REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
25-10-2010

2. REPORT TYPE
Technical Paper

3. DATES COVERED *(From - To)*
OCT 2010 - NOV 2010

4. TITLE AND SUBTITLE
USSS-MITLL 2010 Human Assisted Speaker Recognition

5a. CONTRACT NUMBER
FA8720-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Reva Schwartz, Joseph P. Campbell, Wade Shen, Douglas E. Sturim, William M. Campbell, Fred S. Richardson, Robert B. Dunn, Robert Granville

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NSA
9800 Savage Rd
Ft. Meade, MD 20755

10. SPONSOR/MONITOR'S ACRONYM(S)
NSA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The United States Secret Service (USSS) teamed with MIT Lincoln Laboratory (MlTILL) in the US National Institute of Standards and Technology's 2010 Speaker Recognition Evaluation of Human Assisted Speaker Recognition (HASR). We describe our qualitative and automatic speaker comparison processes and our fusion of these processes, which are adapted from USSS casework. The USSS-MIT/LL 2010 HASR results are presented. We also present post-evaluation results. The results are encouraging within the resolving power of the evaluation, which was limited to enable reasonable levels of human effort. Future ideas and efforts are discussed, including new features and capitalizing on naive listeners.

15. SUBJECT TERMS
Speaker recognition, human assisted, NIST SRE 2010, HASR 2010

16. SECURITY CLASSIFICATION OF: U			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Zach Sweet
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	SAR	4	781-981-5997

MS-50021A

USSS-MITLL 2010 HUMAN ASSISTED SPEAKER RECOGNITION*

Reva Schwartz[†], *Joseph P. Campbell*[‡], *Wade Shen*[‡], *Douglas E. Sturim*[‡], *William M. Campbell*[‡], *Fred S. Richardson*[‡], *Robert B. Dunn*[‡], *Robert Granville*[‡]

[†]United States Secret Service (reva.schwartz@usss.dhs.gov), [‡]MIT Lincoln Laboratory (jpc@ll.mit.edu)

ABSTRACT

The United States Secret Service (USSS) teamed with MIT Lincoln Laboratory (MIT/LL) in the US National Institute of Standards and Technology's 2010 Speaker Recognition Evaluation of Human Assisted Speaker Recognition (HASR). We describe our qualitative and automatic speaker comparison processes and our fusion of these processes, which are adapted from USSS casework. The USSS-MIT/LL 2010 HASR results are presented. We also present post-evaluation results. The results are encouraging within the resolving power of the evaluation, which was limited to enable reasonable levels of human effort. Future ideas and efforts are discussed, including new features and capitalizing on naïve listeners.

Index Terms— Speaker recognition, human assisted, NIST SRE 2010, HASR 2010

1. INTRODUCTION

The United States Secret Service (USSS) teamed with MIT Lincoln Laboratory (MIT/LL) in the NIST Human Assisted Speaker Recognition (HASR) Evaluation. We completed the 15-trial HASR1 Evaluation over the 8-week evaluation period.[1] USSS provided the expert human analyst and MIT/LL provided support, tools, and automatic recognition systems.

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, a performance metric is not defined, and the conditions were not specified. The 15 trials of HASR1 all appear to be in the microphone interview versus telephone conversation condition. The duration of the samples was approximately 3 minutes for the interview[2] and 5 minutes for the telephone conversation (prior to speech activity detection). The samples are provided in two-channel (stereo) format, which allows for analysis of the person of interest (specified via the "channel of interest" by NIST) and the interlocutor in each sample. NIST specifies the samples as the "model segment" and the "test segment", but, consistent with our forensic process, this distinction was ignored and the samples were processed appropriately to produce the required speaker comparison score and decision. NIST granted permission to proceed in this manner and all evaluation rules were strictly followed. MIT/LL also participated in SRE, but no attempt was made to exploit this in HASR (e.g., the file names differed between HASR1 and SRE data and we did not match up the audio during the evaluation or attempt to use additional data available in SRE, such as automatically generated transcripts for SRE).

2. AUDIO PREPROCESSING

First, the samples are acquired for a given trial and prepared for human analysis and for automatic processing. Two samples for a given trial are acquired, per an automated e-mail from NIST, via ftp. The samples are in NIST SPHERE (.sph) format using two-channel G.711 μ-law (8 kHz, 8-bit sampling). The following audio processing chains were used, depending on the recording condition and use.

- Interview recordings for both human analysis and automatic processing:

Source .sph → Peak normalize (90% FS), DC Bias removal → Enhancement[3] → Purification (in stereo) → Extract channel of interest [always channel a (left channel)]

- Telephone recordings for human analysis:

* 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.

[1] The 8-week evaluation period was too short to implement the standard forensic process on the 150-trial HASR2 Evaluation.

[2] The duration of the interview sample ranged from approximately 1¼ to 2½ minutes after purification (and is further reduced by speech activity detection; down to 1 minute in trial 13).

[3] Both channels are enhanced independently. A two-stage enhancement process 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.

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

In the purification step, we (a human) manually remove segments of the interlocutor's speech and regions of overlapped speech. Performing this editing on the two-channel enhanced audio was found to speed the process, likely improve purification accuracy, and reduce fatigue (NIST had apparently added noise to the interviewer's channel and, at times, there was substantial HVAC noise in the interview room).

The FRED system includes telephone network echo cancelation processing, which was deemed unnecessary in the HASR1 trials for human processing because the echo was negligible (and providing those samples would have introduced delay in our grand process).[4] Likewise, the automatic system did not make use of the human-generated transcripts to streamline our processing.

Now these audio samples are ready for our HASR system process.

3. HASR SYSTEM

The Human Assisted Speaker Recognition (HASR) system is an expert-based process adopted from general forensic-phonetics methodology, combined with output from the MIT/LL GMM LFA FRED2 automatic system. The following multistep process is used with the aid of the Super Phonetic Annotation and Analysis Tool [7, 8]:

1. Transcribe audio for speaker(s) on channels of interest.
2. Align transcript with audio (force/correct), creating phone and word 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 Figure 1).
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 0.3 and 0.9 (see Table 2).
12. Combine qualitative score with score from MITLL FRED2 automatic system output.

[4] Also, because of negligible echo, manual purification of the audio was unnecessary on the telephone recordings.

Table 1. Annotation Judgments Scale.

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

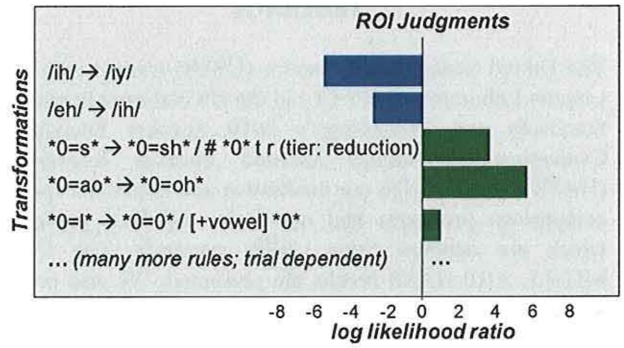

Figure 1. Analysis of ROI Annotation Output. How much more likely is a given feature transformation in a sample than in a reference population? At the phonetic level shown here, e.g., /eh/ → /ih/ in the pen/pin merger.

Table 2. Conclusion Scale (adapted from IAFPA).

Score	Level
0.9	Exceptionally distinctive – the possibility of this combination of features being shared by other speakers is considered to be remote
0.8	Highly distinctive
0.7	Distinctive
0.6	Moderately distinctive
0.5	Not distinctive
0.4	Dissimilar – moderately indistinctive
0.3	Dissimilar – highly indistinctive

4. FRED GMM LFA SYSTEM

The FoRensic Enhanced Detection (FRED) system uses the MITLL GMM-UBM speaker detection system [1] used in the SRE'08 Addendum evaluation for the interview microphone vs. telephone condition [6] with human preprocessing. The main differences this year are:

- A GMM-based speech detector was used as initial speech detector followed by a second stage energy-based speech detector.

- The UBM was trained using Switchboard II and SRE'04 corpora.
- A noise reduction system was used on the microphone channels.
- Audio preprocessing, including human purification on the microphone channels.
- Telephone network echo cancelation on the telephone channels.
- Latent Factor Analysis (LFA) GMM.
- Logistic-regression backend.

The features used were a 19-dimensional mel-cepstral vector extracted from the speech signal every 10 ms using a 20 ms window. The mel-cepstral vector is computed using a simulated triangular filterbank on the DFT spectrum. The log-energy filterbank values are passed through a RASTA filter to remove slowly varying linear channel effects. Bandlimiting is then performed by only retaining the filterbank outputs from the frequency range 300 Hz to 3138 Hz and cepstral coefficients are computed via a DCT transform. Delta cepstral are then computed over a +/-2 frame span and appended to the cepstral vector, producing a 38-dimensional feature vector. Finally the cep+dcep features are mean and variance normalized over the speech segments per file.

To combat additive noise in the microphone channel, the two MIT/LL noise-reduction techniques employed (steady tone removal and wideband noise reduction) were applied in series as a preprocessor step to MFCC feature extraction. The steady tone suppression method used a very long analysis window, 8 seconds, to exploit the coherent integration of the Fourier transform. The wideband noise reduction algorithm used an adaptive Wiener-filter approach directed toward preserving the dynamic components of a speech signal, while effectively reducing noise. Greater detail can be found in [2].

The GMM Latent Factor Analysis (LFA) was based directly on the work presented in [3]. The approach models session variability through a low-dimensional subspace projection in both training and testing. The session variability is modeled as a low-dimensional additive bias to the model means:

$$m_i(s) = m(s) + Ux(s) \tag{1}$$

where $m_i(s)$ and $m(s)$ are supervectors of stacked GMM means [3, 4]. The $m_i(s)$ is the supervector from the i-th session of talker s, $m(s)$ is the session-independent term of talker s, and $x(s)$ is the subspace.

Training of the low-rank transformation matrix U was generated directly, as described in [5] and not iteratively. Z-norm followed by T-normalization was also performed on the scores.

The LFA system was applied gender dependently. Factor analysis was performed using session loading matrices generated with class-variation constrained to be speaker only. However in the presence of a microphone channel, the loading matrix used was one generated with class variation constrained to speaker and session. Additionally, when microphone data was present, the noise-reduction frontend was applied.

For the microphone test conditions, the following configuration was used:

- GMM background model – Trained from Switchboard II and SRE'04 corpora.
- Stacked FA session loading matrix – Trained from 1) NIST SRE Eval'05 microphone data with the class variation to be per speaker-session,[5] 2) Six interview microphone talker dev set provided by NIST before the 2008 evaluation, and 3) NIST SRE Eval'04 using data from speakers with more than 16 enrollment sessions.
- Z-norm test utterances – NIST SRE Eval'04 and switchboard II when testing was on the telephone channel and NIST SRE Eval'05 microphone data when testing on microphone channel.
- T-norm speakers – NIST SRE Eval'04 data set when enrollment was telephone channel and cohorts where chosen from NIST SRE Eval'05 microphone corpus when the enrollment condition was on the microphone channel.
- LFA co-rank was 64.

4.1. Backend Calibration

A logistic regression was trained on the NIST 2008 SRE data for the condition that used 4-wire (stereo) conversational telephone data for enrollment and interview microphone data for verification. Since the target prior probability was not known for HASR, we used an equal prior for target and non-target trials. We used the optimal Bayesian decision threshold for the equal prior and equal cost case of zero for interpreting the output score of the system.

The FRED and FRED2 systems require the interview recording and telephone channel to be specified and the sex of the talker(s) to be specified. These specifications were not given in HASR and are based on human judgment (to be later verified with NIST's keys). The FRED and FRED2 systems differ only in score transformation: FRED uses log-likelihood ratios (λ), whereas FRED2 uses a posterior probability estimate: e^λ/(e^λ+1), assuming flat priors and equal costs, except for Trial 1, which had reversed inputs in the FRED system (ironically, this mistake eliminated a trial error for FRED).

[5] We also explored using a loading matrix to learn variation over microphones and found this to work well on dev data. We elected to not use it for our evaluation system due to concern that it would not generalize well to new microphones.

5. PROCESSING TIME

Automatic processing time was negligible [6]. The total processing time (human plus machine), after our efficiency improved in the later trials, was approx. 8 hours per trial.

6. HUMAN-MACHINE FUSION

An adaptive subjective human weighting is used to combine the human qualitative score with automatic system score. Weights are adapted per trial based on subjective assessments of the following: confidence in the human analysis, how well matched the automatic system is to the conditions, and considering automatic score distributions on development data. This is all highly subjective and we rely on an expert human to make these assessments to adapt the fusion using linear combination:

$$f = wq + (1 - w)s \tag{2}$$

where f is the fused score, q is the qualitative score [0.3, 0.9], s is the automatic system score [0, 1], and w is the weight. We constrained the weight $0.5 \leq w \leq 1$ to limit the automatic system's influence because "difficult trials" were selected for the HASR evaluation [11]. This fusion produces the final overall score f submitted to NIST. NIST also requires a hard decision. The prior probabilities, costs of errors, and performance metric were not specified. In the absence of this information, we chose a balanced operating point. The score f was thresholded at 0.5 to form the hard decision submitted to NIST.

7. RESULTS

NIST reported the results of the HASR1 sites using hard decisions only. There were 15 trials (6 targets and 9 nontargets). Table 3 shows the results for our qualitative method; two automatic systems, FRED and FRED2; and the fusion of the qualitative method with FRED2.

Table 3. NIST SRE 2010 HASR results for USSS and MIT/LL.

System	Misses (out of 6 targets)	False alarms (out of 9 nontargets)
Qualitative	2	2
FRED	2	1
FRED2	3	1
Fusion	2	2
Post Evaluation	0	1

Some of these errors occur on different trials. Following our official submission, we continued improving our system, as shown in the last row of Table 3.

8. FUTURE

This exercise has given us plenty of ideas for tools and methods. Newer automatic systems will be tried in the future (e.g., MIT/LL's IPDF systems). A next-generation Super Phonetic Annotation and Analysis Tool [7, 8] is being developed. It includes an additional processing step and feature for voice quality. We have begun investigations of capitalizing on large-scale human listening (e.g., via Mechanical Turk) [10].

9. CONCLUSION

HASR was exhausting and a valuable learning experience. Having 15 sets of samples to compare in a short period of time really helped crystallize tools and ideas that work, as well as those that do not. HASR is inconsistent with forensic speaker comparison (e.g., w.r.t. scoring and decision making and bias due to selection of "difficult trials," as noted in the Evaluation Plan [11]). Conclusions about forensic performance and human vs. machine performance cannot be fairly drawn here. This exercise is, however, helping to advance the field, as demanded by the National Academy of Sciences [9]. More sites should participate in the future.

10. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] D. E. Sturim, W. M. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proceedings of IEEE ICASSP*, pp. IV–49–IV–52, 2007.

[3] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in Proc *EuroSpeech*, 2006.

[4] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions On Speech And Audio Processing*, vol. 13, no. 3, pp. 345, May 2005.

[5] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

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

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

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

[9] Committee on Identifying the Needs of the Forensic Science Community, National Research Council of The National Academies, *Strengthening Forensic Science in the United States: A Path Forward*, Washington, DC: The National Academies Press, 2009.

[10] W. Shen, J. P. Campbell, D. Straub, R. Schwartz, "Assessing the Speaker Recognition Performance of Naïve Listeners using Mechanical Turk," submitted to special session on Investigations in Combining Human Expertise with Automatic Processing for Successful Speaker Recognition, *Proceedings of IEEE ICASSP*, Prague, May 2011.

[11] C. Greenberg, et al. "Human Assisted Speaker Recognition in NIST SRE10," submitted to special session on Investigations in Combining Human Expertise with Automatic Processing for Successful Speaker Recognition, *Proceedings of IEEE ICASSP*, Prague, May 2011.