DATA SELECTION FOR WITHIN-CLASS COVARIANCE ESTIMATION

Elliot Singer, Tyler Campbell, and Douglas Reynolds*

Massachusetts Institute of Technology
Lincoln Laboratory
{es,dar}@ll.mit.edu
*Rensselaer Polytechnic Institute
tylercampbell@mac.com

ABSTRACT[1]

Methods for performing channel and session compensation in conjunction with subspace techniques have been a focus of considerable study recently and have led to significant gains in speaker recognition performance. While developers have typically exploited the vast archive of speaker labeled data available from earlier NIST evaluations to train the within-class and across-class covariance matrices required by these techniques, little attention has been paid to the characteristics of the data required to perform the training efficiently. This paper focuses on within-class covariance normalization and shows that a reduction in training data requirements can be achieved by proper data selection. In particular, it is shown that the key variables are the total amount of data and the degree of handset variability, with total calls per handset playing a smaller role. The study offers insight into efficient within-class covariance matrix training data collection in real world applications.

Index Terms— channel compensation, i-vectors, within-class covariance, hyperparameter training

1. INTRODUCTION

The recent development of low dimensional vector representations of speech utterances has led to considerable improvement in the performance of speaker recognition systems submitted to NIST's periodic speaker recognition evaluations. In particular, the i-vector method, which models the speech utterance in a total variability subspace, has emerged as the predominant approach due to its state-of-the-art performance, low computational complexity, and compact representation [1]. Of equal importance is the robustness of the recognizer to channel and session variability, an area that has a long history of development. Classical feature domain techniques that have become standard components of many feature extraction front-ends include cepstral mean subtraction, RASTA, feature normalization, feature mapping, and feature warping [2-6]. With the introduction of i-vectors and their low dimensional property, advanced techniques that model the channel/session subspace of utterances, such as nuisance attribute projection (NAP), within-class covariance normalization (WCCN), linear discriminant analysis (LDA), probabilistic linear discriminant analysis (PLDA), and length normalization [7-10] have given researchers effective tools for performing channel and intersession compensation. These latter methods, which require the estimation of a within-class and between-class covariance matrices, rely on the availability of multiple utterances from a large population of speakers. Fortunately, participants in NIST evaluations have access to a vast repository of legacy data from earlier tests that contains large quantities of labeled speaker data, including numerous instances of utterances recorded by a single speaker using the same or different phone numbers. This resource allows participants to evaluate compensation techniques and assess their effectiveness during the development cycle simply by using all or a subset of the available data. The system is thereby tuned for optimum performance, and in subsequent evaluations more legacy data becomes available and the process is repeated.

Although this procedure has been effective, little if any understanding is gained as to the required characteristics of the compensation training data, and the process will not necessarily be sustainable for real world applications where acquisition of such labeled training data may be expensive. To our knowledge, little if any attention has been paid to the problem of the efficient collection and selection of compensation data, and we have thus designed a series of experiments to gain insight into this issue. Results indicate that the key variables are the total amount of data and the degree of handset variability, with total calls per handset playing a smaller role.

The remainder of the paper is organized as follows: Section 2 provides a description of the speaker recognition system, the corpus used, and the experiment design. Section 3 describes a series of experiments designed to reveal underlying desirable properties for a compensation training corpus. Section 4 details results of applying the guidelines to the development data and to a held-out set, and Section 5 presents a summary of the results and suggestions for future work.

2. SYSTEM AND DATA

2.1. Recognizer

The system architecture used in this study consisted of an i-vector generator followed by within-class covariance normalization (WCCN) and cosine scoring. The decision to focus exclusively on WCCN rather than methods requiring both within-class and across-class normalization (e.g., LDA, PLDA) was motivated by the

Distribution A: Public Release, unlimited distribution. 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

desire to avoid the complexity of coupling effects when using both within-class and across-class matrices. In this sense, this paper can be viewed as a preliminary study. In addition, our studies have shown that WCCN alone produces a substantial performance gain, albeit not as great as the gains achieved by using both covariance matrices (e.g., as in PLDA). Support for this choice will be provided in the results section.

Speech segments were first extracted from the utterances using speech activity detection (SAD) based on GMMs trained on telephone speech. Mel-frequency cepstral coefficient (MFCC) features [11] were then computed from the speech signal using a mel-spaced filterbank comprising 20 triangular filters whose center frequencies spanned the range of 300-3140 Hz. Delta cepstra were computed over a ±2 frame span and appended to the cepstral vector, and each component of the resulting 40-dimensional feature vector was normalized to have zero mean and unit variance. Finally, the utterance feature vectors were converted to i-vectors using a 2048-order Universal Background Model (UBM) and a rank-600 total variability (T) matrix.

The estimated within-class covariance matrix was computed via [1]

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_i^s - \overline{w^s})(w_i^s - \overline{w^s})^t$$

where w_i^s is the i^{th} i-vector for speaker s, n_s is the total utterances for speaker s, S is the total number of utterances across all speakers, $\overline{w^s}$ is the average i-vector for speaker s, and $(\cdot)^t$ denotes transpose. Given an i-vector x extracted from an enrollment utterance and an i-vector y extracted from a test utterance, the match score between the two i-vectors is computed as

$$score(x,y) = \frac{x^t W^{-1} y}{\sqrt{x^t W^{-1} x}\cdot\sqrt{y^t W^{-1} y}}$$

The purpose of WCCN is to deemphasize directions of the estimated within-class space with high variance. All directions of the space are maintained, unlike NAP where directions of the space are projected away or LDA which performs a rotation and dimensionality reduction.

2.2. Experiment corpus

Development data for this study consisted of male telephone utterances chosen from previous NIST evaluations. Training data for the UBM and T-matrix was obtained from the NIST Switchboard 2 phases 2-5 [12] and SRE04/05/06 utterances [13] from male speakers having a minimum of eight calls per speaker, totaling 15559 utterances from 1115 speakers. Training sets for the WCCN matrix were selected from the SRE04/05/06 male speaker utterances, totaling 7252 files from 514 speakers. Enrollment and test data utterances were obtained from SRE08 [13] and consisted of 479 male speakers with 3 calls/speaker for enrollment and 895 utterances for test (NIST's SRE08 "3conv-train short3-test" condition). Speaker models were created by averaging the i-vectors of the individual training utterances prior to performing WCCN. Enrollment utterances totaled 1437 files and test utterances included 708 target trials and 8688 nontarget trials. Test utterances were never from the same phone numbers as enrollment utterances.

The availability of metadata information for the NIST utterances allowed for a more detailed study of selection criteria for WCCN training. Among the metadata fields available, the phone number associated with each call appeared to be the most relevant as it could serve as a proxy for handset variability. In the WCCN training data, the number of different phone numbers used by a speaker ranged from 1 to 17, with a median of 5.

Figure 1 shows the histogram of the WCCN training data plotted vs. the number of speakers available with a fixed number of calls. By design, this set includes only those speakers who made 8 or more calls. Generally, the speakers made between 8 and 28 calls, with a couple of outlier speakers who made 33 and 51 calls. In many experiments, multiple draws (between 5 and 30) were taken from the pool of data in order to establish statistical significance in the final result and to illustrate variability.

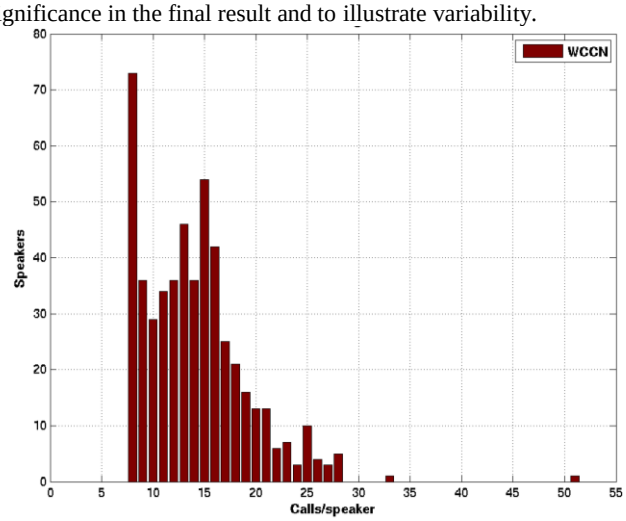

Figure 1: Profile of data available for training the WCCN matrix.

3. EXPERIMENTS

3.1. Amount of Data

A series of experiments was conducted to obtain performance baselines and to measure the sensitivity to varying amounts of WCCN training data. This was achieved by adjusting the total number of speakers used during training while maintaining the histogram profile. For example, for a 50% training run, half the speakers from each of the histogram bins were drawn randomly and all their calls were used in training. Figure 2 shows system performance for percentages between 20% and 90%, where each point represents the equal error rate for scores merged from 30 random draws of the data. The "100%" label represents utilization of all the data and the "0%" label represents using no data (i.e., no WCCN). The plot indicates that for this configuration, WCCN produces a considerable improvement in performance, with additional benefits trailing off when using calls from more than approximately 50% of the available speakers. The result indicates that baseline performance is achieved using 3626 calls from 257 speakers.

Figure 3 shows DET plots using of 100% of the WCCN training data, 50% of the data (3626 utterances from 257 speakers), and no data (0%). The plot for 50% was obtained by merging scores from 20 random draws, with the DETs for the individual draws shown superimposed to illustrate the degree of variability in the result. Figure 3 provides further confirmation that fewer utterances than are available for training are needed to achieve performance equivalent to the 100% baseline, and this observation

encouraged further investigation of factors in the WCCN training data that could influence recognition performance.

3.2. Phone number variability

To examine the impact of handset variability within the WCCN training set, we first constructed an experiment where we limited the number of handsets (phone numbers) per speaker to one. Such a situation may arise when recruited speakers or data from a found

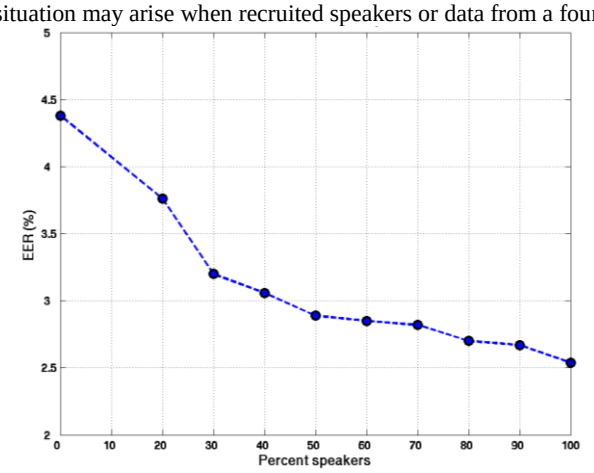

Figure 2: Performance (EER) for varying percentages of WCCN training data. Each point from 20% to 90% is the result of merging runs from 30 random draws of the data. All utterances were used for the selected speakers.

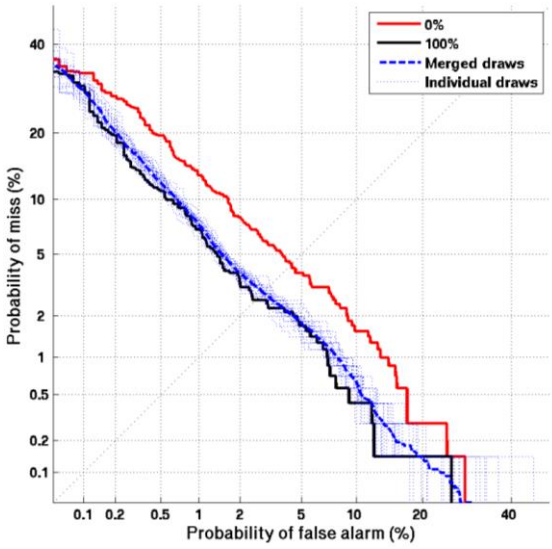

Figure 3: Performance using calls from 20 random draws of 50% of the speakers in the WCCN training set while maintaining the original call/speaker distribution.

corpus were not explicitly controlled to include multiple handsets per speaker; indeed it is most likely that speakers will use a single personal mobile phone. The resulting training list contained 4467 calls (62% of total) from 503 speakers, with same-number calls/speaker ranging from 4 to 19. As is clear from Figure 4, recognition performance is adversely affected in the absence of sufficient diversity in the training data and is considerably worse than using no within-class compensation at all.

It is clear from the previous experiment that insufficient handset variability in WCCN training has a negative effect on speaker recognition performance. To quantify the importance of phone number variability, WCCN training lists were selected from the corpus as follows:

- Choose speakers who made calls from 3 or more phone numbers.

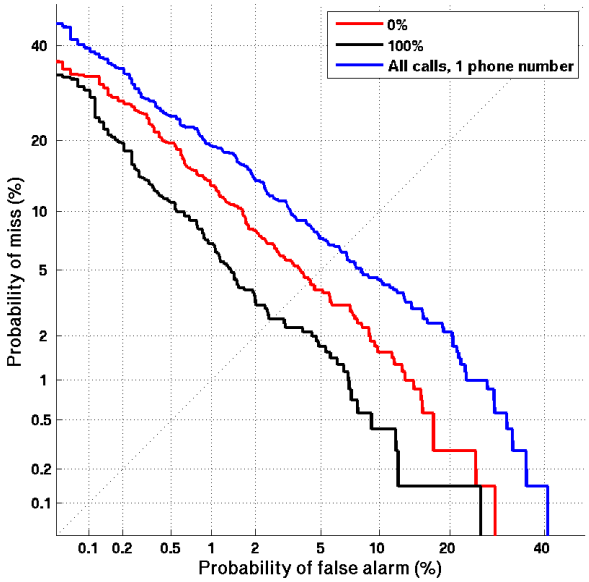

Figure 4: Performance using all calls from a single phone number for each speaker.

- Randomly select a single call from each phone number for each such speaker.
- Create 10 lists using these conditions.

Each list comprised 2122 calls (29% of the total) from 325 speakers (63% of the total). Results using merged scores from the 10 lists indicate that performance nearly equivalent to that of the 100% usage baseline can be achieved over large regions of the DET plot (Figure 5). We note that performance appears to diverge in the low false alarm region, although data in this area is very sparse. It was also found that performance improved marginally when 1-2 calls per phone number were selected (multiple calls were not always available for all phone numbers).

4. PROPOSED SELECTION CRITERIA

The results described above can be summarized as follows:

- Incorporating phone number diversity in the WCCN training data is a key factor in influencing speaker recognition performance. Results indicate that using one call from 3 or more phone numbers is sufficient to achieve baseline performance.
- Given a diverse set of phone numbers, performance is weakly related to the number of calls per phone number, with improvement decreasing rapidly beyond 1 or 2.
- All other factors being equal, a total of 3000-3500 calls appears to be sufficient to achieve baseline results.

These conclusions led to a proposed data selection rule of thumb called the "3-3-3 Rule," which posits that optimum WCCN performance requires training data from 3 or more phone numbers

per speaker, at most 3 calls per phone number, and 3000 total calls. While not called for based on the experiments of Section 3, requiring more than a single call per phone number would likely add robustness in real world applications. With precisely 3 phone numbers per speaker, the total number of speakers required would be 333.

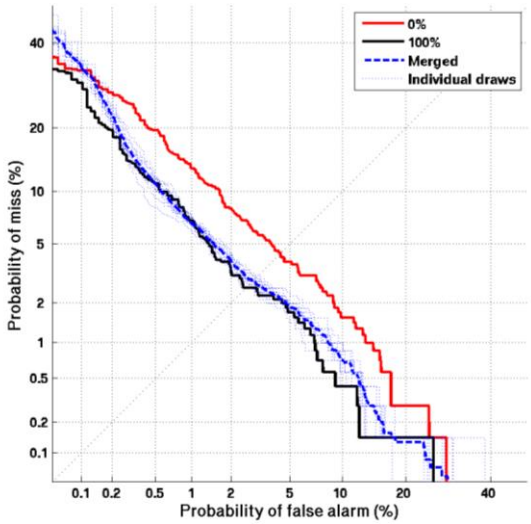

Figure 5: Performance using 10 random draws from the original WCCN training set with 1 call per phone number from speakers with 3 or more phone numbers.

Since there was insufficient WCCN training data of this type to test the guideline, the rule was adjusted to accommodate the limitations of the corpus. Thus, 5 random sets were drawn from the WCCN training set from speakers with 3 or more phone numbers and 1-3 calls per phone number, for a total of 3449 calls (48% of total) from 372 speakers (72% of total). Results, shown in Figure 6, indicate that this selection strategy does indeed achieve performance equal to that of the 100% baseline.

The results in Figure 6 were obtained from the development set. To evaluate the guidelines on a held-out set, the 3-3-3 Rule was applied to the female counterpart of the corpus, none of which was used during the development phase. Characteristics of the female speaker data are as follows:

- Training for UBM and T-matrix: Switchboard 2 and SRE04/05/06, female speakers only, minimum 8 calls/speaker, total 29978 files from 2211 speakers.
- Training for WCCN matrix: As above but SRE files only, total 9961 files from 731 speakers.
- Enroll: 894 speakers, 3 calls/speaker, 2681 files total.
- Test: 1678 calls, total 1444 target trials and 17312 nontarget trials.

Using only speakers with a minimum of 3 phone numbers, calls were drawn such that each speaker contributed 3 calls from 1 phone number and 1 call from the remaining 2 phone numbers. This selection process yielded a total of 2645 calls (27% of the total) from 529 (72%) speakers. Ten random sets corresponding to these conditions were created. Results are shown in Figure 7, again comparing DET plots for the merged random draw scores to those from the 100% and 0% runs on the female data. Results again indicate that baseline performance can be achieved using a significantly smaller amount of carefully selected WCCN training data.

5. CONCLUSIONS

Results of this study lead to the following conclusions:

- Data requirements for WCCN matrix training can be reduced by careful selection of material.
- The key factors in WCCN data selection are speaker and handset variability, with phone number information serving as a proxy for the latter in the current study. Only a few phone numbers are needed, with 1-3 calls per phone number. The total number of calls should be in the vicinity of 3000-3500. In the experiments described in this paper, the total number of speakers was 372 and 529.
- Applying within-class covariance normalization with data from many speakers but little or no phone number diversity can be worse than not applying WCCN at all.

Future work in this area should include evaluating these guidelines on additional held-out data, repeating these studies for other common normalization methods (e.g., PLDA), and investigating these factors jointly across more than one normalization method.

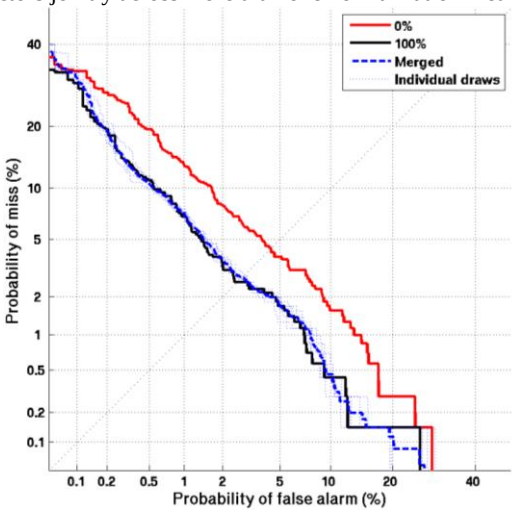

Figure 6: Results obtained by applying the 3-3-3 Rule to the WCCN training data, with modifications dictated by the available data.

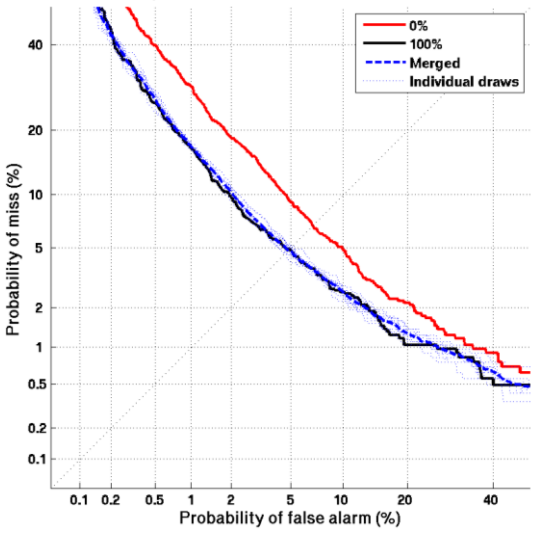

Figure 7: Results obtained by applying the 3-3-3 Rule to the held-out female data.

6. REFERENCES

[1] 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*, vol. 19, no. 4, pp. 788 - 798, May 2010.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[3] H. Hermanksy and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[4] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 113–132, 2000.

[5] Reynolds, D. A., "Channel Robust Speaker Verification via Feature Mapping," *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, pp. II: 53-56, 06-10 April 2003.

[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. IEEE Odyssey: Speaker Lang. Recognition Workshop*, Crete, Greece, 2001, pp. 213–218.

[7] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 97–100, 2006.

[8] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1471–1474, Sep. 2006.

[9] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[10] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 249-252.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.

[12] Switchboard: http://catalog.ldc.upenn.edu/LDC97S62

[13] NIST SRE: http://www.itl.nist.gov/iad/mig//tests/sre