Speech recognition and synthesis

Introduction

R&D in general purpose TTS systems is almost completely directed towards concatenative synthesis. Special purpose systems for speech research, visual speech generation, and small footprint applications still use Articulatory Synthesis or rule based systems

Developping concatenative TTS systems

- A strength is that it produces natural sounding speech from recorded human speech
- A weakness is that its quality totally depends on the qualities of the original recorded voice
- Evaluation must separate voice characteristics and system characteristics

[Boersma and Weenink(2004), Möhler(2005), Black and Lenzo(2003b)]

Recording a voice: Speaker selection

Characteristics of a "good speaker" TM

- Availability and willingness (long recording times)
- Clear voice
- Consistent speaking (variability is bad)
- Will form the personality of the synthesis
- Will sign over all rights to you:
  - free for any use
  - free to distribute to anyone but cannot be used for commercial purposes without further contract.
  - research use only (does this allow public demos?)
  - fully proprietary
- Note: The style of speaking determines the style of the synthesis

[Black and Lenzo(2003b)]

Recording a voice: Diphone database

Diphone lists (≈1600 diphones)

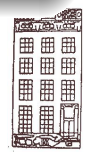
- Choose phoneset
- Construct diphone list in nonsense words, eg [*pau t aa b aa b aa pau*]
- Add special or foreign phonemes and clusters
- Synthesize prompts as sounds for presentation
  - Text is ambiguous
  - Consistent prosody
  - Consistent pronunciation
- Record words under the best of circumstances
- Label and align phones (automatically)
- Extract pitch marks (electroglottogram)
- Build parameter files
- Build and test database itself

[MBROLA(2005)]

Recording a voice: Unit database

Unit selection TTS is based on general speech, prosodic variation is good

- Size: phone, diphone, demi-syllable
- Type: phone, phone+stress, phone+word
- Concatenate units "in context", eg, stressed vs unstressed or word-initial vs -final phones
- Select units that fit requirements best
- Could use general speech corpus, but this generally lacks coverage and consistency
- Best to record a specially designed database

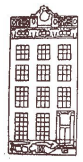


Recording a voice: Constructing a unit database

Use a general language corpus with utterances that cover all relevant phenomena (Festival)

- Design the prompts (greedy algorithms)
- Record the prompts (best of circumstances)
- Autolabel the prompts
- Build utterance structures for recorded utterances
- Extract pitchmarks and build LPC coefficients (electroglottogram)
- Build a unit based synthesizer from the utterances
- Test and tune

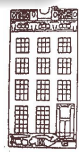
[Black and Lenzo(2003b)]



Processing a voice: Autolabeling

Use the prompts to label and segment

- The prompts have known labeling and segmentation
- Align the prompts to the recordings, eg, dynamic time warping or forced ASR alignment
- Can even be done when synthesized prompts are from a TTS of a different language
- If segmentation goes wrong, verify by hand
- Determine syllable stress and sentence accent from prompt specification
- Feed labels into utterance structures etc.



Speech characteristics: Expressive speech

Consistent pronunciation means little expression. Add different styles (professional speaker/actor)

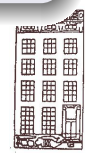
- Use appropriate style for task, eg, news, weather, stories
- Message has more effect in correct emotional state
- Very important when working for children
- Basic states: anger, happiness, sadness and neutral
- Prosodic models must be specific for each emotional state

[Bulut et al.(2002)Bulut, Narayanan, and Syrdal]

Speech characteristics: Changing speaker characteristics (not yet feasible)

Each different "voice" needs a separate speaker. Only what has been recorded can be spoken

- Change voice characteristics to create a different speaker, eg, man to woman to child (Praat allows this as a demo)
- Change voice to a different language variant or style
- Add new (level of) "expressiveness"
- Emotional state can be manipulated to some degree in prosody alone
- Techniques from rule based synthesis are needed to change complex traits, like stress and emotional states, reliably



Evaluation

Evaluation is the mother of progress

- Evaluate modules separately
- Construct rigorous and uniform evaluation procedures and criteria
- Separate diagnostic tests from full system evaluations
- Compare different system
- Standardize external input: Voice, texts, use
- TTS is evaluated by listeners
  - Self selected volunteers (eg, internet)
  - Paid naive listeners (eg, students)
  - Paid target groups (eg, office workers, K12 children)
  - TTS developers (Tit-for-Tat evaluation)
  - External Experts

Evaluation: Criteria

What can be evaluated (full system evaluation)

- Intelligibility at phoneme and word level
- Naturalness and pleasantness
- Intonation and prosody
- Stress positions and breaks
- Long text rendering (eg, intonation variation)
- Task appropriateness, i.e. correct style
- Voice and style selection in multi-speaker story telling (when feasible)

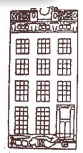
Blizzard challenge 2005

Aim: Find better synthesis techniques by comparing systems on the same data

Evaluating corpus-based speech synthesis on common datasets

- Effort to start international comparative evaluation of TTS systems
- Which approaches work, which don't
- Distribute common unit database, generate full TTS system within two weeks
- Evaluate common texts, 250 sentences from 5 genres
- Prevent "cheating" where needed

[Black and Tokuda(2005), Bennett(2005)]



Blizzard challenge 2005: Speech

Common speech databases

- CMU ARCTIC databases: 2 old + 2 new voices
- 1200 phonetically balanced sentences (5-15 words)
- Project Gutenberg novels (prose style)
- All words in CMUDICT
- Eg, *They were three hundred yards apart.*
- Automatically segmented and labeled

[Project Gutenberg(2005), Festvox(2005)]

Blizzard challenge 2005: Evaluation

5 text genres, 50 sentences each

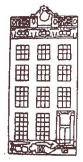
- Novels, same stories as original sentences
  Joe Garland lives like a good fellow.
- News, standard press-wire
  The two countries agreed to resolve any conflict through . . . Interfax said.
- Conversation, human side of spoken dialog system
  Yeah, I guess it will and something downtown please.
- Phonetically confusable sentences
  Now we will say cold/colt again.
- Semantically unpredictable sentences (SUS)
  The unsure steaks overcame the zippy rudder.

Blizzard challenge 2005: Listeners

Listener groups (and number who completed all tests)

- Speech experts, each participant provided 10 local experts (50)
- Volunteers over the web (60, unpaid)
- US undergraduates (58, paid)
- It proved to be difficult to get enough listeners (≈ 100)

[Bennett(2005)]



Blizzard challenge 2005: Test types

Test types

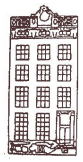
Mean opinion scores on a five point scale for:

- *Novels*
- *News*
- *Conversation*

And Word Error Rate for

- *Phonetically Confusable*
- *Semantically Unpredictable Sentences*

[Bennett(2005)]



Assignment: Week 6 Dynamic Time Warping

Use DTW to match speech samples

- Record or collect different realizations (eg, normal/fast) of the utterances "1 2 3 4 5"
- Use praat (Formant & LPC -, to MFCC...) to create *Mel Frequency based Cepstral Coefficients*
- Generate a dynamic time warp (To DTW..., match start and end and use *no slope restrictions*)
- Paint it
- Use the same technique to select a spoken number from a sequence of numbers (eg, "2 5" from "1 2 3 4 5"). Note that there can be problems from matching the other numbers

Further Reading I

Christina L. Bennett.

Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005.
In *Proceedings of Interspeech 2005, Lisboa, Portugal*, September 2005.
URL `http://festvox.org/blizzard/bc2005/IS052023.PDF`.

Alan W. Black and Kevin A. Lenzo.

Festvox.
Web, 2003a.
URL `http://festvox.org/`.
Festival speech synthesis.

Alan W. Black and Kevin A. Lenzo.

Building Synthetic Voices.
Festvox, 2 January 2003b.
URL `http://festvox.org/bsv/`.
Published on the festvox website.

Alan W. Black and Keiichi Tokuda.

The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets.
In *Proceedings of Interspeech 2005, Lisboa, Portugal*, September 2005.
URL `http://festvox.org/blizzard/bc2005/IS051946.PDF`.

P. Boersma.

Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL `http://www.Praat.org/`.

Further Reading II

P. Boersma and D. Weenink.

Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

Paulus Petrus Gerardus Boersma.

Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives.
PhD thesis, University of Amsterdam, September 1998.
URL http://www.fon.hum.uva.nl/paul/papers/funphon.pdf.

Murtaza Bulut, Shrikanth S. Narayanan, and Ann K. Syrdal.

Expressive speech synthesis using a concatenative synthesizer.
In *Proceedings of ICSLP 2002, Denver, COLORADO*, September 2002.
URL http://www.research.att.com/projects/tts/papers/2002_ICSLP/expressive.pdf.

Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors.

Survey of the State of the Art in Human Language Technology.
Cambridge University Press, 1996.
URL http://cslu.cse.ogi.edu/HLTsurvey/.
ISBN 0-521-59277-1.

Festvox.

Festvox.
Web, 2005.
URL http://www.festvox.org/.

Further Reading III

FSF.

GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

MBROLA.

The MBROLA Project.
Web, 2005.
URL http://tcts.fpms.ac.be/synthesis/.
Synthesis.

Bernd Möbius.

word and syllable models for german text-to-speech synthesis.
In Mike Edgington, editor, *Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS*, 26 November 1998.
URL http://www.slt.atr.co.jp/cocosda/jenolan/Proc/r06/r06.pdf.

Gregor Möhler.

Examples of Synthesized Speech.
Web, 2005.
URL http://www.ims.uni-stuttgart.de/~moehler/synthspeech/.
Good web-site with many examples.

Nextens.

NeXTeNS: Open Source Text-to-Speech for Dutch.
Web, 2003.
URL http://nextens.uvt.nl/index.html.

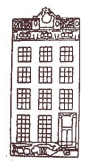
Further Reading IV

Louis C.W. Pols, Jan P.H. van Santen, Masanobu Abe, Alan Black, David House, Mark Liberman, and Zhibiao Wu.
Easy access via a TTS website to mono- and multilingual text-to-speech systems.
In *Proceedings of the Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS*, November 1998.

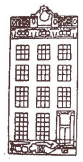
Project Gutenberg.
Project gutenberg free ebook library.
Web, 2005.
URL `http://www.gutenberg.org/`.

Richard Sproat.
ECE 598: Sp eech Synthesis.
Web.
URL `http://catarina.ai.uiuc.edu/ECE598/Lectures/klattlpc.pdf`.

SRL.
Synthesis of Speech.
Web.
URL `http://wagstaff.asel.udel.edu/speech/tutorials/synthesis/`.
Speech Research Lab, A.I. duPont hospital for children and University of Delaware.



Appendix A



Copyright License