Speech Synthesis – Status Quo and Possible Future

Juergen Schroeter
AT&T Labs – Research
http://www.research.att.com/info/jsh

With significant help from the authors of:


Introduction

• **Goal**: Convert arbitrary textual messages to intelligible and natural speech so as to transmit information from a machine to a person; text-to-speech (TTS)

• **Methodology**: Exploit acoustic representations of speech for synthesis; exploit linguistic analyses of text to extract correct pronunciations and prosody of words in context

• **Synthesis Evaluation**: Accuracy of text rendering/pronunciation; intelligibility of resulting voice messages; perceived naturalness of resulting speech

• **Applications**: Automated Telecom Services (e.g., name and address rendering); Network voice server for email, FAX; text previewer for documents; aid in providing information from a machine (directory assistance, business locator service, banking services, help lines)
How TTS works

TTS Engine

Text Analysis
- Document Structure Detection
- Text Normalization
- Linguistic Analysis

Phonetic Analysis
- Homograph disambiguation
- Grapheme-to-Phoneme Conversion

Prosodic Analysis
- Pitch & Duration Attachment

Speech Synthesis
- Voice Rendering

Raw text or tagged text

Tagged text

Tagged phones

Controls

Synthetic Speech out

October 8, 2003 J. Schroeter, Speech Synthesis Positioning Paper VG 3
Where are we?

- Commercial TTS systems have come a long way towards synthesizing highly intelligible, natural sounding speech.
- Non-experts believe that speech synthesis is a “solved problem”.
- Fact is, we can produce “perfect” synthetic speech for very limited applications.
- Shortcomings include:
  - Coverage:
    - Acoustic variability (“holes” in the databases)
    - Pronunciation (e.g., proper nouns/names)
    - Prosody (e.g., “wh” and “yes/no” questions don’t sound like questions)
    - Emotions (other than through recording of separate databases for each)
  - Lack of Control:
    - Prosody (e.g., for conveying discourse, structure, intent, …)
    - Voice Quality, Emotions, Expressiveness
  - Lack of Scientific Usability:
    - Latest synthesizers cannot be used to test theories of Speech Production
How did we get here?

Early Vocoder Technology, manually controlled:
(1939 New York City World’s Fair, The Voder, AT&T)

Formant Synthesis:
(1979 MITtalk)

Concatentative Speech Synthesis (Diphones):
(1985 Bell Labs)

Unit Selection Synthesis:
(2002 AT&T Natural Voices)

Voice Production Pipeline, Special Domain Voices/Synthesizers,
“Signal Processing hurts!” vs. “Lack of Coverage” and “Rare Events”
“Superb Naturalness” vs. “Lack of Fine Control”
Progress in Synthesis

• **Unit-Selection Synthesis** (USEL) can be viewed as an extension of (or progression from) earlier Concatenative Synthesis paradigms.

• USEL is enabled by the increased power of computers and their decreased cost.

• USEL leverages progress made in text analysis & normalization, prosody, and speech input technologies.

• USEL makes use of many of the same algorithms used in ASR:
  – Forced alignment recognition for automatic phonetic labeling
  – Fast search techniques (e.g., FSMs) for unit selection

• Natural Language Understanding is on the verge of introducing “smarts” into TTS systems.
Where do we need to be?

• Depends on our point of view:
  – If we are striving for perfection, we need to pass the Turing test
  – If we have a specific limited application in mind, we need to be just “good enough”

• Questions to ask:
  – How do we evaluate TTS systems? TTS components?
  – How can one maintain high naturalness and allow fine control of prosody and voice quality (“emotions”)?
  – How do we make TTS systems smarter so they seem to understand what they are talking about?
  – How can we transform a large speech database of one speaker to sound like a small speech database of another speaker?
  – How can we take advantage of the vast amount of speech-related knowledge in form of better models for text analysis and pronunciation, prosody, and synthesis?
Key Technical Challenges

• Pronunciation of Rare and New words (e.g., names)
• Prosody, in particular, conveying document structure
• Better ASR tools: Phonetic and Prosodic Auto-Labeling
• Extension to new languages
• Emotions and Expressiveness
• Assuring Acoustic Coverage (or creating it on the fly!)
• Personalization (fast voice creation from minimal data)
• Capture and Synthesis of Paralinguistic Information (for speech-to-speech translation systems)
• Evaluation
• Modularity (to enable collaborative efforts)
• Standards
• Multimodal Extensions
Pronunciation of rare and new words (e.g., names)

- Each day, new names (words) appear in the media. Automatic (zero human-touch) updates to pronunciation dictionaries is a dream.


Prosody, in particular, conveying document structure

• To my knowledge, at present, no TTS system’s prosody module takes into account document structure beyond the current paragraph. Good starting points would be:


• Prosody, as related to speaking style:
Better ASR tools: Phonetic and Prosodic Auto-Labeling

• Performance of Phonetic Aligners benefit from Speech Knowledge


• Prosodic Auto-Labelers are still in their infancy

Extension to New Languages

• There is a chicken-and-egg problem:
  – in order to create TTS systems in new languages quickly, we need ASR tools in the target language
  – in order to create an ASR system in the target system, we need (at least) a pronunciation dictionary and enough transcribed speech for ASR model training


Excursion: Emphasized vs. Non-Emphasized Words


- Examples of raw recordings in database:
  - He **did** then **know** what **had** occurred.
  - **Tarzan** and **Jane** raised **their** heads.

- Example renderings with explicit use of emphasized/non-emphasized tags in Unit Selection
  - This is a short example.
  - **This** is a short example.
  - This **is** a short example.
  - This **is** a short example.
  - This **is** a short example.
  - This is a **short** example.
  - This is a short **example**.
Emotions and Expressiveness

• “Brute force” method is to record databases according to target emotion/expression; problem of eliciting the desired speech data

• Research focus on defining “targets”
Assuring Acoustic Coverage

- Database text / utterance selection / pruning

- Assuring consistency over multi-day recordings

- Which type of units to use (in context)
  **half-phonemes:** A. D. Conkie, “Robust Unit Selection System for Speech Synthesis,” in: Joint Meeting of ASA, EAA, and DAGMA, paper 1PSCB_10, Berlin, Germany, 15-19 Mar., 1999.

- Cautious use of Signal Processing
Choice of units

- One listening test experiment compared phones versus diphones as units for synthesis.
- Each set of stimuli was made from the same database and used the same program.
- Only the cost functions were different.
Unit Selection with Half Phonemes

- Transitional (concatenation) costs are based on acoustic distances [arrow line width]
- Node (target) costs are based on linguistic id of unit [node circle line width]

Optimal Path
Personalization (fast voice creation from minimal data)

- Soon-to-be vocally impaired people like to have TTS speak in their old, healthy, voice
- Ultimately, enabling “Kmart”-style creation of TTS in any given voice, will need a solution to the problem of Voice Transformation
- The idea is to transform a large, existing, database of one (source) voice so it sounds like the smaller set of utterances from another (target) voice
- Issues on all time scales (frame, phoneme, syllable, word, utterance,...)

Female voice:  
Child voice:  

- Linguistic aspects covered in several papers, e.g.,

  
Capture and Synthesis of Paralinguistic Information (e.g., for speech-to-speech translation systems)

• Extra-linguistic Information
  – Gender, age, health

• Paralinguistic Information
  – Intensions, Attitudes, Emotions, and their influence on meaning

Evaluation

- A “hot” topic that is not even close to an accepted general solution.
- Rule of thumb: “Evaluate TTS (components) as close as possible to the intended application and as close as necessary to the module under test.”


Modularity (to enable collaborative efforts)

- Synthesis systems are either “pipelined” (left-to-right sequence of modules) or “hubbed” (loose connection of expert modules that are controlled by scripts)
- For competitive reasons, there is little interest in the industry to standardize on the interfaces between modules and between the TTS engines and the voice databases
- “Festival” and “Euler” were conceived to allow focused research on one or a few modules (and reuse the rest)

http://www.cstr.ed.ac.uk/projects/festival/
http://www.tcts.fpms.ac.be/synthesis/euler/

Standards

- Standards to enable “best of breed” choice for application developers are concerned with text input (interface, mark-up, etc.)
  
  
  
http://www.w3.org/TR/speech-synthesis/

- Simple standard interfaces: VXML, SALT

- In the future, it will be important to create standards compliant front ends that include tags that cover
  - Semantic concepts; vital for Dialog Systems
  - Issues important for translation systems (stress/emphasis; emotions; speech acts)
Multimodal Extensions

- The same way that Natural Language Speech Input Systems (ASR, NLU, DM) are being extended to include non-speech input (pen, graphical), Speech Output systems (i.e., TTS) will be extended to include visual output (“Visual TTS”, VTTS, i.e., “talking heads”; graphical output like highlighting map features, text, etc.)


TTS Future: 5 years out

• Kmart-style voice creation
  – Necessitates more, and more efficient, tools for streamlining the mechanics of recording voices

• Make better use of existing databases in unit-selection
  – Cost functions not yet perceptually based (vowels, consonants)
  – Better trade-offs between dimensions of “quality”, based on solid science
  – Improve the “hit” rate, reduce the “miss” rate
  – More reliable prosody

• Investing in recording large speech and text databases
  – Learning from examples (e.g., emotions, and their acoustic correlates)
  – Dialects, dialect transformations (phonological aspects, signal processing)
  – Identify “holes” in coverage and fill them

• Making use of signal-processing
  – fast and good enough to fill “holes” in the databases (e.g., change pitch) without degrading channel density and quality
Summary and Conclusions

• Unit-Selection Synthesis has caused a sea-change in the field towards more natural-sounding speech synthesis
• Text-to-Speech Synthesis Research is far from “done”, however.
• We have outlined important challenges that lie ahead to create general, flexible, and efficient solutions
• Speech knowledge is still key to help us get there…