Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification

Elizabeth Shriberg[1], *Martin Graciarena*[1], *Harry Bratt*[1], *Andreas Kathol*[1],
Sachin Kajarekar[1], *Huda Jameel*[1], *Colleen Richey*[1], *Fred Goodman*[2]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2]MITRE Corporation, Signal Processing Center, McLean, VA, USA
ees@speech.sri.com

Abstract

We study the question of how intrinsic variations (associated with the speaker rather than the recording environment) affect text-independent speaker verification performance. Experiments using the SRI-FRTIV corpus, which systematically varies both vocal effort and speaking style, reveal that (1) "furtive" speech poses a significant challenge; (2) conversations and interviews, despite stylistic differences, are well matched; (3) high-effort oration, in contrast to high-effort read speech, shares characteristics with conversational and interview styles; and (4) train/test pairings are generally symmetrical. Implications for further work in the area are discussed.

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

1. Introduction

The bulk of the effort in the speaker recognition research community, driven by NIST speaker recognition evaluations (SREs), has been on coping with extrinsic variation. These are factors outside of speech production, including choice of microphone, distance from microphone, room acoustics, background noise, and transmission channel. But a second, major source of variability in real-life spoken data is intrinsic variation, i.e., variation that comes from the talker. Because such variability increases mismatch between train and test samples from the same talker, it can be expected to cause degradations in real-world applications.

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, 5]. In general it is found that it is better to train on a variety of styles to improve performance, or to train a system not to expect speech in the same style as used for background data. A large data collection effort for the speaker verification community is under way at the LDC [6], but the data has not yet been studied in detail with respect to intrinsic variation.

We describe a new, controlled study on speaker verification error rates based on the SRI-FRTIV (Five-way Recorded Toastmaster Intrinsic Variation) corpus. The corpus contains speech from subjects who participated in a variety of elicited conditions that crossed level of vocal effort with speaking style. Even in comparison with the larger corpus in [6], the present study has some unique properties. First, we were interested in "furtive" or very low-effort (but not whispered) speech. Second, we sought to study raised (or high) vocal effort associated with projection over a distance, rather than over noise (as in the Lombard effect). Third, we were interested in an "oration" condition, involving a speech intended to inform or influence listeners. Finally, in order to focus on intrinsic variation, we held constant the microphone and channel (in this case a telephone recording), as well as the subject's position in the room, across all conditions.

Table 1: *Eight conditions within each session in the SRI-FRTIV corpus. Each subject participated in 2 sessions (for a total of 16 recordings per subject). Numbers indicate the temporal order of conditions. "NA" indicates an unnatural condition.*

	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

1.1. Speech data

For the FRTIV corpus we collected data from 30 (15 male, 15 female) native speakers of North American English. Informal experimentation indicated that while most speakers could produce and maintain a furtive level of effort, it was difficult for subjects to speak at a high level of effort. Since we were particularly interested in speakers with the ability to speak at various levels of vocal effort we recruited participants with experience in public speaking. Local "Toastmaster" clubs, in which individuals meet to practice public speaking, provided the perfect opportunity. Toastmasters had available prepared speeches that could be used for the oration condition directly. Each participant was recorded at two different times, separated by an average of two to three weeks.

1.2. Data conditions

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. The order of the resulting eight conditions was chosen to obtain a natural progression of experimenter involvement from strongest to weakest. That is, while the experimenter initially prompted the participant to answer concrete questions in condition 1, he was merely a passive audience member in condition 8. Within each style, normal vocal effort (when present) preceded other vocal effort levels.

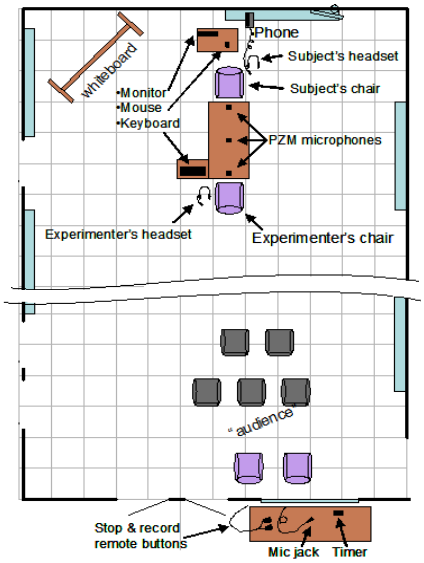

Figure 1: *Recording room (44 by 24 feet). The subject is positioned in the same location for all recordings.*

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, for example:

> **Interviewer**: *Describe what you see when you walk in from the front door of your home.*
> **Subject**: *You walk in with bay windows on both sides. Then there's a doorway into the den . . .*

The interviewer in addition maintained a rapid pace in following the subjects' responses with further questions. To maintain social distance between interviewer and subject, the interview sessions preceded the casual conversations. To further differentiate the two styles, subjects were asked to initiate the phone conversations. Phone conversation topics were chosen by the subject from a list that included movies, news media, holidays, and health and fitness, in a flavor similar to topics in NIST data collections. For example:

> **Subject**: *Hey did you see the news last night?*
> **Experimenter**: *Uh, no, I don't watch T.V.*
> **Subject**: *No? They were talking about the effect of the media on politics ...*

Read speech was included because it was possible to obtain at all three levels of effort, and because it also provides an estimate of automatic word recognition accuracy, for use in later experiments. Subjects selected one of a total of six speeches consisting of excerpts from John F. Kennedy's addresses. The same speech was read in all three speaking styles. For the oration, participants used two speeches already prepared as part of their Toastmaster exercises.

In addition to the experimenter, a human "monitor" was present to signal to the participants (visually) if their effort level drifted away from the target level. The monitor also joined the experimenter in playing the role of audience member for the participants' oration recordings. The second session was conducted in exactly the same manner but with an entirely different content. In addition to giving us more raw data, the two-session collection protocol allows for studies of intersession variability, a major issue in speaker recognition work.

1.3. Recording setup

An unusually large experiment room (44 by 24 feet; see Figure 1) 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 — lower than a quiet office. The ceiling and walls were acoustically treated, resulting in very low reverberation. Five microphones were used to record the subject. The experimenter also wore a close-talking microphone that served additionally as a telephone-like input to the subject in the conversation condition (see below). While we describe results only for telephone recordings in this study, simultaneous recordings on all microphones, including the telephone channel, were made for all conditions. As noted in the introduction, this is an advantage of the FRTIV corpus, which will be made available to the community in the future. Thus for future reference, information on the full set of microphones in the corpus is provided below.

A telephone channel was used to record the subject using two external ATT phone lines (to avoid the internal PBX). The receiving line connected to a Comrex DH-20 digital telephone hybrid, which converted the audio to line level. The telephone sending line used a Plantronics P141N headset attached to a head-mounted boom (the headphone was not used). For the telephone condition, microphone inputs were sent "dry" to the recorder, but the experimenter's version was band-pass filtered to the subject's headphones, to simulate telephone sound. Note that only the subject's speech, which was a true telephone recording, was used in verification experiments. The subject and experimenter each also wore a close-talking Sennheiser HMD-410 microphone, a standard reference microphone for many DARPA funded projects. Three Crown PZM-6D boundary microphones were fixed on the table between the subject and the experimenter, at various distances from the subject (see Figure 1.) Signals were synchronized and digitized up front. All six analog microphone signals were digitized by a Yamaha O1V digital mixer. The mixer synchronized signals with a 48 kHz clock on the PC's audio card, and digitized data was saved on the PC. The mixer and PCI card communicated 8-channel digital data via ADAT lightpipe.

2. Speaker Verification Experiments

A Gaussian mixture model (GMM) system was used to model speaker-specific Mel cepstral (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) [7]. Maximum a posteriori (MAP) adaptation was used to derive a speaker model from the UBM. The GMM has 2048 Gaussian components, and is described in detail in [8]. 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 [9] trained on NIST SRE04 data. The UBM model was trained with a combination of Switchboard and Fisher data.

We trained a speaker-specific GMM for each speaker in each of the 8 conditions (task by vocal effort) described earlier, and 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 any "same words" read conversations, i.e., the conversations that were read from the same reading material within the same session. We also avoided conversations with mismatched gender, since these were too easy for the sys-

EER (%)		TRAIN ON							
		Low Effort			Normal Effort			High Effort	
TEST ON		Inter.	Conv.	Read	Inter.	Conv.	Read	Read	Orat
Low Effort	Inter.	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	Inter.	7.50	5.86	11.67	0.00	0.00	0.23	3.45	2.32
	Conv.	10.00	5.83	11.67	0.08	0.00	0.83	4.17	1.58
	Read	10.80	8.45	9.61	0.08	0.83	0.00	3.33	1.58
High Effort	Read	12.50	11.67	16.93	2.50	2.50	3.42	0.00	2.50
	Orat.	15.09	10.00	16.67	1.67	1.67	0.95	2.50	0.00

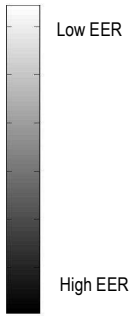

Figure 2: *Speaker verification results by train/test condition. Higher EERs are indicated by darker shading. Low, Normal, High refer to vocal effort level. Speaking styles are interview (Inter), conversation (Conv), reading (Read), and oration (Orat).*

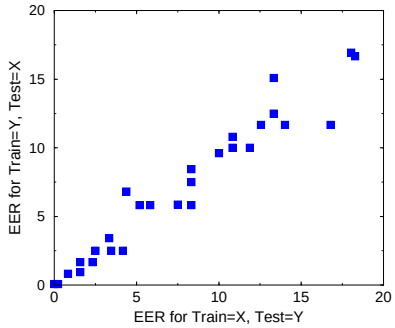

Figure 3: *EERs for mismatched conditions plotted against the corresponding EER for the experiment in which train and test data are reversed, indicating rough symmetry (with a few exceptions).*

tem. The total number of impostor trials (107,520) was about 15 times greater than the number of target trials (6,840). Even though we collected the data using many simultaneous channels, we report here only on the telephone-channel recordings. To mimic NIST SRE conditions, and also to match our background model data, we limited the data length for each condition to 2.5 minutes.

3. Results and Discussion

Figure 2 shows speaker verification results in equal error rate (EER) for all train/test combinations. To aid in visualization, EERs are shaded according to their value; the higher the EER, the darker the shading.

3.1. General observations

A useful first point to draw from Figure 2 is that the matrix is roughly symmetrical. This can be better seen in Figure 3, in which results are plotted for only the nondiagonal (i.e. mismatched) conditions. With a few exceptions, the EER from training on one condition and testing on the other is similar to that when the data sets are reversed, suggesting that it is the degree of mismatch, rather than inherent properties of train and test data, that contributes to error rates. We can thus simplify discussion by referring to condition pairs representing both orderings of train and test sets. For example, "normal/low" will refer to both "train on normal, test on low" and "train on low, test on normal".

A second general observation is that, as we had hoped, "baseline" rates are very low (even 0.00%). The overall baseline for the experiment can be considered the normal-conv/normal-conv experiment, which corresponds most closely to the task and level of effort in NIST SRE data. Baselines for other conditions are the matched cases, along the diagonal. As shown, all but the furtive matched conditions result in EERs of 0.00%. The low rates are useful for our purposes in that our interest is in degradations from these baselines as a function of the various experiment conditions. It is likely that the low rates are made feasible by the quiet room, high-quality recordings, lack of channel variation, and subjects (native English speakers, attentive to task at hand). Overall this means that the degradations seen here are, if anything, an underestimate of what one might expect in real-world scenarios.

Baseline EERs for furtive matched conditions are lowest for read speech and highest for conversations, suggesting that interview speech lies somewhere in between the two in terms of a subject's internal consistency in matched furtive conditions across different recordings.

3.2. Effect of vocal effort

Clearly, the largest effect on EER comes from our vocal effort manipulation, and in particular from experiments that involve furtive speech (in training, testing, or both). This is visible from the diagonal, as noted above, in which matched cases in low-effort speech fare worse than matched cases in normal- or high-effort speech. The degradation is, as might be expected, even more severe for experiments involving mismatch. The highest error rates occur for experiments involving high- and low-effort conditions. In addition to the greater degree of mismatch in effort itself, another likely explanation for the large degradation is that the background model is trained with normal vocal effort data. The high vocal effort data is closer to the normal condition than to the low vocal effort, which results in many true speakers being classified as impostors. We verified this by plotting the true speaker and impostor score distributions. As the vocal effort mismatch increases, both score distributions shift and the overlap increases. The degree of overlap, however, is not sym-

metrical. The overlap of the true speaker distribution within the impostor distriution is larger than the overlap of the impostor distribution within the true speaker distribution.

We were impressed by how furtively our subjects spoke, while avoiding whispering and while maintaining full engagement with the experimenter in the relevant conditions. The combination of a large and quiet room, maintenance of furtive level by the experimenter's own voice, and the human monitoring apparently worked quite well in producing human-intelligible recordings that are extremely challenging for an automatic system. Based on analyses of the low-effort speech, a key issue in improving performance of our system will be to modify the speech/nonspeech segmenter. The speech/nonspeech segmenter uses acoustic models trained with a large corpus of conversational data to extract regions of speech activity. Our system uses the segmenter as a preprocess, and it was not tuned for such low-energy speech. More generally for the design of systems that can handle intrinsic variation, the issue of speech detection needs to be addressed.

3.3. Effect of speaking style

One surprise in Figure 2 is the similarity of the interview and conversational conditions. As noted earlier, we had attempted to create interviews that had a serious tone and task, through the behavior of the experimenter and the use of spatial descriptions. Nevertheless, results in Figure 2, particularly for the normal effort level, suggest that the interview and conversation data are in effect interchangeable in that mode (each produces no errors when paired with the other). Figure 2 contains three additional 2-by-2 matrices comparing these two styles, namely for train-low/test-low, train-low/test-normal, and train-normal/test-low. Taken together, there seems to be no clear pattern to the order of results within the 2-by-2s, suggesting that differences may be random variation and that the two conditions are simply close in style. Further research will investigate this possibility.

Further away in style from conversational and interview speech (than those two are from each other) is read speech, as shown in Figure 2. Overall, the mismatched conditions involving read and conversational speech show higher error rates than do the mismatched conditions involving read and interview speech. This suggests that interviews are closer than are conversations to a read speaking style.

A final interesting finding is that while high-effort read speech is similar to oration (always high effort here), the two are far from identical. As seen in the bottom right quadrant of Figure 2, each matches itself better than the other. Furthermore, a read style, while highly useful for identifying a speaker for when level of effort is held constant between train and test, is not so consistent when the level of effort is varied. As shown in the three rows of cells directly above (and to the left of) the bottom quadrant: oration is consistently better matched (in both train-test and test-train) to all three styles (interview, conversation, and even read) at a normal level than is read speech at a high level. Thus, speeches meant for a third party, or at least those collected here from the Toastmaster subjects, probably contain a broader range of the stylistic characteristics representing a speaker than does read speech. Putting it another way: the speaker characteristics that come out in read speech depend on level of effort. While this is also true for other styles, some of the characteristics of a speaker's conversational and interview speech also seem to be present in their oration despite level-of-effort differences.

4. Conclusions and Future Directions

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 revealed that vocal effort level has a dramatic effect on results, with largest degradations coming from conditions involving furtive speech. Analysis suggests that it will be important to modify speech-nonspeech segmentation in real applications. Speaking style experiments showed similarities between conversational and interview speech, despite attempts to differentiate the two tasks. An oration condition was somewhat similar to read speech at a high level of effort, but better matched to normal-effort conversation, interview, and read speech than was read speech at a high effort.

Intrinsic variation remains an important challenge for speaker recognition systems, and efforts to understand how best to increase robustness of systems to such factors should become a priority as further data resources for such studies become available. In future work we plan to study results for different speaker recognition systems (including those that use higher-level features), analyze results for far-field microphones, and investigate adaptation and compensation approaches for coping with mismatched conditions.

5. Acknowledgments

This work was supported by NGA contract NMA401-02-9-2001 and by NSF IIS-0544682. The views are those of the authors and do not reflect those of the funding agency. We thank ICSI for loaning us the PZM microphones.

6. 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", *in Proc. ICSLP*, pp. 2102–2105, Pittsburgh, PA, Sep. 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, Sep. 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] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification", *Speech Communication*, vol. 31, pp. 255–264, 2000.

[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", *in Proc. Interspeech*, pp. 950–954, Antwerp, Aug. 2007.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 181–202, 2000.

[8] S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST speaker recognition evaluation system", *in Proc. ICASSP*, vol. 1, pp. 173–176, Philadelphia, Mar. 2005.

[9] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification", *Comp. Speech Lang.*, vol. 22, pp. 17–38, Jan. 2008.