Speaker Datasets for Research

OR

FRE 702(b), Daubert and the Problem of Mismatched Conditions

Peter T. Higgins

James L. Wayman

Federal Rule of Evidence Rule 702. Testimony by Expert Witnesses (2013)

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

(a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;

(b) the testimony is based on sufficient facts or data;

(c) the testimony is the product of reliable principles and methods; and

(d) the expert has reliably applied the principles and methods to the facts of the case.

Daubert v. Merrell Dow Pharma (1993)

- "…scientific validity for one purpose is not necessarily scientific validity for other, unrelated purposes."
- "…in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error, see, e.g., United States v. Smith, 869 F.2d 348, 353-354 (CA7 1989) (surveying studies of the error rate of spectrographic voice identification technique)…"

Hypothesis

- Error rates for biometric and forensic methods are highly dependent upon the data
- Forensic case data is collected opportunistically under conditions that cannot completely match any of the research databases
- Error rate measures using research databases cannot be a good predictor of "potential rate of error" in forensic testimony if the research and forensic collection conditions are not the same
- The FRE term "sufficient facts or data" implies that future research databases should be collected under widely varying conditions, including human factors, sensors and protocols.
- Under the best conditions our estimates of "potential rate of error" for case data will be an informed guess.
- If presented in court, error rate guesstimates should be expressed as far less than certain.

Error Rates are Dependent on Data Used

Some evidence from multiple modalities

Fingerprint Error Rates

- NIST Fingerprint SDK Test - 2010
- Morpho results -

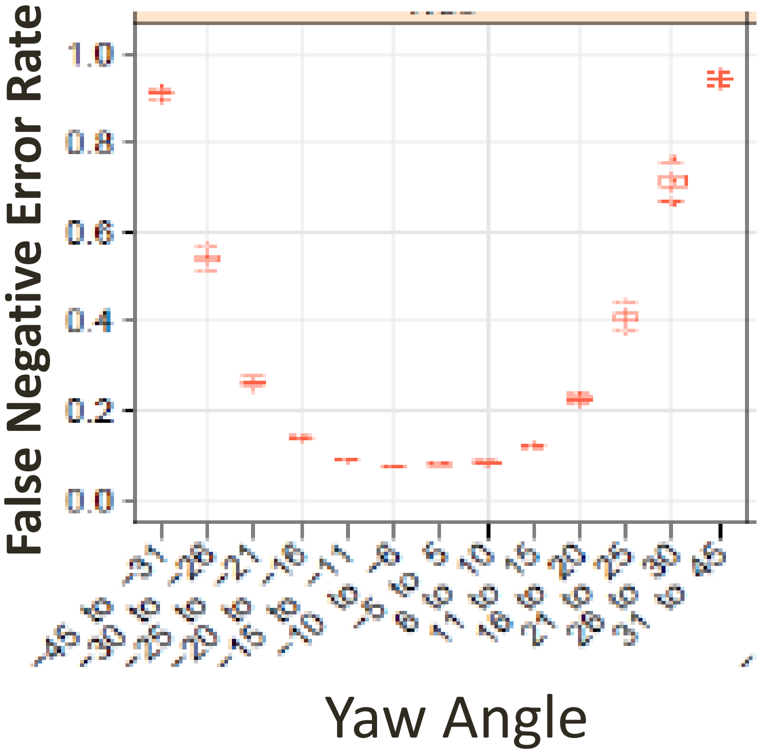
	DHS2	DOS	POE	POEBVA
False negative rate =	0.0079	0.0005	0.0004	0.0003

- At false positive rate of 0.0001

- False negative error rate with DHS2 data is 26 times higher than with POEBVA data

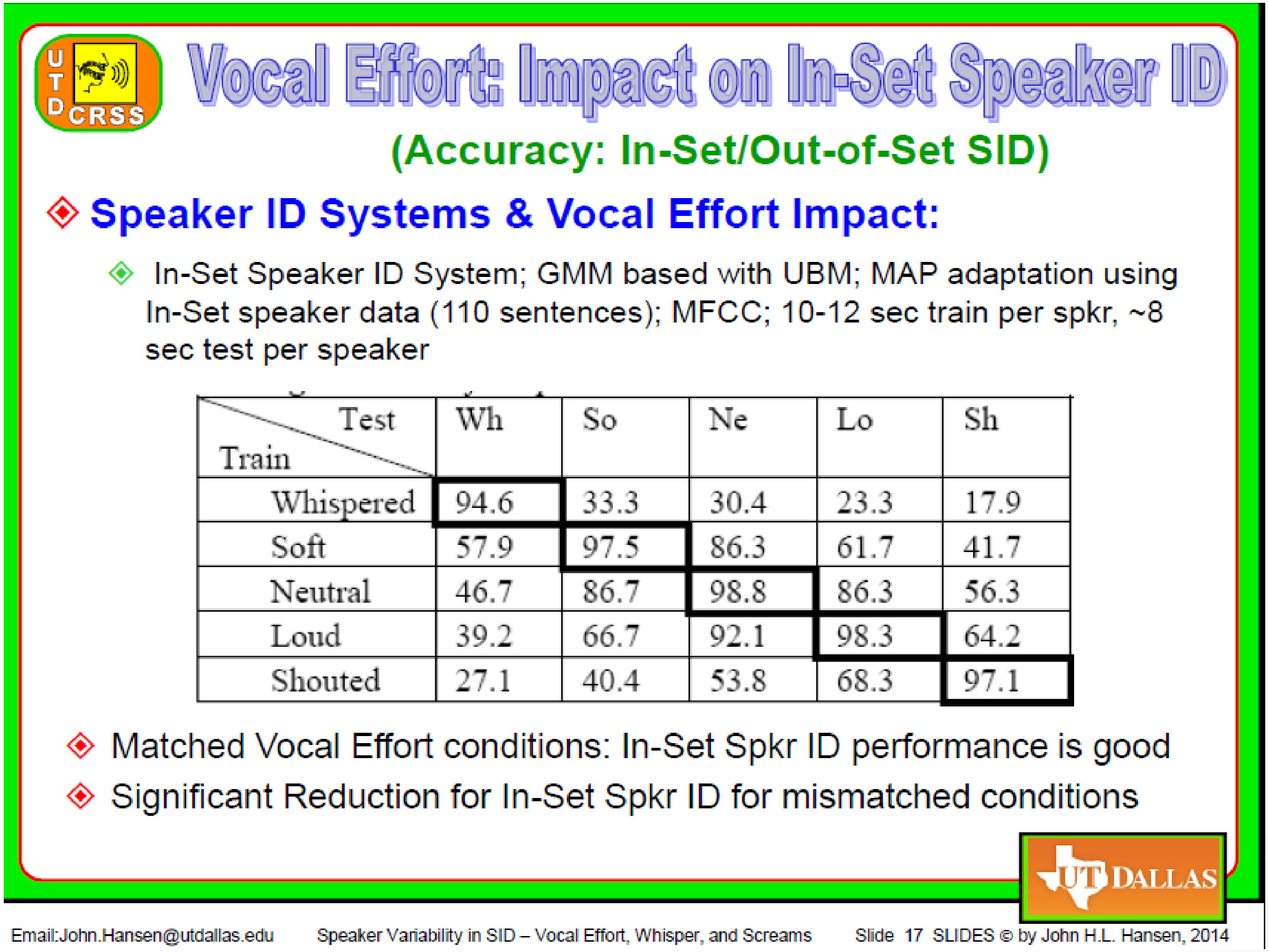
Face Recognition Error Rates

- NIST Multiple Biometrics Evaluation (2011)
- False positive error rate of 0.001



Speaker Recognition Error Rates

(Thank you, Prof. John Hansen)



Vocal Effort: Impact on In-Set Speaker ID

(Accuracy: In-Set/Out-of-Set SID)

Speaker ID Systems & Vocal Effort Impact:

- In-Set Speaker ID System; GMM based with UBM; MAP adaptation using In-Set speaker data (110 sentences); MFCC; 10-12 sec train per spkr, ~8 sec test per speaker

Train \ Test	Wh	So	Ne	Lo	Sh
Whispered	94.6	33.3	30.4	23.3	17.9
Soft	57.9	97.5	86.3	61.7	41.7
Neutral	46.7	86.7	98.8	86.3	56.3
Loud	39.2	66.7	92.1	98.3	64.2
Shouted	27.1	40.4	53.8	68.3	97.1

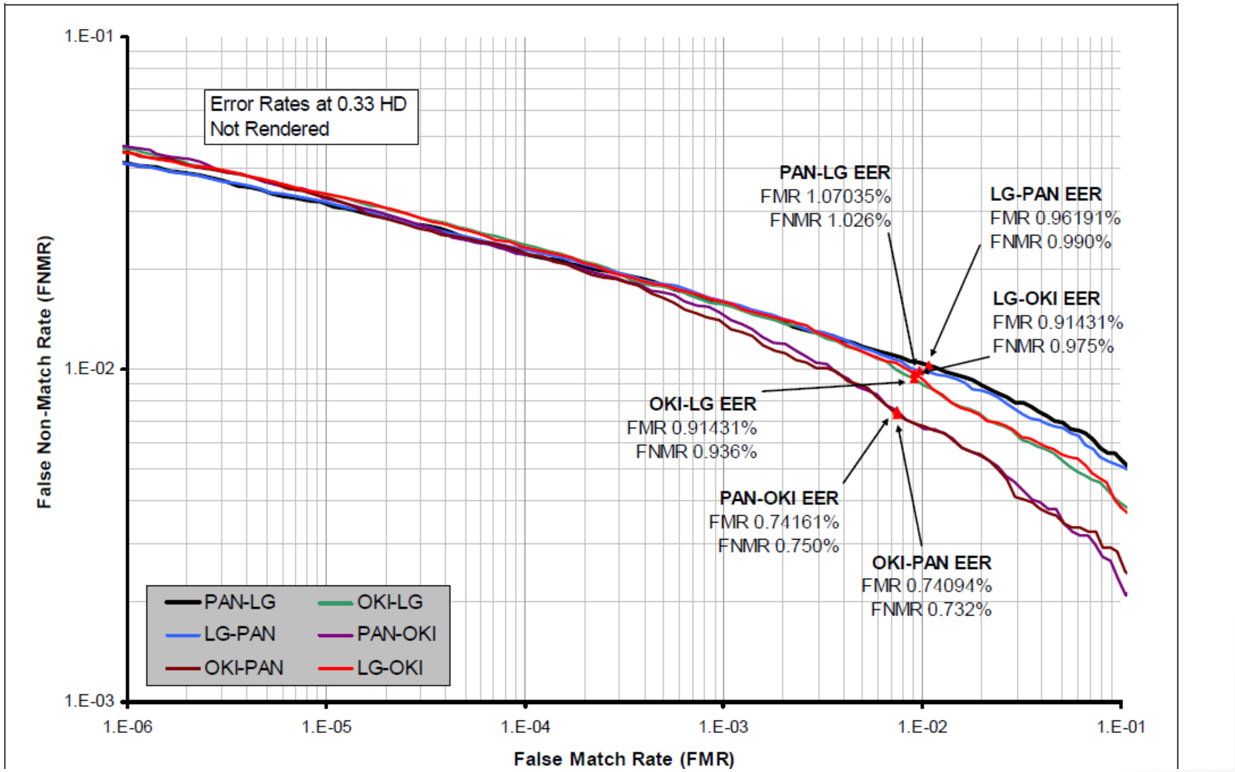
- Matched Vocal Effort conditions: In-Set Spkr ID performance is good
- Significant Reduction for In-Set Spkr ID for mismatched conditions

Email:John.Hansen@utdallas.edu Speaker Variability in SID – Vocal Effort, Whisper, and Screams Slide 17 SLIDES © by John H.L. Hansen, 2014

Iris Recognition Error Rates

- International Biometrics Group, “Independent Testing of Iris Recognition Technology”, Final Report, May 2005

7.8.1 Cross-Device Intra-Visit Enrollment Comparison DETs (Single-Attempt)



Determination of Error Rates through Sufficient Testing for Each Situation is Combinatorially Hard

- “Sufficient” testing requires “sufficient” data in all combinations of all influence variables, most of which are unknown.

A Famous Caution Against Modeling Data without the Data Being Modeled

- Can’t we just collect “clean” data, then use models to distort it to meet any conditions?

- “All models are wrong. Some models are useful … How then is the model builder to know what aspects to include and what to omit so that parsimonious models that are illuminating and useful might result from the modeling building process? We have seen how it is useless to attempt to allow for all contingencies in advance, so in practice, model building must be accomplished by iteration.” -- G.E.P. Box, *Robustness in the strategy of scientific model building*. No. MRC-TSR-1954. Wisconsin Univ-Madison Mathematics Research Center, 1979.
    - By “iteration”, Box means a processing of modifying the model through testing against real-world data.

A Famous Caution Against Use of Numbers in the Courtroom

- “Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not cast a spell over him.” -- **People v**. **Collins**. 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (**1968**)

Future Research Databases

- Wide range of populations, sensors, collection protocols, human factors
- Execute comparisons across conditions
- Models of some forms of forensic conditions
  - High risk endeavor!
- “There is no data like more data” – Robert Mercer (IBM).
- “There is no data like data with more variety” – Peter T. Higgins
- “You can't make bad audio when your starting point is broadcast quality” – Joanna Morley, Met Police, London

Recommendations

- For Speaker Data Collection:
  - Attempts must be made to collect data with variations in :
    - Speaker’s age, native language, sex, level of education, racial / regional accents
    - Microphones (cellphones, land lines, radios, Skype, Sat phone, etc.), acoustic and other environmental conditions (indoors / outdoors, distance to microphone, recording devices, channels, etc.)
    - Tone of speech (whispers, shouts, etc.)
    - Type of speech (interview, read text, recitation of known material
    - Multiple speakers (interrogator / interviewer and subject - maybe a translator) with one and with multiple channels
    - Language changes within a sample set
    - Length of collected samples (seconds to minutes) – with parallel continuous recording of sessions
    - Subject performing mental / physically demanding tasks while speaking – for some samples
  - Unclassified data from real forensic cases where a match was established
    - 911 call centers, court ordered wire taps, other

Conclusions

- Error rates for biometric and forensic methods are highly dependent upon the data
- Forensic case data is collected opportunistically under conditions that cannot completely match any research databases
- Error rate measures using research databases cannot be a good predictor of "potential rate of error" in forensic testimony if the research and forensic collection conditions are not the same
- **The FRE term "sufficient facts or data" implies that future research databases should be collected under widely varying conditions, including human factors, sensors and protocols.**
- Under the best conditions our estimates of "potential rate of error" for case data will be an informed guess.
- If presented in court, error rate guesstimates should be expressed as far less than certain.

peter@higgins-biometrics.com