24.963 Linguistic Phonetics
Fall 2005

24.963

Linguistic Phonetics

Speech Perception: The Problem of Variability

The effects of voicing on vowel formants

- In many accents, the realization of /aɪ/varies as a function of the voicing of a following obstruent.
- Previous studies indicate that the most consistent difference is in the formants of the offglide: F2 is higher and F1 is lower preceding voiceless consonants.
- Kwong and Stevens propose an explanation for this effect based on facilitation of voicing/voicelessness in the following obstruent:
  - Pharyngeal expansion facilitates devoicing.
  - Reducing pharyngeal expansion in the offlgide allows voicing to be facilitated by expansion during stop closure.

The effects of voicing on vowel formants

- The pharynx is expanded in high vowels (and offglides).
- Differences in degree of expansion is expected based on voicing of following obstruent.

Predictions:

- High front vowels: beforce voiceless, F2 is higher and F1 is lower.
- High back vowels: before voiceless, F2 is lower and F1 is higher.
- Lax and non-high vowels: no pharyngeal expansion, hence no difference in formants.

Results

- F2 offset: differences are as predicted

offglide	voiced	voiceless
front	2342 (213)	2607 (145)
back	1695 (166)	1450 (210)
none	2035 (160)	2028 (176)

- F1 offset:
  - Front: no difference (unpredicted)
  - Back: F1 slightly higher before voiceless (unpredicted)
  - Large difference with low vowels (unpredicted)

offglide	voiced	voiceless
front	359 (58)	353 (66)
back	370 (67)	404 (57)
none	473 (113)	648 (184)

Statistical analysis

- Repeated Measures ANOVA: takes into account that multiple data points are collected from each subject (‘repeated measures’ of each subject).
- Data from the same subject are not independent - this must be taken into account in the analysis.
- Taking subject into account also allows us to factor out between subject variability.
- Subjects are a random sample from the population of potential subjects - this must be taken into account if we want to be able to generalize our results to a broader population.

Statistical analysis

- In many experiments the same considerations apply to stimuli (e.g. the words in our experiment):
  - each item is produced by multiple subjects
  - items are often a subset of those that we are interested in (e.g. all words containing high vowels followed by coda /t/).
- In these cases it is necessary to perform a second repeated measures ANOVA with items as the repeated measure.
- The two F-ratios are then combined into a quasi F-ratio *F'* (or a lower bound for it *minF'*- Clark 1973).
- The two ANOVAs are often referred to as Subjects analysis and Items analysis.
- It is common practice in psychology to report both, it is much less common to report *F'* or *minF'*.
- If items are carefully matched (as in our experiment), it is not appropriate to use *minF'* (Raaijmakers et al 1999).

Statistical analysis

What's this sphericity thing? Who are Huynh and Feldt?

- Sphericity is a property assumed in the repeated measures ANOVA model (equality of variances of differences between levels of a factor).
- If you apply repeated measures ANOVA to data that violates the sphericity assumption, it is necessary to correct by reducing the degrees of freedom.
- The Huynh-Feldt epsilon is an estimate of the necessary adjustment - the degrees of freedom are multiplied by epsilon.

ANOVA - F1 offset, high front offglide

Number of obs = 30 R-squared = 1.0000
Root MSE = 0 Adj R-squared =

Source	Partial SS	df	MS	F	Prob > F
Model	108273.467	29	3733.56782		
subject	70051.1333	4	17512.7833	15.36	0.0108
voicing	235.2	1	235.2	0.21	0.6732
subject*voicing	4560.46667	4	1140.11667		
pair	18298.0667	2	9149.03333	10.49	0.0058
subject*pair	6977.26667	8	872.158333		
voicing*pair	4308.2	2	2154.1	4.48	0.0494
subject*voicing*pair	3843.13333	8	480.391667		
Residual	0	0			
Total	108273.467	29	3733.56782		

ANOVA - F1 offset, high front offglide

Repeated variables: voicing*pair

Huynh-Feldt epsilon = 0.7633
Greenhouse-Geisser epsilon = 0.6210
Box's conservative epsilon = 0.5000

Source	df	F	Prob > F Regular	H-F	G-G	Box
voicing*pair	2	4.48	0.0494	0.0692	0.0850	0.1016
subject*voicing*pair	8					

Results

- F2 offset: differences are as predicted

offglide	voiced	voiceless	
front	2342 (213)	2607 (145)	$p < 0.01$
back	1695 (166)	1450 (210)	$p < 0.01$
none	2035 (160)	2028 (176)	$p = 0.85$

  - No significant voicing*pair interactions.
- F1 offset:

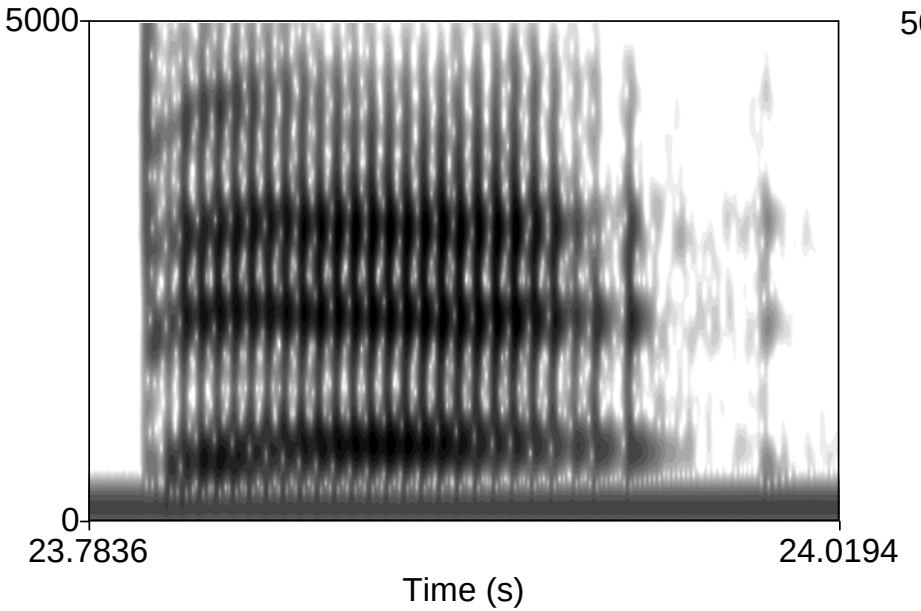
offglide	voiced	voiceless	
front	359 (58)	353 (66)	$p = 0.67$
back	370 (67)	404 (57)	$p < 0.05$
none	473 (113)	648 (184)	$p < 0.01$

  - Marginally significant voicing*pair interaction for front offglides.

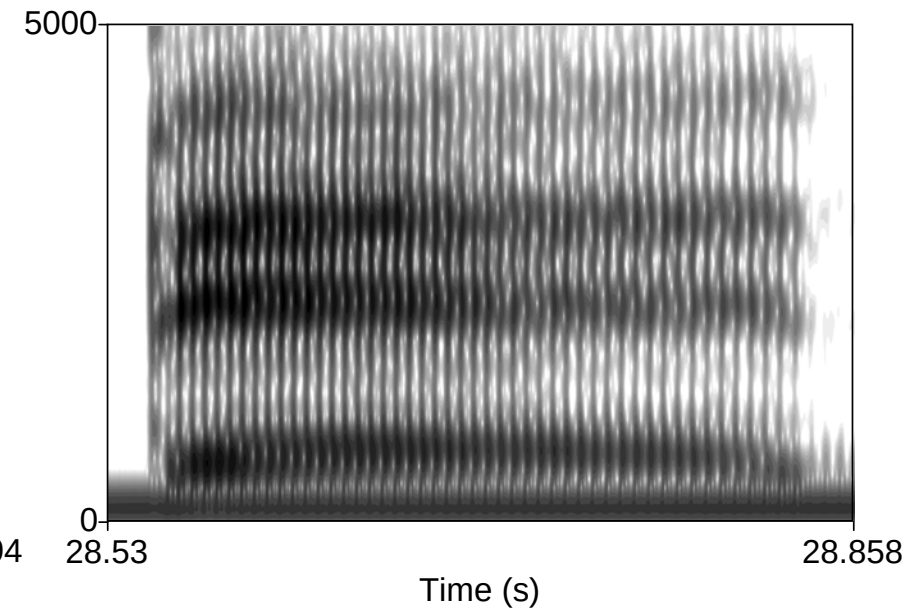
Glottalization

- Final /t/s are often glottalized.
- This probably explains the huge difference in F1 offset of [ɛ] and [æ] before voiced and voiceless - the formant transitions were truncated by glottal closure before significant oral constriction had been achieved.

bet



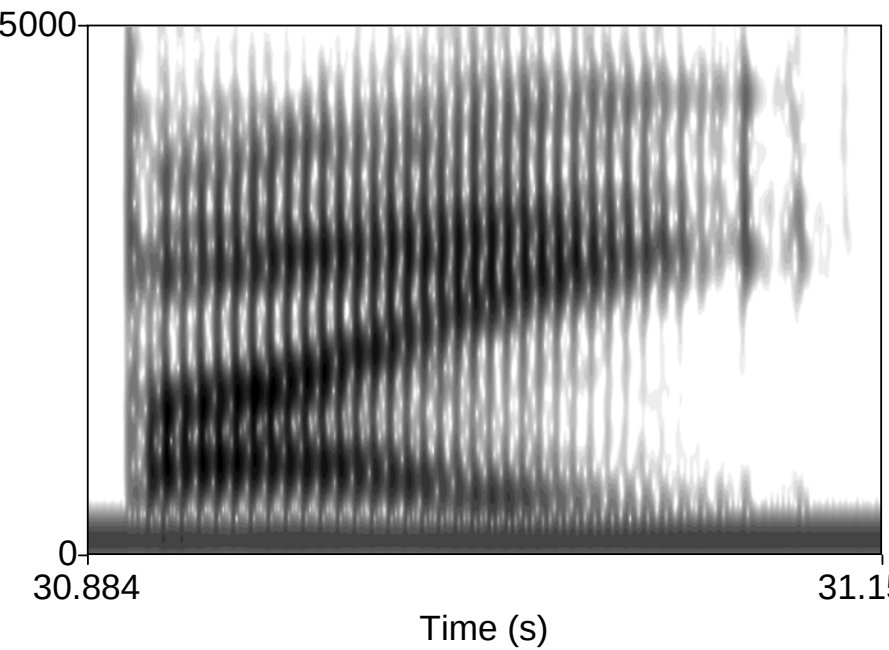
bed



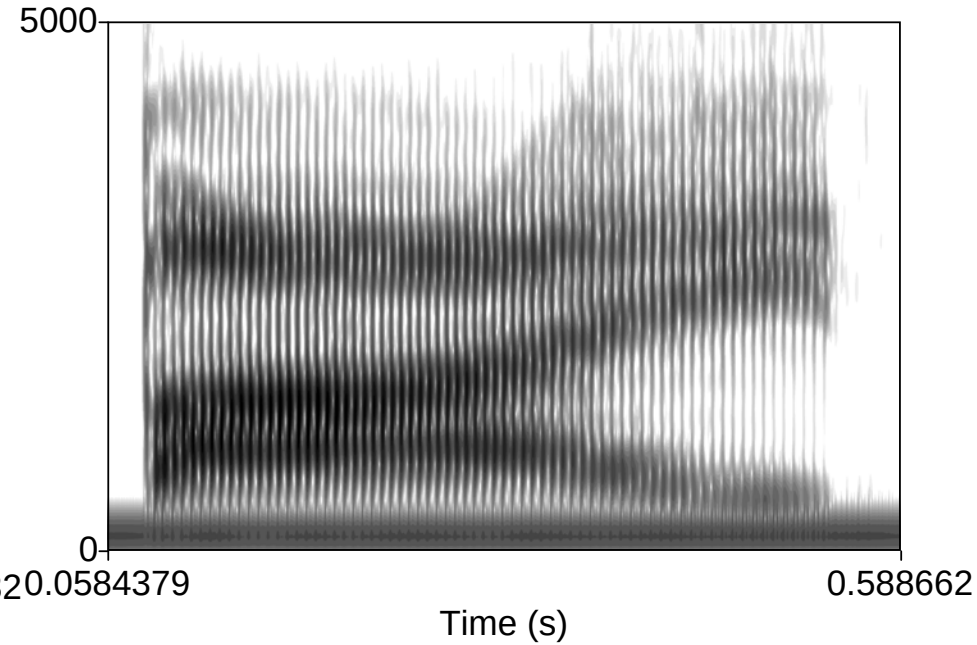
Glottalization

- Glottalization could also be a factor in the observed F2 effects:

bite



bide



Results - steady states

- F2 offset: differences are as predicted

offglide	voiced	voiceless	
front	2293	2433	$p < 0.05$
back	1390	1443	$p < 0.05$
none	2081	2038	$p = 0.01$

  - significant voicing*pair interaction for ‘no offglide’
- F1 steady state:

offglide	voiced	voiceless	
front	542	476	$p < 0.01$
back	531	568	$p = 0.22$
none	661	739	$p < 0.01$

  - Significant voicing*pair interaction for front offglides.

Previous results

- Moreton (2004) did a very similar study to ours.
- Only looked at diphthongs /aɪ, ɔɪ, eɪ, aʊ/.
- Measured F2 max/min in offglide - could help to avoid glottalization problems.
- Found that offglide F1 is lower and F2 is more extreme before voiceless.
- Cites previous studies showing that F1 steady state and offset is higher before voiceless in low vowels [æ, ɑ] (Summers 1987, Crowther and Mann 1992).
- Support for several of Kwong and Stevens's predictions, but additional unexpected effects.
- Moreton: vowels are hyperarticulated more as you move closer to a voiceless consonant.

Speech Perception - The Problem of Variability

- The acoustic realizations of segments and words are highly variable.
- The listener must identify all of these diverse acoustic signals as representing the same thing (at least for words).

Speech Perception and Lexical Access

- The problem faced by the listener: To extract meaning from the acoustic signal.
- It is clear that this task involves the recognition of words
- Schematic model of lexical access (cf. Klatt 1989):

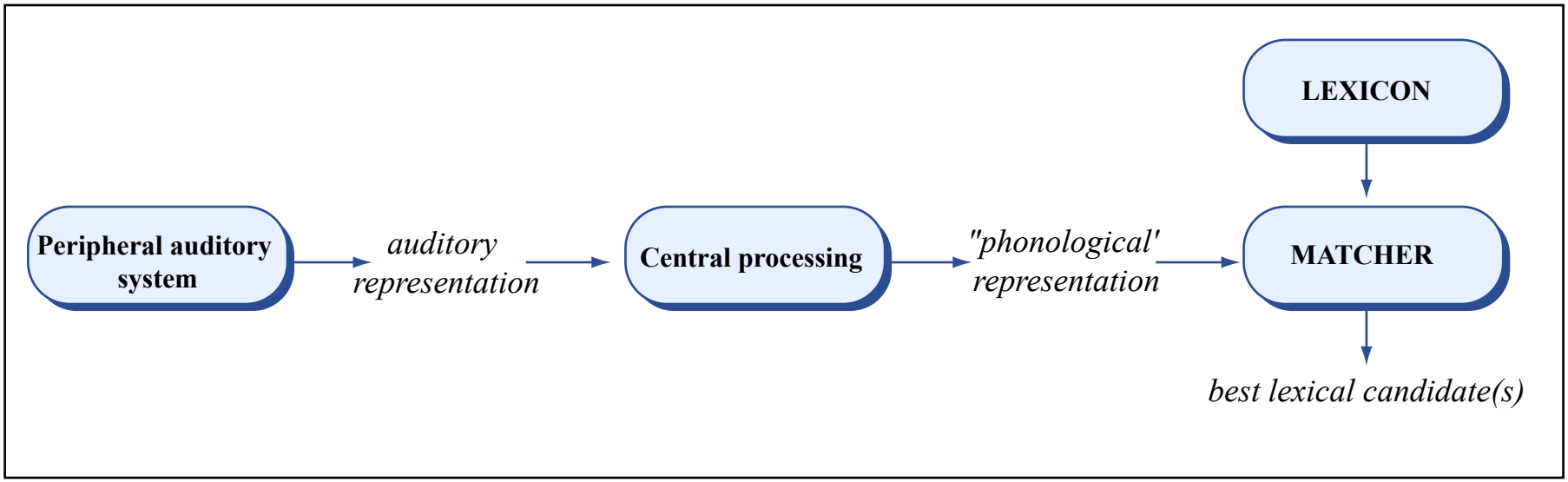


Image by MIT OpenCourseWare. Adapted from Klatt, Dennis H. "Review of Selected Models of Speech Production." In *Lexcial Representation and Process*. Cambridge, MA: MIT Press, 1989.

- Much controversy surrounds the nature of intermediate representations.
- Integration of speech perception and lexical access.

Speech Perception - The Problem of Variability

Sources of variability:

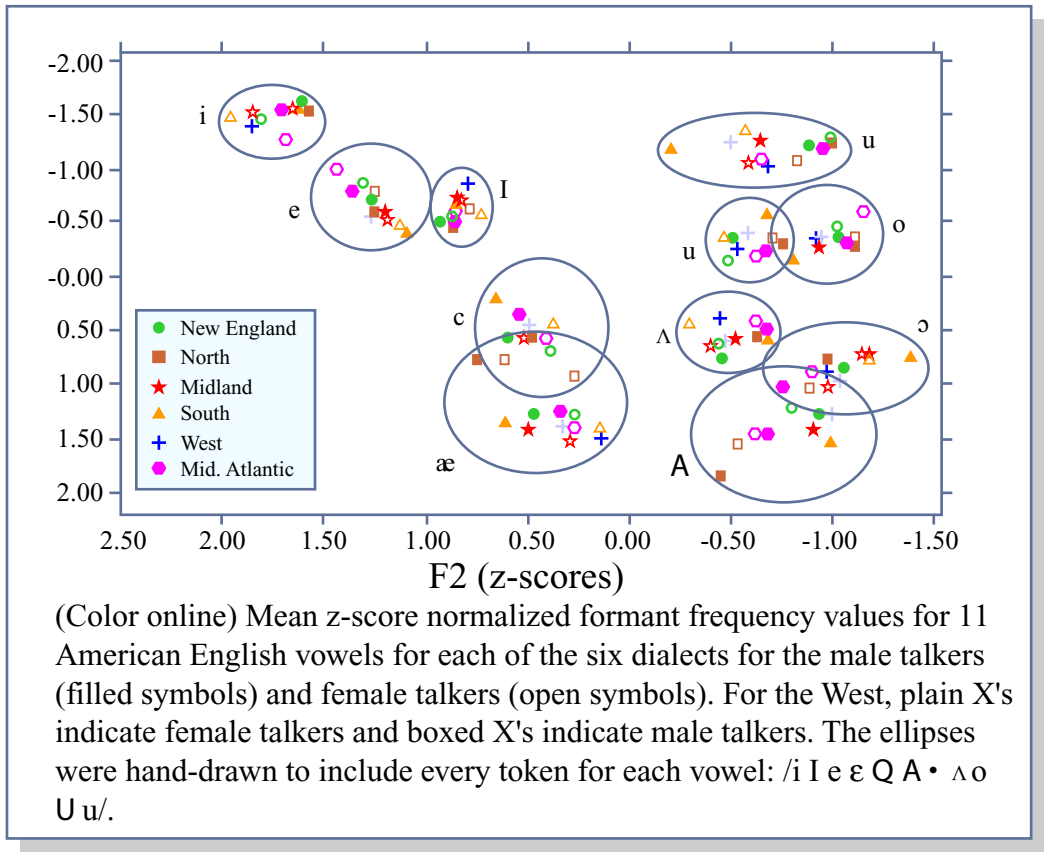
- Environment - background noise, room reverberation etc.
- Cross-speaker - vocal tract size, vocal folds, articulatory habits, dialect, etc.
- Within-speaker - physical and emotional state, etc.
  - Segment - coarticulation, speech rate, register, prosodic position (syllabic, phrasal, stress)
  - Word - cross-word coarticulation, speech rate, register, prosodic position (phrasal, stress)

Sources of Variability

- Cross-speaker - vocal tract size

Graph removed due to copyright restrictions.
Please see Figure 8 in Peterson, G. E., and H. L. Barney. "Control Methods Used in a Study of Vowels." *Journal of the Acoustical Society of America* 24 (1952): 175-184.

Dialect variation



(Color online) Mean z-score normalized formant frequency values for 11 American English vowels for each of the six dialects for the male talkers (filled symbols) and female talkers (open symbols). For the West, plain X's indicate female talkers and boxed X's indicate male talkers. The ellipses were hand-drawn to include every token for each vowel: /i I e ɛ Q A • ʌ o U u/.

Image by MIT OpenCourseWare. Adapted from Clopper, C. G., D. B. Pisoni, and K. J. de Jong. "Acoustic Characteristics of the Vowel Systems of Six Regional Varieties of American English." *Journal of the Acoustical Society of America* 118 (2005): 1661- 1676.

Hypothesis of acoustic invariance

- Blumstein and Stevens 1979: features/segments are identified by invariant (relational) cues in the acoustic signal.
- It is possible to find invariants amid the variability.

Problems:

- Invariants have proven very difficult to identify.
- It is unlikely that there are invariant cues that appear in all casual speech renditions of a segment, (Browman and Goldstein 1990).
  - Other mechanisms required.
- Like most theories of speech perception, it does not address cross-dialect perception.
- Listeners seem to make use of systematic variability

Segmental context-dependence

- Hypothesis of invariance implies that variability is discarded.
- Most contextual variation is systematic and appears to be exploited, not ignored.
- E.g. Mann and Repp (1980).
- But NB relational invariants.

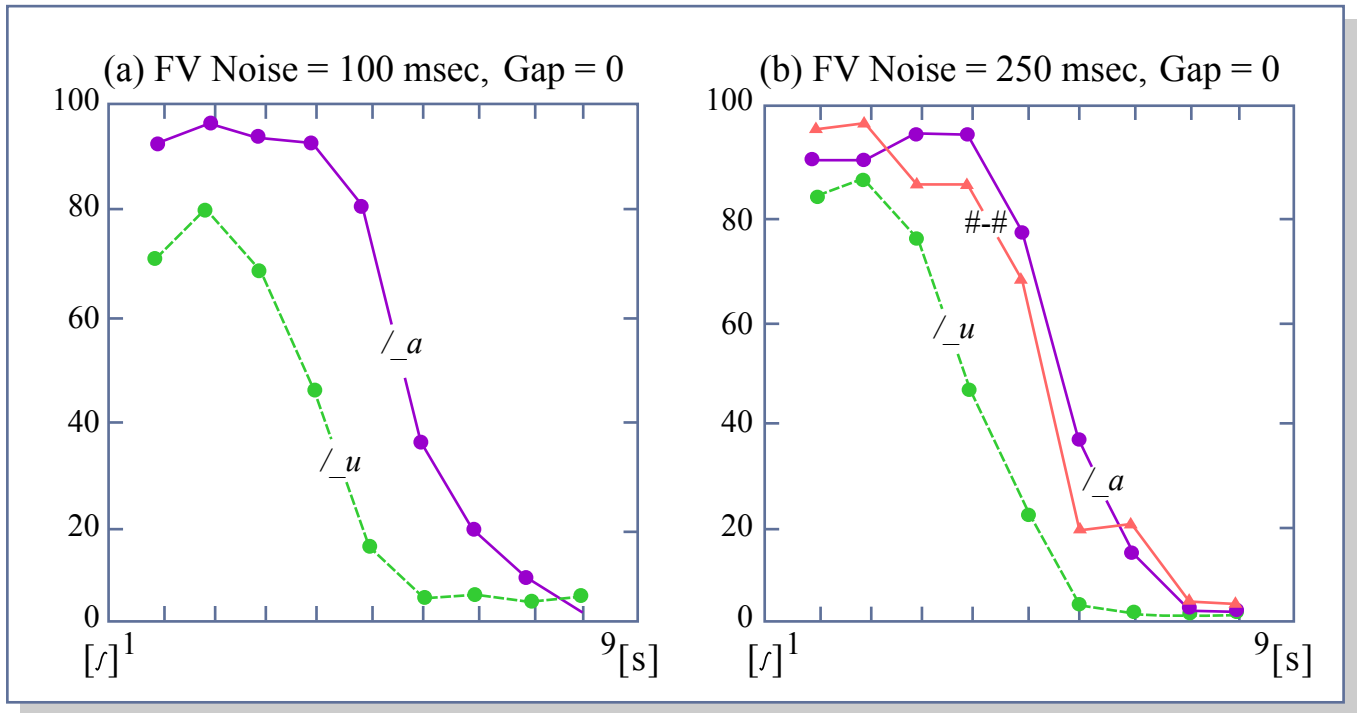


Image by MIT OpenCourseWare. Adapted from Mann, V. A., and B. H. Repp. "Influence of Vocalic Context on the Perception of [ ]-s] Distinction: I. Temporal Factors." *Perception and Psychophysics* 28 (1980): 213-228.

Exploiting lawful variation

Lawful contextual variation might be processed in a variety of ways:

- 'Parsing' - parse signal using rules of contextual variation to recover segments. E.g. compensation for coarticulation (Fowler and Smith 1986), context dependent classification rules (Nearey 1997).
- Analysis-by-synthesis - try to synthesize a match for the input by applying production rules to lexical forms (cf. Halle and Stevens 1959, Stevens and Halle 1964).
- Precompilation - Generate a stored lexicon of alternative realizations for each word using rules (Klatt 1979).
- Exemplar model - store exemplars drawn from many contexts (doesn't directly exploit laws).

The problem of variability

- Variability in the realization of segments due to segmental and prosodic context is not the hardest problem - handled fairly well in Automatic Speech Recognition.
- But these strategies might also be applicable to harder problems:
  - speaker variation (within and across dialect)
  - rate and register variation.

Exemplar-based models of categorization

Prototype models:

- Listeners construct prototypes for categories (words, sounds).
- Categorization proceeds by matching incoming instances to prototypes in memory.

Exemplar models:

- Listeners store categorized instances in memory (exemplars).
- Categorization proceeds by matching incoming instances to the set of exemplars for each category.

- Prototypes are constructed from instances in learning, but only the abstracted prototype is remembered.
- Exemplar models hypothesize that we store instances in considerable detail. Abstraction is performed in the process of matching a new token to the stored exemplars.

Exemplar-based models of categorization

General plausibility of episodic memory:

- There is good evidence for detailed long-term auditory memory.

Palmeri, Goldinger and Pisoni (1993):

- Subjects heard words spoken by 2, 6, 12, or 20 speakers.
- Subjects classified each word as ‘old’ or ‘new’.
- ‘Old’ words are recognized more accurately when spoken by the same voice.
- Goldinger (1996): This advantage persists for at least a day, but is lost after a week.
- ‘Old’ word identification is facilitated if the repetition is in a similar voice.

Speaker normalization

- Speaker variability is a major problem in ASR but apparently unproblematic for people.
- In practice physical differences between speakers often combine with dialectal differences but studies have focused on physiologically-based differences.

Speaker normalization

- Types of normalization models (Johnson 1990):
  - Intrinsic: normalization only uses local information, (e.g. vowels normalized by f0, higher formant frequencies).
  - Extrinsic: non-local properties are used (e.g. vowel formant range, average f0).

  - Direct: normalization information is used directly in constructing perceptual representations of segments.
  - Indirect: normalization information is used to create a frame of reference for the interpretation of segments.

Speaker normalization

Types of normalization models - examples:

- Intrinsic direct: Syrdal and Gopal (1986) - vowels are represented in terms of the differencs F3-F2, F2-F1, F1-F0 in Bark.
- Extrinsic indirect: Gerstman (1968) - vowel formants are normalized with respect to speaker's maximum and minimum F1 and F2. (cf. Lobanov 1971).
- Extrinsic direct: Modify Syrdal and Gopal normalizing F1 w.r.t. long-term average F0.

Extrinsic information in speaker normalization

Evidence that extrinsic information is used:

- Ladefoged and Broadbent (1957).

http://www.jladefoged.com/acousticdemos/acoustics/acoustics.html

  - Identification of an ambiguous 'bit/bet' stimulus is influenced by the formants of the preceding vowels in the sentence.

- Johnson (1990) - perception of vowels with the same f0 is influenced by f0 of the preceding carrier phrase.

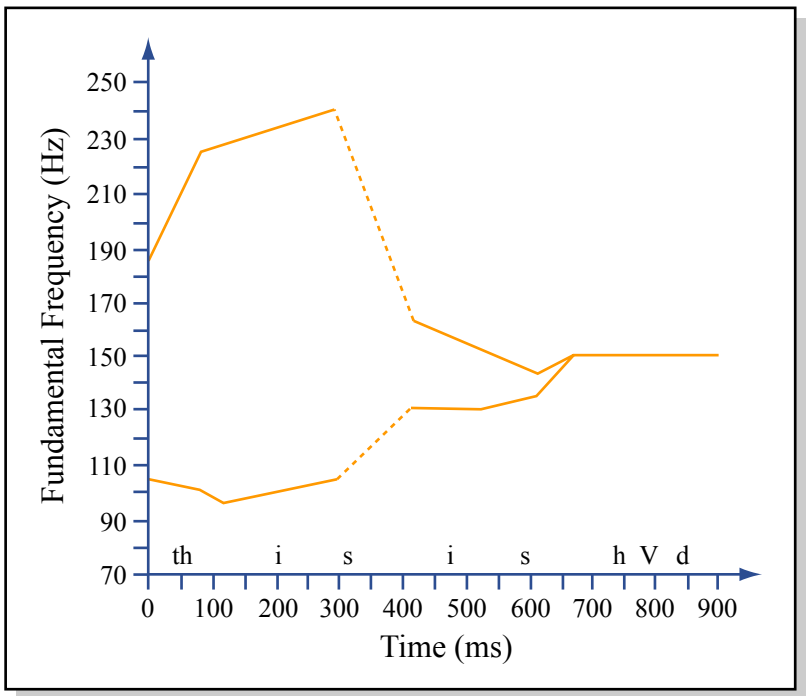


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

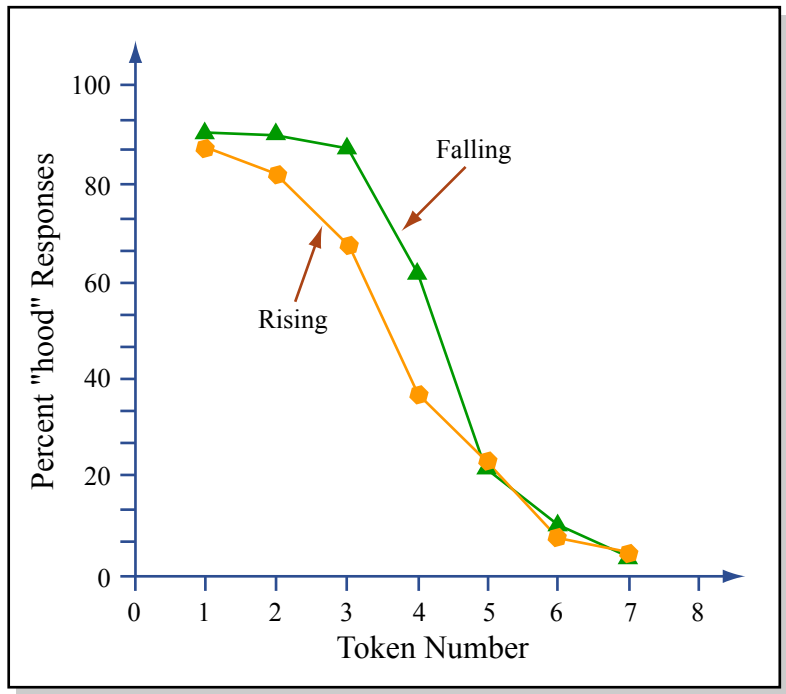


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

Extrinsic information in speaker normalization

- Johnson (1990) - perception of vowels with the same f0 is influenced by f0 of the preceding carrier phrase.
- Shifts in identification functions are better predicted by perceived speaker sex and size than by direct f0.

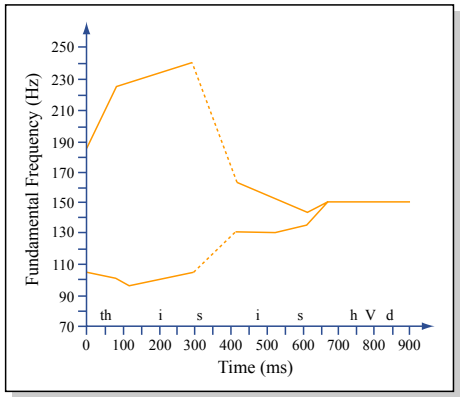


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

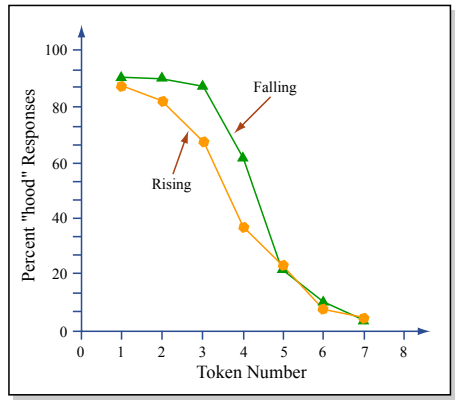


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

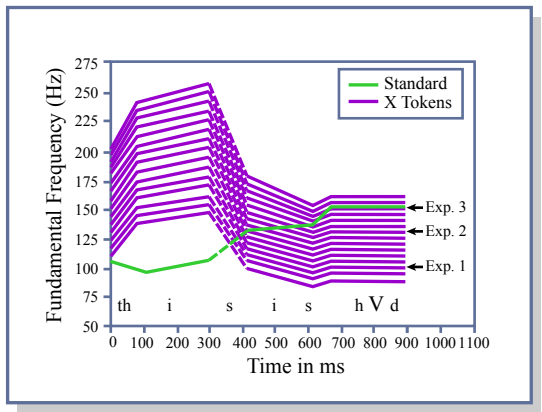


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

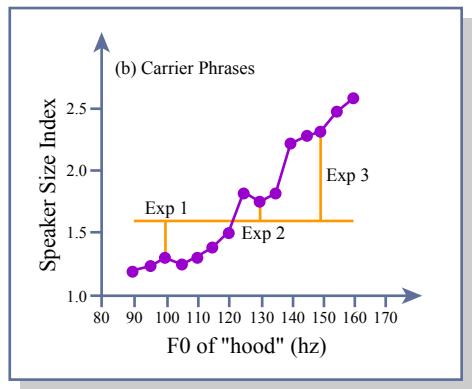


Image by MIT OpenCourseWare. Adapted from Johnson, K. *Journal of the Acoustical Society of America* 88 (1990): 642-655.

Non-speech information in speaker normalization

- Johnson, Strand and d'Imperio (1999) - identification of a hood-HUD continuum is affected by visual information about speaker sex and by instructions specifying speaker sex.
- Johnson's interpretation: listener constructs a representation of the speaker on the basis of available information. This representation provides the basis for expectations concerning the speaker's speech.

- We can also identify speaker characteristics (sex, size, dialect, identity, etc) on the basis of their speech.
- Johnson (1997) and Johnson and Beckman (1996) propose an exemplar-based model of all of these abilities: Labeling exemplars for speaker, sex, dialect etc allows for simultaneous recognition of linguistic content and speaker characteristics.

Cross-dialect speech perception

- Rakerd and Plichta (2003) adapted Ladefoged and Broadbent's experimental method to show that perception of vowels is influenced by dialect information in the preceding context.
- Synthetic [æ, ɑ] continuum (hat-hot, sack-sock)
- Speakers and subjects from Detroit and Michigan Upper Peninsula.
- Detroit accent is characterized by fronting of /ɑ/ and diphthongization of /æ/ (Northern Cities Shift).
- Synthetic words were placed at the end of carrier phrases from Detroit and UP speakers.

Cross-dialect speech perception

- For Detroit listeners identification of continuum shifted as a function of carrier phrase.

Detroit (LM) carrier

UP carrier

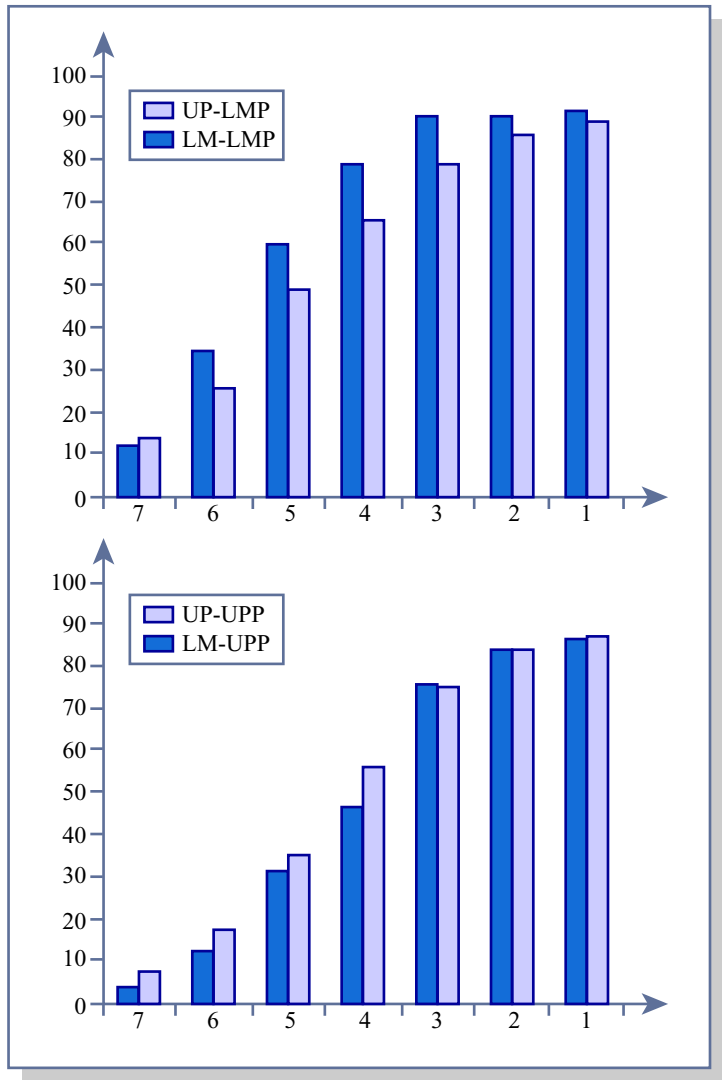


Image by MIT OpenCourseWare. Adapted from Raker, B., and B. Plichta. "More on Perceptions of /a/ Fronting." Paper Presented at the NWAV 32, University of Pennsylvania, 2003.

Register and rate variation

- Minor variations due to rate and register might leave putative invariants intact, but this type of variation can also give rise to substantial changes in pronunciation.
- E.g. palatalization [dɪdju dɪdʒu], [hɪzʃuz, hiʒʃuz]
- Coronal assimilation [foʊnbʊk, foʊmbʊk].
- t,d deletion [bɛnd, bɛn]
- As emphasized by Oshika et al (1975) this variation is still rule-governed. They propose using rules to process this variability by parsing, analysis-by-synthesis or pre-generation of lexical entries.

Evidence for phonological inference

- Evidence that knowledge of casual speech processes is used to infer possible underlying forms in lexical access.
- Gaskell and Marslen-Wilson (1996) priming study of coronal assimilation:
  - Play utterance containing prime word.
  - Present printed word for lexical decision.
  - Repetition priming: lexical decision to visual word is faster when the word has just been heard.
- Primes:
  - Unmodified word: [wɪkɪd]
  - Assimilated word: [wɪkɪb pɹæŋk]
  - Non-assimilatory modification: [wɪkɪb geɪm]
  - Control: unrelated word.
- Assimilated words produce a stronger priming effect than non-assimilated modified words.
- Similar results for ‘legal’ and ‘illegal’ modifications in German (Coenen et al 2001).

Evidence for pre-lexical phonological inference

- It is conceivable that phonological rules are applied after lexical access as a ‘context-checking mechanism’.
- Gaskell and Marslen-Wilson (1998): phoneme monitoring - subjects must indicate when they hear a particular sound.
- Subjects were more likely to report a /t/ in ‘frayp bearer’ than in ‘frayp carrier’.
- Also more likely to report a /t/ in ‘prayp bearer’ than in ‘prayp carrier’ - these are non-words so the effect cannot be a direct consequence of lexical access.
- Can exemplar-based models account for perception of casual speech?
  - Predicts that ability to accommodate casual speech variation should be influenced by experience with the process in the particular word.
  - Can it accommodate casual speech variation in non-words? (Does not allow for rule-based phonological inference).