

Recent advances in automatic speech synthesis

Douglas O'Shaughnessy

INRS-EMT, Montreal, Canada

1. Introduction

Speech synthesis systems:

- simple playback of recorded human voices
- voice-response systems
- digitally-coded, but no loss of quality (unless too compressed)
- over telephone, bandwidth of 4 kHz, at 8000 samples/s (e.g., mu-law log-PCM at 64 kbps)
- for computer applications, rates of 10 000 and 16 000 samples/s
- perceptible energy in the strong sounds fades above 4 kHz
- higher bandwidth raises naturalness

2. Main issues

- Quality (intelligibility and naturalness) of the synthetic voice
- cost (complexity and size (“footprint”))
- delay (real-time or off-line)

3. Brief history

- source-filter model of human speech production:
 - speech modeled as a convolution of an excitation (noisy, periodic, or a combination) and a vocal-tract filter model
- excitation:
 - flat-spectrum noise,
 - line-spectrum of vocal-fold harmonics (pitch)

3.1 Spectral models

- Until the late 1990s, this was the basis for most text-to-speech, e.g., the Klatt formant model - a cascade and parallel combination of second-order digital resonators, each simulating one vocal-tract resonance.
- Also, articulatory models, emulating in three dimensions the form of the vocal tract, transformed into a time-varying digital filter

3.2 Waveform models

Simpler methods of generating synthetic speech:

- concatenating portions of natural speech waveforms directly
- avoids filtering and spectral representations
- early methods yielded unnatural quality
 - abrupt frequent abrupt discontinuities at unit boundaries
 - limited number of units

- Advances in the speed and memory of computers in the 1990s allowed synthesizers to have much larger numbers of stored waveform units – hundreds of thousands (vs. hundreds in the 1980s).
- Earlier waveform concatenation schemes had an inadequate inventory to choose from – leading to jumpy, unnatural speech.
- Newer methods minimize the risk that disparate units are joined together, leading to smoother speech.

Avoiding a synthetic excitation for a vocal-tract filter:

- very natural-sounding speech units are employed in modern synthesizers.

Concatenation synthesis can sound excellent, but weaknesses often appear, as one tries various diverse textual inputs.

4. Basic methodology

- many billions of allowable sentences possible
- stored speech units must be smaller than a sentence
- we concatenate ("join") small sections of actual speech to produce full utterances
- main issues in the joining and smoothing:
 - spectral and temporal continuity
 - prosody (intonation)

Assume text-to-speech (TTS) is the task

- In telephone voice-response systems, the TTS input derives from database text (numbers and words).
- Such data are usually arranged into phrases or sentences by a natural-language processor that assesses the user's wishes
- Each language has a set of approximately 30-50 basic sounds called phonemes (vowels and consonants)

- An average phoneme duration is approximately 80 ms, but varies greatly.
- If TTS joins phoneme-sized speech units, at approximately 12 units/s, we need to smooth at many boundaries per second.
- This is to emulate coarticulation (the vocal tract moves smoothly between phonemes)

5. Elements of the process

- long speech units retain the original coarticulation effects (most natural)
- smoothing methods used so far are simple, e.g., linear interpolation
- ideally, TTS should be able to produce synthetic speech emulating a wide range of human voices of varying styles.
- most systems produce a very limited set of voices, from professional speakers, who record hours of read speech, to produce the library of speech units

6. Trade-offs

- for small databases, record in one session, to get a uniform set of speech units
- uniformity simplifies transitions between units
- common smoothing technique is PSOLA (pitch-synchronous overlap-and-add), which adjusts speech segments of pitch periods to modify intonation
- units of different sizes: thousands of exemplar units for each phoneme (and short sequences of phonemes)

- try to capture the diversity of all coarticulation and prosody possible in speech
- “diphone”: basic unit for TTS that retains much coarticulation
 - last half of one phoneme + first half of the next phoneme

- extend to longer units: quad-phones like /stru/, which number in the millions.

If the database is poorly designed, one can clearly hear annoying discontinuities in the synthetic speech

- possible intermediate units are syllables and words:
 - Approximately 4400 syllables are needed to cover a large majority of English words (most frequent 1370 syllables are used about 93 percent of the time)
- difficult to achieve full coverage, as TTS input may have unpredictable words

- for each unit, one exemplar is normally inadequate, owing to contextual variations, often related to intonation.
- The quality of synthetic speech is often proportional to the average length of the units used to create the synthetic speech

- Uneven quality:
 - long sections of synthetic speech sound very good (corresponding to long units with little modification),
 - other sections have noticeable discontinuities (owing to significantly-modified units)

7. Details of waveform methods

- if synthetic speech simply chains unmodified concatenated units, it resembles two similar people speaking in alternating fashion
- limited modifications at unit boundaries (e.g., for 30-50 ms on each side), are may not be insufficient
- diphones as units presume that coarticulation is limited to about 50 percent of adjacent phonemes

- various aspects of intonation (duration, intensity, F0) must be changed from the stored unit form, according to the wider context of the desired text.
- exception: when units are always used in a particular context (e.g., credit card or telephone number)

- intonation has important roles in speech perception, for stress and grouping; e.g., misplaced stress in a word or phrase often degrades intelligibility.
- adjustments:
 - 1) amplitude (energy must change smoothly and be appropriate for the unit in context)
 - 2) pitch periods (fundamental frequency (F0))
 - 3) duration,
 - 4) other finer aspects of the stored unit

- As waveform methods cannot use spectral parameters (e.g., formants or LPC coefficients), non-intonational adjustments are often less well motivated

The goal of these modifications is to transform each selected stored unit to correspond more closely to what it would be in each phonetic context in normal speech.

8. Database design

- a major area of recent TTS research
- one compensates for a training speaker's variability by being selective in unit storage
- select "desirable" units to retain for the database
- should not simply discard units that correspond to rare sequences
 - combined likelihood of all rare events is a significant percentage of all speech sequences.

8.1 Unit selection synthesis

- large database of recorded speech
- each training utterance: divided into phones, syllables, words, phrases, sentences, as well as other intermediate units
- sometimes with a "forced alignment" speech recognizer (some manual post-correction)
- labeled with the fundamental frequency (pitch), duration, position in the syllable, and neighboring phones

- at runtime, weighted decision tree chooses right unit
- databases can range to gigabytes of recorded data
- dozens of hours of speech

9. Joining of units

- each “join” is evaluated by some similarity or distance measure (the “join cost”), so as to minimize discontinuities.
- successive units can be chosen from diverse sources (varying in time, intonation, and style)

- A suitable distance measure would help both in choosing the units and in their adjustment
- however, most such measures are spectral
- such distances must be repeatedly calculated for each tentative unit in a given context
- need to minimize the needed calculation

- two types of cost: selection cost and join
 - 1) how well the chosen unit matches the specified need
 - 2) how well it smoothes into the flow of the synthetic speech

- 1) likely relates to intelligibility, while 2) controls more the continuity or naturalness of the speech

10. Text processing for TTS

- As the TTS input is text, we must transform text into linguistic parameters, to access the speech database
- natural language processing (NLP) determines
 - which phonemes to pronounce
 - which syllables to stress
 - where to place breaks and pauses
 - other relevant linguistic information
- text processing is language-dependent

- For simple voice-response synthetic speech of sentences, phrases, or individual words, text processing for TTS is a mere table look-up
- intonation (and nuances in phonemic variations): specific to each language
- NLP: text-to-phoneme conversion ("letter-to-sound" rules)

- large computer dictionary with an entry for each word in the desired language
 - word's pronunciation (including which syllables to stress)
 - its syntactic category (e.g., part-of-speech)
 - possible semantic information
- input texts often have words not found in most dictionaries (e.g., mistyped, foreign, newly invented words, and names)

- For some languages (e.g., Korean, Spanish, Chinese), text-to-phoneme task is simple
- others (e.g., Italian, Finnish, German): small set of pronunciation "rules"
- English needs hundreds of such rules
- relatively straightforward task

Name pronunciation: difficult and important

- some systems work by analogy

11. Intonation

- assign intonation - pitch (F0), duration, and intensity - based on an analysis of each input text
- can store suitable speech units with their corresponding intonations
- often useful to record a limited, but large, set of common phrases with their appropriate intonations

- duration varies in ways that are complex, language-dependent, dialect-dependent, and likely even speaker-dependent.
- changes in durations are non-linear, e.g., vowels and stressed syllables tend to expand more than consonants and unstressed syllables do
- much of TTS is distinguishable from human speech owing to unusual durations in synthetic speech

Fundamental frequency: most difficult to predict, owing to its large variation.

- In many languages (including English), lexically stressed syllables of emphasized words have significant F0 changes, which also cue the boundaries of syntactic phrases.
- Many languages have F0 rise at the end of a question asking for a yes-or-no answer
- text has few indicators for syntactic and semantic information, which has great influence on intonation.

- Syntactic boundaries have major effects, e.g., speakers tend to slow their speech right before a major syntactic boundary
- sentences often have many words with no internal punctuation other than commas
- commas do not correspond well to intonation
- recent research: "expressive" synthetic speech with emotions
- most TTS systems output a straightforward, neutral voice

12. Examples

- From SoftVoice, Inc.



- From Ivona:



- How much wood would a woodchuck chuck if a woodchuck could chuck wood? He would chuck, he would, as much as he could, and chuck as much wood as a woodchuck would if a woodchuck could chuck wood.

13. Conclusion

- A major application of the future will be speech-to-speech translation,
- in which a person speaking in one language will be able to converse automatically with someone using another language