Tutorial:
Text-to-Speech Synthesis

Marian Macchi
mjm@bellcore.com
Bellcore
Morristown, NJ USA

ACL/EACL Workshop on Spoken Dialog Systems
Madrid, Spain
July 11-12, 1997

Traditional Text-to-Speech

- Monolog
  - reads news articles, email, books, telephone directories
  - no interaction with listener
- Reads from plain text
  - arbitrary text; can say anything
  - but can only infer intention behind text

Human Speech Production Levels

World Knowledge

Semantic (Answer to *Who read to Dr. Smith?*)

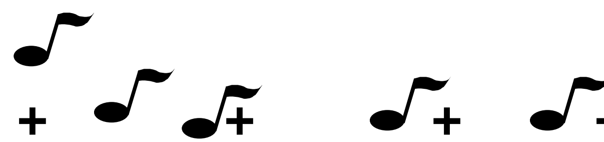
Syntactic

S

NP VP

Word [the][children] [read][to][doctor][smith]

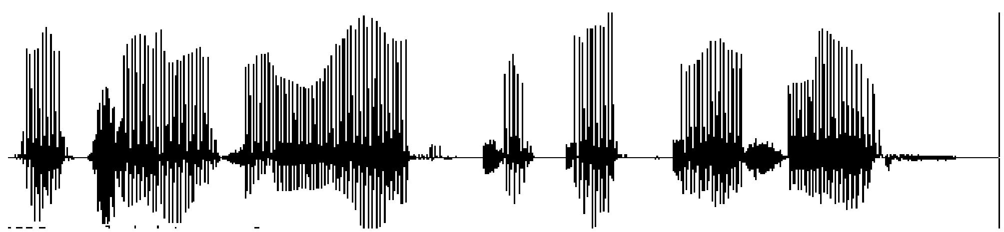
[D][N] [Vpast][P] [H][N]



+ + + +

Phonological %Dx CIldrxn rEd tx daktR smIT %

Acoustic



More Levels

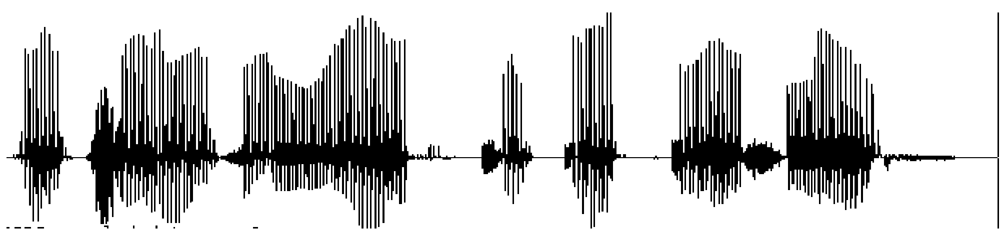
Phonological %Dx CIldrxn rEd tx daktR smIT %

Motor commands

Articulator
movements

Formants,
F0, Amplitude,
Duration

Acoustic



TTS Levels

Orthography ***The children read to Dr. Smith*****.**

World Knowledge

Text Normalization

Semantic

doctor

Syntactic

Front End

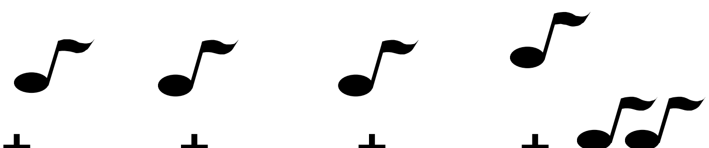
Word Pronunciation

+

dak tR

Word

Intonation

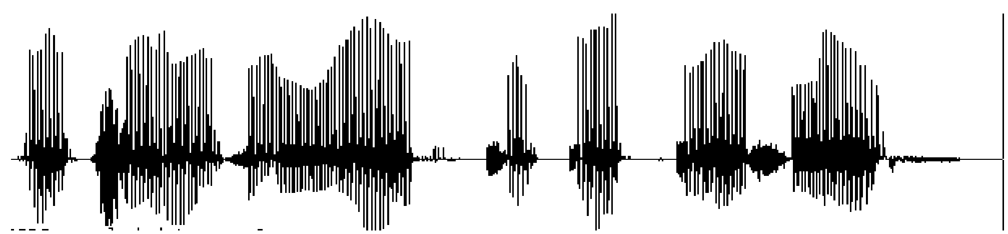


+ + + +

Phonological **%Dx CIldrxn rid tx daktR smIT %**

Back End

Acoustic



Text Normalization

Context-free examples:

Mr.	mister
22	twenty-two
$N	N dollars
NAACP WNEW	N A A C P W N E W
MAACO VISA	“mako” visa

Context-dependent examples:

Dr.	doctor
	drive
St.	saint
	street
1997	nineteen ninety-seven
	one thousand nine hundred ...
3/16	three sixteenths
	March sixteenth

Text Normalization: Methodology

- Abbreviation ambiguity

  St. saint, street

- How to resolve?
  - Application restrictions

    If application is just names = Saint
  - More general: crucial co-occurring items
  - Text context: adjacent words

    St. followed by John = Saint
Joseph
Francis
...
  - Source of list?
    - Intuition
    - Analyze large corpus pre-labeled by human with expansions “Saint”, “Street”

Context-Dependent Text Normalization

- Semantic ambiguity, same part of speech
  - bass bass1 = /bas/, fish; bass2 = /bes/, music

striped bass	bass1
The bass is excellent.	bass1 or bass2
...menu..... The bass is excellent.	bass1

  - Resolve ambiguity by listing disambiguating words:
    - locally, using Ngram statistics
    - globally, using statistics within sentence, paragraph, ...

Context-Dependent Text Normalization

Part of speech ambiguity

+ +

- **convict** **N = /kan vIkt/; V = /kxn vIkt/**
  - **The convict went to jail** **N**
  - **They will convict the criminal** **V**
- **said** **N = /sa id/; V = /sEd/**
  - **Said said hello** **N, V**
- **read** **Vpast = /rEd/; V = /rid/**
  - **They read books** **Vpast /rEd/ or V /rid/**
  - **They will read books** **V /rid/**
  - **They have read books** **Vpast /rEd/**

Examples

– read Vpast = /rEd/; V = /rid/

They read books	Vpast /rEd/ or V /rid/

Local lexical context, possible with Ngrams:

They will read books	V /rid/
They have read books	Vpast /rEd

Global context, possible with syntactic parse or approximation:

They have often read books	Vpast /rEd/:
Have they often read books	Vpast /rEd/
Have the children read books?	Vpast /rEd/
Have the children read books.	V /rid/

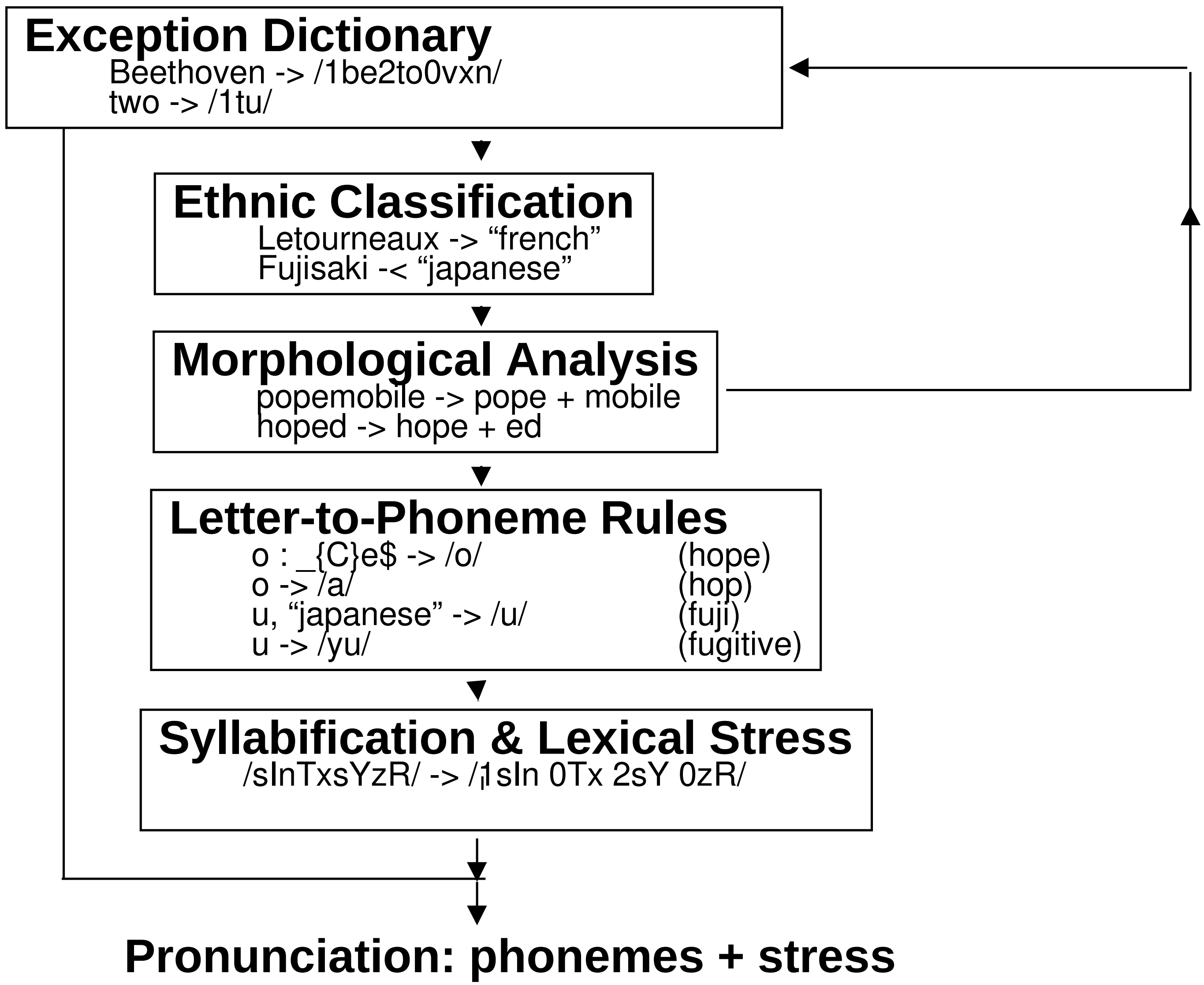
Global context, semantic information, but not possible today:

After dinner they read books	Vpast /rEd/

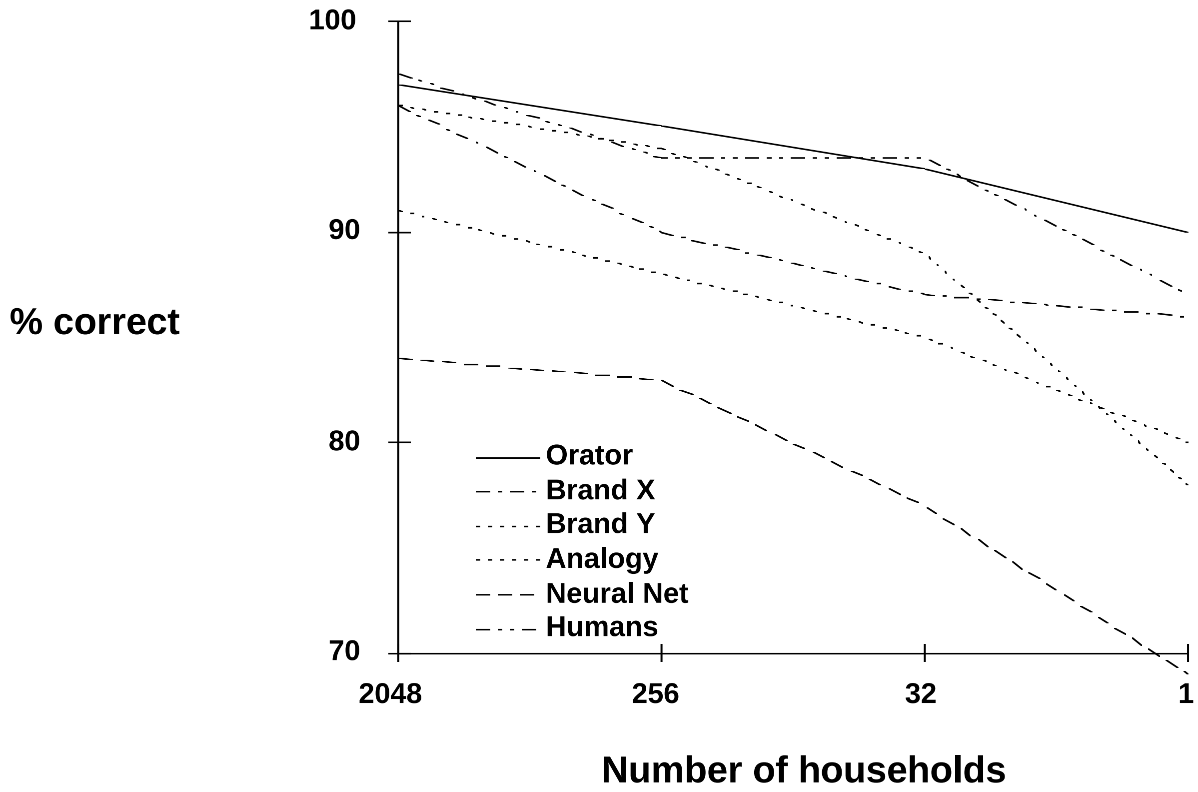
Text Normalization: Summary

- **Application-specific default**
- **Rules based on co-occurring items that resolve ambiguity**
- **Crucial co-occurring items determined by intuition or by statistical analysis of <u>large</u> corpus, pre-labeled by human**
- **Local context: N-grams**
  - **small N (2 or 3)**
  - **words or other forms (part of speech, uninflected words)**
  - **part-of-speech tagger**
- **Global context:**
  - **words within entire text, 100 words, etc.**
- **Syntactic parse**
  - **current systems do not include syntactic parsers, because of computational demands**

Word Pronunciation



Name Pronunciation Accuracy



(Golding, 1991)

Word Pronunciation:
Vowel Reduction and Clitics

Pronunciation of some words, e.g., prepositions, depends on syntactic relationship to adjacent words

• Give the book to John	/tu/
	/tx/
Who did you give the book to?	/tu/
	*/tx/
• Who do you want to see ?	want to
	wanna
Who do you want to join us?	want to
	*wanna

Word Pronunciation: Summary

- Determine phonemes, stress of words
- Much work has been done
- Pronunciation can depend on relation to other words
- Hand-crafted or automatically-learned rules
- Use combination of dictionary and rules, with either:
  - small dictionary, many rules
  - large dictionary, few rules

    Rationale: rules complicated, dictionary builds-in effects of rules
- May perform better than humans

Intonation

- Phrasing: intonational groups, pauses
- Accent: melody, prominence of some words
- Important for intelligibility, naturalness, ease of understanding
  - Ambiguities

    Do you want coffee or tea?

    Bill doesn't drink because he's unhappy.

    John called Mary a Republican and then she insulted him.

  - Even in simple material, like names and addresses

Determining Phrasing

- **Punctuation**
  - **234-5678**
  - **(A + B) * C**
  - **We saw Bill, and Mary saw Bob.**
- **Content/function word: no boundary after function words**
- **Syntactic parse**
  - **We saw Bill and Mary leave.**
  - **We saw Bill and Mary saw Bob.**
- **Other factors**
  - **Length of utterance, other phrases**
  - **Accent**

Phrasing

- Traditional approach: hand-done rules
- Approach: statistical analysis of large corpus of hand-labeled phrase boundaries
- Independent variables:
  - Punctuation
  - Part-of-speech from tagger
  - Syntactic constituency from parser
  - Pragmatic factors: length of utterances, other phrases
  - Accent

Accent

Meaning can determine intonation pattern

Yes-No Question

Continuation

but not a lot known about the meaning-intonation relation

What words get accent?

- **content vs function word**

  Which book did John leave?

  Which car did John leave in?

  But...

  Which paragraph did John leave in?

- **compound words**
- **surface position**
- **information status: contrast, focus, given versus new**

Accent and Word Compounds

Normal case: accent on last word

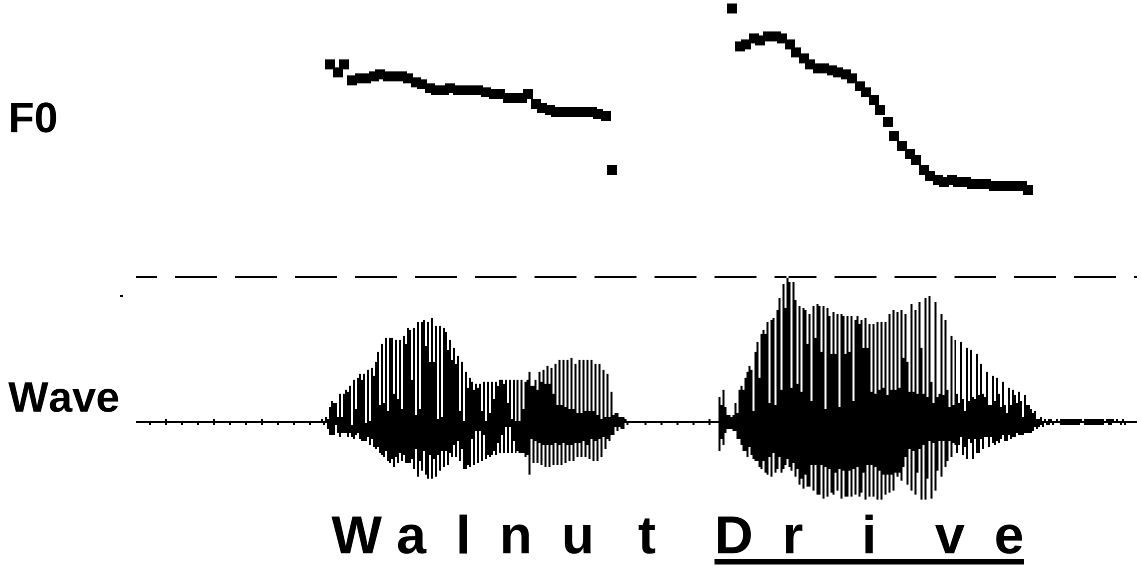
- twenty <u>two</u>
- large <u>number</u>
- long <u>street</u>
- John <u>Smith</u>
- Walnut <u>Drive</u>

Word compounds with accent on first word

- <u>telephone</u> number
- <u>school</u> building
- <u>heart</u> attack
- <u>medical</u> center
- <u>Walnut</u> Street

Identify compounds, store in dictionary

Nuclear Accent



Deaccenting

***given* versus *new* information**

***New* information is accented**

John brought me an apple.

Given information is deaccented.

After John brought me an apple, I ate the apple.

.... *I ate the apple.

How to identify *given* in text?

What is the domain?

- **whole text, paragraph, discourse structure**

Deaccenting: Identifying *given* information

- **Lexical match**

  After John brought me an apple, I <u>ate</u> the *apple*.
- **Morphology**

  John brought me an apple, but I don't <u>like</u> *apples*.

  John likes to swim, but I <u>hate</u> *swimming*.
- **but Compounds**

  After John felt pains near his heart, he took better <u>care</u> of *his heart*.

  After John had a heart attack, he took better care of his <u>heart.</u>
- **Real world knowledge**

  John brought me an apple, but I don't <u>like</u> *fruit*.

  After we met Bill Clinton, we had <u>lunch</u> with *the President*

  John went to Boston on vacation, but I don't <u>like</u> *big cities*.

<u>accent</u>
given

Intonation: Phrasing and Accent

- **Traditional approach: hand-done rules, simple text analysis**
- **Approach: statistical analysis of large corpus of hand-labeled accent and phrase boundaries**
- **Independent variables:**
  - **Punctuation**
  - **Syntactic factors:**
    - **Part-of-speech from tagger**
    - **Syntactic constituency from parser**
    - **Morphological decomposition**
  - **Pragmatic factors: length of utterances, other phrases**
  - **Semantic factors: *given* versus *new*, contrast, focus**
- **Have gained much Insight**

Deaccenting in Text and Dialog

Text:

- **Turn right on Walnut <u>Drive</u>.**
- **After you get to Walnut <u>Drive</u>,**

  turn <u>right</u> on *Walnut Drive.*

  turn right on *Walnut <u>Drive.</u>
- **......., go straight and turn right on Oak <u>Lane</u>.**

  ***   turn <u>right</u> on Oak Lane.**
- **......., go straight and turn right on <u>Oak</u> *Drive*.**

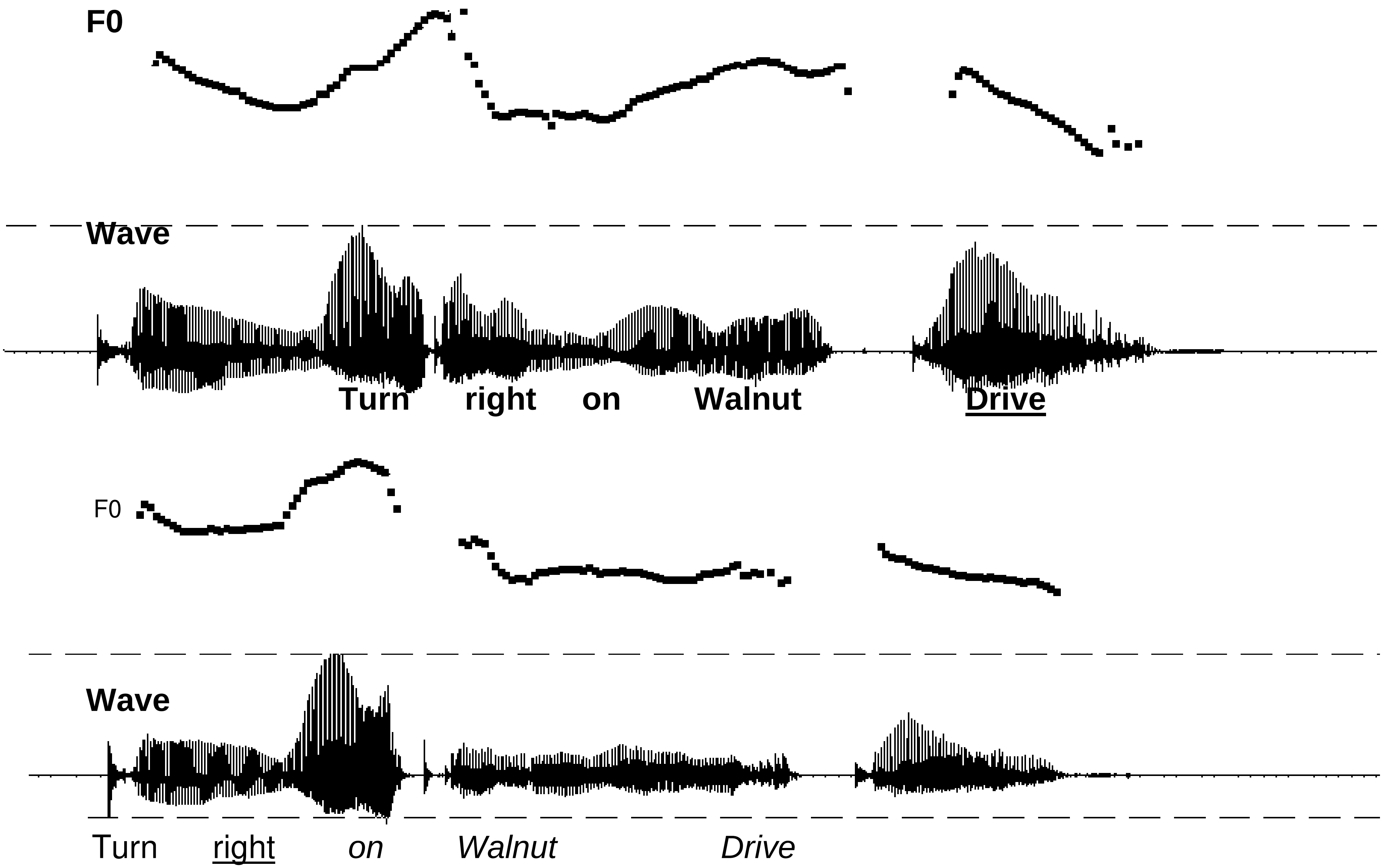
  ***   turn right on Oak *<u>Drive</u>.***

Dialog:

- **Q: I'm coming up on Walnut <u>Drive</u>. What do I do?**
- **A: Turn <u>right</u> on *Walnut Drive.***

  Turn right on *Walnut <u>Drive</u>.

<u>accent</u> *<u>accent</u>*
given
***inappropriate**


F0
Wave
Turn right on Walnut Drive
F0
Wave
Turn right on Walnut Drive

More Accenting: Issues for Dialog

– Q: Should I turn right on Walnut Drive?

– A: Yes *turn right on Walnut Drive*.

Yes, *turn right on Walnut Drive.*

<confirmation: don’t deaccent given>

– A: No, turn left *on Walnut Drive*.

*No, turn left *on Walnut Drive*.

– A: No, don’t *turn right on Walnut Drive*.

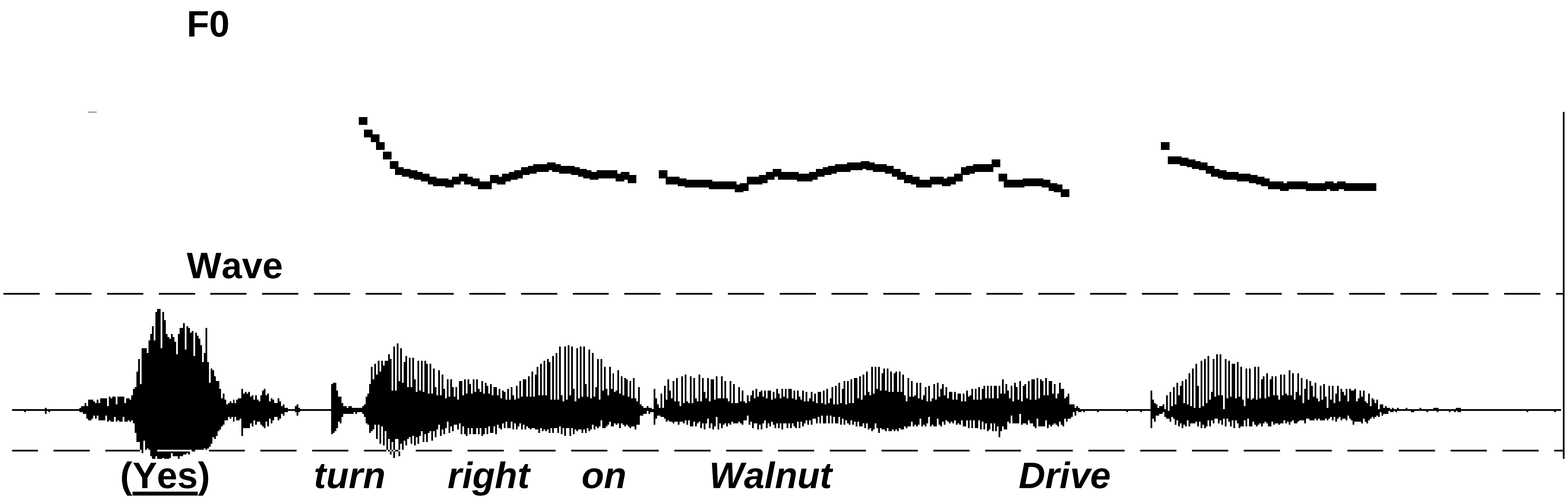
*?No, don’t *turn right on Walnut Drive*.

<ok, but wrong choice for dialog system; implies confirmation>

– A: No, turn right on Oak Lane

No, turn left on Oak Lane

accent *accent given*
*inappropriate


F0
Wave
(Yes)
turn
right
on
Walnut
Drive

TTS Front End
Summary

- Automatic, corpus-based, statistical methods versus hand-crafted rules
- Dictionaries versus rules
- Tap syntax, semantics; within and beyond sentence
- Text analysis tools
  - Part-of-speech tagger
  - Syntactic constituent parse
  - Morphological decomposition
  - Word co-occurrence methods
  - Discourse structure

TTS “Back End”
Phonological to Acoustic

Goal: convert phonological representation into acoustic signal

<u>Phonological</u>

- **Segmental**
  - **Phonemes**
- **Suprasegmental**
  - **Accent**
  - **Phrasing**

<u>Acoustic</u>

- **Spectrum**
- **Pitch (F0)**
- **Duration**
- **Amplitude**
- **….**

Involve phonetic rules and signal processing

Phonetic Component: Segmentals

The problem:

The same phoneme has different acoustic realizations, depending on…..

Nearby phonemes

Stress

...

Approaches:

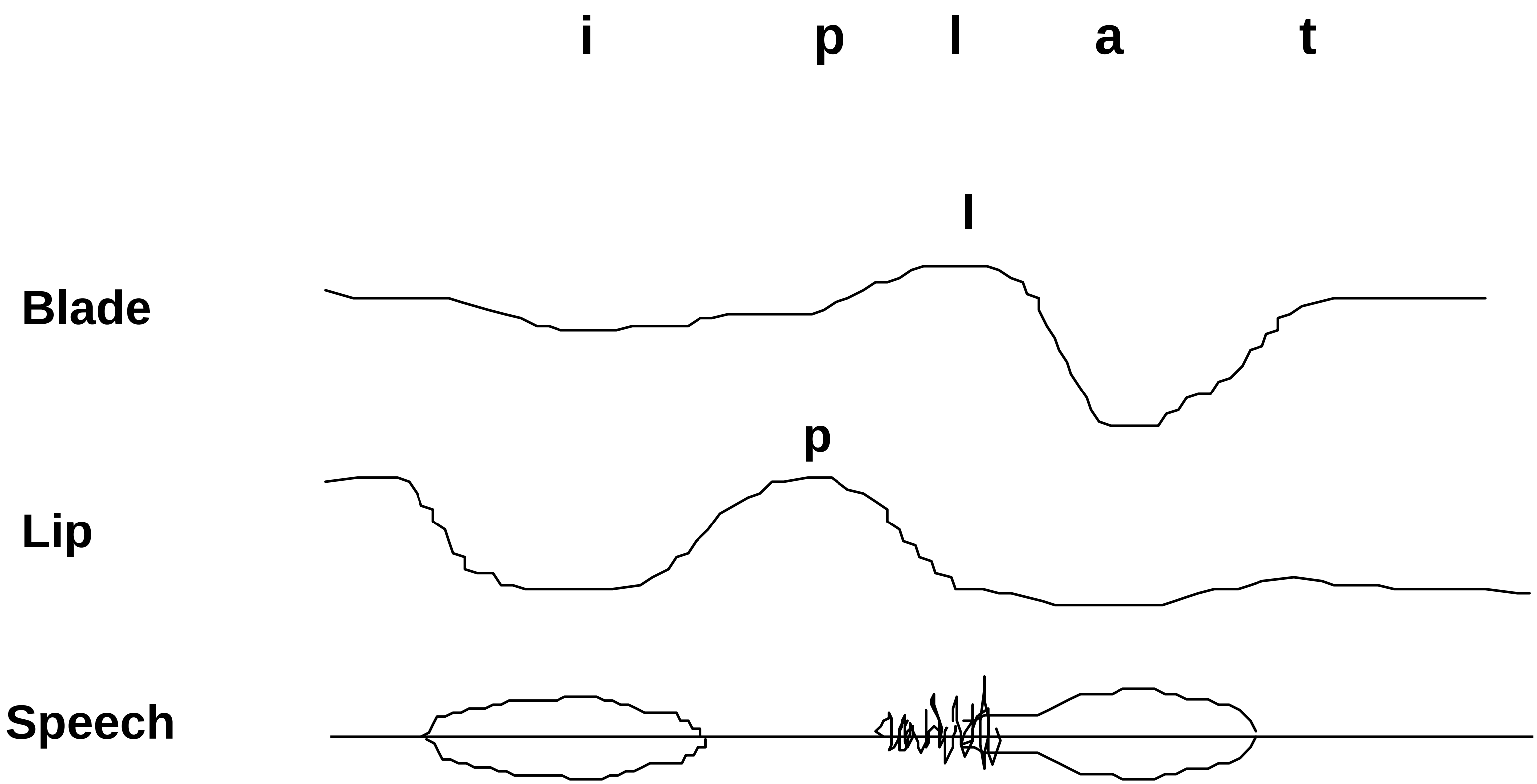
- **Articulatory synthesis**
- **Acoustic (“formant”) synthesis**
- **Synthesis-by-concatenation**

Articulatory Synthesis-by-Rule

- Model of articulators:

  tongue body, tongue tip, jaw, lips, velum, vocal folds
- Rules control timing of movements of different articulators and velocity of articulator movements, etc.
- A phoneme need not have movements specified for all articulators.

Articulatory Synthesis-by-Rule



after Browman & Goldstein 1987

Articulatory Synthesis-by-Rule

Easy to model coarticulation due to independent articulators overlapping in time

Speech quality not natural:

- Details of transform from vocal tract to acoustic signal are still not well understood.
- Human knowledge of all linguistic and articulatory control rules is incomplete.

Acoustic Synthesis-by-Rule

Model of acoustic parameters:

formant frequencies, bandwidths, amplitude of voicing, aspiration, nasal zeros, etc.

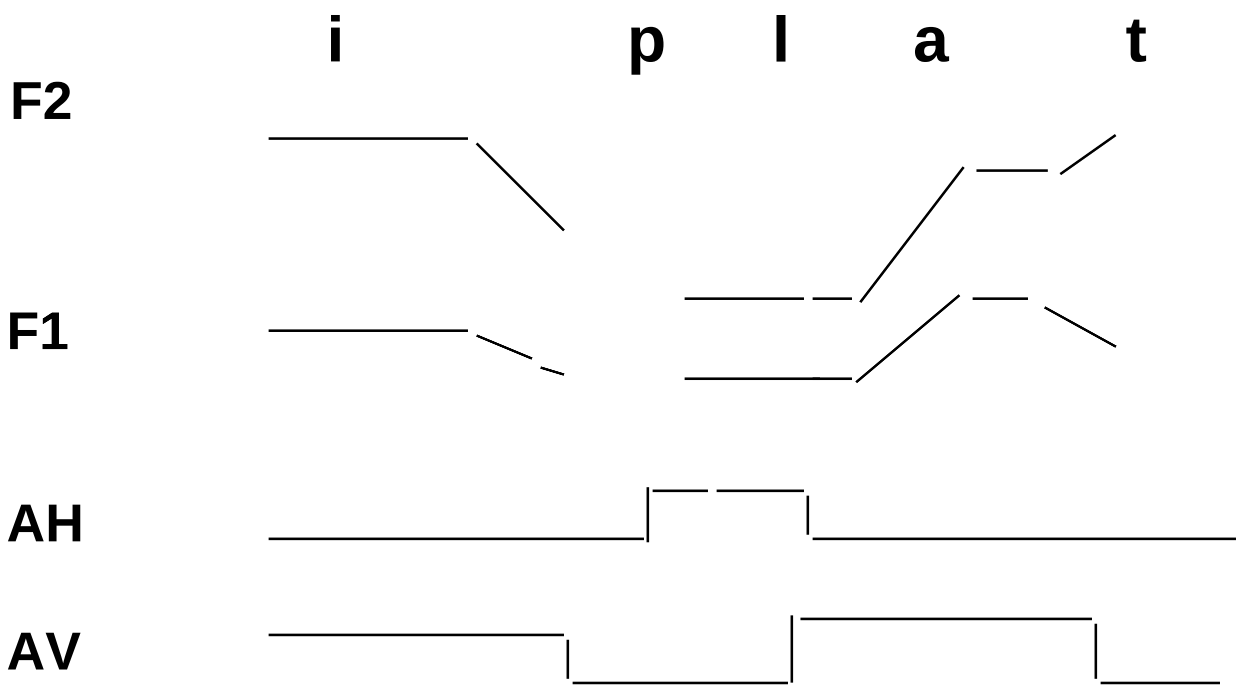
Phonemes or subphonemic units have *target* values specified for parameters

Given a phonemic transcription:

Rules select a sequence of target values

Rules determine the duration of the target values and determine *transitions* between targets

Acoustic Synthesis-by-Rule



after Klatt 1988

Acoustic Synthesis-by-Rule

Gain insight into speech production and perception

Speech quality not natural:

- **Acoustic model is still incomplete.**
- **Human knowledge of all linguistic and acoustic control rules is incomplete.**

Synthesis-by-Concatenation

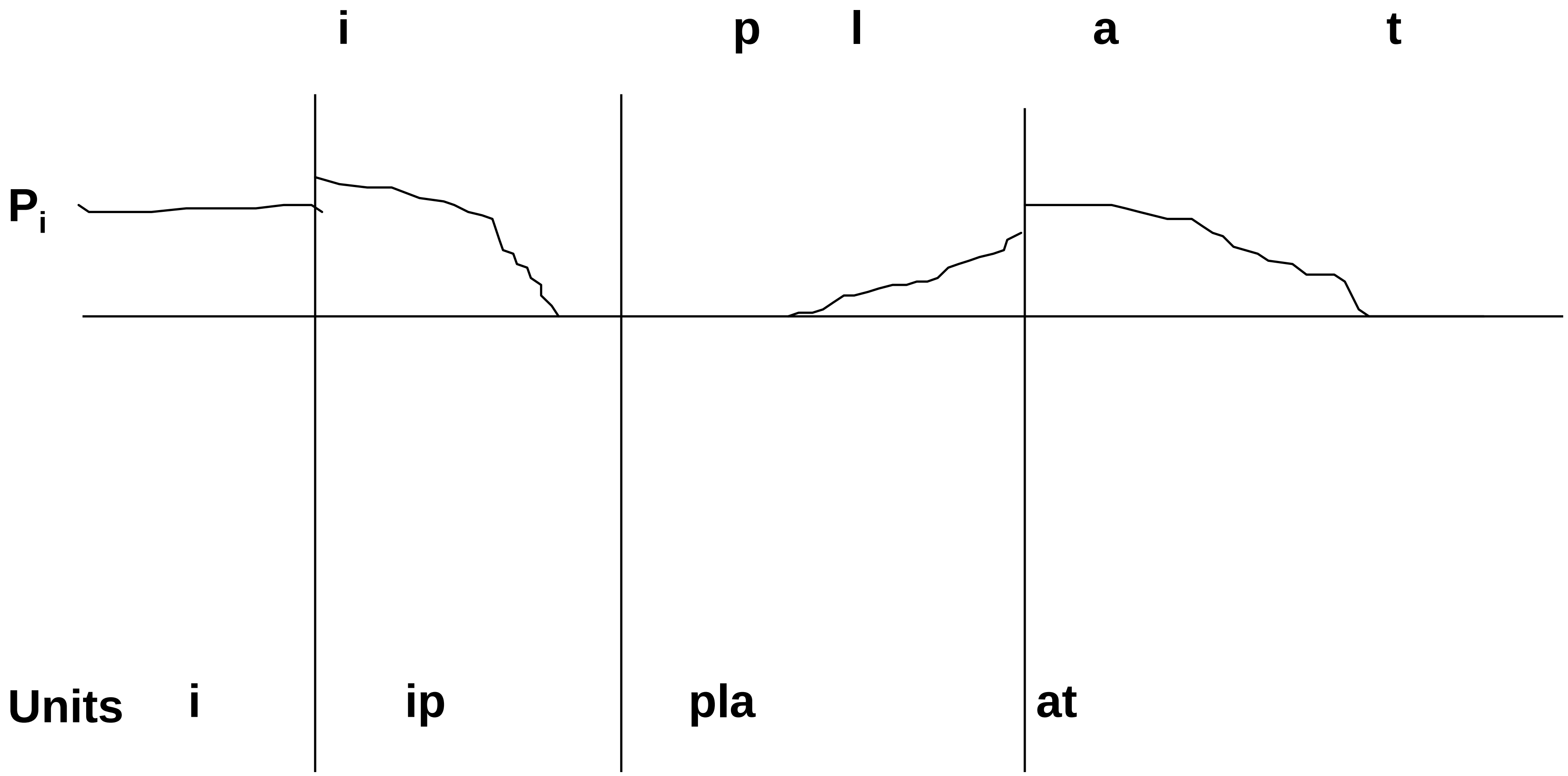
Prerecorded human speech

- Units excised, coded and stored
- Units typically contain more than one phoneme

Given a phonemic transcription:

- Rules select a sequence of units
- Rules concatenate units
- Rules modify duration, amplitude, pitch, source, and smooth the spectrum across the splicing point

Synthesis-by-Concatenation



after Spiegel & Macchi 1985

Synthesis-by-Concatenation: Issues

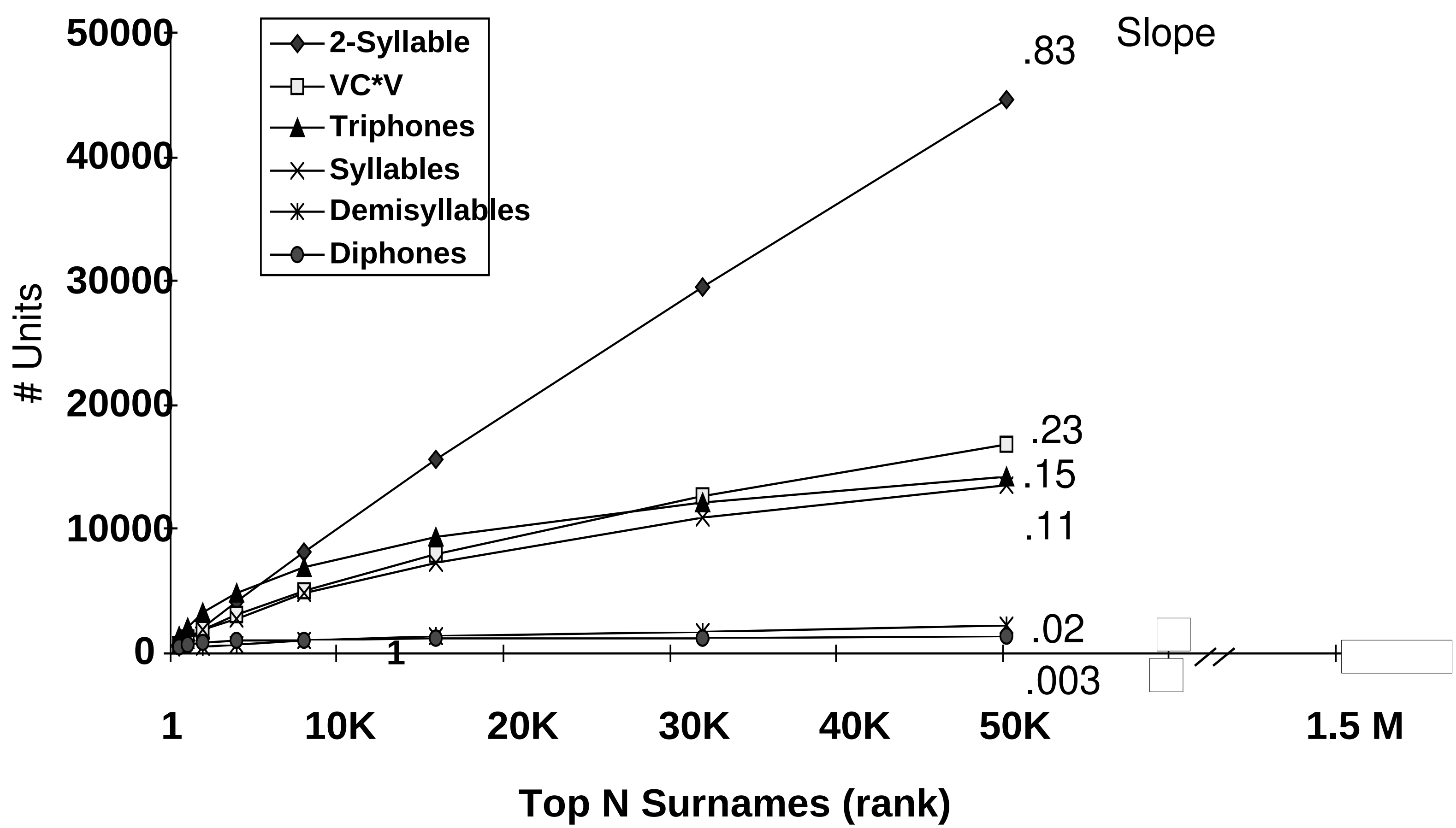
Speech quality depends on:

- **Choice of unit**
  - **Size of unit**
  - **Corpus coverage**
  - **Acoustic coverage**
- **Formulation of rules**
  - **Intonation, phrasing**
  - **Select, concatenate, manipulate templates**
  - **Many rules ….. No rules**
- **Choice of speech coding method**
  - **Coding decomposes acoustic signal into parameters: spectrum, F0, amplitude .... Needed for changing F0, duration of units**

Traditional Segmental Units

Unit Length	Unit	# Units (English)	# Rules	Quality
Short			Many	Low
	Phoneme/Allophone	40/65		
	Diphone	$<40^2$-65^2		
	Triphone	$<40^3$-65^3		
	Demisyllable	2K		
	Syllable	11K		
	VC*V			
	2-syllable	$<11K^2$		
	Word	100K-1.5M		
	Phrase			
	Sentence	∞		
Long			Few	High

Corpus Coverage by Inventory Units



Variable-Length Units

Unit Length	Acoustic Inventory Element		Examples
Short	Phoneme/Allophone		
	Diphone		most phoneme pairs
	Triphone		Cons - schwa - Cons
			Vowel - /h/ - Vowel
	Demisyllable		
	Syllable	N-phones	phones in function words
	VC*V		
	2-syllable		
	Word		
	Phrase		
Long			

Spectral Transitions:

- perceptually significant
- difficult to synthesize by rule
- relatively invariant

After Olive, 1990

Context-Sensitive Units

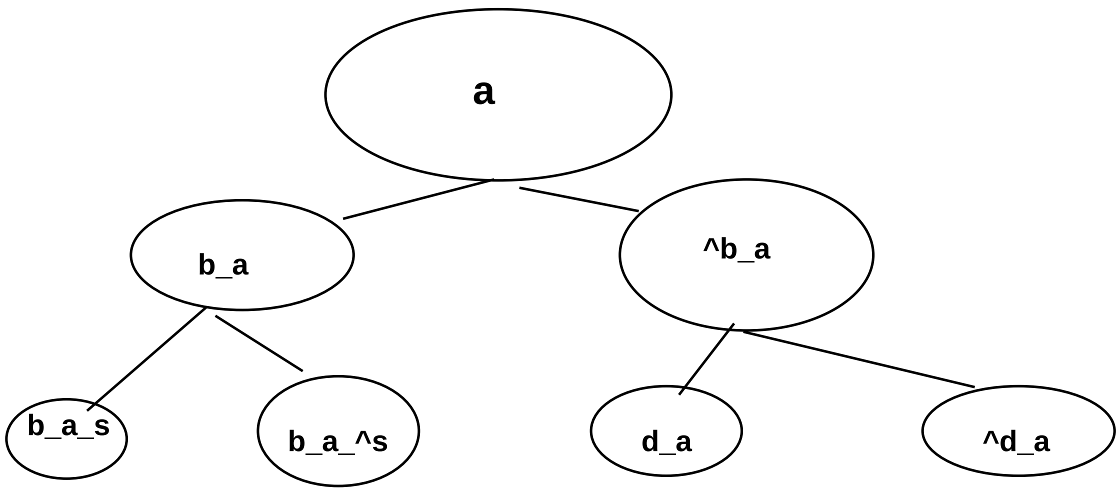
Unit	Contextual Variants Left_Phone, Right_Phone Final/Nonfinal, Stress, ... Pitch
Phoneme/Allophone	. . .
Diphones	/bi/ +stress, /bi/ -stress
Triphone	
Demisyllable	/di/ after front glides, /di/ other
Syllable	/bab/ final, /bab/ nonfinal
VC*V	
2-syllable	
Word	accented, unaccented
Phrase	
Sentence	

Automating Inventory Design

Select units yielding maximal coverage of acoustic space:

Example: Context-Oriented Clustering

Assumption: physical distance = perceptual distance



after Nakajima 1988

Automating Unit Selection

- **Unit selection:**
  Units determined case-by-case from large corpus

- **Concatenation:**
  Splicing point determined case-by-case

Turn right on Walnut Drive

turn right off South
right on Woodland
Units **in Walnut Creek**
Scott Drive

Summary
Trends in Segmentals

- Large sound inventories, incorporating effect of rules:
  - Long units, containing many phonemes
  - Multiple versions of each unit, with different accent and phrasing, to build-in duration and F0 rules and minimize quality distortion due to changing duration and F0.
  - Corpus-based inventory versus phonotactic-based inventory

- Automated techniques for designing, excising and concatenating inventory units, using spectral distance measures

Duration Rules

"Phoneme" duration determined by:

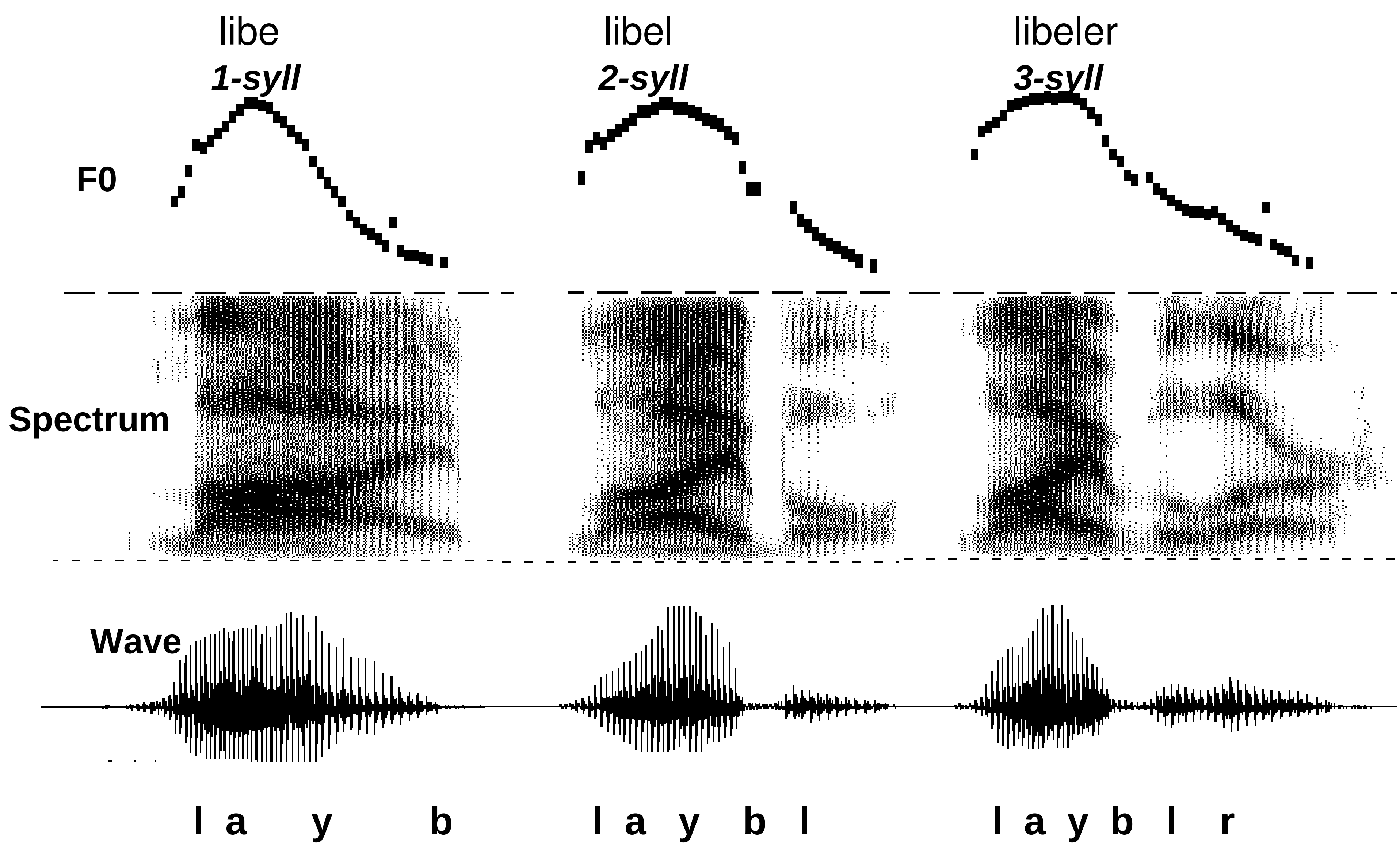
- **identity of nearby phonemes**
- **position within syllable**
- **1-, 2-, 3- syllables**
- **phrasing; utterance-final versus not**
- **stress**
- **speaking rate**

How to predict duration for TTS:

- **rules apply to phonemes, larger units (syllable, foot), smaller units (subparts of phonemes)?**
- **segmental context same as suprasegmental context?**
- **how do rules interact?**
- **some "allophones" (flapping) better as results of duration rules than as allophonic rules?**

Does a TTS system need duration rules at all?

Duration & Pitch for Same Intonation



Pitch Rules

- F0 determined by many factors:
  - accent
  - phrasing
  - phonemic factors

- How to predict F0
  - different intonation models:
    - phonological status
    - contour vs target
  - how to align F0 contour with phonemes
  - how fast F0 falls
  - how high, how low F0
  - how F0 interacts with phonemes

- Does a TTS system need F0 rules at all?

Summary
Trends in Suprasegmentals

- Rules relate accent, phrasing, phonemes to F0, duration, amplitude
- Use of duration, F0 templates rather than rules
- Large sound inventories, directly incorporating effect of suprasegmental effects. Multiple versions of each unit, with different duration and F0.
  - inventory builds-in duration and F0 rules
  - minimize speech quality distortions due to speech coding methods
  - coverage?; speech quality distortions really minimized?
- Automated techniques for formulating rules or preparing large inventory, using corpus of labeled data

Speech Coding Methods for Synthesis

- **Traditional method: LPC (linear predictive coding)**
  - **decompose waveform into parameters representing vocal tract/formant frequencies plus F0, amplitude; simplified representation of glottal excitatiion**
  - **sounds robotic**
- **MPLPC (multipulse LPC), RELP (residual-excited LPC)**
  - **more elaborate representation of glottal excitation**
  - **less robotic-sounding, but distortions when F0, duration change**
- **PSOLA (pitch-synchronous overlap-and-add)**
  - **no waveform decomposition: delete, repeat individual pitch-periods to change duration; overlap pitch-periods to change F0**
  - **distortion with large F0, duration change; sensitive to pitch-period definition**
- **No coding at all: use natural speech**
  - **to avoid distortions of coding methods**
  - **how to blend natural and coded speech in same sentence?**

Synthesizing Different Voices

- Acoustic transform of one voice
  - Traditional approach: raise/lower pitch, formants, vary glottal excitation
  - Fast, easy
  - Results?
- Different sound inventory for each speaker
  - Need to record, excise inventory
  - Was slow, laborious
  - Automated techniques speed up process
  - Results?
- Research needed on characteristics of particular voice

TTS Back End
Summary

- Speech quality still unnatural
  - spectrum, duration, pitch
  - speech coding
- Current approaches:
  - alternatives to linear representaion
  - large inventories
  - replace rules with inventory
  - statistical models
  - automatic methods, borrowed from ASR
  - use natural speech

Evaluation of TTS

- Intelligibility test
  - standard method
  - synthesize nonsense words (“blerk”) and have listeners transcribe them
  - tests segmental intelligibility only
- Naturalness test
  - synthesize words, sentences and have listeners assign subjective numeric rating
  - numeric rating is not absolute; relative to other ratings
- Application-specific tests
  - field trials
  - no standard methodology

Text-to-Speech Summary

- Still not natural; far from Turing test
- Front End and Back End:
  - automatic, corpus-based, statistical methods versus hand-crafted rules
  - specific domains, speaking styles
  - dictionaries, inventory versus rules
- Front End: tap syntax, semantics; beyond sentence
- Back End: natural speech to incorporate rules, minimize coding distortions