Articulatory Speech

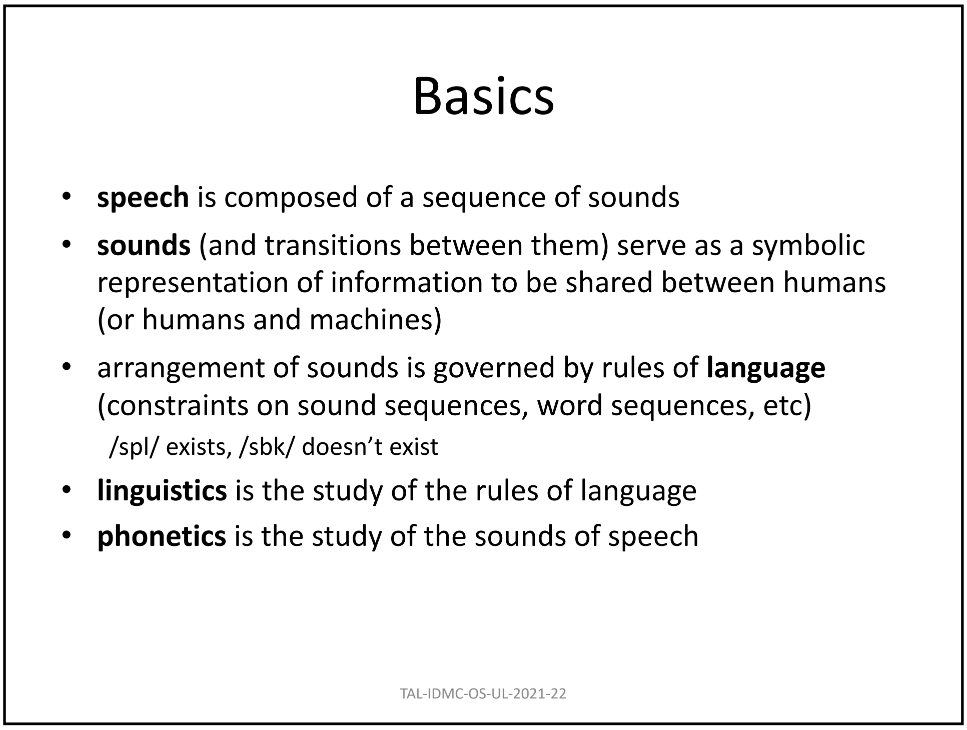
Speech Production

1

Basic speech processing

- Remarkably, humans can decode these sounds and determine the meaning that was intended at least at the idea/concept level (perhaps not completely at the word or sound level); often machines can also do the same task.
  - speech coding: waveform → (model) → waveform
  - speech synthesis: words → waveform
  - speech recognition: waveform → words/sentences
  - speech understanding: waveform → idea

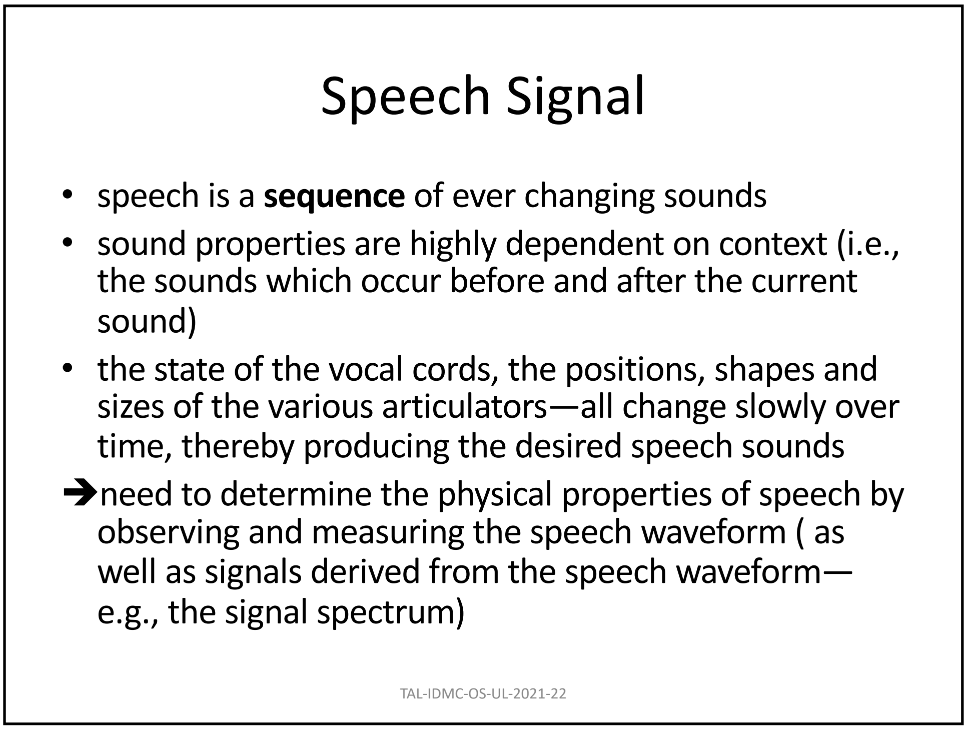
3



Basics

- **speech** is composed of a sequence of sounds
- **sounds** (and transitions between them) serve as a symbolic representation of information to be shared between humans (or humans and machines)
- arrangement of sounds is governed by rules of **language** (constraints on sound sequences, word sequences, etc)
  /spl/ exists, /sbk/ doesn’t exist
- **linguistics** is the study of the rules of language
- **phonetics** is the study of the sounds of speech

4

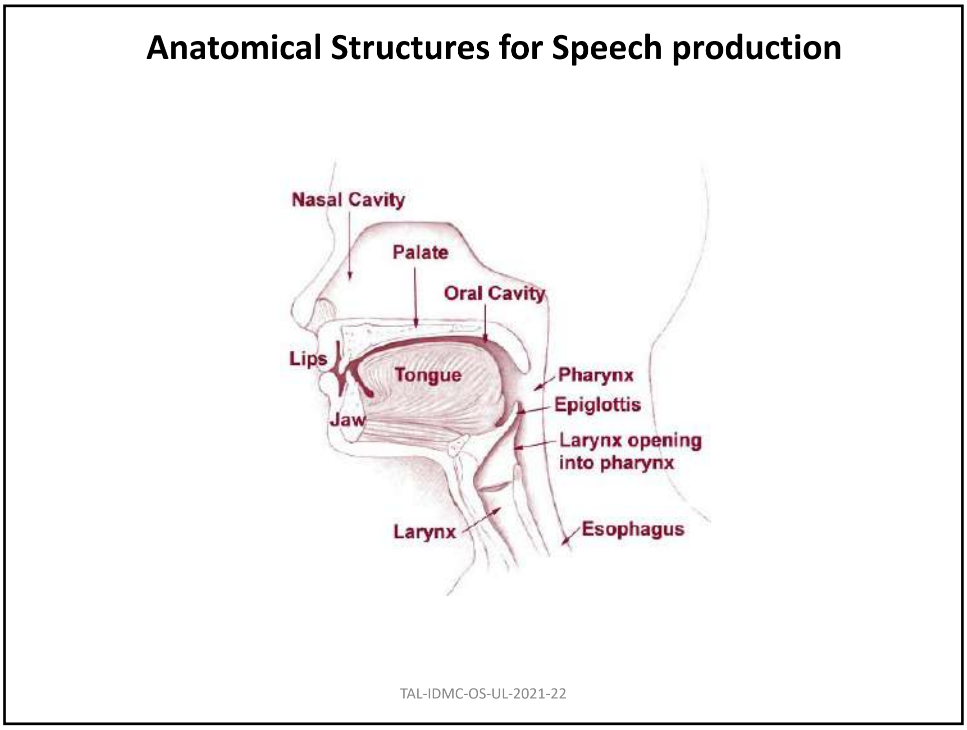


Speech Signal

- speech is a **sequence** of ever changing sounds
- sound properties are highly dependent on context (i.e., the sounds which occur before and after the current sound)
- the state of the vocal cords, the positions, shapes and sizes of the various articulators—all change slowly over time, thereby producing the desired speech sounds

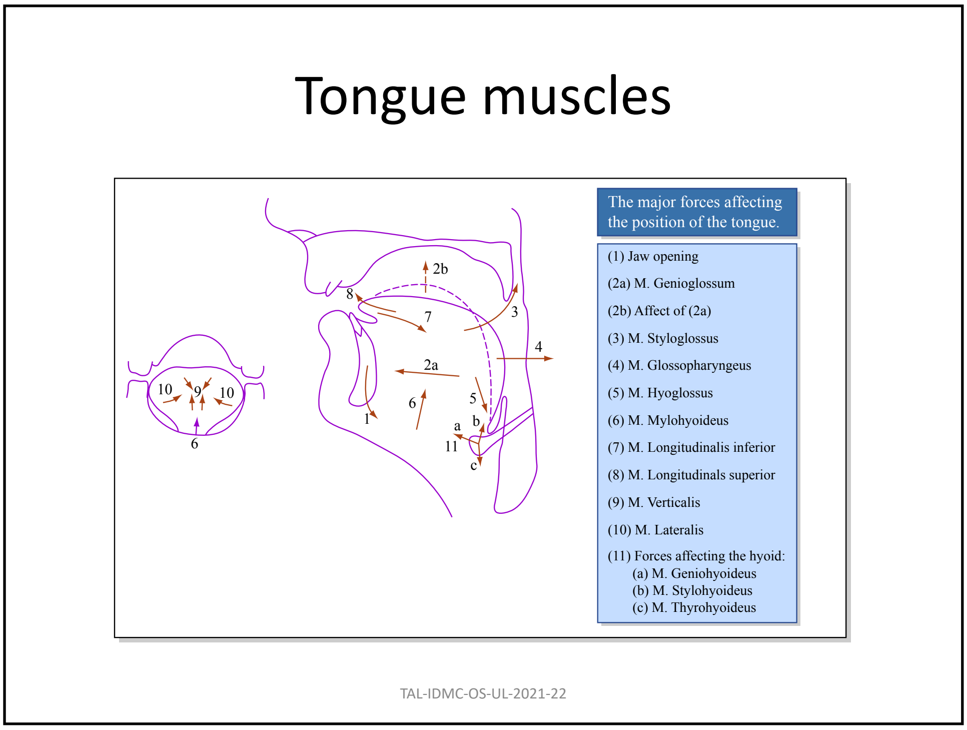
➔need to determine the physical properties of speech by observing and measuring the speech waveform (as well as signals derived from the speech waveform—e.g., the signal spectrum)

5



Anatomical Structures for Speech production

6



Tongue muscles

The major forces affecting the position of the tongue.

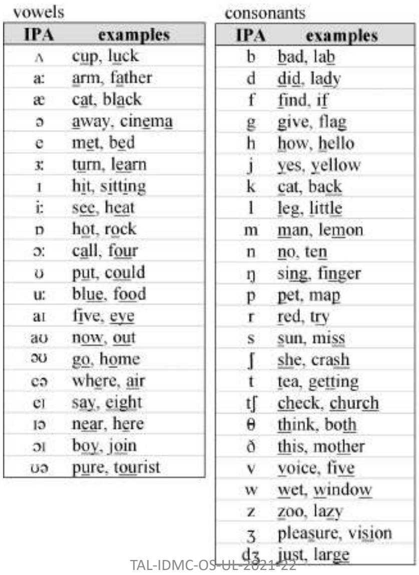
(1) Jaw opening
(2a) M. Genioglossum
(2b) Affect of (2a)
(3) M. Styloglossus
(4) M. Glossopharyngeus
(5) M. Hyoglossus
(6) M. Mylohyoideus
(7) M. Longitudinalis inferior
(8) M. Longitudinals superior
(9) M. Verticalis
(10) M. Lateralis
(11) Forces affecting the hyoid:
(a) M. Geniohyoideus
(b) M. Stylohyoideus
(c) M. Thyrohyoideus

7

Phonemes

- A phoneme is a basic unit of a language,
- When combined with other phonemes, they form meaningful units such as words.
- The phoneme is the smallest contrastive linguistic unit which may change the meaning of a word.
- Example: The difference in meaning between *kill* and *kiss* is a result of the exchange of the phoneme /l/ for the phoneme /s/.
- Two words that differ in meaning through a contrast of a single phoneme are called **minimal pairs**.

Phonemes in American English

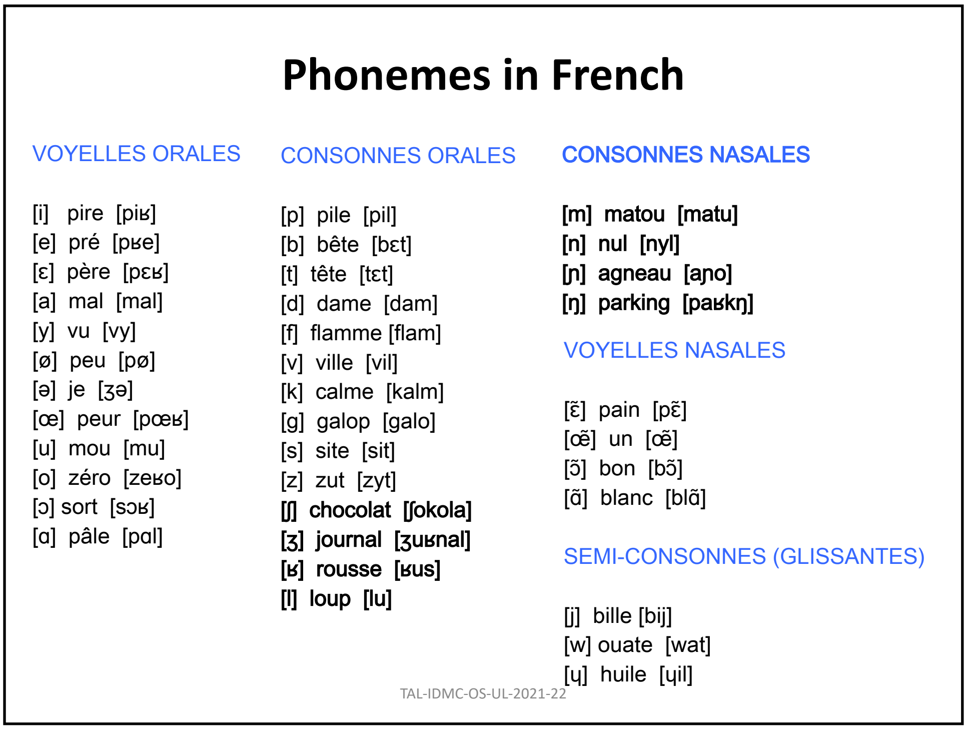


vowels

IPA	examples
ʌ	cup, luck
a:	arm, father
æ	cat, black
ə	away, cinema
e	met, bed
ɜ:	turn, learn
ɪ	hit, sitting
i:	see, heat
ɒ	hot, rock
ɔ:	call, four
ʊ	put, could
u:	blue, food
aɪ	five, eye
aʊ	now, out
əʊ	go, home
eə	where, air
eɪ	say, eight
ɪə	near, here
ɔɪ	boy, join
ʊə	pure, tourist

consonants

IPA	examples
b	bad, lab
d	did, lady
f	find, if
g	give, flag
h	how, hello
j	yes, yellow
k	cat, back
l	leg, little
m	man, lemon
n	no, ten
ŋ	sing, finger
p	pet, map
r	red, try
s	sun, miss
ʃ	she, crash
t	tea, getting
tʃ	check, church
θ	think, both
ð	this, mother
v	voice, five
w	wet, window
z	zoo, lazy
ʒ	pleasure, vision
dʒ	just, large



Phonemes in French

VOYELLES ORALES

[i]	pire	[piʁ]
[e]	pré	[pʁe]
[ɛ]	père	[pɛʁ]
[a]	mal	[mal]
[y]	vu	[vy]
[ø]	peu	[pø]
[ə]	je	[ʒə]
[œ]	peur	[pœʁ]
[u]	mou	[mu]
[o]	zéro	[zeʁo]
[ɔ]	sort	[sɔʁ]
[ɑ]	pâle	[pɑl]

CONSONNES ORALES

[p]	pile	[pil]
[b]	bête	[bɛt]
[t]	tête	[tɛt]
[d]	dame	[dam]
[f]	flamme	[flam]
[v]	ville	[vil]
[k]	calme	[kalm]
[g]	galop	[galo]
[s]	site	[sit]
[z]	zut	[zyt]
[ʃ]	**chocolat**	**[ʃokola]**
[ʒ]	**journal**	**[ʒuʁnal]**
[ʁ]	**rousse**	**[ʁus]**
[l]	**loup**	**[lu]**

CONSONNES NASALES

[m]	**matou**	**[matu]**
[n]	**nul**	**[nyl]**
[ɲ]	**agneau**	**[aɲo]**
[ŋ]	**parking**	**[paʁkŋ]**

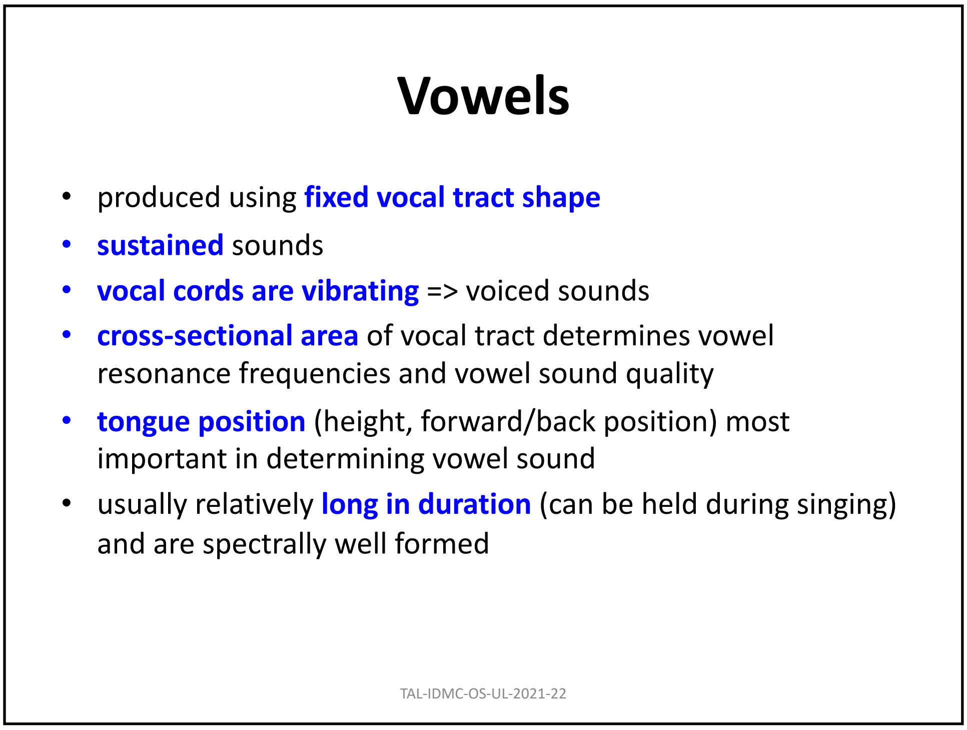
VOYELLES NASALES

[ɛ̃]	pain	[pɛ̃]
[œ̃]	un	[œ̃]
[ɔ̃]	bon	[bɔ̃]
[ɑ̃]	blanc	[blɑ̃]

SEMI-CONSONNES (GLISSANTES)

[j]	bille	[bij]
[w]	ouate	[wat]
[ɥ]	huile	[ɥil]

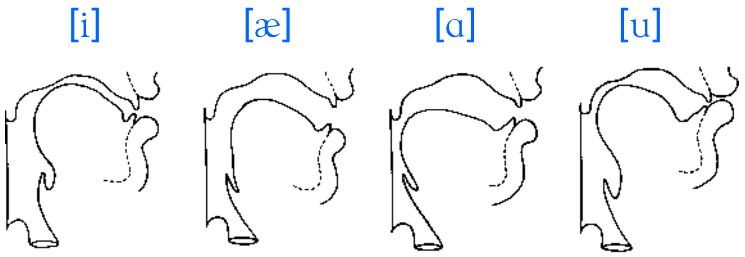
10



Vowels

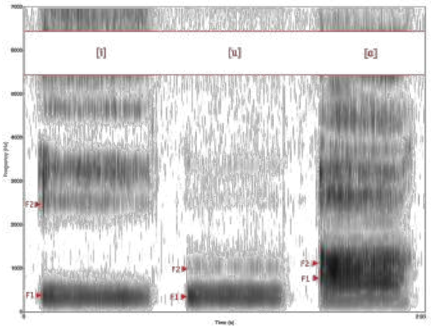
- produced using **fixed vocal tract shape**
- **sustained** sounds
- **vocal cords are vibrating** => voiced sounds
- **cross-sectional area** of vocal tract determines vowel resonance frequencies and vowel sound quality
- **tongue position** (height, forward/back position) most important in determining vowel sound
- usually relatively **long in duration** (can be held during singing) and are spectrally well formed

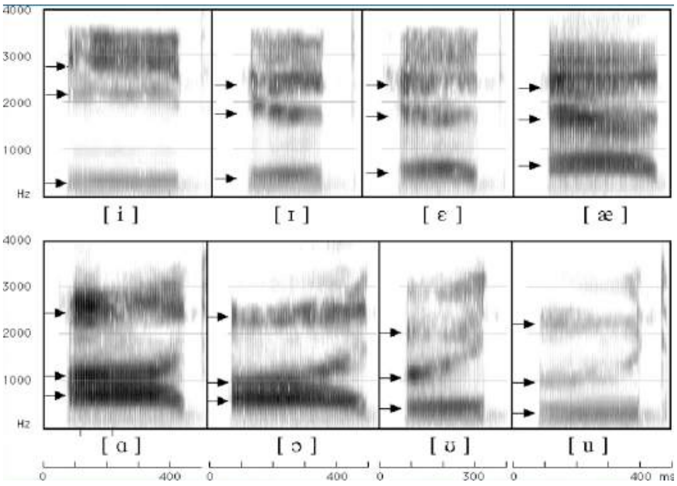
11



Formants

Formants are the distinguishing or meaningful frequency components of human speech and of singing. By definition, the information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds.



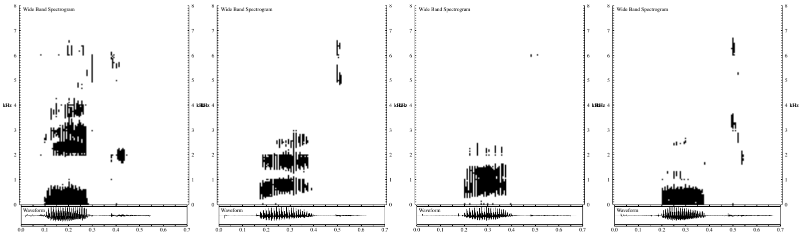


Formant 3

Formant 2

Formant 1

14



beet
/bi^{y}t/

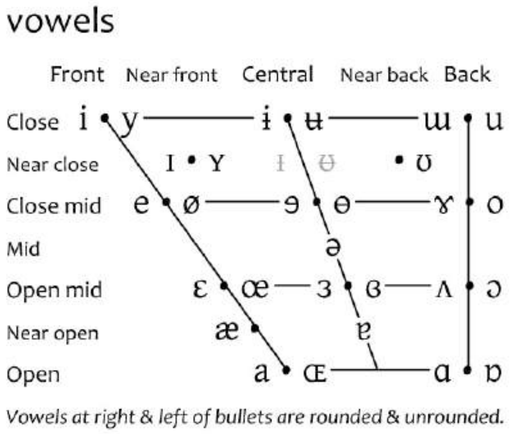
bat
/bæt/

bott
/bɑt/

boot
/but/

15

Vowels



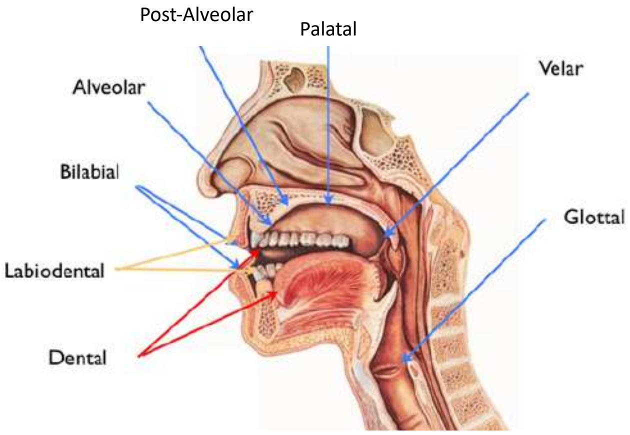
17

Consonants

- A consonant is a speech sound that is articulated with complete or partial closure of the vocal tract.
- Examples
  - [p], pronounced with the lips;
  - [t], pronounced with the front of the tongue;
  - [k], pronounced with the back of the tongue;
  - [h], pronounced in the throat;
  - [f] and [s], pronounced by forcing air through a narrow channel (fricatives);
  - and [m] and [n], which have air flowing through the nose (nasals).

18

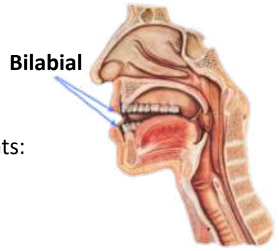
Place of articulation



Bilabial

•*Bilabial consonants* occur when you block/constrict airflow out of the mouth.

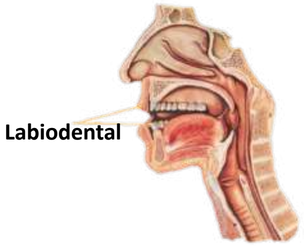
•English contains the following three bilabial consonants:
/p/ as in "**p**urse" and "ra**p**"
/b/ as in "**b**ack" and "ca**b**"
/m/ as in "**m**ad" and "cla**m**"

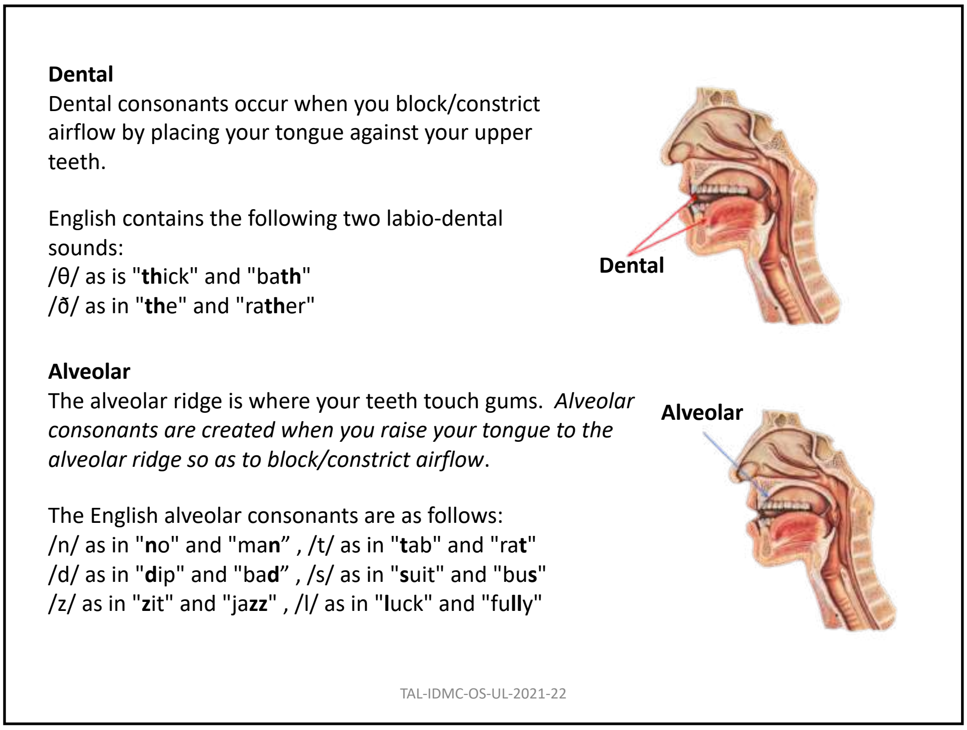


Labiodental

•*Labiodental consonants* occur when you block/constrict airflow by curling your lower lip back and raising it to touch your upper teeth.

•English contains the following two labio-dental sounds:
/f/ as in "**f**ro" and "cal**f**"
/v/ as in "**v**ine" and "ha**v**e"





Dental

Dental consonants occur when you block/constrict airflow by placing your tongue against your upper teeth.

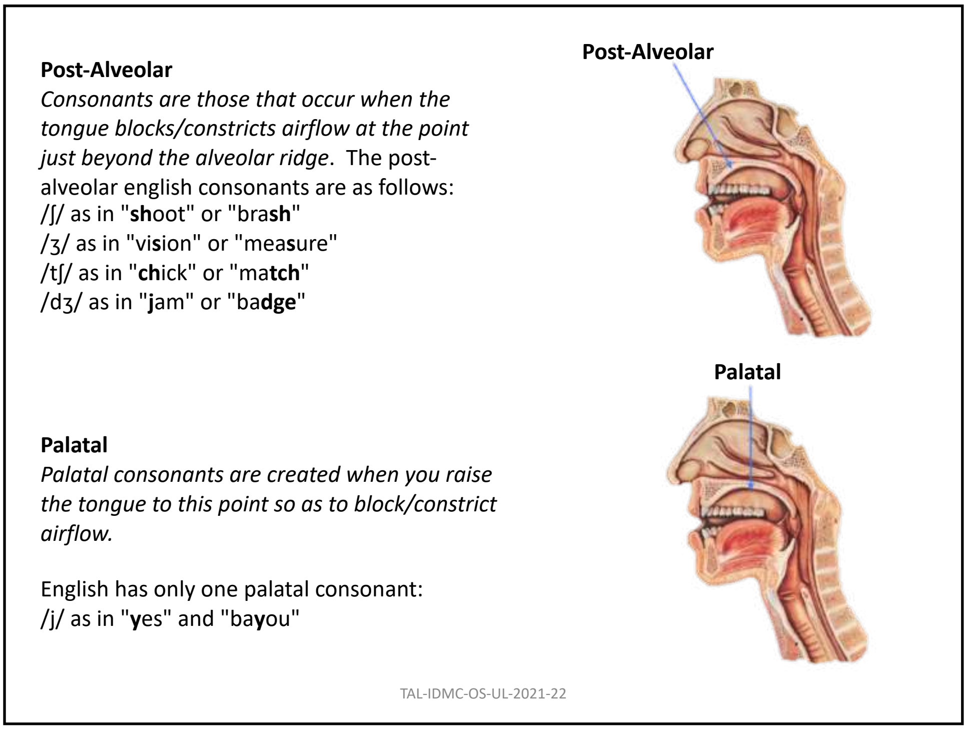
English contains the following two labio-dental sounds:
/θ/ as is "**th**ick" and "ba**th**"
/ð/ as in "**th**e" and "ra**th**er"

Alveolar

The alveolar ridge is where your teeth touch gums. *Alveolar consonants are created when you raise your tongue to the alveolar ridge so as to block/constrict airflow*.

The English alveolar consonants are as follows:
/n/ as in "**n**o" and "ma**n**" , /t/ as in "**t**ab" and "ra**t**"
/d/ as in "**d**ip" and "ba**d**" , /s/ as in "**s**uit" and "bu**s**"
/z/ as in "**z**it" and "ja**zz**" , /l/ as in "**l**uck" and "fu**ll**y"

21



Post-Alveolar

Consonants are those that occur when the tongue blocks/constricts airflow at the point just beyond the alveolar ridge. The post-alveolar english consonants are as follows:
/ʃ/ as in "**sh**oot" or "bra**sh**"
/ʒ/ as in "vi**s**ion" or "mea**s**ure"
/tʃ/ as in "**ch**ick" or "ma**tch**"
/dʒ/ as in "**j**am" or "ba**dge**"

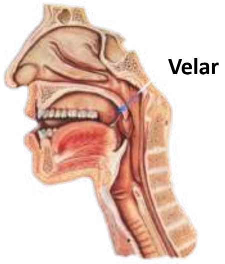
Palatal

Palatal consonants are created when you raise the tongue to this point so as to block/constrict airflow.

English has only one palatal consonant:
/j/ as in "**y**es" and "ba**y**ou"

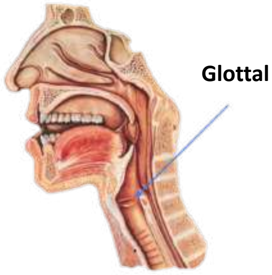
22

Velar
Velar Consonants are created when you raise the back of your tongue to the velum so as to block or restrict airflow.



English has the following velar consonants:
/ŋ/ as in "goi**ng**" and "u**n**cle”
/k/ as in "**k**ite" and "ba**ck**"
/g/ as in "**g**ood" and "bu**g**"
/w/ as in "**w**et" and "ho**w**ard"

Glottal
Glottal consonants aren't really consonants; they just play consonant roles in the language. In English the following things happen at the glottis:
/h/ as in "**h**i" and "Ba**h**amas".



23

Consonants

Place of articulation

- Bilabial (lips)—p,b,m,w
- Labiodental (between lips and front of teeth)-f,v
- Dental (teeth)-th,dh
- Alveolar (front of palate)-t,d,s,z,n,l
- Palatal (middle of palate)-sh,zh,r
- Velar (at velum)-k,g,ng

Manner of articulation

- Glide—smooth motion-w,l,r,y
- Nasal—lowered velum-m,n,ng
- Stop—constricted vocal tract-p,t,k,b,d,g
- Fricative—turbulent source-f,th,s,sh,v,dh,z,zh,h
- Voicing—voiced source-b,d,g,v,dh,z,zh,m,n,ng,w,l,r
- Mixed source—both voicing and unvoiced-j,ch
- Whispered--h

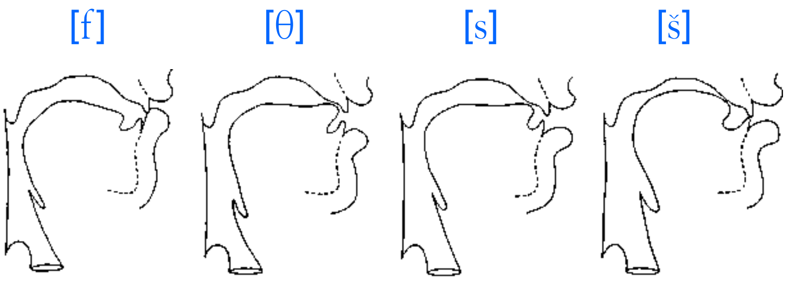
24

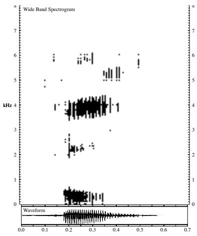
Consonants

- **Fricative**
  - Consonants made with a continuous airflow through the mouth, accompanied by a continuous audible noise.
  - Turbulence produced at narrow constriction
  - Can be held for a long period of time
  - “noisy”, white noise
  - Can be voiced or voiceless

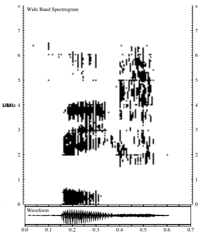
Fricative

Type	Unvoiced			Voiced		
Labial	/f/	f	fee	/v/	v	v
Dental	/θ/	th	thief	/ð/	dh	thee
Alveolar	/s/	s	see	/z/	z	z
Palatal	/š/	sh	she	/ž/	zh	Gigi

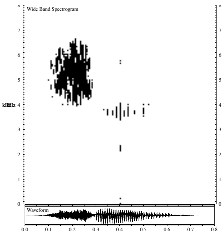




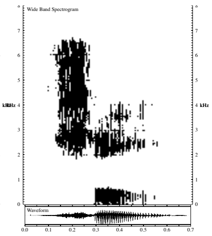
fee
/fi^{y}/



thief
/θi^{y}f/



see
/si^{y}/



she
/ši^{y}/

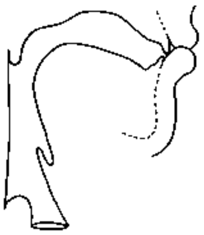
Consonants

- **Stops**
  - consonants made with a complete closure either in the oral cavity or in the glottis.
  - Sudden release of the constriction, turbulence noise.
  - English stops: bilabial, alveolar, velar, glottal
  - **Oral stops**: complete closure in the oral cavity and the velum is raised.
  - **Oral stops**: [p, t, k, b, d, g]
  - **Nasal stops**: complete closure in the oral cavity, but the velum is lowered. Air escapes through the nasal passage.
  - Nasal stops: [m, n, ŋ] - Always voiced

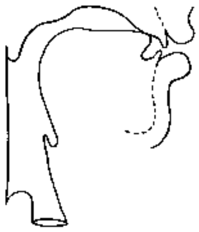
Stops

Type	Voiced			Unvoiced		
Labial	/b/	b	bought	/p/	p	pot
Alveolar	/d/	d	dot	/t/	t	tot
Velar	/g/	g	got	/k/	k	cot

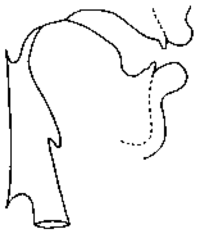
[b]

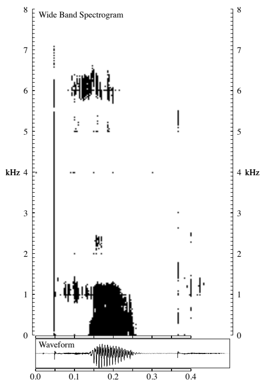


[d]

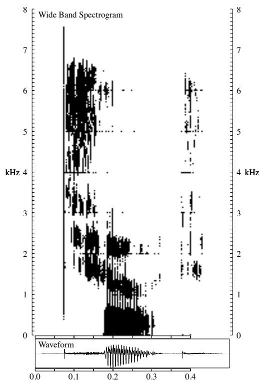


[g]

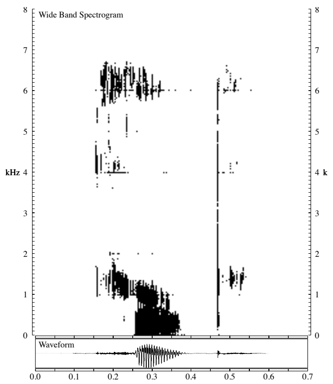




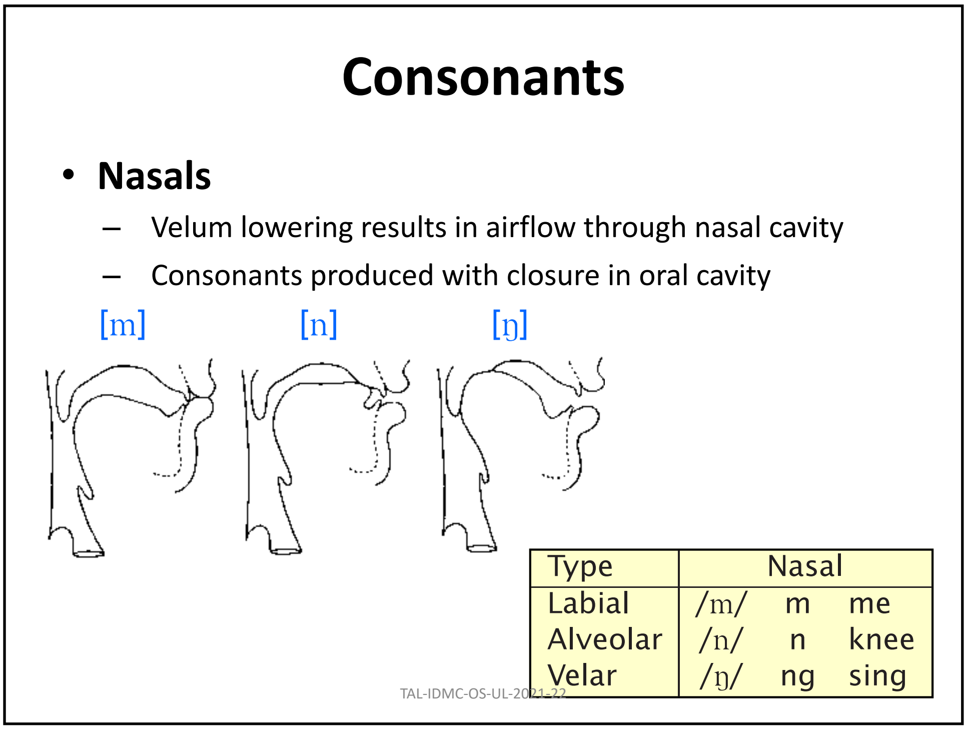
poop
/pup/



toot
/tut/



kook
/kuk/



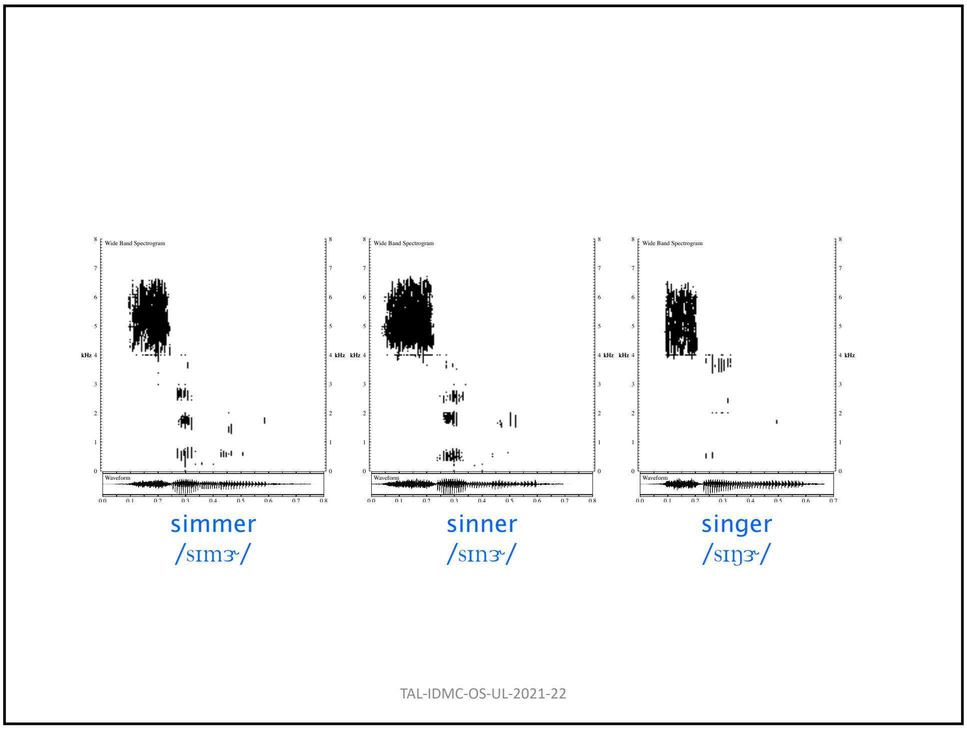
Consonants

- **Nasals**
  - Velum lowering results in airflow through nasal cavity
  - Consonants produced with closure in oral cavity

Type	Nasal		
Labial	/m/	m	me
Alveolar	/n/	n	knee
Velar	/ŋ/	ng	sing

TAL-IDMC-OS-UL-2021-22

31



simmer
/sɪmɝ/

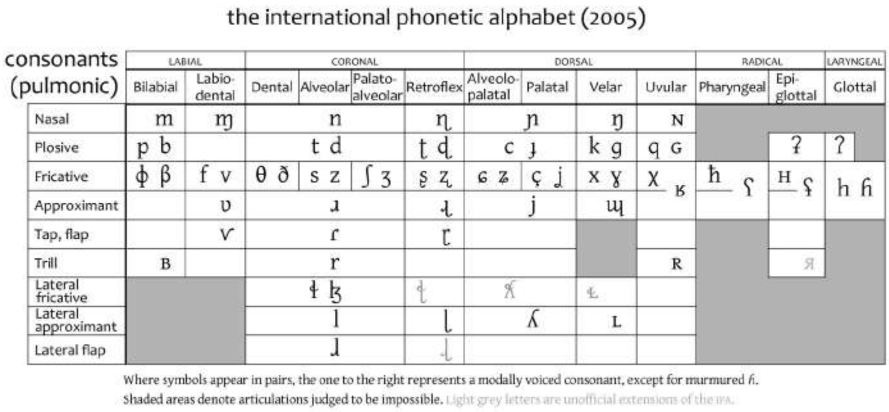
sinner
/sɪnɝ/

singer
/sɪŋɝ/

TAL-IDMC-OS-UL-2021-22

32

Place & manner of articulation

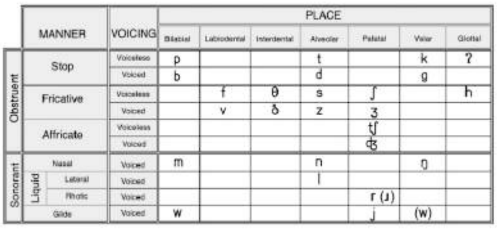


the international phonetic alphabet (2005)

consonants (pulmonic)	LABIAL		CORONAL				DORSAL				RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ		ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ		c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ		j	ɰ				
Tap, flap		ⱱ		ɾ		ɽ							
Trill	ʙ			r						ʀ		я	
Lateral fricative				ɬ ɮ		ꞎ		ʎ̝	ʟ̝				
Lateral approximant				l		ɭ		ʎ	ʟ				
Lateral flap				ɺ		ɺ̢							

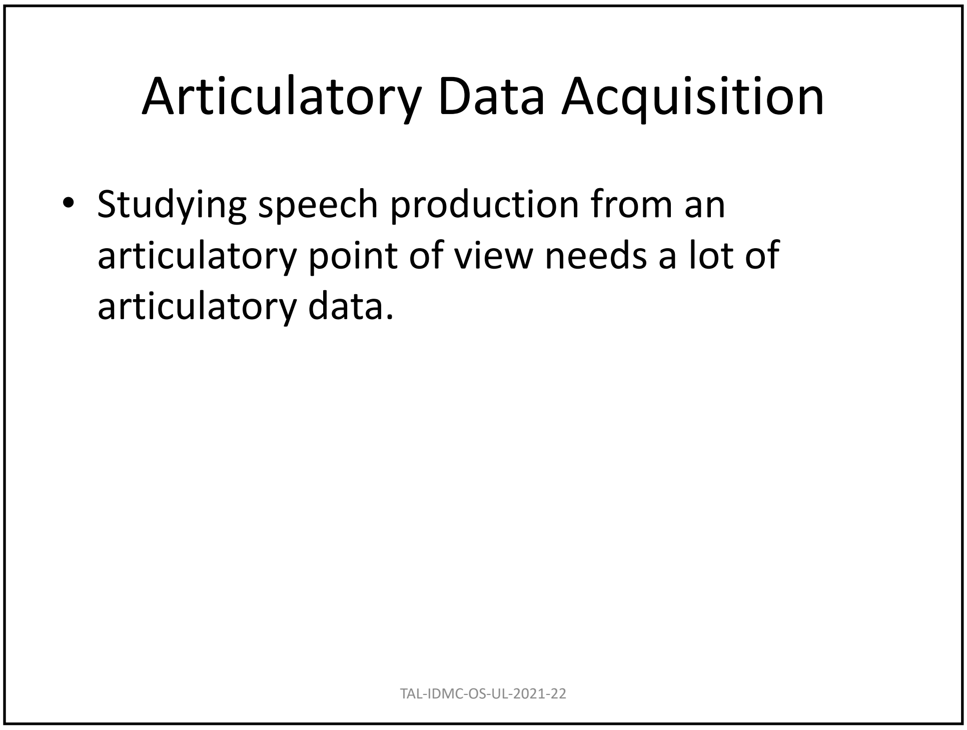
Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

http://www.ipachart.com



	MANNER		VOICING	PLACE						
				Bilabial	Labiodental	Interdental	Alveolar	Palatal	Velar	Glottal
Obstruent	Stop		Voiceless	p			t		k	ʔ
			Voiced	b			d		g	
	Fricative		Voiceless		f	θ	s	ʃ		h
			Voiced		v	ð	z	ʒ		
	Affricate		Voiceless					ʧ		
			Voiced					ʤ		
Sonorant	Nasal		Voiced	m			n		ŋ	
	Liquid	Lateral	Voiced				l			
		Rhotic	Voiced					r (ɹ)		
	Glide		Voiced	w				j	(w)	

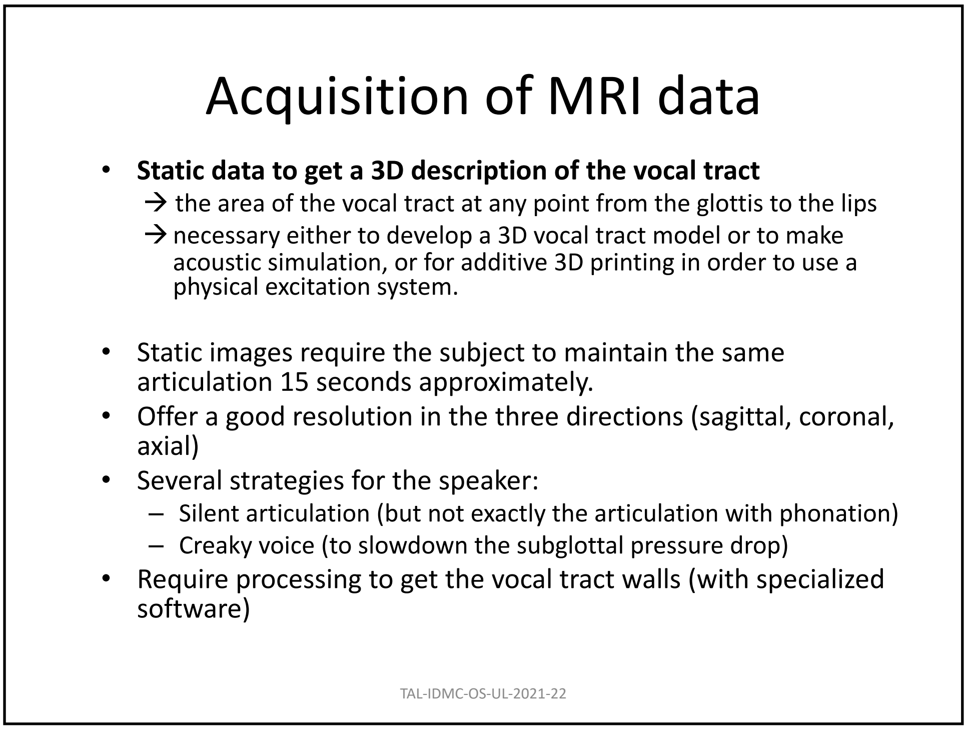
English



Articulatory Data Acquisition

- Studying speech production from an articulatory point of view needs a lot of articulatory data.

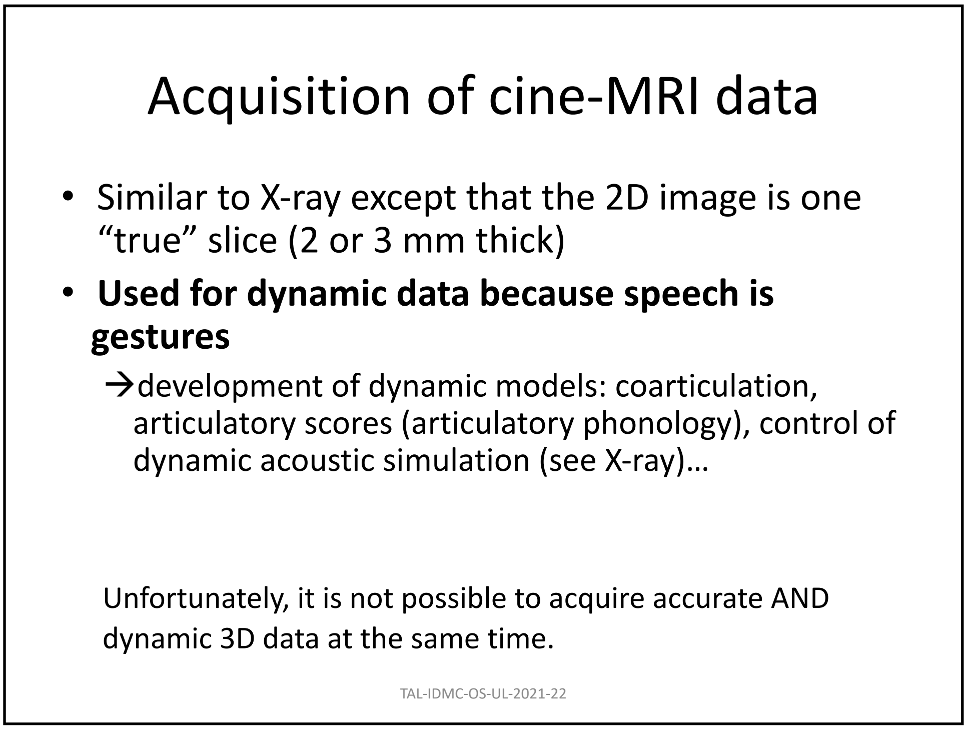
35



Acquisition of MRI data

- **Static data to get a 3D description of the vocal tract**
  → the area of the vocal tract at any point from the glottis to the lips
  → necessary either to develop a 3D vocal tract model or to make acoustic simulation, or for additive 3D printing in order to use a physical excitation system.
- Static images require the subject to maintain the same articulation 15 seconds approximately.
- Offer a good resolution in the three directions (sagittal, coronal, axial)
- Several strategies for the speaker:
  - Silent articulation (but not exactly the articulation with phonation)
  - Creaky voice (to slowdown the subglottal pressure drop)
- Require processing to get the vocal tract walls (with specialized software)

39



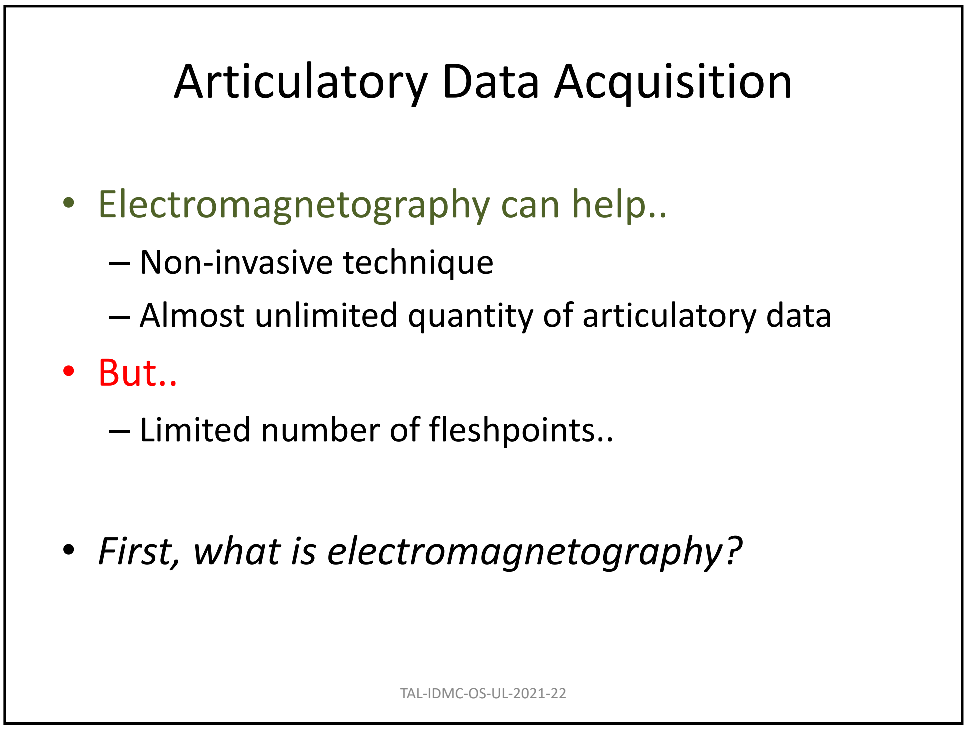
Acquisition of cine-MRI data

- Similar to X-ray except that the 2D image is one “true” slice (2 or 3 mm thick)
- **Used for dynamic data because speech is gestures**

  →development of dynamic models: coarticulation, articulatory scores (articulatory phonology), control of dynamic acoustic simulation (see X-ray)…

Unfortunately, it is not possible to acquire accurate AND dynamic 3D data at the same time.

41



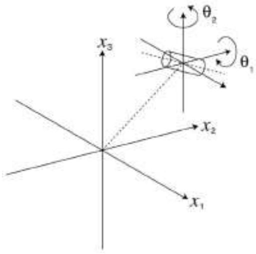
Articulatory Data Acquisition

- Electromagnetography can help..
  – Non-invasive technique
  – Almost unlimited quantity of articulatory data
- But..
  – Limited number of fleshpoints..
- *First, what is electromagnetography?*

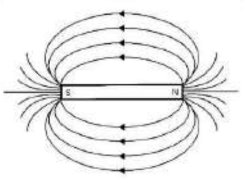
44

Articulatory data acquisition Electromagnetography (EMA)

- A coil immersed into a magnetic fields generates a current.
- Measuring the current provides electromagnetic field strength.
- The magnetic field strength in a receiver is inversely proportional to the cube of its distance from a transmitter.
- The recovery of the location is realized by solving equations from the currents measured in the coils (sensors).
- 3D EMA : 5 degree of freedem sensors (3 coordinates and two angles)



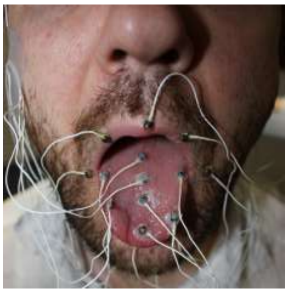
coil (sensor)

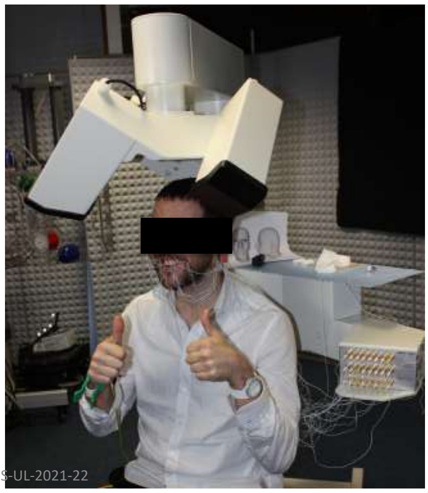


transmitter

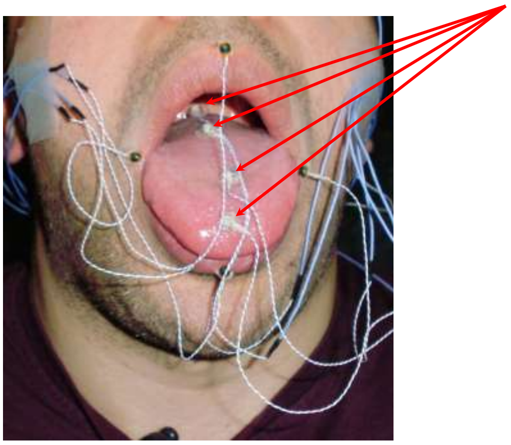
Articulograph Carstens - AG501

- Growing community among speech researchers
- 24 sensors , 250hz up to 1200Hz
- Good accuracy
  - Average error 0.3 mm
- Recording clean audio




[redacted]

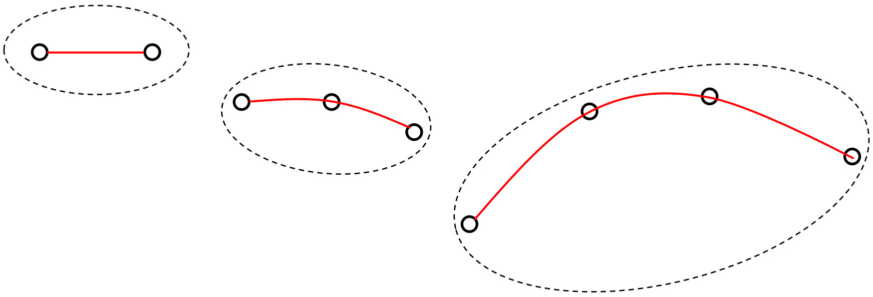
Sensor layout



4 sensors

47

Tongue contour from 4 sensors



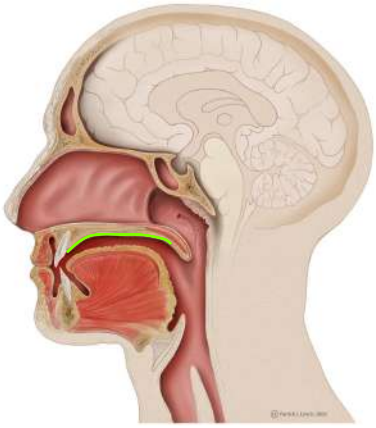
It is possible to predict tongue contour between sensors from their positions (Kaburagi and Honda, 1994)

T. Kaburagi and M. Honda(1994). **Determination of sagittal tongue shape from the positions of points on the tongue surface.** J. Acoust. Soc. Am. V.96, Issue 3, pp. 1356-1366

48

Retrieving the Palate

- Drawing Palate manually using one sensor.
- Some careful attention is needed while interpreting tracings.

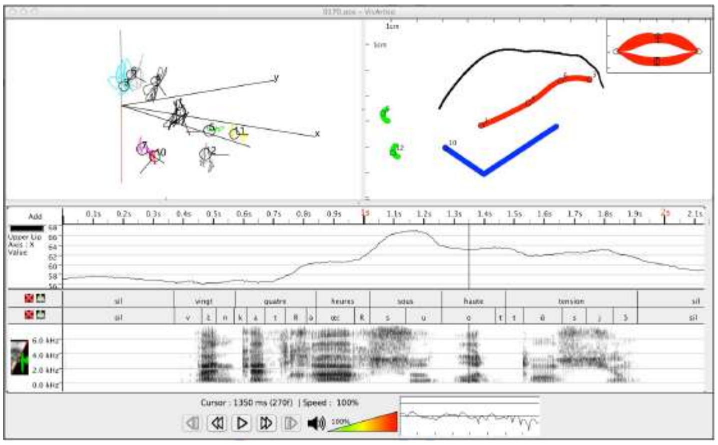


1.b Temporal evolution of the vocal tract shape

- How the vocal tract shape can be measured?
  - Which data ? 2D, 3D, with or without speech signal
  - Which technique? X-ray, MRI, electro-magnetographic articulography, electropalatography
  - Which precision? To be related to the dimension of the constriction which is the order of 1 millimeter and to the duration of a sounds or a fast articulatory event (burst noise for instance).

	X-ray	**Articulography**	**MRI**
+	Reasonable sampling rate (50 fps) Existence of many old databases The whole vocal tract is covered	High sampling rate Good precision in theory Not dangerous	Good precision 3D possible No health hazard
-	Health hazard Average noise Integration along an X-ray (projection)	Limited number of points (max. 24 sensors)	Noise preventing any recording (denoising required) Low sampling rate

Visualizing EMA Data: VisArtico

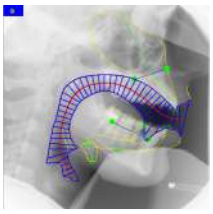


54

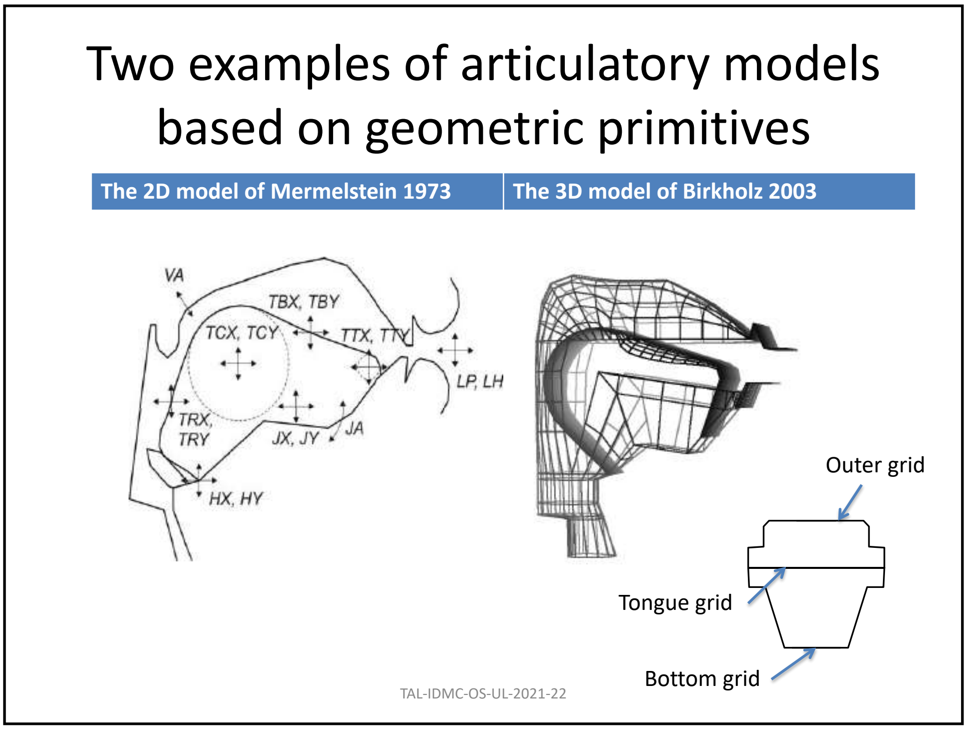
Articulatory models

- Why?

→ The goal is to represent the tongue and vocal tract in a concise form while retaining as much as possible the variability of the tongue shape and vocal tract.



56



Two examples of articulatory models based on geometric primitives

The 2D model of Mermelstein 1973	The 3D model of Birkholz 2003

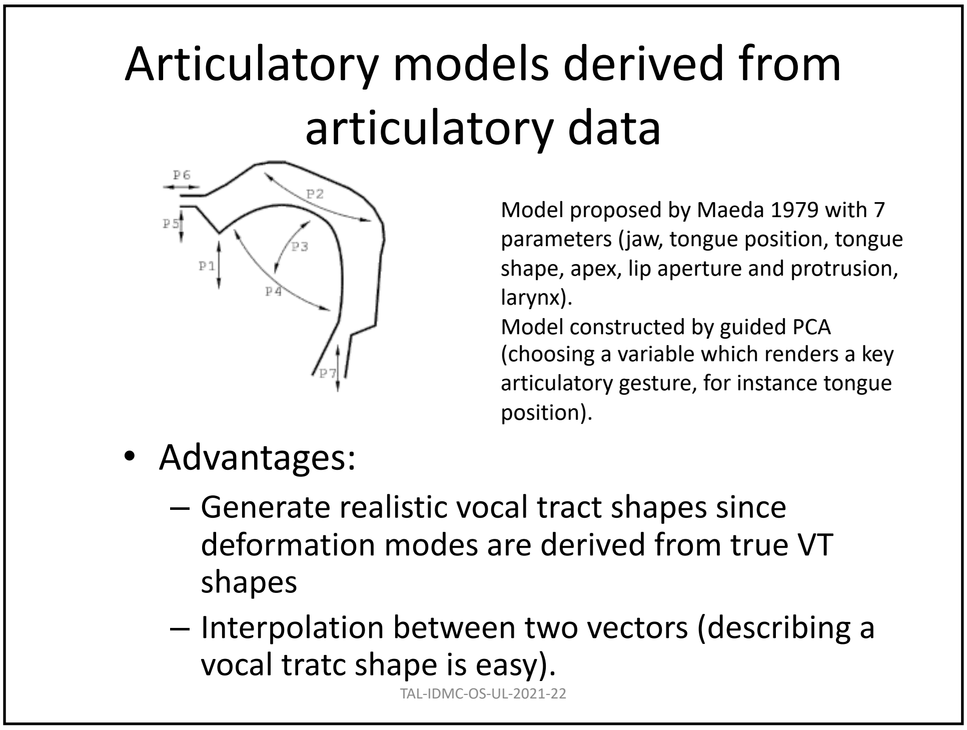
Outer grid

Tongue grid

Bottom grid

TAL-IDMC-OS-UL-2021-22

57



Articulatory models derived from articulatory data

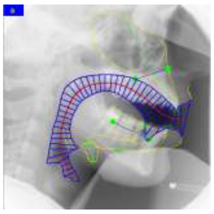
Model proposed by Maeda 1979 with 7 parameters (jaw, tongue position, tongue shape, apex, lip aperture and protrusion, larynx).
Model constructed by guided PCA (choosing a variable which renders a key articulatory gesture, for instance tongue position).

- Advantages:
  - Generate realistic vocal tract shapes since deformation modes are derived from true VT shapes
  - Interpolation between two vectors (describing a vocal tratc shape is easy).

TAL-IDMC-OS-UL-2021-22

58

Articulatory synthesis



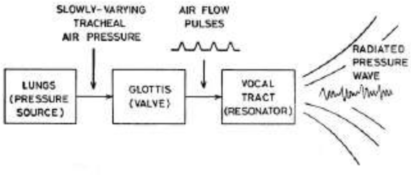


Figure 1. Non-interactive source-tract model for voice production.

- Tube-based analogy

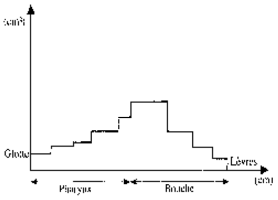


FIG. 1.12 - Fonction d'aire caractérisant le conduit vocal.

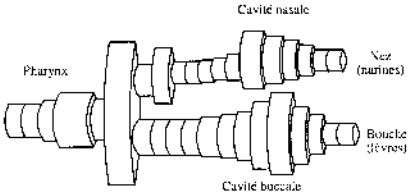


FIG. 1.11 Simulation du conduit vocal par des tubes.

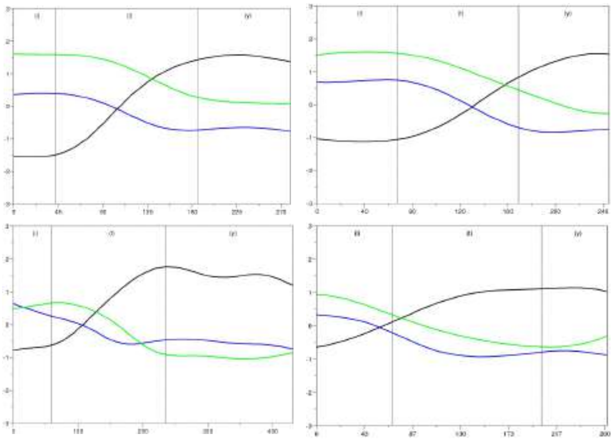
59

Coarticulation

- Speech production is influenced by the surrounding sounds (proceeding and following):
  - It is due to the characteristics (inertia and dynamics) of the articulators and the planning of the tasks of production which aims to minimize the articulatory effort of the speaker. Coarticulation makes automatic recognition as synthesis difficult.
- Coarticulation can be:
  - Anticipatory
  - Carry-over

64

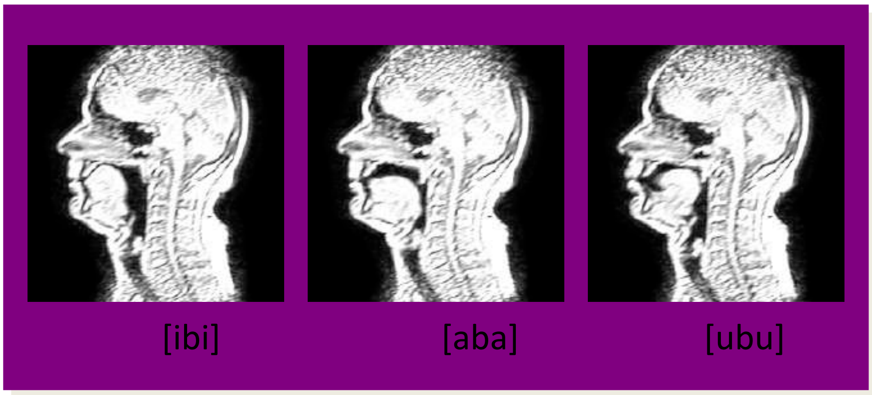
An example: labial anticipation inter-speaker variability for /ity/



protrusion —— opening —— stretching

65

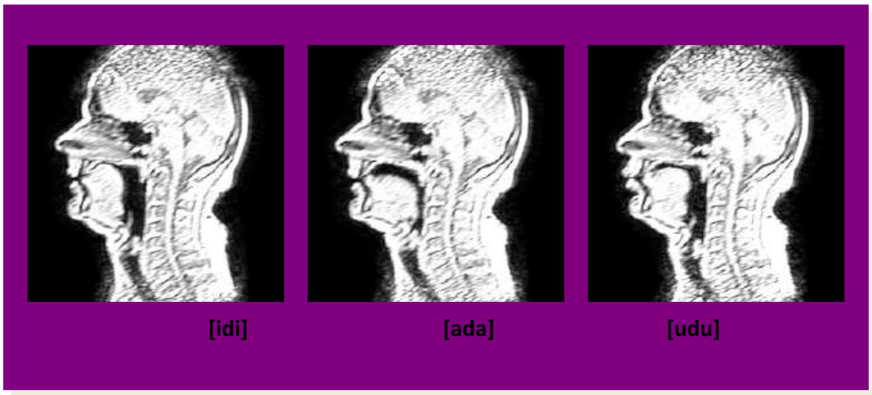
Dynamic MRI-data: [b] (from Kröger)



Strong coarticulatory positioning of the tongue

66

Dynamic MRI-data: [d] (from Kröger)

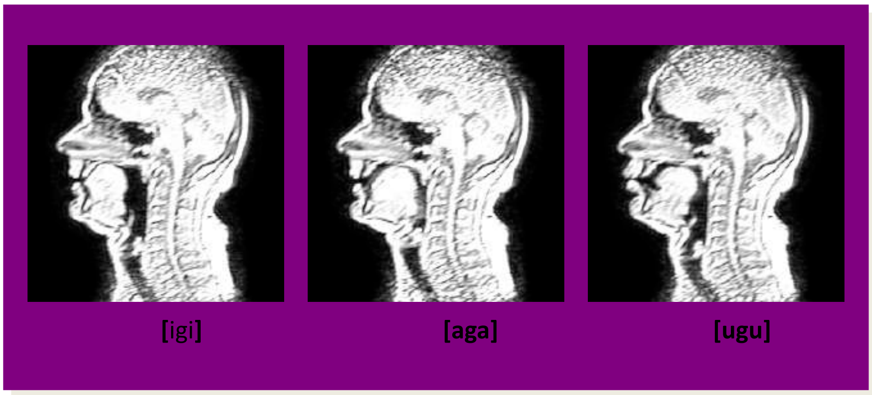


[idi] [ada] [udu]

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

67

Dynamic MRI-data: [g] (from Kröger)



[igi] [aga] [ugu]

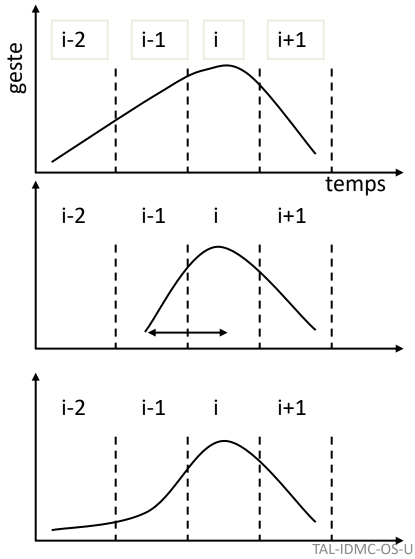
68

Remarks

- There is always marked anticipation.
- The maximum of the articulatory parameter is reached at the beginning of the sound, even a little before.
- Implies very different articulators:
  - Tongue tip (fast and light)
  - The soft palate (light)
  - Lips (fast and light)
  - Body of tongue or jaw (more massive)
- It is assumed that the articulators are independent of each other (eg the lips and the tip of the tongue) which is true as a first approach only.
- There is a great variability inter speaker ... which has often been masked by the small number of subjects studied ;-)
- But speakers remain consistent for their personal coarticulation strategy.
- ... and the data is not easy to acquire.

70

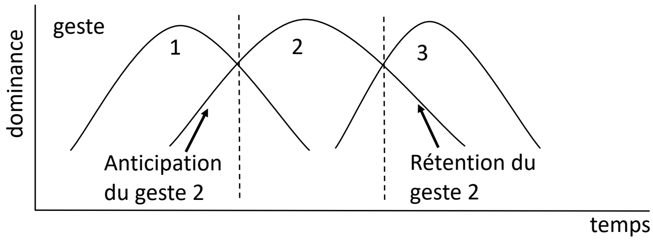
3 modèles testés sur la coarticulation labiale



1. Look-ahead
2. Time-locked
3. Hybrid

71

Time-locked Model



- Physical inspiration model
- The duration of an articulatory gesture is relatively constant (Due to dynamic constraints on speech articulators).
- The anticipation is therefore relatively fixed compared to the beginning of the acoustic sound to achieve.
- Overlapping articulatory gestures of the sound that precedes and the one to come (co-production).

73

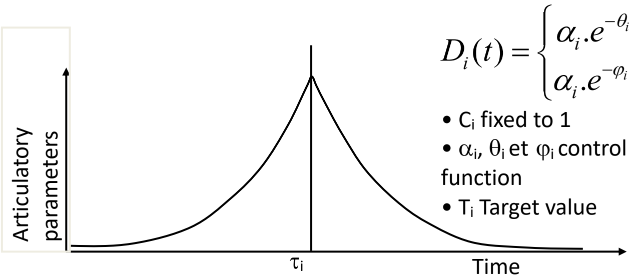
Coarticulation Models

- Cohen et Massaro:
  - uses dominance functions
  - Many used to pilot the labial coarticulation
- Öhman → gives a specific role to consonants and vowels

76

Cohen & Massaro Model

- Inspired by the gestural production model proposed by Löfqvist (recovery of articulatory gestures)
- Use of dominance functions: one per speech segment and per 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 fixed to 1
- α_i, θ_i et φ_i control the decreasing of the function
- T_i Target value

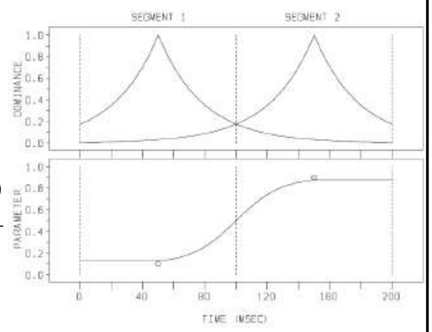
Cohen & Massaro Model

- Dominance functions to articulatory parameters :

Dominance functions

Articulatory parameters

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

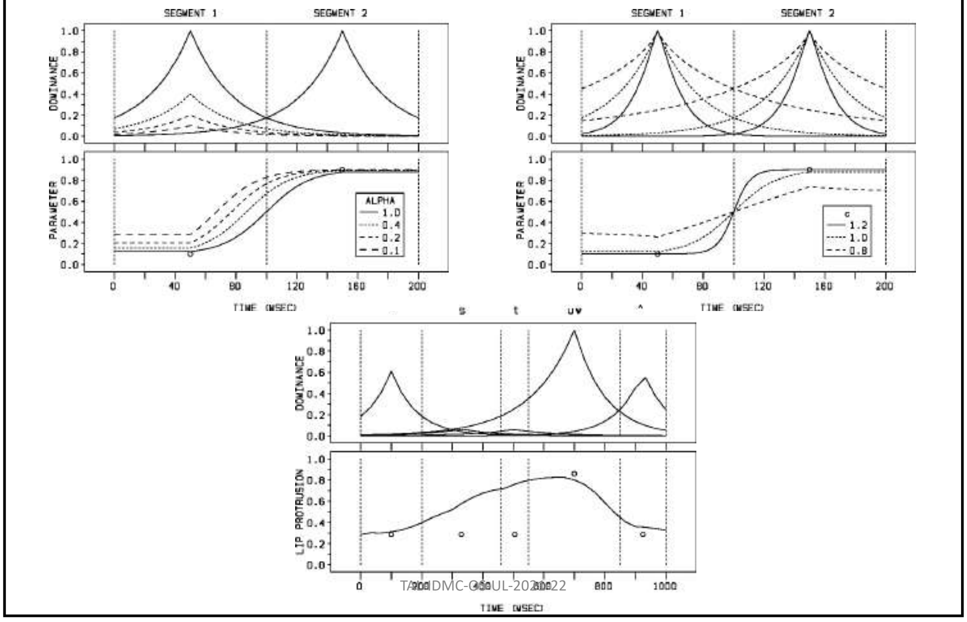


D'après Cohen et Massaro

Cohen & Massaro Model

Implementation

- Requires learning 5 parameters for each segment (viseme or phoneme).
- Often used for labial coarticulation (Beskow 2004 for example)
- It requires an important corpus of learning.
- Learning by optimizing parameters (either directly or using the gradient that can be calculated easily).

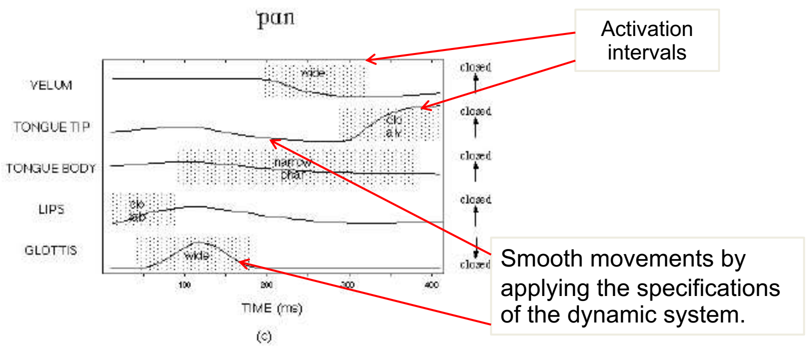


Articulatory Phonology (Catherine Browman and Louis Goldstein-1986)

- Articulatory gestures are the primitives in terms of phonological description and control of the shape of the vocal tract.
- A gesture is a coordination structure between articulators intended to produce a constriction
- The articulators are organized in a group of articulators, and the gestures overlap in time.

Articulatory phonology : gestural partition

- A gestural "partition" is necessary to coordinate the gestures.

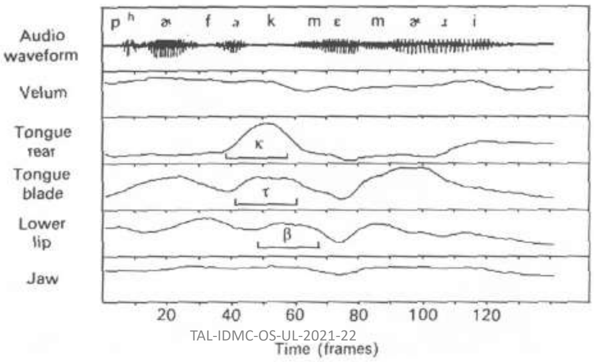


Activation intervals

Smooth movements by applying the specifications of the dynamic system.

Contributions of articulatory phonology

- Interesting for articulatory synthesis
- Gestures may overlap, and some may mask others
- Powerful to explain the disappearance of a sound in a sequence of words: "perfect memory" in the example below



86

Conclusion

1. anticipation of articulatory positions due to the use of articulators that do not move instantly from one point to another
2. geometric constraints imposed by acoustics to preserve the expected phonetic features

87