

# **Issues and tools for creating and annotating a corpus of sociolinguistic field data**

Christopher Cieri

University of Pennsylvania  
Department of Linguistics &  
Linguistic Data Consortium

- Ad hoc system motivated by **sheer laziness**.
- Goal is to support a study is to characterize the phonology of a Regional Italian variety (Aquilano) under the influence of not only Standard Italian but also two local dialects.
- Focal Question: Is the phonological variation observed better modeled as a small number of varieties with inherent variation or a larger number of invariant varieties?
- **Overlap with this workshop**
  - empirical analysis of recorded interview data from
  - live informants speaking in a linguistic variety whose
  - underlying grammatical structure is not fully known &
  - need for infrastructure to support analysis and collaboration

- **Corpus - a body of (raw) data collected and annotated for a specific purpose**
  - Raw Data - naturally occurring data resulting from some linguistic performance
  - Annotation - any process of adding value to a corpus
- **For data originally written, the written text is the raw data. For speech, only the audio is raw data**
- **Annotation encodes either human judgement or automatic processing based on either raw data or on previous layers of annotation.**
- **Transcription and segmentation are special kinds of annotation**
  - transcription encodes subtle human judgements about what was said
  - segmentation defines the granularity of future annotations

- 80 subjects stratified for age, gender, socioeconomic background
- Interviewers both native and non-native; subjects typically interviewed in pairs
- Attempt to capture multiple “styles”; **examine style as a function of time in the interview**
- Objective and subjective analyses:
  - vowels system, intervocalic /v/, /c/ before high vowels
- Need for tools and formats to
  - collect and
  - annotate data
  - manage layers of analysis
  - summarize and
  - share results

- Listen to tape for interesting tokens
- Digitize individual tokens
- Code tokens (using software where appropriate)
- Mark tokens on score sheet
- Reformat data for statistical analysis
- Problems
  - slow, labor intensive
  - high risk of missed tokens
  - tokens typically unbalanced, representation of styles poor
  - time measured poorly
  - effort for reanalysis nearly equal to effort for original
  - only limited opportunities for re-use

- Digitize entire interview & check audio quality.
- Transcribe, segment & check format.
- Query system for items of possible interest.
- Where appropriate, preprocess for segmental analysis.
- Label and analyze segments of interest.
- Summarize.
- **Advantages**
  - fewer misses
  - balanced coverage
  - time measured accurately
  - re-use & reanalysis profits from previous preparation

- Interviews recorded on audio cassette using Sony Walkman Professional stereo recorder and a pair of lavalier microphones.
  - each subject on separate mike
  - interviewer typically off-mike
- Digitized as **two channel**, 16 bit, 32KHz files via a Sony DAT recorder; down-sampled to 16KHz and transferred to computer via a Townshend DAT Link (narecord) Saved in Entropic's .sd format
  - .wav and .sph formats also possible
- Beginning & ending silence trimmed, files demuxed, empty channels removed.
- **Need to incorporate automatic checking of signal quality (sample min/max & long periods of low energy)**

# Transcription & Segmentation

- Orthographic transcription with interesting items & features transcribed phonetically
- Time aligned to audio file via segmentation at the speaker turn level
- Segmentation defines/refines domain of analysis
  - utterance level, word level, segment level (for vowels)
- Initial Segmentation
  - at each speaker turn
  - within long turns at ~8 seconds
  - segmented into breath groups where possible -- though not guaranteed
- Format
  - **start, stop, channel,** speaker, situation, utterance

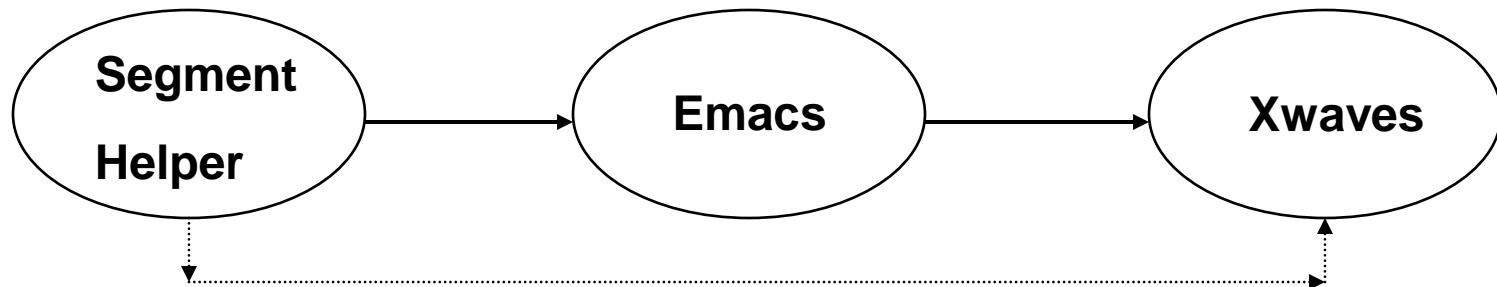


- **Strans**

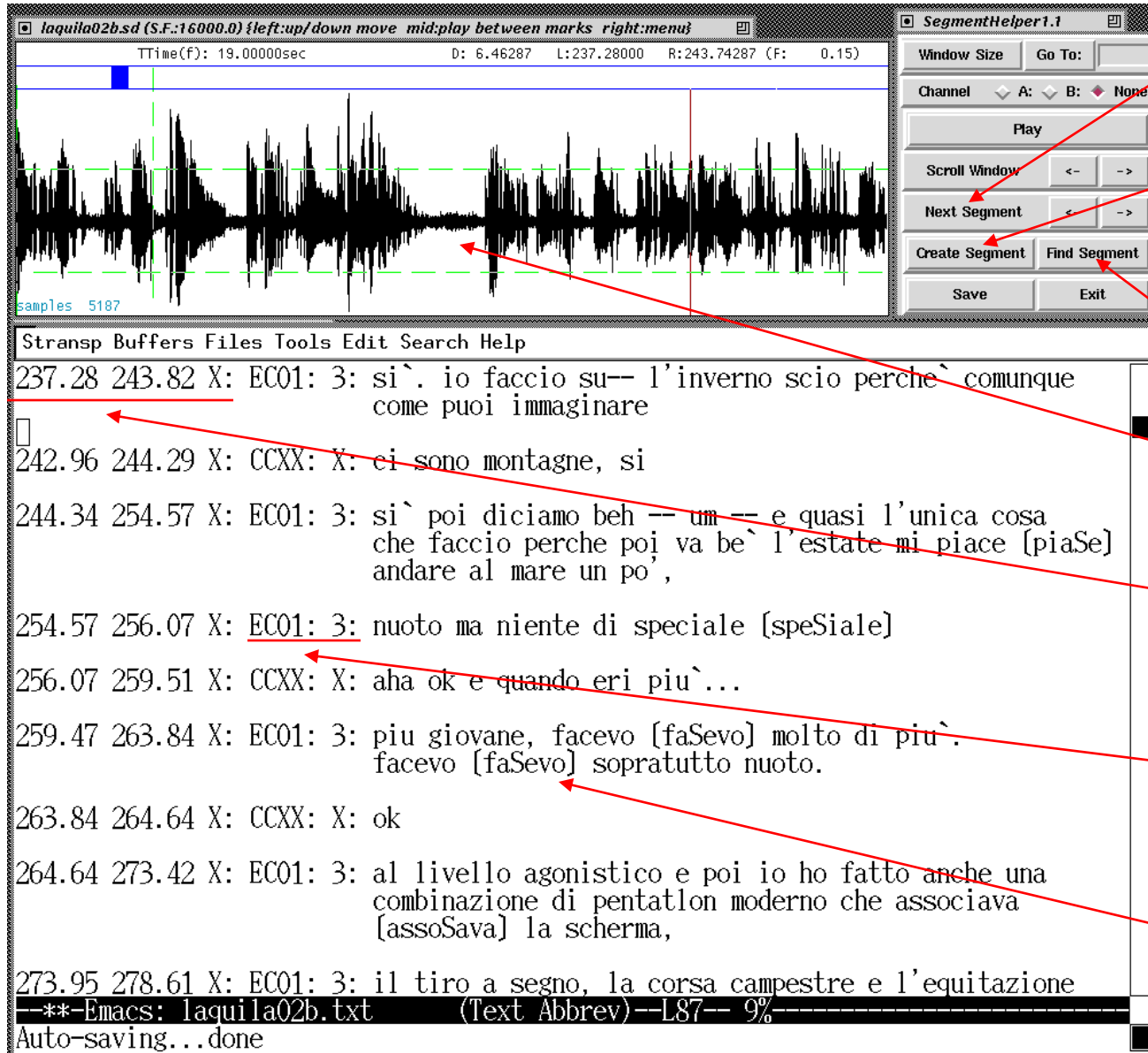
- Emacs with menus modified and macros added to support transcription talking to Xwaves through “send\_xwaves”

- **Segment Helper**

- Emacs running in server mode
- Client writes all commands to stdout where Emacs either acts on them immediately or passes them onto Xwaves.



- Segment Helper & all utilities hereafter written in PerlTK -- free, available on Unix and NT, merges the TK GUI capacity with Perl’s flexibility and flow control.



- Next Segment - shifts display so that 10% of last segment shows
- Create Segment polls Xwaves for left, right cursor positions and writes those as time stamps with channel marker in text
- Find Segment finds position in waveform of segment defined in text
- Monoaural recording with subject on single mike; interviewer off mike.
- Segment defined by start & stop times plus channel marker and written by software based on cursor positions.
- Speaker ID written by human and later normalized. Situation code written semiautomatically and checked by human.
- Interesting feature transcribed phonetically.

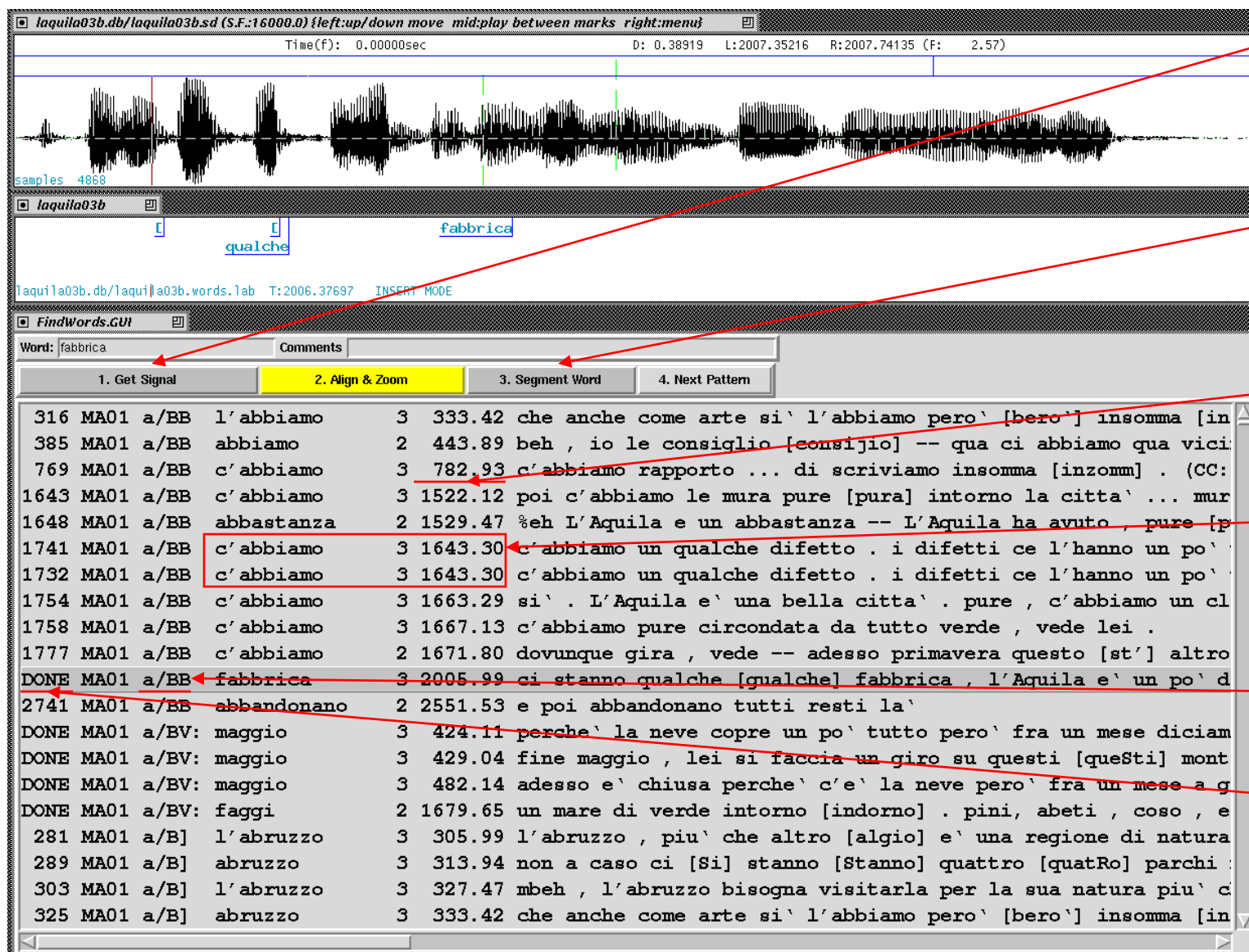
- **Some transcription done initially on foot pedal controlled transcription machine**
  - files subsequently segmented with Strans
- **Many files segmented initially at speaker turn, pause or breath with the segments transcribed subsequently.**
- **As an experiment some files transcribed with help of ASR System**
  - native speaker trained Dragon *Naturally Speaking* Italian
  - listened to tapes via foot-pedal controlled device
  - repeated each utterance to Naturally Speaking & corrected its mistakes

	ASR	Manual
<b>Experiment 1</b>	13.1xRT	13.4xRT
<b>Experiment 2</b>	11xRT	7.8xRT

- **After Segmentation and Transcription, files are checked by a second transcriptionist for**
  - bad segmentation
    - » too much or too little included in the transcript
    - » gap between segments too large
  - inaccurate transcription
  - inaccurate situation code
  - misspellings
  - inaccurate phonetic transcription
- **and by automatic process for**
  - segments too long
  - time stamps out of order or internally inconsistent
  - impossible channel marker, speaker ID or situation code
- **QC catches human formatting errors.**
- **System controls all subsequent processing.**

- FindWord searches reformatted transcript, identifies and numbers any words matching the query. Each hit word is presented to user in context as text and audio
- Software guesses location of word in utterance based on simple assumption that all syllables are of roughly equal length -- does surprisingly well
- Linguist adjusts word boundaries in waveform display, zooms and iterates until satisfied.
- Results saved in new file in SGML format.

```
< hitnum=3      pattern=o/PP word=vent'otto uttnum=1  
  speaker=EC01 situation=3 channel=X  
  ustart=76.85 ustop=79.39  
  utterance=nel vent'otto aprile [abrile] mille  
  novecento [noveSento] sessanta  
>
```



■ **GetSignal** locates and plays utterance, guesses word position and sets cursors

■ **SegmentWord** writes segmentation to new file and marks hit as done.

■ **Retaining times** allows user to balance samples over corpus

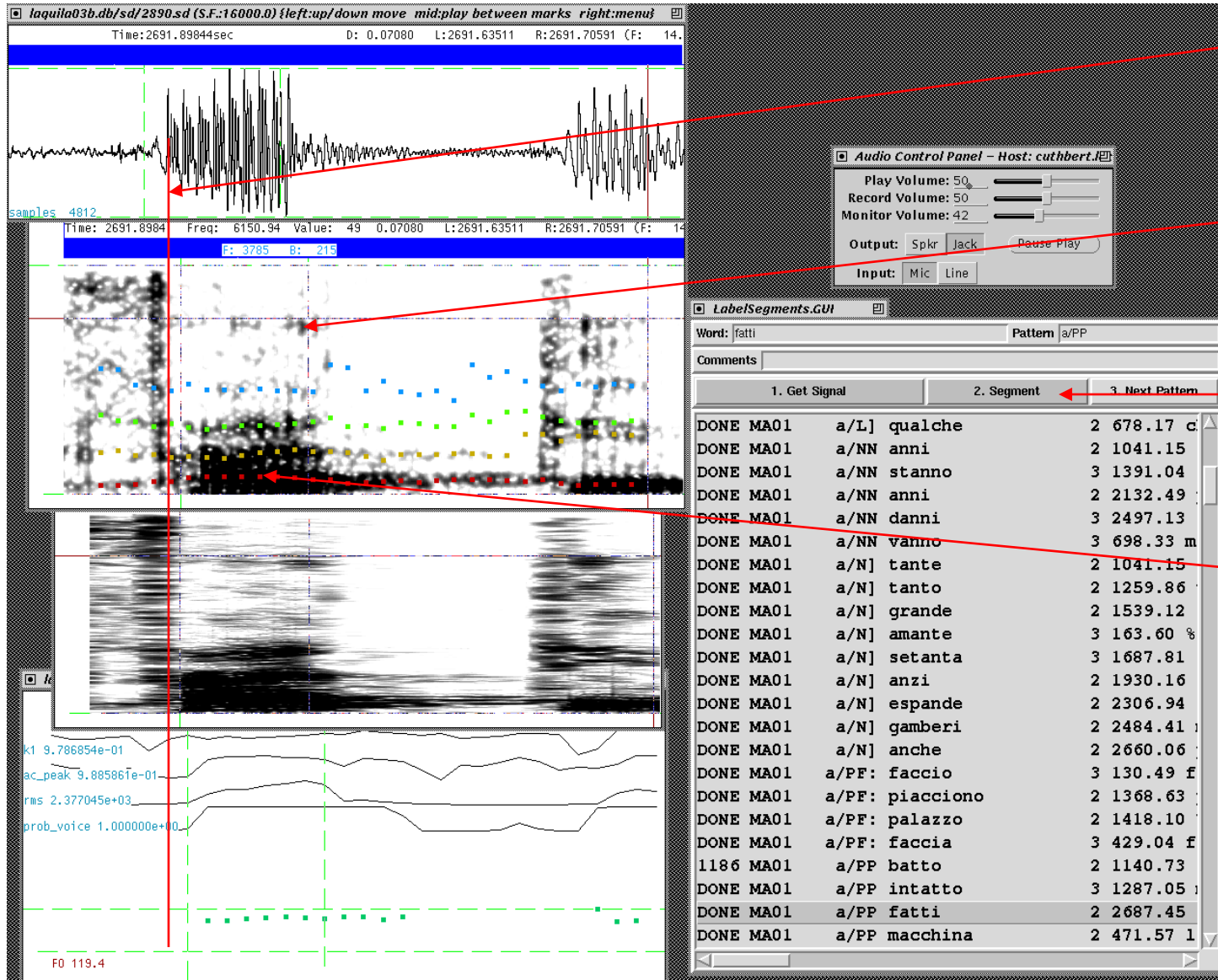
■ **Lexical Item** matching search. May be more than one per utterance

■ **Abstract Label** for Search Pattern

■ **Unique Hit** Number



# SegmentVowels



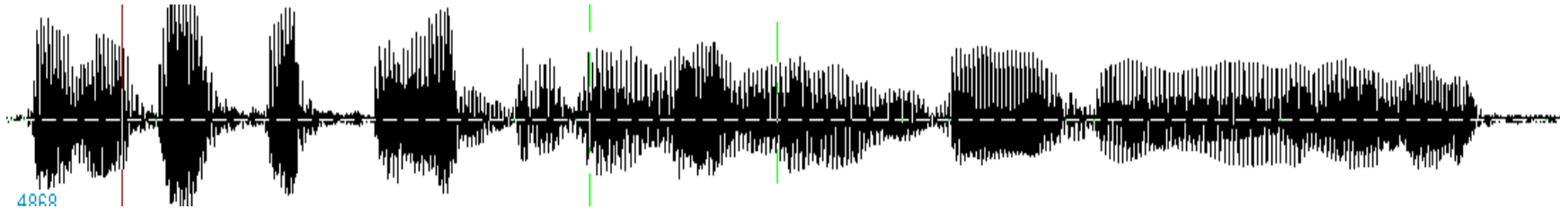
■ Time Aligned displays of waveform, WB and NB spectrogram and F0 characteristics

■ Software guesses position of segment within word.

■ User adjusts segmentation and saves to file.

■ Software estimates formant values automatically

■ All sound files, spectrograms, and F0 files processed ahead of time in batch and saved for later redisplay.



U1 U2		U3						U6 U7
		U4: una donna bella				U5		
			H1: bella					
					S1:E			
					F123			



		Hit		Segment	Analysis
		Hit #	→	Hit #	→ Hit #
	Utterance	Pattern		Segment	F1
	Utterance #	← Utterance #	Lexicon	S Start Time	F2
	U Start Time	Word	→ Word	S Stop Time	F3
	U Stop Time	W Start Time	Expected Pron		
Subject	Channel	W Stop Time			
Speaker	← Speaker	Actual Pron			
Age	Situation				
Sex					
Ed Level					
Profession					
Region					
Location					