

Text-Independent Speaker Verification Using Covariance Modeling

Ran D. Zilca

Abstract—This letter describes speaker verification using a covariance-modeling approach for speaker and world modeling. Two verification methods are suggested: frame level scoring and utterance level scoring. Both methods exhibit extremely low computational and model-storage requirements. The suggested methods are tested on the male segment of the 1999 NIST Speaker Recognition Evaluation corpus, using a single training session, and compared to a Gaussian mixture model (GMM) system. The degradation in accuracy and the computational requirements are estimated. Covariance modeling is seen to be a viable alternative to GMM whenever computational and storage requirements must be traded with verification accuracy.

Index Terms—Covariance modeling, second order statistics, speaker recognition, text independent, utterance level scoring.

I. INTRODUCTION

THE prevailing approach in current speaker verification systems is to score each speech frame separately against the claimant speaker's model and combine frame scores to form an utterance score. The need for scoring frames separately may be associated with the present modeling approaches, where local clusters in the feature space are modeled explicitly, e.g., vector quantization (VQ) [1], Gaussian mixture model (GMM) [2]. The tested utterance includes only a subset of the phonetic space modeled by the text-independent speaker model and therefore cannot be compared in its entirety to the model, and each frame is actually scored mostly against local feature clusters. For example, for VQ models, the frame is compared to the closest codebook vector, and for GMM the frame's score is in effect influenced only by the closest Gaussian components. This multimodal nature of speaker modeling mandates extensive computation since each frame is scored separately against the multimodal models during verification, and training is iterative. Motivated by these observations, we seek a simple unimodal modeling approach that will allow simplified training and verification procedures. Previous work on covariance modeling focused on speaker identification tasks, with no background normalization (e.g., [3], [4]). We suggest a simple extension to such methods that allows performing score normalization for speaker verification.

Manuscript received October 12, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Y. Stylianou.

The author was with the Research and Development Division, Amdocs, Raanana 43000, Israel. He is now with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: ranzilca@ieee.org; ranzilca@hotmail.com).

Publisher Item Identifier S 1070-9908(01)02790-0.

A. Single Gaussian (SG)

The SG method is a reduction of GMM, where a single Gaussian models the speaker's training data. When cepstral mean subtraction is used [5], the mean equals zero, resulting in a covariance-only model. Verification using SG is obtained by calculating the difference between the speaker and world scores, according to the standard GMM verification framework [6].

B. Divergence Shape Ratio (DSR)

The *divergence shape* is an information-theoretic measure between two Gaussian classes, introduced by Campbell as a measure of dissimilarity between a reference speaker utterance and a tested utterance [4]. Let $C_{1,2}$ be the covariances of speech features extracted from the two utterances, the divergence shape equals

$$DS_{1,2} = DS(C_1, C_2) = \frac{1}{2} \text{tr} [(C_2 - C_1)(C_1^{-1} - C_2^{-1})]. \quad (1)$$

Campbell used the divergence shape for text dependent speaker identification, and tested it with clean speech and line spectra pairs (LSP) [4]. We suggest using this measure for *score normalization* as part of a text independent, telephone speaker verification task with mel frequency cepstral coefficients (MFCC). Prior to verification, a "world model," denoted as C_{world} , is computed as the sample covariance of speech features gathered from a diverse speaker population. For each speaker, a model is trained by computing the sample covariance of the speaker's training data, denoted as C_{speaker} . Scoring a test utterance against a claimed speaker is accomplished by calculating its sample covariance matrix C_{utt} and then the DSR, defined as the ratio between the divergence shapes of the utterance with respect to the speaker and the world models, taken with an opposite sign, i.e.,

$$DSR_{\text{score}} = -\frac{DS(C_{\text{utt}}, C_{\text{speaker}})}{DS(C_{\text{utt}}, C_{\text{world}})}. \quad (2)$$

As seen from (2), a speech utterance having small divergence shape with respect to a speaker model, relative to its divergence shape from the world model, will produce a high score and vice versa. The utterance sample covariance is calculated frame by frame, yet unlike the frame likelihood based approach (SG, GMM), each frame calculation is performed regardless of the speaker model. Therefore, once the utterance covariance is obtained, scoring with respect to a model (either speaker or world) involves only a single operation. Speaker verification involves

scoring with respect to more than one model in order to obtain a normalized score, and DSR allows obtaining the normalized score using a single frame level computation.

II. EXPERIMENTS AND CONCLUSIONS

The DET curves [7] in Fig. 1 show the verification results for the male segment of the 1999 NIST Speaker Recognition Evaluation corpus. Training was performed using only one of the two training sessions (session “a”), resulting in 1-min-long training sessions. It should be noted that this task is more difficult than the original NIST-1999 evaluation. First of all, only 1 min of training data is made available instead of 2 min. In addition, the training data originates from a single training conversation, whereas for the original evaluation two training sessions were available, allowing for some intersession variability. We may therefore expect worse results compared to the results of the evaluation [8]. The duration of the test segments varies, mostly between 15 and 45 s. A detailed description of the NIST evaluation data and conditions may be found in [8].

The results are presented for three different mismatch conditions.

- 1) Same number same type (SNST). Verification of true speakers (“target trials”) performed from the same telephone number as the enrollment, using the same handset type (handset types may be either carbon-button or electret). Imposter trials are made from a different number using the same handset type.
- 2) Different number same type (DNST).
- 3) Different number different type (DNST).

All experiments were conducted using mismatch-type dependent world models, trained using all the 3-s male test sessions of the 1998 NIST Speaker Recognition Evaluation data, approximately 2 h of speech for each handset type. Preprocessing included 25 ms framing with 50% overlap, using only voiced frames to extract 18 MFCC per frame and perform cepstral mean subtraction.

The covariance modeling methods were compared to a 512 component Bayesian adaptation GMM, trained using the procedure and adaptation parameters described in [6]. As seen in Fig. 1(a), in SNST conditions, SG and DSR perform similarly, degrading the GMM performance from 10% equal error rate (EER) to 14% (40% relative degradation), though DSR and GMM performance merge in the upper left part (high false rejection, low false acceptance). In DNST conditions [Fig. 1(b)], verification EER drops from 19% to 28% and 33% for SG and DSR, respectively (47% to 74% relative degradation). In DNDT conditions [Fig. 1(c)], the gap between the GMM and covariance models is significantly smaller. The degradation relative to GMM is 30% for SG and 18% for DSR. Interestingly, DSR seems to exhibit similar performance for DNDT and DNST conditions, unlike GMM and SG.

An example that illustrates the CPU time required for enrollment and verification for the different methods, as measured for a single verification trial (utterance *aaaa*, speaker 4309), is shown in Table I. CPU time was measured on a P-III, 450 MHz workstation. To allow a fair comparison GMM verification was performed using a five-component approximation as described

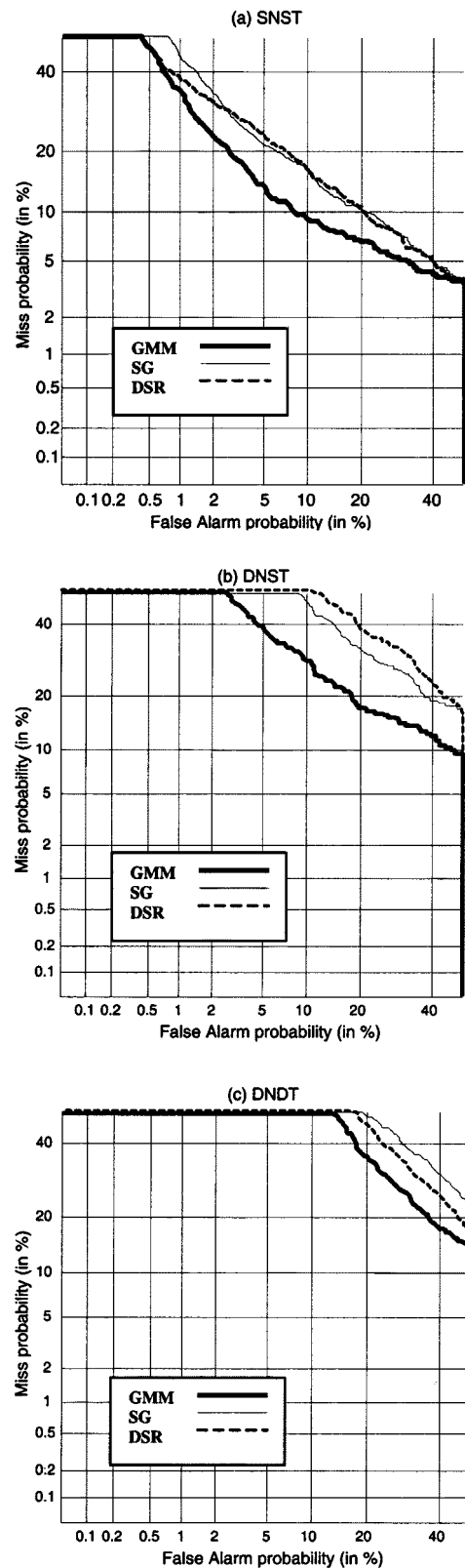


Fig. 1. DET curves comparing DSR, SG, and GMM performance in different mismatch conditions.

in [6]. However, the computational advantage of the covariance modeling methods is evident. Training is four orders of magnitude faster than GMM. SG verification is eight times faster than GMM, and DSR verification is approximately 1300 times

TABLE I
COMPARISON OF CPU TIME [SECONDS], FOR ENROLLMENT AND VERIFICATION.

	Enrollment	Verification
SG	0.07	1.67
DSR	0.07	0.01
GMM	1320	13.44

faster than GMM verification, owing to utterance level scoring. Furthermore, the covariance models require only 171 memory cells, compared to 18 944 for GMM.

The observed verification rates, CPU times, and storage requirements clearly indicate that covariance models allow graceful trading of computational complexity with accuracy requirements. In particular, DSR is shown to be an efficient, promising technique.

The computational advantages of DSR are most useful for speaker recognition systems that involve scoring a single utterance against several models, since the utterance covariance is computed once, frame by frame, but subsequently may be scored against each model in a single operation. For example, in speaker verification systems that use world modeling, as described in this letter, the tested utterance is scored against two models: the speaker's and the world. DSR may be also used with cohort normalization for speaker verification [2], where each utterance is scored against a set of background speakers for nor-

malization. In the case of cohort normalization, the computational advantage would be more significant compared to frame likelihood methods such as GMM.

Future work should focus on tuning and improving the DSR method and evaluating its performance in cohort-based speaker verification.

REFERENCES

- [1] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 387–390, 1985.
- [2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, pp. 91–108, 1995.
- [3] H. Gish, "Robust discrimination in automatic speaker identification," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 289–292, 1990.
- [4] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437–1462, Sept. 1997.
- [5] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition," *IEEE Signal Processing Mag.*, vol. 13, pp. 58–71, Sept. 1996.
- [6] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 963–966.
- [7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 1895–1898.
- [8] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation—An overview," *Dig. Signal Process.*, vol. 10, pp. 1–18, Jan. 2000.