Lecture 12

Speech Synthesis

- Types and applications of speech synthesis
- Stages in speech synthesis
  - Converting non-words to words
  - Converting words to phonemes
  - Prosodic variations

Text-to-Speech Synthesis

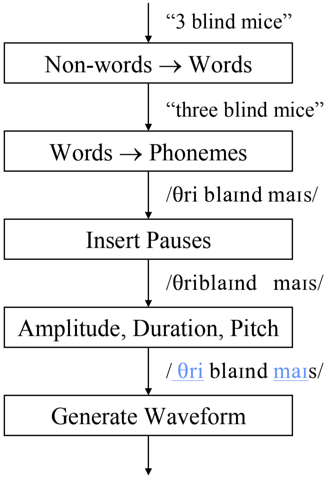
Word or Phrase Concatenation

- Limited Vocabulary
- Natural Sounding provided that entire messages are recorded
- Use for
  - warning messages
  - operating instructions

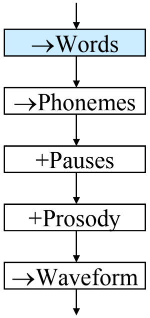
Arbitrary Text-to-Speech

- Unrestricted vocabulary including words never met before
- Use for:
  - reading mail
  - reading a book/newspaper
  - reading back of typed text for proofreading
  - a speech aid for the disabled (though hard to operate in real-time)
  - give information to someone concentrating elsewhere (e.g. a car driver)
- Sounds artificial
  - human readers can insert expression that is not recorded explicitly in the text but is implied by its meaning
- Requires lots of CPU power to sound good

Synthesis Steps



Non-words → Words



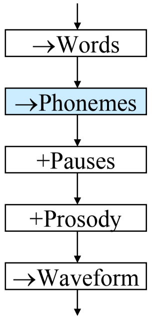
– **Symbols**
- "&" → "and"
- "%" → "percent"

– **Abbreviations**
- "Mr" → "mister"
- "Jan." → "january"
- "St" → "street": "Peter St"
- "St" → "saint": "St Peter"
  – when followed by a capitalised word

– **Acronyms**
- "NATO" → "nayto"
- "HIV" → "aitch eye vee"
- "Henry IV" → "Henry the fourth"
- "Chapter IV" → "Chapter four"

– **Dates and Numbers**
- "The war ended 1945 years ago"
- "The war ended in 1945"
  – when preceded by "in" ...
- "Find a needle in 1945 haystacks"
  – … unless followed by a plural noun
- "An average of 1.945"

Converting Words to Phonemes



The problem:

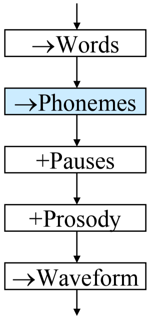
(1) How does "…ough…" sound ?

– "through"
– "though"
– "thought"
– "tough"
– "plough"
– "thorough"
– "hiccough"
– "lough"

(1) How does "bow" sound ?

– "bow" = bend at the waist
– "bow" = knot in a ribbon

Converting Words to Phonemes



We require both of:

– a dictionary: for irregular words
  - "tortilla", "of", "meringue"

– a set of rules: for unknown words

Morphemes

English words are built up from one or more morphemes (sub words):

– "kingdom" = {king} + {dom}
– "choking" = {choke} + {ing}
– "discriminate" = {dis} + {crimin-} + {ate}

We use morphemes rather than whole words because:

– There are fewer distinct morphemes than words.
– Newly coined words are normally made from existing morphemes:
  - "The *windoze* operating system"

The spelling of a morpheme may vary slightly depending on context: a *morph* is a realisation of a *morpheme*.

Dictionary

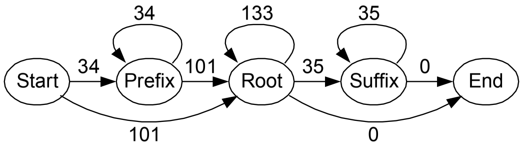
Contains a mixture of whole words and morphemes: 12000 entries in all.

Words → Morphemes

If whole word is in the dictionary, use it, else, try all possible ways of dividing word into morphemes that are in the dictionary.

Each morpheme has a cost: prefixes cost least. roots cost most. Choose the decomposition with lowest cost. Hence prefer:

- – decompositions with common prefixes & suffixes
- – decompositions with few rather than many morphemes
    - “scarcity” = {scarce} + {-ity} rather than {scar} + {city} or {scar} + {cite} + {-y}



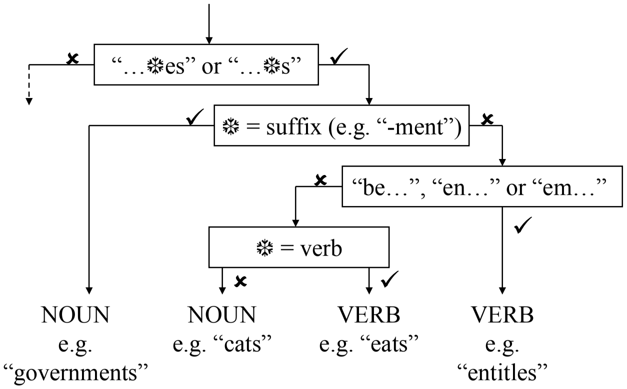
Use depth-first search: rapidly reject bad decompositions.

Include rules for possible mutations. E.g. a suffix beginning with a vowel can alter the end of the previous morpheme:

- Consonant doubling: “fitted” = {fit} + {-ed}
- “y” to “i” mutation: “cities” = {city} + {-es}
- Dropping of silent “e”: “choking” = {choke} + {-ing}

Parts of Speech

Morpheme dictionary entries give phonetic transcription + information on stress position and part of speech (e.g. noun, verb etc)



Letter-to-Phoneme Conversion

Only for words that cannot be decomposed into morphs (<2% of words)

- – Remove common prefixes and suffixes from the root
- – Check suffix/prefix compatibility:
  - “dictatorship” → “dict” + {-ate} + {-or} + {-ship} since
    - – {-ate} is ? → verb
    - – {-or} is verb → noun
    - – {-ship} is noun → noun
  - “astonishing” does not → “aston” + {-ish} + {-ing} since
    - – {-ish} is noun/adjective → adjective
    - – {-ing} is noun/verb → adjective/participle

Letter-to-Phoneme Rules

- – Convert consonant strings to phonemes
  - Convert the long clusters first e.g. “ngth” from “strength”
  - Look up pronunciation in a dictionary
- – Convert vowel strings to phonemes
  - Look up in a dictionary
  - Pronunciation can depend on neighbouring consonants or consonant strings
- – Apply pronunciation rules to prefixes and suffixes
  - Plurals depend on previous syllable (voiced/unvoiced/s) e.g. “dogs”, “cats”, “busses”
  - Past participles depend likewise (voiced/unvoiced/t/d) e.g. “laughed”, “showed”, “panted”, “ridded”
  - /t/ and /s/ can change to /ʃ/ e.g. “retention”, “compression”

Use rules to determine where to stress each word

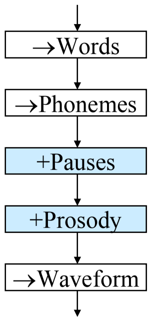
Parsing of Sentence

The sentence structure must be analysed to determine:

- which words are grouped together in phrases: we need a pause after such a group
  - *The satellite radio link transmitter* was green”
- Whether it is a statement, question or order: this affects the pitch contour
- To distinguish between noun/verb pairs with identical spelling
  - “I sur*vey* my kingdom” versus “ I make a *sur*vey of my kingdom”
  - “I ins*ult* you” versus “I give you an *in*sult”

The word-to-morph conversion indicates as a by-product which part of speech each word is. This greatly assists the parsing.

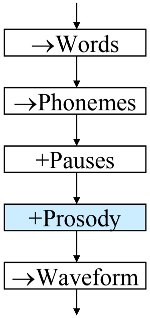
Pauses and Durations



Each phonemes has a standard duration. This is then modified:

- Insert pauses at punctuation
- Speed up at the beginning of long words and phrases
  - “*The wonderful wizard of Oz* hid behind the curtain”
- Speedup for consonant clusters
  - “He was strengthened by spinach”
- Pause after long phrases
- Pause before “and” when it joins two sentences but not when it joins two vowels:
  - “I was tired and I wanted to stay in bed”
  - “He was playing cat and mouse”
- Slow down at the end of sentences

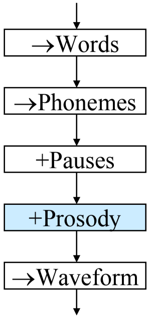
Stress & Rhythm



- Syllables are *stressed* (made louder) to:
  - Distinguish between otherwise identical words: survey, insult
  - Emphasise the important (= least predictable) words in a sentence:
    - "Speech processing is a *fasc*inating subject"
- Information on parts of speech (e.g. nouns/verbs) and importance of words comes from the parsing done earlier.

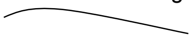
- *Rhythm*: stressed syllables tend to occur at regular time intervals:
  - "The *cat* that sat on the *mat* was *eat*ing a big brown *rat*"
  - Unstressed syllables are squeezed into the available time between stresses

Pitch

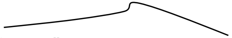


Pitch normally rises to the first stress of a
sentence, then
falls:

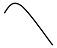
- "He *le*ft an hour ago"



- "How do you like it here"

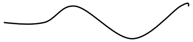


- "Wow"



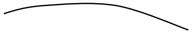
For an incomplete sentence it rises at the end:

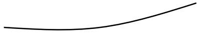
- "As I think about it …"



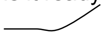
A Yes/No question rises at the end unless it begins with "Wh…" :

- "Today is Wednesday", "Today is Wednesday?"





- "Is it ready?"

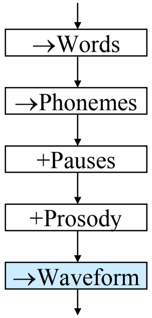


A local pitch rise occurs on all stresses.

Lecture 13

Synthesised Speech Generation

- Concatenative Synthesis
  - concatenating segments of recorded speech
    - **Diphone** synthesis has segments going from the middle of one phoneme to the middle of the next
    - **Demisyllable** synthesis has segments that form the first or second half of a syllable.
- Formant Synthesis
  - generating speech waveforms from scratch



Diphone Synthesis

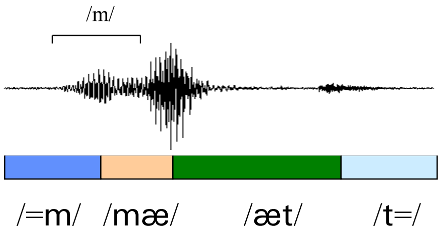
Uses actual recorded speech from a single individual

Stringing together phoneme segments sounds terrible

- get large discontinuities at the boundaries

Diphones go from the middle of one phoneme to the middle of the next

- “mat” = /=m/ + /mæ/ + /æt/ + /t=/
- The middle of a phoneme is chosen to be its most constant and consistent part.



Diphone Recording

43 phonemes ⇒ 43^2 = 1849 diphones.

- some phoneme pairs never occur in English words
- some phoneme pairs vary according to context ⇒ need several versions
- Typically need an inventory of 2000 to 3000 elements

Record all the diphones from a single speaker

Each diphone is placed in the middle of a nonsense word

- "ermater"

Need to vary the pitch, duration and amplitude of each phoneme to fit in with the prosody.

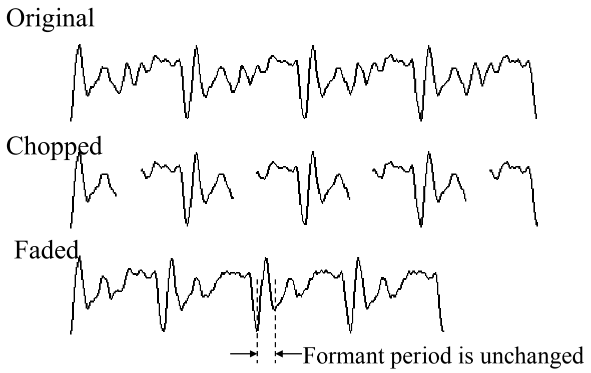
- We must not affect the formant frequencies
- ⇒ we use unaltered chunks of the waveform without stretching or shrinking the time axis.

Pitch Variation

Change the pitch by chopping out or duplicating bits from the lowest energy portion of each pitch cycle

To avoid a discontinuity, we fade gradually from one waveform segment to another

Altering the pitch affects the duration too: fix that next



Duration Variation

Change duration by duplicating or omitting entire pitch cycles:

- To increase the duration by 10% you would repeat every 10th pitch cycle
- To reduce the duration by 10% you would delete every 10th pitch cycle

Snip and join waveforms at the centre of pitch cycles where the energy is lowest.

For unvoiced segments of speech there is no pitch cycle

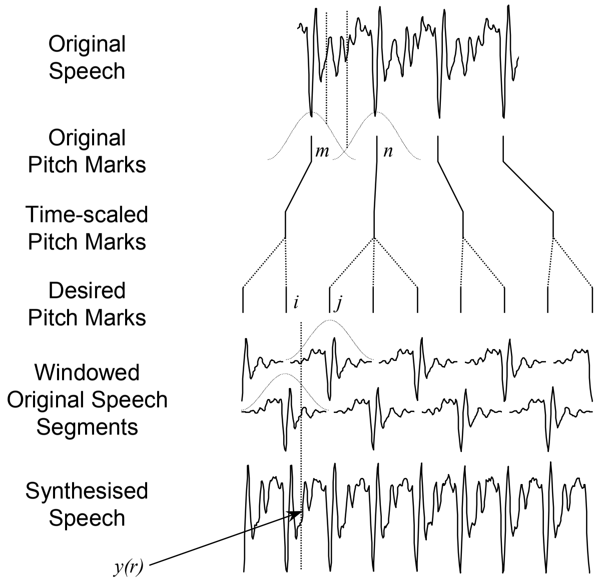
- Find a pseudo pitch by finding the time delay at which the speech is most correlated with itself.
- If there is very little correlation then it doesn’t matter where you snip it

Pitch-Synchronous Overlap-Add (PSOLA)

Detailed procedure for pitch and duration variation:

- Create *pitch marks* for the original speech at glottal closure instants (or every 10 ms for unvoiced speech)
- Translate them onto the new time axis according to whether the duration is to be increased or decreased
- Create new pitch marks for the synthesised speech at whatever fundamental frequency is desired
- Match each new pitch mark to the closest old pitch mark on the time-scaled axis. Each old pitch mark may be mapped onto one, several or none of the new pitch marks.
- For each new pitch mark create a Hanning windowed version of the speech. Window length is twice the minimum of the new and old pitch periods
- Add windowed speech segments together; normalise the energy by dividing each speech sample by the square root of the sum of the squares of the window amplitudes at the corresponding sample number: this assumes random phases in different segments.

PSOLA example



Original Speech

Original Pitch Marks

Time-scaled Pitch Marks

Desired Pitch Marks

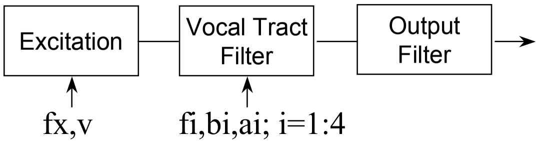
Windowed Original Speech Segments

Synthesised Speech

$$y(r)=\frac{s(m+r-i)w\big((r-i)/k\big)+s(n+r-j)w\big((r-j)/k\big)}{\sqrt{w^2\big((r-i)/k\big)+w^2\big((r-j)/k\big)}}$$

where $k=\min(j-i,n-m)$ and $w(x)=½\big(1+\cos(\pi x)\big)$

Formant Synthesis



From the phoneme sequence, we need to generate the time variation of the fundamental frequency (fx), voiced/unvoiced control (v) and the formant frequencies, bandwidths and amplitudes (fi, bi and ai). These are the *synthesizer control parameters*.

A few phonemes consist of a sequence of sounds (e.g. diphthongs and plosives), these are split into two or more *phonetic elements*.

Each phonetic element has a table of values and a *rank* value. The table of values lists five numbers for each of the synthesizer control parameters (e.g. f2):

Phoneme	Rank	Target	High	Low%	TH	TL
=	63	2100	0	100	600	0
w	10	750	350	50	40	100
e	2	2000	1000	50	40	40
l	11	950	450	50	0	60
z	20	1700	950	50	20	30

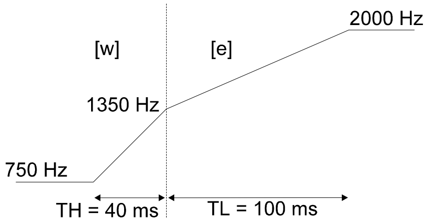
Target and High are in Hz, TH and TL are in ms

Phoneme Transitions

At each phoneme boundary, the ranks are compared. The value of a parameter (e.g. f2) at the boundary is:

HIGH + LOW% × target

where capitalised values come from the higher-ranked phoneme (normally a consonant). The transition times either side of the boundary are taken from the higher-ranked phoneme with TH/TL applying to the high and low rank side respectively. E.g. w→e



Note that short duration phonemes may never attain their targets.

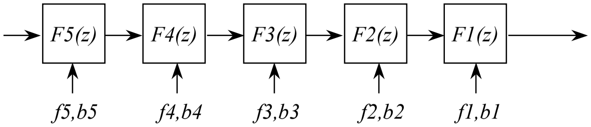
For each phoneme, we get two parameter tracks, one from each of the adjacent phonemes. The combined track is a weighted average of these two tracks with the weights varying linearly from (1,0) at the start of the phoneme to (0,1) at the end.

Cascade Vocal Tract Filter

For vowels, an all-pole filter models the vocal tract well.

We need one complex pole-pair for each formant. If the normalised frequency and bandwidth are f and b Hz, the pole positions are $\exp(-\pi b \pm 2j\pi f)$. We can implement the vocal tract as a cascade of 2nd order sections:

$$\frac{1}{(1-e^{-\pi b+2j\pi f}z^{-1})(1-e^{-\pi b-2j\pi f}z^{-1})}=\frac{1}{1-2e^{-\pi b}\cos(2\pi f)z^{-1}+e^{-2\pi b}z^{-2}}$$



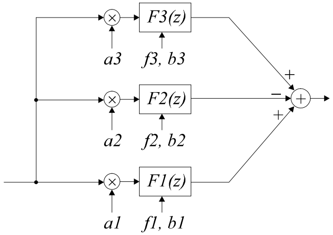
Note that the bandwidth and amplitude of the resonance peaks are not independent:

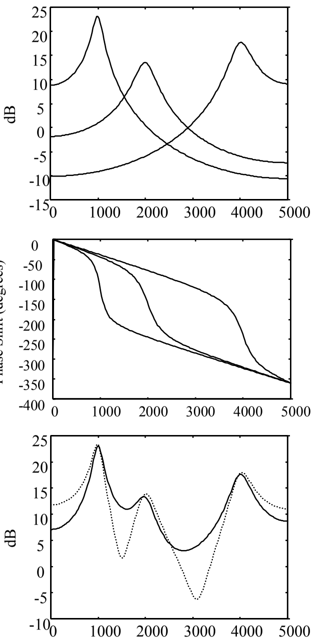
$$\frac{\text{gain at } f}{\text{gain at DC}} \approx \frac{1}{1-r_{pole}} = \frac{1}{1-e^{-\pi b}} \approx \frac{1}{\pi b}$$

Thus high bandwidth ⇒ low gain and vice-versa. In practice, the bandwidths are chosen to give the correct formant amplitudes.

Parallel Vocal Tract Filter

The fixed relationship between formant bandwidths and amplitudes does not hold for consonants. If we use a set of parallel filters instead of a cascade arrangement, we can control the amplitudes and bandwidths independently:





Each filter's phase lag goes up by 180° at resonance.

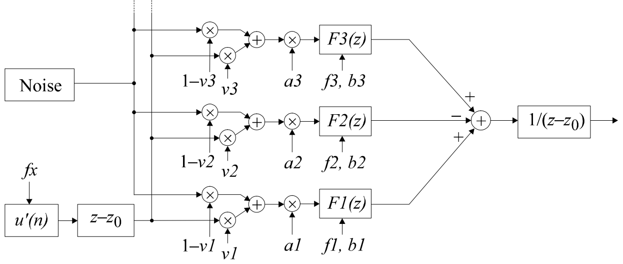
If filter outputs are added directly, the 180° shift gives deep nulls (dashed line)

If filter outputs are added with alternate signs we avoid this cancellation (solid line).

We can add zeros at z=±1 to each filter to improve formant independence.

Full Parallel Synthesiser

We can either have a cascade filter for vowels + a parallel filter for consonants or (simpler) use the same parallel filter for both.



The output deemphasis filter $1/(z–z_0)$ is chosen to give the correct overall spectrum for voiceless sounds. It is cancelled out for voiced sounds by a matching $(z–z_0)$. Choose $z_0 \approx 1–2\pi\times640/f_s$

For 8 kHz bandwidth eight filters are used: upper four can have fixed frequencies and bandwidths; bandwidths of lower four increase when glottis is open.

The v_i voicing controls are caculated by $v_i = 3(v–k_i)$ where ki is a constant for each filter that varies from 0 for f_1 up to 0.7 for f_8. This allows the low formants to be voiced while the upper ones are unvoiced (e.g. for /z/). We force $0\le v_i \le 1$.

Synthesiser is *capable* of giving speech indistinguishable from the real thing but the rules for driving it are not perfect.

Diphone versus Formant Synthesis

<u>Diphone Synthesis:</u>

- Uses real speech
- Much less computation
- Much easier to create the synthesiser
- Ignores allpohonic variations except for extreme cases: for these multiple diphones can be recorded for a particular phoneme pair
- Assumes the centre of a phoneme is independent of its context

<u>Formant Synthesis:</u>

- Uses rules to convert phonemes into sound
- Very hard to create the synthesiser
- Each phoneme is created independently and as a function of its entire context
- Allophonic rules can easily be included if you know what they are
- Smooth and natural speech is possible but requires very complicated rules