VoiceSauce:

A program for voice analysis

Yen-Liang Shue, Patricia Keating,
Chad Vicenik, Kristine Yu

UCLA

Introduction

- Voice measures – relevant in linguistic phonetics, language description, sociophonetics, many other areas, including those represented in this session
- Many acoustic voice measures are used
- How to get lots of measures from lots of speech?
- VoiceSauce is our answer for this
- Source/Sauce: [sɔːs]

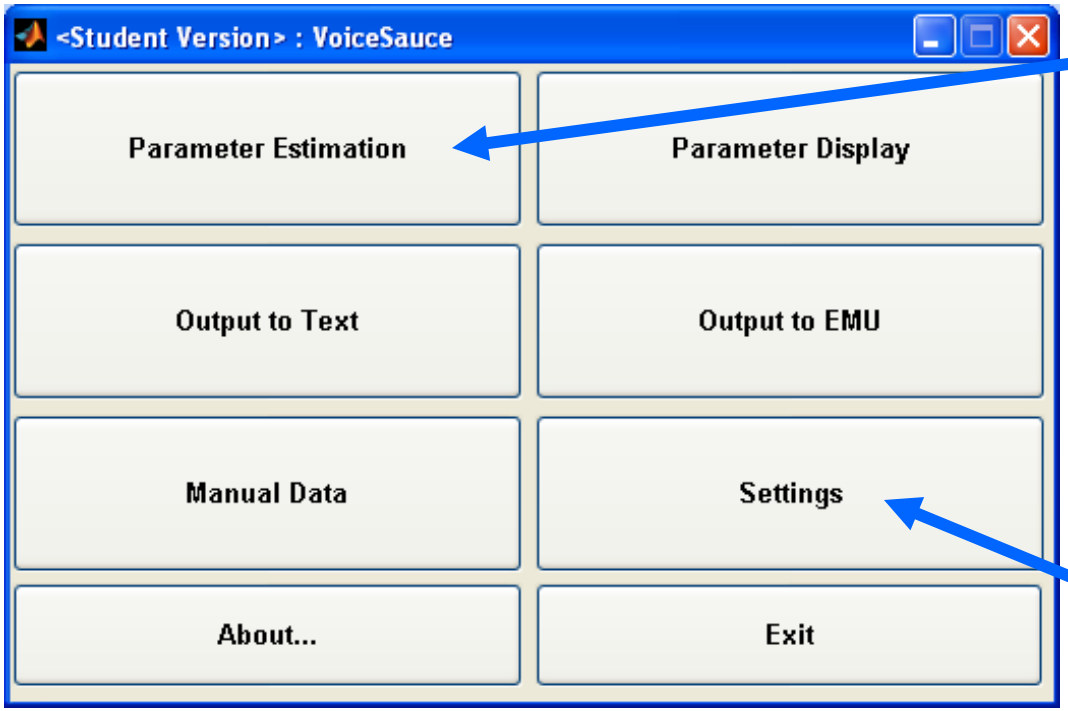
Overview of VoiceSauce

- Available free by downloading
- User-friendly; no scripting needed
- Gives automated voice measurements over time from audio recordings
- Computes many voice measures, including corrections for influence of formants
- Outputs values as text or for Emu database, and can include EGG analysis data too
- Runs in Matlab, or as a freestanding Windows program

Preliminaries: Praat textgrids

- VoiceSauce runs on directories of audio files in WAV format
- If Praat textgrids are available, VoiceSauce can use those to limit analysis to labeled intervals on any tier
- VoiceSauce can also use Praat textgrids to structure its output files
- Thus almost always our first step, before using VoiceSauce, is to provide a textgrid file for each WAV file to be analyzed

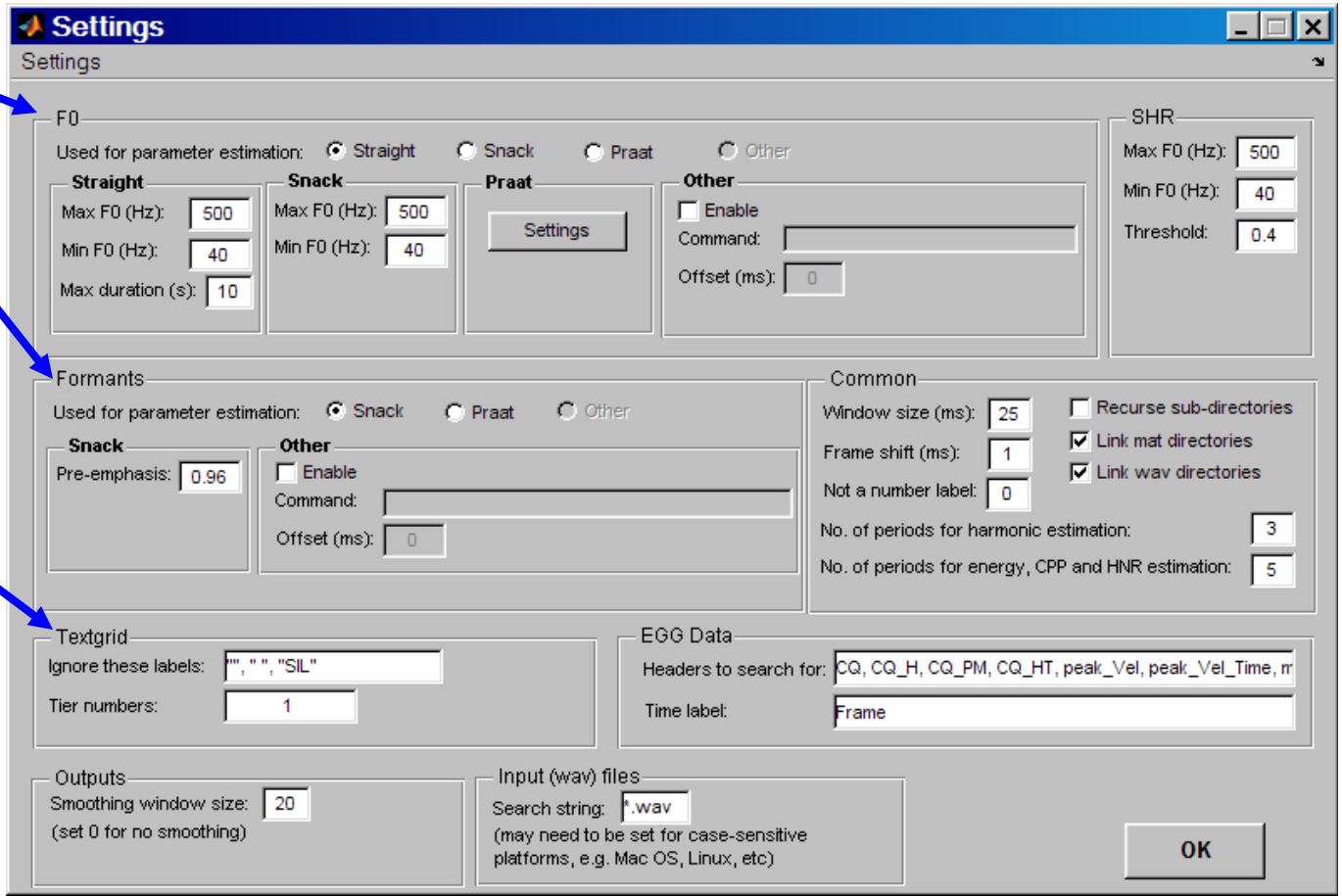
Starting VoiceSauce



- Parameter Estimation to choose which measures to calculate
- Settings to change from defaults

Settings for Parameter Estimation

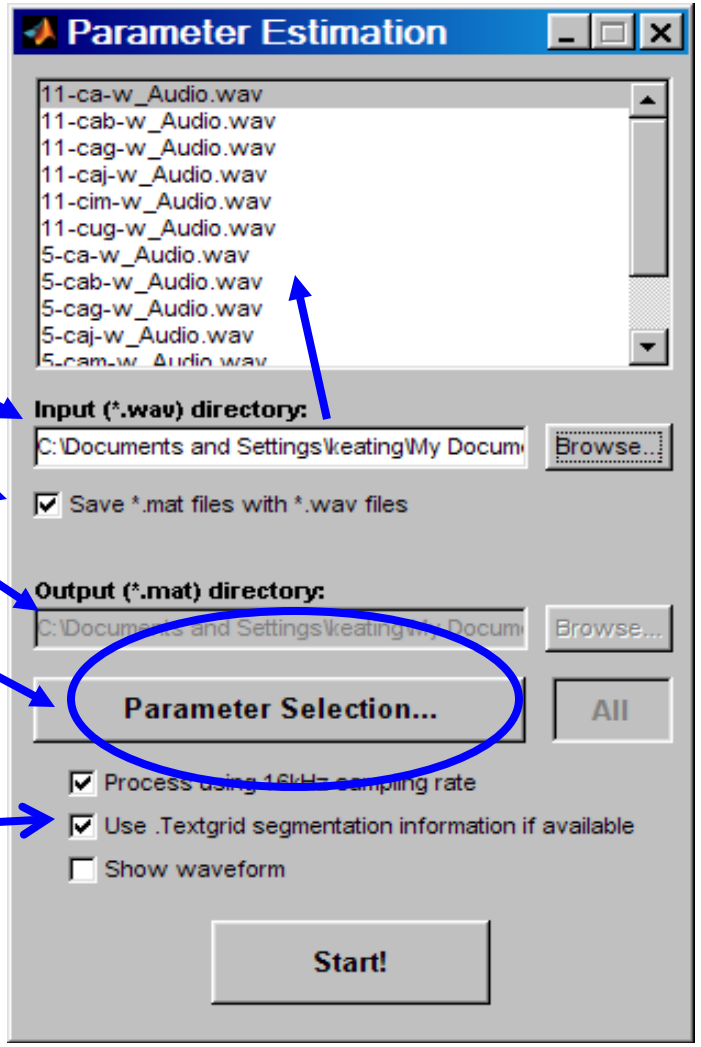
- Choose algorithms to use in calculations, and set their parameters, e.g. for F0 and Formants
- Specify how to use textgrid labels in analysis



Parameter Estimation

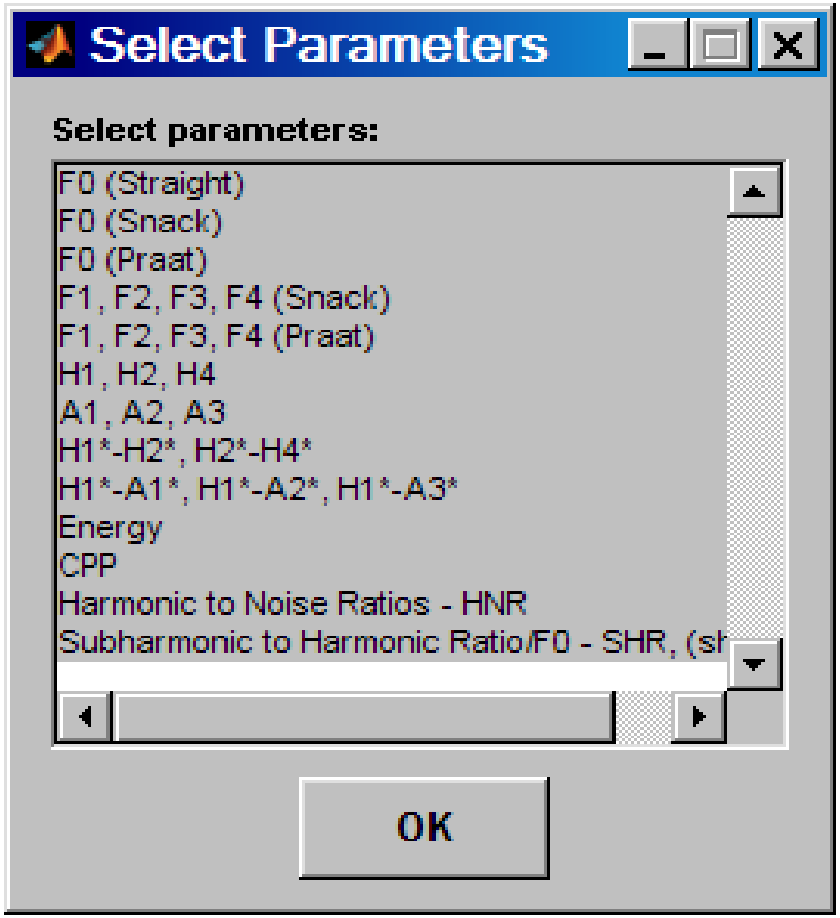
User specifies such things as:

- Where to find audio files to analyze, and where to save results
- Acoustic parameters to be calculated (next slide)
- Whether to limit analysis to intervals in Praat textgrids



Parameter Selection

Default setting is that all available parameters will be calculated:



Summary of measures

- F0 from STRAIGHT, Snack, or Praat
- H1, H2, H4
- F1-F4 and B1-B4 from Snack or Praat
- A1, A2, A3

- All harmonic measures come both corrected (*) and uncorrected

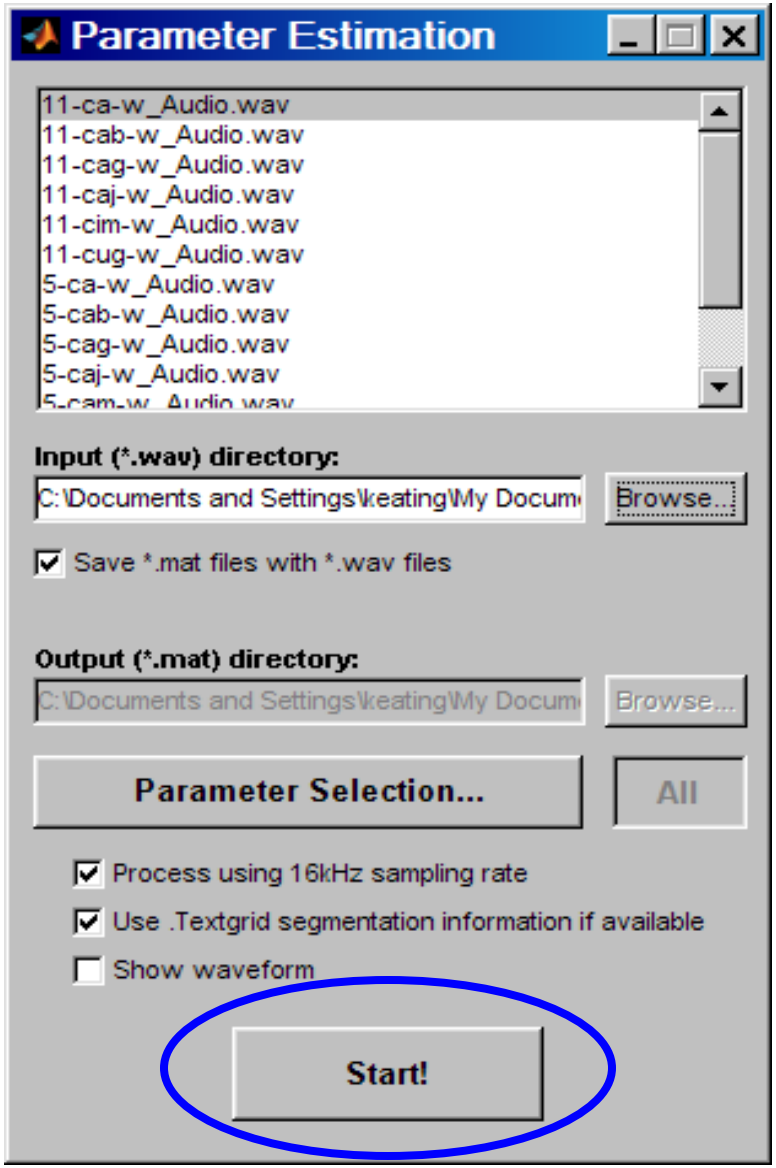
- H1-H2(*)
- H1-A1(*)
- H1-A2(*)
- H1-A3(*)
- H2-H4(*)
- Energy
- Subharmonic to Harmonic Ratio
- Cepstral Peak Prominence
- Harmonic to Noise Ratios (4 freq. bands)

Information about algorithms used in VoiceSauce

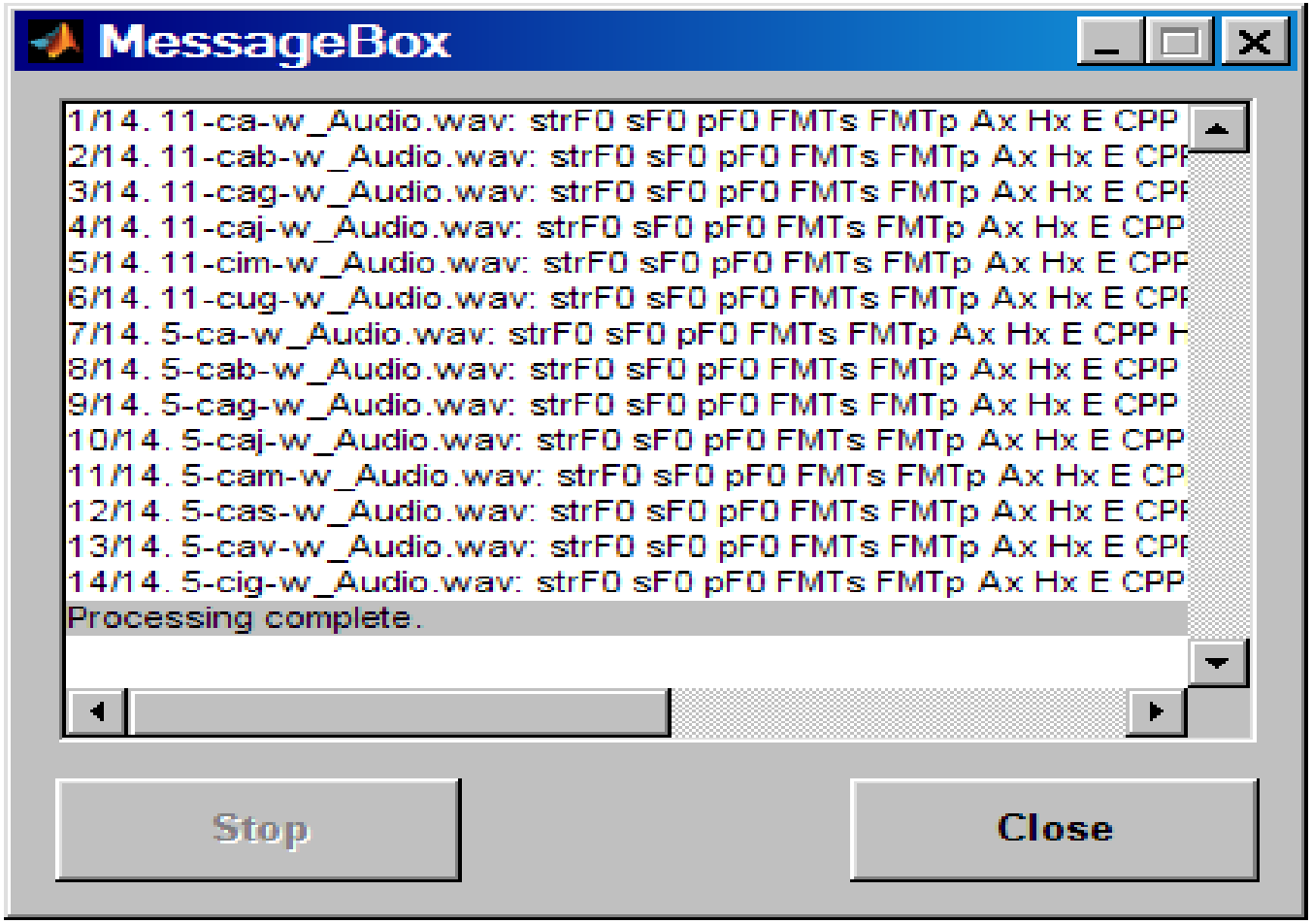
-Available in written paper

-And in question period

Running Parameter Estimation

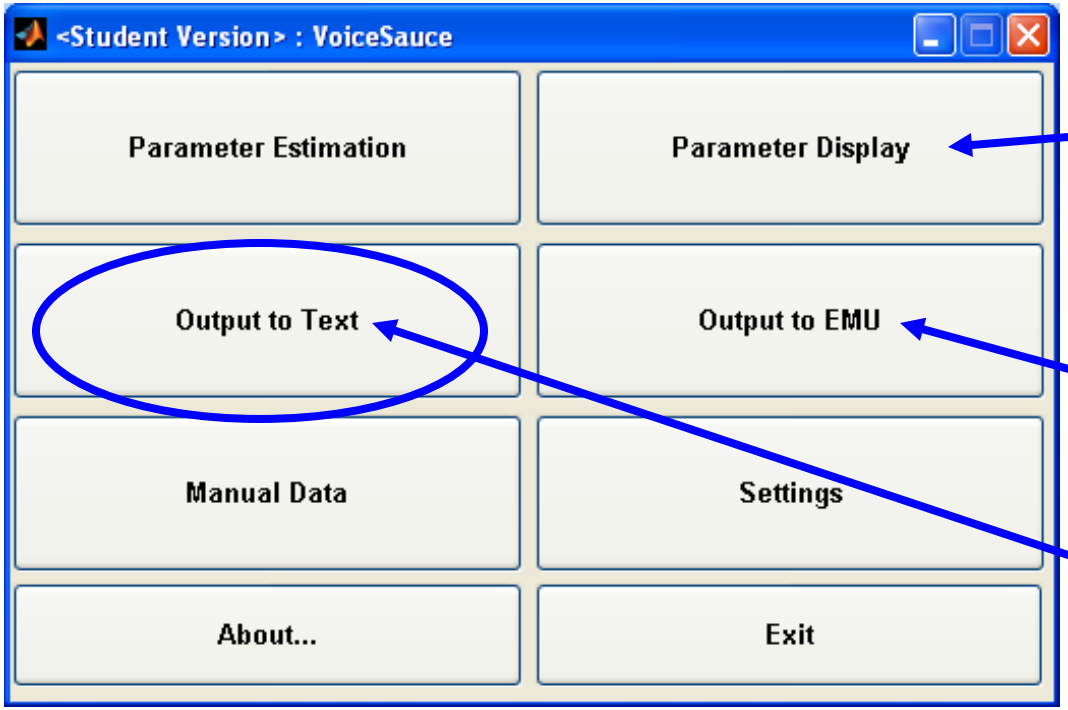


Running Parameter Estimation: onscreen progress report



VoiceSauce outputs

The initial output from VoiceSauce's Parameter Estimation is a set of binary Matlab MAT-files, one per input WAV file. Users can:



- View calculated parameter values
- Output values in Emu format
- Output values to text file

Output to Text

Use textgrids to find labeled “segments”

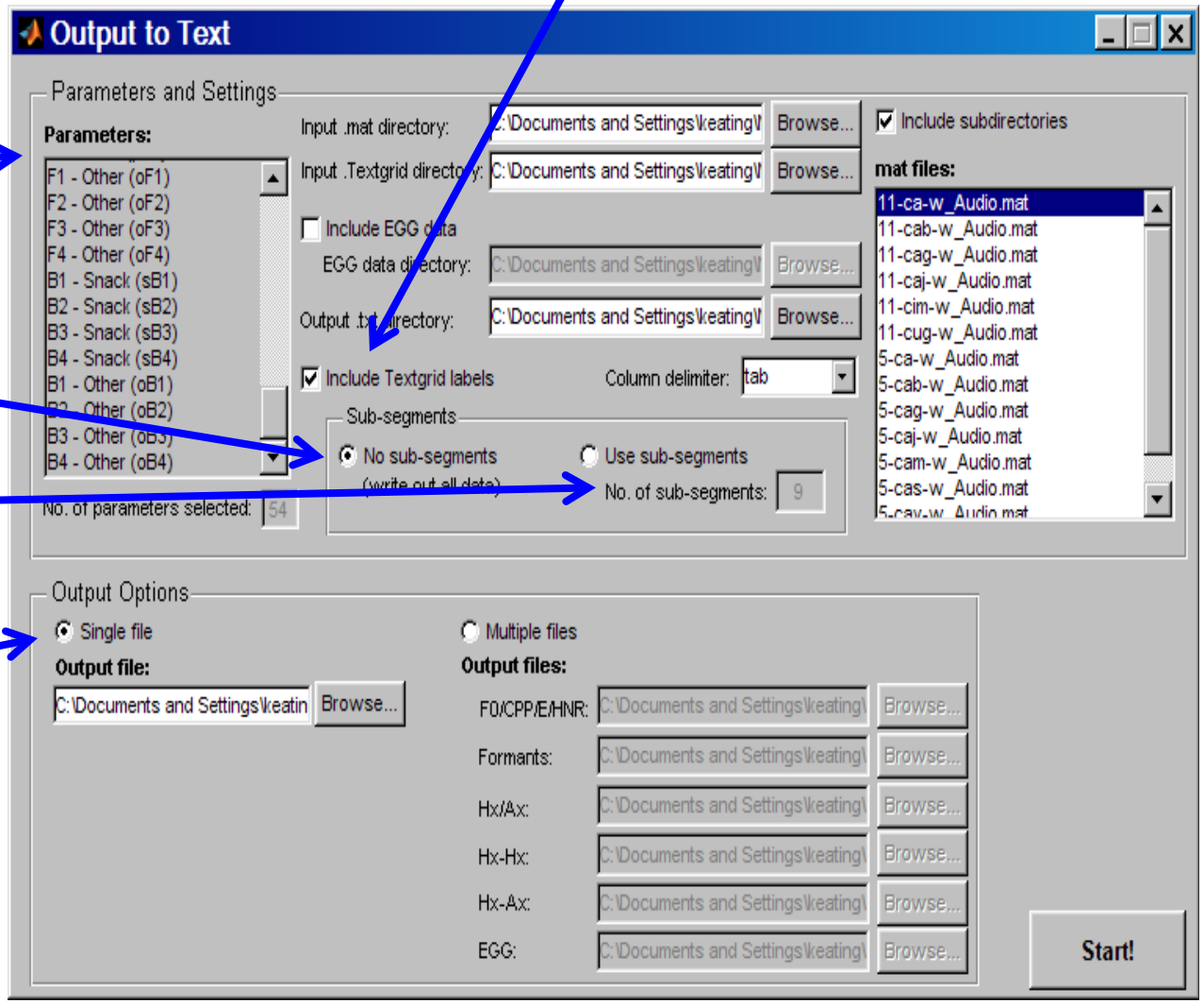
User specifies which parameters to output

Write out the data

- all values (at frame interval)
- **averages over N sub-segments**

(= time-normalization)

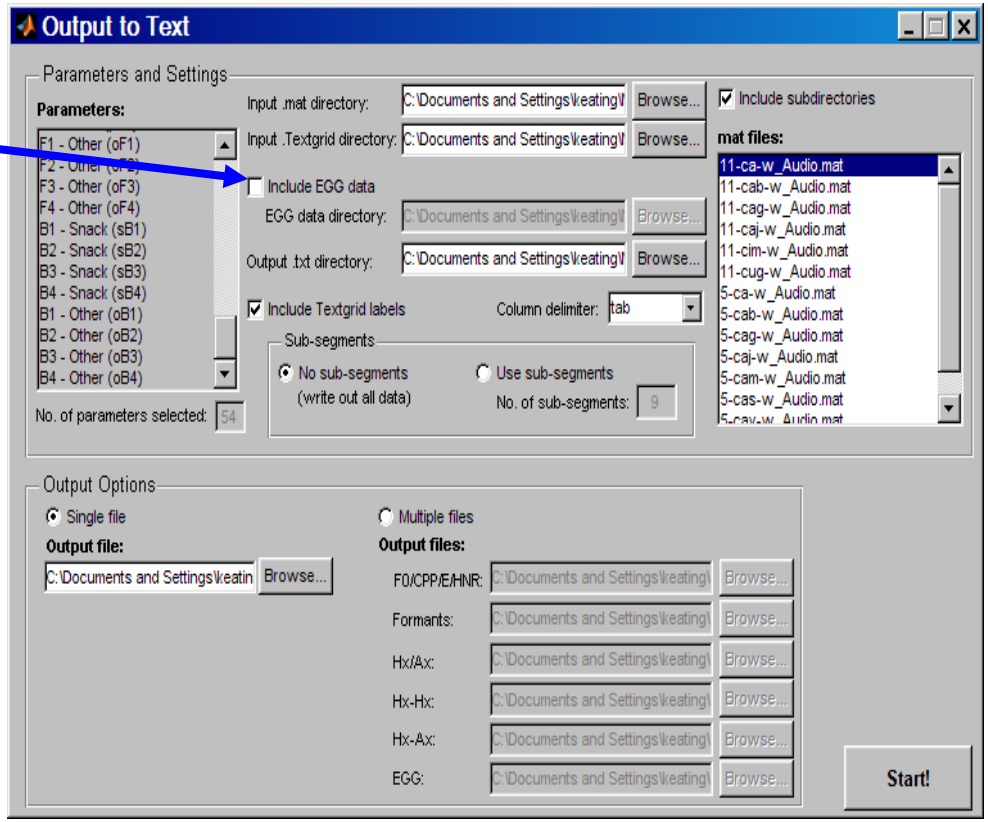
Output can be one giant text file with all parameters, or separate smaller text files with subsets of parameters



ElectroGlottoGraphic data

- EggWorks, a free program by Henry Tehrani, computes several EGG measures from EGG recordings, in batch mode

Its output file can be included as an input to VoiceSauce's output step



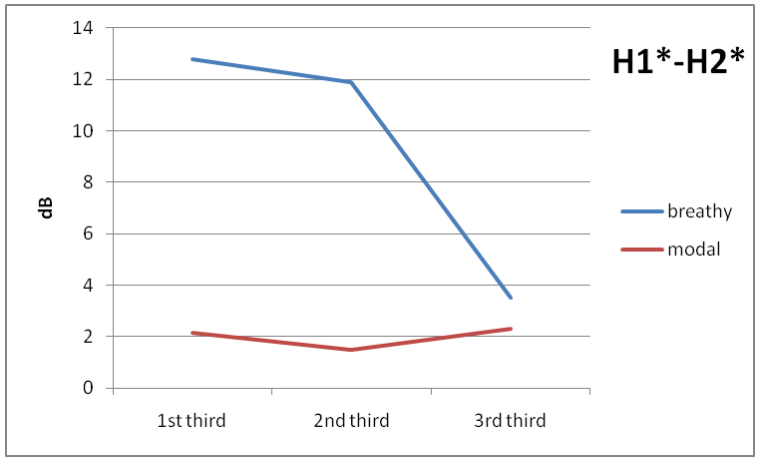
Sample from an output file

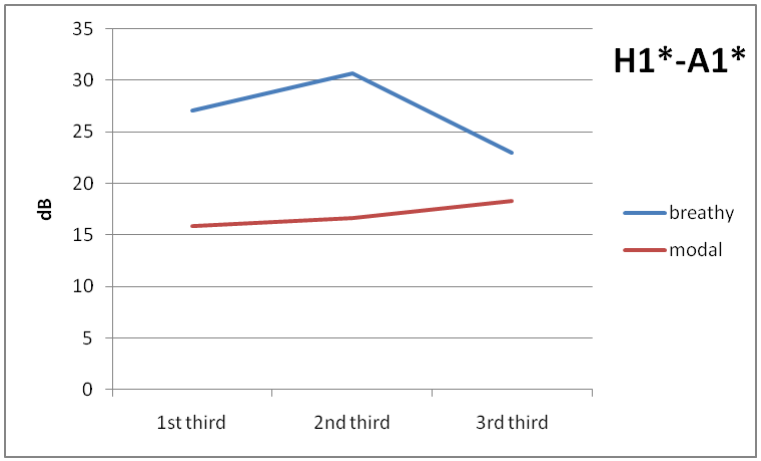
	A	B	C	D	AC	AD	AE	AF
1	Filename	Label	seg_Start	seg_End	H1H2c_mean	H1H2c_means001	H1H2c_means002	H1H2c_means003
2	HmongF-g.mat	a..	802.786	1137.87	9.391	12.799	11.895	3.513
3	HmongF-j.mat	a	1033.287	1320.53	1.993	2.148	1.506	2.311
4	MazM10-LoCr.m	ae1_	268.444	446.972	-2.214	-4.091	-1.439	-1.151
5	MazM10-LoMo.n	ae1	273.524	486.919	3.038	-2.169	4.125	6.877
6	MazM7-LoCr.mat	ae1_	233.71	472.907	-2.875	-5.087	-5.111	1.333
7	MazM7-LoMo.m	ae1	397.892	623.561	-0.652	-3.443	-1.779	3.001
8	MazM8-LoCr.mat	ae1_	185.046	367.817	-4.297	-5.899	-5.818	-1.308
9	MazM8-LoMo.m	ae1	312.611	486.056	0.842	-1.489	-1.993	5.65
10	MazM9-LoCr.mat	ae1_	321.307	574.072	0.11	-0.817	-1.532	2.641
11	MazM9-LoMo.m	ae1	422.9	615.349	-0.722	-1.512	-1.152	0.432
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

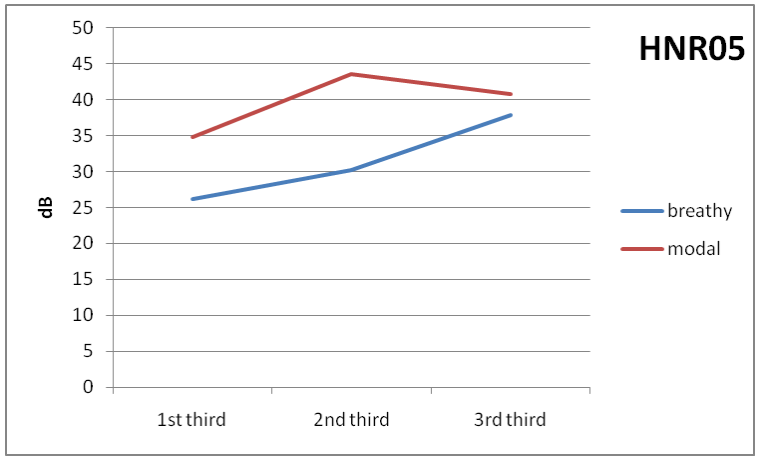
Example of results

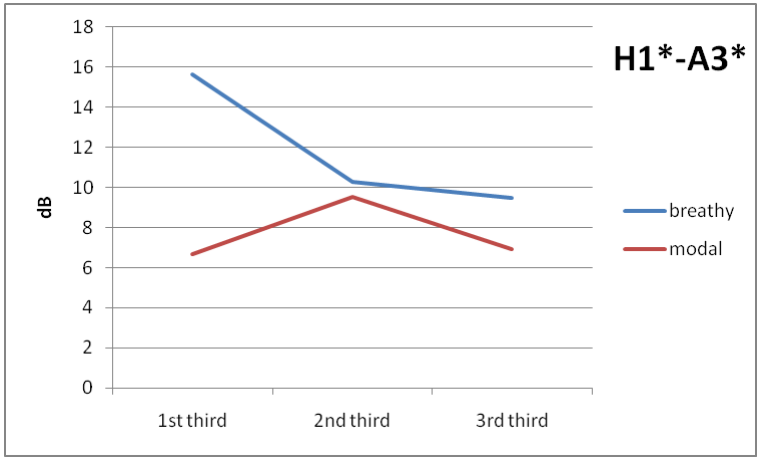
Hmong female, 1 token each

/ca/: breathy vs. modal





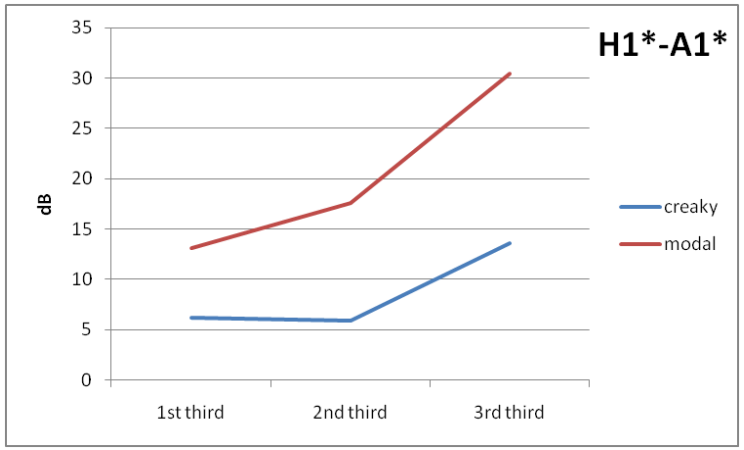


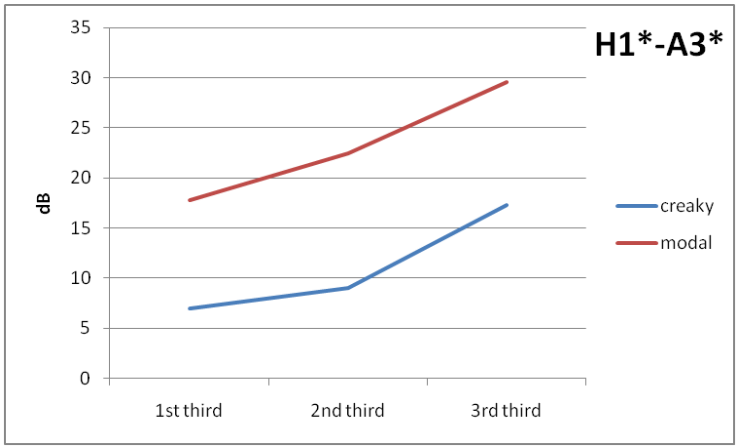


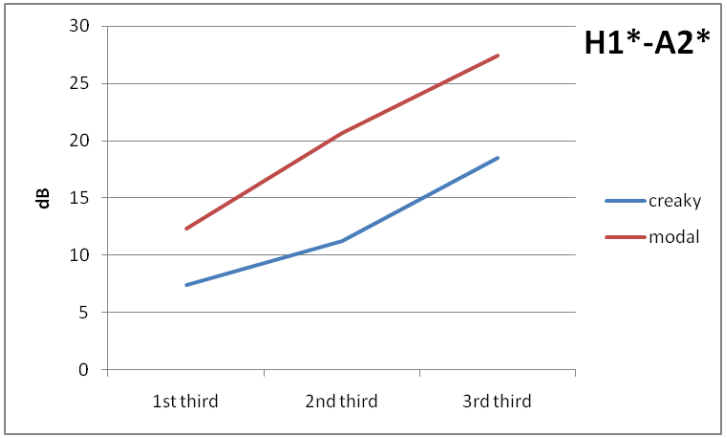
Example of results

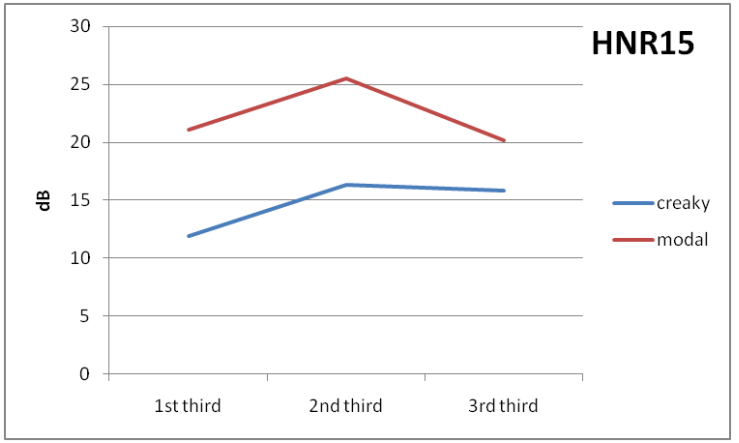
4 Mazatec males, 1 token each

/jæ[1]/: creaky vs. modal









Comparing VoiceSauce to other methods

- Compare VoiceSauce's **H1-H2** to
  - By hand measurements, taken from PCQuirer's FFT spectra

    (traditional method – not a benchmark, but common in the literature)
  - Praat (Boersma & Weenink 2008)
- Same speech materials (from Vicenik 2010) analyzed by all 3 methods

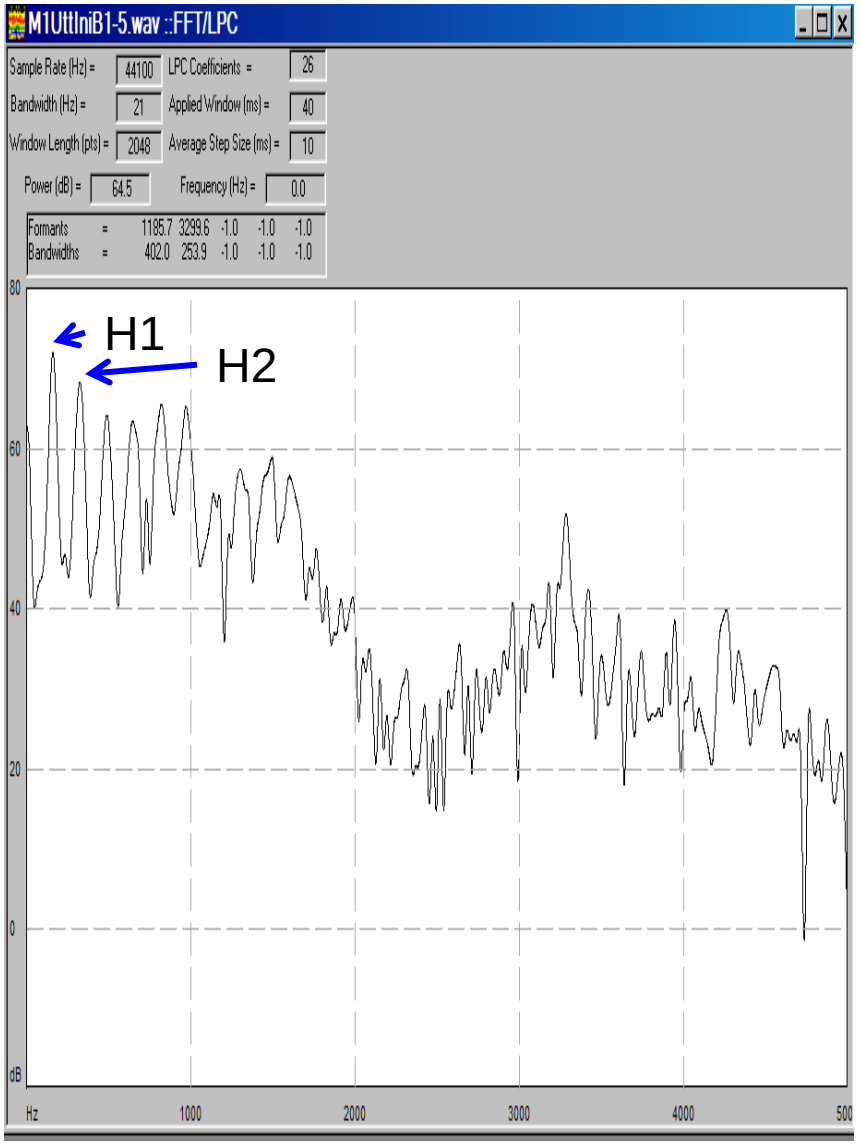
Speech corpus

- 5 speakers of Georgian, middle-aged women from Tbilisi, Georgia
- low vowel [a]
- after 9 Georgian stops
  - Three places of articulation – bilabial, alveolar, velar
  - Three stop types - voiceless aspirated, voiced, ejective – which affect the phonation of the following vowel
- 678 tokens total

H1-H2 by hand

Measured in PCQuirer

- FFT spectrum with 40 ms window (= 21 Hz bandwidth), starting at vowel onset
- Manually marked and logged H1 & H2 using cursor
- *Very* slow
- If spectrum over this window did not show clear H1 & H2, token could not be analyzed



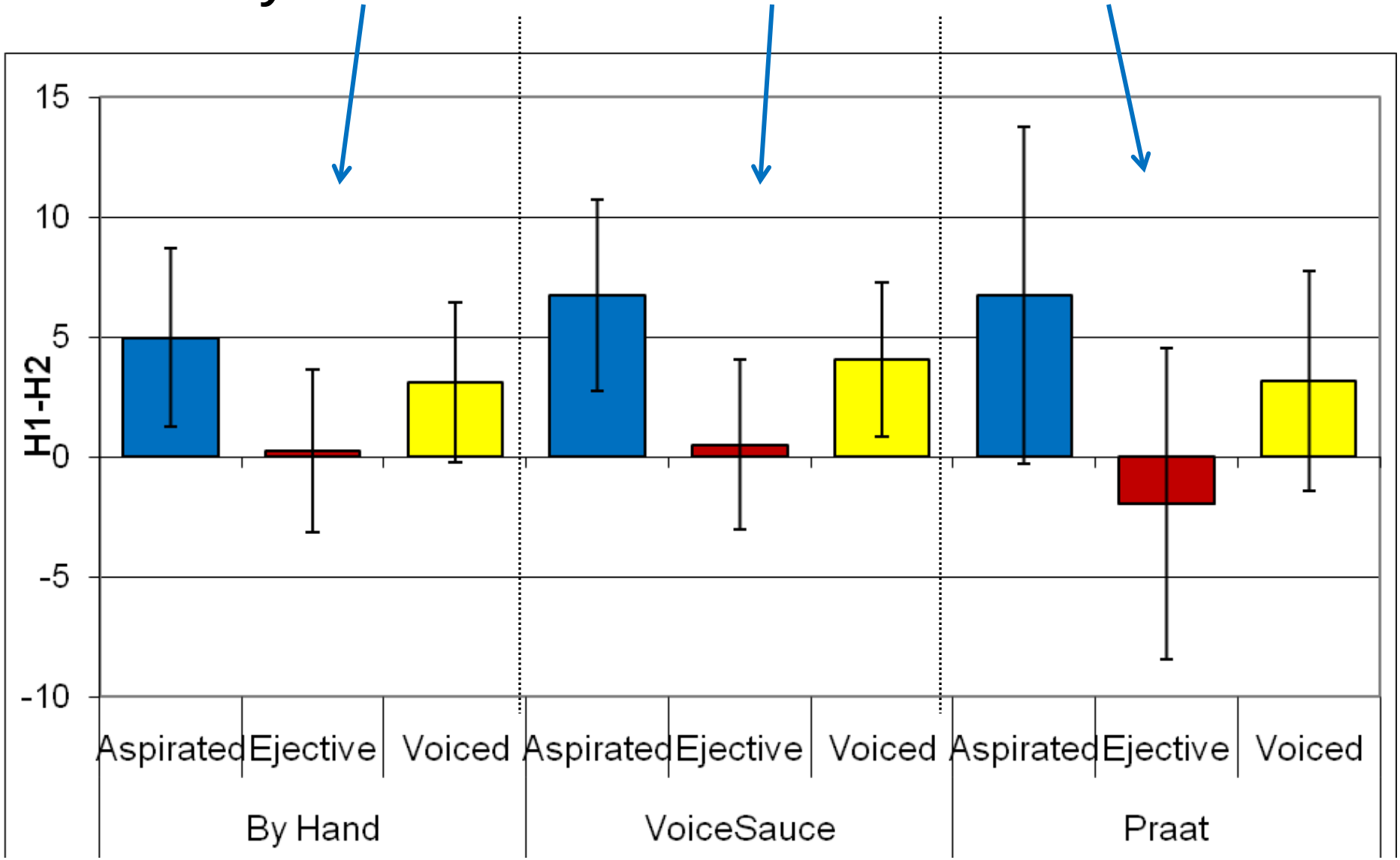
H1-H2 by Praat

- Using a new script by Chad Vicenik based on one by Bert Remijsen – available on our UCLA Praat script page (comparison here holds only for this particular script)
- Value for first third of each vowel (~ 40 ms)
  - 1/3-vowel window FFT > long term average spectrum
  - F0 found at midpoint of interval
  - that F0 (±10%) used to find harmonics in LTAS
- Makes several measurements, but here we extracted only H1 & H2

H1-H2 by VoiceSauce

- Also over first third of each vowel
- Unlike other methods, H1 & H2 calculated every msec, then averaged over the interval (this is *not* a long-term average spectrum)
- Unlike other methods, spectrum window is pitch-synchronous
- Here, *uncorrected* spectral magnitude measures were taken, for comparability with other methods

H1-H2 for three consonant manners
By-hand vs. VoiceSauce vs. Praat



Differences in results

- Overall, results from the 3 methods are similar
- Measurements made by hand have the smallest H1-H2 range; Praat script has largest H1-H2 range (larger category differences)
- BUT Praat script measurements also have greater variation than from VoiceSauce or by hand – about twice as much after ejectives/aspirates

What makes Praat method more variable than VoiceSauce?

- H1 and H2 measures both more variable
- Praat script's F0 and LTAS aren't matched
- Possibly relevant VoiceSauce features:
  - the STRAIGHT pitch-tracker is very good for non-creaky phonation
  - pitch-synchronous window for FT
  - having F0 values every msec avoids discontinuities, improves amplitude estimation
  - harmonic amplitudes are found by optimization, which is equivalent to using a very long FFT window
  - multiple values are averaged over the interval

Summary of comparison

- VoiceSauce maximized cross-category differences while minimizing within-category variability in H1-H2
- VoiceSauce also includes other measures, corrections for formants
- User-friendly, fast, no scripting

Conclusion

- VoiceSauce is available in Matlab and freestanding Windows versions for free download (*http://www.ee.ucla.edu/~spapl/voicesauce/)*
- We hope that VoiceSauce will be a useful and easy-to-use tool for researchers interested in multiple voice measures over running speech - from linguistic phonetics, prosody, sociophonetics, and other areas using speech data.

Acknowledgments

- NSF grant BCS-0720304
- Code contributors: Henry Tehrani and Markus Iseli
- VoiceSauce beta users: Christina Esposito, Marc Garellek, Sameer Khan, Jianjing Kuang, H. Pan
- Co-PIs: Abeer Alwan and Jody Kreiman

Extra slides

VoiceSauce algorithms

VoiceSauce algorithms: F0 estimation

First, F0 is found:

- STRAIGHT algorithm (Kawahara et al. 1998) is used by default, at 1 ms intervals
- Snack Sound Toolkit (Sjölander 2004) and Praat (Boersma & Weenink 2008) can also be used to estimate F0 at variable intervals

VoiceSauce algorithms: Harmonic magnitudes

- Harmonic spectra magnitudes computed pitch-synchronously, by default over a 3-cycle window
  - This eliminates much of the variability in spectra computed over a fixed time window
- Harmonics found using standard optimization techniques to find the maximum of the spectrum around the peak locations as estimated by F0
  - This enables a much more accurate measure without relying on large FFT calculations

VoiceSauce algorithms: Formant estimation

- Snack Sound Toolkit is used to find the frequencies and bandwidths of the first four formants, using as defaults the covariance method, pre-emphasis of .96, window length of 25 ms, and frame shift of 1 ms (to match STRAIGHT).
- Praat's Burg algorithm can also be used

VoiceSauce algorithms: Formant corrections

- Following Hanson (1997) and Iseli et al. (2007), spectral magnitudes can be corrected for the effect of formants (frequencies and bandwidths)
- Corrected every frame using the measured formant frequencies, and estimates of bandwidths (Hawks & Miller 1995)

VoiceSauce algorithms: Energy

- Root Mean Square (RMS) energy is calculated at every frame over a variable window equal to five pitch periods.
- The variable window effectively normalizes the energy measure with F0 to reduce the correlation between them.

VoiceSauce algorithms: Subharmonic to Harmonic Ratio

- Proposed by Sun (2002) to quantify the amplitude ratio between subharmonics and harmonics
- Derived from the summed subharmonic and harmonic amplitudes calculated in the log domain using spectrum shifting
- Implemented using Sun's algorithm and code
- May be especially relevant for characterizing speech with alternating pulse cycles

VoiceSauce algorithms: Cepstral measures

Harmonic to Noise Ratios

- De Krom (1993)
- Variable window length equal to five pitch periods
- Energy of harmonics is compared with noise floor
- 3 frequency ranges:
  - 0-500Hz
  - 0-1500Hz
  - 0-2500Hz
  - 0-3500Hz

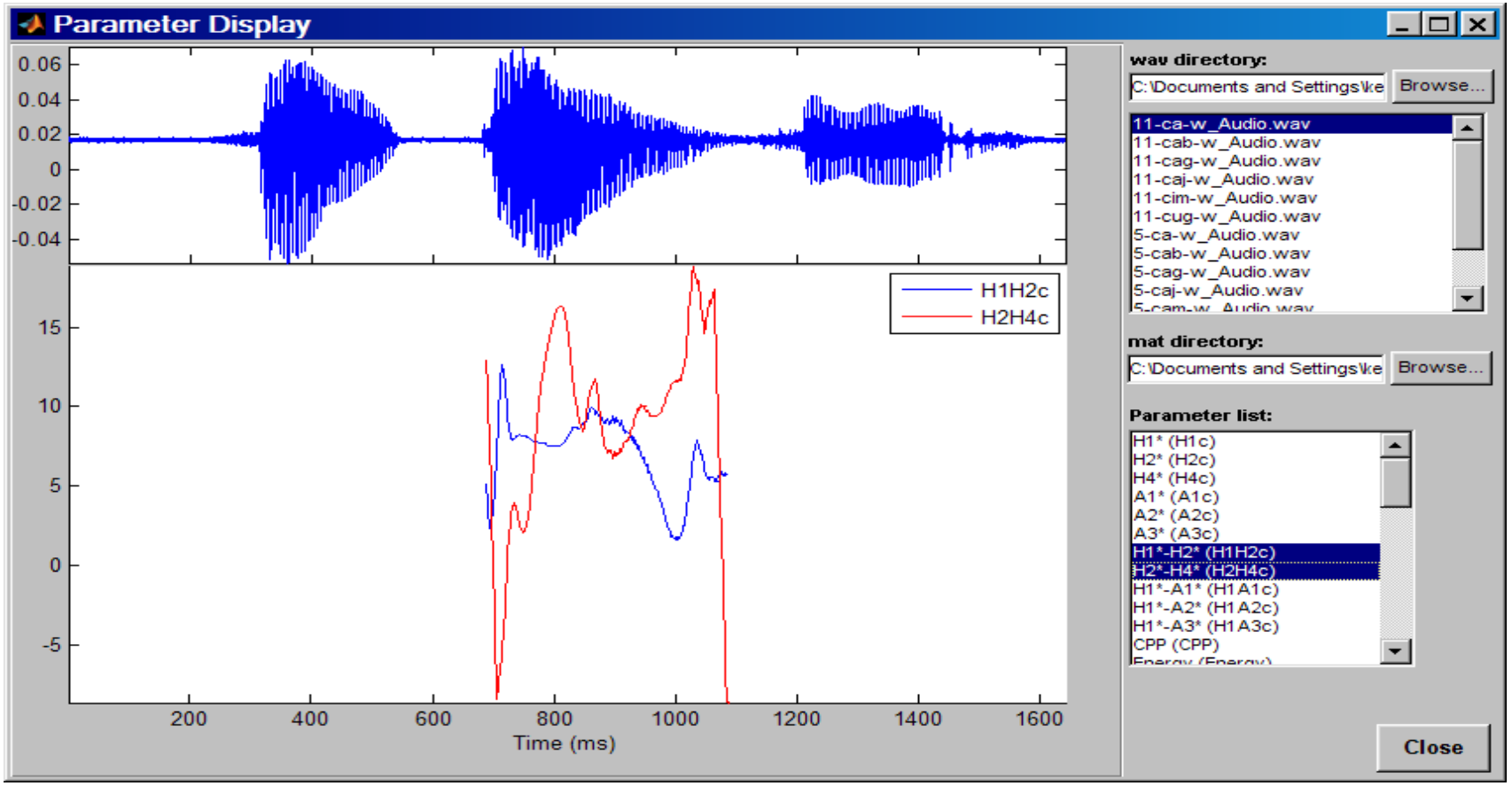
Cepstral Peak Prominence

- Hillenbrand et al. (1994)
- Variable window length equal to five pitch periods
- Cepstral peak is normalized to a linear regression line between 1 ms and the maximum quefrency
- Entire frequency range

Other extra slides

Parameter display

Displays (multiple) parameters with the waveform of a single audio file, for quick visual checks of data

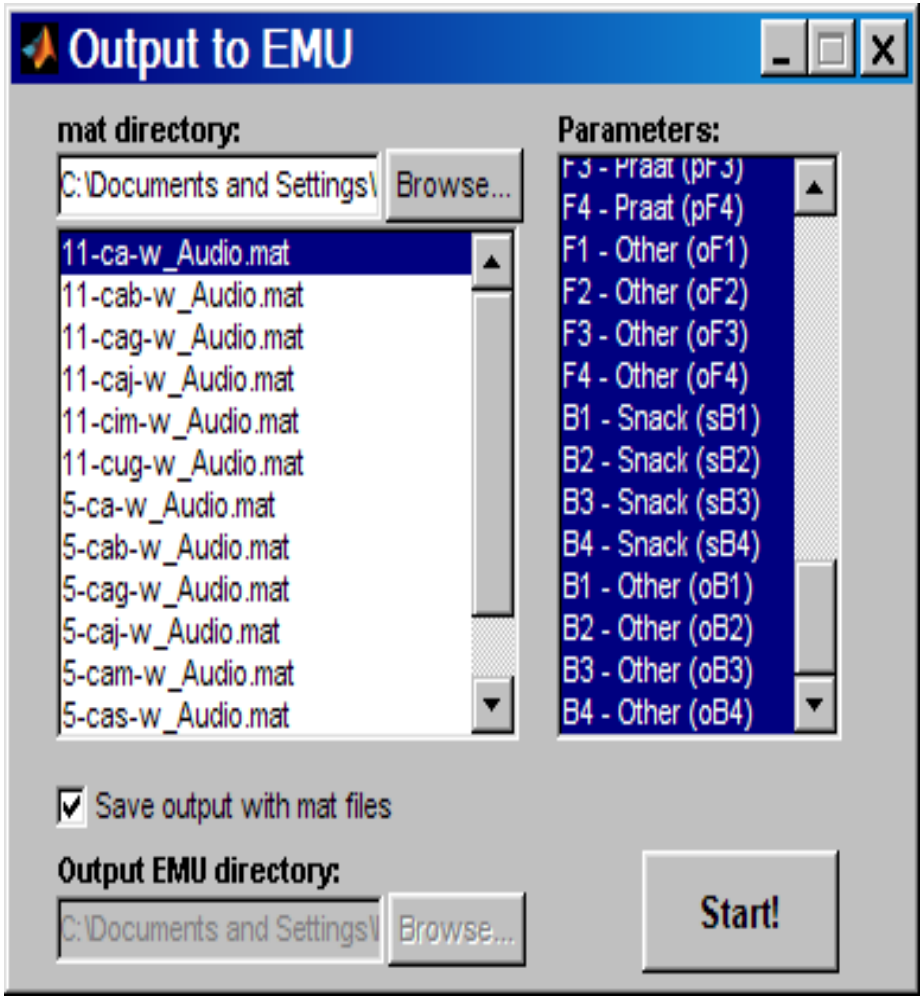


Manual Data option

- Most voice measures depend on F0 and formant measures, but these are often problematic
- Users can try different algorithms for calculating these (e.g. for 1 of the sample files, Snack's F1 was wrong, so Parameter Estimation was re-run using Praat's formants)
- Alternatively, users can provide hand-corrected measure(s) in a new data file which is loaded into VoiceSauce

Output to Emu

- For use in Emu speech databases (Harrington, 2010)
- Outputs Emu's trackdata files in SSFF format, 1 track file per parameter per audio file
- Can view, query, analyze in Emu, or in R using Emu library



Sample display in Emu