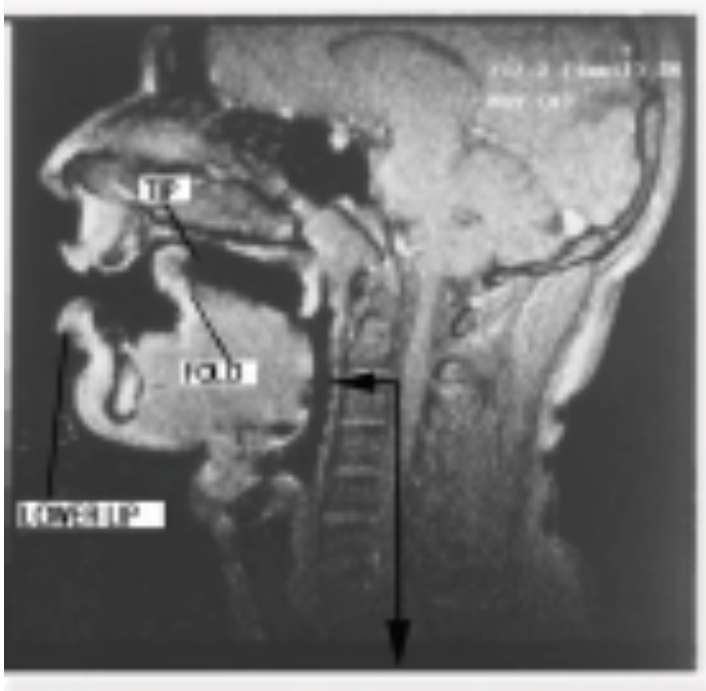
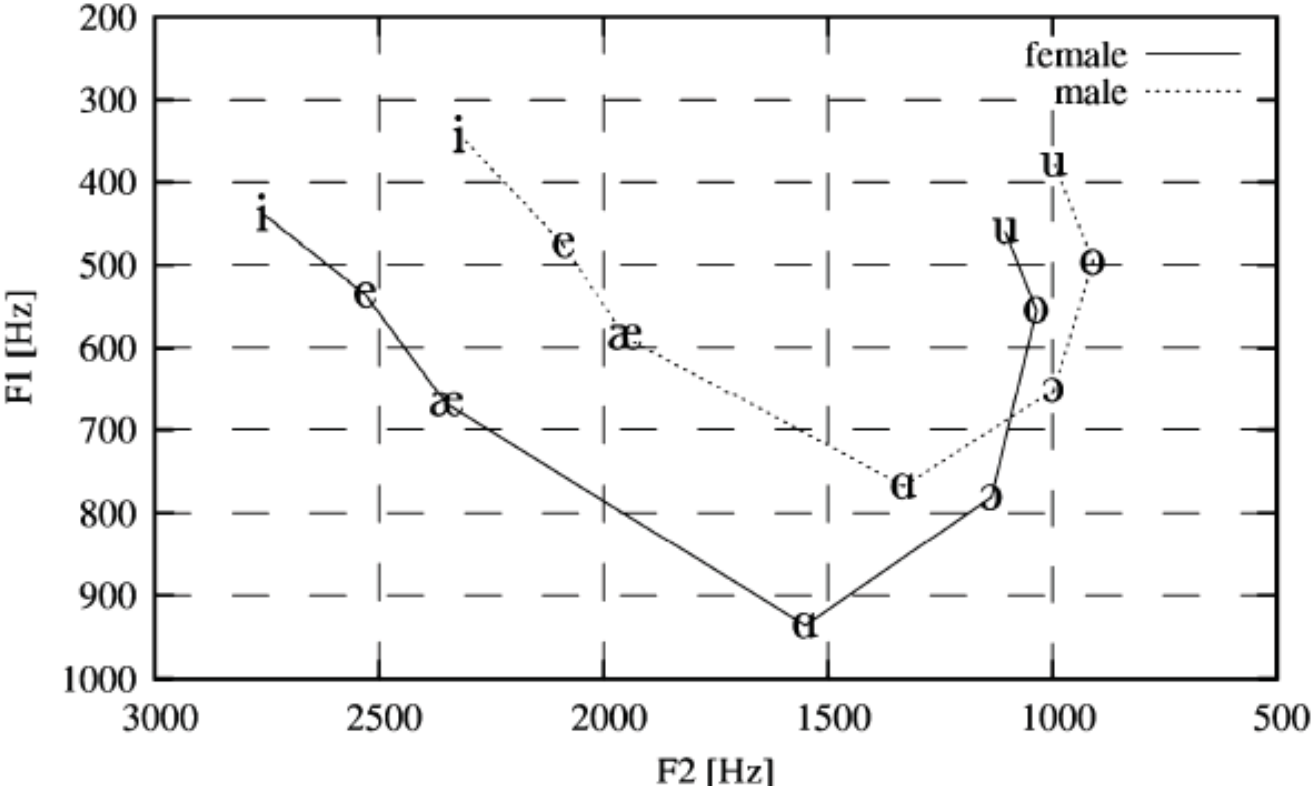
24.963

Linguistic Phonetics

Speech Production



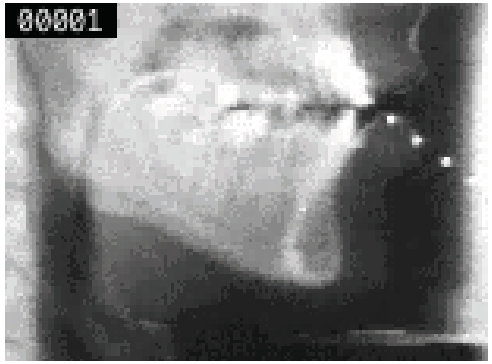
Reproduced from Srikanth, Byrd and Kaun. (1999) "Geometry, kinematics, and acoustics of tamil liquid consonants." The Journal of the Acoustical Society of America, with the permission of the Acoustical Society of America.



Hillenbrand et al (1995)

Speech Production

- Speaking is a very complex motor task, involving the coordination of many articulators.



Speech Movements

- Consider the movements of each of these structures
- Approximate number of muscle pairs that move the
  - Tongue: 9
  - Velum: 3
  - Lips: 12
  - Mandible: 7
  - Hyoid bone: 10
  - Larynx: 8
  - Pharynx: 4
- NB: The respiratory system

- Observations:
  - A large number of degrees of freedom
  - A very complicated control problem

Slide from Joe Perkell

Speech Production - basic questions

- So one of the central questions is ‘What are the control parameters in speech production?’
  - muscle tensions?
  - lengths and shortening velocities of muscles?
  - vocal tract shape?
  - acoustic/perceptual properties?
  - all of the above?
- Timing/coordination: Speaking involves coordinating movements in time.
  - How are the control parameters varied over time?
  - How are changes in control parameters coordinated?

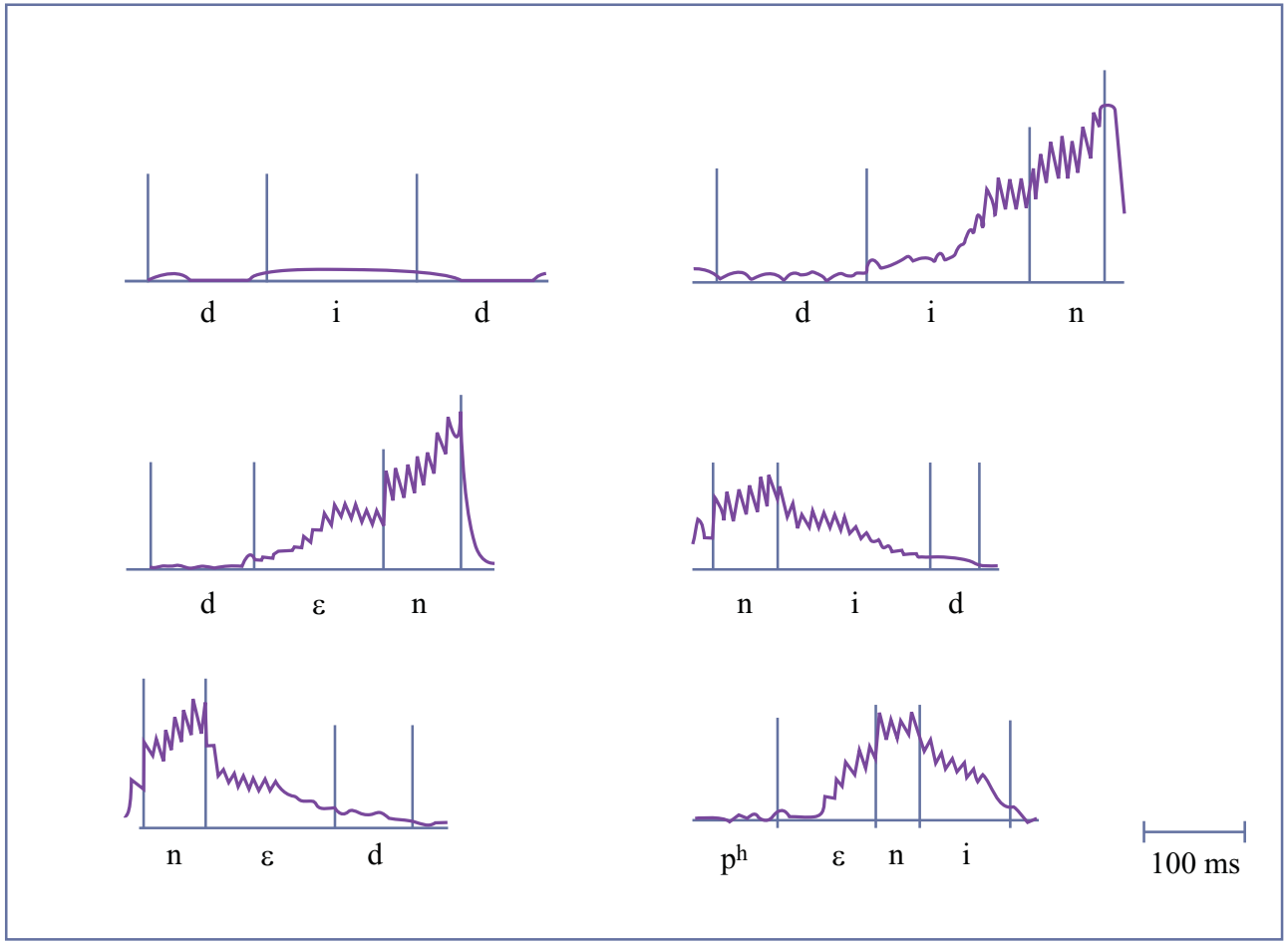
A simple model of speech production: the ‘beads on a string’ model

- Idea: Speech production involves concatenating a temporal sequence of targets corresponding to phonological segments.
- Targets are vocal tract shapes.
- Speech production involves concatenating a sequence of vocal tract shapes in time, and coordinating the muscles to move between these shapes.

- We see that this model is too simple when we consider data on coarticulatory variation in the realization of segments.

Coarticulation

- The influence of segmental context on the articulatory/ acoustic realization of a target segment.



Nasal airflow in English (Cohn 1990)

Image by MIT OCW.
Adapted from Cohn, A. "Nasalization in English: Phonology or phonetics?" *Phonology* 10 (1993): 43-81.

Coarticulation

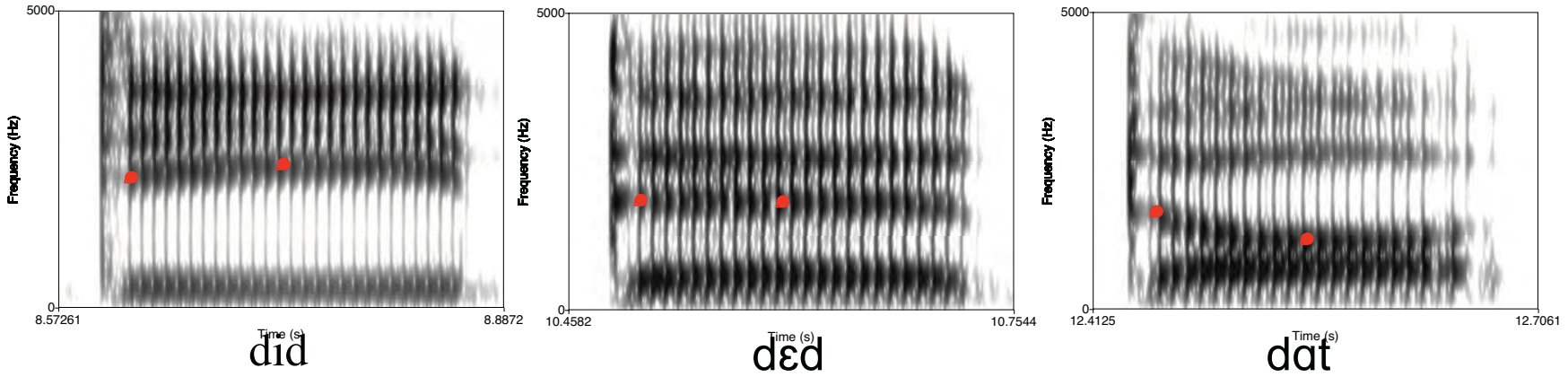
- Data on coarticulatory variation have been important in the development of models of speech production.
- We need to account for the types of influence that one segment has on another, and for the temporal extent of the influence of a segment on its neighbours.
- The simplest 'beads on a string' model leads us to expect that coarticulatory variation results solely from the transitions between segments (cf. Delattre et al's (1955) theory of acoustic loci for consonants, Liberman 1957).
- In fact coarticulation is considerably more complex than this.
  - Long range coarticulation effects.
  - Variation in targets as well as transitions.

Target variation or target undershoot

- Simple 'beads on a string' model implies that segment targets are invariant - variation is restricted to transitions.
- In a CV sequence,
  - F2 at the consonant (and therefore vocal tract shape) varies according to the following vowel (locus equation),
  - F2 in the vowel varies according to the adjacent consonants (vowel undershoot).

Target variation or target undershoot

- CV coarticulation - F2 frequency at the release of a stop varies depending on the following vowel.
  - Reflects assimilation towards the tongue body and lip position of the following vowel.



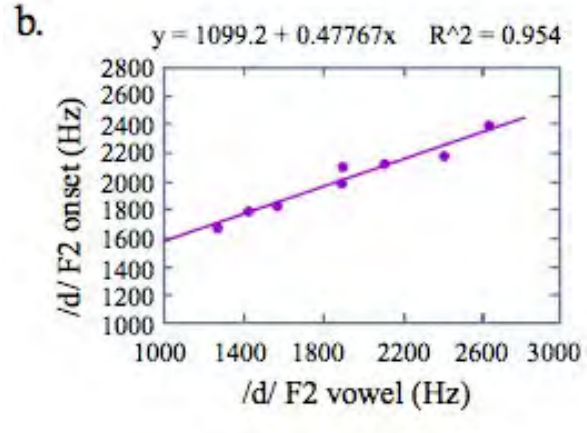


Image by MIT OCW.
Adapted from Fowler, C.A. "Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation." Perception and Psychophysics 55, no. 6 (1994): 597-610.

Target variation or target undershoot

- There are vowel-dependent differences in tongue body and lip position even in the middle of stops.
- Tracings of frames from X-ray movies (Öhman 1966):

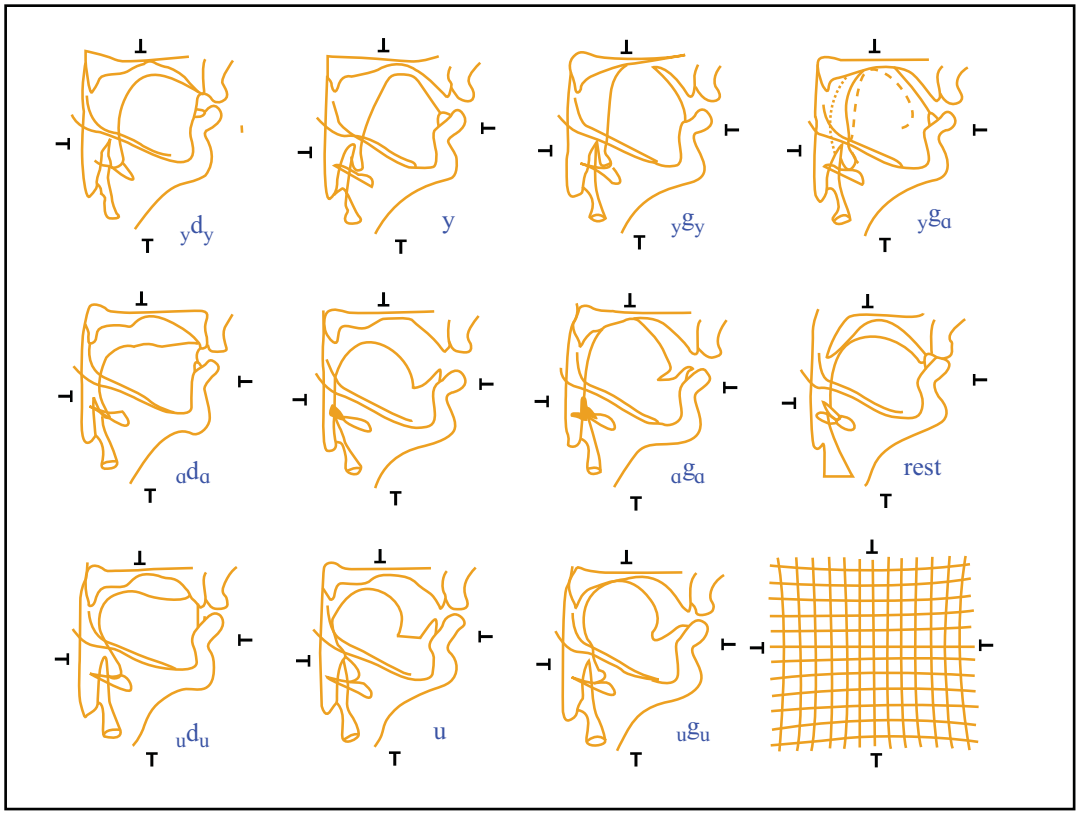
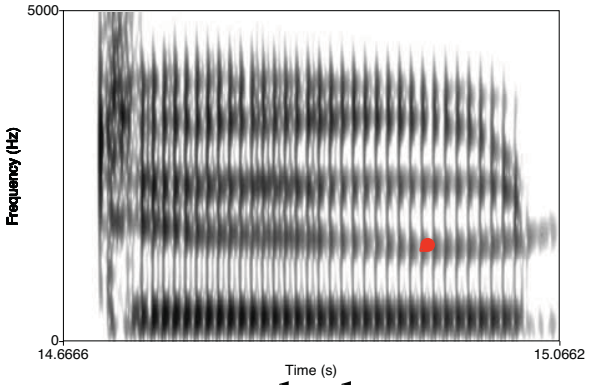


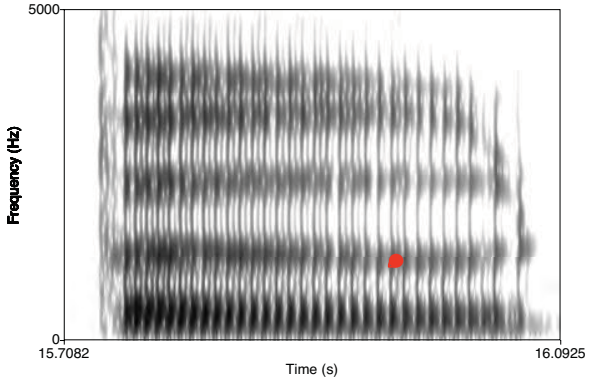
Image by MIT OCW.

Adapted from Ohman, S.E.G. "Coarticulation in VCV utterances: Spectrographic measurements." *Journal of the Acoustical Society of America* 39 (1966): 151-168.

- CV coarticulation - F2 frequency at the steady state of the vowel in turn depends on consonant context.
  - Vowels assimilate to surrounding consonants.

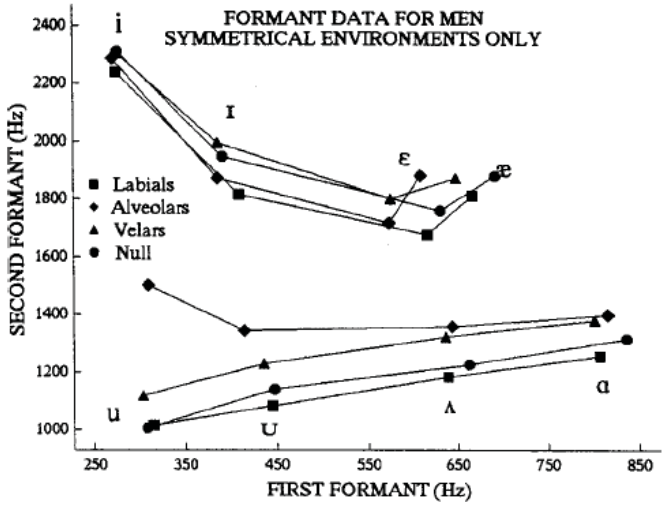


dud



bud

- Hillenbrand, Clark & Nearey 2001



Coarticulation between non-adjacent segments

Lip-rounding: Lip-rounding for rounded vowels has been reported to begin substantially before the onset of the vowel itself:

- 'Coarticulation of lip protrusion extends over as many as four consonants preceding the vowel /u/' (Daniloff and Moll 1968) - e.g. [sku], [ist#tu].
- Benguerel and Cowan (1974) report coarticulation of lip-rounding across seven segments in French.
  - 'une sinistre structure' [istrsty] vs.
  - 'une sinistre stricture' [istrsti]
- Perkell (1969) reports that protrusion starts at the beginning of English nonsense words like [hətu] (cf. [həti])

Coarticulation between non-adjacent segments

- Coarticulation between vowels across intervening consonants has been well-known since Öhman (1966).
  - Swedish VCV sequences

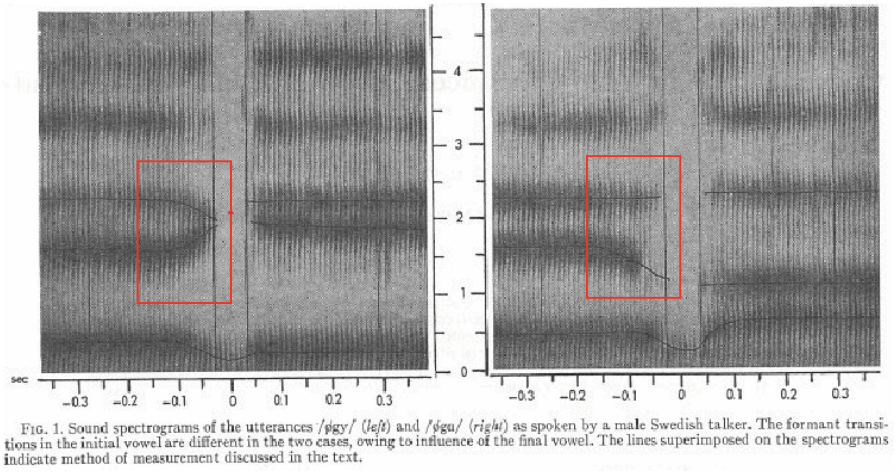


FIG. 1. Sound spectrograms of the utterances /ɸgy/ (*left*) and /ɸgu/ (*right*) as spoken by a male Swedish talker. The formant transitions in the initial vowel are different in the two cases, owing to influence of the final vowel. The lines superimposed on the spectrograms indicate method of measurement discussed in the text.

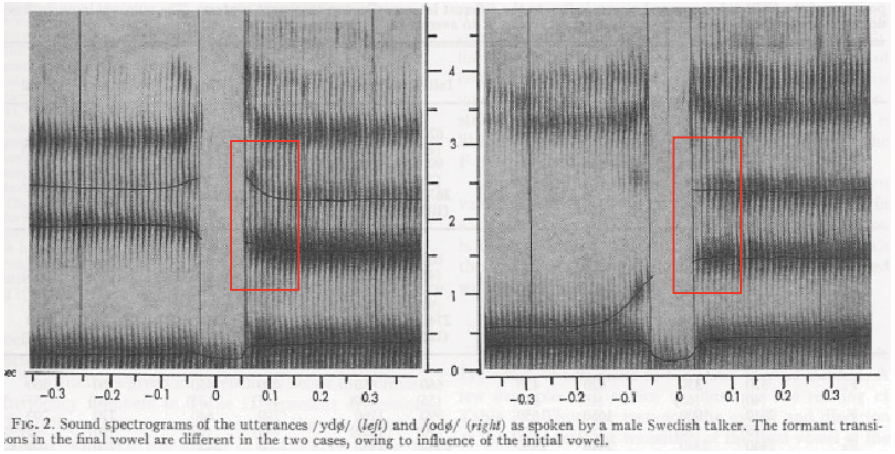


FIG. 2. Sound spectrograms of the utterances /ydø/ (*left*) and /odø/ (*right*) as spoken by a male Swedish talker. The formant transitions in the final vowel are different in the two cases, owing to influence of the initial vowel.

øgy øgu ydø odø

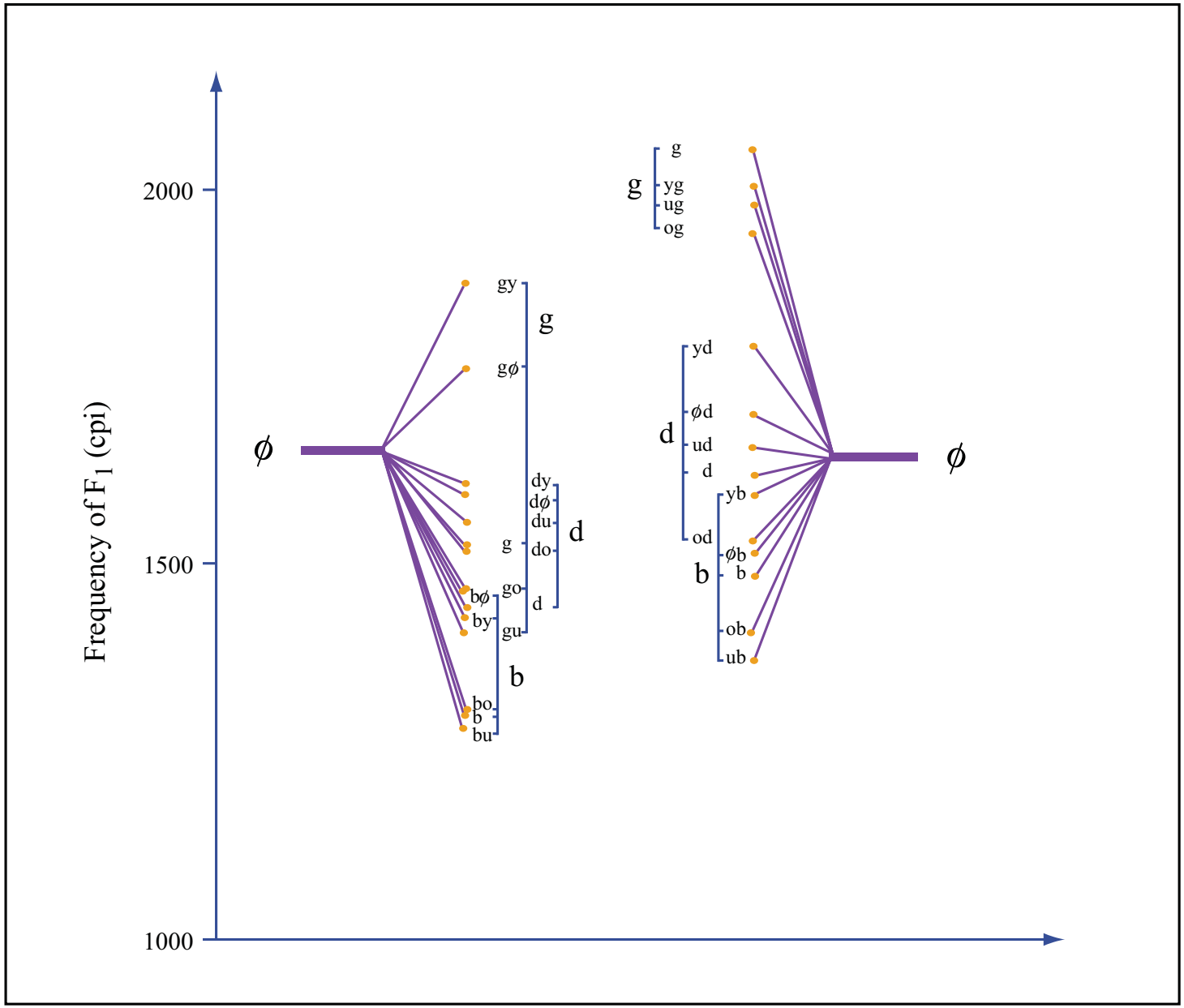


Image by MIT OCW.

Adapted from Ohman, S.E.G. "Coarticulation in VCV utterances: Spectrographic measurements." *Journal of the Acoustical Society of America* 39 (1966): 151-168.

Coarticulation between non-adjacent segments

- Öhman (1966)

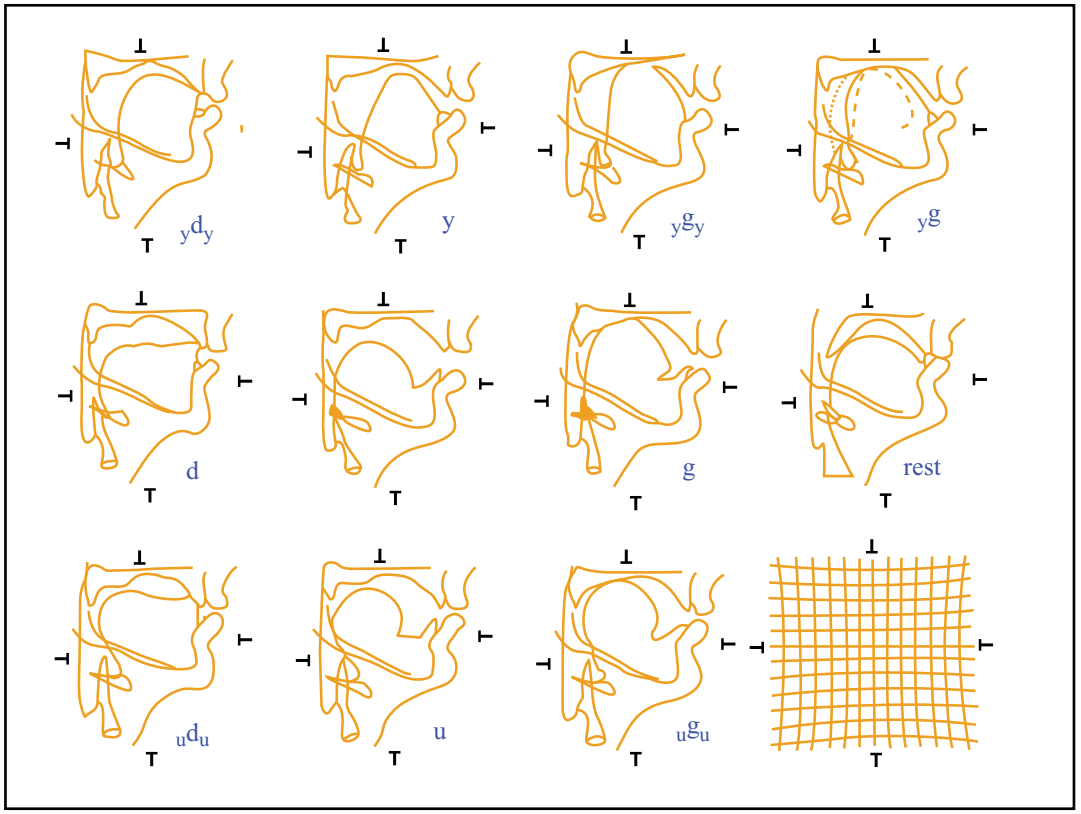


Image by MIT OCW.

Adapted from Ohman, S.E.G. "Coarticulation in VCV utterances: Spectrographic measurements." *Journal of the Acoustical Society of America* 39 (1966): 151-168.

Target variation

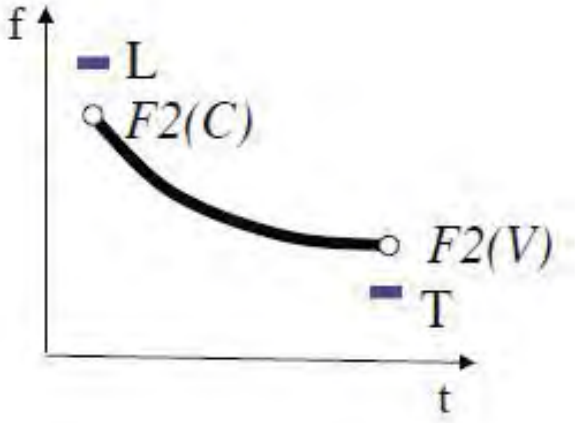
- Target variation suggests that we need a less rigid notion of a target, e.g. a range (Keating's windows) or a violable target (Lindblom 1963, Flemming 2001, Browman and Goldstein).

Violable targets:

- These kinds of target variation have been conceptualized in terms of undershoot: targets are consistent but are not always reached (e.g. Lindblom 1963).
- The basic reason for failure to achieve targets is hypothesized to be a dispreference for the effort involved in rapid transitions (minimization of effort).

CV coarticulation - an analysis

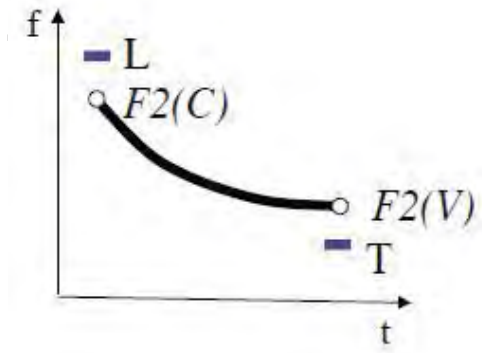
- F2 transitions are a compromise between:
  - achieving the F2 targets for consonant (*L*) and vowel (*T*)
  - avoiding fast movement between the two.
- Minimization of effort: movements with higher peak velocity are more effortful, other things being equal (Nelson 1983, Perkell 1997).
  - Peak velocity is proportional to displacement (e.g. Kent & Moll 1972)
  - Constraint: $F2(C) = F2(V)$



CV coarticulation - analysis

- Given *L*, *T*, select *F2(V)*, *F2(T)* so as to minimize violation of the following constraints (Flemming 2001):

	Constraint	*Cost of violation*
IDENT(C)	$F2(C) = L$	$w_c(F2(C)-L)^2$
IDENT(V)	$F2(V) = T$	$w_v(F2(V)-T)^2$
MINIMISEEFFORT	$F2(C) = F2(V)$	$w_e(F2(C)-F2(V))^2$



- These constraints conflict where *L* and *T* differ.
- The actual F2 transitions are a compromise between the constraints.
- Resolving conflict - minimize summed constraint violations:

  $cost = w_c(F2(C) - L)^2 + w_v(F2(V) - T)^2 + w_e(F2(C) - F2(V))$

  – w_i are positive weights.

  – one value of w_c for each C. How many values of w_v?

Finding optimal values

- Given the form of the constraints, the cost function is smooth and convex.
  - optimum lies at the bottom of a 'bowl'.
- So optimum can be found using simple search algorithms (e.g. steepest descent).
- In this case cost function is simple enough to derive a closed form solution.

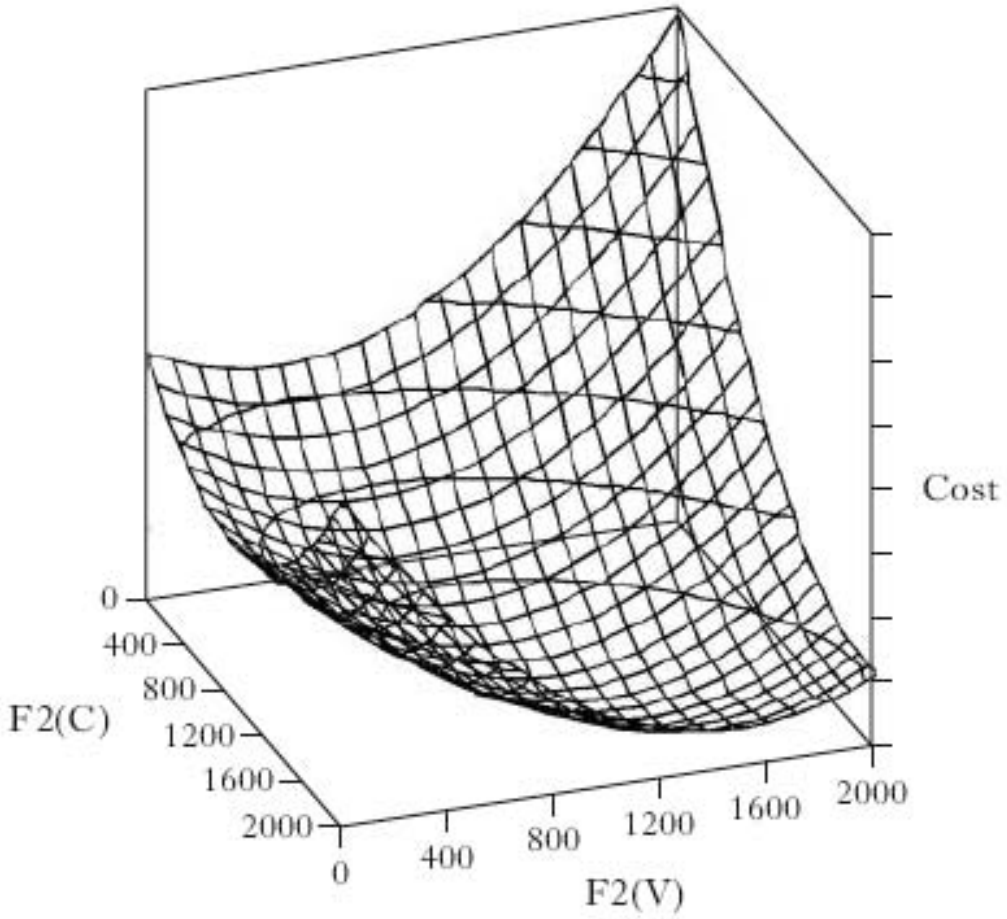


Figure 3
Cost plotted against $F2(C)$ and $F2(V)$, with $L = 1700$ Hz, $T = 1000$ Hz, and all weights set to 1. The minimum is located at $F2(V) = 1233$ Hz, $F2(C) = 1467$ Hz.

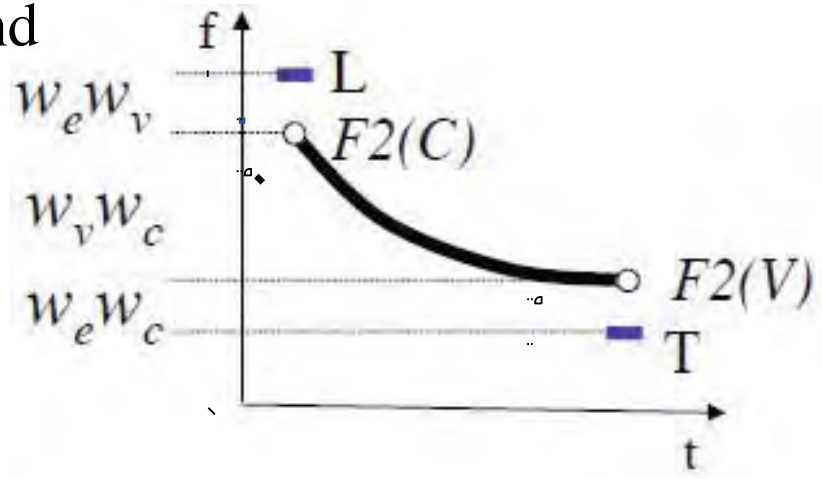
CV coarticulation - analysis

- Optimal values for *F2(C)*, *F2(V)* as a function of *L*, *T*:

$$F2(C) = -u_c(L - T) + L \qquad u_c = \frac{w_e w_v}{w_e w_c + w_v w_c + w_e w_v}$$

$$F2(V) = u_v(L - T) + T \qquad u_v = \frac{w_e w_c}{w_e w_c + w_v w_c + w_e w_v}$$

- The interval between *L* and *T* is divided into three parts by *F2(C)* and *F2(V)*
  - C undershoot
  - V undershoot
  - transition
- In the proportions $w_e w_v : w_e w_c : w_v w_c$



CV coarticulation - analysis

- Optimal value for *F2(C)* is a linear function of *F2(V)*, as observed empirically:

$$F2(C) = \frac{w_e}{w_c + w_e} F2(V) + \frac{w_c}{w_c + w_e} L$$

$$F2(C) = \frac{w_e}{w_c + w_e} (F2(V) - L) + L$$

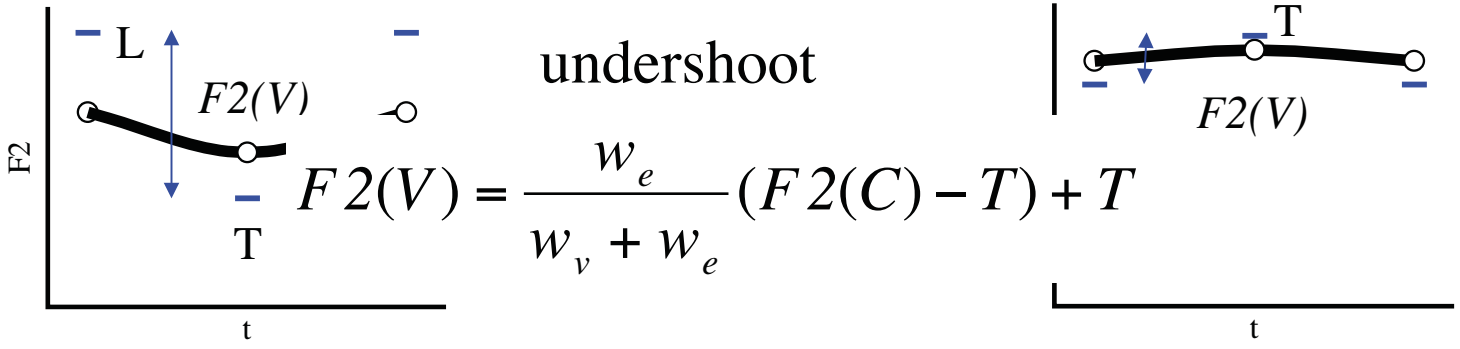
Figure removed due to copyright restrictions.
Source: Figure 1, Fowler, Carol A. "Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation." Attention, Perception, & Psychophysics 55, no. 6 (1994): 597-610.

- Vowel undershoot is proportional to the distance between *L* and T, for a given consonant context (Lindblom 1963, Broad & Clermont 1987):

$$F2(V) = \underbrace{u_v(L - T)} + T \quad (u_v \leq 1)$$

undershoot

[tut]



$$F2(V) = \frac{w_e}{w_v + w_e} (F2(C) - T) + T$$

[tit]

Estimating model parameters from the data

$$F2(C) = \underbrace{\frac{w_e}{w_c + w_e} F2(V)}_{\text{slope}} + \underbrace{\frac{w_c}{w_c + w_e} L}_{\text{intercept}}$$

- Weights for English vowels, based on Fowler (1994):

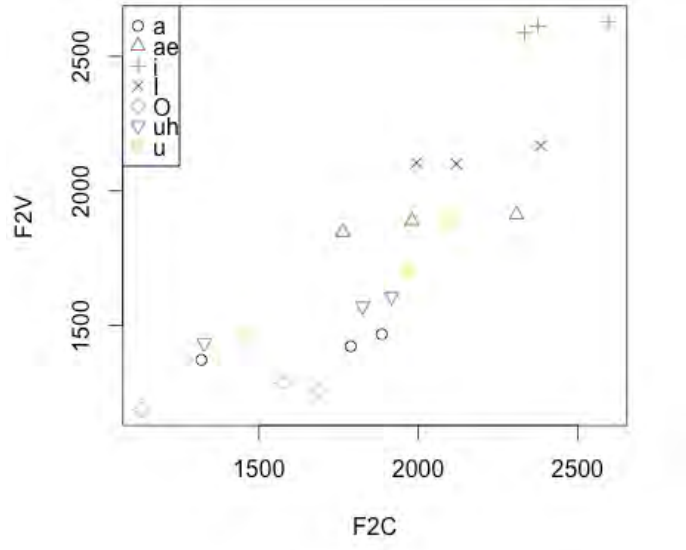
	slope	int	w_c	L	$w_e = 1$
b	0.80	228	0.25	1140 Hz	
d	0.48	1099	1.09	2113 Hz	
g	0.71	779	0.40	2709 Hz	

- This is not a general method for parameter estimation – some constraint models are more complex (as we will see).

Estimating model parameters from the data - vowels

$$F2(V) = \underbrace{\frac{w_e}{w_v + w_e}}_{\text{slope}} F2(C) + \underbrace{\frac{w_v}{w_v + w_e}}_{\text{intercept}} T$$

- We need one value of w_c for each C because slope and intercept differ for each C. How would we know if we need one value of w_v for each V? What difference does it make?



Estimating model parameters from the data - vowels

$$F2(V) = \underbrace{\frac{w_e}{w_v + w_e}}_{\text{slope}} F2(C) + \underbrace{\frac{w_v}{w_v + w_e} T}_{\text{intercept}}$$

- Rough estimates of weights and targets for English vowels, based on Fowler (1994):

	slope	int	w_v	T	$w_e = 1$
ɪ	0.13	2287	8.5	2638 Hz	
ɪ	0.18	1740	6.6	2116 Hz	
æ	0.12	1649	9.7	1864 Hz	
ʌ	0.29	1052	4.5	1478 Hz	
ɑ	0.15	1174	7.7	1379 Hz	
ɔ	0.16	1008	7.1	1204 Hz	
u	0.63	528	2.6	1427 Hz	

CV coarticulation - analysis

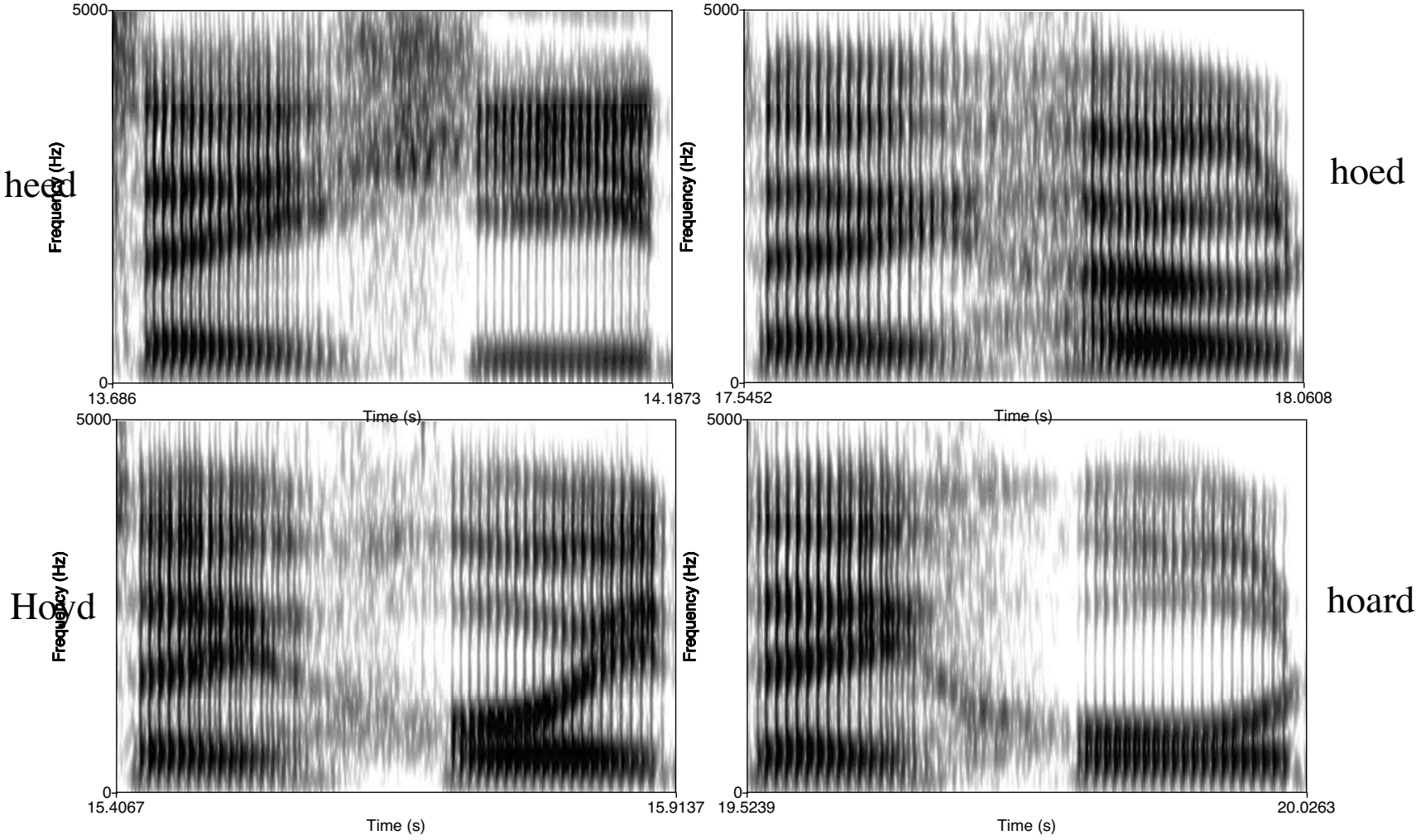
- Cross-linguistic variation in locus equations for similar consonants can be analyzed as variation in w_c and L_c
  - Thai [d̪] $F2(C) = 0.3F2(V) + 1425$ (0.24-0.33)
  - Urdu [d̪] $F2(C) = 0.5F2(V) + 857$ (0.43-0.57)
  - Sussman et al (1993).
- Fix $w_e = 1$
- Thai: $w_d = 2.3, L_d = 2036$ Hz
- Urdu: $w_d = 1.0, L_d = 1714$ Hz
- This is only the beginnings of a typological analysis:
  - Where does *L* come from?
  - What are the limits on variation in constraint weights?

Keating's Window Model

- An alternative analysis of 'target variation' is to propose that targets specify a range of permissible values and that the observed variation falls within these target ranges.
  - Implies that there is no undershoot.
- Keating's window model of coarticulation develops this approach.
- Originated as a refinement of an earlier proposal that segments could lack targets on some dimensions ('phonetic underspecification') (Keating 1988).

Keating (1988)

- Example of underspecification: Argues that [h] lacks specifications for oral features, based on data like the following:



Keating's (1990) 'Windows' model

- Phonetic underspecification á la Keating (1988) allows only inviolable targets on a parameter, or no target at all (freely variable).
- Keating (1990) argues that this is too simplistic - targets may vary in degree of specificity.
- Implemented by replacing point targets with 'windows' specifying a range of acceptable values on a parameter.

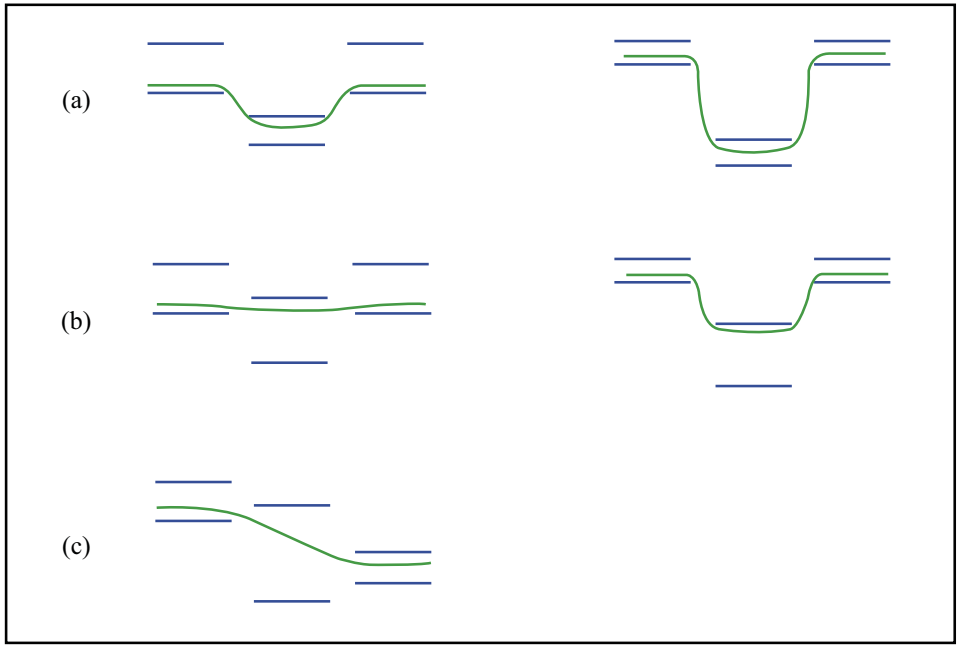


Image by MIT OCW.

Adapted from Keating, P. A. "The window model of coarticulation: articulatory evidence." In *Papers in LaboratoryPhonology I: Between the Grammar and Physics of Speech*. Edited by John Kingston and Mary E Beckman. New York, NY: Cambridge University Press, 1990, pp. 451-470. ISBN: 9780521368087.

Keating's (1990) 'Windows' model

- Motivated by evidence for segments that are exhibit substantial, but bounded, contextual variability on a parameter. E.g. velum height in

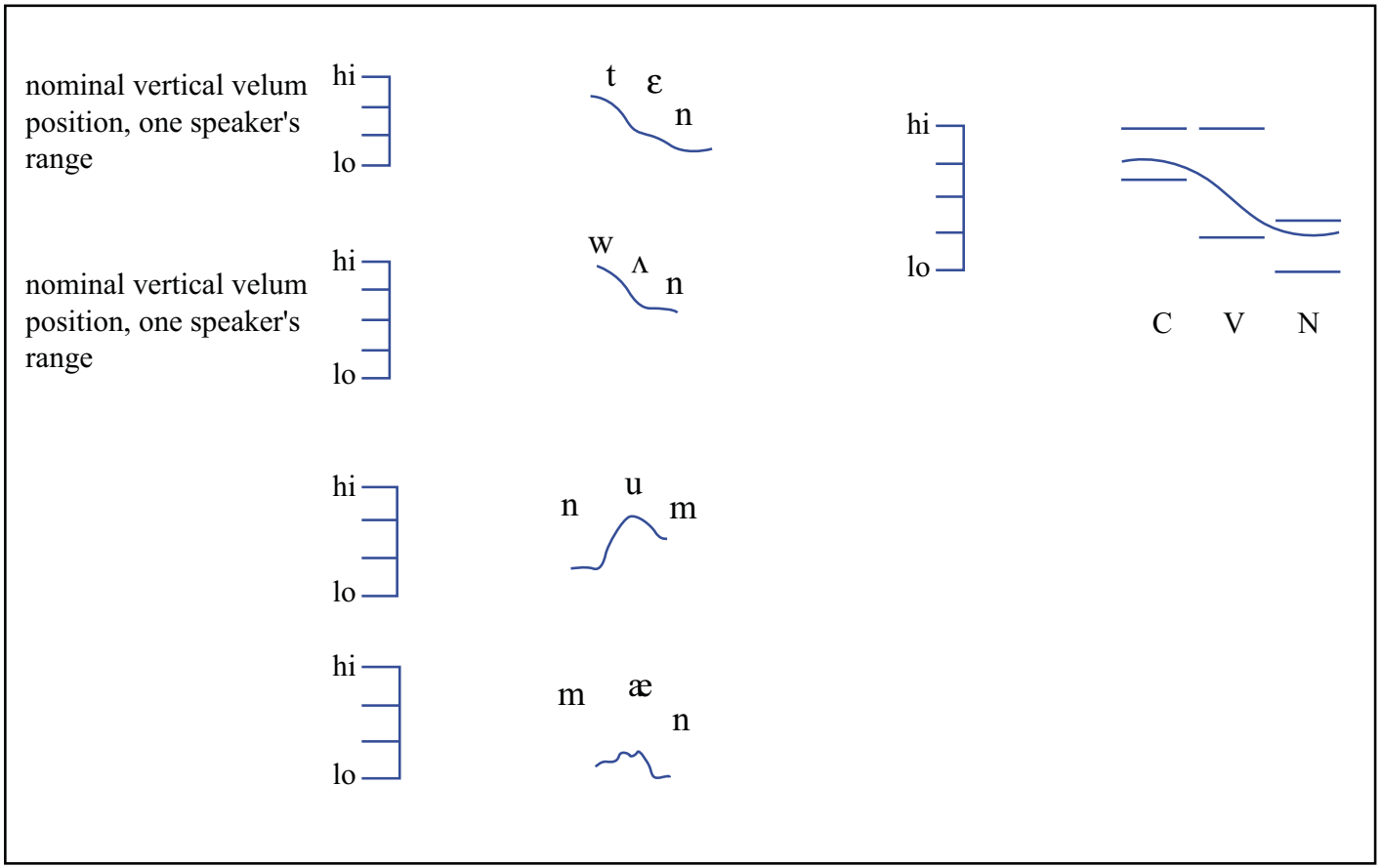
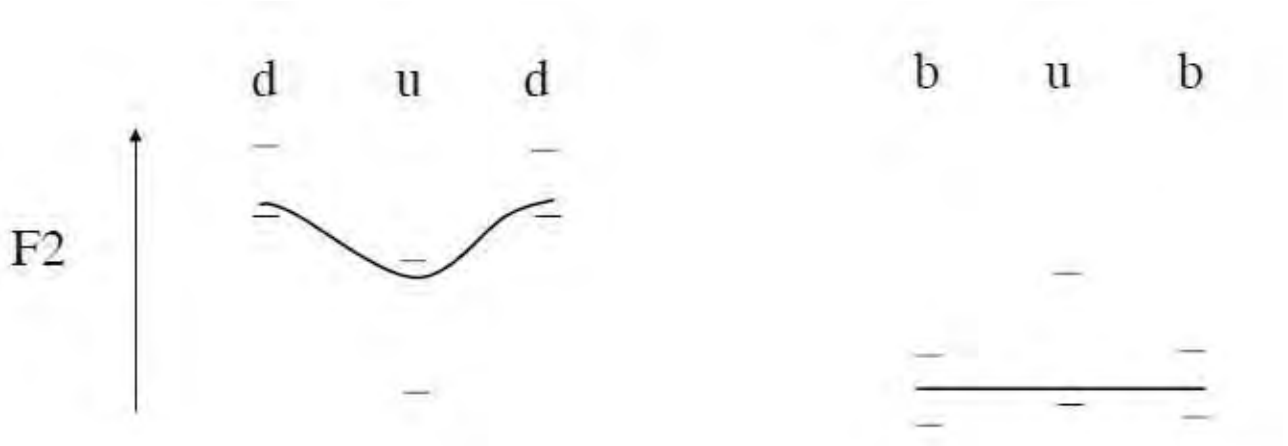


Image by MIT OCW.

Adapted from Keating, P. A. "The window model of coarticulation: articulatory evidence." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Edited by John Kingston and Mary E Beckman. New York, NY: Cambridge University Press, 1990, pp. 451-470. ISBN: 9780521368087.

Modeling C-V coarticulation: Windows model



- [u] has a wide window for F2 (or tongue body backness).
- Optimal trajectory minimizes peak velocity (Keating 1990:456)
- So the optimal trajectory passes through differents parts of the [u] window, depending on context (coarticulation).

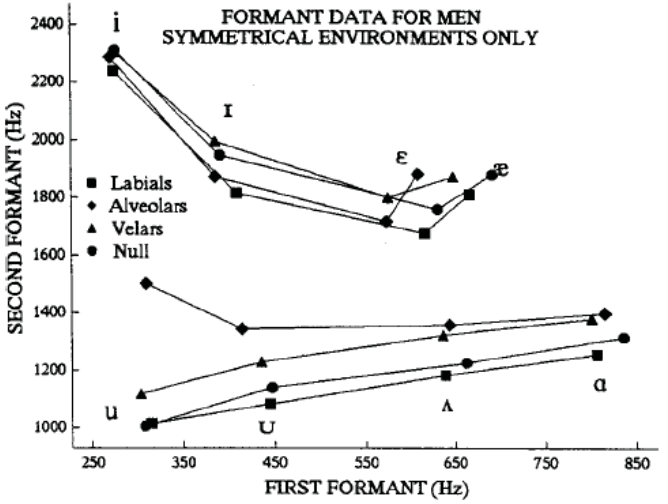
Violable targets vs. windows

- Window model treats all realizations that fall within a window as equally good.
- In the undershoot model, deviations from the target are dispreferred.
- Evidence from CV coarticulation supports the undershoot model.

Violable targets vs. windows

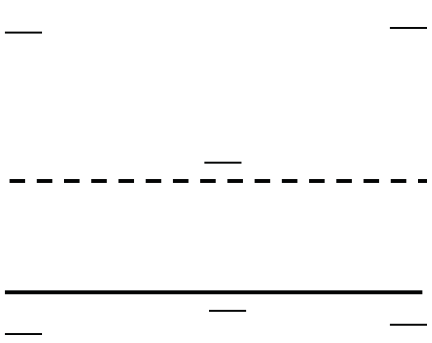
- [b] must have a wide window for F2/tongue body position
- [u] must have a relatively wide F2 window to account for [bub]/[dud] variation

Figure removed due to copyright restrictions.
Source: Figure 2, Flemming, Edward. "Scalar and categorical phenomena in a unified model of phonetics and phonology." Phonology 18, no. 01 (2001): 7-44.



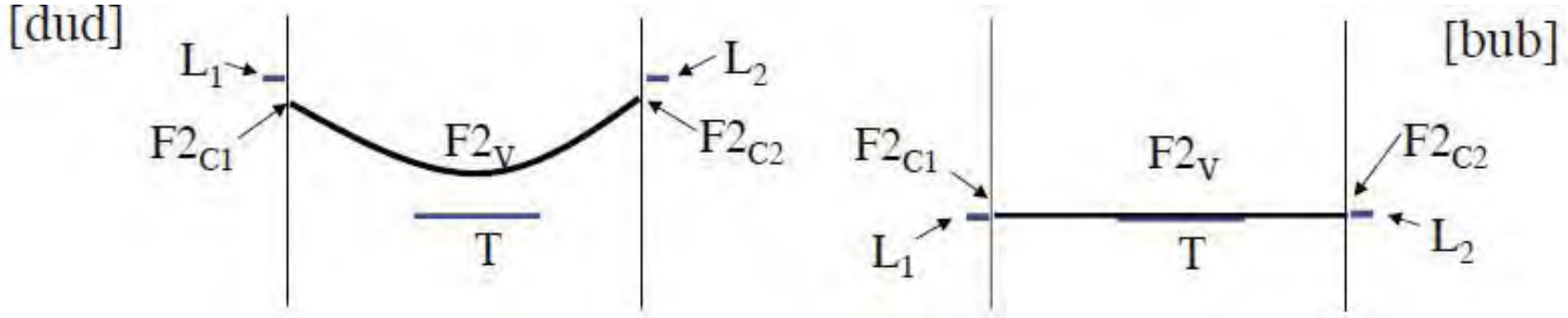
Violable targets vs. windows

- So a sequence like [bub] consists of three wide windows.
- When all windows are wide, the optimal trajectory is underdetermined because there is a range of flat (minimum effort) trajectories that pass through all the windows.
- So the windows analysis leads us to expect free variation.
- In fact we observe a low F2 trajectory in [bub].



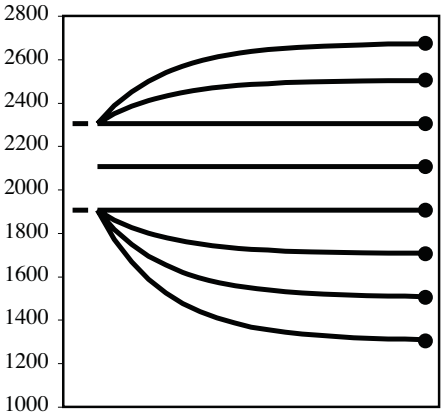
Violable targets vs. windows

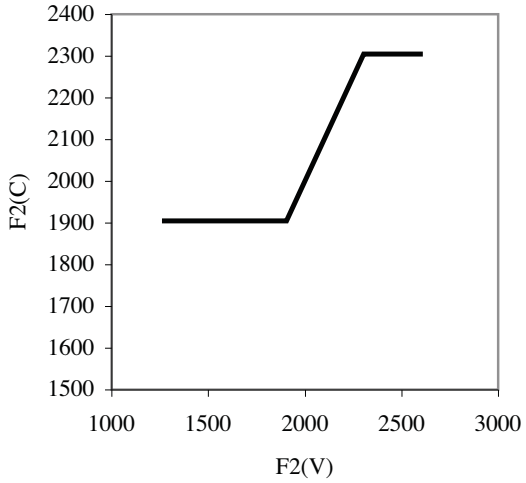
- In fact we observe a low F2 trajectory in [bub].
- This follows from the weighted targets model:
  - [u] has a low F2 target which is undershot in [dud] due to the distance between [d] and [u] targets and their relative weights.
  - [b] has a lower-weighted F2 target (hence the contextual variability of F2 adjacent to [b]), so [b] assimilates to [u] and [u] is realized faithfully.
  - E.g. to fit the Fowler data, $w_e = 1$ (only ratios of weights matter),
    - [b]: $L_b = 1140$ Hz, $w_{c(b)} = 0.8$
    - [d]: $L_d = 2098$ Hz, $w_{c(d)} = 1.1$



Violable targets vs. windows

- The windows model predicts that there should be a sharp distinction between realizations that fall inside and outside a target window (good vs. impossible).
- Predicts discontinuities in coarticulatory variation at window edges.
- E.g. [d] would have a window for F2.
  - Expect total assimilation to vowels whose F2 is within the window range.
  - No assimilation to vowels outside this range.
- Actual coarticulatory variation is a smooth function of vowel F2.
  - Derived by weighted targets model.





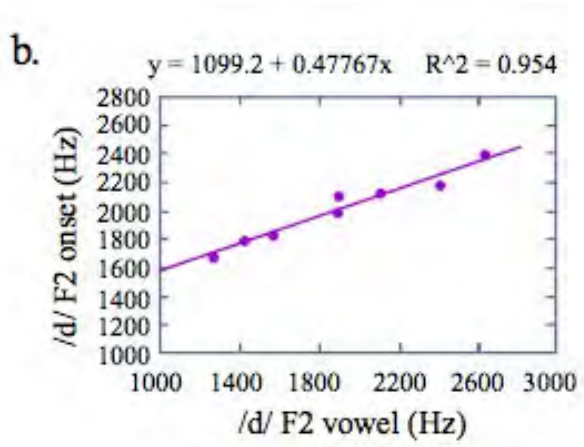


Image by MIT OCW.

24.915 / 24.963 Linguistic Phonetics

Fall 2015