

STATISTICAL MODEL MIGRATION IN SPEAKER RECOGNITION

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

ABSTRACT

In large-scale deployments of speaker recognition systems the potential for legacy problems increases as the evolving technology may require configuration changes in the system thus invalidating already existing user voice accounts. Unless the entire database of original speech waveform were stored, users need to reenroll to keep their accounts functional, which, however, may be expensive and commercially not acceptable. We define model migration as a conversion of obsolete models to new-configuration models without additional data and waveform requirements and investigate ways to achieve such a migration with minimum loss of system accuracy. As a proof-of-concept, an algorithm for statistical migration in the Maximum A-Posteriori framework is studied and evaluated experimentally using the NIST SRE-03 dataset. The migration step is discussed in a wider conceptual framework of Conversational Biometrics.

1. INTRODUCTION

A variety of real-world challenges arise in today's voice authentication technology accompanied by an increasing business demand for security in telephone applications. While most of the current challenges are directly or indirectly related to the accuracy, robustness and computational efficiency, future field deployments with an ever growing number of voice-enabled user accounts will bring new sets of practical issues with them. The still dynamically evolving area of speaker recognition promises one particular such challenge: legacy issues in voiceprint model maintenance. It is reasonable to expect that 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. Because the parametric structure of the user models is dictated by the underlying algorithms used to produce them, significant legacy issues (incompatibilities) can be introduced into existing large-scale databases of users. Consequently, algorithmic changes rendering existing accounts obsolete put infrastructure providers before new problems and decisions. For instance, assume a service provider maintains several hundreds of thousands of voice-enabled accounts including users' voiceprint models in their database. These models have a close relationship to data and algorithmic components inherent to the system. With a new generation of updated (improved) components the provider is faced with the question of how to keep the existing accounts usable (voice-

enabled). Among the few possibilities to address this problem 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. Obviously, each solution builds on different assumptions (e.g. existence of an original waveform), has different consequences (e.g. increased complexity due to multiple system versions), and degrees of practicability (e.g. having users call to re-enroll which may be unacceptable). Acknowledging the suitability of those approaches in specific conditions, we aim our focus at the lastly mentioned way of automatically converting an obsolete user model (based on a legacy system configuration) to a new model (compatible with the updated system) with minimum compromise in accuracy. In this paper, this process 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 short of losing the client.

The primary quality criterion for any model migration technique is its performance at preserving system accuracy. The degree to which individual techniques will perform largely depends on the type of mismatch between the obsolete and the new technology, the background (system) data mismatch, as well as configuration mismatch. While it is desirable to have seamless migration independent of the modeling and features used, it is reasonable to focus first on the more common scenario involving a conversion between two GMM models of different size sharing common feature space. While we believe that most migration cases can be addressed and solved to a certain degree, this paper concentrates on mismatch in the configuration rather than in the underlying technology type. In particular, we study the case of mismatched sizes between GMM models that are adapted from some system-internal model structures (e.g. universal background models with differing data composition) and refer to these structures as *substrates*. With a change of the substrate every user model (now obsolete) needs to be migrated to the new substrate.

The rest of the paper describes a method of achieving such a migration in a statistical fashion (Section 2), presents an experimental study carried out on the cellular task of the 2003 NIST Speaker Recognition Evaluation (Section 3), and discusses the migration scenario in a wider conceptual framework of Conversational Biometrics (Section 4).

2. MODEL MIGRATION

We restrict our considerations to the task of speaker verification and user models having a GMM structure with mean parameters adapted via the *Maximum A-Posteriori* (MAP) method from a Universal Background Model (UBM). We propose a statistical method 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 . We assume both substrate UBMs are trained in a feature space identical up to a linear transform, however were composed from different data sets, and, in general, $N_0 \neq N_1$. Possibilities of overcoming the feature space assumption are outlined in Section 4.

The following one-iteration algorithm implements the MAP principle based on the means of the obsolete model and given the parameters of the new substrate.

1. Reconstruct the obsolete sample means $\hat{\mu}_{0i}$ of the target speaker from the adapted M_0 and W_0 for each Gaussian i . For the typical adaptation formula $\mu_{0i} = \frac{n_i}{n_i+r}\hat{\mu}_{0i} + \frac{r}{n_i+r}m_{0i}$ ([1]) this can be easily achieved knowing the global relevance factor r , the vector softcount n_i (which is assumed available along with the obsolete model), and using the obsolete UBM mean m_{0i} .
2. Calculate the set of posterior probabilities of Gaussian i of the new UBM accounting for the obsolete sample mean $\hat{\mu}_j$

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

where π_{1i} denotes the prior probability and $p_{1i}(\cdot)$ the observation probability of Gaussian i of model W_1 .

3. Calculate new sample mean estimates on the new substrate:

$$\hat{\mu}_{1i} = \frac{\sum_{k=1}^{N_0} n_k \gamma_{ik} \hat{\mu}_{0k}}{\sum_{k=1}^{N_0} n_k \gamma_{ik}} \quad 1 \leq i \leq N_1 \quad (2)$$

Note that each component contribution is weighted by the original sample size n_k attributed to each mixture component to reflect the natural proportionality of the data. This is identical to multiplying by an obsolete prior probability π_{0k} which, however, is typically not used in mean-only MAP adapted systems and is typically replaced by π_{1k} .

4. Compute new (migrated) mean parameters via adaptation:

$$\begin{aligned} \mu_{1i} &= \alpha_i \hat{\mu}_{1i} + (1 - \alpha_i) m_{1i} \\ \alpha_i &= \frac{\sum_{k=1}^{N_0} n_k \gamma_{ik}}{\left(\sum_{k=1}^{N_0} n_k \gamma_{ik} + r \right)} \end{aligned} \quad (3)$$

$$1 \leq i \leq N_1 \quad (4)$$

The above algorithm can be interpreted as one providing a new MAP estimate based on the new substrate model W_1 of a feature vector sequence comprised of the individual

obsolete mean vectors in their original proportional representation. Further elaborating this idea, the described migration is identical to using the original training vector sequence, however, without its original causal ordering, quantized into N_0 different codebook vectors via the obsolete substrate model W_0 .

In case of differences in feature spaces due to invertible linear transforms, such as the Maximum Likelihood Linear Transform (MLLT) [2], a straightforward transformation can be applied as follows

$$\hat{\mu} = A_1(A_0)^{-1}\hat{\mu}' \quad (5)$$

with $\hat{\mu}'$ the mean in the space of A_0 , and A_1 the transform into the space of W_1 .

Note that in computing the posteriors in Step 2, only the obsolete mean parameters are used. An alternative approach to further include the obsolete covariances into the softcount computation utilizes the symmetric KL divergence of two Gaussians defined as

$$\begin{aligned} D_{KL}(\mathcal{N}_p||\mathcal{N}_q) &= \frac{1}{2}tr \left[(\Sigma_p - \Sigma_q)(\Sigma_p^{-1} - \Sigma_q^{-1}) \right] + \\ & \quad \frac{1}{2}tr \left[(\Sigma_p^{-1} + \Sigma_q^{-1})(\mu_p - \mu_q)(\mu_p - \mu_q)^T \right] \end{aligned} \quad (6)$$

This measure becomes 0 iff $\mu_p = \mu_q$ and $\Sigma_p = \Sigma_q$, as opposed to the quadratic term in the exponent of the Gaussian function, $(\mu_p - \mu_q)^T \Sigma_p^{-1} (\mu_p - \mu_q)$ which becomes 0 already with $\mu_p = \mu_q$. Utilizing the KL divergence can prevent overemphasizing particular Gaussian pairs in γ_{ij} due to pure mean similarity. To obtain a proper scale we use (6) in an exponential form as follows:

$$\begin{aligned} \gamma_{ij} &= \sqrt{\pi_{1i}\pi_{0j}} \exp(-D_{KL}(p_{1i}||p_{0j}))/C_j \\ & \quad 1 \leq i \leq N_1, 1 \leq j \leq N_0 \end{aligned} \quad (7)$$

with C_j the normalizing term to satisfy $\sum_i \gamma_{ij} = 1$. The above form of γ is used in Step 2 in place of (2).

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) [3]. 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 differed as follows: 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 [4]. For experimental purposes substrate models with varying sizes between 256 and 2048 Gaussian components were created using the techniques described in [4]. The two models each had a different linear transform applied, which for the purpose of model migration was compensated for via (5).

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.

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 [3].

3.3. Naive Baseline

Given the previously stated assumption of the original waveform being unavailable, it may be debatable what other realistic baseline solution should be considered to compare to the described migration. We consider a simplistic solution using the obsolete model as a target likelihood generator while calculating the background likelihood on the new substrate. Later on, this baseline is referred to as naive, noting however, that in many real cases even such a simplistic solution is not practical as even a difference in the GMM size may not allow for a proper integration.

3.4. Ideal Baseline

Migration results will further be compared to an 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.

3.5. Results

Experiments were carried out on varying sizes of the substrate (and consequently target) models in both the obsolete domain (i.e. size N_0 of W_0) and the new domain (N_1 , W_1). Of interest is the behavior of the migrated models with respect to the difference in size during the migration process as well as in absolute sense of the size.

An initial overview of the essential migration configurations and the two baselines for a particular case of a $2048 \rightarrow 256$ mixture component conversion is shown in Figure 1. Between the two delimiting DET curves, i.e. the naive baseline, which turns out to provide a rather poor performance (EER of 40%), and the ideal baseline that includes the T-Norm (EER of 10.2%), three variants of a migrated system are plotted: 1) converted models without score normalization, 2) converted models with scores normalized using the new system’s T-Norms, and 3) converted models normalized using a set of T-norms migrated consistently along with the obsolete targets (EER=12.3%). The latter normalization appears to have a significant positive impact on the performance of migrated models resulting in an overall loss of about 2% absolute points in the EER to the ideal baseline, therefore the consistent T-Norm is applied in all subsequent experiments.

An exhaustive set of experiments with migrating models from and to various sizes ranging between 256 and 2048 Gaussian components for unnormalized and T-normed systems is summarized in Table 1 and Table 2, respectively. Each row in a table corresponds to a particular obsolete size N_0 (or waveform in case of 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 .

Several trends can be readily extracted from these results. First, compared to the ideal baseline the performance no-

Table 1. DCF/EER results for migrated systems without T-Norm

Original Size (N_0)	Target size (Number of Gaussians N_1)			
	256	512	1024	2048
256	80.2/22.2	94.8/32.0	97.3/35.6	99.1/39.3
512	85.2/23.3	76.6/21.5	90.0/29.1	96.3/34.7
1024	79.6/20.5	81.2/22.8	64.9/18.7	87.7/29.7
2048	70.3/17.4	73.1/18.8	68.4/19.7	64.7/20.2
Ideal Bsl	46.6/11.4	42.6/10.8	39.9/10.1	37.1/9.4

Table 2. DCF/EER results for migrated systems with T-Norm

Original Size (N_0)	Target size (Number of Gaussians N_1)			
	256	512	1024	2048
256	49.0/13.5	61.7/16.3	74.2/19.8	85.6/23.0
512	50.1/13.8	45.2/12.2	58.0/15.7	75.8/20.4
1024	48.1/13.3	47.4/13.0	41.0/11.1	63.0/16.6
2048	46.4/12.3	46.1/12.4	44.6/12.0	47.2/12.4
Ideal Bsl	35.6/10.2	33.5/9.3	32.4/8.8	31.9/8.4

tably degrades with larger differences in the sizes N_0 and N_1 while the least relative loss occurs around the diagonal elements where $N_0 = N_1$. In those particular cases, the migration challenge is reduced to aligning two substrate models of same size but of different data composition, while in the other cases the additional task of one-to-many or many-to-one Gaussian mapping comes to play. Second, as can be expected, a migration from larger to smaller substrates tends to preserve more accuracy than the vice versa case. This is easily explained from the viewpoint of vector quantization where smaller obsolete substrates relate to a coarser quantization and consequently a greater unrecoverable loss of information.

Overall, in terms of the EER performance the described migration technique including a consistent score normalization causes an average loss of about 3-4% for larger substrates (1024+ Gaussians) with conversions between reasonably close sizes (half or double the size). The least overall loss with respect to the ideal baseline occurs with migrating the largest (2048) to any other smaller system.

Table 3 compares the migration algorithm using observation probability to compute softcounts as in (2) with using the alternative symmetric KL divergence as in (7) on two selected migration cases. Although the KL-based migration outperforms the standard observation probability in the unnormalized case, the same does not hold when applying the T-Norm. An explanation for the degradation in conjunction with the T-Norm requires further study.

Table 3. DCF/EER results for migration based on probabilities versus the KL divergence

Migration	512 \rightarrow 256	2048 \rightarrow 256
Prob. (plain)	85.2/23.3	70.3/17.4
KL (plain)	62.8/16.8	60.5/15.6
Prob. (T-Norm)	50.1/13.8	46.4/12.3
KL (T-Norm)	56.4/16.0	52.2/14.0

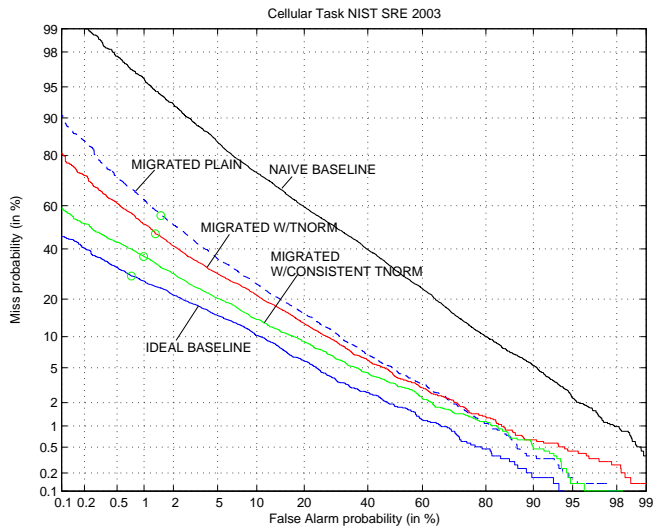


Figure 1. Migration example from a 2048- to a 256-Gaussian system with and without T-Norm. The naive baseline uses obsolete models while the ideal baseline recreates models from the original waveform.

4. DISCUSSION

The experimental results obtained using the described migration technique appear promising particularly because of the fact that neither of the presented baselines may in fact be a practical solution to the legacy problem. While the naive baseline by far does not provide a necessary level of accuracy, the ideal baseline will not be existent as per initial assumption.

Although a system migrated via the MAP method entailing about 20% relative performance degradation for the converted accounts could well be considered usable, we also argue that the acoustic performance can fully be recovered (even improved) in a wider conceptual authentication framework of Conversational Biometrics (CB). The CB approach relies on two information sources conveyed through voice, namely the acoustic voiceprint reflecting the related biometrics and personal knowledge extracted from a dialog, thus vastly increasing its robustness and flexibility [5, 6]. Thanks to the double footing, a CB-based system governed by a policy manager [7], i.e. a component determining reliance on the two individual information sources, is capable of compensating for an expected degradation in the acoustic biometrics by temporarily emphasizing the knowledge extraction component. The newly acquired acoustic sample can subsequently be used to adapt a freshly migrated account to further improve its accuracy. Due to the CB process, such an adaptation can be carried out in an unsupervised fashion and securely. More detail on adjusting CB policies can be found in [7].

As initially pointed out, focusing on more common cases of system mismatches involving configuration model changes and data updates covers a first proof of concept for model migration. However, the class of more dramatic system mismatches still remains to be studied, including examples of incompatible feature spaces and classifiers. In general, we argue that any waveform modeling system can be inverted to produce the original information less a loss incurred by the modeling. For example, using specialized

algorithms waveform signals can be synthesized from cepstral features [8]. Thus, it is reasonable to assume a migration between any two systems through such synthesis is possible, however with varying degradation. This area surely remains open and is likely to be addressed in the future with wide-spread deployments of voice authentication technology.

5. CONCLUSIONS

The presented experimental results show that 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. In compatible feature spaces, the migrated system incurred a relative degradation in EER and DCF performance ranging between 15 - 30% in most cases, dependent on the degree of mismatch in model size. This loss is compared to an ideal baseline that uses the original waveform signal to recreate the models. In absence of the waveforms, a naive baseline using the obsolete models directly provides error rates of about four times higher than the proposed migration algorithm. The use of consistent T-Norms was shown as beneficial to the migrated system. We outlined the use of model migration in a framework of Conversational Biometrics in which the relative performance loss is compensated for by increasing the emphasis on knowledge verification with a subsequent acoustic adaptation step carried out on the migrated model.

REFERENCES

- [1] 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.
- [2] U. Chaudhari, J. Navrátil, and S. Maes, "Multi-grained modeling with pattern-specific maximum likelihood transformations for text-independent speaker recognition," *IEEE Trans. Speech and Audio Processing*, 2002.
- [3] (URL), "<http://www.nist.gov/speech/tests/spk/index.htm>."
- [4] 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.
- [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] U. Chaudhari, J. Navrátil, G. Ramaswamy, and R. Zilca, "Future speaker recognition systems: Challenges and solutions," in *Proc. of AUTOID-2002*, (Tarrytown, NY), March 2002.
- [7] G. Ramaswamy, R. Zilca, and O. Aleksovich, "A programmable policy manager for conversational biometrics," in *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, (Geneva, Switzerland), September 2003.
- [8] 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.