

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Speech databases for phonetic research and speech synthesis development: presentation and critique of a possible suite of tools

Briony Williams
Language Technologies Unit

Canolfan Bedwyr
University of Wales, Bangor, UK

1. Phonetic transcription stage

Procedure adopted in e.g. Spoken Dutch Corpus:-

1.1 Initial automatic segmentation

Requirements:

- Easy to use for non-speech-technologists.
- Free of charge.
- Will run on a fairly low-spec machine.

Autosegmentation software

- Some kind of HMMs, e.g. HTK: (learning curve involved, requires specialist expertise).
- Festival's "make_labs" script: only applicable when there's an existing synthetic voice (unlikely for less-resourced languages).

1.2 Manual post-editing of autosegmentation output

Various possibilities: Praat, WaveSurfer, SFS, Emu.

- Display waveform, spectrogram, 1 or more tiers.
- Playback file/current selection/current segment.
- Free to download!

WaveSurfer vs Praat, SFS, and Emu

- WaveSurfer is better for manual segmentation.
- WS & Emu (not Praat, SFS): Vertical synch of cursor across tiers: easier to place boundaries accurately when the vertical line is draggable.
- Right-click menu (in WS and Emu, but not SFS or Praat) provides symbols. Less scope for typos.



- WS has good import/export facilities (ESPS, HTK, etc). SFS can export ESPS labelfiles, but cannot import any form of label file.
- WS & Praat handle IPA symbols easily if desired.
- Praat comes into its own with scripting features (which WaveSurfer lacks). But these are not needed for manual segmentation/editing.

1.3 User requirements

- For speech recognition: This is sufficient to create a DB for training/testing acoustic models.
- For speech synthesis: This is sufficient for labelling a DB for basic (context-free) diphones. But if CART trees (Classification & Regression Trees) needed from this data (for duration, F0) then higher linguistic levels must also be labelled.
- For acoustic phonetic research: Higher linguistic levels must also be labelled.

2. Higher-level transcription stage

- Given: (broad?) phonetic transcription, soundfile.
- Then: Annotate at higher linguistic levels, e.g. syllable, word, phrase, intonation unit(s).
- Possible software: Emu, Praat

Common features of Emu and Praat:

- Multiple tiers (indefinite number).
- Tiers time-aligned with speech signal.
- Spectrogram and waveform display.

Emu-specific features:

- Emu can import from/export to Praat TextGrid (Praat can import ESPS labelfile).
- Emu: not only tiers but also branching structure between tiers (1-to-1, 1-to-many, many-to-many).
- **Importantly:** Emu has built-in DB query language with "wizard" to minimise learning curve.
- Queries can be structured across multiple tiers simultaneously (e.g. "all cases of /a/ that are word-final but not phrase-final").
- Output of database queries can form input to the "R" statistics suite, for statistics on e.g. "freq of F1 and F2 for all cases of /a/ that are word-final before a word-initial vowel".
- Possible to "explode" annotated levels into separate label files (ESPS format). These can be used for e.g. training duration parameters in a speech synthesis system.

3. Tools for using the corpus

Statistics software: R statistics suite

- Free to download.
- User-friendly graphical interface.
- Can be applied to trackfiles as well as labelfiles: formant analysis, F0 statistics, power, durations.
- Emu output integrates into R.

Typical use: basic descriptive statistics of a language

- Phonetic attributes of segments (formant frequencies & bandwidths, power, duration, F0).



- Prosodic patterns (F0 variations over time).

Emu/R vs Praat:

- Integrated structure of Emu/R saves much time compared to using user-written Praat scripts.
- Emu/R graphical wizard reduces learning curve.

4. Some issues

- ESPS labelfiles: deprecated standard/still useful?
- With less-resourced languages, need to consider learning curve (most native speakers not expert phoneticians/computer scientists). Compromises needed between tool power and usability (e.g. Emu has no scripting facility, but for its specific task has more built-in functionality than Praat).
- Some LRL's may be candidates for speech tech work. Need to make it easy to re-use earlier data.

