Feature-Based and Channel-Based Analyses of Intrinsic Variability in Speaker Verification

Martin Graciarena[1], *Tobias Bocklet*[2], *Elizabeth Shriberg*[1], *Andreas Stolcke*[1], *Sachin Kajarekar*[1]

[1] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2] University of Erlangen-Nuremberg, Germany
martin@speech.sri.com

Abstract

We explore how intrinsic variations (those associated with the speaker rather than the recording environment) affect text-independent speaker verification performance. In a previous paper we introduced the SRI-FRTIV corpus and provided speaker verification results using a Gaussian mixture model (GMM) system on telephone-channel speech. In this paper we explore the use of other speaker verification systems on the telephone channel data and compare against the GMM baseline. We found the GMM system to be one of the more robust across all conditions. Systems relying on recognition hypotheses had a significant degradation in low vocal effort conditions. We also explore the use of the GMM system on several other channels. We found improved performance on table-top microphones compared to the telephone channel in furtive conditions and gradual degradations as a function of the distance from the microphone to the speaker. Therefore distant microphones further degrade the speaker verification performance due to intrinsic variability.

Index Terms: speaker recognition, vocal effort, speaking style, intrinsic variation, furtive speech, interview speech, read speech, oration

1. Introduction

An underexplored issue in the field of speaker recognition is the impact of *intrinsic variability*, or variability which comes from the test speaker him- or herself. A limited number of studies have looked at the issue of intrinsic variation and recognition by humans or by machine, e.g., [1, 2, 3, 4].

In an earlier paper [5] we studied the effects of level of effort and speaking style on speaker verification performance, using the newly collected SRI-FRTIV corpus. Results of a cepstral GMM-UBM system on an all-out pairing of train/test conditions using telephone channel data revealed that vocal effort level has a dramatic effect on results, with the largest degradations coming from conditions involving furtive (low vocal effort) speech. In this paper we explore the use of other types of speaker verification systems on the telephone channel data and compare against the GMM baseline. We also explore the use of the GMM on several other channels. We found an important impact of the microphone quality in the speaker verification performance, especially in the furtive condition. We also found that distant microphones further degrade the speaker verification performance due to intrinsic variability.

2. Data Collection

The SRI-FRTIV corpus, as detailed in [5], contains data from 30 (15 male, 15 female) native speakers of North American English. A novel aspect of the corpus is that the subjects were recruited from local “Toastmaster” clubs, to facilitate recording of an “oration” condition. A main focus of the collection was to carefully control the level of vocal effort while varying the speaking condition. Each participant was recorded at two different times, separated by an average of two to three weeks.

Each session included recordings in four different speaking styles and at three levels of vocal effort, as shown in Table 1. It was found in pilot experiments that interviews and phone conversations were highly unnatural at a high vocal effort, and that oration was unnatural at low and normal vocal efforts. Thus, those conditions were not recorded. Read speech was recorded at all three vocal effort levels. Interviews were intended to be more “serious” and more contextualized than those of the MIXER-5 collection [6]. To this end, interview topics were designed to elicit spatial descriptions. Phone conversation topics were chosen by the subject from a list that included movies, news media, holidays, and health and fitness, similar to topics in NIST data collections. For read speech (John F. Kennedy addresses) it was possible to obtain speech at all three levels of effort. Read speech also allows estimation of automatic speech recognition performance (used in the higher-level systems explored) without the need to transcribe reference data. For the oration condition, participants used two speeches already prepared as part of their Toastmaster activities; speeches were not read.

Table 1: *Eight conditions within each session in the SRI-FRTIV corpus. Each subject participated in two sessions, for a total of 16 recordings per subject. Numbers indicate the temporal order of condition within a session. “NA” indicates an unnatural condition that was not included in the collection.*

	Normal Effort	Low Effort	High Effort
Interview (5 min.)	1	2	NA
Conversation (5 min.)	3	4	NA
Reading (2.5 min.)	5	6	7
Oration (5 min.)	NA	NA	8

A large experiment room (44 by 24 feet) was used. The room was acoustically isolated from the surrounding environment, and was therefore very quiet, with a sound pressure level (SPL) measured at 39.8 dB. Five microphones were used to record the subject. The experimenter was also wearing a close-talking microphone that served additionally as a telephone-like input to the subject in the conversation condition.

A telephone channel was used to record the subject using two external phone lines (to avoid the internal telephone switch). The subject and experimenter each also wore close-talking Sennheiser HMD-410 microphones, a standard reference microphone for many speech data collections. Three Crown PZM-6D boundary microphones were fixed on a table between the subject and the experimenter, at various distances from the subject. We have assigned names to these three microphones depending on their position in the table: “PZM subject”, “PZM mid”, and “PZM interviewer”.

3. System Descriptions

Because different speaking styles and vocal effort conditions impact different features in different ways, we were interested in discovering how well different types of speaker recognition features perform under the different types of variability in the SRI-FRTIV corpus. Below we describe experiments based on four different types of systems used in the SRI submission to the NIST 2008 Speaker Recognition Evaluation (SRE08) [7]. The systems were selected from a larger list of SRI systems because together they represent a wide range of feature types.

3.1. Baseline GMM System

A Gaussian mixture model (GMM) system was used to model speaker-specific Mel frequency cepstral coefficient (MFCC) features. The system is based on the GMM-UBM model paradigm, in which a speaker model is adapted from a universal background model (UBM). Maximum a posteriori (MAP) adaptation was used to derive a speaker model from the UBM. The GMM has 2048 Gaussian components. The cepstral GMM system uses the standard telephone bandwidth (200-3300 Hz) and includes gender/handset normalization and utterance level mean and variance normalization. It also incorporates session variability normalization [8] trained on NIST SRE04 data. The UBM model was trained with a combination of Switchboard and Fisher data. In NIST SRE evaluations this is one of the best performing systems [7].

3.2. Constrained GMM System

A new, “constrained” cepstral GMM system [9] makes use of automatic syllabification of phone alignments from automatic speech recognition (ASR). The constrained system combines scores from five subsystems, each of which uses features only from frames that satisfy a specific constraint. The five constraint specifications are (1) syllable nuclei, (2) syllable onsets, (3) syllable codas, (4) the phone [N], and (5) one-syllable words. The combiner was trained with data from the SRI-FRTIV corpus. Background models are trained on SRE04 English telephone data. A 512-component GMM is used in every subsystem except constraint subsystem (5), which uses 1024 Gaussians. Eigenchannel matrices for each subsystem are trained using data from SRE04 and alternate microphone data from SRE05. This system performed extremely well in the SRE08 evaluation [7,9].

3.3. MLLR-SVM System

The MLLR-SVM systems use speaker adaptation transforms as features [10]. The MLLR (maximum likelihood linear regression) reference models use 52-dimensional perceptual linear prediction (PLP) features normalized with VTLN, truncated at 39 dimensions after using LDA+MLLT, and a speaker-adaptive CMLLR (SAT) transform. A total of 16 affine 39x40 transforms are used to map the Gaussian mean vectors from speaker-independent to speaker-dependent speech models; eight transforms each are estimated relative to male and female recognition models. The transforms are estimated using MLLR. The transform coefficients form a 24,960-dimensional feature space. The rank-normalized MLLR features are then subjected to nuisance attribute projection (NAP) estimated on SRE04 and SRE05-alternate microphone data, using 32 nuisance dimensions. The projected feature vectors are then modeled by support vector machines (SVMs) using a linear kernel. The impostor set for SVM training comes from SRE04. No score normalization was applied. This system has a competitive performance in NIST speaker recognition evaluations [7].

3.4. Word N-Gram System

This system uses the relative frequencies of word unigrams, bigrams, and trigrams extracted from the final 1-best ASR output and forms a sparse vector of these frequencies for SVM speaker modeling. The framework combines lexical speaker characterization with SVM modeling using a simple linear kernel and rank-normalization of N-gram frequencies. The impostor training set for this system was a selection of 5297 conversation sides from SRE04 and Fisher Phase 2 corpora. The feature space was given by the 126,663 most frequent N-grams from this set. No score normalization was applied. In NIST SRE evaluations this is one of the weaker systems but provides gain in combination [7].

4. Telephone Channel Experiments

For speaker verification experiments, we trained a speaker-specific model for each speaker in each of the eight conditions (task and vocal effort) described earlier, for each session, for a total of 16 different models per speaker. We then tested each speaker model on the other conditions (task by vocal effort by session combinations). In doing so we avoided cases involving the same reading material. We also avoided comparing data with mismatched gender, since these were too easily rejected by the system. The total number of impostor trials (107,520) was about 15 times greater than the number of target trials (6,840). To mimic NIST SRE conditions, and also to match our background model data length, we limited the data length for each condition to 2.5 minutes.

In Table 2 we present the equal error rate (EER) of the baseline GMM system on the telephone channel data [5]. In Tables 3, 4 and 5 we present the EER differences with respect to those in Table 2 for the constrained GMM system, the MLLR-SVM system, and the word N-Gram system, respectively. Negative differences show an improvement over the baseline GMM system, whereas positive differences reveal degradation. In each cell in Tables 3-5, a white background means the difference is not significant, a green background means a significant error reduction, and a red background means a significant error increase. The numbers shown in Tables 3-5 are absolute differences, whereas the significance are computed using relative differences.

In Table 3 we observe that the constrained GMM results behave similar to the baseline GMM in the conv/conv normal condition. It degrades mostly when testing in the low condition however this behavior is not symmetric. Some of the degradations appear in read style in normal and high vocal efforts. Further analysis will assess the degradation of each subsystem.

The MLLR-SVM results in Table 4 show comparable performance in normal and high vocal effort conditions. It shows a much higher degradation in low vocal effort conditions, probably due to the degraded word hypotheses used for adaptation. The Decipher large-vocabulary speech recognition system developed for NIST SRE2008 was used to obtain the word hypotheses. The average word error rates (WER) computed in the telephone channel, read condition (the only condition for which we had transcriptions) and low, normal, and high vocal efforts were 81.4%, 20.0%, and 16.6%. The WER on normal vocal effort is comparable to the WER obtained in NIST evaluations using telephone conversations.

Table 2: *Baseline telephone channel* ***GMM*** *system EER results. Darker (closer to blue) boxes indicate low EER values and lighter (closer to red) boxes indicate higher EER values.*

EER (%)		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	3.72	5.83	4.38	8.33	11.88	10.83	13.33	13.33
	Conv	5.83	6.67	5.18	7.50	8.33	8.33	14.02	10.83
	Read	6.82	5.83	1.67	12.56	16.79	10.00	18.04	18.27
Normal Effort	Intr	7.50	5.86	11.67	0.0	0.0	0.23	3.45	2.32
	Conv	10.00	5.83	11.67	0.0	0.0	0.83	4.17	1.58
	Read	10.80	8.45	9.61	0.0	0.83	0.0	3.33	1.58
High Effort	Read	12.50	11.67	16.93	2.50	2.50	3.42	0.0	2.50
	Orat	15.09	10.00	16.67	1.67	1.67	0.95	2.50	0.0

Table 3: *Difference between the* ***Constrained GMM*** *EER and the baseline telephone channel GMM EER. A green (light) cell indicates a significant error reduction, a red (dark) cell indicates a significant error increase and a white cell indicates difference is not significant.*

EER Diff to Baseline GMM		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	4.61	1.67	5.62	2.5	2.29	0.84	2.5	0.12
	Conv	0.66	0	3.3	1.58	2.32	4.17	2.65	2.5
	Read	4.01	5	11.37	3.15	0.71	3.33	1.96	-0.77
Normal Effort	Intr	-0.83	-0.86	0.83	1.28	0.68	0.69	0.63	1.01
	Conv	0.12	1.52	2.5	0.17	0.05	-0.15	-0.12	0.77
	Read	-0.8	-1.66	2.06	0.2	-0.6	1.67	2	0.92
High Effort	Read	1.67	3.33	1.52	1.67	1.67	3.25	1.34	0.03
	Orat	-1.76	0	-0.06	0.71	-0.75	0.72	0	0.05

Table 5 shows that the word N-Gram results are always significantly worse than the GMM for telephone data. Nevertheless, the word N-Gram results on normal vocal effort are similar to those in NIST evaluations for similar type of data. The degradation is fairly consistent across different conditions.

5. Varying Channel GMM Experiments

We want to study the impact of different channels other than the telephone given intrinsic variability conditions in the speaker verification performance. We present speaker verification experiments using the baseline GMM system on multiple microphones from the SRI-FRTIV data. We present results using matched train and test conditions, i.e., the speaker model and the test utterance come from the same microphone. The background model was trained using telephone data. We used constraints on test and train combinations similar to those for the telephone channel.

Table 4: *Difference between the* ***MLLR-SVM*** *EER and the baseline telephone channel GMM EER.*

EER Diff to Baseline GMM		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	4.40	3.40	8.95	6.67	4.79	5.96	5.93	2.50
	Conv	3.34	-3.34	11.49	2.50	2.50	7.44	0.15	5.00
	Read	6.42	10.99	15.23	12.44	8.21	16.46	8.63	9.23
Normal Effort	Intr	7.50	4.97	14.37	0.00	0.00	-0.09	3.22	-0.65
	Conv	7.50	5.87	14.31	-0.03	0.00	-0.12	-1.46	-0.75
	Read	6.70	6.55	18.57	0.00	-0.12	1.62	4.65	-0.75
High Effort	Read	5.15	2.50	9.74	3.33	0.83	3.63	0.09	-0.83
	Orat	0.74	7.38	10.83	0.06	-0.84	-0.12	-0.83	0.00

Table 5: *Difference between the* ***Word N-Gram*** *EER and the baseline telephone channel GMM EER.*

EER Diff to Baseline GMM		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	33.13	29.17		34.88	27.14			31.52
	Conv	29.14	21.84		30.15	21.88			34.38
	Read								
Normal Effort	Intr	34.17	29.17		36.29	32.45			37.68
	Conv	27.59	22.53		32.63	28.28			40.03
	Read								
High Effort	Read								
	Orat	29.26	33.39		37.50	40.06			38.28

Tables 6, 7, and 8 show EER differences with respect to the telephone channel GMM EER, for the GMM system using the following microphones: Sennheiser, PZM subject, and PZM mid, respectively.

From Table 6 we observe that the Sennheiser results are almost always significantly better than the telephone conditions. The largest differences occur in the low vocal effort conditions. It is interesting to note that the quality of the microphone makes a considerable difference when capturing low-amplitude speech signals (even though such a microphone might not be available in real world scenarios).

The results in Table 7 show that the PZM subject microphone improves over the telephone microphone in the low/low, low/normal, and normal/low conditions. The main reason for this difference may be a telephone channel “gating effect” that blocks out very-low amplitude signals. This was confirmed in listening experiments. We additionally observed that the PZM subject is worse than the telephone channel in normal and high conditions. However in real-world conditions we expect the PZM microphone to perform worse than the telephone channel because it is more sensitive to environmental conditions.

In Table 8 we see that the PZM mid results are almost always significantly worse than the telephone results. The main differences occur in the mismatched conditions low/normal and low/high. This reveals that using a far-field microphone, such as the PZM mid, will further increase the effect of mismatched vocal effort and style.

Table 6: *EER Difference between the* ***Sennheiser*** *based GMM and the telephone channel GMM.*

EER Diff to GMM Telephone		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	-3.72	-5.63	-4.3	-5.83	-7.86	-7.5	-5	-7.5
	Conv	-5.83	-6.47	-4.5	-5.83	-5.77	-4.34	-7.2	-6.45
	Read	-6.62	-4.34	-1.62	-6.73	-9.47	-6.67	-8.04	-9.1
Normal Effort	Intr	-5.98	-5.03	-9.02	0	0	-0.23	-0.86	-1.31
	Conv	-7.59	-4.88	-6.7	0	0	-0.83	-1.76	-0.75
	Read	-7.47	-4.94	-6.28	0.14	-0.03	0	0	-0.75
High Effort	Read	-5.62	-6.61	-10	-1.01	-1.76	-1.69	0.08	-0.83
	Orat	-9.26	-5.95	-11.5	-0.99	-1.44	-0.87	-0.83	0

Table 7: *EER Difference between the* ***PZM subject*** *based GMM and the telephone channel GMM.*

EER Diff to GMM Telephone		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	-2.05	-2.50	-1.02	-3.45	-5.21	-3.33	3.34	2.38
	Conv	-3.27	-3.34	-2.00	-0.83	-1.63	-2.50	3.42	1.67
	Read	-2.65	-1.75	1.66	-2.56	-2.62	-1.76	0.26	-3.92
Normal Effort	Intr	-3.21	-1.69	-3.37	0.03	0.78	0.66	2.23	2.68
	Conv	-3.27	0.84	-0.84	0.78	0.00	0.84	1.66	0.98
	Read	-3.30	-2.62	-1.54	0.75	0.84	1.62	0.24	0.92
High Effort	Read	4.17	2.50	0.00	1.67	1.61	1.58	0.33	-0.83
	Orat	0.74	0.83	-2.50	3.33	0.21	1.55	-0.83	0.24

6. Conclusions

We have explored the use of several speaker verification systems on the telephone channel data from the recently collected SRI-FRTV intrinsic variations database. We found the GMM system to be one of the more robust across all conditions. Constrained GMM system produced some gains over the baseline GMM however those gains were not consistent when swapping train and test conditions. Degraded recognition performance in furtive (low vocal effort) conditions severely affected the MLLR-SVM and word N-Gram systems, therefore work needs to be done to improve ASR performance in furtive conditions. Additionally speech segmentation needs to be improved in furtive conditions. One hope is that a combination of these systems may result in a gain over the GMM system. However this combination may benefit from conditioning on the vocal effort and/or style. We also explored the use of the GMM systems on several microphones other than the telephone channel. We first found that the quality of the microphone significantly impacts performance in furtive conditions. We also found improved performance on table-top microphones compared to the telephone channel in furtive conditions and a gradual degradation as the distance between the microphone and the speaker increases. Channel compensation techniques should be used to assess the improvements in all conditions.

7. Acknowledgements

This work was supported in part by contracts NMA401-02-9-2001 and by NSF IIS-0544682. The views are those of the authors and do not represent the views of the funding agencies. We thank the International Computer Science Institute (ICSI) for making the PZM microphones available. We want to thank Fred Goodman for his valuable input to this project.

Table 8: *EER Difference between the* ***PZM mid*** *based GMM and the telephone channel GMM*

EER Diff to GMM Telephone		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Intr	Conv	Read	Intr	Conv	Read	Read	Orat
Low Effort	Intr	8.36	10.00	12.94	8.19	9.79	11.61	14.17	13.31
	Conv	9.17	6.66	10.65	5.71	5.84	8.34	10.15	9.17
	Read	13.03	11.67	16.66	11.61	9.88	10.00	13.63	11.02
Normal Effort	Intr	8.12	5.81	5.83	0.36	1.44	3.91	5.72	4.35
	Conv	10.00	6.73	11.54	1.59	1.56	2.50	2.50	3.42
	Read	11.55	6.64	10.39	3.25	3.34	3.28	1.67	1.84
High Effort	Read	10.83	10.00	8.07	7.50	3.33	3.13	1.41	-0.03
	Orat	11.58	8.33	7.50	4.16	1.66	1.76	-0.83	1.62

8. References

[1] I. Shahin, “Enhancing Speaker Identification Performance under the Shouted Talking Condition using Second-Order Circular Hidden Markov Models”, *Speech Communication*, vol. 48, pp. 1047–1055, 2006.

[2] W. Wu, T. F. Zheng, M. Xu, and H. Bao, “Study on Speaker Verification on Emotional Speech”, *Proc. ICSLP*, pp. 2102–2105, Pittsburgh, PA, 2006.

[3] D. Brungart, K. Scott, and B. Simpson, “The Influence of Vocal Effort on Human Speaker Identification”, in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. Eurospeech*, pp. 747–750, Aalborg, Denmark, 2001.

[4] I. Karlsson, T. Banziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, and K. Scherer, “Speaker Verification with Elicited Speaking Styles in the Verivox Project”, *Speech Communication*, vol. 31, pp. 121–129, 2000.

[5] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey and F. Goodman, “Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification”, *Proc. Interspeech*, pp. 609–612, Brisbane, Australia, 2008.

[6] C. Cieri, L. Corson, D. Graff, and K. Walker, “Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora”, *Proc. Interspeech*, pp. 950–954, Antwerp, 2007.

[7] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, “The SRI NIST 2008 Speaker Recognition Evaluation System”, *Proc. ICASSP*, Taipei, Taiwan, 2009.

[8] R. Vogt and S. Sridharan, “Explicit Modeling of Session Variability for Speaker Verification”, *Computer Speech and Language,* vol. 22, pp. 17–38, Jan. 2008.

[9] T. Bocklet and E. Shriberg, “Speaker Recognition Using Syllable-Based Constraints for Cepstral Frame Selection”, *Proc. ICASSP*, Taipei, Taiwan, 2009.

[10] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, “MLLR Transforms as Features in Speaker Recognition,” *Proc. Eurospeech*, pp. 2425-2428, Lisbon, 2005.