Foundations of Data and Audio-Visual Analytics ($FODA^2VA$)

Mark Hasegawa-Johnson, Kai-Hsiang Lin, Xiaodan Zhuang, Camille Goudeseune, Sarah King, Thomas Huang, and Hank Kaczmarski

University of Illinois

FODAVA Review Meeting, December 9, 2010

Definition of Visual Analytics

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces." (*Thomas and Cook, Illuminating the Path*)

The goal of our research is to make
AUDIO-VISUAL DATA
available to
VISUAL ANALYTICS.

Motivation

"People use visual analytics to synthesize information and derive insight from massive, dynamic, ambiguous and often conflicting data; detect the expected and discover the unexpected; provide timely, defensible, and understandable assessments; and communicate assessment effectively for action." (*ibid*)

One Equation

The core research problem in audio-visual analytics can be summarized in one equation:

$$f^* = \arg\max_f \ \mathcal{I}(Y, \psi(f(X))) \tag{1}$$

$$\text{s.t.} \quad M(f) \leq M_{max}$$

Datum. X is an audio spectrogram.

Labels. Y is what the analyst "should" notice.

Physical Image. $f(X)$ is displayed.

Perceived Image. $\psi(f)$ is what the user sees.

Information. $\mathcal{I}(Y, \psi)$ is the information the user derives from ψ.

Memory Consumption. $M(f)$ is the RAM required.

1. Audio Event Detection: Motivation

The target labels, Y, are words in English text. If it is possible to compute $f(X) = Y$, then we should do so: no other representation has higher mutual information.

Audio Event Detection: Results (Zhuang et al., 2010)

	ap	cl	cm	co	ds	kj	kn	kt	la	pr	pw	st	Average
MFCC	**78.3**	**26.9**	29.5	24.2	56.3	**39.9**	7.7	0.0	39.0	35.2	14.1	28.7	28.2
FB	34.5	21.8	25.4	24.9	38.9	27.2	11.7	0.0	49.1	13.8	11.7	28.1	27.8
Adaboost	44.4	25.5	31.3	31.2	57.3	33.2	13.5	1.9	51.3	36.7	17.6	36.8	34.0
Adaboost+T	52.6	21.9	37.2	**51.3**	**63.0**	29.6	11.5	0.0	54.2	**42.7**	25.8	34.6	35.3
Adaboost+S	44.4	25.0	33.7	31.2	56.6	33.2	**20.9**	35.5	51.3	36.7	19.2	41.3	37.5
Adaboost+T+S	52.6	21.5	**37.4**	47.9	**63.0**	29.6	13.6	**44.8**	**58.6**	**42.7**	**26.7**	**44.4**	**41.2**

AED-Accuracy (%). Columns are different types of event.

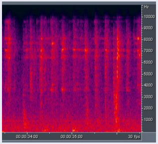
- Adaboost = soft Bayes (1a)
- T = tandem nnet+HMM
- S = GMM supervector (1b)

2. Spectrograms: Motivation

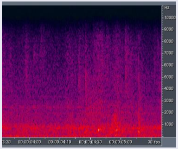
- Analysts like spectrograms and waveform plots; they know how to get information from them.
- Even naïve human subjects prefer a time-frequency plot to any other display.

Problem

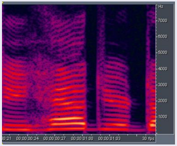
Spectrogram settings for non-speech audio are non-obvious.



Key Jingle



Footsteps



Speech

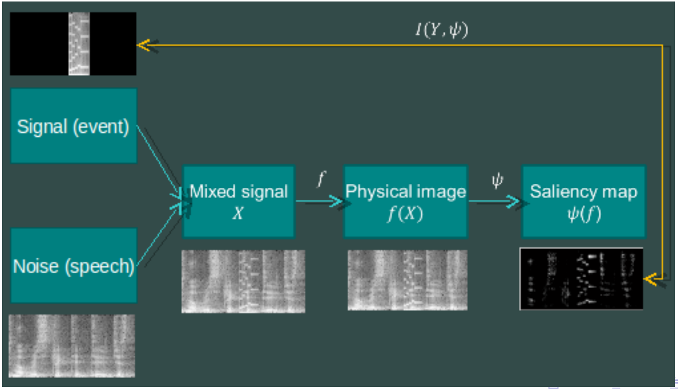
Experimental Solution

Multi-day audio timeliner:

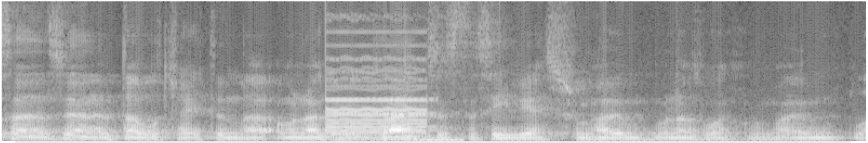
- Load coarse features into RAM
- Allow user to zoom continuously from full-day view to millisecond view

3. Explicitly Optimized Features: Motivation

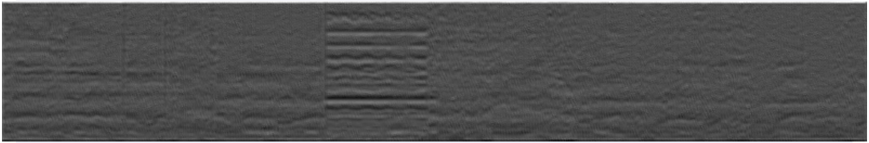
- Start with the spectrogram ($f(X) = X$), because analysts and computer gamers love spectrograms.
- Adjust $f(X)$ in order to maximize information.



Input: Non-speech “Easter Eggs” hidden in speech



Output: Speech is attenuated, “easter egg” emphasized



Contributions to the FODAVA Community

The goal of our research is to make
AUDIO-VISUAL DATA
available to
VISUAL ANALYTICS.

Specific Contributions

- Meeting room data annotated with **audio salience annotations**, http://isle.illinois.edu/sst/data/salientevents/
- **Timeliner** and **Milliphone** audio visualization tools
- Publications: two in *Pattern Recognition Letters*, several in conferences, presentations at NVAC and NIPS
- Currently in development: **Audio-Visual Analytics Homepage** (expected 12/31/2010)