14 October 2010
22:45

Speech Technology in Computer-Assisted Language Learning (CALL)

Zöe Handley
(zoe.handley@education.ox.ac.uk)

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 1980s
    Communicative CALL
  - 1990s onwards
    Integrative/Multimedia CALL

2

Speaking and Listening in CALL

Storyboard exercises:
http://www.lapasserelle.com/lm/exercices/storyboard/index.html

3

Speaking and Listening in CALL

***Tracy Talk – The Mystery* by CPI**

4

Speaking and Listening in CALL

A la Rencontre de Philippe
http://web.mit.edu/fll/www/projects/Philippe.shtml

5

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

6

Speech Technology

- Speech synthesis
  - Text-to-Speech (TTS) synthesis
- Speech recognition
  - Automatic Speech Recognition (ASR)

- Introduction
- Language learning
- Benefits and challenges
- Effectiveness

7

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 Bezooijen and van Heuven, 1997: 709)

8

Synthetic speech

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

(Hattori and Iverson, 2007)

9

Speech Synthesis

http://www.speaknspell.co.uk/speaknspell.html

10

Text-to-Speech Synthesis

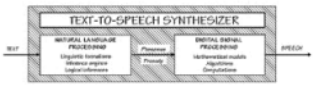
- The man (and he certainly was one!) just said, "Maybe. I'll see. I can't promise."
- Dr. Jones lives at 11 School Dr. and works on the corner of St. James St.
- Challenges

  Segmentation, text normalisation, ambiguity, heterophones

11

Text-to-Speech Synthesis: Architecture

- Two tasks, two modules
  - Natural Language Processing (NLP) module
    - Text-to-Phonemes (TTP) module
    - Input: Text
    - Output: Narrow phonetic transcription, augmented with information for the generation of intonation (an unambiguous representation)
  - Digital Signal Processing (DSP) module
    - Phoneme-to-Speech (PTS) module
    - Input: Narrow phonetic transcription
    - Output: Speech / wave form



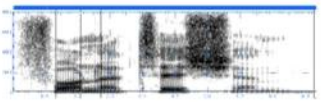
Dutoit (1997)

12

Zoe Handley
Department of Education, University of Oxford

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 formats are the frequencies of the different resonant cavities of the vocal tract

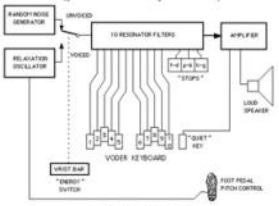


Spectrogram for "phonetician"

13

Phoneme-to-speech module

- Formant synthesis (parametric synthesis)



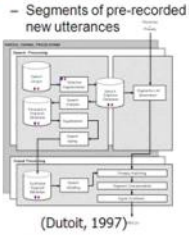
Dudley's Voder

http://www.haskins.yale.edu/featured/heads/SIMULACRA/voder.html

14

Phoneme-to-speech module

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



Units of Synthesis

http://www.telecom.tuc.gr/~ntsourak/tutorial_synthesis.htm

(Dutoit, 1997)

15

Concatenative synthesis

Acapela Text to Speech Demo

http://www.acapela-group.com/text-to-speech-interactive-demo.html

- Nuance RealSpeak
  - http://www.nuance.com/realspeak/
- Cepstral
  - http://www.cepstral.com/demos/
- CereProc
  - http://www.cereproc.com/support/live_demo

16

Quality of Concatenative Synthesis

- Overall quality
  - Varies from very high to very low
  - Only produces possible human speech sounds
- Segmental quality
  - Bad vowels, good consonants
  - 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

17

Text-to-Speech Synthesis Applications

- Talking toys
- Reading machines for the blind
- Augmentative and Assistive Communication (AAC) –
  - e.g. Stephen Hawking

18

Zoe Handley
Department of Education, University of Oxford

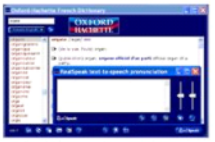
CALL Applications of TTS synthesis

- Reading machine
- Pronunciation models
- Conversational partner

(Handley and Hamel, 2005; Handley 2009)

19

Reading machine



Oxford Hachette 4 French Dictionary on CD-ROM

- Talking dictionary
- Talking text
- Talking word processor
- Talking conjugator
- Dictation
- Grapheme⇔phoneme exercises

20

Pronunciation Model

SAFexo (Hamel, 1998; 2003)

- 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)

21

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:

- *SCIL (Spoken Conversational Interaction for Language Learning)* (Seneff et al., 2004)
- *Let's Go SDS (Spoken Dialogue System)* (Raux and Eskenazi, 2004)

Mr Smoketoomuch Monty Python sketch (KTH, 1999)

http://www.speech.kth.se/

22

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** (Sherwood, 1981) and therefore the automatic generation of feedback, conversational turns, and exercises with speech models
- **Adds value** to CALL, i.e. brings about new possibilities such as provision of interactive conversations

23

Is speech synthesis ready for use in CALL?

1. Basic research evaluation of TTS synthesis for use in CALL
   - Viability 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 Hamel (2005) is a combination of the levels of evaluation recommended by Chapelle (2001) for the evaluation of CALL activities and by ELSE (1999) for the evaluation of Speech and Language Technologies (SALT).

24

Zoe Handley
Department of Education, University of Oxford

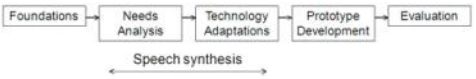
Is speech synthesis ready for use in CALL?

2. Technology evaluation of TTS synthesis for use in CALL
   - Stratil et al (1987) evaluated the quality of a Spanish TTS synthesis chip for the presentation of grammar exercises in a language laboratory
   - Handley and Hamel (2005) investigated the requirements of CALL in an exploratory evaluation of a French research TTS synthesis system
   - Handley (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 the three different roles in which TTS is being used in CALL: (1) reading machine, (2) pronunciation model, (3) conversational model
   - Kange et al. (2008) in an evaluation involving Japanese learners of English compared the intelligibility of a commercial English TTS system with that of natural speech
5. Usage evaluation of the teacher-planned activity

   Process oriented
   - 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 synthesis

(Holland and Fisher, 2008)

26

Speech Recognition

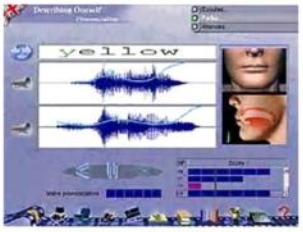
= Automatic Speech Recognition (ASR)

"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).

27

Waveform displays



***Tell Me More* from Auralog**

28

Speech Recognition: Challenges

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

29

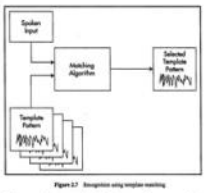
Speech Recognition

- 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
  - Are speaker independent at the outset and over time adapt to the user through user training
  - Example application: IBM ViaVoice

30

Speech Recognition

Isolated word recognition

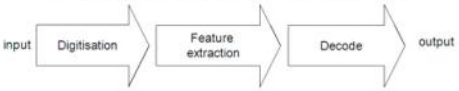


Pattern matching / template-based speech recognition (Rodman, 1997)

31

Speech Recognition

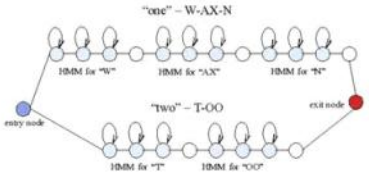
Large Vocabulary Continuous ASR (LVCASR)



32

Speech Recognition

Statistical models

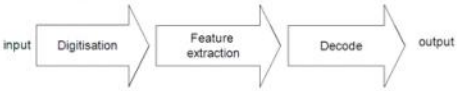


http://cmusphinx.sourceforge.net/sphinx4/

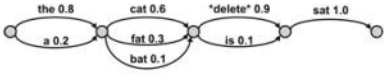
33

Speech Recognition

Large Vocabulary Continuous ASR (LVCASR)



Statistical models



Find the best path through the network of hypotheses

34

Speech Recognition in CALL

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

35

Vocabulary Tutors Integrating Speech Recognition

- TriplePlayPlus Bingo (aka Smart Start from Syracuse)
  - Variant of traditional classroom activity
  - You must pronounce a word correctly to fill in a square
  - 3 acceptance levels reflecting different levels of proficiency
- Limitations
  - Does not discriminate between the different proficiency levels
  - Recognition errors – false positives

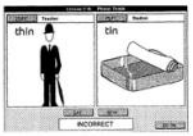
Target	Accepted
Quatro (esp: four)	Quando (esp: when)
Gracias (esp: thank you)	No grass here
Mais (fr: corn)	My niece

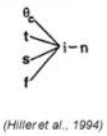
36

Zoe Handley
Department of Education, University of Oxford

Pronunciation Tutors Integrating Speech Recognition

- SPELL Project





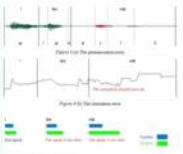
(Hiller et al., 1994)

- *Vocabulary Builder* from *Hyperglot*
  - Pronunciation scoring
    - Red (tourist)
    - Yellow (intermediate)
    - Green (native speaker)

37

Pronunciation Tutors Integrating Speech Recognition

- Multimedia English Learning (MEL) System



(Lai et al., 2009)

38

Reading Tutors Integrating Speech Recognition

- **Project LISTEN Reading Tutor (CMU)**

(Mostow and Aist, 2001; Mostow et al., 2003)

39

Conversational Tutors Integrating Speech Recognition

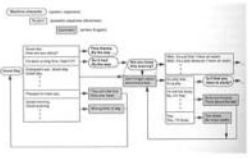
- **TriplePlayPlus Bubble Dialogues**

- **CPI Canned Conversations and Conversations with Traci**

40

Conversational Tutors Integrating Speech Recognition

- **Subarashii (Interactive Spoken Language Education (ISLE) project)**



(Bernstein et al., 1999)

41

GrammarTutors Integrating Speech Recognition

- **DISCO project**

(Strik et al., 2009)

42

Zoe Handley
Department of Education, University of Oxford

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 (Dalby and Kewley-Port, 1999)
- Uncertainty
- False positives and false negatives

43

Working Within the Limitations of Speech Recognition in CALL

- Admit that the software can be "fooled"
  (*TriplePlayPlus Bingo Vocabulary Game*)
- Input verification
  - "I heard X. Is that what you said?"
    (*DynEd Question Formation*)
- Personality of the conversational agent
  - "Traci appears to be a bit absent-minded; she simply does not listen" (Wachowicz and Scott, 1999)
    (*CPI Canned Conversations and Conversations with Traci*)
- Predict possible mispronunciations
  (*SPELL Project*)
- Constrain the dialogues
  - Multiple choice (*Conversations with Traci; DISCO project*), predictable responses, high quality images (*Subarashii*)

44

Is speech recognition ready for use in CALL?

1. Basic research evaluation of speech recognition for use in CALL
   - Viability and potential benefits of the use of speech recognition in CALL
2. Technology evaluation of speech recognition for use in CALL
   - Adequacy of speech recognition 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 Hamel (2005) is a combination of the levels of evaluation recommended by Chapelle (2001) for the evaluation of CALL activities and by ELSE (1999) for the evaluation of Speech and Language Technologies (SALT).

45

Is speech recognition ready for use in CALL?

2. Technology evaluation of speech recognition for use in CALL
   - Have focused on evaluating the recognizers ability to detect errors in learners' speech
   - Dalby and Kewley-Port (1999) compared the accuracy of word identification and the validity of pronunciation scores for template-based and statistically-based speech recognition systems
   - Rypa and Price (1999) compared speech recognition based pronunciation scoring with that of human raters (interrater correlation = 0.68; human – ASR correlation = 0.61)
5. Usage evaluation of the teacher-planned activity

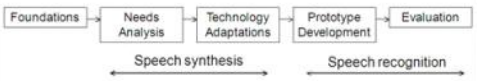
Outcome oriented

- Poulsen et al. (2007) evaluated the Project LISTEN Reading Tutor with a group of grade 2-4 Hispanic learners of English
- Neri et al. (2008) evaluated the effects of the PARLING CAPT system on the pronunciation accuracy of a group of 11 yr old Italian learners of English
- Lai et al. (2009) evaluated the effects of the MEL pronunciation tutor on the phonemic awareness, and spelling and reading abilities of a group of 3rd grade Taiwanese learners of English

46

Speech recognition in CALL: Summary

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



Speech synthesis ↔ Speech recognition

(*Holland and Fisher, 2008*)

47

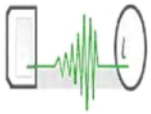
References

- Speech synthesis in CALL
  - Handley, Z (2009). Is Text-to-Speech Synthesis Ready for Use in Computer-Assisted Language Learning? *Speech Communication*. 51 (10): 906-919
    Special issue on speech technology in education
- Speech recognition in CALL
  - Wachowicz , K. A. and Scott, B. (1999). Software that Listens: It's not a Question of Whether, It's a Question of How. *CALICO Journal*. 16 (3): 253-276
    Special issue on speech recognition in language learning
- Speech technology in CALL
  - Holland, V. M., Fisher, F. P. (2008) (eds.) *The Path of Speech Technologies in Computer-Assisted Language Learning: From Research Toward Practice*. London: Routledge.

48

Zoe Handley
Department of Education, University of Oxford

Thank you!



Questions?

Zoe Handley
Department of Education, University of Oxford