Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

The 2011 BEST Speaker Recognition Interim Assessment

Craig Greenberg, Alvin Martin, Mark Przybocki

National Institute of Standards and Technology
Gaithersburg, Maryland, USA
craig.greenberg@nist.gov, alvin.martin@nist.gov, mark.przybocki@nist.gov

Abstract

In the fall of 2011, NIST conducted an interim assessment of speaker recognition technology developed as part of the Intelligence Advanced Research Project Activity (IARPA) Biometric Exploitation Science and Technology (BEST) program. The goal of the first phase of the BEST program was to advance the state of the art in biometric technology and to provide direction for future phases of the program. Robustness to intrinsic, extrinsic, and parametric variations was of particular interest to the BEST program, and therefore measuring performance across such variations was a focus of the assessment. This included the use of data with simulated room acoustics and additive noise. A new, simple, and intuitive performance measure was utilized. Improvement in performance compared to a baseline system was observed in all conditions examined to date.

1. Introduction

In December, 2009, IARPA launched the BEST program with the goal of advancing the state of the art in biometrics, including speaker recognition as one of the targeted technologies [1]. Phase 1 of the BEST program was specifically interested in achieving high performance in face-to-face interview scenarios and telephone scenarios and in obtaining robustness to various speaker recognition challenges. The primary challenges addressed were:

- **Intrinsic variations**: internal speaker variability issues, specifically speech style, in particular interview speech and conversational telephone speech, and vocal effort variability, namely high or low vocal effort induced by the recording conditions in contrast to "normal" vocal effort.

- **Extrinsic variations**: sources of signal variability due to sources other than the speaker him/herself. Extrinsic variations include differences in room acoustics, noise level, sensor differences and speech coding. A key example is the cross-domain situation, as when the training sample is extracted from a telephone transmission channel, and the test sample is collected over a microphone channel in an interview room (or vice versa).

- **Parametric variations**: sources of variability due to evaluation conditions not included in the above. Variation by language is included in this category. Language independence, the need for algorithms to perform well regardless of the language being spoken was a key BEST Program goal. Also desirable, though less key, was that algorithms perform well when training and test speech samples are produced in different languages (as in the case of bilingual speakers). "Aging" of the vocal tract (the change in speech over extended time periods) was included here, though it is arguably an intrinsic variation. Finally, variations in performance based on the number of training sessions or their duration is also included in this category.

NIST conducted an assessment of the speaker recognition technology developed for the BEST program at the end of the program's first phase in order to provide direction for future phases of the program. The data utilized in the BEST interim assessment was produced by the Linguistic Data Consortium (LDC) and included some data from previously collected corpora, including Switchboard [2], Mixer [3-5], and Greybeard [6], as well as a new corpus, Mixer-7. In addition, the interview data collected as part of Mixer-7 was altered to simulate different room acoustics and noisy environments.

The intent of this paper is to describe the BEST interim assessment and to share some of the initial results. The objectives of the BEST program are both aggressive and diverse, and so the assessment was necessarily large and complex. It included more than 1,000 speakers, 82,000 audio segments, and 41,000,000 trials, making it the largest NIST conducted evaluation of speaker recognition technology to date. There is a wealth of possible analysis, and we have only begun to scratch the surface.

2. Evaluation Task and Conditions

The BEST interim assessment was limited to the broadly defined task of speaker detection. The task was to determine whether a specified speaker is speaking during a given segment of speech. This task was performed in the context of interview style speech or of conversational telephone style speech. The speech included was recorded over multiple types of telephone or room-microphone channels.

It was specified to systems whether each segment was recorded over a telephone type channel or a room microphone type channel; however, information was not provided about the particular type of telephone or room microphone channel over which a segment was recorded. The sex of the target speaker was also specified to systems, and no cross-sex trials were included in the evaluation.

2.1. Training and Test Conditions

The majority of trials included in the BEST interim assessment consisted of trials defined by a single training and a single test segment. In addition, two multi-session training conditions were included. These involved multiple (generally eight) interview segments on which to train and testing on either a telephone or interview segment. It should be noted that these training interviews consisted of recordings over four channels of two interviews on the same day. In both cases the target speaker was the speaker in the designated channel of interest for each of the training sessions.

Table 1: Descriptions of each of the nine conditions of particular interest identified in the evaluation.

Training Data	Test Data
Phone calls recorded over a telephone channel	Phone calls recorded over a telephone channel
Phone calls recorded over a microphone channel	Phone calls recorded over a telephone channel
Phone calls recorded over a microphone channel	Phone calls recorded over a microphone channel
Interviews recorded over a microphone channel	Interviews recorded over a microphone channel
Interviews recorded over a microphone channel	Phone calls recorded over a microphone channel
Phone calls recorded over a telephone channel including languages other than English	Phone calls recorded over a telephone channel including languages other than English
Phone calls recorded over a microphone channel including languages other than English	Phone calls recorded over a microphone channel including languages other than English
Multiple microphone recorded interviews	Interviews recorded over a microphone channel
Multiple microphone recorded interviews	Phone calls recorded over a microphone channel

There were nine conditions identified to be of particular interest to the evaluation. These are described in Table 1. It should be noted that, with the exception of the two conditions involving language, the systems were provided with sufficient information to determine to which of these conditions a trial belonged.

3. Data

Past NIST evaluations have utilized corpora collected by the LDC. These have included various Switchboard and Mixer corpora, and have featured collections of phone calls, usually on assigned topics, between two people who did not know one another. In the case of the recent Mixer collections they have also included interview sessions in a room at a specified location (often the LDC) involving a subject and an interviewer. These sessions, mainly conversational in nature, have been recorded over a range of in-room microphones. Also in recent Mixer collections, some telephone calls were collected at the LDC and recorded over room-microphones as well as telephone channels, and other telephone calls were collected remotely, with the speaker calling in from outside the LDC.

The BEST interim assessment included data from these previously used corpora. Thus some target speakers appeared in past evaluations. In particular, the assessment utilized five corpora, one of which was newly collected to support BEST.

All excerpts were two-channel, with the channel of interest designated. For telephone conversational data, the other channel was the telephone channel of the other speaker in the conversation. For interview data, the other channel was the interviewer’s close-talking microphone channel. The telephone conversational excerpts were of approximately five minutes total duration (with the speaker in the channel of interest speaking on average around half the time) as in recent NIST Speaker Recognition Evaluations (SRE’s)[7]. The interview excerpts were approximately 11 minutes in total duration.

3.1. Greybeard

The recently collected Greybeard Corpus [6] contains telephone conversations of about 180 speakers who participated in the various Switchboard and Mixer Corpora previously collected by the LDC. For each such speaker the corpus contains both conversations included in the prior corpus and newly collected conversations. Thus it is designed to test the effect of aging on speaker recognition performance. The time interval between the old and new conversational data of each speaker ranges from a couple of years to as many as twelve.

The Greybeard Corpus data was previously used in the NIST SRE10 evaluation. However, it should be noted that Greybeard data was not included in the SRE10 keys that were released.

3.2. Mixer 1/2

Mixer-1 and Mixer-2 [3] (together referred to as Mixer-1/2 throughout) is a multi-language collection of telephone conversations used in SRE05 [8] and SRE06 [9]. The speakers include over 100 bilingual speakers of English, plus Spanish, Arabic, Mandarin, or Russian, as well as English-only speakers. Each speaker generally completed four or more calls in each language spoken.

3.3. Mixer-5

Mixer-5 [4] is a collection of interview sessions and telephone conversations involving about 300 different speakers. About half of these speakers were used in the 2008 NIST evaluation.

The BEST assessment included interview and telephone data of approximately 150 speakers that were not used in the 2008 evaluation.

3.4. Mixer-6

Mixer-6 [5] is also a collection of interview sessions along with telephone conversations, though in Mixer-6 some phone calls were made at the LDC (referred to as “internal calls”) and others made externally. There are generally three internal calls per subject, one low vocal effort, one high vocal effort, and one normal vocal effort call made using a cell phone (see section 4.2.2). This was the primary corpus used in SRE10 [10].

A small number of the speakers of this corpus were not utilized in SRE10, and only these speakers were included from the Mixer-6 corpus in the BEST assessment. Speech data of these speakers included some microphone channels used in SRE10 as well as other channels not used in the earlier evaluation.

3.5. Mixer-7

Mixer-7 is a newly collected corpus utilized for BEST. Like Mixer-5 and Mixer-6 it features both interview sessions and telephone calls involving a large set of newly recruited speakers. Some of these speakers are bilingual and had interview sessions in Spanish as well as in English.

Generally, each subject participated in eight interview sessions at the LDC on four separate dates. The interviewer's role was to initiate conversation on various topics and to encourage open expression by the subject. Interview sessions were 15-20 minutes in duration and took place in multiple rooms. There were variations in the acoustics between rooms; in particular, one room was altered to be more reverberant than the other. Each interview was recorded over multiple room-microphone channels, including a close-talking microphone worn by the interviewer.

Speakers were also asked to take part in multiple phone conversations each day they came in for interviews. As with Mixer-6, these include calls designed to elicit high or low vocal effort. The subject sides of these calls were recorded over multiple room-microphone channels and over a telephone channel. Subjects also were encouraged to make eight or more external calls from home or another location using the LDC's robotic system for pairing callers who do not know one another to speak on an assigned topic.

4. Evaluation Factors

We group the factors examined in the assessment into three types: those that are intrinsic to the speakers involved, those extrinsic to them, and those that do not fit well into the intrinsic/extrinsic dichotomy, but instead rely on some orthogonal parameter. We begin by discussing parametric factors, followed by intrinsic factors, and then extrinsic factors.

4.1. Parametric

The inclusion of Mixer-1/2 and greybeard, and the design of Mixer-7 supported examination of the effects of several parametric factors, including language (and dialect), aging, and number of training sessions.

4.1.1. Language/Dialect

One of the primary objectives of the assessment was to evaluate system performance robustness to language independence objective. The central comparison made was between performance for English-language trials and those trials involving a single other language. The degree to which performance for non-English trials fell short of that for English trials can be viewed as an important indicator of the extent to which systems may have been developed to be dependent upon specific features of English or to utilize English language specific knowledge.

It was also of interest to compare performance on native English speech with that on non-native English speech. This again could indicate system dependence on features of fluent English, but it is also possible that the English spoken by many non-native speakers may prove to have useful speaker-specific features.

Performance was also observed for trials involving different languages (or dialects) in training and test, and for non-target trials involving native and non-native English in training and test. It could be expected that performance would be enhanced for such non-target trials. Of some interest also was how recognizable bilingual speakers proved to be when their training and test segments involved different languages.

In comparing cross language conditions, it was important to separately examine the effects of language differences on target trials and on non-target trials, as these effects could operate in opposite directions.

4.1.2. Aging

The primary comparison was between performance involving target trials with training and test data recorded close in time (training and test on previously collected corpora, training and test on the newly collected corpus) and target trials separated in time (training on the previously collected corpora, test on the newly collected corpus or, possibly, vice versa).

Unfortunately, it should be noted that Greybeard is a small corpus with a limited number of speakers. The Greybeard Corpus data that was used was also used in the NIST SRE10 evaluation (without key release for this data), and little aging effect was apparent in the results. This result was reexamined in the BEST interim assessment.

4.1.3. Multiple Training Sessions

As noted in section 2.1, two test conditions, each involving multiple, generally eight, training sessions for each target speaker was included. The eight conversation training condition has been included in recent NIST SRE's and has in the past given considerable performance improvements over one conversation training.

Multiple sessions of training has enhanced performance considerably more than simply longer duration training, perhaps adding robustness by reflecting a subject's speech variation over time. As a result of having eight interviews, in general, per speaker in Mixer-7, the BEST assessment was the first evaluation that supported a multiple interview training condition. With interview segments of approximately 11 minutes, the total training duration was almost an hour and a half. The effect of this amount of training, compared with single interview training, was examined.

4.2. Intrinsic Factors

While language and aging, discussed above, could be viewed as intrinsic factors, here we consider two factors that are clearly intrinsic to the speaker: speech style and vocal effort.

4.2.1. Speech Style

The two speech styles that were contrasted were interview and conversational telephone speech. Both were largely conversational, but the former involved face to face dialogue with a person in the same room with whom the subject has to some extent become familiar, while the latter involved an unseen person the subject did not know. The phone conversations were also frequently on an assigned topic for which the subject may have limited interest or knowledge.

Style comparisons included train and test on interview, train and test on phone conversation, and a mixed train/test condition. These comparisons were made over trials involving a common set of room-microphone channels. These included same or different microphones in train and test and the same or different collection rooms.

4.2.2. Vocal Effort

The Mixer-7 Corpus, like the Mixer-6 Corpus, included phone calls made by each speaker while at the LDC and wearing headsets, thus providing feedback intended to induce low vocal effort (LVE) or to induce high vocal effort (HVE). Additional calls made by each speaker at the LDC involved normal vocal effort (NVE). This supported examining the effect of vocal effort in training or test on performance with other conditions largely fixed. For more details on how the vocal effort data were collected see [5, 11].

The primary conditions compared were those trials where the training was on NVE and the test on LVE, NVE, or HVE. Other conditions involving vocal effort were included for examination as well.

4.3. Extrinsic Factors

The Mixer-6 and Mixer-7 corpora were designed to support, among others experiments, investigation of extrinsic factors relating to channel types, telephone transmission types, and room setup and acoustics. The effects of these factors were examined. In addition, noise and room reverberation factors were investigated by artificially adding noise or reverberation to some of the collected data, as discussed further below.

4.3.1. Channel Type

Channel type refers to the contrast between data received over telephone channels as opposed to data received over room-microphone channels.

The internal phone conversations of Mixer-7 were collected over both types of channels. The comparison involved is a binary one like that of speech style (section 4.2.1), and was handled similarly, examining the two matched train/test conditions and the unmatched condition.

4.3.2. Microphone Type

Mixer-7 interviews and calls, like those of Mixer-5 and Mixer-6, were collected over multiple room-microphone channels. A limited number of the available channels were included, however, in evaluating the effects of microphone type. This was due to a desire to include as many training and test segments as possible for each microphone train/test combination being examined, and thus to obtain a sufficiently significant total number of trials, particularly non-target trials, while limiting the numbers of segments and trials included in the evaluation. The channels were selected to reflect a range of microphone quality and microphone distance from the subject. All matched train/test microphone combinations were included in the trials, along with selected unmatched microphone pairs.

A similar strategy was used with respect to all of the unexposed Mixer-5 and Mixer-6 interview speakers included in the evaluation.

The Mixer-6 microphones used included both ones previously exposed and ones not exposed previously.

4.3.3. Telephone Type

Telephone type may refer to transmission type (e.g., landline, cellular) or the type of telephone instrument (e.g., hand-held, head-mounted).

These factors were examined separately in a way that avoids conflation with the effects of vocal effort.

4.3.4. Interviewer/Subject Distance

The effect of distance between interviewer and subject was examined based on Mixer-7 data. The intent of changing the distance between the interviewer and subject was to collect vocal effort variation that occurred more naturally than in the case where the variation was induced with headphones (see section 4.2.2). Both matched and mismatched distances in training and test were included.

4.3.5. Reverberation

Mixer-7 interviews were collected in different rooms with different reverberation properties. The effects of using different room with different reverberation characteristics in training and test were studied.

A procedure was proposed by MITRE and was implemented by MIT Lincoln Laboratory to transform collected signals to have the reverberation qualities of arbitrarily specified rooms over a range of possible dimensions and surface conditions. A summary description of these data is given in Table 2.

These experiments used Mixer-7 interview sessions recorded over a high quality microphone close to the subject. The train/test combinations examined were training without

Table 2: A description of the simulated reverb conditions, including the measured RT30, measured RT60, estimated RT30. The abbreviations should be understood as follows: hp – heavy-plate glass, gy – gypsum wallboard, pl – plywood, uc – unpainted concrete, pe – percent 50 (an imaginary material with low reflection properties), 1(2,3,4,5,6) – one (two, three, four, five, six, respectively) surfaces are made of the preceding material. E.g., pl5uc1 is a simulated room with five surfaces made of plywood and one made of unpainted concrete

	Meas. RT30	Meas. RT60	Est. RT30	Reverb Perception
hp2gy4	0.792	1.313	0.773	Unintelligible at times
gy4pl2	0.403	0.989	0.498	Very noticeable
gy2pl4	0.357	0.682	0.337	Very noticeable
pl5uc1	0.211	0.614	0.183	Noticeable
pl3uc3	0.166	0.374	0.110	Noticeable
uc6	0.081	0.240	0.058	Slightly noticeable
uc4pe2	0.062	0.161	0.032	Slightly noticeable

additive reverb and test on each of the simulated room conditions, and train/test on matched simulated room conditions. Systems were not provided any development data with simulated reverberation.

4.3.6. Additive Noise

As with reverb, additive noise experiments involved Mixer-7 interview sessions recorded over a high quality microphone close to the speaker. Noise at two different levels (6 dB-A and 15 dB-A) and of two different types (speech shaped noise and noise typical of heating, ventilation, and air conditioning systems (HVAC)) were then added to these recorded excerpts and included in the evaluation.

The train/test combinations examined were training without additive noise and testing on each of the additive noise conditions and train/test on matched noise conditions. Systems were not provided any development data with added noise.

5. PERFORMANCE MEASURES

Each trial of each test was independently judged as “true” (the model speaker spoke in the test segment) or “false” (the model speaker did not speak in the test segment), and the correctness of these decisions was tallied. This resulted in the determination of a miss rate P_{Miss} and a false alarm rate $P_{FalseAlarm}$:

$$P_{miss} = \frac{\#\ of\ target\ trials\ decided\ "false"}{\#\ of\ target\ trials}$$

$$P_{FalseAlarm} = \frac{\#\ of\ non-target\ trials\ decided\ "true"}{\#\ of\ non-target\ trials}$$

Further, each trial was assigned a score, where higher scores indicated greater belief or probability that “true” was the correct answer. In assessing performance these scores were pooled across all trials involving all target speakers. Thus these scores had to be normalized to be independent of the target speaker. The range of possible operating points was determined by thresholding based on these scores.

For additional submissions requested of participants beyond their primary system, the inclusion of decisions for each trial was optional. The ordering of the scores is all that matters for computing the application-based performance measures discussed below (sections 5.1 and 5.3) and for plotting DET curves (section 5.4). But the scores become more informative, and can be used to serve multiple applications, if they represent actual probability estimates. Participants were therefore asked to provide scores that may be interpreted as estimated log likelihood ratio values (using natural logarithms). In terms of the conditional probabilities for the observed data of a given trial relative to the alternative target and non-target hypotheses the likelihood ratio (LR) is given by:

$$LR = \frac{prob\ \ data\ target\ hypothesis}{prob\ \ data\ non-target\ hypothesis}$$

5.1. Primary (Official) Metric

The primary and official metric chosen for this evaluation is the value of PFA at the decision threshold (operating point) for which P_{Miss} is 0.1. This metric is new to NIST run evaluations of speaker recognition technology and was chosen since it is simple and intuitive, and is relevant across a range of applications.

5.2. DET Curves

The trial scores were also used to produce Detection Error Tradeoff (DET) curves[12], showing the range of possible system operating points and how misses may be traded off against false alarms. DET curves serve as a useful means of presenting system performance for the various conditions of interest.

6. Results

As described in the introduction (section 1) and should be clear from reading up to this point, the BEST interim assessment was large and complex. We share below some of the initial results of the evaluation, with the hope of being able to share further results in future publications.

6.1. Parametric Factors

6.1.1. Language

As described in section 4.1.1, there were several language conditions analyzed as part of the BEST interim assessment and train and test on same language was of particular interest.

Figure 1 compares the performance on trials where the train and test were both in either English or Spanish, and were external phone calls recorded over a telephone channel as part of Mixer-7 (MX7). In this plot, the trials consist of speech in the native language of the speakers included.

Performance is similar for English and Spanish for both the baseline[1] and BEST system, suggesting some level of robustness to language. Spanish seems to be an exceptional language, however, since, for this system, all languages other than Spanish fared worse than English, as we can see in Figure 2. It should be noted, however, that this system outperformed the baseline on all languages tested.

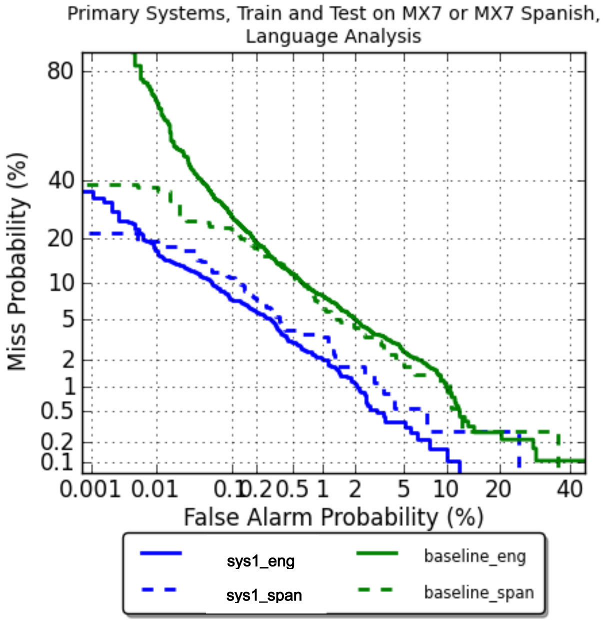

Figure 1: A DET-plot of same language telephone trials for one BEST system and a baseline system.

[1] The baseline was a Joint Factor Analysis system meant to represent the current state of the art at the start of the BEST program.

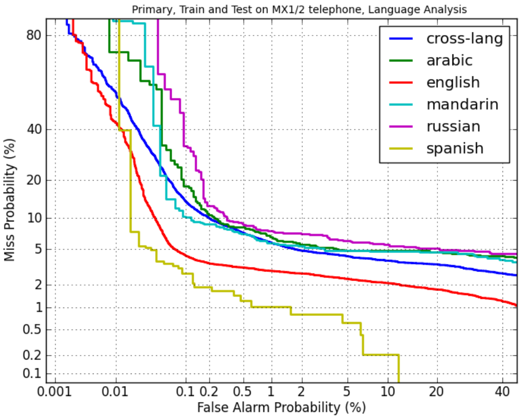

Figure 2: A DET-Plot showing a BEST system's performance on Mixer-1/2 external telephone trials with train and test on same language or, for cross-language, on English and another language.

It is worth mentioning that the corpora used in Figure 1 and Figure 2 (Mixer-7 and Mixer 1/2, respectively), are very different from one another, and so corpus effects are quite possible.

6.1.2. *Aging*

As in SRE10, the effect of target speaker aging was examined by utilizing the Greybeard corpus.

There were two periods of collection in the greybeard corpus, recent (where the data is referred to as "new") and past (where the data is referred to as "old").

Figure 3 shows performance on the greybeard corpus for a BEST system. The condition where the train data is "old" and the test data "new" is considered the aging condition. The similarity between the aging condition and test on old suggests the possibility of a corpus effect which would be a possible explanation for the train and test on new condition exhibiting the best performance. These results are similar to those observed in SRE10 [11].

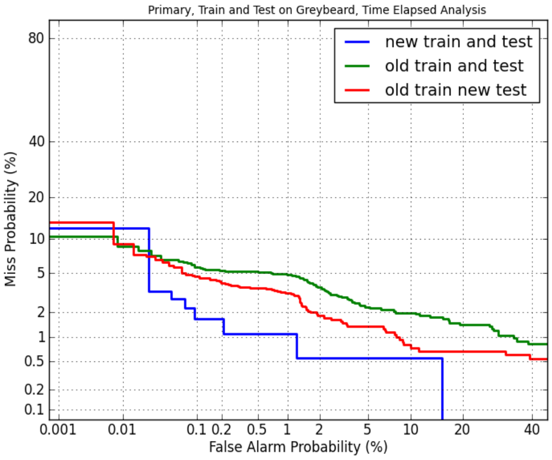

Figure 3: A DET-plot showing aging results from the Greybeard corpus for a BEST system.

6.2. Intrinsic Factors

6.2.1. *Speaking Style*

Figure 4 shows performance for a BEST system on data from the Mixer-7 corpus, fixing the speakers and training and test on the same microphone across conditions and varying the speaking style.

The train and test on interview condition demonstrated best performance, and, somewhat surprisingly, the cross-condition (train on interview and test on phone call) outperformed the matched train and test on phone call condition. One possible way to explain these results is that the interviews were typically 11 minutes in duration and relatively dense in speech of interest, while the phone calls were typically only 5 minutes in duration, only half of which was expected to be speech of interest. Another possible explanation is that the phone calls include HVE and LVE data and the interviews do not (note that this remains the case for the train on interview test on phone call condition).

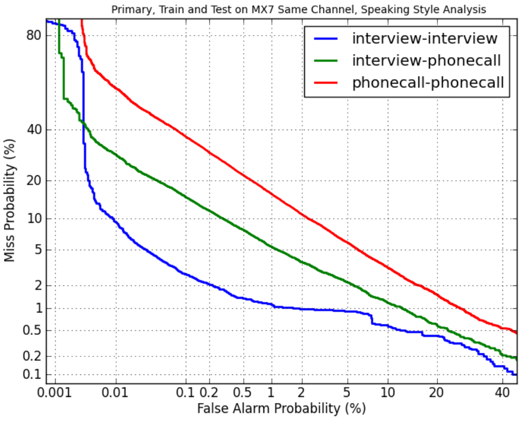

Figure 4: A DET-Plot showing a BEST system's performance on Mixer-7 same microphone data, either train and test on interview speech, train on interview test on phone call speech, or train and test on phone call speech.

6.2.2. *Vocal Effort*

The Mixer-7 corpus included internal phone calls collected while attempting to elicit either high or low vocal effort from the subject. The process was similar to that used in the Mixer-6 collection, as described in [5].

Similar to the results observed in SRE 10 [11], we see in Figure 5 that train and test on high vocal effort performed worst, as may be expected, but train and test low vocal effort out performed all other vocal effort conditions. In particular, the results for train and test on low vocal effort were so good that we did not observe enough errors to calculate the primary metric for this condition. Despite being consistent with our finding in SRE10, the relatively excellent performance on low vocal effort speech remains surprising and deserves still further exploration.

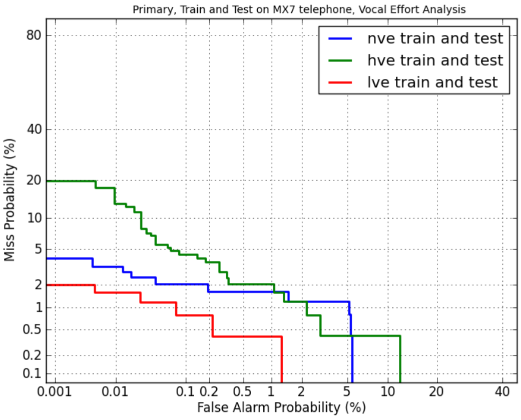

Figure 5: A DET-Plot showing a BEST system's performance on Mixer-7 internal phone calls, with matched train and test on low vocal effort (lve), normal vocal effort (nve), and high vocal effort (hve).

6.3. Extrinsic Factors

6.3.1. Reverb

The use of simulated reverb was new to the BEST interim assessment, and the amount of reverb used ranged from barely noticeable to so reverberant the speech was largely unintelligible.

As we can see in Figure 6, the system performance varied widely as a function of reverb level. It may be surprising to note that two of the (lesser) reverb conditions outperformed the (matched) no reverb condition over a range of operating points. This result, as well as the odd shape of the no reverb DET curve, seems worth exploring further.

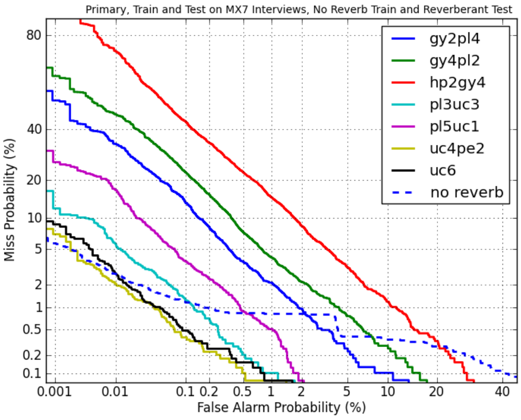

Figure 6: A DET-Plot showing a BEST system's performance on simulated reverb data. Each condition shown is train without any simulated reverb and test with the simulated reverb level listed in the legend. See Section 4.3.5 for description of the reverb data.

6.3.2. Additive Noise

Also new to the BEST interim assessment was the use of additive noise. Two different noise types (speech spectrum and HVAC) each at two different noise levels (6 dB-A and 15 dB-A SNR) were used.

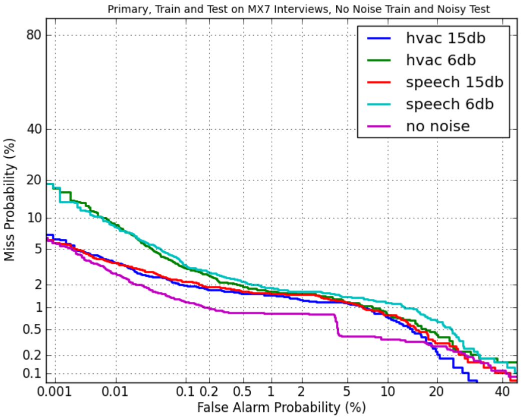

Figure 7: A DET-Plot showing a BEST system's performance on simulated noise data. Each condition shown is train without any added noise and test with the noise type and level listed in the legend.

Figure 7 shows the performance of a BEST system across additive noise conditions where there was no noise added to the training data. There was little difference observed between the noise conditions in the lower miss region. In the lower false alarm region, there was little difference observed between no noise and 15 dB noise of either type, while some degradation can be seen when testing on 6 dB of noise.

7. Future Work

Due to the aggressive and diverse nature of the goals of the BEST program, the BEST interim assessment was large and complex, and therefore offers an extraordinary opportunity for analysis. We have thus far only scratched the surface of the possible analysis. Below we offer a few suggestions for future work we hope to accomplish.

In the greybeard analysis, the non-target trials did not have the same corpus division as the target trials. It may be of interest to compare performance on trials from the same epoch and corpus (training and test from one of the old corpora included or from the newly collected conversations) versus trials with training and test from different epochs. Note that each Greybeard speaker might be viewed as two speakers, one from an older corpus, and one from newly collected data.

Additional metrics were defined for use in the BEST interim assessment. These include a non-linear cost function designed to strongly penalize mis-calibration, as well as $C_{LLR,}$ [13] an information theoretic metric designed to be application independent. We would like to explore the results using these metrics and to measure system calibration, which is not captured with the primary metric.

We would like to further explore the results with limited reverb applied, as well as the low vocal effort condition results. Suggested avenues for exploration of the reverb and low vocal effort results are described in sections 6.3 and 6.2.2, respectively.

8. Disclaimer

These results are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

9. References

[1] IARPA. *Biometrics Exploitation Science and Technology: Broad Area Announcement*. Available: https://www.fbo.gov/utils/view?id=50d1282e99d4fb552ecb3d6723dcf19b

[2] D. Graff and S. Bird, "Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies," in *Second International Language Resources and Evaluation Conference*, Athens, Greece, 2000.

[3] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data," in *Fourth International Conference on Language Resources and Evaluation*, Lisbon, Spain, 2004.

[4] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora," in *Interspeech*, Antwerp, Belgium, 2007.

[5] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Mixer 6," in *Seventh International Conference on Language Resources and Evaluation* Valletta, Malta, 2010.

[6] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Greybeard Longitudinal Study," in *Seventh International Conference on Language Resources and Evaluation* Valletta, Malta, 2010.

[7] NIST. *NIST Speaker Recogntition Evaluations*. Available: http://www.nist.gov/itl/iad/mig/sre.cfm

[8] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST Speaker Recognition Evaluation Chronicles - Part 2," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 2006, pp. 1-6.

[9] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer Corpora - 2004, 2005, 2006," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, pp. 1951-1959, Sep 2007.

[10] A. F. Martin and C. S. Greenberg, "The NIST 2010 Speaker Recognition Evaluation," in *Interspeech*, Makuhari, Chiba, Japan 2010, pp. 2726-2729.

[11] C. S. Greenberg, A. F. Martin, B. N. Barr, and G. R. Doddington, "Report on Performance Results in the NIST 2010 Speaker Recognition Evaluation," in *Interspeech*, Florence, Italy, 2011, pp. 261-264.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Eurospeech*, Rhodes, Greece, 1997, pp. 1895-1898.

[13] N. Brummer, "Application-Independent Evaluation of Speaker Detection," in *Odyssey2004 - The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 33-40.