# IMPROVED SPEAKER MODEL MIGRATION VIA STOCHASTIC SYNTHESIS

Jiří Navrátil, Ganesh N. Ramaswamy
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
e-mail: {jiri,ganeshr}@us.ibm.com

## ABSTRACT

Model migration in speaker recognition is a task of converting parametrically-obsolete models to new structures and configurations without the requirement to store the original speech waveforms or feature vector sequences along with the models. The need for model migration arises in large-scale deployments of speaker recognition technology in which the potential for legacy problems increases as the evolving technology may require configuration changes thus invalidating already existing user voice accounts. A migration may represent the only alternative to otherwise costly user re-enrollment or waveform storage and, as a new research problem, presents the challenge of developing algorithms to minimize the loss in accuracy in the migrated accounts. This paper reports on further enhancements of a statistical migration technique based on Gaussian Mixture Models, introduced previously. The present approach is based on a stochastic synthesis of feature sequences from obsolete models that are subsequently used to create the new models. Here, in addition to Gaussian means and priors, as utilized in the previous contribution, also the covariances are included resulting in significant performance gains in the migrated models, compared to the mean-only method. Overall, measured on the NIST 2003 cellular task, the described algorithm achieves a model migration incurring a loss in performance of 8-20% relative to a full re-enrollment from waveforms, dependent on the type of mismatch between the obsolete and the new configuration. The inclusion of the covariance information is shown to reduce the loss of performance by a factor of 3-4 as compared to the baseline mean-only migration technique.

## 1. INTRODUCTION

The task of model migration was introduced in [1] as a problem arising in field deployments with an ever growing number of voice-enabled user accounts. We expect that in the still dynamically evolving area of speaker recognition, legacy issues in the model maintenance will occur, as the average life span of a user account is likely to last longer than an innovation cycle of the underlying authentication technology. In other words, for a voice-enabled account including the user's voice model representation, the particular implemented algorithm that created the model may change one or several times during the overall period of using the account. Since the parametric structure of the user models is dictated by the underlying algorithms used to produce them, significant incompatibilities can be introduced into existing large-scale databases of users. Consequently, algorithmic or data-related changes rendering existing accounts obsolete put infrastructure providers before new problems and decisions on how to address them. Among the few possibilities are: 1) have users actively re-enroll into the new system, 2) automatically re-enroll users from stored original waveform, 3) keep multiple system versions on-line to support obsolete as well as new accounts, 4) automatically convert obsolete models to the new configuration. As discussed in [1], each solution builds on different assumptions and has different degrees of practicability. The process 4) is referred to as *model migration* and builds on the assumption that the obsolete model is the sole information available for the account, i.e. that no original waveform exists. The obvious merit of a well-performing model migration method is the fact that it may be the only alternative to requiring all users to re-enroll into the system.

Building on [1] we focus on the common scenario involving a conversion between two GMM models in a UBM-MAP framework [2], with different GMM sizes and a different composition of the background data, however sharing a common feature space. Throughout the paper the UBM structure from which a speaker model is created via the MAP adaptation is referred to as *substrate*. With a change of the substrate every user model (rendered obsolete) needs to be migrated to the new substrate.

The rest of the paper describes the new migration algorithm based on a stochastic synthesis with inclusion of Gaussian covariances as an enhancement of the baseline mean-only migration formula. Experiments carried out on the cellular task of the 2003 NIST Speaker Recognition Evaluation using the new algorithm are presented and several variants for types of covariance information as well as mismatch configurations are studied.

## 2. MODEL MIGRATION

Considering user models having a GMM structure with mean parameters adapted via the *Maximum A-Posteriori* (MAP) method from a Universal Background Model (UBM) [2], we proposed a statistical method [1] to migrate the user mean parameters from an obsolete model, $M_0$, that were adapted from an obsolete substrate, $W_0$, of size $N_0$ Gaussians to a new user model $M_1$ consistent with a new substrate, $W_1$, of size $N_1$. Both substrate UBMs are assumed in

a feature space identical up to a linear transform, however were composed from different data sets, and, in general, $N_0 \neq N_1$.

A MAP-based algorithm for migrating the mean parameters of the obsolete system to a new substrate in [1] involves computation of a posterior probability of Gaussian $i$ of the new UBM accounting for the obsolete sample mean $\hat{\mu}_j$

$$\begin{aligned} \gamma_{ij} &= \Pr(i|\hat{\mu}_{0j}) \\ &= \frac{\pi_{1i} p_{1i}(\hat{\mu}_{0j})}{\sum_{k=1}^{N_1} \pi_{1k} p_{1k}(\hat{\mu}_{0j})} \\ & \quad 1 \leq i \leq N_1, 1 \leq j \leq N_0 \end{aligned} \tag{1}$$

with $\pi$ the UBM component weights, $p(\cdot)$ the UBM Gaussian density function, and subscripts $0, 1$ denoting the obsolete and the new system, respectively. This is followed by a single iteration of MAP estimation to obtain a new mean

$$\begin{aligned} \mu_{1i} &= \alpha_i \hat{\mu}_{1i} + (1 - \alpha_i) m_{1i} \\ \alpha_i &= \sum_{k=1}^{N_0} n_k \gamma_{ik} / \left( \sum_{k=1}^{N_0} n_k \gamma_{ik} + r \right) \\ \hat{\mu}_{1i} &= \sum_{k=1}^{N_0} n_k \gamma_{ik} \hat{\mu}_{0k} / \sum_{k=1}^{N_0} n_k \gamma_{ik} \\ & \quad 1 \leq i \leq N_1 \end{aligned} \tag{2}$$

where $m_{1i}$ denotes the $i$-th mean vector of the new UBM, $n_k$ the original vector count in the $k$-th obsolete Gaussian (which is either stored in the obsolete model or can be approximated via the $\pi$ parameter), and $r$ the adaptation relevance factor [2].

From a different viewpoint, the above algorithm can be interpreted as providing a new MAP estimate based on a *synthesized* feature vector sequence comprised of the individual obsolete mean vectors in their original proportional representation, distributed via $\gamma_{ij}$ into Gaussians of the new substrate $W_1$.

A drawback of the above migration formula is that the covariance information, be it $\Sigma_{0i}$ of the obsolete UBM Gaussian $i$ or the sample covariance $\hat{\Sigma}_{0i}$ of the speaker data, is not utilized. Therefore we adopt the synthesis interpretation of (3) and generalize it to a 0-th order stochastic process with a Gaussian output distribution to perform the migration task. This *stochastic synthesis* process is formulated as follows: to create a migrated speaker model a sequence of feature vectors is used as input to the appropriate enrollment procedure in the new system. The synthesized sequence

$$X = \{X^1, ..., X^{N_0}\}$$

is composed of blocks

$$X^i = \{x_{it}\}_1^{n_i},$$

whereby each block is generated by an i.i.d. stochastic process with distribution parameters of the $i$-th obsolete Gaussian. Let $Y \sim \mathcal{N}(0, I)$ be a $d$-dimensional normal random variable, then each block $X^i$ is viewed as a sample of size $n_i$ of a random variable $Y_i$:

$$Y_i = A_0^{-1}(Y \hat{\Sigma}_{0i}^{\frac{1}{2}} + \hat{\mu}_{0i}) \tag{3}$$

where $\hat{\mu}_{0i}$ is the sample obsolete mean as used in (3), $\hat{\Sigma}_{0i}$ is the sample obsolete covariance of Gaussian $i$, and $A_0$ accounts for any global linear feature-space transform applied in the obsolete system (e.g. the MLLT as described in [3]). In mean-only MAP-adapted systems the speaker sample covariance may not be available, therefore in our experiments the effect of replacing $\hat{\Sigma}_{0i}$ by the UBM (speaker-independent) covariance $\Sigma_{0i}$ was also studied (see Section 3.). Note that this procedure with $\hat{\Sigma}_{0i} = 0$ is equivalent to the mean-only migration formula (3).

The sequence $X$ serves as input to the new system that performs regular enrollment.

## 3. EXPERIMENTS

### 3.1. Database

The performance of the described method was evaluated using data from the cellular part of the Switchboard (SWB) telephone corpus, as defined by NIST for the 1-speaker cellular detection task in the 2003 Speaker Recognition Evaluations (SRE) [4]. The 2003 set consists of 356 speakers, and a total of 37664 verification trials.

The 2001 cellular SRE, the 1996 SRE landline-quality dataset and an internal cellular-quality data collection served as the data for the estimation of two substrate models (UBMs) and score normalization via T-Norms.

### 3.2. System Setup

The data composition in the two substrate models was designed to differ as follows [1]: while the 2001 SRE data were used in both models, the "obsolete" substrate set included also the 1996 SRE data set, and the "new" substrate model included the internal data set as well as the SRE 1996 but postprocessed by a GSM transcoder [3]. For experimental purposes substrate models with varying sizes between 256 and 2048 Gaussian components were created using the techniques described in [3]. The two models each had a different linear transform applied, which for the purpose of model migration was compensated for as described in [1].

In the detection phase, log likelihood ratio scores are calculated given each test utterance, target model and the corresponding substrate model. Furthermore, the T-Norm score normalization technique is applied. A total of 234 speakers from the 2001 cellular SRE served as T-Norm speakers in both systems and are used in a gender-matched fashion in the test. Note that in all migrated configuration the T-Norm models undergo the same migration procedure as the target models.

The system performance was measured at two operating points, namely in terms of the Equal-Error Rate (EER) and the minimum Detection Costs Function (DCF) as defined in the evaluation plan [4]. Note that the DCF values reported in this paper are scaled by $10^3$.

### 3.3. Two Baselines

Migration results obtained using the stochastic synthesis will be compared to the baseline mean-only migration technique of [1] and also to the ideal achievable performance obtained by using the original waveform to recreate the target models. Although this baseline is idealistic, as it

goes beyond our original assumption of waveform absence, it provides a necessary performance reference point.

### 3.4. Results

Experiments were carried out on varying sizes of the substrate (and consequently the target) models in both the obsolete domain (i.e. size $N_0$ of $W_0$) and the new domain ($N_1$, $W_1$).

Table 1 and Table 2 summarize results obtained with migrating models from and to various sizes ranging between 256 and 2048 Gaussian components, for unnormalized and T-normed systems, respectively. Each row in a table corresponds to a particular obsolete size $N_0$ (or waveform in case of the ideal baseline) with each corresponding column showing performance in terms of the DCF and the EER after a migration to its specific new substrate size $N_1$. The mean-only baseline ("Bsl") and the full stochastic synthesis ("New") are labeled correspondingly.

**Table 1. DCF/EER results for migrated systems without T-Norm. The stochastic synthesis method is labeled "New," and "Ideal Bsl" refers to re-enrollment from waveforms**

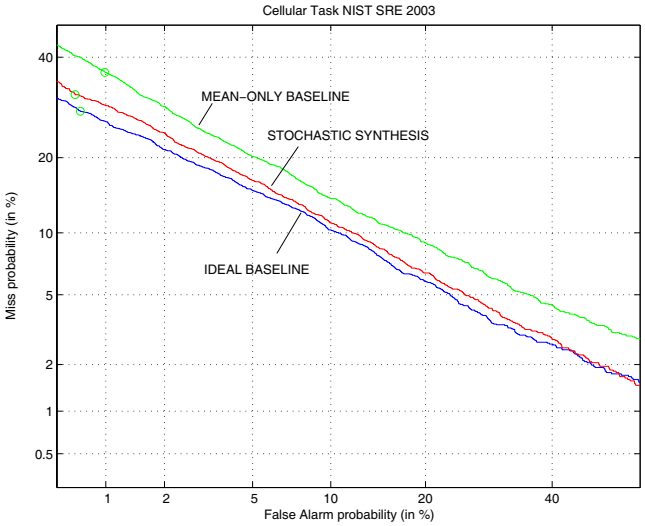
| Original Size ($N_0$) | Target size (Number of Gaussians $N_1$) 2048 | 1024 | 512 | 256 |
|---|---|---|---|---|
| 2048-Bsl | 64.7/20.2 | 68.4/19.7 | 73.1/18.8 | 70.3/17.4 |
| 2048-New | 47.9/12.2 | 48.6/12.3 | 51.7/12.5 | 55.3/12.6 |
| Ideal Bsl | 37.1/9.4 | 39.9/10.1 | 42.6/10.8 | 46.6/11.4 |

**Table 2. DCF/EER results for migrated systems with T-Norm. The stochastic synthesis method is labeled "New."**

| Original Size ($N_0$) | Target size (Number of Gaussians $N_1$) 2048 | 1024 | 512 | 256 |
|---|---|---|---|---|
| 2048-Bsl | 47.2/12.5 | 44.6/12.0 | 46.1/12.4 | 46.4/12.3 |
| 2048-New | 38.6/9.9 | 37.0/9.6 | 37.1/9.9 | 38.5/10.6 |
| 512-Bsl | 75.8/20.4 | 58.0/15.7 | 45.2/12.2 | 50.1/13.8 |
| 512-New | 60.2/15.8 | 46.1/12.7 | 36.0/9.9 | 41.4/10.8 |
| Ideal Bsl | 31.9/8.4 | 32.4/8.8 | 33.5/9.3 | 35.6/10.2 |

Besides the general migration performance trends in the various size configuration already observed in [1], an improvement averaging 20% relative in reduction of the DCF as well as the EER can be seen in the new stochastic synthesis method including the covariance information. This seems to hold consistently across the various conditions and thus appears to apply independently of the model size and the degree of mismatch in the migration. As can be expected, a migration from the larger 2048 to smaller substrates tends to preserve more accuracy than the vice versa case (from 512). Explained from the viewpoint of vector quantization, smaller obsolete substrates relate to a coarser quantization and consequently a greater unrecoverable loss of information.

**Figure 1. Migration example from a 2048- to a 256-Gaussian system with T-Norm for the stochastic synthesis including covariance information and its mean-only baseline**




As compared to the ideal baseline, in the $N_0 = 2048$ case, the new migration algorithm causes a relative performance degradation of 8-20% in DCF and 4-17% in EER, dependent on target size $N_1$, comparing to 30-50% in DCF and 20-50% in EER incurred in the mean-only baseline. thus effectively reducing the loss caused by the mean-only baseline by a factor of about 3-4.

A corresponding DET plot for the T-Normed ideal, mean-only migrated, and covariance-migrated system is shown in Figure 1 indicating that the relative performance seems to behave uniformly across the entire operating range.

Based on the above results it can be concluded that the covariance information is highly relevant to the statistical migration procedure. As mentioned above, however, in some systems (e.g. with mean-only MAP-adapted models) the original speaker covariance may not be available as the covariance parameters were not adapted. To investigate the importance of speaker-specific covariance versus speaker-independent covariance information replacing the former, experiments with the $\hat{\Sigma}_{0i}$ parameters replaced by the UBM $\Sigma_{0i}$ parameters were carried out and the corresponding results are summarized in Table 3. The replacement causes a degradation of about 4-5% relative to the speaker-dependent covariance case, which can be viewed rather minor considering the relative improvement compared to the mean-only case. Similarly, DET plots for both covariance cases are shown in Figure 2.

## 4. CONCLUSIONS

The presented experimental results show that 1) statistical model migration is a viable way of converting models, that were rendered obsolete by system configuration changes, to new models compatible with a new system, 2) the migration process between UBM-GMM configurations benefits significantly from inclusion of the covariance information, which

**Figure 2. Migration from a 2048- to a 256- Gaussian system using speaker-dependent and speaker-independent covariance**

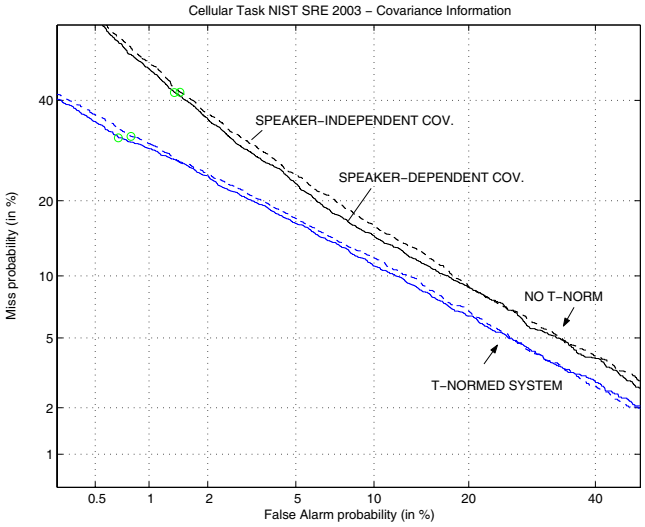



**Table 3. DCF/EER results for stochastic synthesis migration with original speaker covariance and with a replacement by the speaker-independent UBM covariance**

| Mgr. 2048 → 256 | Speaker-Σ | UBM-Σ | Mean-Only |
|---|---|---|---|
| Plain | 55.3/12.6 | 56.2/13.3 | 70.3/17.4 |
| w/T-Norm | 38.5/10.6 | 40.0/11.1 | 46.4/12.3 |
| Ideal Bsl. w/T-N. | 35.6/10.2 | | |

was achieved by the described algorithm of synthesizing feature sequences according to mean and covariance parameters of the obsolete model, and 3) useful covariance information may be retrieved from both the speaker-dependent as well as speaker-independent parameters.

Compared to an ideal baseline involving re-creation of the speaker models from original waveforms, the described method achieved a migration incurring a loss of 4-20% relative, dependent on the degree of mismatch and the UBM size configuration. Across all configurations, the proposed stochastic synthesis algorithm gained a 20% relative improvement over the method in [1], which corresponds to a reduction of the abovementioned migration loss of performance of factor 3-4.

As outlined in [1], the remaining inaccuracies due to model migration can be addressed by subsequent adaptation, specifically in the framework of Conversational Biometrics [5, 6] in which a knowledge-based verification with a subsequent acoustic adaptation is carried out on the migrated model.

Major technical challenges still remain in systems defined in different feature spaces and employing different classifier types, such as a migration from a discriminative-classifier operating in a space of LPCC features to a generative-model classifier operating on MFCCs. A possible solution path could lead via waveform synthesis from the source system followed by an appropriate re-enrollment. Techniques for such PCM synthesis from features (e.g. the MFCCs) are known [7] and open a direction for further research in speaker model migration.

## REFERENCES


[1] J. Navrátil, G. Ramaswamy, and R. Zilca, "Statistical model migration in speaker recognition," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, (Jeju Island, South Korea), October 2004.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, January/April/July 2000.

[3] G. Ramaswamy, J. Navrátil, U. Chaudhari, and R. Zilca, "The IBM system for the NIST 2002 cellular speaker verification evaluation," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Hong Kong), IEEE, April 2003.

[4] (URL), "http://www.nist.gov/speech/tests/spk/index.htm."

[5] S. Maes, J. Navrátil, and U. Chaudhari, *E-Commerce Agents, Marketplace - Solutions, Security Issues, and Supply Demand*, ch. Conversational Speech Biometrics. LNAI 2033, Springer Verlag, 2001.

[6] G. Ramaswamy, R. Zilca, and O. Alecksandrovich, "A programmable policy manager for conversational biometrics," in *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, (Geneva, Switzerland), September 2003.

[7] D. Chazan, G. Cohen, R. Hoory, and M. Zibulski, "Speech reconstruction from Mel-frequency cepstral coefficients and pitch frequency," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, June 2000.