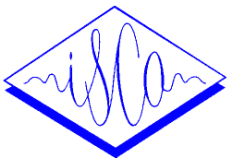
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

On-Line Unsupervised Adaptation in Speaker Verification: Confidence-Based Updates and Improved Parameter Estimation

Larry Heck, Nikki Mirghafori

Nuance Communications
heck@nuance.com, nikki@nuance.com

Abstract

This paper presents the second part of a new approach to on-line unsupervised adaptation in speaker verification. The new approach extends previous work in the literature 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 show a 61% (rel.) overall reduction in EER using the new on-line adaptation approach even with a significant impostor attack rate, and a 24% improvement in EER due to the new confidence-based adaptation scheme for those speaker models corrupted by impostor utterances.

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 sub-networks 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, 4].

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 [1]. 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[5]. 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 is the second part of our previously published work in [2], which completes the presentation of a new approach for on-line unsupervised adaptation of speaker verification models. 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 (1) degrading the performance on the enrollment handset-type when adapting on new handset-types, (2) increasing the size of the speaker model, and (3) significantly corrupting the speaker models due to impostor attacks.

In the first paper [2], we focused on the first two items above. The paper showed that the new approach not only avoided degrading the performance on the enrollment handset-type when adapting on new handset-types, but the performance actually improves across all handsets when adapting on any handset-type. This paper focuses on some additional aspects of the second item above, and on the third item. Specifically, an expanded description is presented on a "variable-rate smoothing" technique that increases the *effective* complexity and fidelity of the speaker model without increasing the storage requirements of the speaker model. In addition, this paper describes a technique to minimize the adverse affects of impostor attacks in an unsupervised adaptation setting. The new approach relies on a confidence-based weighting scheme where speaker models are adapted more or less aggressively depending on how confident the sys-

tem is on the validity of the identity claim. Expanded experimental results are presented to demonstrate the effectiveness of the variable-rate smoothing approach and the confidence-based weighting scheme.

2. Approach

The foundation for the approach used in this work is described in [2]. Summarizing, the speaker models are initialized during enrollment by adapting a handset-dependent, gender-dependent, and speaker-independent Gaussian Mixture Model (GMM) using a Bayesian approach. Speaker model synthesis (SMS) is used to synthetically create speaker models on channels not seen during enrollment. These new "synthetic" 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.

On-Line adaptation is implemented on top of SMS to update the channel-dependent speaker models (synthetic or real) in an unsupervised fashion using data observed during in-field verification attempts. An "inverse SMS" approach was developed and used to satisfy the constraints of no growth in the size of the speaker model with adaptation. The inverse SMS approach relies on the fact that the transformations between channels described in [2] 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 inverse SMS 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.

2.1. On-Line Adaptation Update Equations

To update the speaker model statistics, we use the following equations:

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

where $\Sigma_i(\underline{x})$ and $\Sigma_i(\underline{x^2})$ are the first and second-order sufficient statistics 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, $\Sigma_i(.)^{[j]}$ 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 below in (1). The terms $(\beta_\mu, \beta_\sigma, \beta_w)$ are Bayesian smoothing factors. 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.

2.2. Confidence-Based Updating of Statistics

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 a verifier score Λ, we form the adaptation weight $\mathcal{W}$ by utilizing a nonlinear function of the verifier score based on a cumulative Rayleigh distribution

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

where τ is the acceptance threshold of the verifier, and b is the Rayleigh coefficient which controls the smoothness of the function. The verification score for the work in this paper is given as

$$\Lambda(X|s) = \frac{1}{T}\sum_{t=1}^{T}\,[\,\log p(\underline{x}_t|\lambda_s) - \log p(\underline{x}_t|\lambda)\,] \qquad (2)$$

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:

$$p(\underline{x}_t|\lambda) = \sum_{i=1}^{L}\,w_i\,p(\underline{x}_t|b_i) \qquad (3)$$

with mixture weights w_i, and Gaussian densities $p(\underline{x}_t|b_i)$, parameterized by a collection of mixture weights, means, and covariances.

The Rayleigh function used to compute the confidence weight simply increases the weighting of the adaptation update as the score increases, starting with a weight of zero at the acceptance threshold and increasing to a weight of 1 as the score tends to infinity. Large values of the Rayleigh coefficient more conservatively update the speaker model statistics, whereas small values cause the adaptation to be completed more aggressively. A value of $b = 0$ indicates any utterance that scores above the acceptance threshold should have maximum contribution to updating the speaker model statistics (i.e., $\mathcal{W}(\Lambda) = 1$).

2.3. Variable Rate Smoothing

Variable rate smoothing (VRS) employs separate smoothing factors $(\beta_\mu, \beta_\sigma, \beta_w)$ 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 smoothing factors are defined as

$$\beta = \frac{n_i}{\alpha}, \qquad \text{with } \alpha = \frac{n_{i,s}}{n_{i,s} + \gamma} \tag{4}$$

where n_i and $n_{i,s}$ are the speaker independent and dependent probabilistic occupancies of the data in the i-th Gaussian, and γ is a parameter that can be varied to provide more or less smoothing between the speaker independent and dependent models.

3. Experiments

In the experiments presented, we will refer to the three new approaches developed in this paper as "SMS", "SMS+Inverse", and "SMS+Inverse+VRS", where all three use the Speaker Model Synthesis (SMS) approach but the first stores the new adaptation statistics separately with each channel, and the latter two approaches transform the new statistics back to the enrollment channel (with the difference between the last two consisting of the last using the variable-rate smoothing technique). 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 was described in detail in [2]. Summarizing, the 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. 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. Results

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, 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.

Iteration	**Supervised**	**SMS**	**SMS +Inv.**	**SMS +Inv.+VRS**
Enroll	5.67%			4.33%
Iter 1	3.01%	4.58%	4.50%	4.09%
Iter 2	2.27%	4.23%	3.76%	3.54%
Iter 3	1.98%	4.13%	3.35%	3.23%
Iter 4	1.67%	4.05%	3.15%	3.02%
Iter 5	0.94%	3.57%	2.84%	2.78%
Iter 6	0.97%	2.80%	2.32%	2.28%
Iter 7	0.91%	2.87%	2.38%	2.27%
Iter 8	-	2.77%	2.35%	2.21%

Table 1: *Improved EERs for the held-out verification test set after each consecutive iteration of unsupervised adaptation. The new "SMS", "SMS+Inverse", and "SMS+Inverse+VRS" unsupervised adaptation techniques improved the EER by 51%, 59%, and 61% over the initial enrollment (Enroll), compared with 84% for (optimal) supervised adaptation. The variable rate smoothing technique gave a 23.5% reduction in EER on the initial enrollment as compared to the baseline, with the improvement becoming less significant after several adaptation iterations.*

Table 1 shows the performance of the unsupervised adaptation approaches developed in this paper, "SMS", "SMS+Inverse", and "SMS+Inverse+VRS". For comparison purposes, the best that can be achieved with online adaptation is if the verifier had perfect knowledge of the identity of the utterance (claimant vs. impostor). This theoretical optimum improvement with supervised adaptation is 84%. The new "SMS", "SMS+Inverse", and "SMS+Inverse+VRS" unsupervised adaptation techniques improved the EER by 51%, 59%, and 61% respectively. After the initial enrollment ("Enroll"), the variable rate smoothing (VRS) technique shows a reduction of 23.5% in EER as compared to the SMS technique without VRS. In these results, the values of $(\gamma_\mu, \gamma_\sigma, \gamma_w)$ in Equation (4) were (4,32,32). As more data became available to reliably train the second-order statistics (after 8 iterations or calls), the additive effect of VRS on top of the SMS-based techniques becomes less significant.

In a second set of experiments with another test cor-

Iteration	Without Rayleigh	With Rayleigh
Baseline	5.37%	
Iter 1	4.69%	4.72%
Iter 2	4.20%	4.19%
Iter 3	3.90%	3.91%
Iter 4	3.68%	3.64%
Iter 5	3.66%	3.45%
Iter 6	3.58%	3.30%
Iter 7	3.60%	3.15%
Iter 8	3.54%	3.00%

Table 2: *After eight iterations of adaptation, the EER of the "With Rayleigh" approach was 15% better than the "Without Rayleigh" approach.*

Without Rayleigh		
EER of Corrupted Models	**EER of NOT Corrupted Models**	**Overall EER**
6.32%	1.86%	3.54%
With Rayleigh		
EER of Corrupted Models	**EER of NOT Corrupted Models**	**Overall EER**
4.83%	2.05%	3.00%

Table 3: *The confidence-based weighting approach for updating the speaker models gives more consistent results across the speaker population (4.83% vs. 6.32% EER for the corrupted models) at the cost of being somewhat more conservative in the model updates (2.05% vs. 1.86% EER for the uncorrupted models).*

pus of the same size and composition as described earlier, tests were completed to determine the effectiveness of the confidence-based weighting scheme for updating the speaker models. Table 2 compares the results of on-line adaptation without and with the new confidence-based approach ("Without and With Rayleigh"). Both experiments were run without the "Inverse" and "VRS" techniques. The acceptance threshold of the verifier, τ in Equation (1), was set to zero, and the Rayleigh coefficient, b was set to 0.4. After eight iterations of adaptation, the EERs of the "Without and With Rayleigh" approaches were 3.54% and 3.00%, representing a 15% improvement when using the confidence-based adaptation scheme.

Table 3 show a more refined breakdown of the results shown in the previous table. Speaker models were separated into two groups: those that had been "corrupted" by adapting on an impostor utterance, and those that had only adapted on true speaker utterances ("not corrupted"). The EERs for these two groups are compared without and with confidence-based weighting ("Without and With Rayleigh"). The results show that the confidence-based weighting scheme provides more consistent results across all speakers (4.83% vs. 6.32% EER for the corrupted models or 23.6% improvement) at the cost of being somewhat more conservative in the model updates (2.05% vs. 1.86% EER for the uncorrupted models).

4. Conclusion

This paper presented the second part of 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 variable-rate smoothing (VRS) algorithm that emphasizes the use of first order parameters in the speaker models until sufficient adaptation data are present to accurately estimate second-order statistics, and (4) developing a new confidence-based adaptation update weight which minimizes the corrupting effects on the speaker models from impostor attacks. Experimental results showed that the new VRS algorithm reduced the EER by 23.5% in the initial enrollment phase, and the SMS+Inverse+VRS techniques gave a combined 61% overall reduction in EER, even with a significant impostor attack rate. Finally, the confidence-based adaptation scheme gave more consistent results across the speaker population, with a and a 23.6% improvement in EER for those speaker models corrupted by impostor utterances.

5. References

[1] 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.

[2] L.P. Heck and N. Mirghafori. Online unsupervised adaptation in speaker verification. In *Proc. International Conf. Spoken Language Processing*, Bejing, China, 2000.

[3] L.P. Heck and M. Weintraub. Handset dependent background models for robust text-independent speaker recognition. *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, 1997.

[4] D.A. Reynolds. Htimit and llhdb: Speech corpora for the study of handset transducer effects. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 1535–1538, 1997.

[5] Aaron E. Rosenberg, Chin-Hui Lee, and Frank K. Soong. Sub-word unit talker verification using hidden markov models. In *Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing*, pages 269–272, New York, NY, 1990.