How can speech technology help linguists understand tone and intonation?

Tutorial 2
Tonal Aspects of Language, Nanjing

Daniel Hirst

Laboratoire Parole et Langage, CNRS and Aix-Marseille*Université
Tongji University, Shanghai
daniel.hirst@lpl-aix.fr

2012-05-27
One world many languages
The curse of Babel

- The language barrier is perhaps the greatest social problem facing modern multicultural societies.
- Language is not just words - non-verbal information is (at least) just as important.
- This is an area where we need speech technology.
- Speech technology for non-verbal information is in its infancy.
What is missing?

Figure: Why can’t we use these to speak to people in other languages?
What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)
- Speech synthesis (Acapela, Google translate)
What have we already got?

- Speech recognition (Dragon dictate, Google translate)
- Translation (Babelfish, Google translate)
- Speech synthesis (Acapela, Google translate)

Figure: My hovercraft is full of eels!
The data explosion

Twenty years ago: common to formulate models on the basis of limited data
Today: linguists expected to take into account large quantities of data - several hours
Specialised laboratories
Main-frame computers
Expensive software
Tools for audio and video

Free and multi-platform:

- Anvil
- Elan
- Praat
- WaveSurfer
- see Llisterri http://liceu.uab.es/ joaquim/
Data annotation

Biggest problem today:
- Not storage of data
- Not analysis
- Annotation
Prosodic Annotation

The explicit characterization of the length, pitch and loudness of the individual sounds which make up an utterance.
Prosodic Annotation

The explicit characterization of the length, pitch and loudness of the individual sounds which make up an utterance. Linguists need tools.
Speech alignment

It takes a linguist several hours to align one minute of speech with a phonetic transcription.
Speech alignment

It takes a linguist several hours to align one minute of speech with a phonetic transcription. They have better things to do...
SPPAS


**Figure:** Sample sentence from Eurom1-ZH corpus ("Because we do not have another room")

Available (GPL license) from: http://www.lpl-aix.fr/~bigi/sppas/.
Speech technology

- current disparity in resources
- small minority of languages - acceptable (?)
- vast majority of languages - primitive
- transfer of resources?
Speech technology resources

- often language specific
- difficult to generalise to:
  - under-ressourced languages
  - different dialects
  - different speaking styles
- speech prosody
The annotation/representation of prosody is crucial for:

- intelligibility "He’s not coming back"
- statement? question? order?
- speaker states "Isn’t this interesting"
- naturalness
  - facilitate cognitive processing
  - cf non-standard, non-native, pathological, or synthetic speech
- limited current use of synthesis for listening tasks but huge potential
Annotation of speech prosody

Current prosodic annotation is too language / theory specific

▶ cross-language annotation
  ▶ - INTSINT (Hirst & Di Cristo 1998)
  ▶ - ToBI (Jun 2005)

▶ interaction between linguists and engineers

▶ Biannual Speech Prosody Conferences

▶ 6th International Speech Prosody Conference, (May 2012 - Shanghai)
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
  - form (HL)

Inter-transcriber agreement (Wightman 2002 “ToBI or not ToBI”)
- functions good
- forms bad

Automatic recognition the opposite
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
  - form (HL)
- Inter-transcriber agreement
  (Wightman 2002 "ToBI or not ToBI")
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
  - form (HL)
- Inter-transcriber agreement (Wightman 2002 "ToBI or not ToBI")
  - functions good
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
  - form (HL)
- Inter-transcriber agreement
  (Wightman 2002 "ToBI or not ToBI")
  - functions good
  - forms bad
Prosodic annotation function vs form

- most prosodic annotation systems don’t distinguish
- ToBI: H* L%
  - function (* %)
  - form (HL)
- Inter-transcriber agreement (Wightman 2002 "ToBI or not ToBI")
  - functions good
  - forms bad
- Automatic recognition the opposite
Automatic annotation of pitch

- Command-Response: Hiroya Fujisaki
- Momel/INTSINT: Daniel Hirst and Robert Espesser
- RFC and Tilt: Paul Taylor
- Stem-ML: Greg Kochanski and Chilin Shih
- Prosogram: Piet Mertens
- Penta: Santitham Prom-On and Yi Xu
- AuTobi: Andrew Rosenberg
Problem for modelling $f_0$

"More news about the Reverend Sun Myung Moon..."

Figure: Two second extract of $f_0$ curve
Problem for modelling $f_0$

"More news about the Reverend Sun Myung Moon..."

Figure: Two second extract of $f_0$ curve

- Raw $f_0$ is discontinuous and not smooth.
Problem for modelling $f_0$

"More news about the Reverend Sun Myung Moon..."

- Raw $f_0$ is discontinuous and not smooth.
- Here beginning and end is continuous and smooth.

Figure: Two second extract of $f_0$ curve
Problem for modelling $f_0$

"More news about the Reverend Sun Myung Moon..."

Figure: Two second extract of $f_0$ curve

- Raw $f_0$ is discontinuous and not smooth.
- Here beginning and end is continuous and smooth.
- Discontinuity is due to microprosodic effect of consonants.
Using sonorants to illustrate intonation patterns

Figure: Finnish intonation (from Iivonen 1998)
Macro- and Micro-melody

Statement intonation

a. A ton papa.
/atÔpapa/

b. A ma maman.
/amamamã/
Macro- and Micro-melody

Question intonation

a. A ton papa ? /atÔpapa/

b. A ma maman ? /amamamã/
General model for $f_0$

Raw $f_0$ is the combination of two components
General model for $f_0$

Raw $f_0$ is the combination of two components

- Macromelodic component: smooth and continuous
  (Underlying intonation pattern)
General model for \( f_0 \)

Raw \( f_0 \) is the combination of two components

- Macromelodic component: smooth and continuous
  (Underlying intonation pattern)
- Micromelodic component: discontinuous
  (Surface effect of phonemes)
Macromelodic profiles

Figure: Macromelodic profile for extract from A01-01
Macromelodic and Micromelodic profiles

Figure: Macromelodic (red) and micromelodic (blue) profiles
Macromelodic profile

Figure: Macromelodic profile for extract from A01-01
Macromelodic profile

Figure: Macromelodic profile for extract from A01-01
Phonetic representation

- Momel/INTSINT
Phonetic representation

- Momel/INTSINT
- Automatic reversible annotation with Momel
Phonetic representation

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
Phonetic representation

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
  - macroprosodic component
    (independent of segmental material)
Phonic representation

- Momel/INTSINT
- Automatic reversible annotation with Momel
- Momel factors raw F0 into
  - macroprosodic component (independent of segmental material)
  - microprosodic component (independent of intonation)
Figure: Momel
Figure: Momel
Figure: Momel
Theory neutral?

- Theory-friendly
Theory neutral?

- Theory-friendly
- Phonetic representation - first step for:
Theory neutral?

- Theory-friendly
- Phonetic representation - first step for:
  - Fujisaki model (Mixdorff)
Theory neutral?

- Theory-friendly
- Phonetic representation - first step for:
  - Fujisaki model (Mixdorff)
  - ToBI (Maghbouleh, Wightman & Campbell, Cho (K-ToBI))
Theory neutral?

- Theory-friendly
- Phonetic representation - first step for:
  - Fujisaki model (Mixdorff)
  - ToBI (Maghbouleh, Wightman & Campbell, Cho (K-ToBI))
  - INTSINT
Surface Phonological Representation

- INTSINT designed as tool for linguists for the symbolic coding of intonation patterns. (Hirst & Di Cristo (eds) 1998)
Surface Phonological Representation

- INTSINT designed as tool for linguists for the symbolic coding of intonation patterns. (Hirst & Di Cristo (eds) 1998)
- Momel and INTSINT are both now implemented as plugin for Praat
Analysis by synthesis
Analysis by synthesis

Figure: The Analysis by Synthesis paradigm
Analysis by synthesis

Figure: The Analysis by Synthesis paradigm
Analysis by synthesis

Figure: The Analysis by Synthesis paradigm
Analysis by synthesis

Figure: The Analysis by Synthesis paradigm
What is science?

Figure: Jean Baptiste Perrin (1870-1942).
What is science?

Figure: Jean Baptiste Perrin (1870-1942).

scientific method: explain visible complexity by invisible simplicity.
(expliquer le visible compliqué par l’invisible simple.)