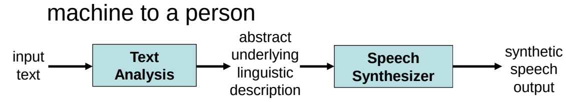
Digital Speech Processing—Lecture 18

Text-to-Speech (TTS) Synthesis Systems

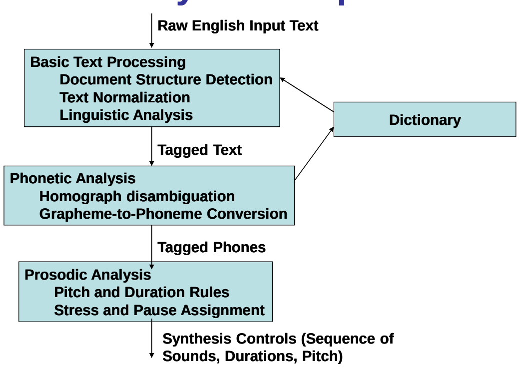
Text-to-Speech (TTS) Synthesis

- **GOAL**: convert arbitrary textual messages to ***intelligible*** and ***natural*** sounding synthetic speech so as to transmit information from a machine to a person



- pronunciation of text => phonemes, stress, intonation, duration
- syntactic structure of sentence (pauses, rate of speaking, emphasis)
- semantic focus, ambiguity resolution (duration, intonation)
  - rules for word etymology (especially names, foreign terms)

Text Analysis Components



Document Structure

- **end of sentence** marked by '.?!' is not infallible
  - "The car is 72.5 in. long"
- **e-mail** and web pages need special processing
  - Larry:

    Sure. I'll try to do it before Thursday :-)

    Ed
- **multiple languages**
  - insertion of foreign words, unusual accent and diacritical marks, etc.

Text Normalization

- **abbreviations and acronyms**
  - Dr. is pronounced either as Doctor or drive depending on context (Dr. Smith lives on Smith Dr.)
  - St. is pronounced either as 'street' or 'Saint' depending on context (I live on Bourbon St. in St. Louis)
  - DC is either direct current or District of Columbia
  - MIT is pronounced as either 'M I T' or 'Massachusetts Institute of Technology' but never as 'mitt'
  - DEC is pronounced as either 'deck' or 'Digital Equipment Company' but never as 'D E C'
- **numbers**
  - 370-1111 can be either 'three seven oh …' or 'three seventy-model 1111' for the IBM 370 computer
  - 1920 is either 'nineteen-twenty' or 'one thousand, nine hundred, twenty'
- **dates, times currency, account numbers, ordinals, cardinals, math**
  - Feb. 15, 1983 needs to convert to 'February fifteenth, nineteen eighty-three'
  - $10.50 is pronounced as 'ten dollars and fifty cents'
  - part # 10-50 needs to be pronounced as 'part number 10 dash fifty' rather than 'part pound sign ten to fifty'

Text Normalization

- **proper names**
  - Rudy Prpch is pronounced as 'Rudy Perpich'
  - Sorin Ducan is pronounced as 'Sorin Duchan'
- **part of speech**
  - read is pronounced as 'reed' or 'red'
  - record is pronounced as 'rec-ard' or 'ri-cord'
- **word decomposition**
  - need to decompose complex words into base forms (morphemes) to determine pronunciation ('indivisibility' needs to be decomposed into 'in-di-visible-ity' to determine pronunciation)

Text Normalization

- proper handling of **special symbols** in text
  - punctuation, e.g., . , : ; ' " - -- _ * & ( ) ^ % @ ! ~ < > ? / \ = +
- **string resolution**
  - 10:20 can be pronounced as either 'twenty after 10 (as a time)' or 'ten to twenty' (as a sequence)

Knowledge Source Confusions

- Let us pray ---------- Lettuce spray
- Pay per view ---------- Paper view
- Meet her on Main St. ---------- Meter on Main St.
- It is easy to recognize speech ---------- It is easy to wreck a nice beach

8

Linguistic Analysis

- part-of-speech (POS)
- word sense
- phrases
- anaphora
- emphasis
- style

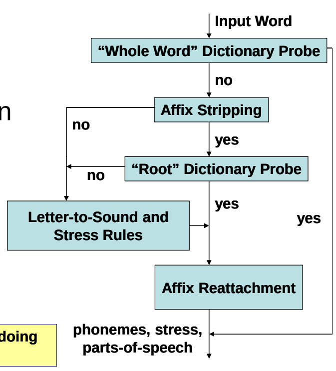
a conventional parser could be used, but typically a simple shallow analysis is done for speed (parsers are not real-time!)

Homograph Disambiguation

- "an absent boy" versus "do you choose to absent yourself?"
- "they will abuse him" versus "they won't take abuse"
- "an overnight bag" versus "are you staying overnight?"
- "he is a learned man" versus "he learned to play piano"
- "El Camino Real road" versus "real world"

Letter-to-Sound (LTS) Conversion

- CART (Classification and Regression Tree) analysis
- conventional dictionary search with letter-to-sound rules

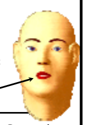


This is only ONE way of doing LTS. FSMs are another.

Prosody

- **pauses**
  - to indicate phrases and to avoid running out of breath
- **pitch**
  - fundamental frequency (F0) as a function of time
- **rate/relative duration**
  - phoneme durations, timing, and rhythm
- **loudness**
  - relative amplitude/volume

Full TTS System



Text/Language → **Text Analysis** → **Letter-to-Sound Rules** → **Synthesis Backend** → Speech

Face Sync

- Numerical expansion (dates, times, $$$, …)
- abbreviations, acronyms
- proper name id

"Dr. Smith lives at 23 Lakeshore Dr."

Prosody

- text and name dictionaries
- timing and duration [rate, stress (position, context)]
- intonation/pitch (type of phrase[list,statement, ??])
- phonetic substitution (vowel reduction, flapping rules)
- loudness/amplitude (phrasal amp., local reduction)

Parsing:
- sentences, phrase and breath groups
- semantic and syntactic accent; stress of compounds
- intonation types; parts of speech

Synthesis model impacts TTS, as well as unit representation.

choice of units:
- words
- phones
- diphones
- dyads
- syllables

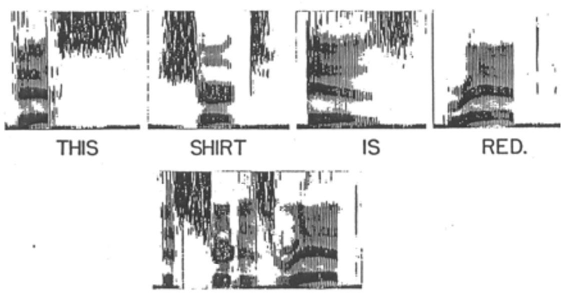
choice of parameters:
- LPC
- formants
- waveform templates
- articulatory parameters
- sinusoidal parameters

method of computation:
- rules
- concatenation

Word Concatenation Synthesis

- words in sentences are much shorter than in isolation (up to 50% shorter) (see next page)
- words cannot preserve sentence-level stress, rhythm or intonation patterns
- too many words to store (1.7 million surnames), extended words using prefixes and suffixes

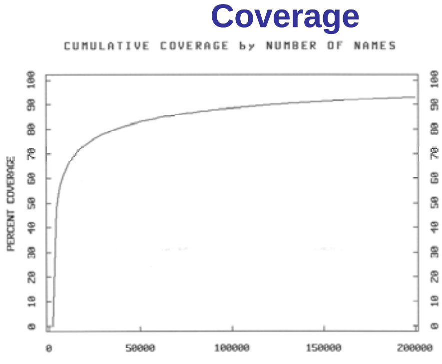
Word Concatenation



Proper Name Statistics

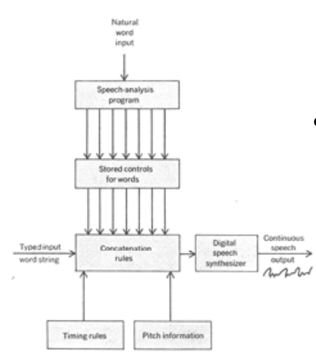
Number of Names	Coverage
10	4.9%
100	16.3%
5,000	59.1%
50,000	83.2%
100,000	88.6%
200,000	93.0%

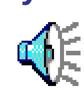
Statistics of Proper Name Coverage



name pronunciation based on etymology

Concatenative Word Synthesis





several examples of concatenated words and full sentences--SBR

word-based synthesis does not work!

Speech Synthesis Methods

- 1939—the VODER (Voice Operated DEmonstratoR)—Homer Dudley
  - based on a simple model of speech sound production
  - select voicing source (with foot pedal control of pitch) or noise source
  - ten filters shaped the source to produce vocal or noise-excited sounds—controlled by finger motions
  - separate keys for stop sounds
  - wrist bar control of signal energy

The VODER

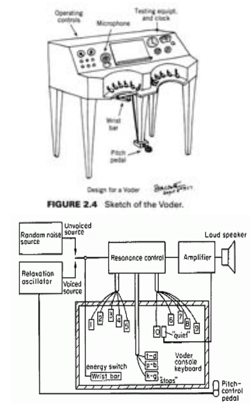


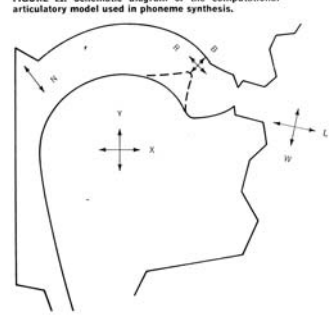
FIGURE 2.4 Sketch of the Voder.

Articulatory Synthesis

- ***in theory*** — can create more natural and more realistic motions of the articulators (rather than formant parameters), thereby leading to more natural sounding synthetic speech
  - utilize physical constraints of articulator movements
  - use X-ray data to characterize individual speech sounds
  - model how articulatory parameters move smoothly between sounds
    - ***direct method***: solve wave equation for sound pressure at lips
    - ***indirect method***: convert to formants or LPC parameters for final synthesis in order to utilize existing synthesizers
  - use highly constrained motions of articulatory parameters

Articulatory Model

FIGURE 21. Schematic diagram of the computational articulatory model used in phoneme synthesis.



Articulatory Parameters of Model:

- **lip *opening*—*W***
- **lip protrusion length—*L***
- **tongue body height and length—*Y,X***
- **velar closure—*K***
- **tongue tip height and length—*B,R***
- **jaw raising (dependent parameter)**
- **velum opening—*N***

Articulatory Synthesis Using Formant Synthesizer Backend

Cecil Coker--teaching computers to talk

Articulatory Synthesis of Speech—Cecil Coker

Articulatory Synthesis by Copying Measured Vocal Tract Data

- fully automatic closed-loop optimization
  - initialized from articulatory codebooks, neural nets
  - Schroeter and Sondhi, 1987

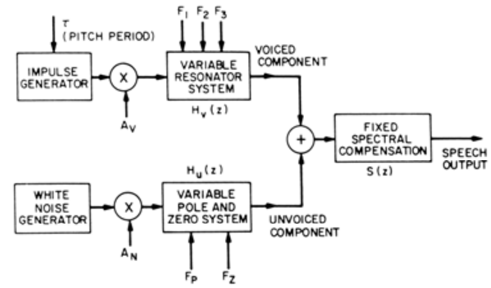
One example: original: re-synthesis:

Articulatory Synthesis Issues

- requires highly accurate models of glottis and vocal tract
- requires rules for dynamics of the articulators

Source-Filter Synthesis Models

- cascade/serial (formant) synthesis model



Serial/Formant Synthesis Model

- **flaws in the serial/formant synthesis model:**
  - **can't handle voiced fricatives**
  - **no zeros for nasal sounds**
  - **no precise control for stop consonants**
  - **pitch pulse shape fixed—independent of pitch**
  - **spectral compensation is inadequate**

OVE 1--Fant **SPASS synthesis** **JSRU Synthesis** **"To Be …"- Bell Labs** **We Wish You …** **Daisy-Daisy with music**

Parallel Synthesis Model

A serial synthesizer is a good approach for open, non-nasal vocal tracts (vowels, liquids). For obstruents and nasals, we need to control the amplitudes of each resonance, and to introduce zeros in addition to the poles.

$$V(z)=\prod_{k=1}^{4}\frac{1-2r_k\cos(\theta_k)+r_k^2}{1-2r_k\cos(\theta_k)z^{-1}+r_k^2z^{-2}}$$

$$=\sum_{k=1}^{4}\frac{A_k-B_kz^{-1}}{1-2r_k\cos(\theta_k)z^{-1}+r_k^2z^{-2}}$$

- use the approximation

$$V(z)\approx\sum_{k=1}^{4}\frac{A_k}{1-2r_k\cos(\theta_k)z^{-1}+r_k^2z^{-2}}$$

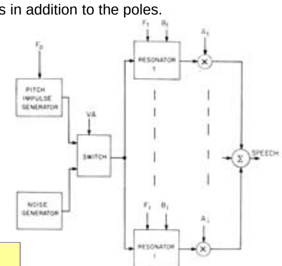


Fig. 7. Parallel terminal-analog synthesizer.

parallel synthesizer provides more flexibility in matching spectrum levels at formant frequencies (via gain controls)—however, zeros are introduced into the spectrum.

Parallel Synthesis

- **issues:**
  - **need individual resonance amplitudes (*A1,…,A4*)—if resonances are close, this is a messy calculation**
  - **phasing of resonances neglected (the B_kz^{-1} terms)**
  - **synthetic speech has both resonances and zeros (at frequencies between the resonances) that may be perceptible**
  - **better reproduction of complex consonants**

Parallel synthesis-Holmes **Parallel synthesis from BYU**

Continuing Evolution (1959-1987)

- Haskins, 1959
- KTH – Stockholm, 1962
- Bell Labs, 1973
- MIT, 1976
- MIT-talk, 1979
- Speak 'N spell, 1980
- BELL Labs, 1985
- Dec talk, 1987

Text-to-Speech Synthesis (TTS) Evolution

Poor Intelligibility; Poor Naturalness →	Good Intelligibility; Poor Naturalness →	Good Intelligibility; Customer Quality Naturalness (Limited Context)
Formant Synthesis	**LPC-Based Diphone/Dyad Synthesis**	**Unit Selection Synthesis**
Bell Labs; Joint Speech Research Unit; MIT (DEC-Talk); Haskins Lab	Bell Labs; CNET; Bellcore; Berkeley Speech Technology	ATR in Japan; CSTR in Scotland; BT in England; AT&T Labs (1998); L&H in Belgium

1962 1967 1972 1977 1982 1987 1992 1997

Year

Speech Synthesis—the 90's

- what changed?
  - TTS was highly intelligible but extremely unnatural sounding
  - a decade of work had not changed the naturalness substantially
  - computation and memory grew with Moore's law, enabling highly complex concatenative systems to be created, implemented and perfected
  - concatenative systems showed themselves capable of producing (in some cases) extremely natural sounding synthetic speech

Concatenation TTS Systems

- key idea: use segments of recorded speech for synthesis
- data driven approach ➔ more segments give better synthesis ➔ using an infinite number of segments leads to perfect synthesis
- key issues:
  - what units to use
  - how to select units from natural speech
  - how to label and extract consistent units from a large database
  - what signal representation should be used for spectrally smoothing units (at junctures) and for prosody modification (pitch, duration, amplitude)

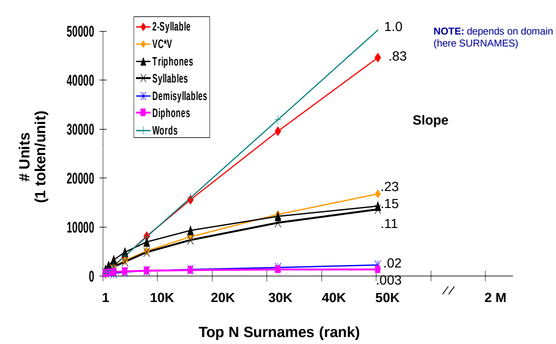
Concatenation Units

- choice of units:
  - **Words**—there are an infinite number of them
  - **Syllables**—there are about 10K in English
  - **Phonemes**—there are about 45 in English
  - **Demi-syllables**—there are about 2500 in English
  - **Diphones**—there are about 1500-2500 in English

Choice of Units

Length	Unit	# Units (English)	# Rules, Necessary Unit Modifications	Quality
Short ↓			Many ↑	Low ↓
	Allophone	60-80		
	Diphone	$<40^2-65^2$		
	Triphone	$<40^3-65^3$		
	Demisyllable	2K		
	Syllable	11K		
	VC*V			
	2-syllable	$<11\ K^2$		
	Word	100K-1.5M		
	Phrase	∞		
Long	Sentence		Few	High

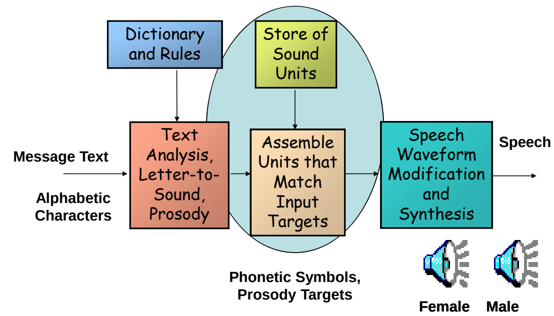
Corpus Coverage by Unit Type



Concatenation Units

- **Words:**
  - no complete coverage for broad domains ➔ words have to be supplemented with smaller units
  - limited ability to modify pitch, amplitude and duration without losing naturalness and intelligibility
  - need huge database to extract multiple versions of each word
- **Sub-Words:**
  - hard to isolate in context due to co-articulation
  - need allophonic variations to characterize units in all contexts
  - puts large burden on signal processing to smooth at unit join points

Block Diagram of a Concatenative TTS System



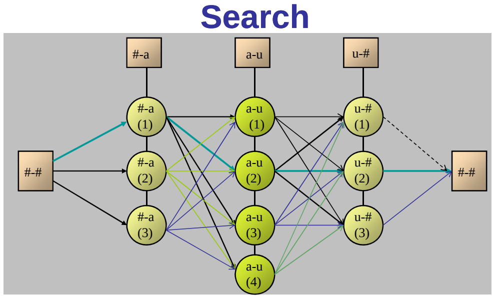
Procedure for Concatenative Synthesis

- off-line inventory preparation:
  - record speech corpus and process with coding method of choice
  - determine location of speech units and store units in inventory
- on-line synthesis from text
  - normalize input text (expand abbreviations, etc.)
  - letter-to-sound (pronunciation dictionary and rules)
  - prosody (melody/pitch&durations, stress patterns/amplitudes, …)
  - select appropriate sequence of units from inventory
  - modify units (smooth at boundaries, match desired prosody)
  - synthesize and output speech signal

Unit Selection Synthesis

- need to optimally match units at boundaries, e.g., fundamental frequency (pitch), and spectrum
- need to automatically and efficiently select optimal sequence of units from database
- issues in Unit Selection Synthesis
  - several examples in each unit category (from 10 to 10^6)
  - waveform modification used sparingly (leads to perceived distortions)
  - high intelligibility must be maintained
  - "customer quality" attained with reasonable training set (1-10 hours)
  - "natural quality" attained with large training set (10's of hours)
  - "unit selection" algorithm must run in a fraction of real time on a state-of-the-art processor

(On-line) Unit Selection: Viterbi Search



- transitional (**concatenation**) costs are based on acoustic distances
- node (**target**) costs are based on linguistic identity of unit

Unit Selection Measures

- USD—Unit Segmental Distortion ➔ differences between desired spectral pattern of target and that of candidate unit, throughout whole unit
- UCD—Unit Concatenative Distortion ➔ spectral discontinuity across boundaries of the concatenated units

Example: source context—want—/w ah n t/
target context—cart—/k ah r t/
USD➔(ah→n versus ah→r)

Modern TTS Systems (Natural Voices from AT&T)

- Soliloquy from Hamlet— NV2005
- Gettysburg Address— NV2005
- Bob Story—
- German female—
- UK British female—
- Spanish female—
- Korean female—
- French male—

Modern COMMERCIAL Systems

- Lucent
- AcuVoice
- Festival
- L&H RealSpeak
- SpeechWorks female
- SpeechWorks male
- Cselt (Actor) - Italian

TTS Future Needs

- TTS needs to know ***how*** things should be said
- context-sensitive pronunciations of words
- prosody prediction➔ emphasis
  - ***I gave the book to <u>John</u> (not someone else)***
  - ***I gave the <u>book</u> to John (not the photos)***
  - ***<u>I</u> gave the book to John (I did it, not someone else)***
- unit selection process ➔ target cost captures mismatch between <u>predicted</u> unit specification (phoneme name, duration, pitch, spectral properties) and <u>actual</u> features of a candidate recorded unit ➔ need better spectral distance measures that incorporate human perception
- better signal processing ➔ compress units for small footprint devices

Business Drivers of TTS

- **cost reduction**
  - **TTS as a dialog component for customer care**
  - **TTS for delivering messages**
  - **TTS to replace expensive recorded IVR prompts**
- **new products and services**
  - **location-based services**
  - **providing information in cars (e.g., driving directions, traffic reports)**
  - **unified Messaging (reading e-mail, fax)**
  - **voice Portals (enterprise, home, phone access to Web-based services)**
  - **e-commerce (automatic information agents)**
  - **customized News, Stock Reports, Sports Scores**
  - **devices**

Reading Email

Example: Old TTS, No Filter

From: Marilyn Walker <walker@research.att.com>
To: David Ross <davidross@home.com>
Subject: Re: Today's Meeting
Date: Tuesday, December 01, 1998 4:25 PM

4:30 is fine for me. See you at the meeting.

Marilyn

-----Original Message-----
From: David Ross <davidross@home.com>
To: Marilyn Walker <walker@research.att.com>
Date: Tuesday, December 01, 1998 2:25 PM
Subject: Today's Meeting

Today's meeting has been changed from 4:00 to 4:30 PM. If the time change is a problem, please send me email at davidross@home.com.

Thanks,

david ross

Reading Email (final)

Example: Enhanced TTS

From: Marilyn Walker <walker@research.att.com>
To: David Ross <davidross@home.com>
Subject: Re: Today's Meeting
Date: Tuesday, December 01, 1998 4:25 PM

4:30 is fine for me. See you at the meeting.

Marilyn Walker

-----Original Message-----
From: David Ross <davidross@home.com>
To: Marilyn Walker <walker@research.att.com>
Date: Tuesday, December 01, 1998 2:25 PM
Subject: Today's Meeting

Today's meeting has been changed from 4:00 to 4:30 PM. If the time change is a problem, please send me email at davidross@home.com.

Thanks,

david ross

TTS Application Categories

Devices	• PDAs, cellphones, gaming, talking appliances
Automotive Connectivity	• driving directions, city and restaurant guides, location services (e.g., "Macys has a sale!")
Consumer Communications	voice control of cell phones, VCRs, TVs • home information access over telephone ("Home Voice Portals")
Enterprise Communications	• information access over the phone such as sales information, HR, internal phonebook, messaging
Voice-assisted E-Commerce	• E-commerce, customer care (e.g., friendly automated talking web agents, FAQs, product information)
Call center Automation	• next-gen "HMIHY" automated operator services