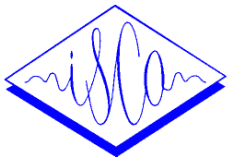

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ON-LINE UNSUPERVISED ADAPTATION IN SPEAKER VERIFICATION

Larry P. Heck *Nikki Mirghafori*
Nuance Communications, 1380 Willow Road, Menlo Park, CA 94025 USA

ABSTRACT

This paper presents a new approach to on-line unsupervised adaptation in speaker verification. The approach extends previous work by (1) improving performance on the enrollment handset-type when adapting on a different handset-type (e.g., improving performance on cellular when adapting on a landline office phone), (2) accomplishing this cross channel improvement without increasing the size of the speaker model after adaptation, (3) employing a count-based, parameter-dependent smoothing algorithm that emphasizes the use of mean parameters in the speaker models until sufficient adaptation data are present to accurately estimate variances, and (4) developing a new confidence-based adaptation update weight which minimizes the corrupting effects on the speaker models from impostor attacks. Experimental results were completed on a gender-balanced database of Japanese digits with 5222 speaker models across mixed channel conditions (landline and cellular). After adaptations on 8 separate phone calls with a single 8-digit utterance per call and a 12.5% impostor attack rate, the EER was reduced by 61% (rel.) using the new unsupervised adaptation approach. This compares favorably to the (optimal) 84% reduction in EER resulting from supervised adaptation.

1. INTRODUCTION

One of the most significant sources of performance degradation in a speaker verification system is the acoustic mismatch between the enrollment and subsequent verification sessions. The acoustic mismatch is a result of differences in the transducer, acoustic environment, and the communication channel characteristics (e.g., varying channels associated with combinations of different subnetworks utilized in a telephone call). Of the factors contributing to acoustic mismatch in telephony applications, it has been shown that the mismatch in transducers of telephone handsets is the most dominate source of performance degradation[3, 5].

To address the acoustic mismatch problem, a variety of approaches for robust speaker recognition have been developed in the past several years. These approaches include robust feature, model, and score-based normalization techniques which are summarized in [2]. These approaches use off-line development data to compensate for the effects of acoustic mismatch that will be present when the system is used on-line. An alternative approach is on-line unsupervised adaptation[1, 4, 6]. On-line unsupervised adaptation can be used to "learn" the unseen channel characteristics automatically while the system is being used in the field. An advantage of this approach is its ability to provide significantly more data for parameter estimation than typically available to the speaker verification system, facilitating more sophisticated modeling approaches and automated parameter tuning. Also, rather than predicting the effects of acoustic mismatch with development data, the effects can be observed directly from this additional data.

This paper presents a new approach to on-line unsupervised adaptation of speaker verification models. Specifically, the new approach automatically updates a speaker model with information from subsequent verification sessions, including user utterances on new handset-types. To address limitations of the previous approaches to on-line adaptation, the updating of the speaker model is accomplished without

- negative effects from impostor attacks,
- increasing the size of the speaker model, and
- degrading the performance on the enrollment handset-type when adapting on new handset-types.

While the actual size of the speaker model remains fixed, the effective complexity of the model increases by employing parameter-specific smoothing factors (i.e., separate rates of smoothing for means and variances). In addition, the approach employs a forgetting factor to allow the speaker model to slowly change with the user over time (months/years). Finally, the approach is implemented in a computationally efficient manner with a negligible increase in computations resulting from the use of on-line adaptation.

2. APPROACH

The approach developed in this paper for on-line unsupervised adaptation is an extension of our model-based transformational approach to robust speaker recognition[7]. Our model-based approach constructs a speaker model by adapting a handset-dependent, gender-dependent, and speaker-independent Gaussian Mixture Model (GMM) using a Bayesian adaptation approach. Speaker model synthesis is used to synthetically create speaker models on channels not seen during enrollment. These new speaker models can be invoked during verification to ensure that all scoring is completed against models that match the handset of the current verification session.

The general form of the verifier used in this work is a likelihood ratio detector, i.e.,

$$\Lambda(X \mid s) = \frac{1}{T}\sum_{t=1}^{T} [\, \log p(\underline{x}_t \mid \lambda_s) - \log p(\underline{x}_t \mid \lambda) \,] \quad (1)$$

where $X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_T\}$ denotes the feature vectors extracted from the utterances by the feature extraction front end, λ_s is the model of the speaker s, and λ is the model of the impostor population. Probability density functions of both speaker and impostor models are estimated with GMMs as follows:

$$p(\underline{x}_t \mid \lambda) = \sum_{i=1}^{L} w_i p(\underline{x}_t \mid b_i) \quad (2)$$

with mixture weights w_i, and gaussian densities $p(\underline{x}_t \mid b_i)$. Each speaker GMM is therefore parameterized by a collection of mixture weights, means, and covariances, i.e., $\lambda = \{\mathbf{W}, \mathbf{M}, \mathbf{\Sigma}\}$.

Multiple GMMs are estimated for each speaker, with each GMM corresponding to a unique channel. Each GMM is constructed by adapting a common root GMM in a way that ensures a correspondence between each gaussian in all GMMs[7]. A model for a channel that was not seen during enrollment is synthetically constructed through a model-based transformation,

$$T_{ab}(w_i) = w_i \left(\frac{w_{b,i}}{w_{a,i}} \right) \quad (3)$$

$$T_{ab}(\underline{\mu}_i) = \underline{\mu}_i + (\underline{\mu}_{b,i} - \underline{\mu}_{a,i}) \quad (4)$$

$$T_{ab}(\underline{\sigma}_i^2) = \underline{\sigma}_i^2 \left(\frac{\underline{\sigma}_{b,i}^2}{\underline{\sigma}_{a,i}^2} \right) \quad (5)$$

where T_{ab} is the transformation between channels a and b, and $w_{a,i}$, $\underline{\mu}_{a,i}$, and $\underline{\sigma}_{a,i}^2$ refer to the mixture weight, mean, and variance of the i-th gaussian of channel a.

To extend the above approach for on-line unsupervised adaptation, several new techniques must be developed that can automatically (1) determine if the data should be used to adapt the speaker model, (2) update the statistics of the speaker model, and (3) enable the system to "track" recent changes in the speaker's voice even if the speaker model has been adapted over long periods of time. The constraints on the design of these techniques are that there will be only one enrollment session, no increase in the size of the speaker model, and no degradation in performance on the initial enrollment channel caused by adapting to new data from other channels.

To determine if the data should be used to adapt the speaker model, we employ a confidence-based weighting scheme that updates the speaker model more aggressively if the verifier is confident of the speaker's identity. Using the verifier score Λ as defined in Equation (1), we form the adaptation weight $\mathcal{W}$ by utilizing a nonlinear function of the verifier score based on a cumulative Rayleigh distribution

$$\mathcal{W}(\Lambda) = 1 - \exp\left[\frac{-(\Lambda - \tau)^2}{2b^2} \right] \quad (6)$$

where τ is the acceptance threshold of the verifier, and b is the Rayleigh coefficient which controls the smoothness of the function.

To update the speaker model statistics, as well as provide a method that "tracks" recent changes in the speaker's voice even after many adaptations, we use the following equations:

$$\begin{aligned} E_i(\underline{x}) &= E_i(\underline{x})^{[0]} (1-F) + \mathcal{W}(\Lambda)\ \beta_\mu\ E_i(\underline{x})^{[1]} \\ E_i(\underline{x}^2) &= E_i(\underline{x}^2)^{[0]} (1-F) + \mathcal{W}(\Lambda)\ \beta_\sigma\ E_i(\underline{x}^2)^{[1]} \\ n_i &= n_i^{[0]} (1-F) + \mathcal{W}(\Lambda)\ \beta_w\ n_i^{[1]} \end{aligned}$$

where $E_i(\underline{x})$ and $E_i(\underline{x}^2)$ are the expected values of the data $\underline{x}$ and $\underline{x}^2$, respectively for the i-th gaussian in the speaker model, n_i is the probabilistic occupancy of the data in the i-th gaussian, $E_i(.)^{[j]}$ is the sufficient statistic of the speaker model for the j-th adaptation iteration (e.g., j-th phone call), and $\mathcal{W}$ is the adaptation weight defined in (6). The terms $(\beta_\mu, \beta_\sigma, \beta_w)$ are Bayesian smoothing factors. Separate smoothing factors are used to enable the system to, for example, rely more heavily on the first-order sufficient statistics until adequate observations have been accumulated to properly estimate the second-order sufficient statistics. Using separate smoothing factors is particularly important for on-line adaptation since it allows the effective complexity of the speaker model to grow with the additional data from new verification attempts, without increasing the actual complexity of the speaker model. The forgetting factor, F, is a number between 0 and 1. Setting $F = 0$ will make the system "remember" statistics from all past utterances completely, and setting $F = 1$ will make the system perfectly track speaker changes but "forget" everything from the past [1].

Finally, to satisfy the constraints of no growth in the size of the speaker model with adaptation while minimizing degradation in performance on the initial enrollment channel, we develop an "inverse synthesis" approach. The inverse synthesis approach relies on the fact that the transformations between channels described in Equation (3-5) are lossless. For each adaptation utterance, channel-dependent statistics are gathered and saved. Instead of storing the statistics of each channel separately on disk, the new approach stores the new statistics along with the original statistics on a single channel by first transforming the new statistics back to the enrollment channel. The statistics (counts) from each channel are added together before storage, resulting in no increase in the size of the speaker model. In addition, the lossless nature of the mappings between channels enables the system to exactly recover the statistics from the new channel on the next verification attempt.

To illustrate the approach, let us consider an example. A user enrolls his voice on an office phone. We will refer to the enrollment channel as channel a, and the resulting speaker model as $\lambda_a^{[0]}$. After enrollment, the user calls the system on a pay phone (channel b). Our goal is to determine the resulting models after the adaptation phone call, i.e., $\{\lambda_a^{[1]}, \lambda_b^{[1]}\}$.

The speaker model for channel b after the first adaptation phone call is given as

$$\begin{aligned} \lambda_b^{[1]} &= \tilde{\lambda_b}^{[0]} \oplus \lambda_b^{[0]} \\ &= T_{ab}(\lambda_a^{[0]}) \oplus \lambda_b^{[0]} \end{aligned}$$

where $\tilde{\lambda_b}^{[0]}$ is the *synthetic* speaker model for channel b determined through a forward transformation T_{ab} of the speaker model from channel a (the symbol $\oplus$ indicates that the addition is completed on the counts rather than the model parameters directly). To store these new statistics from channel b, we could save both models $\lambda_a^{[0]}$ and $\lambda_b^{[0]}$, but this would effectively double the storage space required in the speaker model database. Instead, we can map the new statistics, $\lambda_b^{[0]}$, back onto channel a, add them to the existing statistics from $\lambda_a^{[0]}$, and then save a single model, $\lambda_a^{[1]}$.

$$\lambda_a^{[1]} = \lambda_a^{[0]} + T_{ab}^{-1}(\lambda_b^{[0]})$$

By following this procedure, the size of the speaker model does not grow, and the model on channel a "learns" from the call made on channel b. Also, it can be shown that the speaker model for channel b can be exactly reconstructed since the transformations are one-to-one and invertible, i.e., $T_{ab}^{-1} T_{ab} = I$ where I is the identity matrix. This is shown as

$$\begin{aligned} \tilde{\lambda_b}^{[1]} &= T_{ab}(\lambda_a^{[1]}) = T_{ab}(\lambda_a^{[0]} \oplus T_{ab}^{-1}(\lambda_b^{[0]})) \\ &= T_{ab}(\lambda_a^{[0]}) \oplus \lambda_b^{[0]} = \lambda_b^{[1]} \end{aligned}$$

[1] Note that the adaptation update equations only affect the speaker models and not the impostor models. After the statistics for the speaker model are updated, the final speaker model used for verification (λ_s in Equation (1)) is constructed by smoothing the updated speaker model with the impostor models as described in [7].

3. EXPERIMENTAL RESULTS

In the experiments presented, we will refer to two approaches developed in this paper as "SMS" and "SMS+Inverse", where both use the Speaker Model Synthesis (SMS) approach but the former stores the new adaptation statistics separately with each channel, and the latter transforms the new statistics back to the enrollment channel. Since we are using three channel types in these experiments (electret, carbon-button, and cellular handsets), the SMS approach yields speaker models that are triple the size of the SMS+Inverse approach.

3.1. The Database

We used a database of Japanese digit strings. This database contains a gender-balanced set of 40 speakers, each of whom made four calls (two landline and two cellular). In each call, three repetitions of ten unique 4-digit strings were spoken. In other words, each unique 4-digit string was uttered 12 times by each speaker, half in a landline and half in a cellular condition.

To extend the data, multiple speakers models (between 130 to 150 on average) were built for each of the 40 speakers. Care was taken to assure that each of the speaker models were distinct, by training each on a unique combination of two 4-digit strings. Roughly half of the models were trained on landline and half on cellular data.

For all the experiments reported, the data is divided into three disjoint subsets: a training set for building the speaker models, a verification test set, which is run after each iteration of adapting the models, and an adaptation subset.

3.2. Unsupervised vs. Supervised Adaptation

For the unsupervised adaptation performance results of this paper, 5222 speaker models were trained on three repetitions of an 8-digit utterance (each utterance was constructed by concatenating 2 4-digit utterances from the database). For verification testing, a total of 66899 mixed-gender impostor trials and 12258 true-speaker trials were performed (an average of 2.3 true-speaker and 12.8 impostor trials per model), each composed of one 8-digit utterance. The adaptation set was composed of eight utterances (8-digits each) for each speaker model, of which seven utterances were from the true-speaker and one utterance was by an impostor. The number of impostor attempts were uniformly distributed across all trials, such that **1)** in every adaptation iteration, 1/8 of the speaker models were attacked by an impostor, and **2)** 1/8 of the total trials for each speaker model were by an impostor.

The outline of the experiment was as follows: first, the speaker models were trained on the enrollment data. A verification test was run on all models to establish a baseline performance. Next, each model was adapted on one adaptation utterance. The adaptation step was followed by a round of verification testing to track the improvement in performance. The last two steps (adapt & test) were repeated eight times. The diagram in Figure 1 outlines the experimental flow.

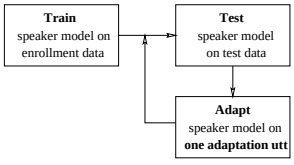

Figure 1. The experimental flow.

Table 1 shows the performance of the two unsupervised adaptation approaches developed in this paper, "SMS" and "SMS+Inverse". These approaches are compared to two separate baselines, the first being the (unadapted) initial enrollment speaker models (Baseline), and the second being the optimal (Supervised) adaptation case. The new SMS and SMS+Inverse unsupervised adaptation techniques improved the EER by 55% and 61%, respectively, over the recently presented state-of-the-art baseline system[7]. Also, this performance compares favorably to the optimal improvement of 84% for supervised adaptation.

Iteration	Supervised	SMS	SMS+Inverse
Baseline	5.67% ($\pm$ 0.2%)		
Iter 1	3.01%	4.20%	4.09%
Iter 2	2.27%	3.90%	3.54%
Iter 3	1.98%	3.86%	3.23%
Iter 4	1.67%	3.71%	3.02%
Iter 5	0.94%	3.36%	2.78%
Iter 6	0.97%	2.49%	2.28%
Iter 7	0.91%	2.61%	2.27%
Iter 8	-	2.58%	2.21%

Table 1. Improved EERs for the held-out verification test set after each consecutive iteration of unsupervised adaptation. The new SMS and SMS+Inverse unsupervised adaptation techniques improved the EER by 55% and 61% over Baseline, compared with 84% for (optimal) supervised adaptation).

3.3. Channel-Specific Performance

The goal of our second set of experiments was to study the performance of the adaptation algorithms for every enrollment/test/adaptation channel combination. In order to avoid obfuscating the results with possible model corruption, we chose to run the experiments in a more controlled manner by using supervised adaptation throughout.

A total of 5,935 speaker models (3,090 landline and 2,845 cellular) were each trained on three 8-digit utterances. The test set for each model was composed of two true-speaker utterances (one landline and one cellular) and 12 same-gender impostors (six landline and six cellular). The adaptation set was made up of four true-speaker utterances (two landline and two cellular).

In the top section of Table 2, the baseline EERs are reported. As expected, the matched conditions of landline and cellular (with 3.81% and 5.27%, respectively) are significantly better than the mismatched training/testing conditions of cellular/landline (with 7.80%) and finally, landline/cellular (with 8.33%). The table reports the EER for every enrollment/test channel combination. After four iterations of supervised adaptation, the SMS+Inverse algorithm improves the baseline from 7.14% to 2.67%, a 63% improvement. The mismatched and matched condition performance has improved by 55-71% and 64-85%, respectively, compared to the baseline. Also, the overall results have improved by 13% (from 3.07% to 2.67%) compared to the SMS algorithm. The performance of the most common condition of landline/landline is at a low EER of 0.56%.

To see the change in performance after every consecutive iteration of adaptation, the top bar chart in Figure 2 shows the performance of the SMS adaptation algorithm in a supervised mode. Each group of five bars pertains to a given enrollment/test condition. For example, consider the group of bars second from the left. The training data for these speaker models was from a landline and the test utterances were from a cellular environment. The black bar represents the baseline EER, before any adaptation (8.33%). After one iteration of adaptation on cellular data, the EER on cellular test data improved to 5.51%. As expected, adapting on landline data did not improve the performance on cellular data for the second and third iteration. This is because there is no information being shared between channels

	Enrollment Channel	
Testing Channel	**Cellular**	**Landline**
Baseline		
Cellular	5.27%	8.33%
Landline	7.80%	3.81%
Overall	7.14% ($\pm 0.24\%$)	
SMS		
Cellular	2.49%	4.39%
Landline	2.31%	1.10%
Overall	3.07% ($\pm 0.16\%$)	
SMS+Inverse		
Cellular	1.92%	3.71%
Landline	2.24%	0.56%
Overall	2.67% ($\pm 0.15\%$)	

Table 2. The EER for every enrollment/test channel combination. After four iterations of supervised adaptation, the SMS+Inverse algorithm improves the baseline by an average of 63% for mismatched and 75% for matched conditions.

with the SMS technique (no inverse) [2]. In the fourth and last iteration, the model was adapted on cellular data and the EER improved to 4.39%.

Now, consider the same group of five bars (second from the left) in the bottom chart of Figure 2. Note that, unlike SMS, the performance on cellular test data for the SMS+Inverse adaptation algorithm has improved after adaptation on landline data (represented by the two middle dark gray bars). The EER after four iterations of adaptation was 3.71% compared to 4.39%. This result illustrates the power of the inverse synthesis approach: all channels are improved from each new adaptation utterance regardless of the channel type of the utterance.

4. CONCLUSION

This paper presented a new approach to on-line unsupervised adaptation in speaker verification. The approach extended previous work by (1) improving performance on the enrollment handset-type when adapting on a different handset-type (e.g., improving performance on cellular when adapting on a landline office phone), (2) accomplishing this cross channel improvement without increasing the size of the speaker model after adaptation, (3) employing a count-based, parameter-dependent smoothing algorithm that emphasizes the use of mean parameters in the speaker models until sufficient adaptation data are present to accurately estimate variances, and (4) developing a new confidence-based adaptation update weight which minimizes the corrupting effects on the speaker models from impostor attacks. Two sets of experimental results were completed on a gender-balanced database of Japanese digits. The first experiments showed that the new approach can improve the performance of the verification system after initial enrollment by 61% (rel.) in an unsupervised mode under stressed impostor attack conditions. The second set of experiments showed that by mapping statistics from all new adaptation utterances through an inverse transform back to the enrollment channel, the speaker model synthesis on-line adaptation approach can be improved by an additional 13%.

[2] The performance did not remain constant as compared to the baseline condition for this example due to channel labeling errors of the verifier

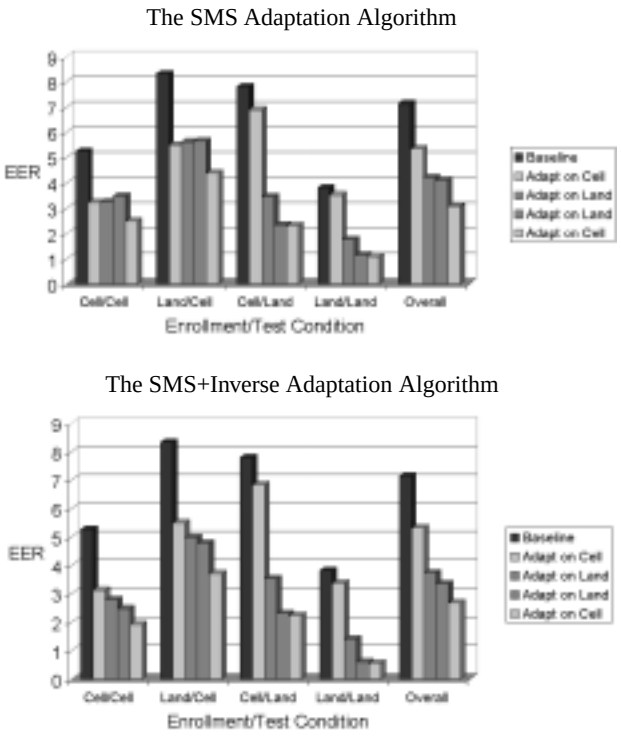

Figure 2. The figure shows EER for all training/adaptation/testing combinations of (mis)matched channels for the SMS and SMS+Inverse algorithms. Each group of five bars represent four consecutive iterations of adaptation for a given enrollment/test condition. Note that with the SMS+Inverse adaptation algorithm, when one channel is adapted, the performance on all channels improve. This is not the case for the SMS (no inverse) algorithm.

Acknowledgments

The authors would like to thank Matthieu Hebert, Dominique Genoud, and Remco Teunen for many stimulating and fruitful discussions related to this work.

REFERENCES

[1] C. Fredouille, J. Mariéthoz, C. Jaboulet, J. Hennebert, J.-F. Bonastre, C. Mokbel, and F. Bimbot. Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. *ICASSP*, Istanbul, Turkey, 2000.

[2] L.P. Heck, Y. Konig, M.K. Sönmez, and M. Weintraub. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. *Speech Communications*, 2000.

[3] L.P. Heck and M. Weintraub. Handset dependent background models for robust text-independent speaker recognition. *ICASSP*, Munich, Germany, 1997.

[4] William Mistretta and Kevin Farrell. Model adaptation methods for speaker verification. *ICASSP*, Seattle, Washington, 1998.

[5] D.A. Reynolds. HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects. *ICASSP*, Munich, Germany, 1997.

[6] Aaron E. Rosenberg, Chin-Hui Lee, and Frank K. Soong. Sub-word unit talker verification using hidden markov models. *ICASSP*, New York, NY, 1990.

[7] R. Teunen, B. Shahshahani, and L. Heck. A model-based transformational approach to robust speaker recognition. *ICSLP*, Beijing, China, 2000.