Speech recognition and synthesis

Copyright ©2007-2008 R.J.J.H. van Son, GNU General Public License [FSF(1991)]

Introduction: The problem

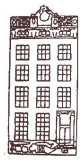
Both Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) are based on collecting and manipulating speech corpora

- ASR and TTS can be seen as clever speech databases
- Both compare the target, *input* or *output*, utterance to a speech model
- Select the speech model that best fits the target utterance
- The model speech is constructed from stored examples
- Two questions:
  - How to create a model of the target utterance?
  - How to compare a model to the target utterance?

Introduction: Basic problem

How to build TTS and ASR

- Store speech data in an (abstract) model description
- Create model utterances
- Compare these models to the target utterance
- Select the best fitting model utterance
- Example: Mandarin tones for student practise



Introduction

Problems Teaching Madarin

- Mandarin Chinese is a tone language
- Every syllable in a word has one of 4 (5) tones which determines the meaning of the word
- Using the wrong tone makes a word incomprehensible (cf, English *bad* and *bat*, Dutch *boot* and *bot*)
- Mastering the production and recognition of tones is a major stumbling block in learning Mandarin Chinese
- Direct interaction with a highly proficient speaker, usually the teacher, is needed to practise tone pronunciation

Introduction

A consequence of the difficulty of learning tones

- Classes must be kept small to allow for ample student-teacher interaction
- Speaking and listening proficiency improves very slowly
- High drop-out rates of demotivated students
- Speaking is neglected in favor of writing

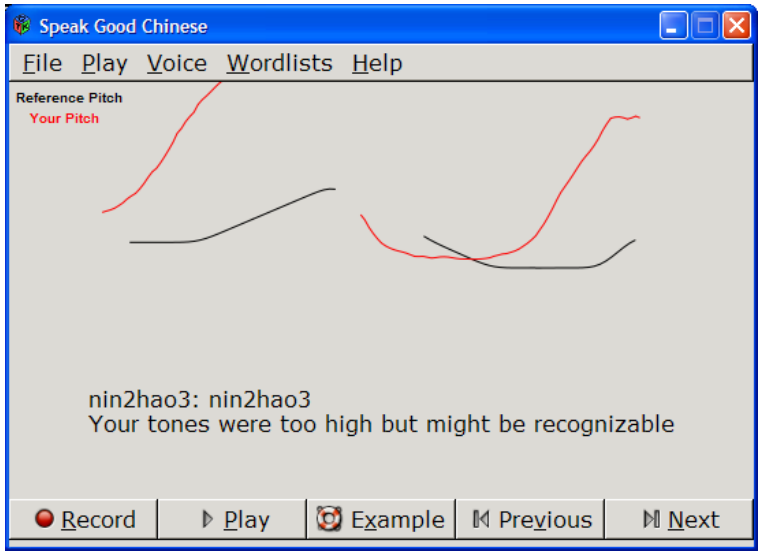
Introduction

Computer Assisted Language Learning (CALL)

- Language learning requires practise
- Teachers are scarce and expensive
- Use computer technology to help students practise
- Reading and Writing: texts, spelling and grammar checkers
- TTS: Read aloud texts, generate examples
- ASR: Judge student pronunciations and give feedback

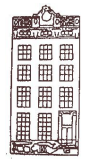
SpeakGoodChinese

An aid for practising Mandarin tones.



http://www.SpeakGoodChinese.org/

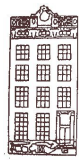
- All mono- and bisyllabic words
- Automatic Tone Recognition
- Graphical Tone Presentation
- A written analysis of tone pronunciation.
- Hummed (TTS) or pre-recorded examples
- Replaying recorded student pronunciation
- Automatic student evaluation (hidden)



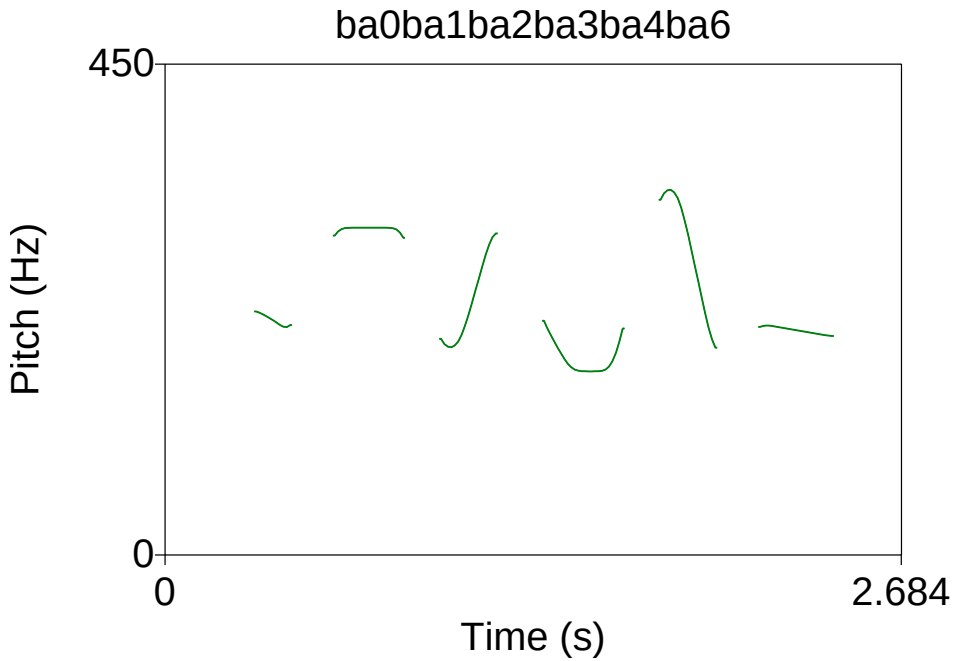
SpeakGoodChinese

Pinyin to Tone synthesis as TTS

- Pinyin phonetic transcription system (eg, *ni3hao3*)
- Each syllable has a number 1-4 or the neutral tone 0
- Split pinyin word into syllables (on tone number)
- Split pinyin syllable into Unvoiced initial and voiced final
- Tone contour is realized on voiced part only



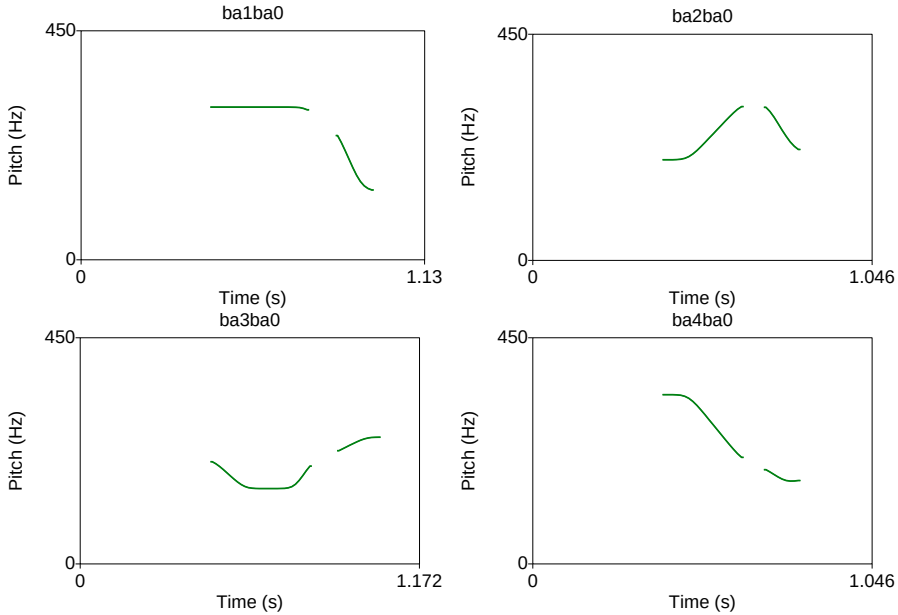
Tone models: All tones



SpeakGoodChinese tone models

- Neutral tone, 0, tones 1-4, and garbage model 6
- Tones change in "context"

Tone models: Assimilation of neutral tone



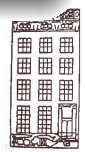
Examples

- Neutral tone continues from previous tone
- Returns to "neutral" position
- Fourth tone seems exception

Tone synthesis: Pinyin to syllables and tones

Mandarin syllables, eg, *zhong1*

- Syllable: Optional Initial (*zh*) + Obligatory Final (*ong*)
- Initial is always a single phoneme (*zh* = /dʒ/)
- Initial can be voiced and voiceless
- Final is always voiced
- Final is a vowel and an optional nasal /nmŋ/
  (rarely an /r/)
- Vowel can be a monophthong, /e/, diphthong, /ei/, or triphthong, /iau/
- Tones are realized on the voiced part of the syllable



Tone synthesis: Initials and finals

	a	ei	ong	ia	iong	uan
b	ba	bei				
d	da	dei	dong			
zh	zha		zhong			zhuan
r						
j				jia	jiong	
g	ga	gei	gong			guan

Durational model

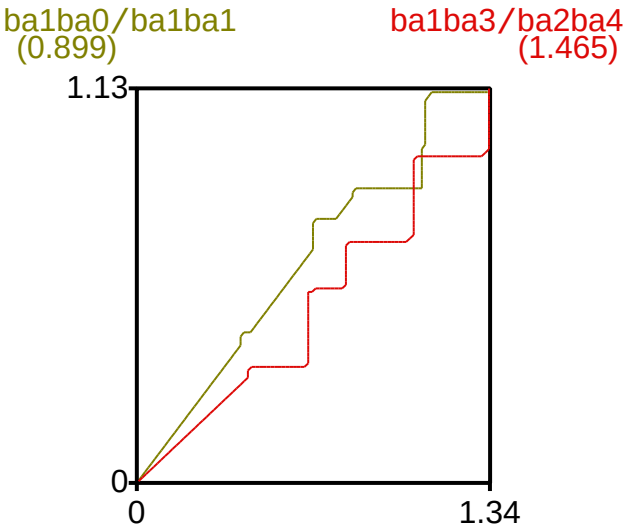
- Estimate durations of Initial and Final
- Crude model: Fixed duration $+ \delta\cdot$ number of symbols (iao=3)
- Adapt duration to tone: $3 > 1 > 2 \approx 4 \gg 0$

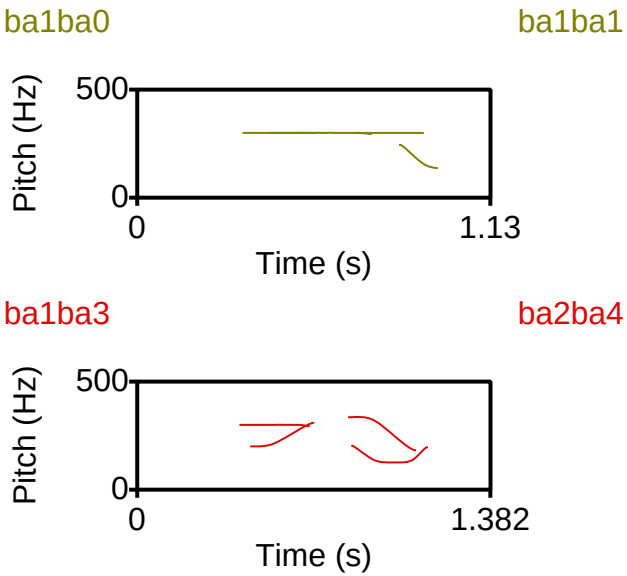
Tone recognition

Tone recognition: Was student correct?

- Extract utterance pitch contour (F_0)
- Pinyin-to-Tone synthesis for all tones (correct and *in*correct)
- Compare student utterance to all possible tone contours using Dynamic Time Warping
- Pick best matching model $\Rightarrow$ Recognition
- Construct possible countours from theoretical tone model
- Limited to two syllables (combinatorial explosion)
- Student pitch register must be known

Tone recognition: Dynamic Time Warping

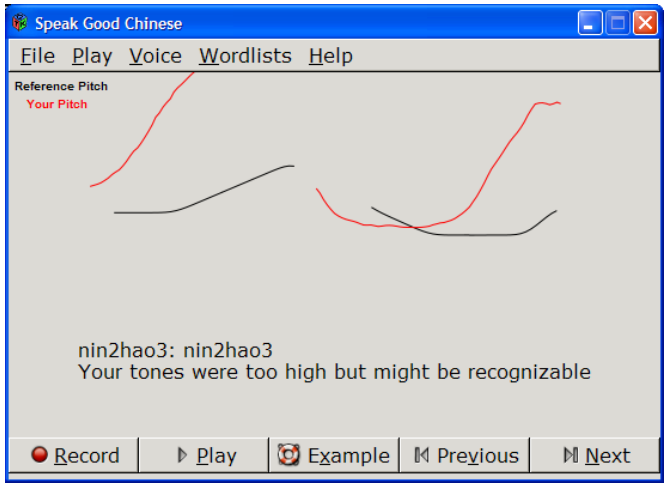




Align time points and "sum" distances ⇒ shortest path

- *ba1ba0* and *ba1ba1* very much alike (0.899)
- *ba1ba3* and *ba2ba4* more different (1.465)
- Do this for all combinations, effective for bisyllabic words

Tone recognition: Pitch height and movements



A good tone has correct pitch height and movements

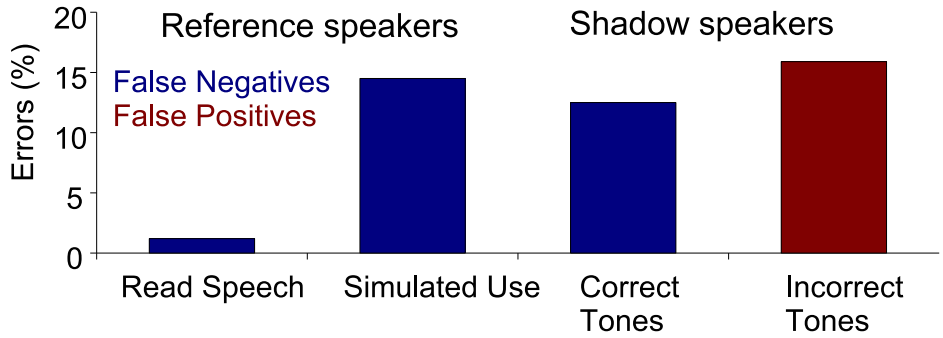
- If *top pitch* deviates from model, flag an error
- If *pitch range* deviates from model, flag an error
- Students will exagerate tones, punish exagerations less
- Flag error if 3 semitones too low or too narrow
- Flag exageration if 6 semitones too high or too wide

Tone recognition: Heuristic rules

Model tones do not model enough variation

- Duration rules currently very bad
- Current tone models do not capture variation
- Use "heuristic" rules to capture common confusions
- Eg, tones 2 and 3 merge before another tone 2 or 3
- Eg, tones 2 and 4 often misidentified as tone 0 in DTW but tone 0 would have been flagged by tone height and movement

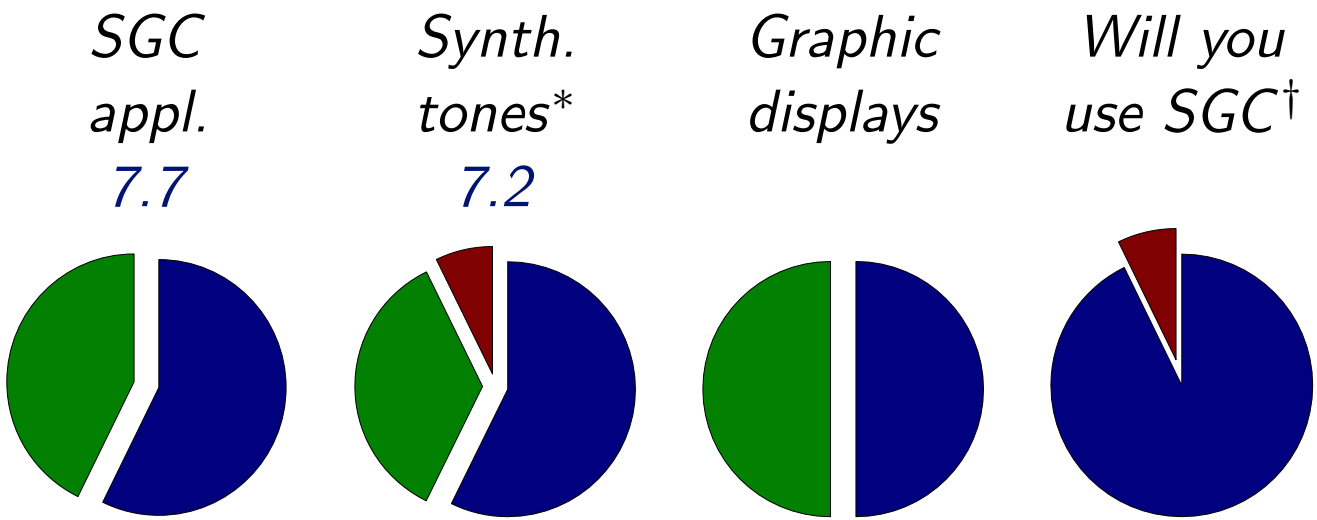
Evaluation: Recognizer False *rejects* and *accepts*



Reference speakers and Students

- Correct Tones
  - Read Speech: R read aloud 6 words: *cha2, dian4hua4, duo1shao3, gong1zuo4, jie2hun1, shi2jian1*, *83* tokens.
  - Simulated Use: R free word choice, *358* tokens
  - Shadowed Correct Speech: R and S shadowed 6 words, *160* tokens
- Incorrect tones
  Shadowed Incorrect Speech: R and S shadowed 6 words, *320* tokens

Evaluation: Usefulness and grade 1-10



Legend: Not useful/No - Useful - Very useful/Yes

* One subject couldn't hear the tones clearly

† One subject preferred to practice with family members

Questionnaire to 14 students

- Tested RAD Tcl/Tk GUI with functional recognition
- Responses used to design User Interface

Evaluation: Usage data

Does SpeakGoodChinese improve tone pronunciation?

- Single Female student (13)
- Tried out SpeakGoodChinese in 7 session of a few hours
- In total she utterred 1531 words
- Each session started and ended with test runs without audio feedback
- Pretest and Posttest $\approx$ 30 words
- Practise $\approx$ 83-389 words
- Automatically determined error rate ($^*p < 0.002, X^2$)
  - Overall: 28% (including practise)
  - Pretest: 39% *
  - Posttest: 24% *
- Real progress awaits human judgment

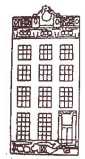
Assignment: Week 4 Dynamic Programming

function MIN-EDIT-DISTANCE(*target*, *source*) **returns** *min-distance*

$n \leftarrow$ LENGTH(*target*)
$m \leftarrow$ LENGTH(*source*)
Create a distance matrix *distance[n+1,m+1]*
distance[0,0] $\leftarrow 0$
for each column i **from** 0 **to** n **do**
  for each row j **from** 0 **to** m **do**
    $distance[i,j] \leftarrow \text{MIN}($ $distance[i-1,j] + ins\text{-}cost(target_i)$,
      $distance[i-1,j-1] + subst\text{-}cost(source_j, \text{target}_i)$,
      $distance[i,j-1] + del\text{-}cost(source_j))$

- Implement the Minimal Edit Distance (see figure)
- See 5.6, pages 153-156 + figure 5.6, of Jurafsky & Martin
- Use your favorite programming/scripting language
- Include back-pointers to allow tracing the best alignment
- NOTE: there are (initialization) errors in the pseudo-code
- Test it on several character strings

See blackboard for more information.



Further Reading I

BNC.
Britisch National Corpus.
Corpus, 1997.
URL http://www.natcorp.ox.ac.uk/.

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

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

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

James L. Hieronymus.
Ascii phonetic symbols for the world's languages: Worldbet.
Web, 1994.
URL http://www.ling.ohio-state.edu/~edwards/worldbet.pdf.

Further Reading II

IDS.
COSMAS.
Corpus, 2005.
URL http://corpora.ids-mannheim.de/~cosmas/.

NTU.
Spoken Dutch Corpus (CGN).
Corpus, 2004.
URL http://www.tst.inl.nl/cgn.htm.
Metadata (MPI) - http://www.mpi.nl/world/ISLE/overview/Overview_CGN.html Contents - http://www.elis.ugent.be/cgn/ Descriptions and references - http://lands.let.ru.nl/cgn/ehome.htm.

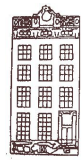
Roeland Ordelman.
Twente Nieuws Corpus (TwNC).
Corpus, 2002.
URL http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html.

R.J.J.H. Van Son.
IFA corpus 1.0.
Corpus, 2003.
URL http://www.fon.hum.uva.nl/Service/IFAcorpus.

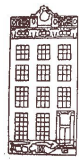
D. Weenink, G. Chen, Z. Chen, S. de Konink, D. Vierkant, E. van Hagen, , and R.J.J.H. van Son.
Learning tone distinctions for Mandarin Chinese.
In *Proceedings of INTERSPEECH 2007*, pages 950–953, Antwerp, Belgium, August 2007.
URL http://www.fon.hum.uva.nl/rob/Publications/p950I07_WeeninkEtAl2007.pdf.

Further Reading III

M Wynne, editor.
Developing Linguistic Corpora: a Guide to Good Practice.
Oxford: Oxbow Books, 2005.
URL `http://ahds.ac.uk/linguistic-corpora`.
Accessed 2007-09-04.



Appendix A



Copyright License

Copyright ©2007-2008 R.J.J.H. van Son, GNU General Public License [FSF(1991)]

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.
This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.
You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.