Standoff Speaker Recognition: Effects of Recording Distance Mismatch on Speaker Recognition System Performance

Mike Fowler[1], *Mark McCurry*[2], *Jonathan Bramsen*[3]
Kehinde Dunsin[3], *Jeremiah Remus*[3]

[1]Department of Mathematics, Clarkson University, Potsdam, New York
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
[3] Department of Electrical and Computer Engineering, Clarkson University, Potsdam, New York

Abstract

Speech can potentially be used to identify individuals from a distance and contribute to the growing effort to develop methods for standoff, multimodal biometric identification. However, mismatched recording distances for the enrollment and verification speech samples can potentially introduce new challenges for speaker recognition systems. This paper describes a data collection, referred to as the Standoff Multi-Microphone Speech Corpus, which allows investigation of the impact of recording distance mismatch on the performance of speaker recognition systems. Additionally, a supervised method for linear subspace decomposition was evaluated in an effort to mitigate the effects of recording distance mismatch. The results of this study indicate that mismatched recording distances have a consistent negative impact on performance of a standoff speaker recognition system; however, subspace decomposition techniques may be able to reduce the penalty observed with mismatched recording distances.

Index Terms: Speaker recognition, far-field speech, beamforming, standoff biometrics

1. Introduction

There is growing interest in the use of biometric signatures collected in standoff scenarios to identify individuals from a distance. There are a number of scenarios, including surveillance and defense applications, where it would be advantageous to remotely identify an individual, possibly without cooperation of the individual or alerting them that biometric data is being collected. To enable biometric identification at a distance, the growing consensus is that a multi-modal approach for collecting biometric information is needed [1, 2, 3]. There has been substantial progress towards the development of robust speaker recognition techniques, and significant potential for fusion with other biometrics systems (e.g. iris, face, physiological). Therefore, in addition to measuring traditional biometric information, it may be necessary to consider other signatures that can be easily gathered from individuals at a distance, such as speech, that may contain useful identifiers.

With increasing interest in the collection of biometrics at a distance, it would be beneficial to have a clearer understanding of the sensitivity of speaker recognition systems to the changes in captured speech when recording speech at a distance. It is reasonable to expect that greater distance from the recording device will degrade the signal- to-noise ratio and introduce more room acoustic artifacts; however most investigations of audio quality and its effect on speaker identification performance have focused on the channel quality (e.g. telephone lines or mobile handsets). A previous speaker recognition investigation [4] using the MultiRoom data set initiated interest in the effect of recording distance mismatch. The data set contained recordings in different rooms at three distances (microphones at 1 ft, 1/3 of room length, and 2/3 of room length). However, only a subset of the data was made available, restricting analysis of the specific effects of distance mismatch. A larger investigation with more recordings in a greater diversity of conditions will allow for a more complete investigation of the effect of recording distance mismatch.

This paper describes the Standoff Multi-Microphone Speech Corpus, which was collected to allow investigation of the effects of recording distance on speaker recognition system performance. The remainder of the paper is organized as follows. In Section 2, related work in the area of speaker identification is reviewed. Section 3 describes the custom platform for synchronized capture of multi-microphone recordings, experiment setup, and data collection. Sections 4 and 5 present baseline results that illustrate the effects of distance mismatch on speaker recognition, as well as an initial effort to mitigate the effects of recording distance mismatch. Discussions and conclusions are presented in Section 6.

2. Relation to Prior Work

While there have been significant efforts within the speaker recognition research community to develop methods for handling session-to-session speaker variability or variations introduced by different microphones (e.g. [5, 6]), it is unclear how well these solutions can address the problem of speech recorded at a distance, particularly when trying to match speakers using recordings measured at different distances. There have been several studies that combine beamforming and speaker recognition (e.g. [7]), and some studies of far-field or distant speech for speaker identification (e.g. [8, 9]), but the effect of distance mismatch and specific methods to mitigate its effects have not been thoroughly investigated. Partial least squares (PLS) has recently been introduced to the speaker recognition community [10, 11] as a viable tool. However, this approach has also not yet been investigated as a method for mitigating the effects of distance on speaker recordings.

3. Standoff Multi-Microphone Speech Corpus

The Standoff Multi-Microphone Speech Corpus was designed and collected to provide a data set that allows direct examination of the effect of recording distance on speaker recognition system performance. The data set contains 46 speakers, with two sessions at five different distances: 1.5m, 2.4m, 4m, 6.4m, and 10.4m (5ft, 8ft, 13ft, 21ft, and 34ft). The experiment protocol was designed with the goal of providing recordings with at least 90 seconds of speech after silence removal. In the collected data, the average length of the parsed segments is 125 seconds with a standard deviation of 12 seconds.

The data collection contains 18 channels of synchronized, recorded speech. Sixteen of the eighteen channels are configured to serve as an 8x2 element microphone array, allowing the use of beamforming techniques to enhance signal-to-noise ratio in the processed speech. The component microphones in the beamform array are omnidirectional electret condenser microphones produced by CUI, Inc. (part number CMA-4544PF-W). After 40 dB of amplification, the beamform array microphones are sampled at 48 kHz and 16-bit resolution. Horizontal spacing of the beamform elements is 5 inches (12.7 cm), with an overall array width of 35 inches (88.9 cm). The vertical spacing between the two rows of eight microphones is twelve inches (30.5 cm). The platform housing the beamform array also contains a studio condenser microphone (Audio Technica AT2020USB) and a supercardioid directional microphone (Rode VideoMic). The data collection array is shown in Figure 1. It should be noted that, in this work, the beaform array was simply used as a method of data collection. Utilizing beamforming algorithms to mitigate classification error stemming from distance and channel mismatch is saved for future work.

The data collection was performed in a 41 ft by 23 ft classroom that had background HVAC noise typical of an office environment. The microphone array was placed 3 ft from the wall (in the same spot for each recording session), and subjects were seated at the appropriate distance from the microphone for each recording condition. The study participants received prompts via headphones and were instructed to repeat (in their own voice and speaking style) what they heard. Sentence stimuli were assembled from various corpora including the Hearing in Noise Test (HINT), CID Everyday Speech Sentences, Harvard Sentences, and sound clips from recorded lectures. The presentation order of the sentences was randomized for each study participant without repetition of any sentences. The headphone output is available in the raw data waveforms as the 19th audio channel. The order of the distances were also randomized, with subjects completing one pass through all five distances 5ft, 8ft, 13ft, 21ft, 34ft before taking a short break and repeating the process. The first and second pass through each of the five distances are referred to as seg1 and seg2 in the dataset.

The entire data set, as full-session-length unparsed WAV files as well as parsed WAV files separated by channel, distance, and subject number, along with associated documentation, are available via SFTP download. Information for accessing the dataset is available at http://people.clarkson.edu/∼fowlermj/smmsc.html

4. Effects of Recording Distance and Channel Mismatch

The Standoff Multi-Microphone Speech Corpus was analyzed to assess the impact of recording distance mismatch and cross-

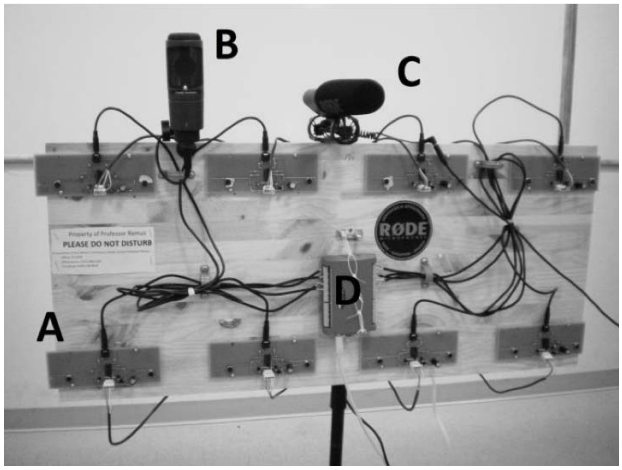

Figure 1: Microphone array platform used in the data collection. Labeled in the photo are (A) one of the eight microphone array circuit boards, containing two electret microphones, (B) the studio condenser microphone, (C) the supercardioid microphone, and (D) the 16-channel DAQ.

microphone conditions on speaker recognitions system performance. The baseline speaker recognition system was implemented using the ALIZE open-source toolbox [12] and based on a framework using GMM-UBM supervectors. The 45-dimensional cepstral-based features (15 MFCC, 15 delta, 15 delta-delta) were extracted from 20 millisecond windows with 50% overlap and normalized before adapting a 1024-component universal background model (UBM). The supervectors were generated from the GMM-UBM component mean vectors, and normalized by the UBM model. A total of 2185 supervectors were available after processing.

Baseline speaker recognition results were generated using the raw high-dimensional supervectors. A nearest neighbor classifier using Pearson correlation as a distance measure was used to generate similarity scores between supervectors in the training (i.e. enrollment) and testing (i.e. verification) data sets. The setup of the classification experiment was designed to focus on the evaluation of the effects of recording distance mismatch between training and testing speech samples. There are 25 pairs of distance train/test conditions (including training and testing at the same distance). For example, training with speech recorded at a distance of 2.4m and testing with speech reordered at 10.4m yields a distance mismatch of 8m. The other 8m data point comes from training at 10.4 m and testing at 2.4m. The equal error rate (EER) was calculated from the performance curve found for each pair of train/test conditions. To isolate the effects of distance mismatch and avoid potentially conflating distance mismatch and cross- microphone speaker identification, distance mismatch was evaluated within matched microphone conditions (for all microphones). Decision metrics were then aggregated across the different runs of the classifier for each microphone to allow generation of a single performance curve from which the EER was calculated.

Figure 2 shows the results of the speaker recognition experiments focused on analyzing the effects of recording distance. In the top subplot, the equal error rate is shown for the cases where recording distance was matched in training and testing. As hypothesized, there is some observed increase in EER when

using speech recorded at greater distances, a likely reflection of the decreased signal-to-noise ratio. The bottom subplot in Figure 2 plots the equal error rates for all 25 pairs of distance train/test conditions as a function of the distance mismatch between the training and testing data. The scatter plot reveals a distinct linear trend between distance mismatch and EER. A slope of 0.0063 was found using a linear regression fit to the data scatter.

5. Subspace Decomposition to Mitigate Effects of Condition Mismatch

The results presented in Figure 2 reveal a substantial penalty when the training and testing speech samples are recorded at different distances, with larger mismatch leading to higher error rates. A preliminary attempt was made at mitigating the effect of this recording distance mismatch using a linear subspace decomposition method to project the high- dimensional supervectors into a lower-dimensional subspace where recordings from the same speaker recorded at different distances are mapped to the same point. The subspace projection method used in this initial study was partial least squares decomposition.

Partial Least Squares (PLS) can be used as a dimension- reducing technique that finds projections of the supervectors into lower dimensional subspaces that account for most of the variation in the data. Our assumption is that most of the variability in our data is associated with the speaker and that the variability stemming from channel, session and distance differences will be mitigated by projecting into the subspace found by PLS. In this sense PLS is similar to Principal Components Analysis (PCA). One of the key differences is that while PCA is an unsupervised learning technique, PLS uses the class labels as well as the supervectors to project into a subspace where data from the desired speaker is well separated from that of imposters. A detailed discussion of PLS can be found in [13].

If given m supervectors x_k with corresponding class labels y_k from d different speakers, an m by n data matrix X can be formed by vertically concatenating the n-dimensional supervectors x_k. The m by d binary response matrix Y has entries y_{ij} = 1 if x_i comes from speaker j and y_{ij} = 0 otherwise. The PLS decomposition is given as

$$X = TP^T + E \quad (1)$$
$$Y = UQ^T + F \quad (2)$$

where T and U are the m by p ($p < n$) projections of X and Y, respectively, and P (n by p) and Q (d by p) are the loading matrices, and E (m by n) and F (m by d) are the residual matrices. The dimensionality p of the projection subspace is a parameter whose value can be optimized using cross- validation techniques. The matrices T and U are constructed by finding a series of weight vectors w_i and c_i such that

$$[\mathrm{cov}(t_i, u_i)]^2 = \max_{||w_i||=||c_i||=1} [\mathrm{cov}(Xw_i, Yc_i)]^2 \quad (3)$$

where $\mathrm{cov}(t_i, u_i)$ denotes the sample covariance between vectors t_i and u_i. In the design of the speaker recognition system, it is necessary to have a set of development data that are used to learn the PLS projection loadings. In this study, the development dataset was specific to the microphone used in each training/test pair. To mitigate distance mismatch, four PLS projection matrices were constructed (one for each microphone in the dataset) using microphone-specific development datasets. However, the development data did not contain any recordings at the two distances present in the training and test data; these were excluded to prevent overfitting and to provide a more robust result. Further possibilities regarding the development data set used to learn the PLS projection loadings are discussed in Section 6.

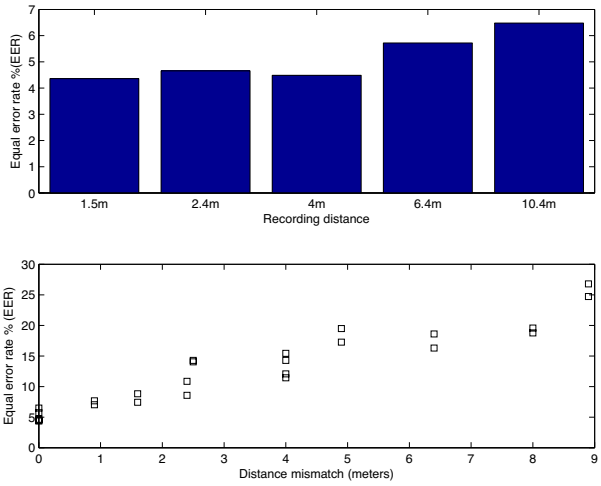

Figure 2: (Top) Equal error rates for the speaker recognition system at each of the five distances. (Bottom) Scatter plot of equal error rate versus mismatch in the recording distances of the enrollment and verification data. There is one marker per train/test data pair (25 markers in total).

The experiment design was nearly identical to the train/test classification setup used in Section 3, with the only difference being the projection of the raw supervectors to a 25-dimensional subspace using PLS prior to running the nearest neighbor classifier. Results using PLS to mitigate the effect of recording distance mismatch are presented in Figure 3. For additional clarity the EER rates for the classifier with and without the use of PLS are given in tables 1 and 2, respectively. The scatter plot from Figure 2 showing EER versus distance mismatch is reproduced to compare the effect of PLS decomposition on classification performance. Figure 3 shows that the effect of distance mismatch on system EER has been reduced significantly by the PLS decomposition. A linear least squares fit to the data scatter determines a slope of 0.003, substantially less than the change in EER vs. distance mismatch when using the raw supervectors for classification. The results shown in Figure 3 suggest that the PLS projection, intended to find a subspace where recordings of the same speaker at different distances map to the same point, is able to reduce the variability in the supervector features that results from mismatched recording distances.

6. Discussion and Conclusions

This paper outlines a study of the effects of recording distance mismatch on performance of a speaker recognition system, with the objective of furthering the development and use of multimodal biometrics for identification at a distance. As part of this study, a data set was collected and will be provided to other researchers who may be interested in investigating techniques for beamforming or the effects of recording distance and different microphones in speaker recognition.

The results of this study reveal that there is a clear, strong

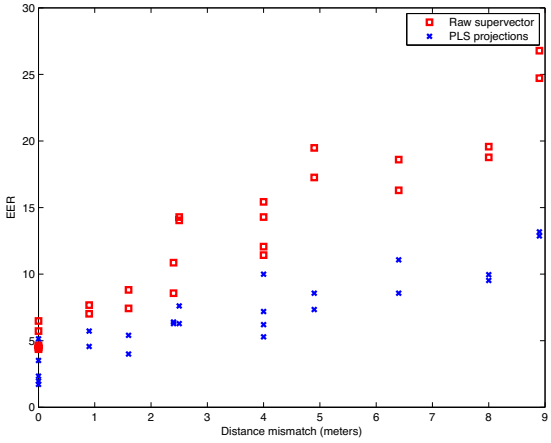

Figure 3: Scatter plot of the equal error rates (EER) versus the mismatch in recording distance for the enrollment and verification speech samples. Speaker recognition results are shown for both the raw supervector features (square markers) and PLS-decomposed supervectors (x markers). There are 25 markers for each data set; one marker per train/test data pair.

trend in the degradation of performance with distance mismatch (over the range of distances considered), though not quite as severe as in the case of cross-microphone speaker recognition. This penalty for distance mismatch will hinder the development of systems for successful standoff speaker recognition if it is not addressed. The proposed method of partial least squares decomposition, which is capable of exploiting development data to learn a supervised decomposition of the high-dimensional supervectors, showed a significant reduction in the effect of distance mismatch. However, further study is necessary to determine whether other techniques that are popular within the speaker recognition community for managing channel and session variability will perform similarly on the distance mismatch challenge.

A significant question with the use of partial least squares decomposition is the choice of development data for learning the projection loadings. It is not clear whether development data should be as specific as possible for the intended task, or more general to potentially increase robustness of the learned projection. For example, in the current study, the PLS projections to mitigate distance mismatch were microphone-specific since each iteration of the classification experiment included training and test data from only a single microphone. The development data for PLS could have instead included data from all microphones and distances, and only a single PLS projection matrix would have been necessary (and more appropriate if the training and testing data used different microphones, in addition to being recorded at different distances). There are a number of combinations of data that can be included in the development data set for learning the PLS projection, and the best choice may be the one that reflects the potential use scenarios for the speaker recognition system that is being designed.

Table 1: EER using GMM Supervectors

	Training Distance				
Test Distance	1.5	2.4	4	6.4	10.4
1.5	4.36	7.67	14.29	19.48	24.71
2.4	7.02	4.66	8.82	15.43	19.57
4	14.04	7.42	4.49	10.86	16.29
6.4	17.26	12.07	8.57	5.72	11.43
10.4	26.78	18.77	18.60	14.29	6.48

Table 2: EER using PLS on GMM Supervectors

	Training Distance				
Test Distance	1.5	2.4	4	6.4	10.4
1.5	1.96	5.72	6.28	8.57	13.17
2.4	4.56	2.34	4.00	6.20	9.52
4	7.61	5.40	1.72	6.28	8.57
6.4	7.34	5.29	6.40	3.52	7.19
10.4	12.87	9..96	11.07	10.00	5.14

7. References

[1] J. Fei and I Pavilidis. “Thermistor at a distance: Unobtrusive measurement of breathing”. IEEE Trans. Biomed. Eng. 2009.

[2] Z. Liu and S. Sarkar. “Outdoor recognition at a distance by fusing gait and face”. Image and Vision Comp. 25(6), pp. 817-832. 2007.

[3] C. Fancourt, L. Bogoni, K. Hanna, Y. Guo, R. Wildes, N. Takahashi and U. Jain. “Iris Recognition at a distance”, Lecture Notes in Computer Science 3564, Springer, Ed. 2005.

[4] J.J. Remus, J.M. Estrada, and S.A.C. Schuckers. “Mitigating effects of recording condition mismatch in speaker recognition using partial least squares”, INTERSPEECH. 2012

[5] N. Dehak P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet. “Front-end factor analysis for speaker verification”, IEEE Transactions on Audio, Speech, and Language Processing 19(4), pp. 788-798. 2011.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. “Joint factor analysis versus eigenchannels in speaker recognition”, IEEE Transactions on Audio, Speech, and Language Processing 15(4), pp. 1435-1447. 2007.

[7] I.A. McGowan, J. Pelecanos, and S. Sridharan. “Robust speaker recognition using micrphone arrays”, Presented at 2001: A Speaker Odyssey-the Speaker Recognition Workshop. 2001.

[8] Q. Jin, R. Li, Q. Yang, K. Laskowski and T. Schultz. “Speaker identification with distant microphone speech. Presented at 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2010.

[9] Q. Jin, T. Schultz, and A. Waibel. “Far-field speaker recognition”, IEEE Transactions on Audio, Speech, and Language Processing 15(7), pp. 2023-2032. 2007.

[10] B. Srinivasan, D. Garcia-Romero, D. Zotkin, and R. Duraiswami. “Kernel partial least squares for speaker recognition”, INTERSPEECH. 2011.

[11] B.V Srinivasan, D. N. Zotkin, and R. Duraiswami. “A partial least squares framework for speaker recognition,” ICASSP. 2011

[12] J. F. Banastre, N. Scheffer, D. Matrouf, C. Fredouilee, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B Fauve, J. Mason. “ALIZE/SpkDet: A state-of-the-art open source software for speaker recognition”, in Proceedings of Odyssey. 2008.

[13] R. Rosipal and N. Kramer. “Overview and recent advances in partial least squares”, Subspace, Latent Structure and Feature Selection, pp. 34-51. 2006.