

Speech synthesis

Speech synthesis

- What is the task?
 - Generating natural sounding speech on the fly, usually from text
- What are the main difficulties?
 - What to say and how to say it
- How is it approached?
 - Two main approaches, both with pros and cons
- How good is it?
 - Excellent, almost unnoticeable at its best
- How much better could it be?
 - marginally

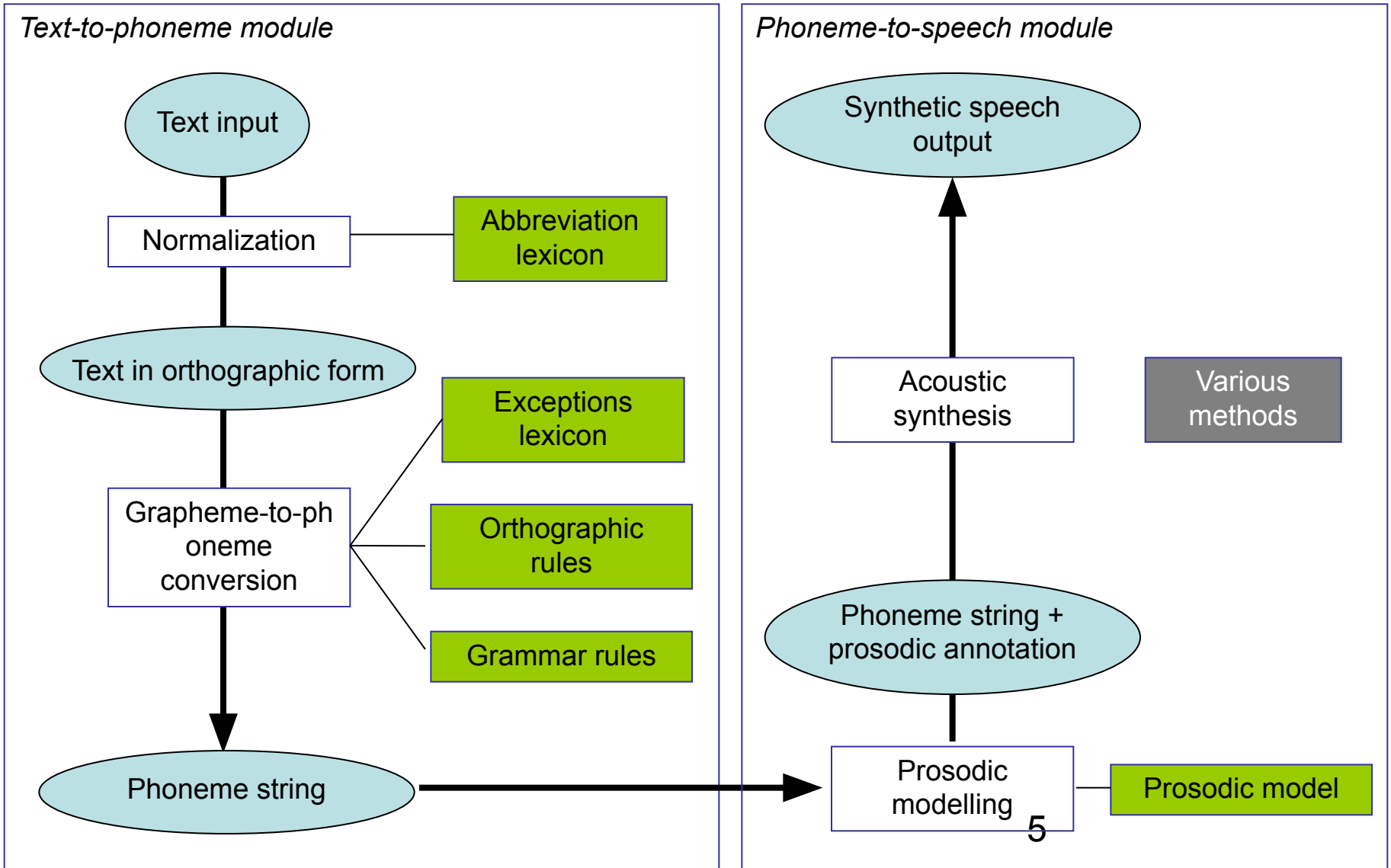
Input type

- Concept-to-speech vs text-to-speech
- In CTS, content of message is determined from internal representation, not by reading out text
 - E.g. database query system
 - No problem of text interpretation

Text-to-speech

- What to say: text-to-phoneme conversion is not straightforward
 - *Dr Smith lives on Marine Dr in Chicago IL. He got his PhD from MIT. He earns \$70,000 p.a.*
 - *Have you read that book? No I'm still reading it. I live in Reading.*
- How to say it: not just choice of phonemes, but allophones, coarticulation effects, as well as prosodic features (pitch, loudness, length)

Architecture of TTS systems



Text normalization

- Any text that has a special pronunciation should be stored in a lexicon
 - Abbreviations (Mr, Dr, Rd, St, Middx)
 - Acronyms (UN but UNESCO)
 - Special symbols (&, %)
 - Particular conventions (£5, \$5 million, 12°C)
 - Numbers are especially difficult
 - 1995 2001 1,995 □236 3017 233 4488

Grapheme-to-phoneme conversion

- English spelling is complex but largely regular, other languages more (or less) so
- Gross exceptions must be in lexicon
- Lexicon or rules?
 - If look-up is quick, may as well store them
 - But you need rules anyway for unknown words
- **MANY** words have multiple pronunciations
 - Free variation (eg *controversy*, *either*)
 - Conditioned variation (eg *record*, *import*, weak forms)
 - Genuine homographs

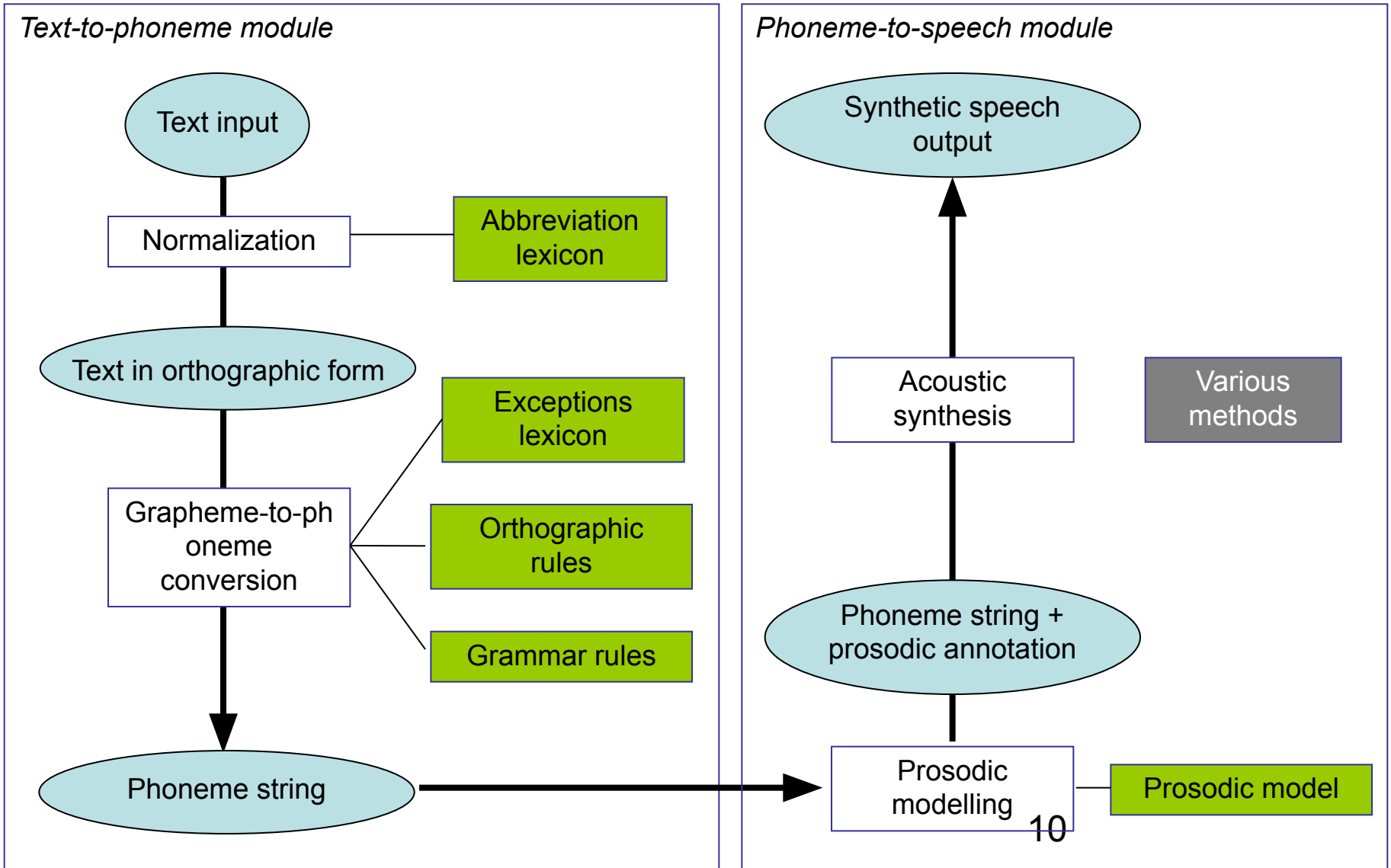
Grapheme-to-phoneme conversion

- Much easier for some languages (Spanish, Italian, Welsh, Czech, Korean)
- Much harder for others (English, French)
- Especially if writing system is only partially alphabetic (Arabic, Urdu)
- Or not alphabetic at all (Chinese, Japanese)

Syntactic (etc.) analysis

- Homograph disambiguation requires syntactic analysis
 - *He makes a record of everything they record.*
 - *I read a lot. What have you read recently?*
- Analysis also essential to determine appropriate prosodic features

Architecture of TTS systems



Prosody modelling

- Pitch, length, loudness
- Intonation (pitch)
 - essential to avoid monotonous robot-like voice
 - linked to basic syntax (eg statement vs question), but also to thematization (stress)
 - Pitch range is a sensitive issue
- Rhythm (length)
 - Has to do with pace (natural tendency to slow down at end of utterance)
 - Also need to pause at appropriate place
 - Linked (with pitch and loudness) to stress

Acoustic synthesis

- Alternative methods:
 - Articulatory synthesis
 - Formant synthesis
 - Concatenative synthesis
 - Unit selection synthesis

Articulatory synthesis

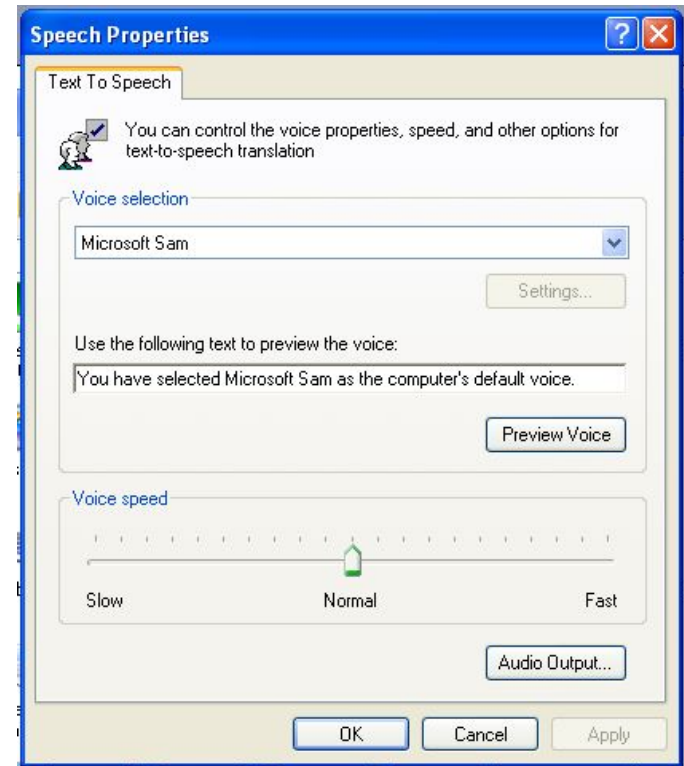
- Simulation of physical processes of human articulation
- Wolfgang von Kempelen (1734-1804) and others used bellows, reeds and tubes to construct mechanical speaking machines
- Modern versions simulate electronically the effect of articulator positions, vocal tract shape, etc.
- Too much like hard work

Formant synthesis

- Reproduce the relevant characteristics of the acoustic signal
- In particular, amplitude and frequency of formants
- But also other resonances and noise, eg for nasals, laterals, fricatives etc.
- Values of acoustic parameters are derived by rule from phonetic transcription
- Result is intelligible, but too “pure” and sounds synthetic

Formant synthesis

- Demo:
 - In control panel select “Speech” icon
 - Type in your text and Preview voice
 - You may have a choice of voices

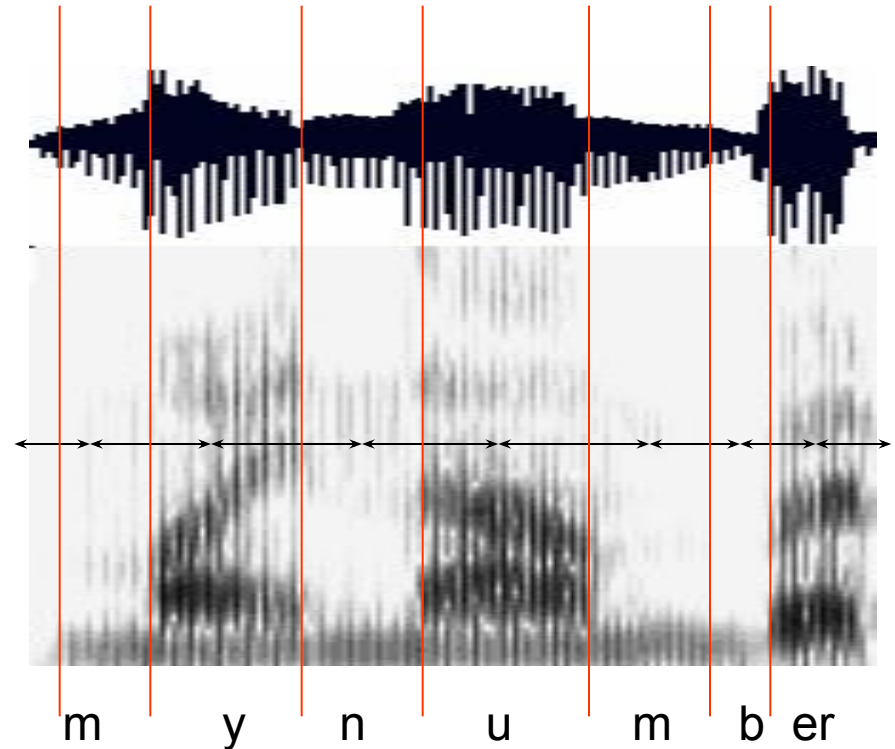


Concatenative synthesis

- Concatenate segments of pre-recorded natural human speech
- Requires database of previously recorded human speech covering all the possible segments to be synthesised
- Segment might be phoneme, syllable, word, phrase, or any combination
- Or, something else more clever ...

Diphone synthesis

- Most important for natural sounding speech is to get the transitions right (allophonic variation, coarticulation effects)
- These are found at the boundary between phoneme segments
- “diphones” are fragments of speech signal cutting across phoneme boundaries
- If a language has P phones, then number of diphones is $\sim P^2$ (some combinations impossible) – eg 800 for Spanish, 1200 for French, 2500 for German)



Diphone synthesis

- Most systems use diphones because they are
 - Manageable in number
 - Can be automatically extracted from recordings of human speech
 - Capture most inter-allophonic variants
- But they do not capture all coarticulatory effects, so some systems include triphones, as well as fixed phrases and other larger units (= USS)

Concatenative synthesis

- Input is phonemic representation + prosodic features
- Diphone segments can be digitally manipulated for length, pitch and loudness
- Segment boundaries need to be smoothed to avoid distortion

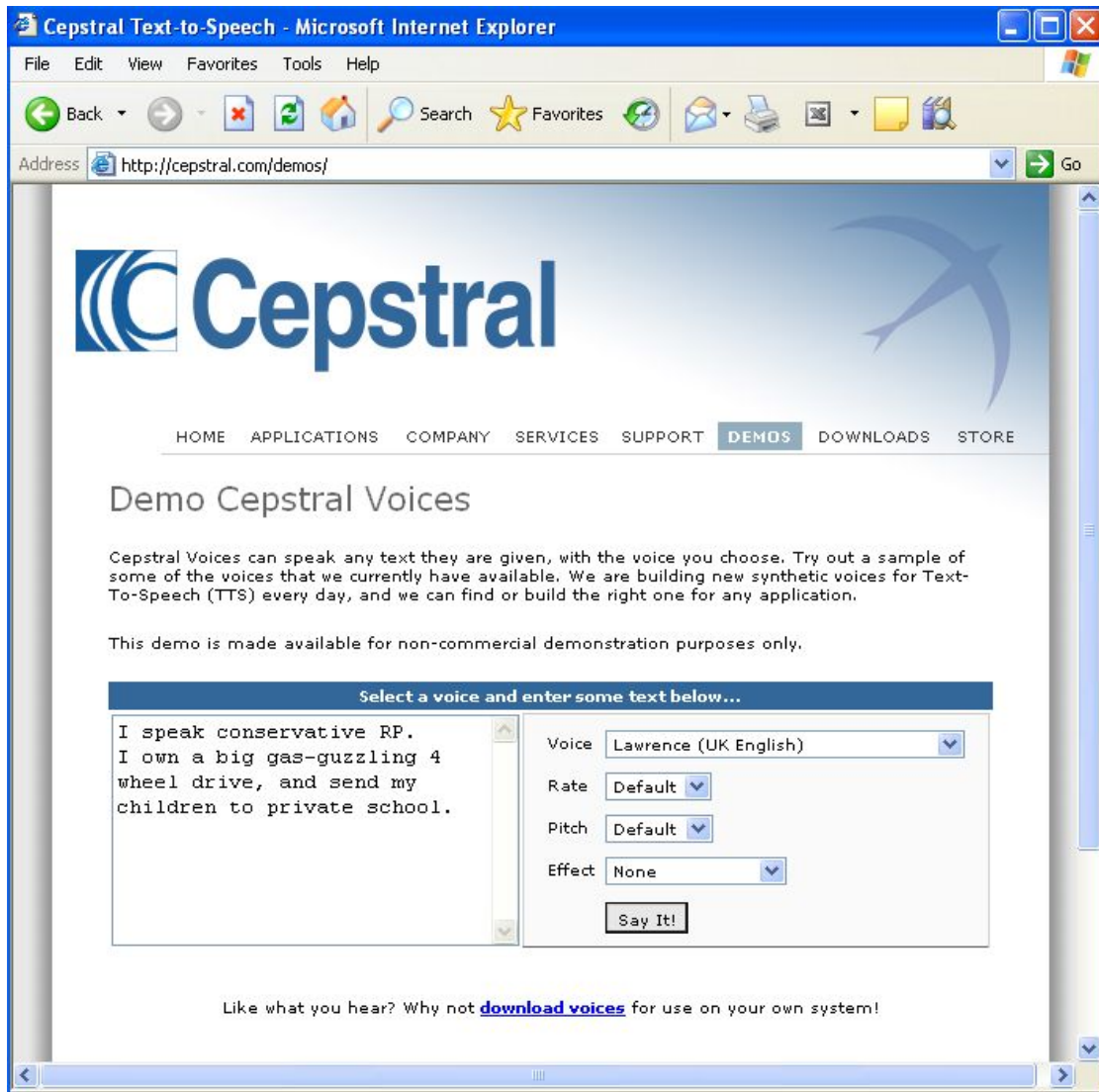
Unit selection synthesis (USS)

- Same idea as concatenative synthesis, but database contains bigger variety of “units”
- Multiple examples of phonemes (under different prosodic conditions) are recorded
- Selection of appropriate unit therefore becomes more complex, as there are in the database competing candidates for selection

Speech synthesis demo

The screenshot shows a Microsoft Internet Explorer browser window titled "AT&T Labs Text-to-Speech: Demo - Microsoft Internet Explorer". The address bar contains the URL "http://www.research.att.com/~ttsweb/tts/demo.php". The page header features the AT&T logo and the text "AT&T Labs, Inc. - Research". Below the header, the main heading is "Text-To-Speech (TTS) -- Our Demo Speaks Your Text". A navigation menu includes links for "Home", "> Demo", "FAQ", "Publications", "Contact", "Wizzard Software", "AT&T Natural Voices", and "TV Commercial". The main content area is a blue box with three steps: "STEP 1 Voice & Language:" with a dropdown menu set to "Audrey UK English"; "STEP 2 Text:" with a text input field containing "I hope you are enjoying this course unit, L E L A 3 O 4 3 1"; and "STEP 3 Click:" with "SPEAK" and "DOWNLOAD" buttons and a note "[restrictions apply*]". A speaker icon is visible in the bottom right corner of the browser window.

Speech synthesis demo



The screenshot shows a Microsoft Internet Explorer browser window titled "Cepstral Text-to-Speech - Microsoft Internet Explorer". The address bar displays "http://cepstral.com/demos/". The page features the Cepstral logo and a navigation menu with links for HOME, APPLICATIONS, COMPANY, SERVICES, SUPPORT, DEMOS, DOWNLOADS, and STORE. The main heading is "Demo Cepstral Voices". Below this, a paragraph explains that Cepstral Voices can speak any text with a chosen voice, and they are building new synthetic voices for Text-To-Speech (TTS) every day. A note states that the demo is for non-commercial demonstration purposes only. The interface includes a text input area with the sample text: "I speak conservative RP. I own a big gas-guzzling 4 wheel drive, and send my children to private school." To the right of the text area are four dropdown menus for "Voice" (set to "Lawrence (UK English)"), "Rate" (set to "Default"), "Pitch" (set to "Default"), and "Effect" (set to "None"). A "Say It!" button is located below these settings. At the bottom of the page, a link is provided to "download voices" for use on a personal system.