SPEAKER VERIFICATION WITH DATA FUSION AND MODEL ADAPTATION

Kevin R. Farrell

T-NETIX Incorporated, SpeakEZ Division
9100 Greenwood Plaza Boulevard
Englewood, Colorado 80112

Kevin.Farrell@speakezinc.com

ABSTRACT

This paper presents methods for adapting models in a data fusion-based speaker verification system. The models that are used in the data fusion system are the neural tree network (NTN), dynamic time warping (DTW), and hidden Markov model (HMM). The models provide information based on discriminant information, distortion measurements, and probabilistic evaluation, respectively. The parameters of these models are updated during the adaptation process using verification data. This allows the models to track changes in the users voice over time and additionally allows the technology to supplement the typically limited data obtained at enrollment. Experiments are performed on voice data collected within landline telephony, wireless telephony, and multimedia environments. Additionally, the adaptation algorithms are evaluated for both cases where the data is known to come from the correct user (supervised) and not known to come from the correct user (unsupervised). For the case where the adaptation data is not known to come from the correct user, threshold criteria is used for determining if the adaptation should occur or not. The adaptation leads to a 20% relative reduction on the equal error rate for the unsupervised scenario and a 40% relative reduction in equal error rate for the supervised scenario.

1. INTRODUCTION

Speaker verification consists of determining whether or not a voice sample provides sufficient match to that of a claimed identity. Speaker verification has matured to the point where the technology has been commercially deployed in corrections, telecommunications, and finance. One critical aspect attributing to the success of speaker verification systems is the robustness to intersession variability and aging. Intersession variability refers to differences of the recorded voice from one day to the next, be they due to subtle changes in speaking patterns or variations in the communications channel. For substantial periods of time, such as several months to years, the effects of voice aging can also degrade system performance. Whereas the spectral variation of a speaker may be small when measured over a several week period, as time passes this variance will grow [1].

Due to the effects of intersession variability and model aging, a model created solely from the enrollment utterances of a single session can experience problems for some users over extended periods of time. This problem is alleviated by creating a model with data from several enrollment sessions, however, this can pose a burden on the user. Another option that achieves intersession robustness while not posing a burden to the user is to use model adaptation.

Model Adaptation consists of adjusting model parameters based on verification data encountered after enrollment. Model adaptation methods have been extensively researched in the field of speech recognition. Some of these advances have been extended to applications in speaker recognition. For example, methods have been proposed to adapt speaker-independent hidden Markov models (HMMs) to create speaker-dependent models that could be used for speaker verification [2]. Methods have also been evaluated for adapting dynamic time warping (DTW) approaches by averaging the new observation with the original template [3]. Methods have also been explored for adapting Gaussian mixture models (GMMs) and neural tree networks (NTNs) [4]. This work, however, assumed that the true label of the verification data is known, i.e., the adaptation is supervised. The case where the true label of the verification data is unknown, i.e., unsupervised adaptation, is much more pertinent to speaker verification applications. In [5], a study was performed to evaluate the adaptation of GMMs using unsupervised adaptation. The current work evaluates a data fusion-based speaker verification system that uses a combination of HMM, DTW, and NTN models where the adaptation is supervised and unsupervised.
2. SPEAKER VERIFICATION SYSTEM

The speaker verification system evaluated in this paper uses several different modeling techniques to contribute to the overall decision. The overall system is illustrated in Figure 1. Digitized speech data is first processed with feature extraction to yield pole-filtered cepstral coefficients [6]. The feature data is then applied to a DTW template created during enrollment that aligns the spoken password to the enrollment template. A byproduct of this alignment is a DTW score that provides a distortion measurement of speaker match. The aligned features are then segmented using a blind segmentation algorithm [7] so that they can be subsequently processed by the NTN at a sub-word level [8]. The blind segmentation algorithm utilizes a HMM and outputs a statistical level of speaker match. The sub-word NTN models output a discriminant measure of speaker match. As these metrics for speaker verification are based on different criteria, their errors tend to be uncorrelated [9] and their combination outperforms that of the models used individually. The three modeling methods used in the proposed speaker verification system can all be updated using model adaptation.

3. MODEL ADAPTATION

The DTW, HMM, and NTN models as illustrated in Figure 1 can all be updated with model adaptation. The details of adapting each model component are described in the following subsections.

3.1. DTW Adaptation

The dynamic time warping (DTW) template is a speaker verification modeling approach that provides a distortion measurement between the feature data average obtained during enrollment and feature data obtained at verification. The DTW template does not use statistical parameters as does the HMM, but however relies more on the feature data itself to provide a measurement of similarity based on the distortion or distance. The DTW template is trained by first warping the feature data all onto the feature set resulting from a selected utterance so that the resulting feature sets are all the same length. These warped features are then averaged to produce a new feature set that is stored in the DTW model. Adaptation of the DTW template simply consists of warping new feature sets onto the feature set within the DTW model and then computing a new weighted average to use as the adapted template.

The DTW template will now reflect the additional speaker information from the verification utterance. However, as the DTW template is updated to reflect more and more speaker data, the distribution of future DTW scores will shift. Specifically, the mean for true speaker scores will increase. This will cause the threshold for a specific operating point to drift and require future adjustments. This drift can be prevented by adapting a scaling term in the DTW normalization. Currently, the DTW score is normalized to a scale of 0 to 1 for subsequent combination with other models by applying a scaled negative distortion to an exponential, i.e., $y = e^{-\alpha x^2}$. Here, the scaling parameter $\alpha$ can also be adapted such that the resulting distributions are similar to those prior to adaptation.

3.2. HMM Adaptation

The hidden Markov model (HMM) provides a statistical model of the feature data obtained from an enrollment. This statistical data consists of a sub-word decomposition of the features where the multivariate mean and variance is measured for each sub-word and additionally the transition probability between sub-words. Given an utterance to be used for adaptation, the method used for HMM adaptation is as follows.

A new HMM is first trained using the provided feature data. On a parameter-by-parameter basis, each of the original HMM parameters is scaled by adding in the new HMM parameter with an adaptation rate. For example, for each transition probability $T[i]$, use $T_{new}[i] = (1.0f - \text{adaptation rate}) \times T_{original}[i] + (\text{adaptation rate}) \times T_{adapt}[i]$, where adaptation rate is a scalar value between 0.0 and
where $T_{\text{new}}[]$ is the array of adapted transition probabilities, $T_{\text{original}}[]$ is the array of original values, and $T_{\text{adapt}}[]$ is the array of transition probabilities from the HMM trained on the adaptation utterance. An adaptation rate of 1.0 will correspond to the case where the new HMM consists solely of information from the adaptation utterance and an adaptation rate of 0.0 corresponds to no adaptation. This process is repeated for the mean and variance parameters.

As with the DTW model, adapting the HMM will also cause a shift in the distribution of future verification scores. Adaptation will cause the speaker and imposter scores to be higher for a given model and require the threshold to be updated to yield the same operating point. This issue is addressed in a similar manner to that of the DTW model. Basically, the HMM output is scaled to be on a scale of 0 to 1 by applying it to a sigmoid function, i.e., $y = 1 / (1 + e^{-\alpha x + \beta})$. The $\beta$ term in the sigmoid can be also adapted to negate the shift in scores that will result from the adapted HMM. The $\alpha$ value does not have to be changed.

### 3.3. NTN Adaptation

The neural tree network (NTN) provides a discriminative measurement based on enrollment data from a given user and additional voice data from other users. This discriminative measurement provides information as to which portions of a person’s spoken password are unique to that person and which parts are more confusable with speech from other people. The primary information within the NTN used to make this measurement is the counts of true user observations and imposter observations at each leaf of the NTN. By adapting these counts based on new observations, the NTN performance can be improved.

The adaptation process for the NTN is illustrated in Figure 2. Each feature vector for the adaptation utterance is applied to the NTN and the probabilities at the leaves are updated based on the new relative frequency estimates. For example, in the right-most leaf there are two speaker observations and two imposter observations, so the leaf probability is 0.5. When the new feature vector of the adaptation utterance reaches this leaf, there are a total of three speaker observations and two imposter observations, so the new speaker probability at this leaf is 0.6.

As with the DTW and HMM models, the NTN adaptation will also cause a shift in scores. The NTN score is scaled with a forgetting factor to compensate for this shift in score distribution. This successfully maintains the operating point.

### 4. DATABASE

The adaptation algorithms were evaluated with ten separate databases. These databases are broken down as one wireless database, four landline databases, and five multimedia databases. For the multimedia databases, the input data was collected with a PC microphone. The statistics for each of the database categories are summarized in the following table.

<table>
<thead>
<tr>
<th>Database</th>
<th># Users</th>
<th>True trials</th>
<th>Imposter trials</th>
<th>Total trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landline</td>
<td>80</td>
<td>435</td>
<td>12,099</td>
<td>12,534</td>
</tr>
<tr>
<td>Wireless</td>
<td>56</td>
<td>273</td>
<td>6,825</td>
<td>7,098</td>
</tr>
<tr>
<td>Multimedia</td>
<td>100</td>
<td>545</td>
<td>4,846</td>
<td>5,391</td>
</tr>
<tr>
<td>All</td>
<td>236</td>
<td>1,253</td>
<td>23,770</td>
<td>25,023</td>
</tr>
</tbody>
</table>

The wireless and landline data collections were both performed using 8 kilohertz sampling with 8 bit u-law samples. The multimedia data collection used 16 bit linear samples with a 11.025 kilohertz sampling rate. In the above table, the row labeled “All” corresponds to the case where all of the verification scores from all three-database categories are concatenated into a single file and the statistics are measured from this data. For all imposter trials, the correct password was spoken.

### 5. EXPERIMENTAL RESULTS

The databases described above were used to evaluate the model adaptation algorithms. The results are reported for the baseline performance along with the adaptation performance. The adaptation performance is broken down into the unsupervised and supervised categories. The “unlabeled” category corresponds to the “real-world” case where it is not known if the data comes from the correct user or an imposter and a threshold must be applied. The condition for adaptation here is that the verification score is above 0.65 on a scale of 0 to 1. The
supervised category corresponds to the case where the data is known to come from the correct user. This performance metric provides more of a "best-case" scenario.

<table>
<thead>
<tr>
<th>Database</th>
<th>Baseline (EER)</th>
<th>Unsupervised adaptation (EER)</th>
<th>Supervised adaptation (EER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landline</td>
<td>1.23</td>
<td>1.15</td>
<td>0.65</td>
</tr>
<tr>
<td>Wireless</td>
<td>5.87</td>
<td>3.97</td>
<td>3.23</td>
</tr>
<tr>
<td>Multimedia</td>
<td>1.18</td>
<td>1.08</td>
<td>1.05</td>
</tr>
<tr>
<td>All</td>
<td>2.73</td>
<td>2.13</td>
<td>1.57</td>
</tr>
</tbody>
</table>

From the above table it is shown that the overall EER is reduced by 22% for the unlabeled adaptation and 42% for the labeled adaptation. This represents a sizeable reduction in error rate.

Another point to be made is that the adaptation results from our benchmark databases are measured from a scenario consisting mainly of imposter attempts. For example, the “All” case corresponds to 25,023 verification attempts that are composed of 23,770 from imposters and 1,253 from true users. As most fielded applications would probably have a reversal of this distribution of true user and imposter attempts, it is likely that larger reductions in error rate could be expected.

4. CONCLUSIONS

Model adaptation was evaluated for a data fusion-based speaker verification system. The system used three separate modeling approaches, namely the HMM, DTW, and NTN in comprising the final speaker score. All of these models were adapted with the methods discussed in this paper. The unlabeled and labeled adaptation methods were able to reduce the equal error rate by 22% and 42%, respectively.

5. REFERENCES