USSS-MITLL 2010 Human-Assisted Speaker Recognition Evaluation System*

Reva Schwartz[†], Joseph Campbell[‡], Wade Shen[‡], Douglas Sturim[‡], William Campbell[‡], Fred Richardson[‡], Robert Dunn[‡], Robert Granville[‡]

[†] United States Secret Service
Reva.Schwartz@usss.dhs.gov

[‡] MIT Lincoln Laboratory
jpc@ll.mit.edu

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 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).

* 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. DISTRIBUTION STATEMENT F. Further dissemination only as directed by NSA/R66E (4/2/2010) or higher DoD authority.

[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).

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 [a]

- Telephone recordings for human analysis:

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

In the purification step, we (a human) manually removing segments of the interlocutor's speech and regions 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 target speakers
2. Align transcript with audio, creating .phns and .words tier 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

[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.

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

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 target audio samples
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 .30 and .90 (see Table 3)
12. Combine qualitative score with score from FRED2 automatic system output from MIT LL core system. The automatic system was fused at a weight from 0% to 100%

Table 1: annotation judgments	
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	

Table 2: Sample analysis output
Per tier: substitution : k = 0.07, q = 0.05, log(LR) = -0.3882 (nK = 118, nQ = 307) reduction : k = 0.26, q = 0.20, log(LR) = -0.2975 (nK = 91, nQ = 97)
Per rule: A: [iy] : B: [ih] : CONTEXT: ___ [#] ::: *0=iy* -> *0=ih* / *0* # (tier: substitution) : k = 0.00, q = 0.08, log(LR) = 4.3820 (nK = 16, nQ = 25)
A: [l] : B: [0] : CONTEXT: { +vowel } ___ [#] : *0=l* -> *0=0* / [+vowel] *0* # (tier: reduction) : k = 0.19, q = 0.18, log(LR) = -0.0970 (nK = 18, nQ = 17)

Table 3: Conclusion scale (adapted from IAFPA for specific use on HASR).

- Normally this scale is from “Exceptionally distinctive to “Not distinctive”. But to carry out analysis for HASR the scale was adapted 1) to add numerical values to each level of the scale, so the score could be reported and/or fused with automatic system output, and 2) to add levels below “Not distinctive/.50”
- 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 and a decision was made as to the level of similarity based on distinctiveness of features. This is adapted from the general process used in FP

casework.	
Numeric score	
.90	Exceptionally distinctive – the possibility of this combination of features being shared by other speakers is considered to be remote
.80	Highly distinctive
.70	Distinctive
.60	Moderately distinctive
.50	Not distinctive
.40	Dissimilar – moderately indistinctive
.30	Dissimilar – highly indistinctive

4 GMM LFA System "FRED"

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 mic 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 SRE04 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 is 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 – 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 two noise reduction techniques were employed, 1) steady tone removal and 2) wideband noise reduction, were applied in series as preprocessor step to MFCC feature processing. 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:

$$m_i(s) = m + U(s) \quad (1)$$

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

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 be 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 SRE04 corpora
- Stacked FA session loading matrix – Trained from 1) NIST SRE Eval05 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 Eval04 using data from speakers with more then 16 enrollment sessions.
- Z-norm test utterances – NIST SRE Eval04 and switchboard II when testing was on the telephone channel and NIST SRE Eval05 microphone data when testing on microphone channel
- T-norm speakers – NIST SRE Eval04 data set when enrollment was telephone channel and cohorts where chosen from NIST SRE Eval05 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

[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.

4w 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 Processing Time

There are multiple elements done by humans and machines that contribute to the 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. The total processing time (human plus machine), after our efficiency improved in the later trials, was approximately 6 hours per trial (this could be further improved).

6 Future

Newer automatic systems will be tried in the future (e.g., MIT/LL's SRE10 JFA system). This exercise has given us plenty of ideas for tools and methods. A next generation Super Phonetic Annotation and Analysis Tool [7, 8] is in the works. We have begun investigations of capitalizing on large-scale human listening (e.g., via Mechanical Turk) for improved performance.

7 Summary

HASR was exhausting and a great learning experience. Having 15 sets of samples to compare in a short period of time has really helped crystallize the tools/ideas that work, as well as those that do not work. There is nothing like letting the engineers stew in their own juices to bring about some rapid tool improvements and rigor! Although HASR is not quite consistent with forensic speaker comparison (e.g., w.r.t. scoring and decision making and potential bias due to selection of "difficult trials" noted in the Evaluation Plan), this exercise is helping to advance the field, as demanded by the National Academy of Sciences [9]. Hopefully more forensic speaker comparison organizations will participate in the future.

8 Acknowledgements

Thanks to Douglas Reynolds and Tom Quatieri for their speaker recognition and speech enhancement contributions. We appreciate discussions with and the support of the Lincoln Human Language Technology group. Thanks to NIST, LDC, and their sponsors for this evaluation.

9 References

[1] Douglas 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 ICASSP*, 2007, pp. IV–49–IV–52.

[3] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *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 June 2008.

[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.

[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.