

Speech Synthesis – Status Quo and Possible Future

Juergen Schroeter

AT&T Labs – Research

<http://www.research.att.com/info/jsh>

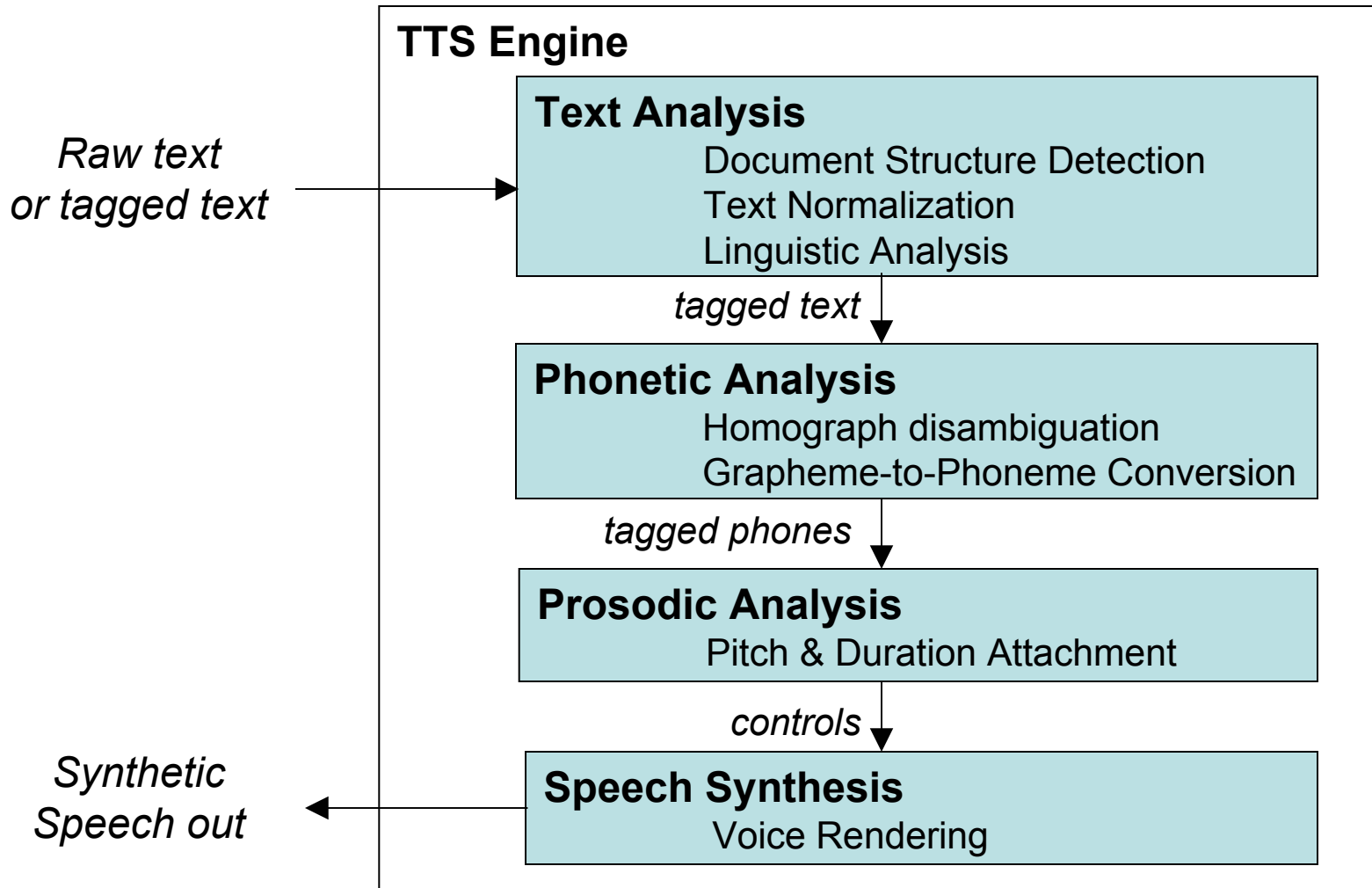
With significant help from the authors of:

- **A. W. Black**, “Perfect Synthesis for all of the People all of the Time,” Keynote Paper in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.
- **G. Baily, N. Campbell, B. Möbius**, “ISCA Special Session: Hot Topics in Speech Synthesis,” in: Proc. Eurospeech 2003, Geneva, 37-40, Sept. 1-4, 2003.
- **J. Schroeter, A. Conkie, A. Syrdal, M. Beutnagel, M. Jilka, V. Strom, Y.-J. Kim, H.-G. Kang, and D. Kapilow**, “A Perspective on the Next Challenges for TTS Research,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

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



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

M. F. Spiegel, “Proper Name Pronunciations for Speech Technology Applications,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

L. Galescu, J. F. Allen, “Pronunciation of Proper Names with a Joint N-Gram Model for Bi-Directional Grapheme-to-Phoneme Conversion,” in: Proc. ICSLP2002, Denver, Colorado, Session TuA4p.4, 109-112, 16-20 Sept. 2002.

J. Meron, “Using Rules to Improve Letter to Sound Conversion of Names,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, “Normalization of non-standard words.” *Computer Speech and Language*, 15(3), 287–333, 2001.

R. Sproat, “Pmtools: A pronunciation modeling toolkit”, Paper 104 in: Fourth ISCA ITRW on Speech Synthesis, August 29 - September 1, 2001, Perthshire, Scotland, 2001.

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:

J. Hirschberg, "Communication and Prosody: Functional Aspects of Prosody," Speech Communication: Special Issue on Dialogue and Prosody: 36:31–43, ed. J. Terken and M. Swerts, 2002.

J. Hirschberg and C. Nakatani, "Acoustic Indicators of Topic Segmentation," ICSLP-98, Sydney, 1998.

J. Venditti and J. Hirschberg, "Intonation and Discourse Processing," Proceedings of ICPHS 2003, Barcelona.

A. W. Black, N. Campbell, "Predicting the intonation of discourse segments from examples in dialogue speech," in: Proceedings of ESCA Workshop on Spoken Dialogue Systems, Vigso, Denmark, 1995, pp. 197-200.

- **Prosody, as related to speaking style:**

J. Hirschberg, "Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech," Proceedings of ICPHS-95, Stockholm, August, Vol. 2, pp. 36–43, 1995.

Better ASR tools: Phonetic and Prosodic Auto-Labeling

- Performance of Phonetic Aligners benefit from Speech Knowledge

S. Greenberg, “Strategies for Automatic Multi-Tier Annotation of Spoken Language Corpora,” in: Proc. Eurospeech 2003, Geneva, 45-48, Sept. 1-4, 2003.

J.-P. Hosom, “Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling,” in: Proc. ICSLP2002, Denver, Colorado, Session TuB9p.15, 357-360, 16-20 Sept. 2002.

J. Kominek, C. L. Bennett, A. W. Black, “Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis,” in: Proc. Eurospeech 2003, Geneva, 313-316, Sept. 1-4, 2003.

Y.-J. Kim, A. Conkie, “Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction,” in: Proc. ICSLP2002, Denver, Colorado, Session TuA4p.13, 145-148, 16-20 Sept. 2002.

- Prosodic Auto-Labelers are still in their infancy

A. K. Syrdal, J. Hirschberg, J. McGory, and Mary Beckman, Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody, *Speech Communications* **33**, 135-151, 2001.

I. Bulyko and M. Ostendorf, “A Bootstrapping Approach to Automating Prosodic Annotation for Constrained Domain Synthesis,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

V. Strom, “From Text to Prosody Without ToBI,” in: Proc. ICSLP2002, Denver, Colorado, Session ThC46o.3, 2081-2084, 16-20 Sept. 2002.

X. Ma, W. Zhang, Q. Shi, W. Zhu, L. Shen, “Automatic Prosody Labeling Using Both Text And Acoustic Information,” in: Proceedings ICASSP 2003, Paper P4.8, Hong Kong, 2003.









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

A. W. Black and A. F. Llitjós, “Unit Selection without a Phoneme Set,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

J. Tian, J. Häkkinen, O. Viikki, “Multilingual Pronunciation Modeling for Improving Multilingual Speech Recognition,” in: Proc. ICSLP2002, Denver, Colorado, Session TuC11p.3, 497-500, 16-20 Sept. 2002.

Excursion: Emphasized vs. Non-Emphasized Words

- (Borrowed from: **A. W. Black**, “Perfect Synthesis for all of the People all of the Time,” Keynote Paper in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.)
- 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**. 

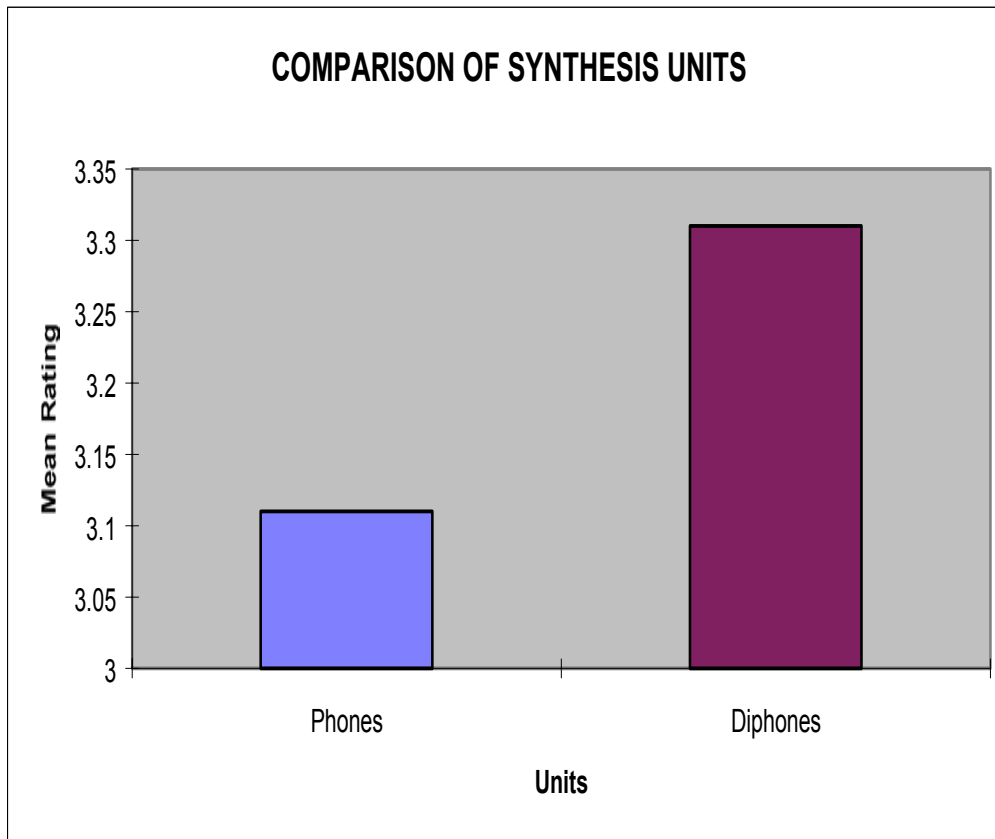
Emotions and Expressiveness

- “Brute force” method is to record databases according to target emotion/expression; problem of eliciting the desired speech data
G. Dogil, and B. Möbius, "Towards a model of target oriented production of prosody," in: Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark), 1:665-668, 2001.
V. Aubergé, N. Audibert, A. Rilliard, “Why and How to Control the Authentic Emotional Speech Corpora,” in: Proc. Eurospeech 2003, Geneva, 185-188, Sept. 1-4, 2003.
N. Campbell, “Towards a Grammar of Spoken Language: Incorporating Paralinguistic Information,” in: Proc. ICSLP2002, Denver, Colorado, Session TuC15p.14, 673-676, 16-20 Sept. 2002.
- Research focus on defining “targets”
L. Yang, N. Campbell, "Linking form to meaning: The expression and recognition of emotions through prosody", Paper 141 in: Fourth ISCA ITRW on Speech Synthesis, August 29 - September 1, 2001, Perthshire, Scotland, 2001.
C. Gobl, E. Bennett, and A. N. Chasaide, “Expressive Synthesis: How Crucial is Voice Quality?” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.
M. Bulut, S. S. Narayanan, A. K. Syrdal, “Expressive Speech Synthesis Using a Concatenative Synthesizer,” in: Proc. ICSLP2002, Denver, Colorado, Session WeC28o.2, 1265-1268, 16-20 Sept. 2002.

Assuring Acoustic Coverage

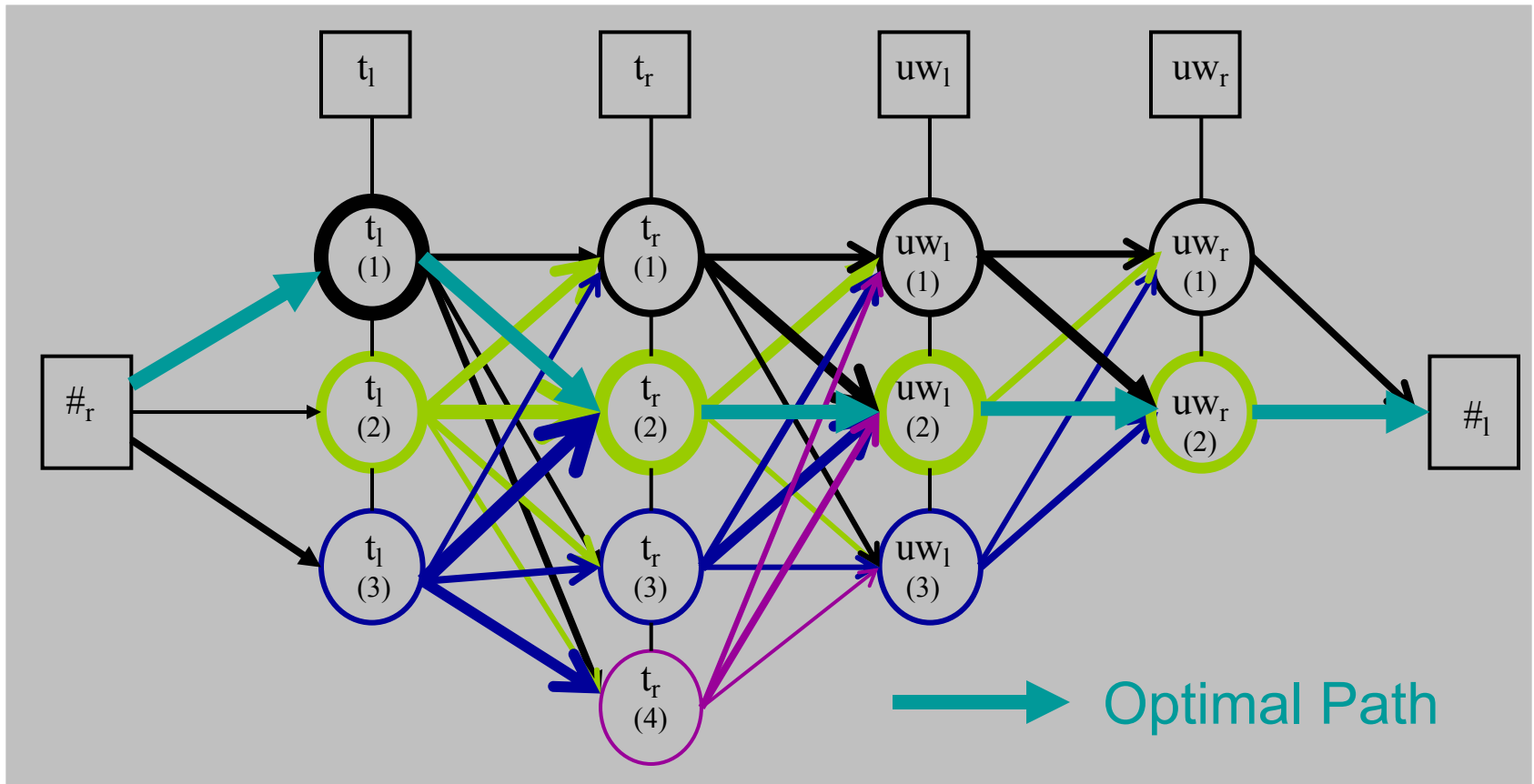
- **Database text / utterance selection / pruning**
A. W. Black, K. A. Lenzo, "Optimal data selection for unit selection synthesis", Paper 129 in: Fourth ISCA ITRW on Speech Synthesis, August 29 - September 1, 2001, Perthshire, Scotland, 2001.
- **Assuring consistency over multi-day recordings**
H. Kawai and M. Tsuzaki, "A Study on Time-Dependent Voice Quality Variation in a Large-Scale Single Speaker Speech Corpus used for Speech Synthesis," in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.
- **Which type of units to use (in context)**
phonemes: **A. Hunt and A. W. Black**, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," in Proc. ICASSP-96, pp. 373-376, 1996.
half-phonemes: **A. D. Conkie**, "Robust Unit Selection System for Speech Synthesis," in: Joint Meeting of ASA, EAA, and DAGA, paper 1PSCB_10, Berlin, Germany, 15-19 Mar., 1999.
variable-length units: **B. Bozkurt, T. Dutoit, R. Prudon, C. D'Alessandro, V. Pagel**, "Improving Quality of Mbrola Synthesis for Non-Uniform Units Synthesis," in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.
HMM-states: **J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi**, "A Context Clustering Technique for Average Voice Model in HMM-Based Speech Synthesis," in: Proc. ICSLP2002, Denver, Colorado, Session TuA4p.10, 133-136, 16-20 Sept. 2002.
some comparisons, e.g.: **S.P. Kishore, A. W. Black**, "Unit Size in Unit Selection Speech Synthesis," in: Proc. Eurospeech 2003, Geneva, 1317-1320, Sept. 1-4, 2003.
- **Cautious use of Signal Processing**
J. Wouters and M. Macon, "Spectral modification for concatenative speech synthesis," in: Proc. ICASSP, Istanbul, Turkey, 941-944, 2000.
J. van Santen and X Niu, "Prediction and Synthesis of Prosodic Effects on Spectral Balance of Vowels," in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

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]

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.,

B. J. Williams and S. Isard, "A Keyvowel Approach to the Synthesis of Regional Accents of English," in: Proc. Eurospeech 97, 1997.

C. Olinsky and F. Cummins, "Iterative English Accent Adaptation in a Speech Synthesis System," in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

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

N. Campbell, “Towards a Grammar of Spoken Language: Incorporating Paralinguistic Information,” in: Proc. ICSLP2002, Denver, Colorado, Session TuC15p.14, 673-676, 16-20 Sept. 2002.

E. Eide, “Preservation, Identification, and Use of Emotion in a Text-to-Speech System,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

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.”

<http://www.slt.atr.co.jp/cocosda/jenolan/> (1998)

Y. V. Alvarez, M. Huckvale, “The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-Speech Systems,” in: Proc. ICSLP2002, Denver, Colorado, Session TuB9p.8, 329-332, 16-20 Sept. 2002.

A. Rilliard, V. Aubergé, “Prosody evaluation as a diagnostic process: subjective vs. objective measurements”, Paper 140 in: Fourth ISCA ITRW on Speech Synthesis, August 29 - September 1, 2001, Perthshire, Scotland, 2001.

M. Viswanathan and M. Viswanathan, “Comparison of Measures of Speech Quality for Listening Tests of Text-to-Speech Systems,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

J. Xu, C. Guan, and H. Li, “An Objective Measure for Assessment of a Corpus-Based Text-to-Speech System,” in: Proceedings, IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

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/>

A. W. Black, A., and K. A. Lenzo, “Building voices in the Festival speech synthesis system,” <http://festvox.org> , 2000.

Standards

- Standards to enable “best of breed” choice for application developers are concerned with text input (interface, mark-up, etc.)

P. A. Taylor and A. Isard, "SSML: A Speech Synthesis Markup Language", *Speech Communication*. 21, pp. 123-133, 1997.

R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, "SABLE: A Standard for TTS Markup", in *Third ESCA Workshop on Speech Synthesis*, pp. 27-30, 1998.

<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.)

E. Cosatto, J. Ostermann, H. P. Graf, J. Schroeter, “Lifelike Talking Faces for Interactive Services,” Proceedings of the IEEE, Vol. 91, No. 9, 1406-1429, 2003.

<http://www.research.att.com/projects/MultimodalAccessToCityHelp/>

G. Bailly, “Audiovisual Speech Synthesis. From Ground Truth to Models,” in: Proc. ICSLP2002, Denver, Colorado, Session Spec5Ao.3, 1453-1456, 16-20 Sept. 2002.

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...