Thisl progress report

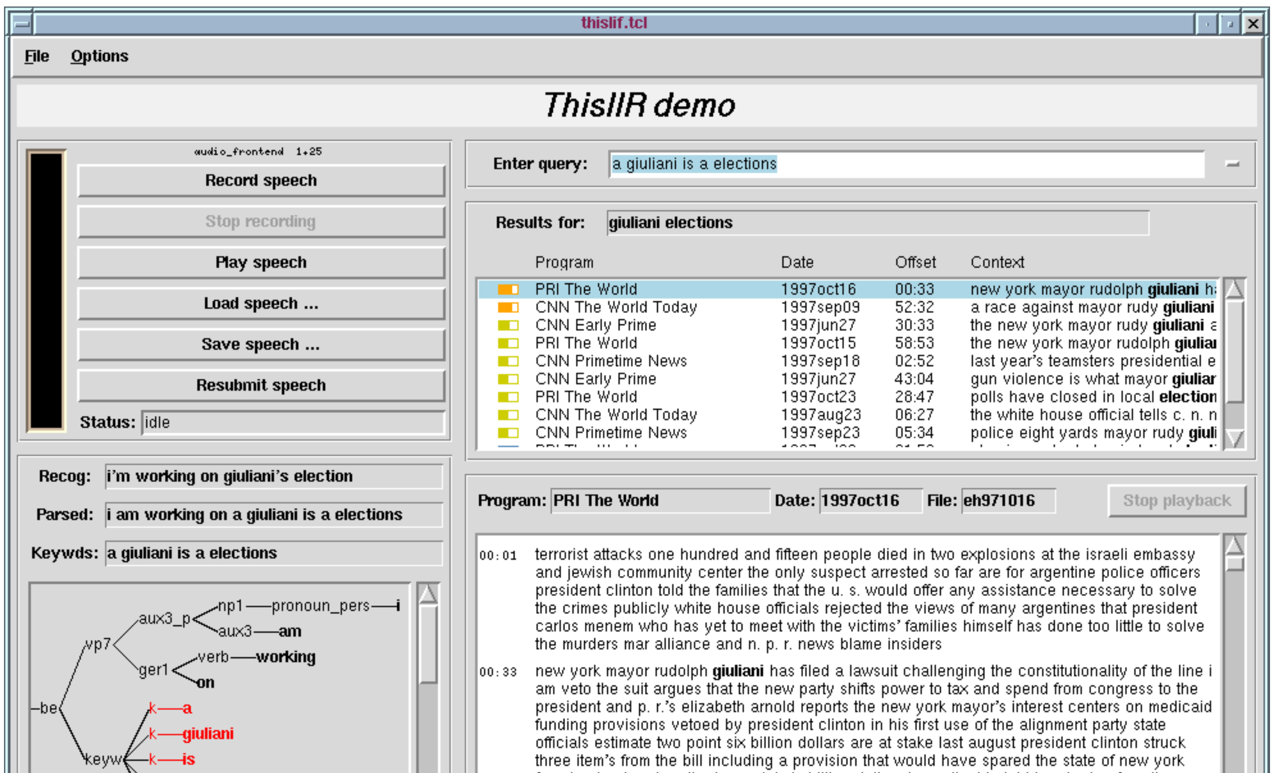
Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. **ThislGui enhancements**
2. **Thisl MSG-MLP acoustic model**
3. **MLP-based speaker ID**
4. **Speech/nonspeech discrimination**
5. **Quicknet enhancements etc.**

1 ThislGui enhancements

- **Spoken query input:** SPRACHdemo/AbbotDemo
- **NLP integration:** Thomson's prolog lattice parser
- **Faster:** SRT files parsed & saved (+ Tcl8)
- **Better:** supports Real Audio, more status feedback



MSG-MLP acoustic model

- **From SPRACH: combined models good**
  - especially plp-RNN and msg-MLP
  - e.g. WER: 27.2% + 29.7% $\rightarrow$ 24.9%
- **Obtained 50hr BBC training set from Softsound**
  - trained (28x9):8000:42 multi-layer perceptron
  - used TetraSPERT = 175h train, 375 MCUPs
  - feature calc: 0.2 xRT; fwd pass: 0.3-1.6-2.1 xRT
- **Results (euro99Eval test set):**
  - RNN baseline: 29.2%
  - msg1N-8k alone: 35.5%
  - Posterior combination: 28.7%
- **Why less benefit than SPRACH?**
  - not enough training data?
  - no msg-based realignment?
  - less telephone-bandwidth data?
  - (bugs?)

MLP-based speaker ID

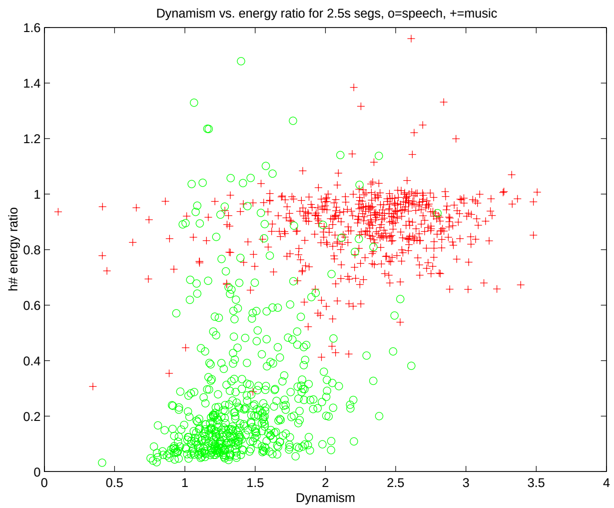
(par Dominique GENOUD, ex-IDIAP)

- **How to use hybrid systems for speaker ID?**
  - train speaker-dependent nets? too little data
  - specialize SI nets with a little SD data and compare posteriors?
- **But nets are *discriminative*...**
  - speaker-detection nets have *two* outputs per phone: speaker's phone, rest-of-world phone
  - train on 15 min of speaker + 15 min of others
  - sum posteriors on Viterbi path for each half
  - EER on 12 speakers (from BN) ~ 9% (c/w ?)
- **Applications**
  - speaker ID - but have to gather training data
  - speaker-adapted recognition:
    SD-trained nets have ~ 20% RER reduction

Speech/nonspeech discrimination

(with Gethin Williams of SU)

- **Posteriors features to detect 'decodable' segs**
- **4 statistics based on acoustic model outputs:**
  - avg per-frame entropy
  - $|\text{first-order diff}|^2$
  - energy ratio of h#
  - phone var'ce template
- **Test on Scheirer/Slaney music+speech data:**
  - classif err: 0% (15s segs); 1.3% (2.5s segs)
  - use to discard non-speech before decode



Quicknet enhancements

- **New release, quicknet v0_97**
- **MultiSPERT support integrated**
  - general client-server structure for other CPUs?
- **Online delta calculation**
  - saves disk space
  - waiting on Torrent-native convolution
- **Online per-utterance normalization**
  - two-pass - bad with online deltas
- **Also:**
  - new RLE-compressed ilab label format (1/30th)
  - full support for pre/lna input
  - bug fixes

Other news

- **Fabio CRESTANI visiting ICSI**
  - ex-Glasgow IR
  - IR for spoken documents, PDA applications
- **Multimedia indexing project with UCB EE (Avideh Zakhor)**
  - Build indexes for video URLs found on the web
  - testing BN recognizers on decompressed audio
- **Segmentation-as-decoding**
  - MDL-style criterion: next segment merges into this model, or starts a new model?
  - long, stable segments become easier to test
  - incremental decoder-like algorithm:
    lookahead window, alternate hypotheses