Evaluation of Text-to-Speech synthesis (the 3rd ESCA/COCOSDA SSW)

Minkyu Lee
Bell Laboratories, Lucent Technologies
Murray Hill, New Jersey
minkyul@research.bell-labs.com
Presented by Minsoo Hahn (ICU)

Minkyu Lee's Prelude

• Prof. Y.J. Lee asked me to summarize TTS evaluation activities in the U.S. I thank Prof. Lee for giving me this opportunity.

• I felt that presenting the recent TTS evaluation experiment at the 3rd ESCA Speech Synthesis workshop is more appropriate. The evaluation was more global rather than activities limited in the U.S.

• I apologize for not being able to attend the meeting. Also, I could not prepare this presentation in Hangul because I don't have the Hangul powerpoint.

• I also thank Prof. M.S. Hahn for presenting this on behalf of me.

Background

- After EuroSpeech97 in Rhodos, Greece, several key members in the TTS community expressed their concern about the non-scientific way of demonstrating TTS output.
  - Canned demonstration can be misleading.
  - It is hard for researchers to compare different TTS algorithms.
  - Not many TTS systems are on the web for open experiment.
- Need standardized methods of evaluating TTS.
  - TTS evaluation is much more complicated than the Mean Opinion Test (MOS) test.

Background - continued

- The ESCA workshop on Speech Synthesis,
  - The 1st: Autrans, France (Sep. 1990)
  - The 2nd: Mohonk Mountains, New Jersey, USA (Sep. 1994).
  - The 3rd: Blue Mountains, Australia (Nov. 1998).
- TTS evaluation in the Speech Synthesis Workshop
  - The first trial of semi-formal evaluation of TTS system.
  - Total of 68 systems from 15 countries participated.
  - 10 language/gender groups.
  - Thee text material categories : telephone directory entries, newspaper sentences, and semantically unpredictable sentences.
  - 18 different evaluation categories.

Goals

- To give workshop participants a through and honest impression of current TTS systems.

- To stimulate discussion of formal TTS evaluation procedures.

- To provide valuable feedback to system developers and researchers about their systems.

- It is NOT for potential customers who wants to purchase a TTS. Also it is NOT a competition between systems (the results are not published).

Evaluation Procedure

- Text material is not know prior to the test.

- Text material is created by an automated method.

  – Random selection from text corpora owned by the *Linguistic Data Consortium (http://www.ldc.upenn.edu)*, which has no formal or informal ties to any participating TTS systems.

- System specific results are NOT to be published because

  – the subject population are not representative of the population at large.
  – statistics are based on very small numbers of observation.
  – listening situation is not acoustically optimal. (used a headphone from a PC in a noisy conference room)

Text material

- Telephone directory entries :

  – format : <first name>, <last name>, <street address>, <city>, <country, province or state>, <telephone number>

- Newspaper sentences - easy sentences :

  – generated by *minimum word frequency based selection method*,?which guarantees that all words in a selected sentence have a frequency of occurrence above a pre-defined threshold.
  – the sentence is relatively easy for grapheme-to-phoneme conversion.

Text material - continued

- Newspaper sentences - difficult sentences :

  – uses successive letter triples as a basic unit.
  – maximizes the diversity of triphones in the selected text.
  – word frequency is not considered.
  – it tests many components of a TTS system including grapheme-to-phoneme conversion, acoustic units, and prosody generation.

- Semantically unpredictable sentences :

  – common syntactic structures with words randomly selected from lexicons.
  – All words are meaningful, the grammar is correct but the sentence is meaningless. (e.g. 밪he tree ran through the green night?
  – it tests the segmental intelligibility of a TTS at the sentence level.

Test items and scoring

- Telephone directory entries : listeners has to

  – type in what he/she understood (dictation) .
  – for each element (name, address, etc.), select either 뱊o? 빽inor? or 빽ajor?problem.

- Newspaper and semantically unpredictable sentences :

  – global term (3 items) : Text analysis, Prosody, and Signal Processing. select either 뱊o? 빽inor? or 빽ajor?problem.
  – detail term (10 items) : Text analysis (mispronunciation, wrong syllable stress, wrong focal stress, phrase boundary), Prosody (phone duration, F0 contour, amplitude distribution), Signal Processing (voice quality, discontinuities, unclear or wrong phones). select either 뱊o? 빽inor? or 빽ajor?problem.

Test items and scoring - continued

- Listener's general opinion in the following categories:

  – Overall Impression, Intelligibility(comprehension), Naturalness of Prosody, Overall Voice Quality. Five point scale from *poor* to *excellent*
  –Overall Speech Rate. Five point scale from *too slow* to *too fast*

Participating Languages

- Initial applicants by deadline : language(number of systems)

  – Chinese(5), Dutch(3), English [Australia(1), UK(5), US(10)], French(7), Galician(1), German(11), Italian(3), Japanese(7), Korean(1), Portuguese(1), Romanian(1), Russian(1), Spanish [Basque(1), Catalan(2), Iberian(5), Mexican(3)]

- For final evaluation, each group must have at least 3 systems to compare with.

- Final evaluation groups :

  – US English (male and female), UK English, German(male and female), French, Chinese, Japanese, Iberian Spanish, Dutch,

Listener qualification

- There are three categories of listeners : native speaker, fluent speaker, and others.

- A listener can participate in the test for a language as

  – a native speaker? if he/she is a native speaker of the language.
  – a fluent speaker? if he/she has lived in the country more than 10 years.
  – an observer? if none of the above.

Listening process

• A listener listens to the output of each TTS with the same input text. And a listener listens to each TTS system equally often.

• A given listener hears different text items on each presentation in order to avoid potentially biasing text repetition effects.

• Across listeners, each TTS system is presented exactly once with each text item.

• Acoustics:

  – speech files have sampling rate of 11.025 kHz, 16 bit, mono, linear.
  – listening test over a headphone.

Experiment

- Standard experimental design.
  - For more detail, please refer to perceptual experiments for diagnostic testing of text-to-speech systems,?*Computer Speech and Language, 7, 1993, pp.,49-100,* by Jan van Santen.

- It provides unbiased?score estimates. But, many of the listeners are the system developers. Therefore, it is likely that familiarity with their own system leads to a biasing effect.

- Must correct scores by eliminating responses by a listener on trials involving a system the listener is affiliated.

Conclusion

• The evaluation gave many valuable feedback to the developers.

• The experiment revealed several flaws in the current evaluation procedures. It will be reflected at the next evaluation and hopefully for standardized TTS evaluation procedure. (for example, more systematic intelligibility test is needed.)

• ESCA is planning to hold the 2nd TTS evaluation at the 4th ESCA Speech Synthesis Workshop (scheduled, Year 2002).

•The committee encourage more systems/languages to join the next TTS evaluation.