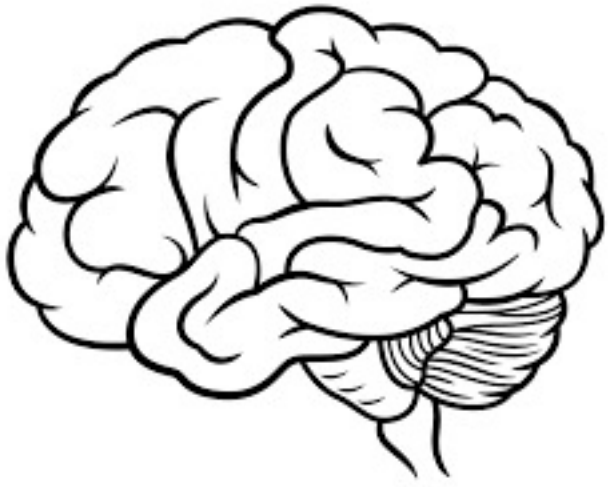
The acoustic dimension of words

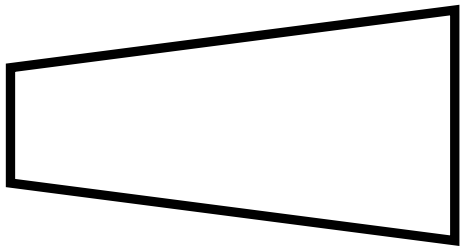
Eva Maria Luef

SS 2025

Phoneme-by-phoneme basis

- Most speech recognition models assume serial phonemic analysis to *assemble the target word (*Cohort Model, Neighborhood Activation Model…)
  - /k/ + /æ/ + /t/
- Perception depends on recognition of each segment and correct order of segments





k + ɜ + r + t + ə + n + z

Speech recognition

/ki/

/krɔl/

/kru/

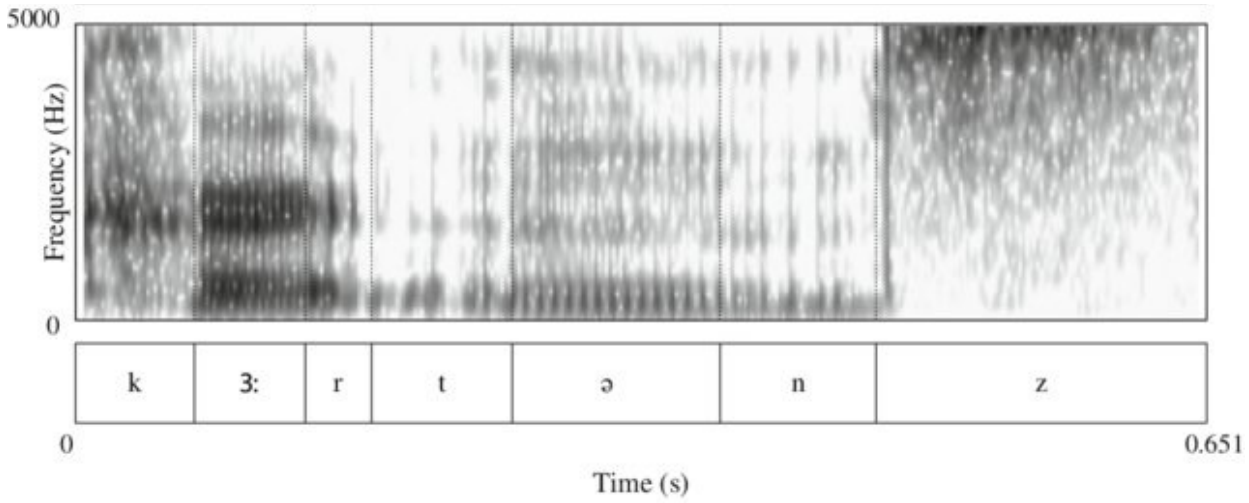
/ˈtruθfəl/

/ˈrisənt/

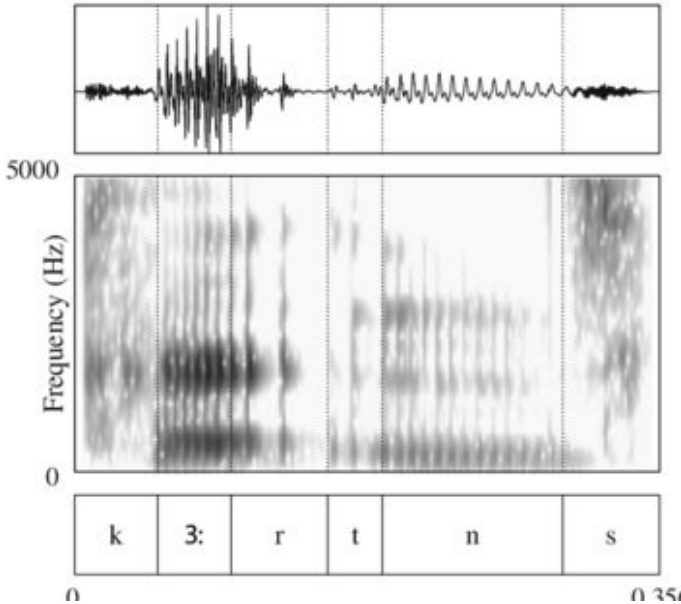
/briz/

Acoustic processes in word recognition

- Actual realization of ‘acoustic word’ can be quite different from its canonical form
- But word recognition is not impaired
  - In first languages
  - But can be in L2
- What are differences in the “curtains” example?



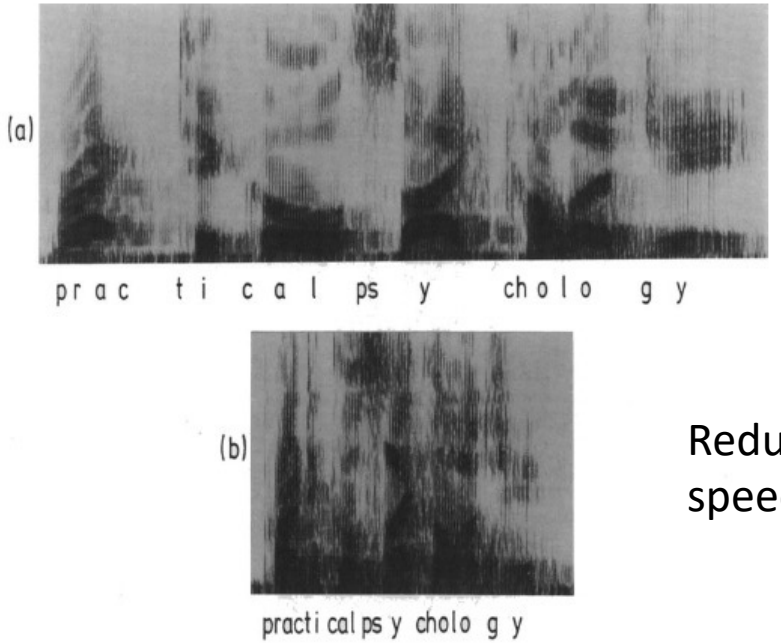
(a)



curtains

Some peculiarities of acoustic words

- Depend on
- Type of speech
  - Laboratory speech, clear speech
    - Read text or single sentences/ words
    - Utter words in isolation (some experiments)
  - Spontaneous, natural speech
    - Completely spontaneous
    - Interviewer-structured



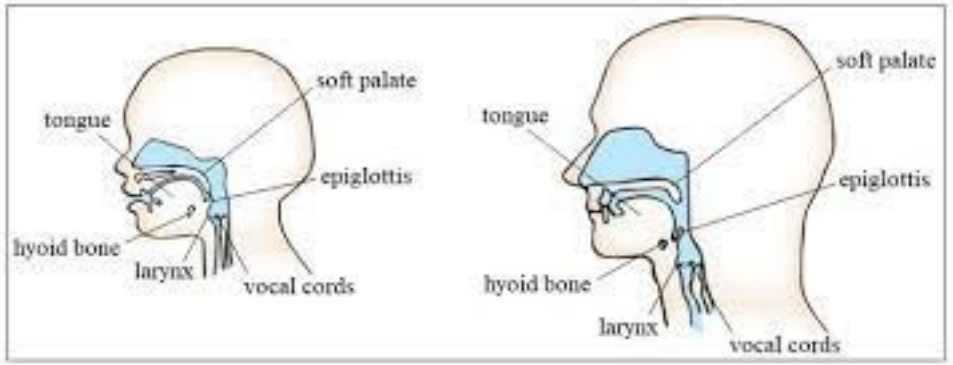
Hyper-articulation

Reduced speech

- *Which type of speech elicits the form closest to the canonical phonetic form?*
- In psycholinguistics: words are seen as strings of phonemes
- In acoustics: no clear delineation between phonemes in a string

Perceptual constancy in the face of acoustic variability

- Ability to understand different speakers
- Size and shape of vocal tract determines many acoustic speech characteristics
  - Children, women, men (of different sizes)
    - Small vocal folds → more widely spaced harmonics
  - Idiosyncratic speech
  - Accents, dialects



- Also: speech is rarely heard in quiet environment
- Challenges to Automatic Speech Recognition

Phonemic restoration effect

- Perceptual illusion effect (Warren, 1970, 1971)
- Hear a phoneme that is not there
  - Brain tries to guess at missing information
- when a phoneme is obscured, listeners rely on their understanding of language structure and meaning to infer what was likely said
- demonstrates that perception is not solely based on raw auditory input but is influenced by contextual cues that inform expectations
- https://www.youtube.com/watch?v=ZyvyGMkzNQc

- Top-down processing, rather than bottom-up
  - Chunks of phonemes
  - rather than individual phonemes

Ganong Effect

- Humans show tendency to perceive an ambiguous speech sound as a phoneme that would complete a real word, rather than completing a nonsense/fake word (Ganong, 1980)
  - E.g., sound heard as either /g/ or /k/ is perceived as
    - /g/ when followed by “ift”
    - but perceived as /k/ when followed by “iss”
  - Or ambiguous sound at the end of “car” is likely to be perceived as /d/
- higher-level activation of lexical representations directly affects sublexical components (e.g., phoneme categories)
- phonetic categories are flexible and perception of even individual speech features depends critically on the surrounding signal (Repp & Liberman, 1987)

Phonetic reduction

- Well-known characteristic trait of casual speech
  - words are sometimes pronounced distinctly, e.g. [ˌfoʊnəˈtɪʃən] phonetician, and sometimes more reduced, e.g. [fənˈtɪʃn̩] or various intermediate forms
- Can affect any sound
  - Commonly involves vowels (“vowel reduction”, or even elision)
    - Schwa: e.g., [ˌfoʊnəˈtɪʃən] → [fənˈtɪʃn̩]
- Causes of reduction:
  - High lexical frequency (“house” as opposed to “hose”)
  - High contextual probability (”he vaccuumed in the house”)
  - Strong collocation (”vacuum cleaner”)
  - Repetition
  - …

Reduced speech

- https://sites.arizona.edu/nwarner/reduced-speech-examples/

Homophonous strings of words

- Most speech production models predict same strings of segments to be produced identically
  - /taɪd/ → tide – tied
  - /friz/ → freeze - frees
- but: morphological information may influence the phonetic properties of words
  - for example acoustic duration/ reduction
- Seyfarth et al. (2017):
  - stems of words ending in [s, z] have longer durations if these are inflected words
  - corresponding strings of segments in mono-morphemic words ending in [s, z] have shorter durations
    - frees vs. freeze, tied vs. tide

Morpheme boundaries

- Word (=free morpheme) boundary [meɪkɔf] – represented by #
  - Make # off
  - May # cough
  - Makoff #
- Bound morpheme boundary – represented by +
  - “odd + ity” [ɑdəti]
  - “post + al” [poʊstəl]
  - “opac + ity” [oʊpæsəti]
  - “acid + ic” [əsɪdɪk]
- Morpheme boundary is invisible to phonology
  - “may name” vs. “main aim”
    - Can be difficult to tell apart
    - Identical phonology

Near-homophones and morphemic boundaries

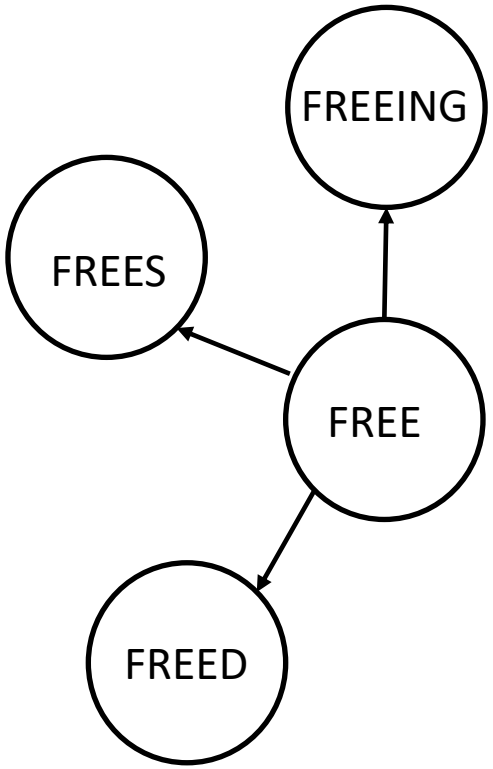
- *freeze – free+s* (”they freeze the meat”, “he frees the whales”)
- *need – knee+d* (“I need something”, “they kneed in church”)

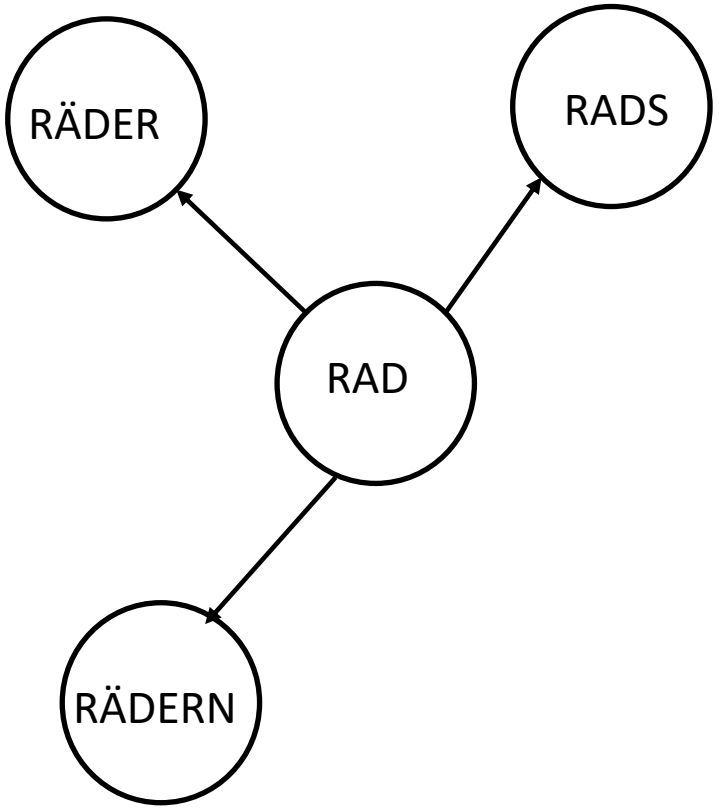
- Stems → freeze, free; need, knee
  - Stems + affixes = (near, apparent) homophones
  - Words = free morphemes, affixes = bound morphemes
  - Morphologically complex “frees” vs. morphologically simple “freeze”
- Word phonetics depends on morphemic status of a word (Seyfarth et al., 2017)
  - Vowels are lengthened before morpheme boundary
    - Vowel in “frees” is longer than in ”freeze”
    - Also see “brewed”/ “brood”, “eyeful”/ “Eiffel” (Bermúdez-Otero, 2006; McMahon, 1991)

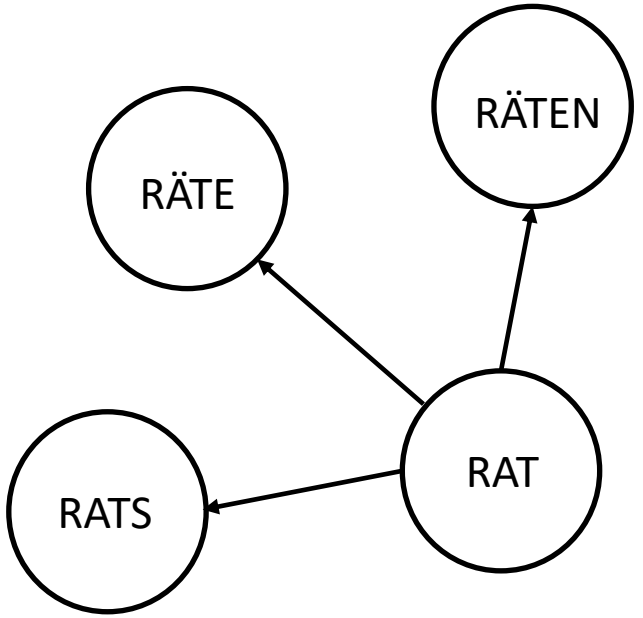
Phonetic paradigm uniformity

- Words/ inflections belonging to the same paradigm show similar phonological/ phonetic patterns
  - **dance** - **danc**ing
  - Morphological families affect production phonology
  - f**ree** – **free**s
    - Homophones: “freeze” vs. “frees”
    - In German: “Rat” and “Rad” are apparent homophones [ʁa:t] (due to word-final devoicing)
    - “Rad” is morphologically related to “Räder”, “Rat” is not
- Spreading activation among wordforms
- Word retrieval does not only activate a target word but also its inflectional neighbors
  - German “Rad” activates “Räder”
  - “Rat” does not activate “Räder”
    - Fine-grained phonetic differences between “Rat” and “Rad” (Roettger et al., 2014)

Activation spreading: the morphological family







Paradigm uniformity

- effects arise from morphological paradigm uniformity
  - morphological paradigm = set of words that have common lemma
  - stem plus all inflections (*free, frees, freed, freeing*…)

- the stem of an inflected word like ”*frees”* is influenced by its morphologically simple paradigm member “*free*”
  - “free” is an open syllable at the end of prosodic boundary → lengthening effect on [i]
    - this acoustic features is transferred to all inflections, e.g., “frees”, “freeing”

Near-homophones

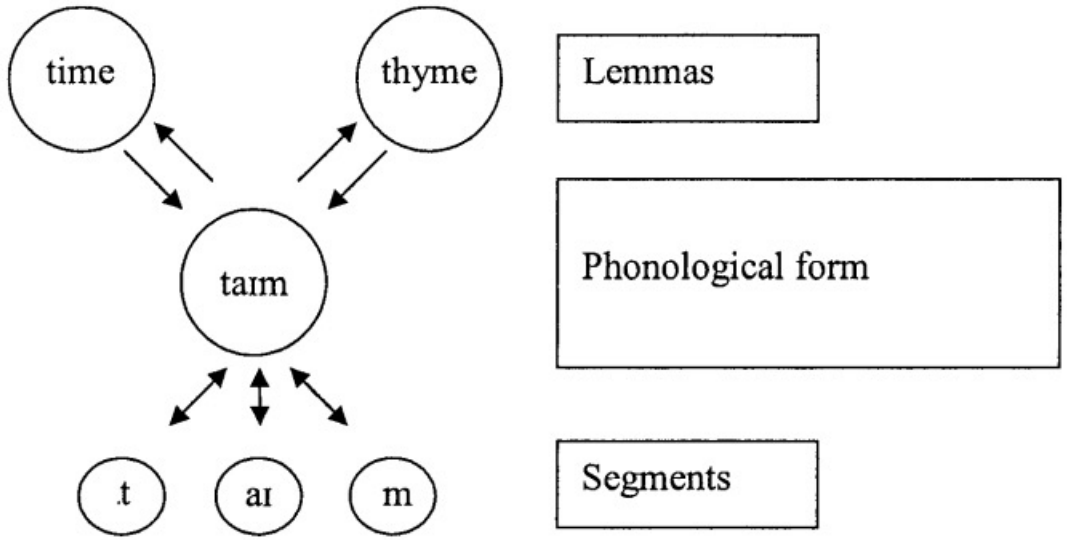
- the stem of an inflected word like ”*frees”* is influenced by its morphologically simple paradigm member “*free*”
- thus, “freeze” and “frees” are not homophonous
  - even thought they represent same string of sounds /friz/
    - same for *tax vs. tacks, lapse vs. laps, duct vs. ducked…*

“I’m sorry, I hate to be the bear of bad news.”

Frequency effects in homophony

- Gahl (2008)
  - Homophones show subtle acoustic variation caused by frequency effects
    - i.e., phonetic reduction in high-frequency words
      - time vs. thyme

  - Lemma = semantic, syntactic, morphological properties
  - Phonological form



Frequency inheritance

→ "thyme" inherits its frequency from "time"

→ sum of all frequencies of a wordform (Dell, 1990)

→ no inheritance of frequency in English homophones (Gahl, 2008)

Thus different phonetic realizations

Frequency effects in speech errors

- low-frequency words are more vulnerable to speech errors than high-frequency words (Dell, 1990)
- low-frequency words with high-frequency twins are less vulnerable to errors than low-frequency words without such twins
  - Low-frequency “thyme” is less prone to speech errors because of its high-frequency homophone/twin “time”
  - Compare to: hose-hoes, shoe-shoo

- high-frequency homophones protect against speech errors

Co-articulation

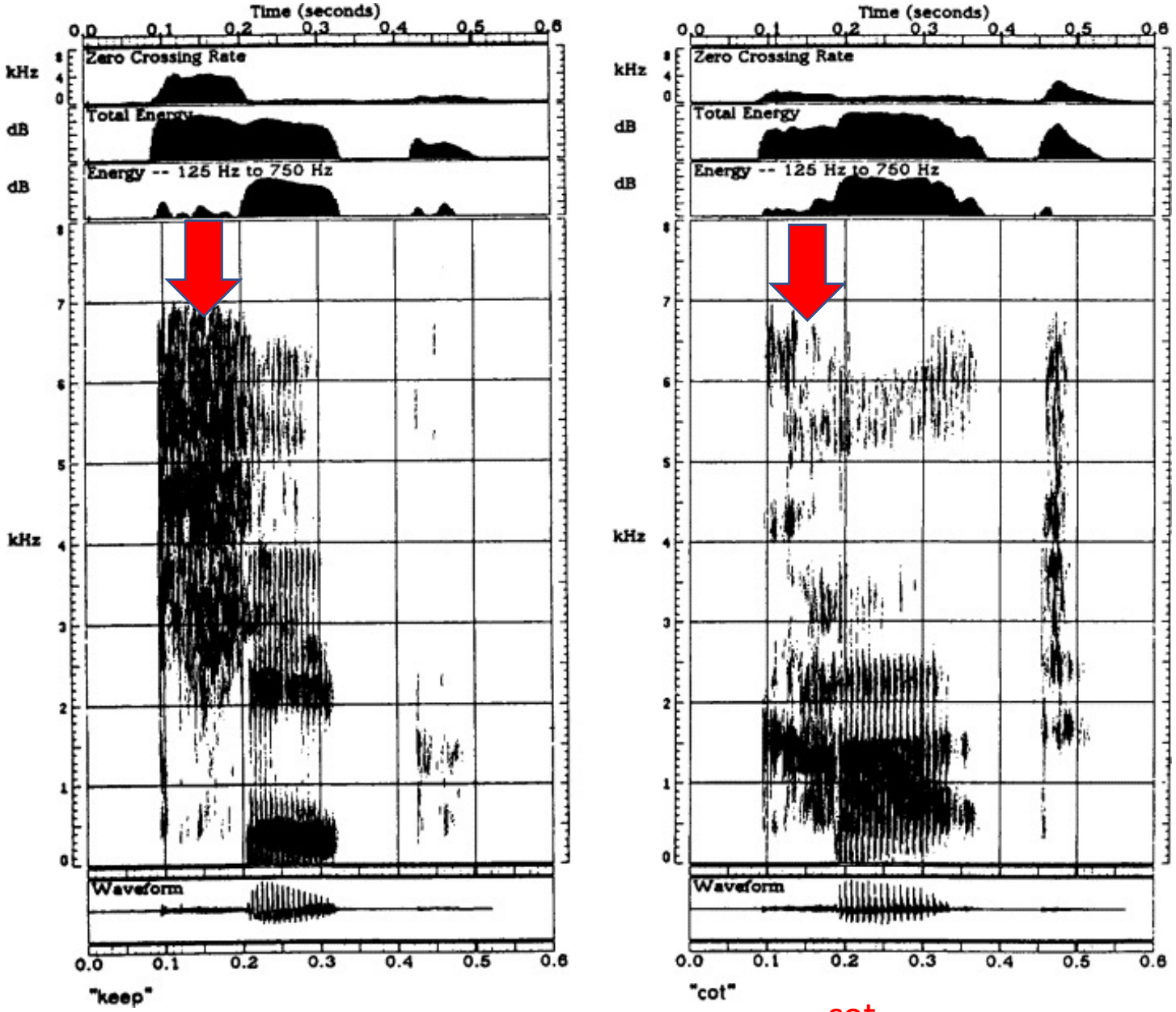
- acoustic analysis reveals that speech does not consist simply of a sequence of sound segments corresponding to phonemes, syllables or words
- often no clear acoustic boundaries exist between segments
  - the phonetic features change relatively smoothly over time
- occurrence of phonetic/ articulatory features does not coincide consistently with boundaries between phonemes, syllables or words
  - When you say “tulip” – your lips are rounded at the /t/
    - each sound planned only after the preceding sound was produced, → rounding would only occur after “t” was uttered
    - “Anticipatory lip rounding”

Co-articulatory effects

- smooth changes between successive segments have an important consequence:
  - Carry-over effects of acoustic-phonetic features
- the acoustic form of a particular speech sound will depend upon the context in which it occurs
  - Transfer to previous segment → lip rounded /t/ in “tulip”
  - Transfer to following segment → frication on /r/ in “try”
- this contextual influence is known as coarticulation
  - inevitable consequence of the way in which vocal tract articulators move to avoid successional discontinuities

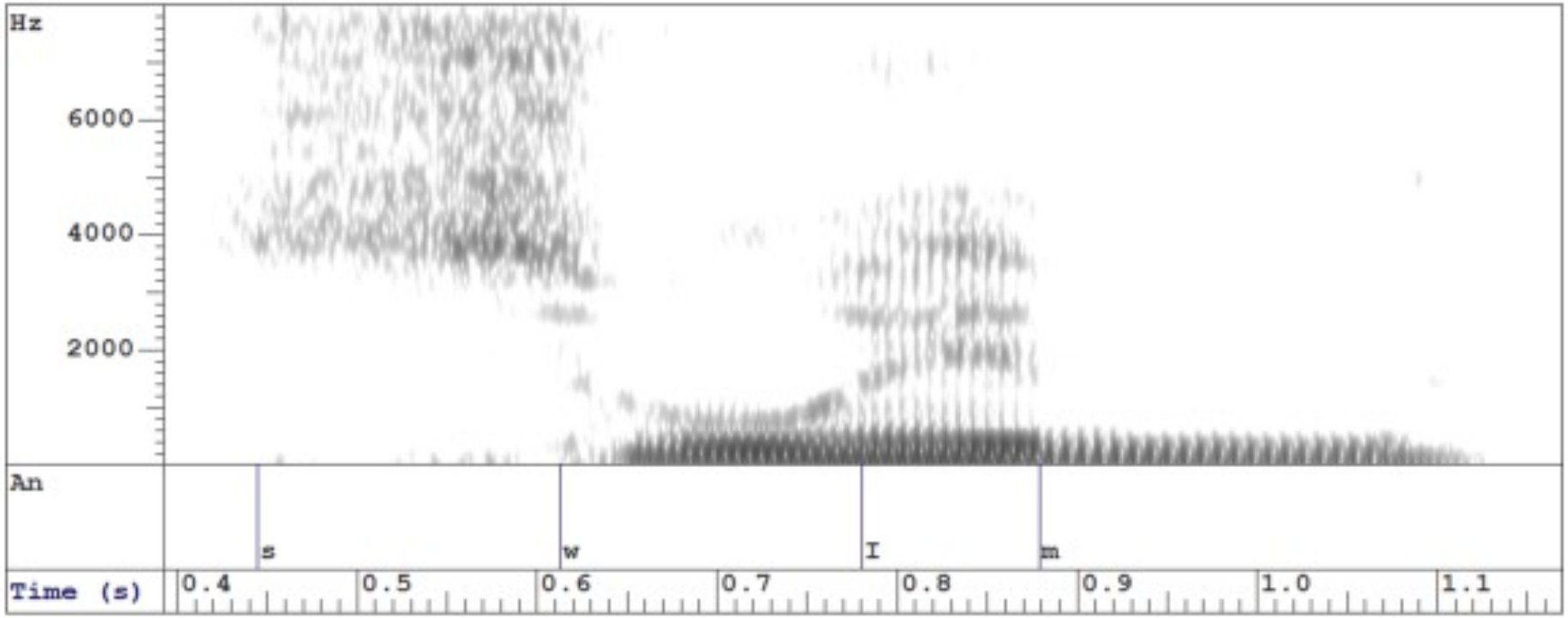
Co-articulation

- reflects a crucial equilibrium between speaker efficiency and listener comprehension
  - in language development, appropriate co-articulatory overlap indicates mature, adult-like speech
  - Also in L2 (Jang et al., 2022)
- In difficult communication (degraded speech, background noise…) speakers tend to co-articulate less



keep cot

”swim”



rounding

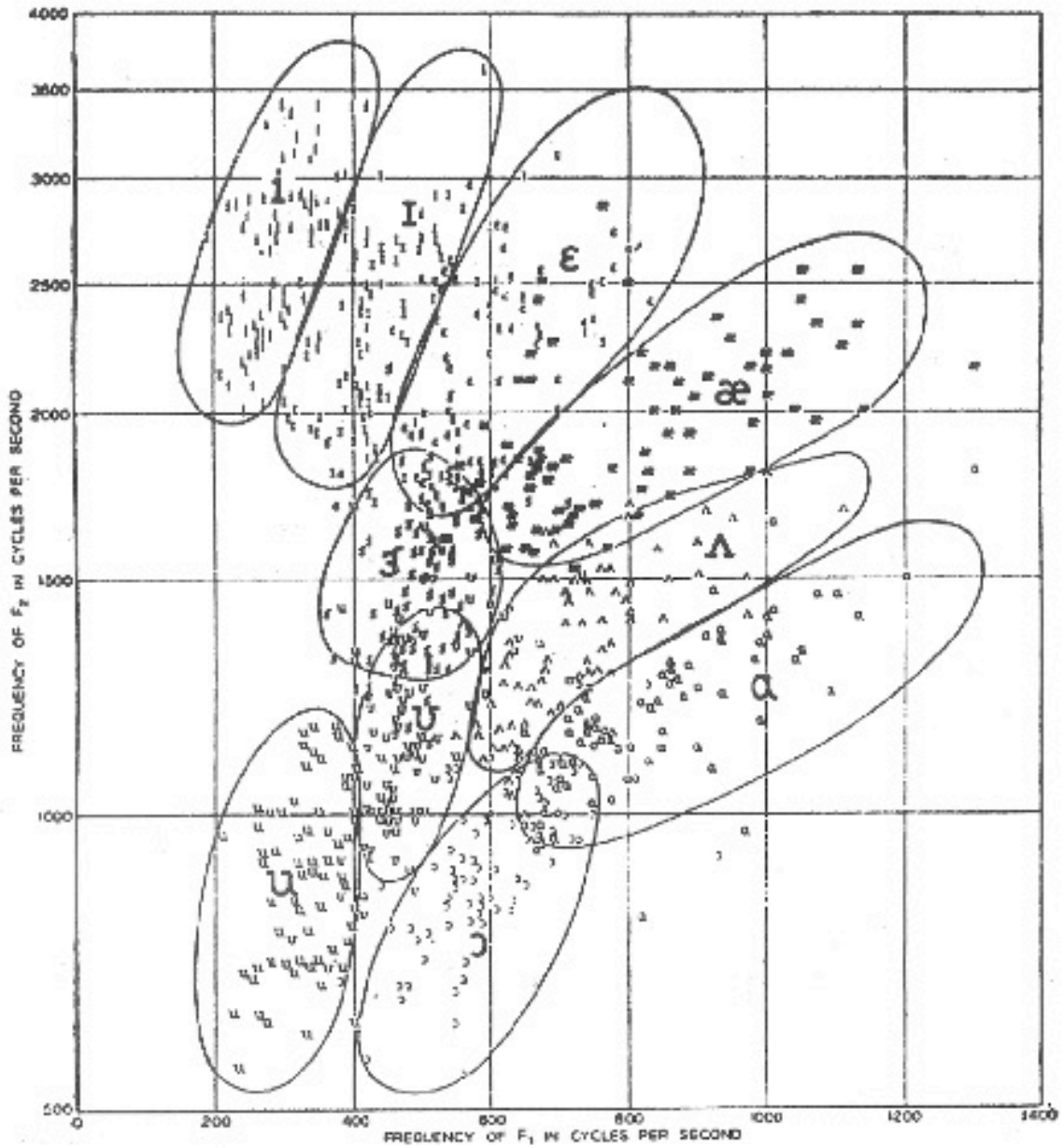
rounding

devoicing

nasality

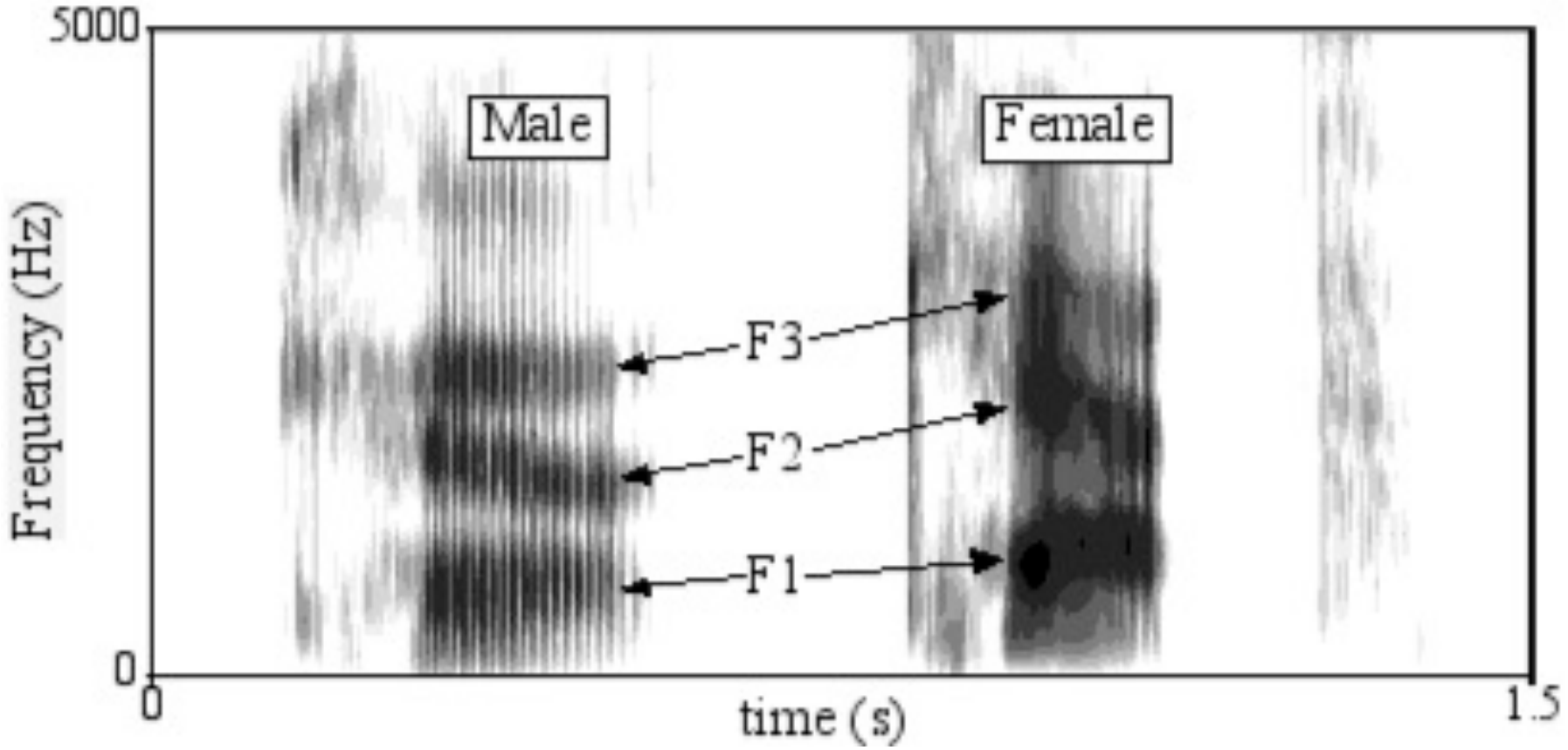
Vowel perception

- Extensive variation in speech
- How come that listeners agree in their perception of vowels?
- Which information is utilized?



Scatter plot of first and second formant values of American English vowels (Peterson & Barney, 1952)

Vowels are relative patterns, not absolute frequencies



Spectrogram of a man and a woman saying “cat”.
The three lowest vowel formants are marked as F1, F2 and F3) (Johnson, 2004)

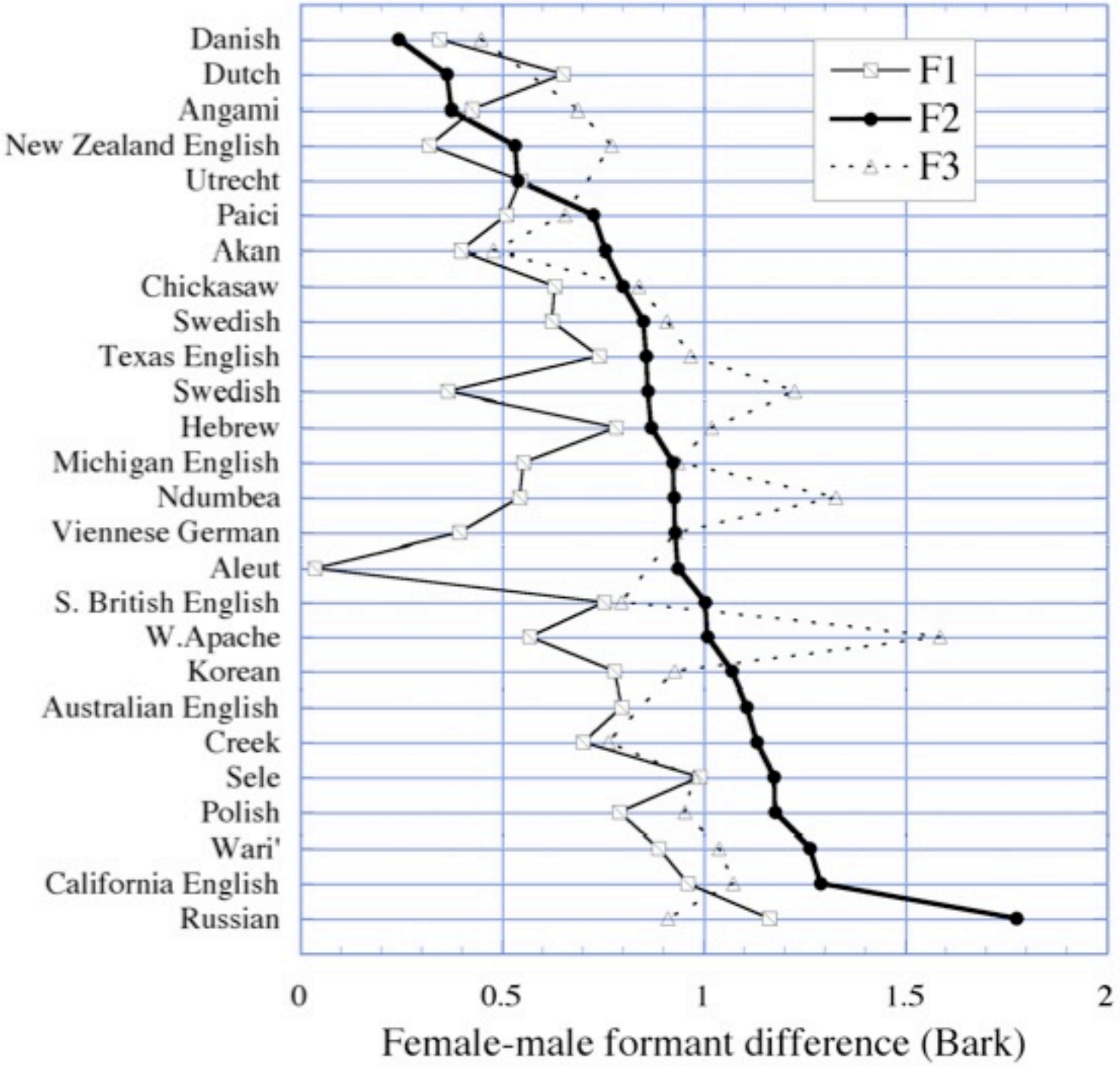
Gender difference

Differences between men and women vary from language to language

→ Cultural factors are involved in defining and shaping male or female speech

→ Anatomy does not completely determine the vowel formant frequencies

(Bladon, Henton and Pickering, 1984)



Normalization

Talker normalization is an active process:

Listener adaptation to talker voice (Kato & Kakehi, 1988)

→ Increase in recognition accuracy over the course of 5 stimuli presented in noise

"In this approach, cognitive categories are represented as collections of the stored cognitive representations of experienced instances of the category,

rather than as normalized abstract representations from which category-internal structure has been removed"

(Johnson 2004)

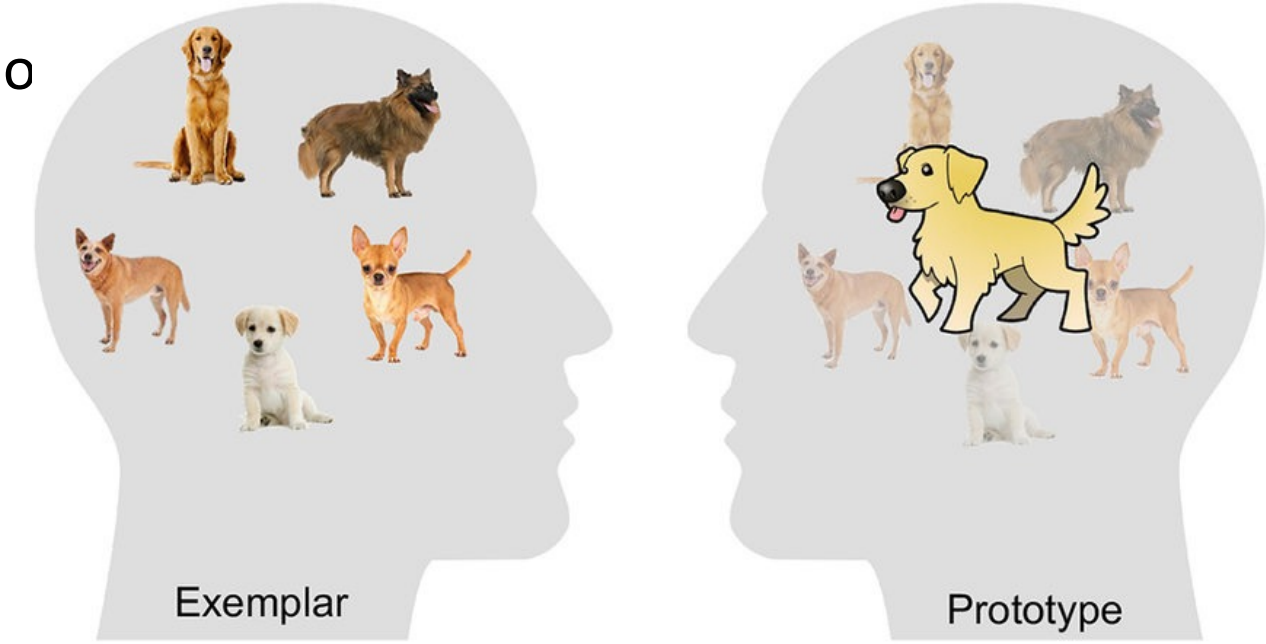
Exemplar theory

- Sea change in phonological and phonetic theory
- Strict separation between lexicon and sublexical (phonemic) knowledge abandoned
  - Usage-based phonetic lexicon
  - Memory integrates phonetics, pho

Exemplar theory = family of theories in cognitive psychology

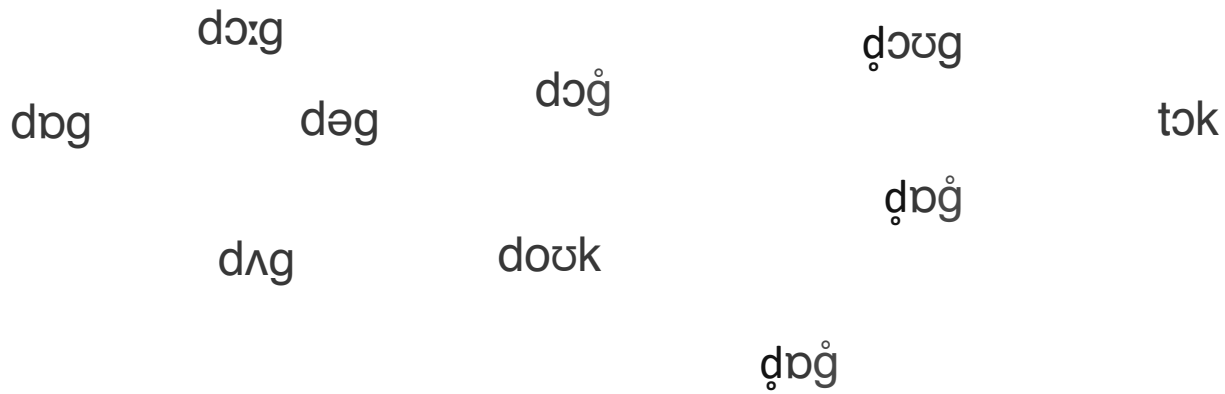
→ We store each encounter of a concept as an exemplar of a larger cognitive category

→ E.g., “dog”



Phonetic exemplars

- We store each encounter of a pronunciation variant
- *Linguistic knowledge does not consist of abstract generalizations but rather of a large number of specific remembered linguistic experiences (“exemplars”)*
- Phonetic spaces of “dog”
  - ”clouds of exemplars” associated with each word
    - High-frequency words have more exemplars
  - Perception = Incoming word is matched to exemplar

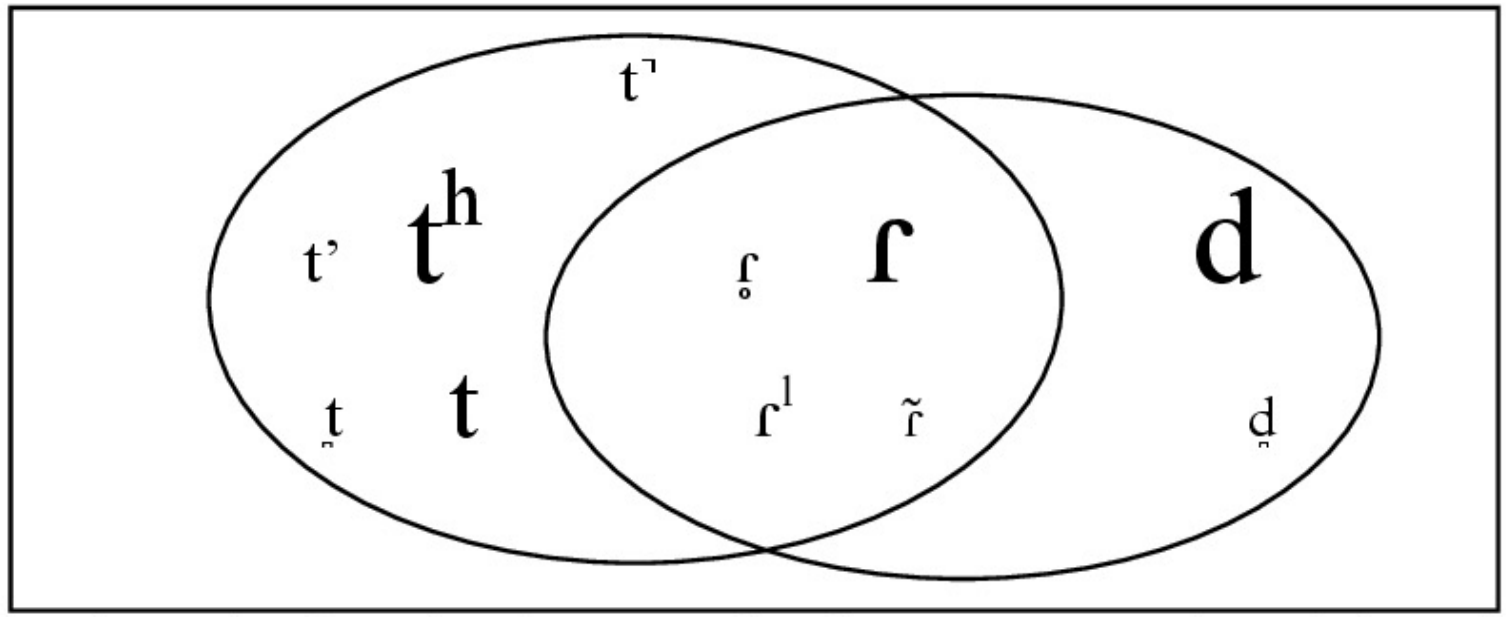


- Different voices
  - Age (frequency)
  - Creak…
- Different accents

- Highly individualistic

→ depending on personal experience

Overlapping exemplar clouds



Apical alveolar consonants in English

References

Bladon, R.A., Henton, C. G. & Pickering, J. B. (1984) Towards an auditory theory of speaker normalization. *Language Communication 4, 59-69.*

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6(1), 110–125.

Hillenbrand, J. M. & Neary, T. M. (1999) Identification of synthesized /hVd/ utterances: Effects of formant contour. *J. Acoust. Soc. Am. 105, 3509-3523.*

Ladefoged, P. & Broadbent, D. E. (1957) Information conveyed by vowels. *J. Acoust. Soc. Am. 29, 98-104*

Leather, J. (1983) Speaker normalization in the perception of lexical tone. *Journal of Phonetics 11, 373-382*

Johnson, K., Strand, E. A. & D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics 27, 359-384*

Johnson, K. (2004) Speaker normalization in speech perception. *Ohio State University*

Johnson, K. (1990) The role of percieved speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am. 88 642-654*

Kato, K & Kakehi, K. (1988) Listener adaptability to individual speaker differences in monosyllabic speech perception. *J. Acoust. Soc. Of Japan 44, 180-186*

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. Science, 167, 392-393.

Warren, R. M., & Obusek, C. (1971). Speech perception and phonemic restorations. Perception & Psychophysics, 9, 358-363.