Human Speech Recognition

Julia Hirschberg

CS4706

(thanks to Francis Ganong and John-Paul Hosum for some slides)

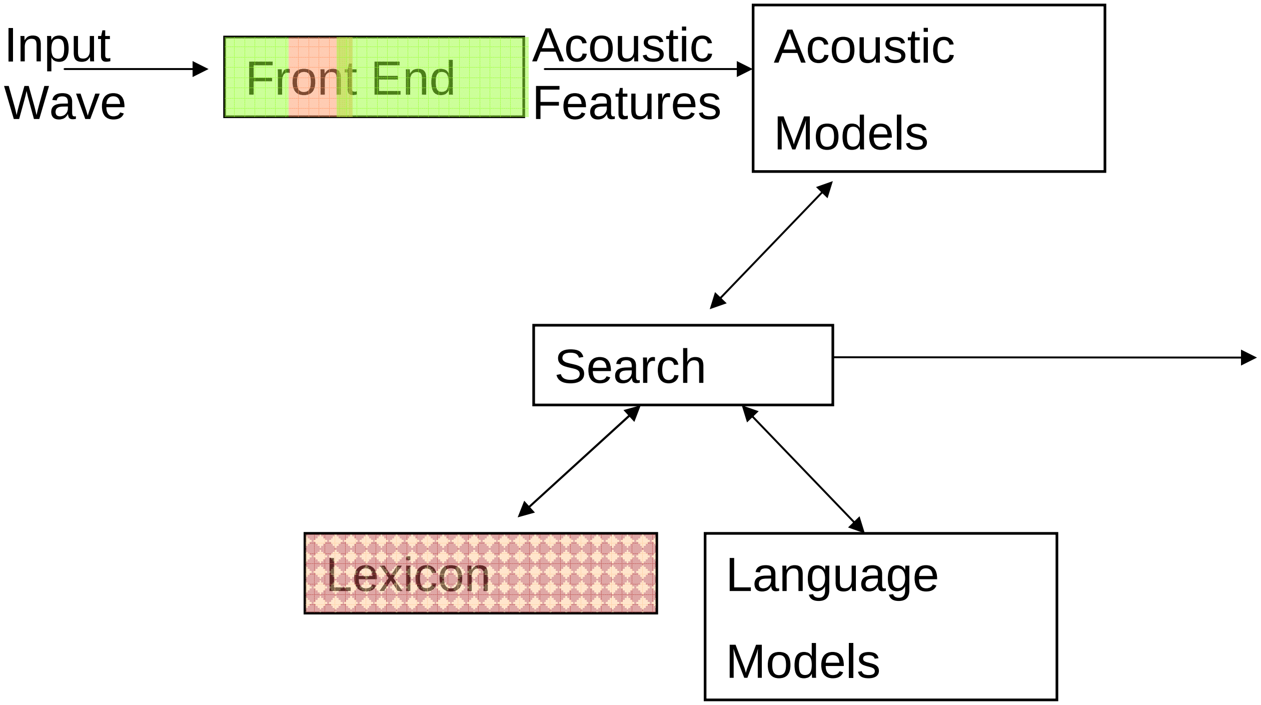
Linguistic View of Speech Perception

- Speech is a sequence of articulatory gestures
  - Many parallel levels of description
    - Phonetic, Phonologic
    - Prosodic
    - Lexical
    - Syntactic, Semantic, Pragmatic
- Human listeners make use of all these levels in speech perception
  - Multiple cues and strategies used in different contexts

ASR Paradigm

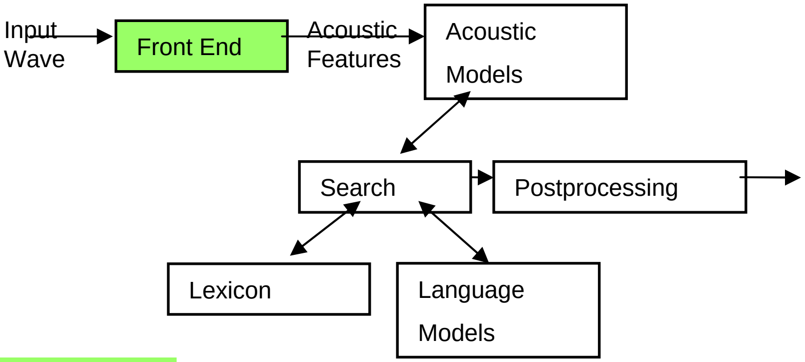
- Given an acoustic observation:
  - What is the most likely sequence of words to explain the input?
    - Using
      - Acoustic Model
      - Language Model
- Two problems:
  - How to score hypotheses (Modeling)
  - How to pick hypotheses to score (Search)

So….What’s Human about State-of-the-Art ASR?



N1

Front End: MFCC

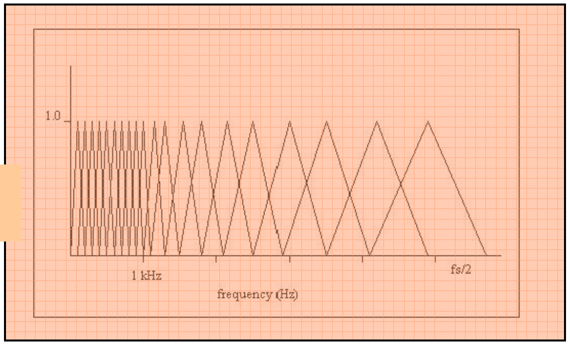


Input
Wave

Sampling, Windowing

FastFourierTransform

Mel Filter Bank:



cosine transform
first 8-12 coefficients

Stacking, computation of
deltas:Normalizations: filtering, etc

Linear Transformations:dimensionality
reduction

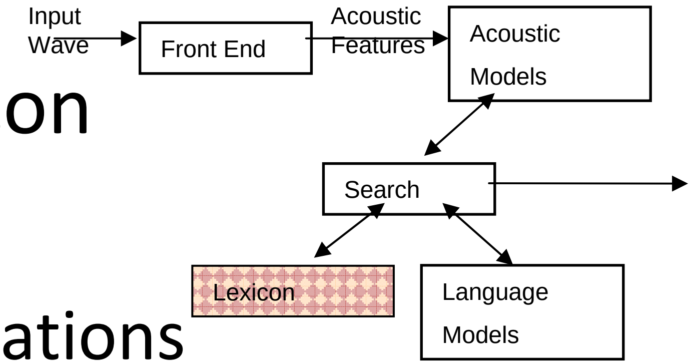
Acoustic
Features

Slide 5

N1 change color of 2nd box to pink; first 1/3 only

Nuance, 3/7/2010

Basic Lexicon



- A list of spellings and pronunciations
    - Canonical pronunciations
    - And a few others
    - Limited to 64k entries
  - Support simple stems and suffixes
- Linguistically naïve
  - No phonological rewrites
  - Doesn’t support all languages

Lexical Access

- Frequency sensitive, like ASR
  - We access high-frequency words faster and more accurately – with less information – than low frequency
- Access in parallel, like ASR
  - We access multiple hypotheses simultaneously
- Based on multiple cues

How Does Human Perception Differ from ASR?

- Could ASR systems benefit by modeling any of these differences?

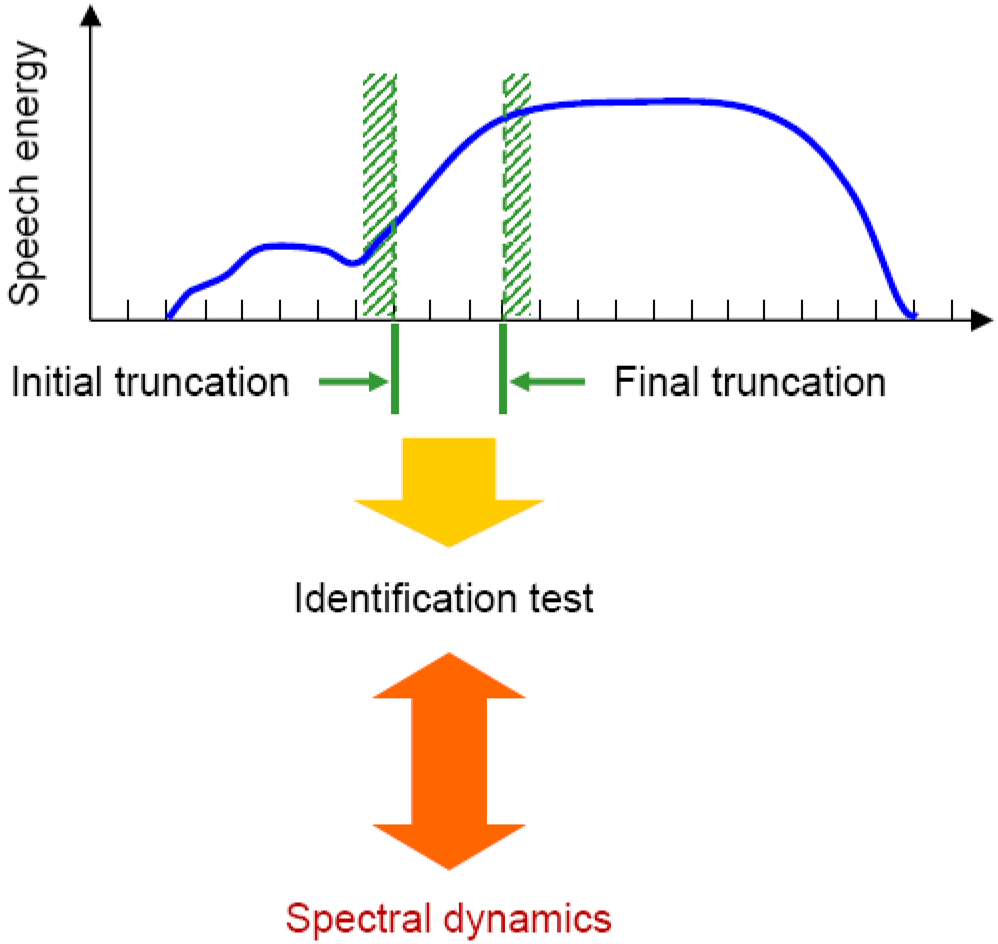
How Do Humans Identify Speech Sounds?

- Perceptual Critical Point
- Perceptual Compensation Model
- Phoneme Restoration Effect
- Perceptual Confusability
- Non-Auditory Cues
- Cultural Dependence
- Categorical vs. Continuous

How Much Information Do We Need to Identify Phones?

- Furui (1986) truncated CV syllables from the beginning, the end, or both and measured human perception of truncated syllables
- Identified “perceptual critical point” as truncation position where there was 80% correct recognition
- Findings:
  - 10 msec during point of greatest spectral transition is most critical for CV identification
  - Crucial information for C and V is in this region
  - C can be mainly perceived by spectral transition into following V

Analysis of relationships between spectral dynamics and syllable perception



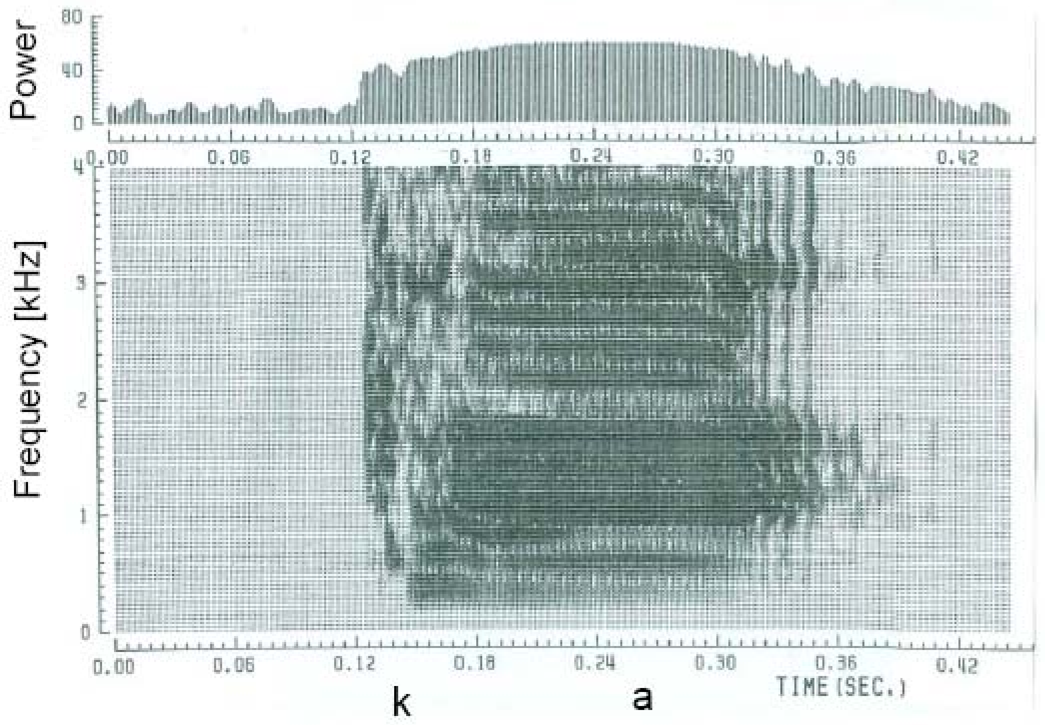
Initial truncation

Final truncation

Identification test

Spectral dynamics

Role of spectral transition for speech perception



Periods that can be truncated

↑

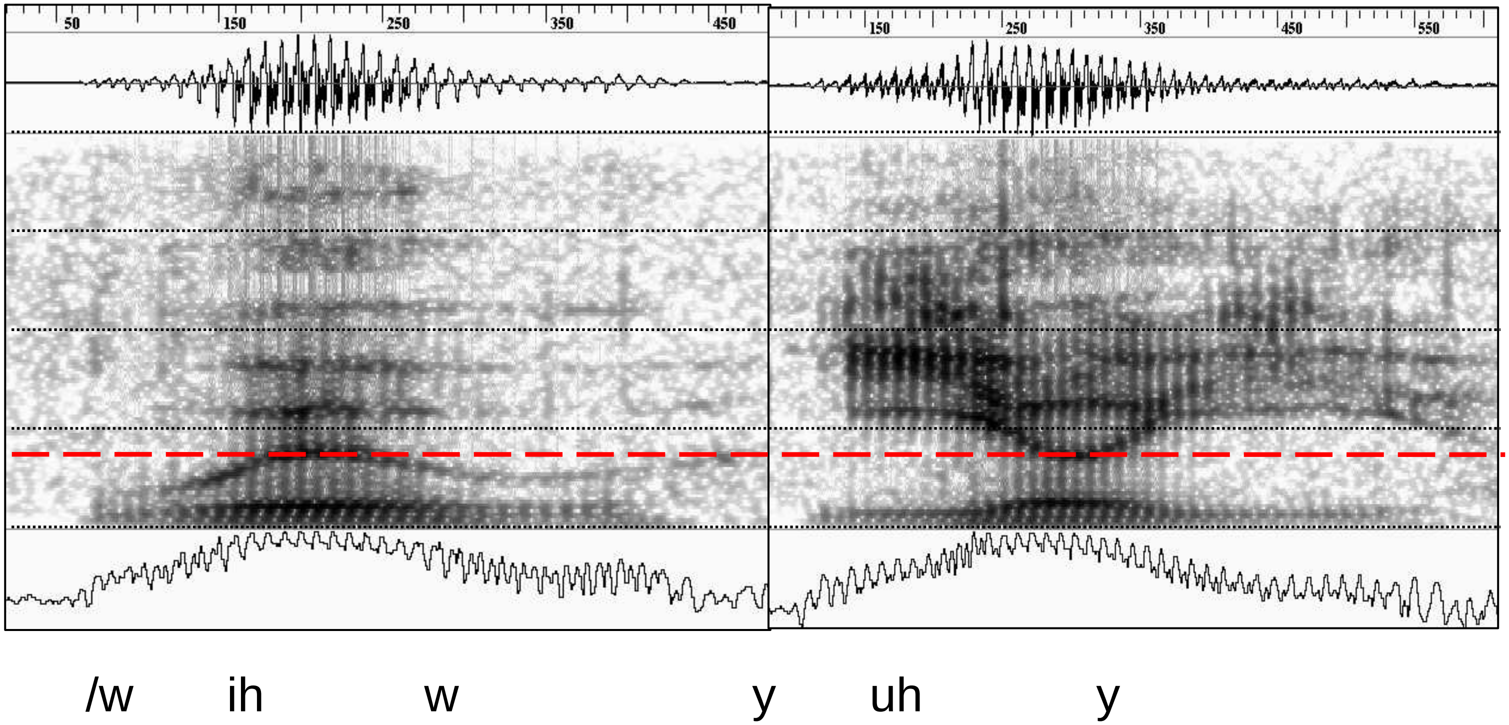
Maximum spectral change period: essential for syllable perception

Can this help ASR?

Target Undershoot

- Vowels may or may not reach their ‘target’ formant due to coarticulation
  - Amount of undershoot depends on syllable duration, speaking style,…
  - How do people compensate in recognition?
- Lindblom & Studdert-Kennedy (1967)
  - Synthetic stimuli in wVw and yVy contexts with V F2 varying from high (/ih/) to low (/uh/) and with different transition slopes from consonant to vowel
  - Subjects asked to judge /ih/ or /uh/

- Boundary for perception of /ih/ and /uh/ (given the varying F2 values) different in the wVw context and yVy context
- In yVy contexts, mid-level values of F2 were heard as /uh/, and in wVw contexts, mid-level values of F2 heard as /ih/



Perceptual Compensation Model

- Conclusion: subjects relying on direction and slope of formant transitions to classify vowels
- Lindblom’s PCM: “normalize” formant frequencies based on formants of the surrounding consonants, canonical vowel targets, syllable duration
- Application to ASR?
    - Determining locations of consonants and vowels is non-trivial

Can this help ASR?

Phoneme Restoration Effect

- Warren 1970 presented subjects with
  - “The state governors met with their respective legislatures convening in the capital city.”
  - Replaced [s] in legislatures with a cough
  - Task: find any missing sounds
  - Result: 19/20 reported no missing sounds (1 thought another sound was missing)
- Conclusion: much speech processing is top-down rather than bottom-up

Perceptual Confusability Studies

- Hypothesis: Confusable consonants are confusable in production because they are perceptually similar
  - E.g. [dh/z/d] and [th/f/v]
  - Experiment:
    - Embed syllables beginning with targets in noise
    - Ask listeners to identify
    - Look at confusion matrix

Is there confusion between voiced and voiceless sounds?

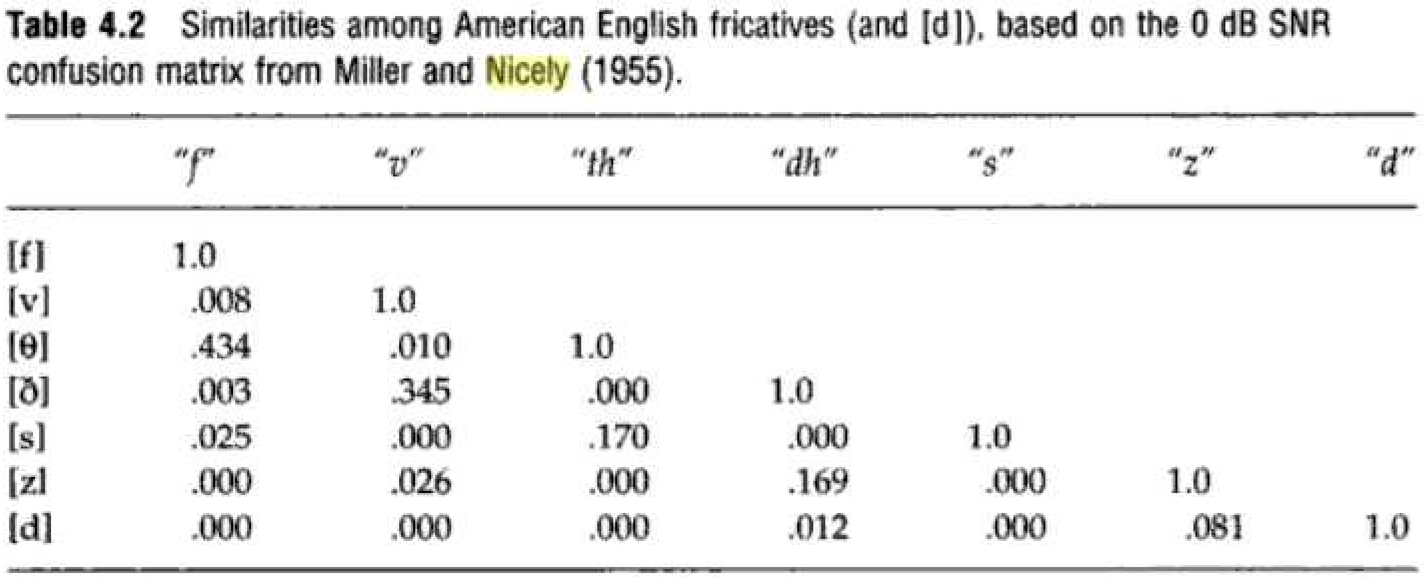


Table 4.2 Similarities among American English fricatives (and [d]), based on the 0 dB SNR confusion matrix from Miller and Nicely (1955).

	“f”	*“v”*	*“th”*	*“dh”*	*“s”*	*“z”*	*“d”*
[f]	1.0						
[v]	.008	1.0					
[θ]	.434	.010	1.0				
[ð]	.003	.345	.000	1.0			
[s]	.025	.000	.170	.000	1.0		
[z]	.000	.026	.000	.169	.000	1.0	
[d]	.000	.000	.000	.012	.000	.081	1.0

- Shepard’s similarity metric

$$S_{ij} = \frac{P_{ij} + P_{ji}}{P_{ii} + P_{jj}}$$

Can this help ASR?

Speech and Visual Information

- How does visual observation of articulation affect speech perception?
- McGurk Effect (McGurk & McDonald 1976)
  - – Subjects heard simple syllables while watching video of speakers producing phonetically different syllables (demo)
  - – E.g. hear [ba] while watching [ga]
  - – What do they perceive?
  - – Conclusion: Humans have a ***perceptual*** map of place of articulation – different from auditory

Can this help ASR?

Speech/Somatosensory Connection

- Ito et al 2008 show that stretching mouth can influence speech perception
  - Subjects heard head, had, or something on a continuum in between
  - Robotic device stretches mouth up, down, or backward
  - Upward stretch leads to ‘head’ judgments and downward to ‘had’ but only when timing of stretch imitates production of vowel
- What does this mean about our perceptual maps?

Can this help ASR?

Is Speech Perception Culture-Dependent?

- Mandarin tones
  - High, falling, rising, dipping (usually not fully realized)
  - Tone Sandhi: dipping, dipping → rising, dipping
    - Why?
      - Easier to say
      - Dipping and rising tones perceptually similar so high is appropriate substitute
- Comparison of native and non-native speakers tone perception (Huang 2001)

- Determine perceptual maps of Mandarin and American English subjects
  - Discrimination task, measuring reaction time
    - Two syllables compared, differing only in tone
    - Task: same or different?
    - Averaged reaction times for correct ‘different’ answers
    - Distance is 1/rt

Mandarin					American			
	High [55]	Rising [35]	Dipping [214]	Falling [51]	High [55]	Rising [35]	Dipping [214]	Falling [51]
High [55]								
Rising [35]	563				615			
Dipping [214]	579	**683**			536	**706**		
Falling [51	588	548	545		600	592	608	

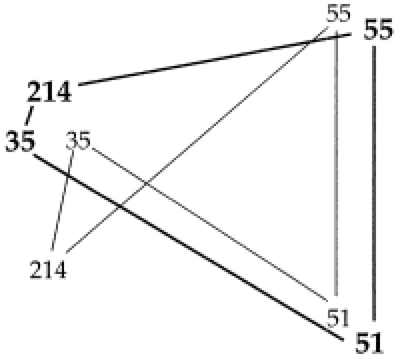


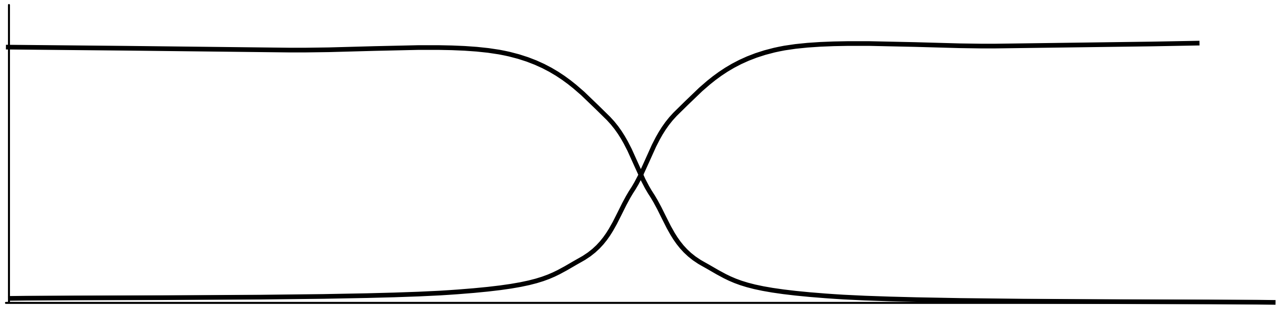
Figure 4.7 Maps of the perceptual tone spaces for Mandarin Chinese listeners (heavy lines, large font) and American English listeners (light lines, small font). These maps were produced by linear multi-dimensional scaling of the data in table 4.5. The MDS dimensions were scaled by the regression coefficent relating actual and predicted distance so that distance in the MDS solution is proportional to distance in the raw data. The lines drawn in this figure show some selected perceptual distances – e.g. the distance between tone [55] and tone [214] is shown as a line for both the Mandarin and American English listeners, while the [55]–[35] distance is not directly drawn in for either group.

Can this help ASR?

Is Human Speech Perception Categorical or Continuous?

- Do we hear discrete symbols, or a continuum of sounds?
- What evidence should we look for?
    - Categorical: There will be a range of stimuli that yield no perceptual difference, a boundary where perception changes, and another range showing no perceptual difference, e.g.
        - Voice-onset time (VOT)
            - If VOT long, people hear unvoiced plosives
            - If VOT short, people hear voiced plosives
            - But people don’t hear ambiguous plosives at the boundary between short and long (30 msec).

- Non-categorical, sort of
  - Barclay 1972 presented subjects with a range of stimuli between /b/, /d/, and /g/
  - Asked to respond only with /b/ or /g/.
  - If perception were completely categorical, responses for /d/ stimuli should have been random, but they were systematic
  - Perception may be continuous but have sharp category boundaries, e.g.



Can this help ASR?

Where is ASR Going Today?

- 3->5
  - Triphones -> Quinphones
  - Trigrams -> Pentagrams
- Bigger acoustic models
  - More parameters
  - More mixtures
- Bigger lexicons
  - 65k -> 256k

- Bigger language models
  - More data, more parameters
- Bigger acoustic models
  - More sharing
- Bigger language models
  - Better back-offs
- More kinds of adaptation
  - Feature space adaptation
- Discriminative training instead of MLE to penalize error-producing parameter settings
- Rover: combinations of recognizers
- Finite State Machine architecture to flatten knowledge into uniform structure

But not…

- Perceptual Linear Prediction: modify cepstral coefficients by psychophysical findings
- Use of articulatory constraints
- Modeling features instead of specific phonemes
- Neural Nets, SVM / Kernel methods, Example-Based Recognition, Segmental Models (frames->segments), Graphical Models (merge graph theory/probability theory)
- Parsing

No Data Like More Data Still Winning

- Standard statistical problems
  - Curse of dimensionality, long tails
  - Desirability of priors
- Quite sophisticated statistical models
  - Advances due to increased size and sophistication of models
- Like Moore’s law: no breakthroughs, dozens of small incremental advances
- Tiny impact of linguistic theory/experiments

Next Class

- Newer tasks for recognition