INTERSPEECH 2009 BRIGHTON

# Does Session Variability Compensation in Speaker Recognition Model Intrinsic Variation Under Mismatched Conditions?

*Elizabeth Shriberg, Sachin Kajarekar, Nicolas Scheffer*

SRI International, Menlo Park, CA, USA

{ees,sachin,scheffer}@speech.sri.com

## Abstract

Intersession variability (ISV) compensation in speaker recognition is well studied with respect to *extrinsic* variation, but little is known about its ability to model *intrinsic* variation. We find that ISV compensation is remarkably successful on a corpus of intrinsic variation that is highly controlled for channel (a dominant component of ISV). The results are particularly surprising because the ISV training data come from a different corpus than do speaker train and test data. We further find that relative improvements are (1) inversely related to uncompensated performance, (2) reduced more by vocal effort train/test mismatch than by speaking style mismatch, and (3) reduced additionally for mismatches in both style and level. Results demonstrate that intersession variability compensation does model intrinsic variation, and suggest that mismatched data may be more useful than previously expected for modeling certain types of within-speaker variability in speech.

**Index Terms:** speaker recognition, channel compensation, intersession variability compensation, intrinsic variation, speaking style, vocal effort.

## 1. Introduction

Significant progress in the field of automatic speaker recognition has been made by addressing the effects of ***extrinsic variability***, or variability associated with factors outside the speaker, such as channel, handset, and environmental noise. The variability has been addressed by a variety of techniques, including feature transformation [1], eigenchannel compensation [2], and more recently joint factor analysis (JFA) [3]. These techniques have led to large performance improvements by modeling variability associated with the recording session. As the names suggest, the initial model was dominated by across channel (or handset) variability. Later studies broadened the scope to intersession variability (ISV). The techniques are often referred to loosely as "intersession variability" or "channel" compensation techniques. Within joint factor analysis, the corresponding factor is commonly referred to as the "channel" factor.

In spite of this terminology, it is not clear what type of variability is being modeled in practice. Since these approaches do not require data labeled according to specific variables in training or testing, they are capable of handling, in principle, any type of between-session variation. In other words, the variation being modeled is implicit in the data. Nevertheless, it is the case that recent NIST speaker recognition evaluations (SREs) have focused on channel variability by virtue of the data chosen (e.g., by varying the microphone type and placement in evaluation data).

In this paper we ask to what extent ISV compensation may be modeling ***intrinsic variability***, i.e., variability associated with the speaker rather than with the channel or acoustic environment. Forms of intrinsic variation include the speaking style, emotion, level of vocal effort, cognitive state, state of health, and combinations of these and other factors. From a theoretical perspective, the question is interesting because acoustically, external variation and human-based variation have different characteristics and constraints. Furthermore, relatively little is currently known about intrinsic variation and speaker recognition. It is also a question with practical relevance, because of the large benefit usually obtained from session variability compensation. If a speaker known only from one intrinsic speaking "mode" appears in another mode in testing, to what effect can we expect to get a benefit from session variability compensation trained only on the known mode?

We ask this question using a small but highly controlled corpus of intrinsic variation, the SRI-FRTIV (Five-way Recorded Toastmaster Intrinsic Variation) corpus, in which speakers varied both level of vocal effort and speaking style. Previous work [4] that did not examine the effect of session variability compensation showed that intrinsic variation results in various degrees of degradation relative to a NIST SRE-like baseline (i.e. relative to conversational speech at a normal effort level, in both training and testing, and matched channel). In this study we examine the relative error reduction associated with session variability compensation.

Because we did not have matched data to train the session variability compensation, we expected minimal benefit. Results, however, show surprisingly good error reductions on intrinsic variation data. We report these results and further ask how the relative performance improvement from session compensation is related to (1) baseline performance for a particular train/test condition, (2) mismatch in vocal effort between train and test, (3) mismatch in speaking style between train and test, and (4) mismatch in both vocal effort and speaking style.

## 2. Intrinsic Variation Corpus

The SRI-FRTIV corpus [4] is a highly controlled corpus designed to support the study of intrinsic variation in speech. The corpus varies level of vocal effort (normal, low, high), and speaking context or "style" (conversational, interview, read, oration), as shown in Figure 1. Session variability can be studied because each subject participated in the eight conditions on two different occasions, on average two to three weeks apart. Although the corpus is small (30 speakers—15 male, 15 female), it is highly controlled, and thereby provides a unique opportunity for studying session variability compensation for intrinsic variation. Further details on the corpus are given in [4]. Two important characteristics of the

corpus, however, deserve mention. First, background noise and channel variability across different sessions are minimal; thus results from session compensation can be largely attributed to intrinsic rather than extrinsic variation. Second, the intrinsic variation was elicited under carefully designed and monitored conditions, to ensure that subjects actually maintained the specified speaking style and level of vocal effort over each condition.

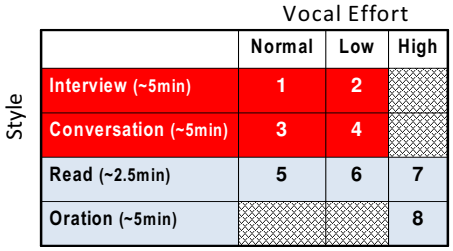


Figure 1. Within-subject vocal effort level and speaking style conditions in the SRI-FRTIV corpus. Numbers indicate collection order within a session; 1-4 are dialogs; 5-8 are monologs. Hatched cells are unnatural conditions and were not collected. Dialog duration is total time, which was divided over the two talkers.

Subjects were recorded in a large (44x24 foot), acoustically isolated room 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. Extrinsic session variability was also minimized because across conditions and sessions, the same microphones, microphone placements on subject and in room, subject placement in room, room and physical setup, experimenter (and interlocutor for the interview and conversation conditions), and calibration procedures were used.

The normal level of effort was that obtained with no special effort modification instructions to subjects. Low vocal effort (or "furtive" speech) was induced by telling subjects not to whisper, and to be loud enough for the interlocutor across the table to hear, but quiet enough that a human monitor present in the room 36 feet away from the subject could not make out what was said. High vocal effort was designed to capture level increases based on communicating over a distance (speaking to hearers on the other side of the very large room) rather than over background noise. In the interviews and conversations, the same experimenter (for all subjects and all conditions) acted as interlocutor; he performed interviews before conversations (because the reverse order would have made interviews too casual). Subject and experimenter sat across a table for the interview. For the conversation, the subject stayed in the same location and the experimenter went to a different room. In the read mode, the subject read a prepared text (texts were varied across sessions). The oration condition was included to represent the style of a speech of personal importance to the talker, designed to have an effect on others. The natural mode for the speeches was a high level of vocal effort. The corpus subjects, all from local Toastmasters clubs, performed this condition using their own prepared speeches. The topics in all four speaking style conditions were varied within subject from the first to the second session.

For each condition, the subject was simultaneously recorded over two close-talking microphones (a Sennheiser channel and a telephone channel) and three far-field microphones. The experiments reported in the present study use the telephone channel, which is a true telephone recording but on a fixed line and using a fixed and consistently placed "handset". Two external ATT phone lines were used. 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). The subject wore this microphone for all conditions (interview, conversation, read, oration, at all levels of effort). Since the line and the handset were fixed, the channel effects were minimal.

# 3. Experimental Setup

## 1.1. Task

We created a NIST SRE-like task from the FRTIV corpus. We trained a model from each recording of the subject. We tested it against all the other recordings from the same subject to create target trials, and against all recordings from other subjects of the same gender to create impostor trials, with one exception: we omitted cases in which a trial would have involved the same read text (assuming they would have been too "easy"). Scores were split according to the train and test mode, and the equal error rate (EER) was computed for each resulting condition. We trained on the subject's first visit and tested on the second, and vice versa, and averaged the results. Impostor test trials were drawn from the same speaking mode as target test trials. The number of trials per condition was 1740 (60 target, 1680 impostor) for read conditions and 1800 (120 target, 1680 impostor) for all other conditions.

## 1.2. Speaker recognition system

Speaker recognition experiments were performed using the JFA paradigm [3]. More specifically, ISV compensation (ISVC) used a particular case of this framework called eigenchannel [5-7]. The JFA framework uses the distribution of an underlying Gaussian mixture model (GMM), the universal background model (UBM) of mean $m_0$ and diagonal covariance $\Sigma_0$. Let the number of Gaussians of this model be $N$ and the feature dimension in each Gaussian be $F$. A supervector is a vector of the concatenation of the means of a GMM: its dimension is $N * F$.

In JFA [8], the basic assumption is that a speaker supervector m can be decomposed into a sum of two supervector compononents: the speaker supervector $s$ and the nuisance (or channel) supervector $c$.

$$m = s + c$$

In the eigenchannel framework, the speaker supervector is obtained by adapting the means of the UBM using a standard maximum a-posteriori (MAP) adaptation [9]. It can be expressed as

$$s = m_0 + Dz,$$

where D is well defined as $D^2 = \Sigma_0 / \tau$; $\tau$ is the regulation factor that controls the prior distribution for MAP adaptation. The nuisance supervector distribution lies in a low-dimensional subspace of rank $R$, and is assumed to be distributed according to $c = Ux$. The matrix $U$, the eigenchannels (or channel loadings), has a dimension of $NF * R_c$. The loadings $U$ are estimated from a

sufficiently large data set while the latent variables $x$, $z$ are estimated for each utterance.

Our baseline system employs gender-independent 512-Gaussian UBMs. Cepstral features are mel frequency cepstral coefficients (MFCCs) composed of 13 cepstra and energy, adding derivatives of first, second, and third order (for a total dimension of 52). The rank of the channel space is 100. To train the matrices, several iterations of the expectation maximization (EM) algorithm of the factor analysis framework are used. An alternative minimum divergence estimation (MDE) is used at the second iteration to scale the latent variables to a $N(0,1)$ distribution. To train a speaker model, the posteriors of $x$ and $z$ are computed using a single iteration (via the Gauss-Seidel method as in [10]).

For the baseline system, the verification score for each trial was a scalar product between the speaker model mean offset and the channel-compensated first-order Baum-Welch statistics centered on the UBM. This scalar product was found to be simple yet very effective [11, 12] and was subsequently adopted by the community. The speaker verification system is gender independent with a gender dependent score normalization (ZT norm). The channel loadings were trained with 2004 NIST SRE data, using 301 speakers and about 4500 sessions. Gender dependent score normalization was performed with 2004 and 2005 NIST SRE data. Performance is reported as percentage EER.

## 4. Results

We first examine results for the eight matched conditions. In Figure 2, total bar height indicates performance before ISV compensation; the height of the red (dark) bar indicates performance after ISV compensation (implemented as Factor analysis (FA) with only channel factors (U)). "Cn", or the conversational style at a normal level of vocal effort, can be viewed as a rough point of comparison with the speaking mode in many NIST evaluation conditions.

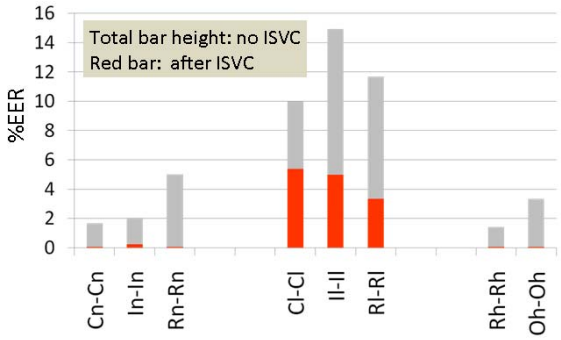


Figure 2. Effect of speaking style (**C**onversation, **I**nterview, **R**ead, **O**ration) and vocal effort level (**n**ormal, **l**ow, **h**igh) on results before and after ISV compensation (ISVC). Conditions are [train style][train level]-[test style][test level].

A first observation from the plot is that low vocal effort is a challenging condition. Further discussion is found in [4], including explanations based on the reduced rate of speech frames found in speech/silence segmentation for such very low level speech. What is new in these results is the degree to which ISV compensation reduces error rates. For all but the low effort conditions, compensation eliminates nearly 100% of the errors. And although the reduction is less dramatic for the furtive speech, it is still remarkably good. Compared with the effect of vocal effort, the effect of style variation on relative error reduction is relatively small.

One possible interpretation of these results is that they reflect qualitative differences in what is being modeled by the compensation approach. Another hypothesis is that the degree of error reduction from ISV compensation is roughly predicted from baseline (uncompensated) performance. The latter possibility makes sense in the case of very poor baseline results, since in such a case, estimation of the factored model should suffer. Whether it should hold for better-performing experiments is not clear. We plotted results for both matched (same as in Figure 2) and mismatched train/test conditions, by baseline EER. Results are shown in Figure 3.

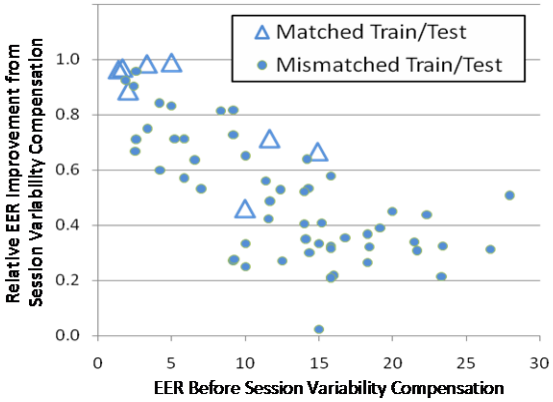


Figure 3. Relationship between baseline EER and relative improvement from ISV compensation, for both matched and mismatched train/test experiments.

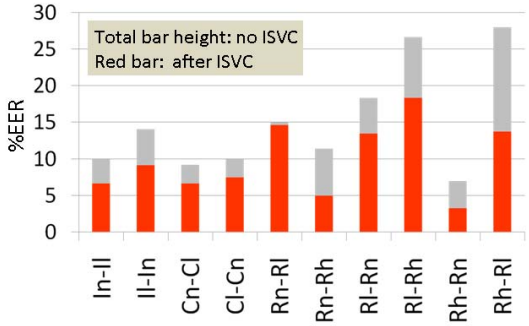


Figure 4. Effect of mismatched vocal effort level in train and test, by speaking style. See Figure 2 for abbreviations for conditions.

As shown, there is a general inverse relationship between baseline error rates and the relative improvement from ISV compensation. However, the relationship is not perfect and there are some clear exceptions. Overall, Figure 3 suggests that the degree of improvement from session compensation for intrinsic variation data depends on both quantitative (baseline performance) and qualitative (which speaking modes are involved) factors.

When train and test samples are mismatched in level, results look appreciably different from those for the matched samples shown in Figure 2. Figure 4 shows that not only do baseline

error rates increase significantly compared with those in Figure 2 (note the different scales), but also that ISV compensation is less successful. When furtive speech is involved in either train or test, the benefit is generally limited.

An interesting observation is that read speech shows almost no gain from ISV compensation when training on normal effort and testing on low effort. Read speech in general shows higher error rates. A possible explanation is that read speech is under-represented in the NIST data used to train variability compensation.

Figure 5 shows performance before and after compensation for the case of speaking style mismatch. As can be discerned, style mismatch is mainly a problem in the furtive speaking mode. At a normal speaking level, the conversational and interview modes are nearly interchangeable in terms of effect on compensation performance. Read speech, however, behaves differently. Although baseline error rates for style mismatches involving read speech are similar to those involving conversational and interview speech, ISV compensation is less effective for mismatches involving read speech. A possible reason, as suggested earlier, is that read speech is not well represented in the NIST data used to train the ISV model.

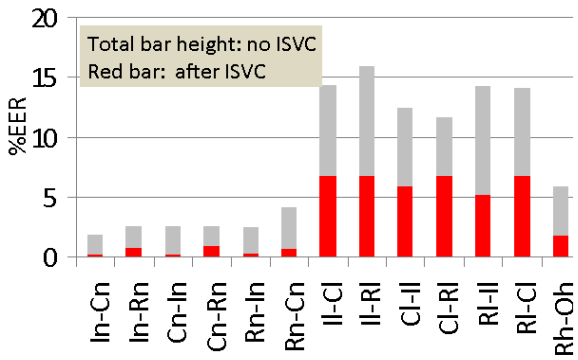


Figure 5. Effect of mismatch in speaking style before and after ISVC. See Figure 2 for abbreviations.

When both the style and effort mismatch are present (plot not shown, but results inferable from the majority of points in Figure 3), baseline error rates generally increase relative to single-mode mismatch conditions. Accordingly, the relative benefit of compensation generally decreases. Such rates probably better reflect real-life intrinsic variation conditions than do single-mode mismatches, and thus constitute an important area for further study.

## 5. Summary and Conclusion

We found that despite a mismatch in ISV training and speaker recognition evaluation data, ISV compensation gave significant improvements for speaker verification on a corpus of intrinsic variation that is highly controlled for channel. Because of this control, we infer that an ISV technique originally designed for channel compensation is indeed modeling the intrinsic variation represented in the data. We further find that relative improvements are (1) inversely related to uncompensated performance, (2) reduced more by vocal effort train/test mismatch than by speaking style mismatch, and (3) reduced additionally for mismatches in both style and level.

An important goal for future research is to better understand the issue of data mismatch in session variability compensation training. Results suggest that a significant amount of the variability obtained in an elicited corpus of intrinsic variation is also present in data that was not collected explicitly to elicit such variation. We hope that further work can shed light on this unexpected finding. In the meantime, a practical implication is that mismatched ISV compensation data may be more useful than previously expected for modeling certain types of within-speaker variability in speech.

## 6. Acknowledgements

We thank Martin Graciarena, Andreas Stolcke, Luciana Ferrer, Harry Bratt, Andreas Kathol, and Fred Goodman for valuable input and infrastructure, and Kristin Precoda and three anonymous reviewers for comments. This work was supported in part by contract 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.

## 7. References

[1] D. Reynolds, "Channel robust speaker verification via feature mapping," Proc. of ICASSP, Hong Kong, China, 2003.

[2] A. Solomonoff, C. Quillen, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," Proc. of ICASSP, Philadelphia, USA, 2005.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Improvements in factor analysis based speaker verification," Proc. of ICASSP, Toulouse, France, 2006.

[4] 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. of Interspeech, Brisbane, Australia, 2008.

[5] R. Vogt, B. Baker, and S. Shridharan, "Modeling session variability in text-independent speaker verification," Proc. of Eurospeech, Lisbon, Portugal, 2005.

[6] D. Matrouf, N. Scheffer, B. Fauve, and J. F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," Proc. of Interspeech, 2007.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 1435-1447, 2007.

[8] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 345-354, 2005.

[9] D. Reynolds, "Speaker identification and verification using Gaussian mixture models," in *Speech Communication*, vol. 17, 1995, pp. 91-108.

[10] R. Vogt and S. Shridharan, "Experiments in session variability modelling for speaker verification," Proc. of ICASSP, Toulouse, France, 2006.

[11] N. Brummer, "SUN SDV system description for the NIST SRE 2008," Proc. of NIST Speaker recognition evaluation workshop, 2008.

[12] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," Proc. of ICASSP, Taipei, 2009.