1 to V2)

Articulatory phonology (Catherine Browman and Louis Goldstein-1986)

- Articulatory gestures are primitives in terms of phonological description and control of the vocal tract shape.
- One gesture is a coordination structure between articulators intended to produce a constriction.
- Spring mass system to simulate the movement of an articulator: $m\ddot{x} + b\dot{x} + k(x - u) = 0$

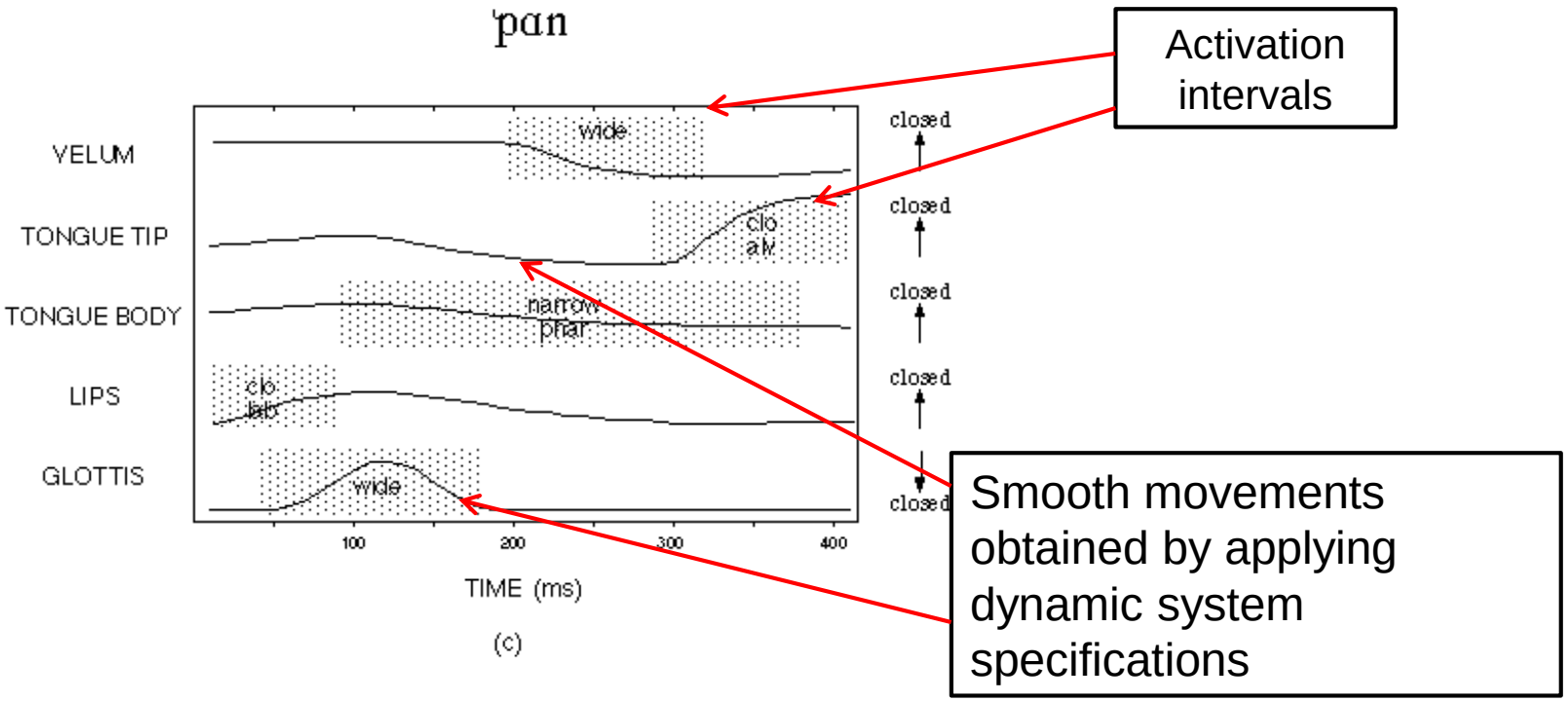
  x is the position of the mass (articulator), u the new position
- Articulators are organized in groups of articulators and gestures overlap in time.

Constrictions utilized by la articulatory phonology

- Coordinative structures according to the constriction:
  - Lips (protrusion and opening)
  - Tongue tip (place and degree of constriction)
  - Tongue body (place and degree of constriction)
  - Velopharyngeal port (opening)
  - Glottis (glottis opening)

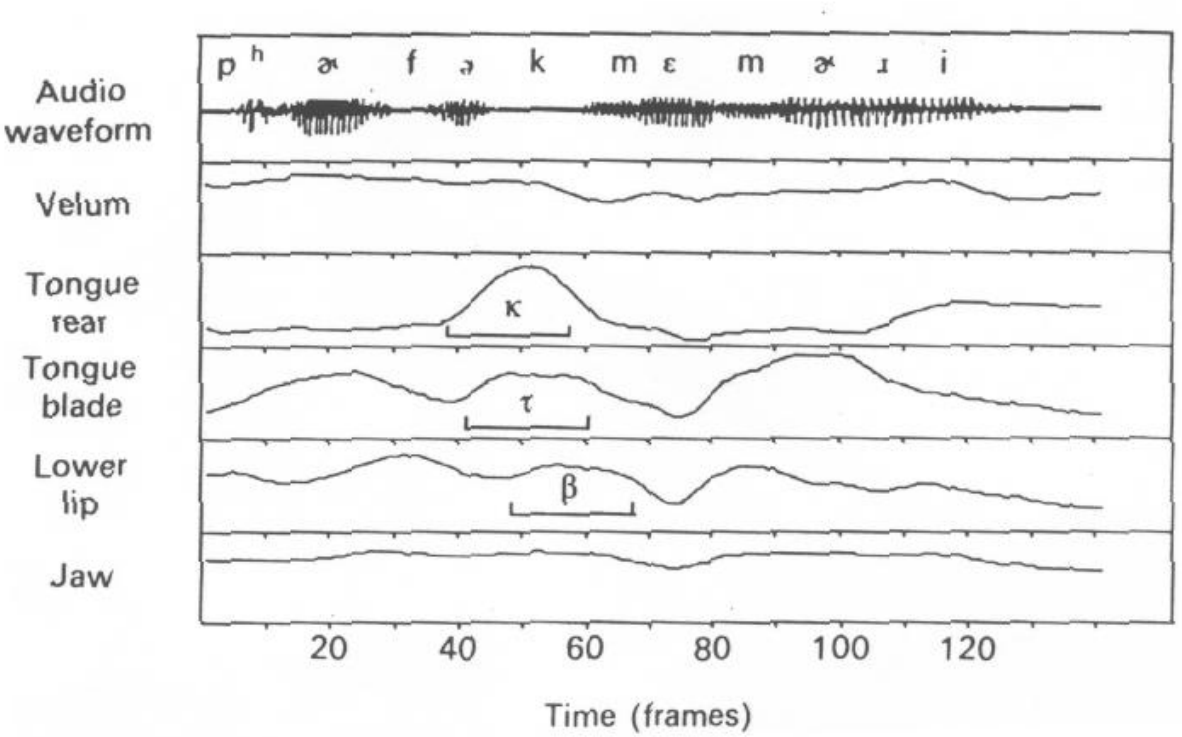
Articulatory phonology: gestural score

- A gestural score is required to coordinate gestures. Here /pan/ from a Browman et Goldstein paper



Contribution of the articulatory phonology

- Interesting from the point of view of articulatory synthesis
- Gesture can overlap, and some can hide other gestures
- A powerfull tool to explain the disparition of one sound in a sequence of words "perfect memory" in the following example



To conclude about coarticulation

- Two big principles:

1. **anticipation** of articulatory positions due to the utilization of articulators which cannot move instantaneously from one point to the next.
2. **geometrical constraints imposed by acoustics** so as to guarantee expected phonetic features.