Speech Technology in CALL

Speech Technology in Computer-Assisted Language Learning (CALL)

Zoe Handley

Feb 2010, Applied Linguistics, Department of Education, University of Oxford

Speaking and Listening in CALL

- A brief history of CALL
  - 1950s - 1970s
    Behaviouristic CALL
  - 1970s to 1990s
    Communicative CALL
  - 1990s onwards
    Integrative/Multimedia CALL

Speaking and Listening in CALL

- Storyboard exercises:

Speaking and Listening in CALL

- Tracy Talk - The Mystery by CPI

Speaking and Listening in CALL

- Computer-controlled audio cassettes and branching programs
- Digitised speech
- Multimedia (video and animation)
- Computer-Mediated Communication including audio chat and audio and video conferencing

A la Rencontre de Philippe

http://web.msu.edu/~billars/philippe.html

Zoe Handley
Department of Education, University of Oxford
Speech Technology

- Speech synthesis
  - Text-to-Speech (TTS) synthesis
- Speech recognition
  - Automatic Speech Recognition (ASR)
  - Introduction
  - Language learning
  - Benefits and challenges
  - Effectiveness

Speech Synthesis

“Systems that allow the generation of novel messages, either from scratch (i.e. entirely by rule) or by recombining shorter pre-stored units”
(van Bezoogen and van Heuven, 1997: 709)

Synthetic speech

- Speech synthesis is not
  - Recorded speech (CD, MP3)
  - Waveform manipulation
- Waveform manipulation

(Hattori and Krenn, 2007)

Text-to-Speech Synthesis

- The man (and he certainly was one!) just said, “Maybe, I'll see. I can't promise.”
- Dr. Jones lives at 11 School Dr. and works on the corner of St. James St.
- Challenges
  - Segmentation, text normalisation, ambiguity, heterophones

Text-to-Speech Synthesis: Architecture

Zoe Handley
Department of Education, University of Oxford
Speech Technology in CALL

Phoneme-to-speech module
- Formant synthesis (parametric synthesis)
  - Electronically models the features of the acoustic signal which are necessary from the point of view of perception
  - The formants are the frequencies of the different resonant cavities of the vocal tract.

Spectrogram for "phonetician"

Phoneme-to-speech module
- Concatenative synthesis
  - Segments of pre-recorded human speech are combined to generate new utterances

Units of Synthesis
(Dubat, 1997)

Concatenative synthesis
- Nuance RealSpeak
- Capstral
- CorePerc

Quality of Concatenative Synthesis
- Overall quality
  - Varies from very high to very low
  - Only produces possible human speech sounds
- Segmental quality
  - Better than formant synthesis for consonant clusters and vowel durations in stressed syllables
- Coarticulation
  - No segment provides a perfect account
- Prosody
  - More natural than formant synthesis
- Flexibility
  - A new database is needed for each new voice and style of speech
  - Only permits the modification of duration, pitch and amplitude within a narrow range.

Text-to-Speech Synthesis Applications
- Talking toys
- Reading machines for the blind
- Augmentative and Assistive Communication (AAC)
  - e.g. Stephen Hawking

Zoe Handley
Department of Education, University of Oxford
Speech Technology in CALL

CALL Applications of TTS synthesis
- Reading machine
- Pronunciation models
- Conversational partner

Reading machine
- Talking dictionary
- Talking text
- Talking word processor
- Dictation
- Grapheme-to-phoneme exercises

Oxford-Hachette 4 Frech Dictionary on CD-ROM

Pronunciation Model
- Practice of individual and combined phonemes
  - Auditory discrimination (listening)
  - Repetition (pronunciation)
- Practice of intonation and prosody (the music of speech)
  - Auditory discrimination (listening)
  - Repetition (pronunciation)

Conversational Partner
- In combination with automatic speech recognition, speech understanding, the generative power of TTS synthesis can be harnessed to provide learners with interactive speaking practice, i.e. a dialogue partner
- Examples:
  - SCS (Spoken Conversational Interaction for Language Learning) (bene et al., 2004)
  - Let’s Go BSS (Spoken Dialogue System) (Raum and Ekman, 2004)

Mr. Smokozoom, Monty Python sketch (KTH, 1998)
http://www.speech.kth.se

Benefits
- Easy creation and editing of speech samples
- Simultaneous presentation of text and speech
- Low storage requirements
- Non-human and therefore perceived as non-judgemental
- Improves on possibilities other media provide, but does not add value, i.e. bring about new possibilities
- Generation of examples on demand, as in Berufskunde, 1951, and therefore the automatic generation of feedback, conversational turns, and exchanges with speech models
- Adds value to CALL, i.e. brings about new possibilities such as provision of interactive conversations

Zoe Handley
Department of Education, University of Oxford

Is speech synthesis ready for use in CALL?
1. Basic research evaluation of TTS synthesis for use in CALL
   - Validity and potential benefits of the use of TTS synthesis in CALL
2. Technology evaluation of TTS synthesis for use in CALL
   - Adequacy of TTS synthesis for use in CALL
3. Judgemental evaluation of the CALL application
   - Potential of the CALL program to provide ideal conditions for SLA
4. Judgemental evaluation of the teacher-planned activity
   - Potential of the planned activity to provide ideal conditions for SLA
5. Usage evaluation of the teacher-planned activity
   - Learner's performance in the planned activity

This framework presented in Handley and Handley (2005) is a combination of the levels of evaluation recommended by Chaplin (2001) for the evaluation of CALL, et al. (1994) for the evaluation of Speech and Language Technologies (SALT).
Is speech synthesis ready for use in CALL?

1. Technology evaluation of TTS synthesis for use in CALL
   - Shoval et al. (1997) evaluated the quality of a Spanish TTS synthesis chip for
     the presentation of grammar exercises in a language laboratory.
   - Hambly and Hamer (2005) investigated the requirements of CALL in an
     exploratory evaluation of a French research TTS synthesis system.
   - Holland (2009) asked a group of French teachers to evaluate the quality
     of the speech generated by a range of French TTS systems with respect to
     their use in three different roles in which TTS is being used in CALL: (1)
     reading machine, (2) pronunciation model, (3) conversational model
   - Kang et al. (2009) in an evaluation involving Japanese learners of English
     compared the intelligibility of a commercial English TTS system with that of
     natural speech.

2. Usage evaluation of the teacher-planned activity
   - Cohen (1993) evaluated the use of a talking word processor to support
     literacy activities, namely writing stories, for young learners of French

Speech synthesis in CALL: Summary

- Despite the potential benefits of the use of TTS
  synthesis in CALL, namely the unique capacity to
  generate speech models on demand, research is still
  in its infancy.

Speech Recognition

- Automatic Speech Recognition (ASR)
  “Speech recognition is the process of converting an
  acoustic signal, captured by a microphone, or
  telephone, to a set of words.”
  (Zue, Cole et al., 1996: 4).

Waveform displays

Tell Me More from Auralog

Speech Recognition: Challenges

- Creativity
- Continuous speech
- Co-articulation
- Ambiguity
  - Homophones: to, too, two
  - Word boundaries: grey tape vs. great ape
- Variation
  - Inter-speaker
  - Contextual
  - Intra-speaker
- Environment
  - Ambient noise
  - Microphone

Speech Recognition

- Speaker-dependent ASR
  - Recognizes the speech of only one speaker
  - Example application: Command and control, e.g. UK RAF uses
    such a system to control cockpit functions (see Eurofighter
    Typhoon).
- Speaker-independent ASR
  - Recognizes the speech of a variety of speakers
  - Example application: National Rail Enquiries 08457 48 49 50
- Adaptive ASR
  - Anti-speaker independent at the outset and over time adapt to the
    user through user training.
  - Example application: IBM ViaVoice

Zoe Handley
Department of Education, University of Oxford
Speech Technology in CALL

Speech Recognition

- Isolated word recognition

  [Diagram of isolated word recognition]

- Large Vocabulary Continuous ASR (LVCSR)

  [Diagram of LVCSR]

Speech Recognition

- Statistical models

  [Diagram of statistical models]

Speech Recognition in CALL

- Applications
  - Vocabulary
  - Pronunciation
  - Reading
  - Conversation
  - Grammar

Vocabulary Tutors Integrating Speech Recognition

- TriplePlayPlus: Sings
  - Video enhances learning activities
  - Visual aids provide a solid context for LVCSR
  - Encourages learners to reflect different aspects of pronunciation

- Limitations
  - Speech proficiency varies between different proficiency levels
  - Recognition accuracy varies

<table>
<thead>
<tr>
<th>Target</th>
<th>Accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close (esp. flour)</td>
<td>3rd person when</td>
</tr>
<tr>
<td>Crochet (esp. thank you)</td>
<td>no gloss here</td>
</tr>
<tr>
<td>Mail (S. form)</td>
<td>My niece</td>
</tr>
</tbody>
</table>

Zoe Handley
Department of Education, University of Oxford
Speech Technology in CALL

Pronunciation Tutors Integrating Speech Recognition
- SPELL Project
- Vocabulary Builder from Hypergloss
  - Pronunciation scoring
    - Red (novice)
    - Yellow (intermediate)
    - Green (native speaker)

Conversational Tutors Integrating Speech Recognition
- Project LISTEN Reading Tutor (CMU)
- Subversive (Interactive Spoken Language Education (ISLE) project)
- DISCO project

Zoe Handley
Department of Education, University of Oxford
Speech Technology in CALL

Speech Recognition in CALL: Limitations

- General purpose speech recognition systems are designed to accept a wide range of pronunciations.
- Phonetic discrimination of minimally contrasting word pairs is more challenging than transcribing whole utterances (Duffy and Kealey-Pot, 1999).
- Uncertainty
- False positives and false negatives

Working Within the Limitations of Speech Recognition in CALL

- Admit that the software can be "fooled" (e.g., compare the accuracy of word identification in a vocabulary game).
- Input verification: "I heard you, is that what you said?"
- Personality of the conversational agent: TAT (Wachowicz and Scott, 1999)
- Predict possible mispronunciations (e.g., spell or predict)
- Constrain the dialogues: Multiple choice (e.g., with TracDISCO project).

Is speech recognition ready for use in CALL?

1. Basic research evaluation of speech recognition for use in CALL
   - Validity and potential benefits of the use of speech recognition in CALL
2. Technology evaluation of speech recognition for use in CALL
   - Evaluation of speech recognition technologies for CALL
3. Judgemental evaluation of the CALL application
   - Potential of the CALL program to provide ideal conditions for SLA
4. Judgemental evaluation of the teacher-planned activity
   - Potential of the planned activity to provide ideal conditions for SLA
5. Usage evaluation of the teacher-planned activity
   - Learner's performance in the planned activity

This framework, presented in Handley and Hemes (2005), is a combination of the three levels of evaluation recommended by Chapelle (2001) in the evaluation of CALL activities and in the SLA framework (1999) for the evaluation of speech and language technologies (SALT).

Speech recognition in CALL: Summary

- Despite the fact that many commercial CALL applications integrate speech recognition, few evaluations of its effectiveness have been conducted.

References

- Speech synthesis in CALL
  - Special issue on speech technology in CALL
- Speech recognition in CALL
  - Special issue on speech recognition in language learning
- Speech technology in CALL

Zoe Handley
Department of Education, University of Oxford
Thank you!

Questions?