USSS-MITLL 2010 Human-Assisted Speaker Recognition: Evaluation System (Part I)

Reva Schwartz (USSS); Joseph Campbell, Wade Shen, Douglas Sturim, William Campbell, Fred Richardson, Robert Dunn, Robert Granville (MIT-LL)

jpc@ll.mit.edu

NIST Speaker Recognition Evaluation Workshop
Brno, Czech Republic

25 June 2010

This work was sponsored by the Department of Defense under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- Introduction
- Audio Preprocessing
- HASR System
- Qualitative
- Automatic
- Fusion
- Results
- Mechanical Turk
- Summary

Introduction

- **United States Secret Service (USSS) teamed with MIT Lincoln Laboratory (MIT/LL) in NIST HASR1 Evaluation***
- **Completed 15-trial HASR1 Evaluation over the 8-week evaluation period**
  - **Too short an evaluation period to conduct a forensic-like process on the 150-trial HASR2 Evaluation**

- **USSS provided expert human analyst and MIT/LL provided support, tools, and automatic recognition systems**
  - **USSS: qualitative method**
  - **MIT/LL: audio preprocessing (speech enhancement and echo canceling) and automatic speaker id**
  - **USSS + MIT/LL: sample purification and fusion consultation**

* C. Greenberg, et al., "Human Assisted Speaker Recognition in NIST SRE10," *Odyssey 2010: The Speaker and Language Recognition Workshop,* Brno, Czech Republic, 28 June – 1 July 2010.

Introduction (cont.)

- Unlike conventional NIST SRE, HASR
  - Audio samples are provided one trial at a time
  - Listening to data is allowed
  - The sex of the talker(s) is not provided
  - The prior probability of a match is not provided (or inferable)
  - Costs of errors are not provided
  - Performance metric is not defined
  - ASR transcripts are not provided
  - Conditions are not specified (other than subset of SRE10)

- HASR was not designed to be an evaluation of forensic speaker comparison methods, but
  - Serves to inform forensic, and other, applications
  - Forensic laboratories were invited to participate
    - How to increase participation in the future?

HASR Audio Samples

- **15 HASR1 trials were found to be all microphone interview (~3 min*) versus telephone conversation (~5 min*) condition****
  - **Interview microphones (~15) vary over trials**
  - **Interview rooms (2) vary over trials**
  - **Noise (varying over trials) electrically added to interviewer channel**
    - **LDC's HVAC provides varying acoustic noise**

- **Samples provided in 2-channel format**
  - **Allows for analysis of person of interest (specified via "channel of interest" by NIST) and interlocutor**
    - **Conversational analysis**
  - **Enables improved speaker purification**

*Duration is reduced after purification; 65% reduction in some samples

*Brandschain. et al., "Mixer 6", *Proc. LREC'10*, Malta, 19-21 May 2010, *http://www.lrec-conf.org/proceedings/lrec2010/summaries/792.*

Audio Preprocessing

- **Prepare samples for human analysis and for automatic processing, for each sequential trial given in two-channel G.711 μ-law (8 kHz, 8-bit sampling) SPHERE format**

- **Interview recordings for both human analysis and automatic processing:**

  Source .sph → Peak normalize (90% FS), DC Bias removal → Enhancement* → Purification (in stereo) → Extract [a]**

- **Telephone recordings for human analysis:**

  Source .sph → Peak normalize (90% FS), DC Bias removal → Extract channel of interest [a or b]

*Both channels are enhanced independently. Two-stage enhancement is run on the individual channels. First, MIT/LL's stationary narrowband noise reduction (RemTones) is run. Next, MIT/LL's stationary wideband noise reduction is run (LLEnhance). Various settings of these algorithms were tried, but the default settings worked well throughout all the HASR1 trials.
**Purify by manually removing segments of the interlocutor's speech and regions overlapped speech. Editing he two-channel enhanced audio speeds the process, likely improves purification accuracy, and reduces fatigue (NIST had apparently added noise to the interviewer's channel and, at times, there was substantial HVAC noise in the interview room).

Outline

- Introduction
- Audio Preprocessing
- HASR System
- Qualitative
- Automatic
- Fusion
- Results
- Mechanical Turk
- Summary

HASR Expert-Based Process

- **Our Human-Assisted Speaker Recognition (HASR) system is an expert-based process**
  - **Adopted from general forensic-phonetics methodology**
  - **Expert conducts qualitative analysis**
    - **Considerable variation in human speaker recognition abilities***
  - **Expert combines qualitative analysis with output from the MIT/LL GMM LFA FRED2 automatic system**
- **The following multistep process is used**
  - **Aided by Super Phonetic Annotation and Analysis Tool [7, 8]**
  - **SPAAT focuses on pronunciation differences, also measured prosody and word choice (voice quality module in future)**

*Astrid Schmidt-Nielsen, Thomas H. Crystal, "Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data," *Digital Signal Processing,* Volume 10, Issues 1-3, January 2000, Pages 249-266

[7] Schwartz, R., Shen, W., Campbell, J., Paget, S., Vonwiller, J., Estival, D., Cieri, C., "Construction of a Phonotactic Dialect Corpus using Semiautomatic Annotation," *Interspeech 2007,* Antwerp, Belgium. August 27, 2007.

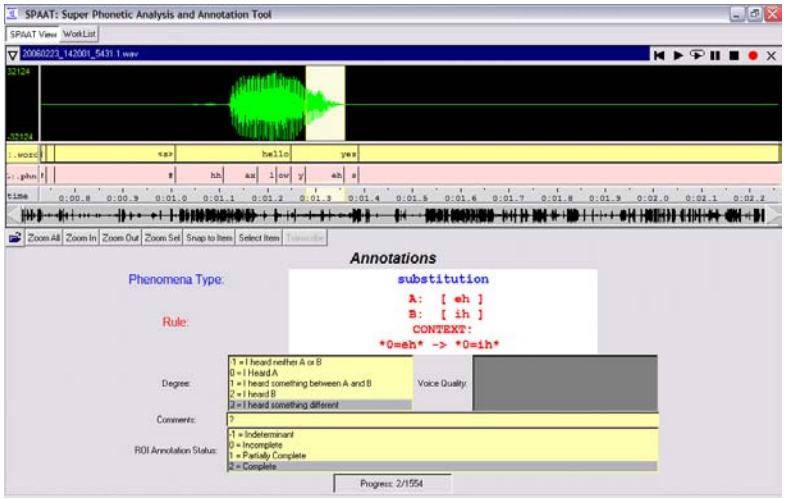
[8] Schwartz, R., Shen, W., Campbell, J., Granville, R., "Measuring Typicality of Speech Features in American English Dialects: Towards Likelihood Ratios in Speaker Recognition Casework," *5th European Academy of Forensics Science,* Glasgow, Scotland, Sept. 8, 2009.

USSS-MITLL HASR Protocol (12 Steps)

1.	Transcribe audio for speaker(s) on channels of interest
2.	Align transcript with audio (force/correct), creating phones and words tiers for annotation
3.	Create “rules” file for phonetic annotation of features. Rules are developed on a per-set basis depending upon dialect and vocabulary and articulatory feature content
4.	Generate phonetic-based regions of interest (ROIs) from applying rules to aligned audio/transcript file sets
5.	Expert annotation of regions of interest at phonetic level within each ROI (see Table 1)
6.	Analysis of ROI annotation output (see Table 2)
7.	Generate prosodic analysis of speaker(s) on channels of interest
8.	Generate acoustic analysis (if applicable)
9.	Vocabulary/word usage analysis (SVM)
10.	Final critical listening for various features
11.	Discern level of similarity and distinctiveness between target speakers, with output as numerical score between 0.30 and 0.90 (see Table 3)
12.	Combine qualitative score with score from MITLL FRED2 automatic system output

Annotation Judgments (Step 5)

- **Expert annotation of regions of interest (ROI) within each ROI**
- **Lincoln/USSS Super Phonetic Analysis and Annotation Tool***
  - **Driven by transformation rules**
  - **Annotate if a particular transformation occurred and to what degree...**
  - **Accelerates process**
- **How common are the features in use found in a proper reference population?**
  - **Typicality effort****



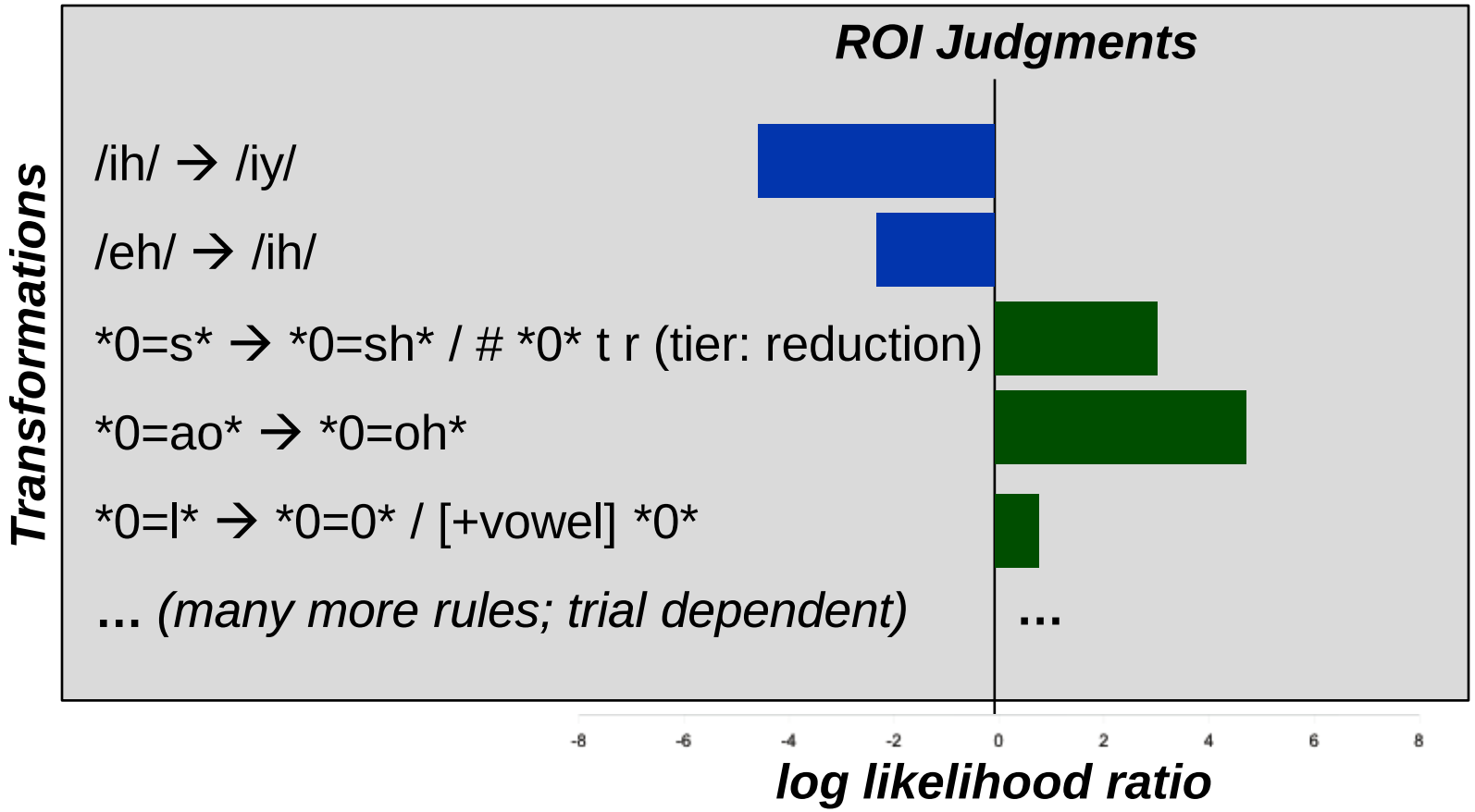
A: Feature transformation did *not* occur	B: Feature transformation did occur
1. Sounds like A	
2. In between A and B	
3. Sounds like B	
4. Something else entirely	
5. Impossible to judge	
6. This ROI is wrong	

*Schwartz, R., Shen, W., Campbell, J., Granville, R., "Measuring Typicality of Speech Features in American English Dialects: Towards Likelihood Ratios in Speaker Recognition Casework," *5th European Academy of Forensics Science,* Glasgow, Scotland, Sept. 8, 2009.

**Cieri, et al., "Bridging the Gap between Linguists and Technology Developers: Large-Scale, Sociolinguistic Annotation for Dialect and Speaker Recognition," *Language Resources and Evaluation Journal*, Springer

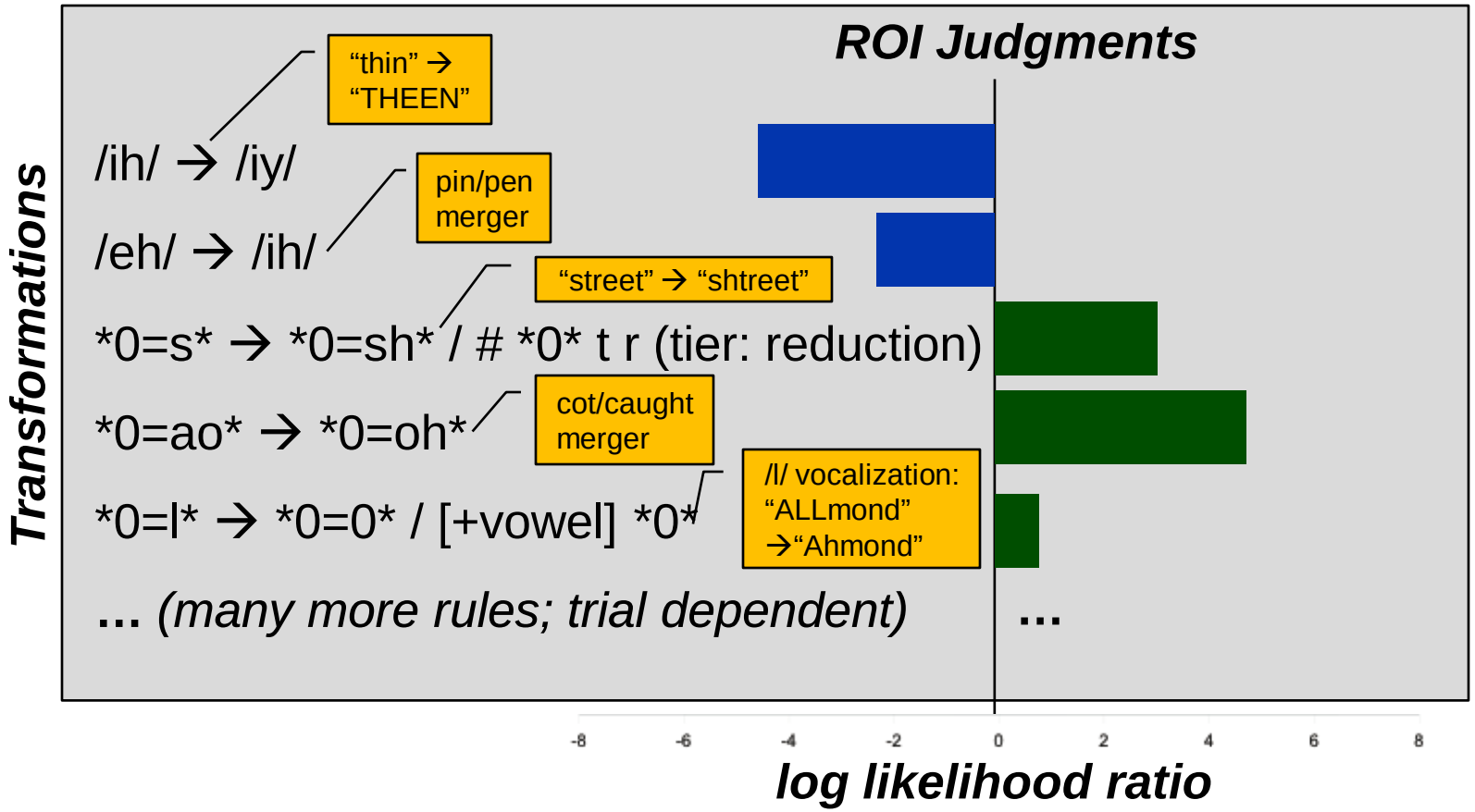
Analysis of ROI Annotation Output (Step 6)

- **How much more likely is a given feature transformation in a sample than in a reference population? E.g., phonetic level**



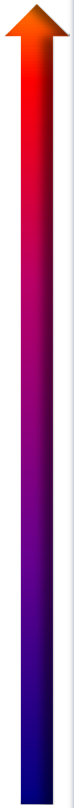
Analysis of ROI Annotation Output (Step 6)

- **How much more likely is a given feature transformation in a sample than in a reference population? E.g., phonetic level**



Conclusion Scale (Step 11)
Adapted from IAFPA for HASR

- Normally this scale is from "Exceptionally distinctive" to "Not distinctive"
- Adapt scale for HASR
  - Associate scores with levels
  - Score is reported and fused with automatic system output
  - Add levels below "Not distinctive" = 0.5
- Normally, a decision is made as to whether the speech samples are consistent or not
  - In HASR, even if the samples were not consistent, all files went through the entire analysis process
  - A decision was made as to the level of similarity based on distinctiveness of features



Score	Level
0.90	Exceptionally distinctive – the possibility of this combination of features being shared by other speakers is considered to be remote
0.80	Highly distinctive
0.70	Distinctive
0.60	Moderately distinctive
0.50	Not distinctive
0.40	Dissimilar – moderately indistinctive
0.30	Dissimilar – highly indistinctive

Human Decision Making: Jell

- How do humans make good decisions?
- Make decisions after allowing time for thoughts to jell
  - Take a break from this type of analysis for about an hour
  - Postpone final decision until next morning

    Common practice among forensic phoneticians
  - The literature supports this practice in complex decision making tasks, e.g.,

On Making the Right Choice: The Deliberation-Without-Attention Effect

Ap Dijksterhuis,* Maarten W. Bos, Loran F. Nordgren, Rick B. van Baaren

Contrary to conventional wisdom, it is not always advantageous to engage in thorough conscious deliberation before choosing. On the basis of recent insights into the characteristics of conscious and unconscious thought, we tested the hypothesis that simple choices (such as between different towels or different sets of oven mitts) indeed produce better results after conscious thought, but that choices in complex matters (such as between different houses or different cars) should be left to unconscious thought. Named the "deliberation-without-attention" hypothesis, it was confirmed in four studies on consumer choice, both in the laboratory as well as among actual shoppers, that purchases of complex products were viewed more favorably when decisions had been made in the absence of attentive deliberation.

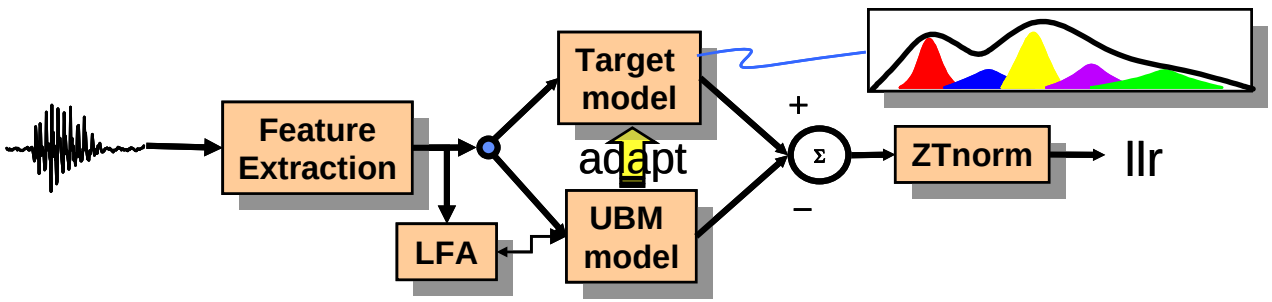
Science, 17 Feb 2006

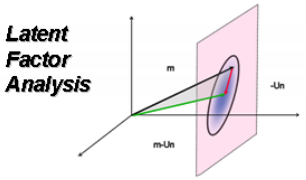
Outline

- Introduction
- Audio Preprocessing
- HASR System
- Qualitative
- Automatic
- Fusion
- Results
- Mechanical Turk
- Summary

Automatic System: FRED
FoRensic Enhanced Detection

- GMM Latent Factor Analysis (LFA) models session variability through a low-dimensional subspace projection in both training and testing [6]
  - Used in SRE'08 Addendum evaluation for interview mic vs. telephone condition
- GMM speech detector followed by energy-based speech detector
- UBM trained using Switchboard II and SRE04 corpora
- Noise reduction system used on interview microphone channels
- Human purification on interview microphone channels
- Telephone network echo cancelation on the telephone channels
- Logistic-regression backend





[6] W. Campbell, et al., "MITLL 2007 Speaker Recognition Evaluation System Description," *NIST SRE Workshop,* Montreal, Canada, 17-18 June 2008.

Fusion

- **Adaptive subjective human weighting is used to linearly combine the human qualitative and automatic system scores**
  - **Weights adapted per trial based on subjective assessments of**
    - **Confidence in the human analysis**
    - **How well matched the automatic system is to the conditions**
    - **Considering automatic scores on dev data**

$$f = wq + (1 - w)s$$

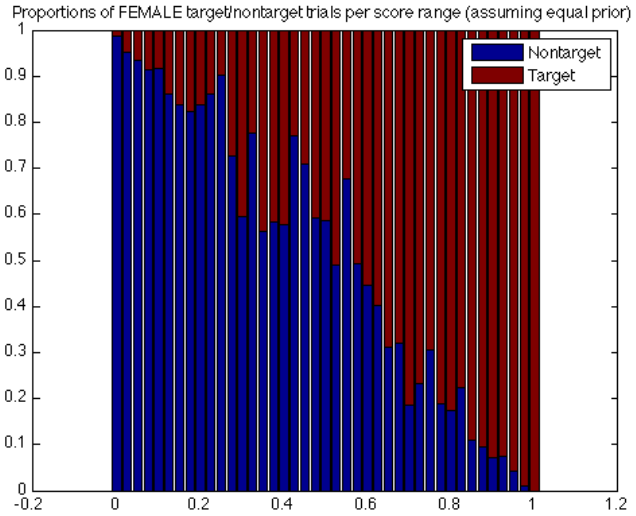
$$where: f = \text{fused score}$$

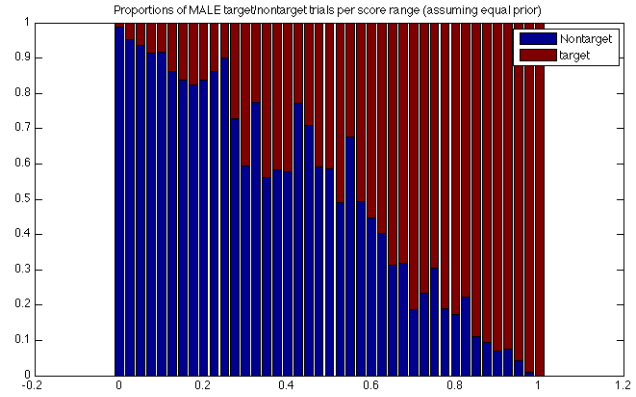
$$q = \text{qualitative score } [0.3, 0.9]$$

$$s = \text{automatic system score } [0, 1]$$

$$\text{and } 0.5 \leq w \leq 1$$

    - **Given these are "difficult trials," we limited automatic system weight**
- **Better ways?**





Results

Trial	Ans	Reva	Fuse	Qual	w %	jpc.3 "FRED"		jpc.4 "FRED2"		Run As	Err/20
1	T	FALSE	0.294	0.5	50	2.66	TRUE	**0.087**	FALSE	male	8
2	F	FALSE	0.29	0.5	50	-2.44	FALSE	**0.08**	FALSE	female	13
3	F	FALSE	0.26	0.5	50	-3.54	FALSE	**0.028**	FALSE	female	8
4	F	FALSE	0.251	0.5	50	-6.1	FALSE	**0.002**	FALSE	male	8
5	T	TRUE	0.75	0.75	100	-2.26	FALSE	**0.094**	FALSE	female	8
6	F	FALSE	0.46	0.56	50	-0.56	FALSE	**0.363**	FALSE	female	11
7	T	FALSE	0.3	0.3	100	3.529	TRUE	**0.972**	TRUE	male	11
8	F	TRUE	0.8	0.8	100	-1.14	FALSE	**0.243**	FALSE	female	7
9	F	FALSE	0.21	0.35	50	-2.66	FALSE	**0.065**	FALSE	female	9
10	T	TRUE	0.85	0.7	50	5.274	TRUE	**0.995**	TRUE	male	2
11	F	TRUE	0.85	0.75	50	3.003	TRUE	**0.953**	TRUE	female	15
12	F	FALSE	0.21	0.3	50	-2.06	FALSE	**0.113**	FALSE	female	7
13	F	FALSE	0.21	0.4	50	-4.07	FALSE	**0.017**	FALSE	female	8
14	T	TRUE	0.89	0.8	50	4.593	TRUE	**0.99**	TRUE	male	4
15	T	TRUE	0.8	0.8	100	-1.5	FALSE	**0.183**	FALSE	male	13

Post Evaluation Results

Trial	Ans	Reva	Fuse	Qual	w %	jpc.3 "FRED"		jpc.4 "FRED2"		Run As	SRE10 Post Eval		Err/20
1	T	FALSE	0.294	0.5	50	2.66	TRUE	**0.087**	FALSE	male	4.87	TRUE	8
2	F	FALSE	0.29	0.5	50	-2.44	FALSE	**0.08**	FALSE	female	-1.8	FALSE	13
3	F	FALSE	0.26	0.5	50	-3.54	FALSE	**0.028**	FALSE	female	-4.69	FALSE	8
4	F	FALSE	0.251	0.5	50	-6.1	FALSE	**0.002**	FALSE	male	-0.11	FALSE	8
5	T	TRUE	0.75	0.75	100	-2.26	FALSE	**0.094**	FALSE	female	0.987	TRUE	8
6	F	FALSE	0.46	0.56	50	-0.56	FALSE	**0.363**	FALSE	female	-1.37	FALSE	11
7	T	FALSE	0.3	0.3	100	3.529	TRUE	**0.972**	TRUE	male	0.955	TRUE	11
8	F	TRUE	0.8	0.8	100	-1.14	FALSE	**0.243**	FALSE	female	-1.78	FALSE	7
9	F	FALSE	0.21	0.35	50	-2.66	FALSE	**0.065**	FALSE	female	-1.79	FALSE	9
10	T	TRUE	0.85	0.7	50	5.274	TRUE	**0.995**	TRUE	male	6.323	TRUE	2
11	F	TRUE	0.85	0.75	50	3.003	TRUE	**0.953**	TRUE	female	8.773	TRUE	15
12	F	FALSE	0.21	0.3	50	-2.06	FALSE	**0.113**	FALSE	female	0.778	FALSE	7
13	F	FALSE	0.21	0.4	50	-4.07	FALSE	**0.017**	FALSE	female	-4.75	FALSE	8
14	T	TRUE	0.89	0.8	50	4.593	TRUE	**0.99**	TRUE	male	9.982	TRUE	4
15	T	TRUE	0.8	0.8	100	-1.5	FALSE	**0.183**	FALSE	male	3.288	TRUE	13

- **Who are these people!?**

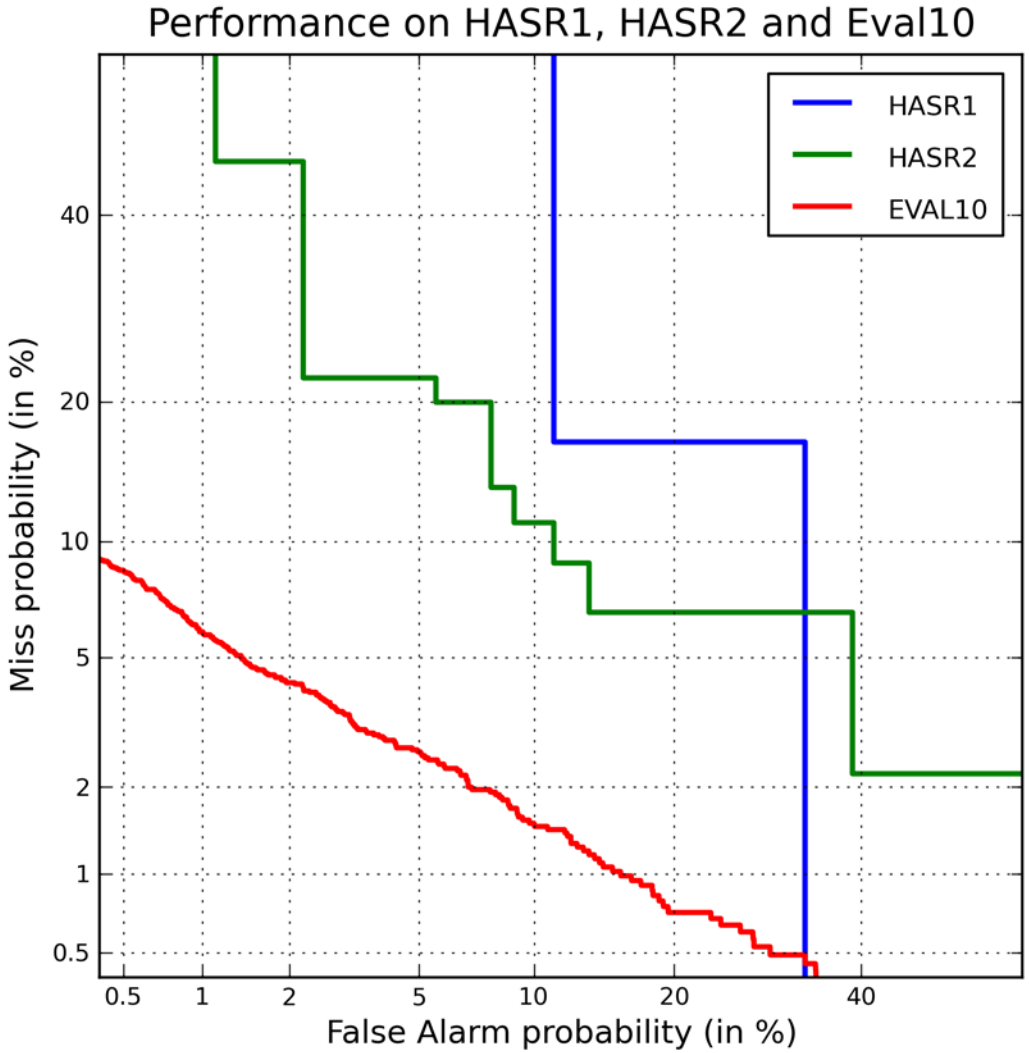

model 11


test 11


model 7


test 7

MITLL SRE10 Automatic Systems Fusion



- **HASR1 ⊂ HASR2 ⊂ EVAL10 Core Test**
- **Error bars on EER get big with small HASR subsets**
- **Note that our systems are not tuned for low EER**

Fusion System	Lower EER	EER	Upper EER
EVAL10	2.8%	**3.2%**	3.7%
HASR2	3.8%	**11%**	21%
HASR1	0.4%	**17%**	62%

Processing Time

- Multiple elements done by humans and machines contribute to the total processing time
- For the automatic FRED system, the processing time is 0.6 times real time, as measured on an Intel Xeon CPU running at 2.00 GHz with 4 MB of cache and 8 GB of memory [6]
- The human processing times, including the manual audio preprocessing and intersite coordination, were somewhat variable depending on the analysis
- Total processing time (human plus machine), after our efficiency improved in the later trials, was 6-8 hours per trial (not including jell time)
  - Faster than 6-8 hour throughput, but HASR is sequential
  - Much different than naïve listeners, but so are results!

Outline

- Introduction
- Audio Preprocessing
- HASR System
- Qualitative
- Automatic
- Fusion
- Results
- Mechanical Turk
- Summary

Human & Machine: Human vs. Machine?

- Finding: machines assist humans in speaker comparison work
  - *Human-Assisted Machines Assisting Humans in Speaker Comparison*
- Inconclusive: human vs. machine performance
  - Sociolinguistic mic interview vs. telephone conversation condition
    - Automatic systems highly tuned to this (couple SREs on int vs tel)
    - Other conditions, styles, stresses, and states in forensics
  - Considerable variation in human performance and approaches
    - Naïve vs linguistically/phonetically-trained listeners
    - Brief aural comparison vs formal qualitative analysis
  - Human?
    - HASR2 impractical for human involvement over 8 weeks – extend?
  - Bias
    - HASR1 selects most difficult SRE10 trials wrt machines and humans
      - Selection introduces biases – any unexpected?
      - Understandable wrt research, but not representative of forensics
  - HASR not intended to be a forensic domain proficiency/competency test
    - Required modification of USSS forensic protocol

Summary

- **HASR was an excellent experience!**
  - **Rapid tool and process development**
    - **Engineers stewed in their own juices!**
- **Highlights the need for caution [1]**
- **Increase participation?**
- **HASR 2010 not quite forensic, but it's helping advance the field**
  - **The National Academy of Sciences forensic sciences report calls for independent and rigorous scientific evaluation [2]**
- **Wish to see future HASR evaluations!**

[1] Campbell, J.P.; Shen, W.; Campbell, W.M.; Schwartz, R.; Bonastre, J.-F.; Matrouf, D "Forensic Speaker Recognition," *IEEE Signal Processing Magazine,* Special Issue on Digital Forensics, vol 26, issue 2, March 2009, p. 95-103, available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=4806187&arnumber=4806209&count=23&index=13

[2] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, "Strengthening Forensic Science in the United States: A Path Forward", National Academies Press, 2009, available: http://www.nap.edu/catalog.php?record_id=12589