

Automated Speaker Recognition in Real World Conditions: Controlling the Uncontrollable

HirotaKa Nakasone

Federal Bureau of Investigation
Quantico, VA, USA
hnakasone@fbiacademy.edu

Abstract

The current development of automatic speaker recognition technology may provide a new method to augment or replace the traditional method offered by qualified experts using aural and spectrographic analysis. The most promising of these automated technologies are based on statistical hypothesis testing methods involving likelihood ratios. The null hypothesis is generated using a universal background model composed of a large population of speakers. However, techniques with excellent performance in standardized evaluations (NIST trials) may not work perfectly in the real world. By defining and controlling the input speech samples carefully, we show quantitative differences in performance for different factors affecting a speaker population, and discuss on-going efforts to improve the accuracy rate for use in real world conditions. In this paper we will address two issues related to the factors that affect the system performance, namely the speech signal duration and the signal-to-noise ratio.

1. Introduction

Any automatic speaker recognition technique must deal with many adverse effects when applied under real world conditions. The input speech is provided to the Federal Bureau of Investigation (FBI) in recorded format, usually in analog form, or more increasingly in digital form. Once it leaves the speaker's vocal tract, the speech signal is modified by a variety of conditions. These conditions include the acoustic environment in which the signal is propagated, additive noise, transduction from acoustics to electronic signals by a microphone, telephone, amplification and filtering, and the recording process. In addition, the questioned voice, or test voice sample, could be extremely animated, excited, angry, agitated, distressed, joyous, sad, erratic, or could even be fearful.

The system performance depends on the quality and quantity of information available in the signal. One way of reducing the potential error rate is to identify those voice quality measures that have a significant impact on the performance and are algorithmically quantifiable. Our daily experience shows that the signal duration, signal-to-noise ratio (SNR), and the signal bandwidth measured from the input file have the most impact on the system performance. As with the spectrographic method of speaker recognition [1,2,3], the most often asked question is: "Is the evidence of sufficient length and quality?" There are other equally important factors to be investigated, but we will confine our report to the study results about the impact of signal duration and signal-to-noise ratio on automatic speaker recognition system performance. The Forensic Automatic Speaker

Recognition (FASR)¹ system used by the FBI is described briefly as follows [4]. Short time signal processing methods are used to extract acoustic feature measurements on digitized voice signals. The input signals are first windowed into short time overlapping frames. Each frame represents 50 milliseconds of speech, and the frames are overlapped by 80%, with one frame sliding every 10 milliseconds. Each window is transformed into a set of Mel-warped cepstral coefficients. Additionally, 14 delta cepstral coefficients are computed for each frame over a spread of five frames. These features are then compared with multiple speaker models and a background model using a Gaussian mixture model (GMM) classifier [5].

2. Experiments on Speech Duration and SNR

We know that the measured speech duration and SNR can directly influence the FASR system performance. If the speech duration is too short or SNR is too poor, it is normally a good practice to terminate any further recognition processing. However, what we do not yet know are: how short the signal duration could be, and how poor the SNR could be before we can justifiably terminate the recognition process. We conducted the experiment to measure the FASR system performance as a function of speech signal duration and as a function of speech SNR. The similar research on the effects of duration was also conducted by [6].

2.1 Experiments on Duration

For the experiment on duration we used clean, high fidelity, voice samples from the FBI Voice Database [7]. The microphone data, which is both text-independent and spontaneous, was used. We then evaluated the GMM recognition performance for a given set of training and test set signal durations. The FASR program was used to remove silence from the data, and to calculate the GMM scores.

2.1.1 Procedures for Duration Experiment

Step 1: We constructed a training set of signal-only data from the original Microphone training set. The data was resampled at 8 KHz in Microsoft WAV format. The training set included one spontaneous speaking mode file from each of 45 speakers. Each training file was limited to exactly 16 seconds of purified signal-only speech. Step 2: We constructed a test set of signal-only data from the original microphone training set. The test set included 5 files per speaker in the spontaneous speaking mode.

¹ Developed by the U.S. Air Force Research Laboratory, Rome Laboratory, Rome, New York, USA.

Each test file was also limited to exactly 16 seconds of purified signal-only speech. Step 3: We used FASR to generate speaker models, and then to calculate the GMM scores for each test file. We calculated the Rank1 ID measure and the Equal Error Rate (EER) performance for this batch of GMM scores. Care was taken so that the performance results would correspond closely to the original evaluation results [1] of Rank1 ID correct=100% and EER=1.3%. Step 4: We created test sets of increasingly shorter duration. The initial set included the following signal durations in seconds: [16, 12, 10, 8, 6, 4, 2, 1, .5]. We measured the FASR performance against the training set. Then we performed a group evaluation on the sufficiency of this set of signal durations. Step 5: We created training sets of the same durations as the test set. Step 6: We measured the FASR performance against the test set. We evaluated all the remaining combinations of training/test set durations.

2.2 Experiments on SNR

The purpose of this experiment was to quantify the performance of automatic speaker recognition using controlled homogenous speech energy to noise ratios. Speech signals can be characterized as highly non-stationary signals with time varying energy amplitude. Due to these dynamic characteristics and the fact that signal and noise sections are never observed separately in recorded speech samples, there is no simple definition of speech SNR. However, a histogram of a sufficiently large number of short time speech file energy segments will show two modes – one for the noise distribution and one for the signal plus noise. Therefore, we designed a procedure that gives consistent results over a wide range of speech signals, file lengths, and SNRs.

2.2.1 Procedures for SNR Experiment

The first step in the procedure was to subtract the DC energy, or mean, from the time series data. The time series data was then segmented into 10 millisecond non-overlapping frames. The energy was calculated for each frame and normalized by dividing by the maximum value over all frames. A histogram of the normalized frame energy was calculated, and the maximum value determined. The center point of the bin containing the maximum histogram value was set as the decision threshold. Frames with normalized energy less than the threshold were considered to be noise only. All other frames were considered to be signal plus noise. The variance of the noise-only frames and the variance of the signal plus noise frames were then calculated and used to determine the SNR ratio as shown by the equation below.

$$SNR = 10 \log_{10} \left(\frac{\sigma^2_{S+N} - \sigma^2_N}{\sigma^2_N} \right)$$

The SNR experiment used the same microphone data set as the duration experiment. Similar to the duration experiment, the first file for each speaker was used for speaker model training. The remaining files were used for speaker testing. There were 50 files originally identified as training files and 243 files originally identified as test files. The SNR was calculated for each of these files and all files with an SNR less than 20dB were

discarded. This left 38 training files and 186 test files. The remaining training and test data were used to generate WAV files at specified SNRs. This was done by adding a scaled time series of filtered (colored) noise to the audio speech file. The noise was generated according to a procedure based on the EIA-549-1988 standard from the USA Standards Institute (now ANSI). The specified SNR was achieved using an iterative process of adjusting the noise scale factor and generating a new noise time series until the calculated SNR was close to the desired SNR. The original FBI Database microphone speech files had calculated SNRs ranging from 50 dB to 12 dB. Thirty-eight speakers were used to create a new training corpus with a homogenous SNR of 20dB. FASR was used to create speaker models using this training corpus. Using the 186 test files, new speaker testing corpora were generated at eleven different SNRs ranging from 20dB to 0dB in 2dB steps. Channel normalization and noise removal were used with a Gaussian Mixture Model background model. The background model used in this experiment was the MALE_BAL universal background model with 1024 mixtures and generated from the NIST 1999 corpus.

3. Results

3.1. Results of Duration Experiments

Table 1 shows the results of speaker identification (ID) performance as a function of both test and training signal duration. The ID results are measured in terms of the percentage of correct matches between the test signal and the closest speaker model (Rank 1 ID match). Since we started out with very high quality speaker data, the ID performance for both 16 and 14 second duration speech samples were all 100%. The performance gradually decreased with decreasing duration, until a sharp decrease for durations below four seconds. For the 0.5 second training and 0.5 second test durations, the ID performance was essentially random. Since there were forty speakers in the ID performance, theoretically for random guessing is 2%.

Table 2 shows the results of speaker verification performance as a function of training and test duration. The measure we used for speaker verification performance for comparing different durations was the EER expressed in percent. The EER is the point at which the missed detection rate equals the false alarm rate. Although the entire detection error tradeoff (DET) curve (or ROC curve) gives the most accurate picture of detection performance, the EER performance was a convenient point to use for comparison purposes. As in the ID performance, the EER remained relatively low until the duration for both training and test durations dropped to four seconds.

Figure 1 shows ID performance in terms of percent correct matches for a fixed 16 second training set length. The ID performance is relatively flat and stays fairly stable between 16 seconds and 8 seconds. At the 6-second duration, the ID performance starts to drop off, showing poor performance for test signals less than 2 seconds. Figure 2 shows ID performance in terms of percent correct matches for a fixed 16-second test data length. Again, the ID performance is relatively flat and fairly stable from 16 seconds down to 6 seconds. At the 4-second, the ID performance starts to drop off, showing poor performance for training signal less than 2 seconds.

Table 1: ID Performance as a Function of Duration

| Test Set Duration in Sec | Training Set Duration in Sec | | | | | | | | | | |
|--------------------------|------------------------------|------|------|------|------|------|------|------|------|------|--|
| | 16 | 14 | 12 | 10 | 8 | 6 | 4 | 2 | 1 | 0.5 | |
| 16 | 100 | 100 | 99.2 | 98.8 | 99.2 | 97.7 | 91.9 | 74.4 | 43.8 | 20.5 | |
| 14 | 100 | 100 | 99.2 | 99.2 | 97.7 | 96.1 | 90.7 | 76 | 42.6 | 19.8 | |
| 12 | 100 | 100 | 98.8 | 96.5 | 95 | 95 | 87.6 | 70.5 | 37.2 | 17.8 | |
| 10 | 99.2 | 98.8 | 98.1 | 97.7 | 93.8 | 91.5 | 84.9 | 69 | 39.5 | 16.7 | |
| 8 | 96.1 | 94.6 | 93.4 | 91.9 | 88.8 | 86.1 | 77.5 | 61.2 | 36.1 | 18.2 | |
| 6 | 96.1 | 93.4 | 91.1 | 89.9 | 86.8 | 83.3 | 75.2 | 60.1 | 31 | 14.3 | |
| 4 | 87.6 | 86.8 | 83.7 | 81.8 | 78.7 | 75.2 | 68.6 | 52.3 | 25.2 | 12.4 | |
| 2 | 75.2 | 72.9 | 70.9 | 68.6 | 64.3 | 64.3 | 58.5 | 47.7 | 20.5 | 12 | |
| 1 | 50.4 | 47.7 | 47.3 | 46.5 | 45 | 45.4 | 43 | 43.4 | 12.8 | 5.4 | |
| 0.5 | 25.6 | 24 | 24.8 | 24 | 23.6 | 22.9 | 21.3 | 22.1 | 7 | 3.9 | |

Table 2: EER Performance as a Function of Duration

| Test Set Duration in Sec | Training Set Duration in Seconds | | | | | | | | | | |
|--------------------------|----------------------------------|------|------|------|------|------|------|------|------|------|--|
| | 16 | 14 | 12 | 10 | 8 | 6 | 4 | 2 | 1 | 0.5 | |
| 16 | 3.9 | 3.9 | 4.2 | 5.1 | 5.4 | 7.6 | 10.1 | 18.6 | 28.7 | 39.1 | |
| 14 | 5.2 | 5.7 | 5.5 | 6.6 | 7.8 | 9.6 | 11.6 | 19.6 | 29.4 | 40.3 | |
| 12 | 8.3 | 8.1 | 9.1 | 8.7 | 10 | 13.1 | 14 | 20.5 | 31.8 | 42.8 | |
| 10 | 9.3 | 9.3 | 10.3 | 10.2 | 12.7 | 14.2 | 15.6 | 23.8 | 32.6 | 43.8 | |
| 8 | 14.3 | 15.1 | 16.1 | 16.5 | 17.8 | 19.2 | 22.6 | 26.2 | 35.3 | 44.7 | |
| 6 | 19.8 | 19.4 | 20.6 | 21.3 | 22.5 | 23.3 | 23.6 | 27.3 | 37.2 | 46.1 | |
| 4 | 27.4 | 28.3 | 27.9 | 29.8 | 29.6 | 29.8 | 29.5 | 30.6 | 39.8 | 46.6 | |
| 2 | 18.4 | 38.6 | 38.4 | 36.4 | 37.1 | 37 | 35.3 | 33.6 | 42.6 | 48.5 | |
| 1 | 44.6 | 45.1 | 45.3 | 44.1 | 45 | 43.7 | 42.2 | 39.3 | 46.1 | 49.6 | |
| 0.5 | 48.5 | 48.7 | 48.4 | 48.6 | 48.1 | 47.3 | 47.8 | 45.1 | 50 | 50.8 | |

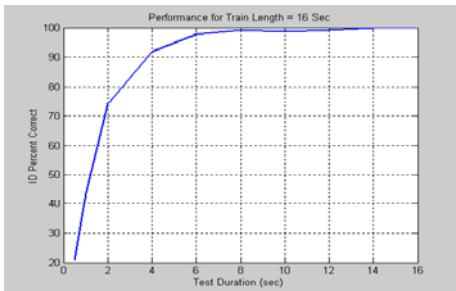


Figure 1: ID Performance for Variable Test Length, Fixed Training Length

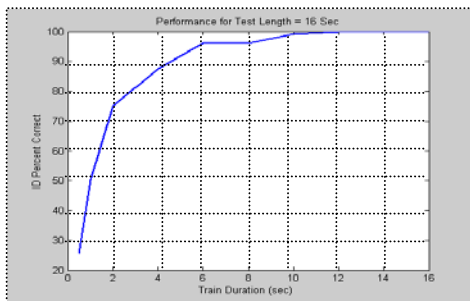


Figure 2: ID Performance for Variable Training Length, Fixed Test Length

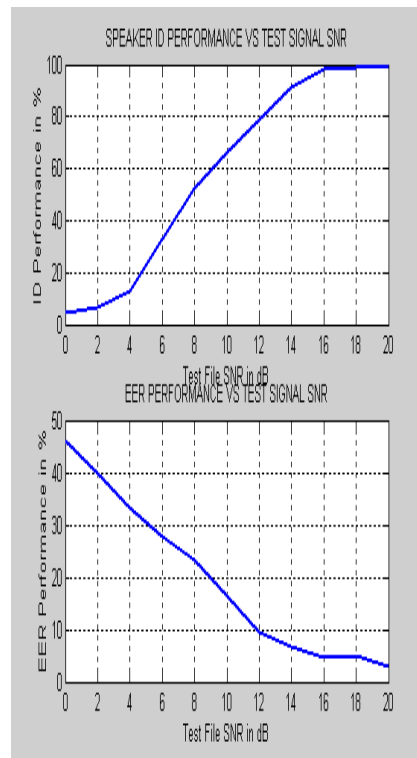


Figure 3: Speaker ID (top) and EER (bottom) Performance as a function of Test Signal SNR.

3.2. Results of SNR Experiments

Figure 3 shows speaker identification and EER performance as a function of test signal SNR in dB. Note that as the SNR decreases, the ID performance decreases and the EER increases. There is also a sudden decrease in performance for test signals with less than 14 dB SNR. Figure 4 (a-c) shows the probability density function for true and false LLR scores for: (a) test files = 20 dB SNR, (b) test files = 12 dB SNR, and (c) test files = 0 dB SNR. Training files were all 20 dB SNR. Figure 4(a) shows a clear separation between true and false distributions when both SNRs of the training and test files are 20 dB. As shown in Figure 4(b), when SNR of test files decreases to 12 dB, there is increasing overlap between true and false distribution. As shown in Figure 4(c), when the SNR of test files is 0 dB, we can almost see a complete overlap between true and false distributions.

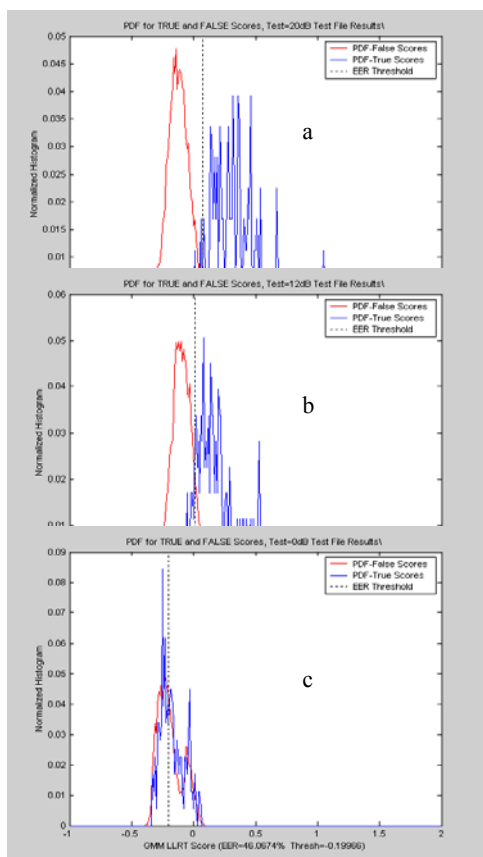


Figure 4 (a-c). Probability Density Function of True and False Scores derived from SNR experiment.

4. Conclusions

We showed the effect of both training signal duration and test signal duration based on a microphone quality voice data on automatic speaker recognition performance. The ID

performance and the EER are excellent when both training and test signal duration are at least in the range of 12 seconds to 16 seconds. The results also indicate that the signal duration of either the training and/or test files of 6 seconds or less yielded poor recognition performance. We also showed that ID performance is clearly affected as a function of the test signal SNR, revealing obvious degradation at approximately 16 dB.

These two voice quality experiments provide insight about setting up cautionary thresholds below which unreliable and misleading scores may result. Baseline knowledge of voice signal duration and SNR may be used to facilitate a set of safeguards against potential misuse of automated speaker recognition systems. The present experiments were conducted with a clean microphone voice database, therefore the effects of these two quality measurements do not represent the realistic forensic conditions.

Additional experiments are needed to study the joint effects of signal duration and SNR upon ASR performance. The present study on speech duration and SNR must be validated by larger speaker populations.

5. Acknowledgements

The author appreciates individuals who contributed to this manuscript including Steve Beck, Donald Wallace, Somit Mathur, Carson Dayley of BAE Systems; and Maria Mimikopoulos, Barbara Snyder, and Artese Kelly of the Forensic Audio, Video and Image Analysis Unit of the FBI.

6. References

- [1] Bolt, R., Cooper, F., David, E., Denes, P., Pickett, J., and Stevens, K. "Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes", *Journal of Acoustical Society of America*, Vol. 47, NO. 2 (Part 2), pp.597-612, 1970.
- [2] Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J, and Nash, E., "Experiment on Voice Identification", *Journal of Acoustical Society of America*, Vol. 51, No. 6, pp. 2030-2043, 1972.
- [3] Committee on the Evaluation of Sound Spectrograms, National Academy of Sciences, "On the Theory and Practice of Voice Identification", 1979, Washington, DC.
- [4] Nakasone, H., Beck, S., "Forensic Automatic Speaker Recognition", *Proc. Of Odyssey 2001 Speaker Recognition Workshop*, pp. 139-144, Crete, Greece, 2001.
- [5] Reynolds, D., Quatieri, T., Dunn, R. "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, 19-41, 2000
- [6] Pfister, B., "Estimating the Weight of Evidence in Forensic Speaker Verification", in this Proceedings, Special Session on Forensic Speaker Recognition, *Eurospeech 2003 – Switzerland (Interspeech 2003)*, September 1-4, 2003, Geneva, Switzerland.
- [7] Linguistic Data Consortium, "Cross-Channel Forensic Speech for Automatic Speaker Recognition", LDC 2003 S04, pending a public release in May 2003.