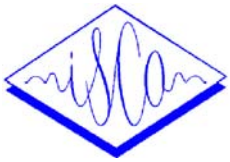
Unsupervised Online Adaptation for Speaker Verification over the Telephone

Claude Barras, Sylvain Meignier, Jean-Luc Gauvain



ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Unsupervised Online Adaptation for Speaker Verification over the Telephone

Claude Barras, Sylvain Meignier, Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
{barras,meignier,gauvain}@limsi.fr

Abstract

This paper presents experiments of unsupervised adaptation for a speaker detection system. The system used is a standard speaker verification system based on cepstral features and Gaussian mixture models. Experiments were performed on cellular speech data taken from the NIST 2002 speaker detection evaluation. There was a total of about 30.000 trials involving 330 target speakers and more than 90% of impostor trials. Unsupervised adaptation significantly increases the system accuracy, with a reduction of the minimal detection cost function (DCF) from 0.33 for the baseline system to 0.25 with unsupervised online adaptation. Two incremental adaptation modes were tested, either by using a fixed decision threshold for adaptation, or by using the a posteriori probability of the true target for weighting the adaptation. Both methods provide similar results in the best configurations, but the latter is less sensitive to the actual threshold value.

1. Introduction

Automatic speaker verification systems generally have to deal with limited enrollment data per speaker, which limits the accuracy of the system. Not only the amount of data is important, but also the diversity of acoustic and channel conditions. For a fixed amount of speech, multiple enrollment sessions improve significantly the performance. Furthermore, voice is known to evolve over time; taking into account recent sessions is needed to couterbalance the model aging (see e.g. [1]). For real applications of speaker verification, unsupervised adaptation of the speaker models is thus a very useful feature, but the risk of corrupting the models with impostor data must be carefully controlled.

In this paper we report on unsupervised adaptation of a speaker verification system. Experiments are carried out on cellular speech data from the NIST one-speaker detection task [6]. The baseline speaker verification system is a standard text-independent Gaussian-mixture models (GMM) system [2]. A specificity of the test set used is the very high proportion of impostor trials (above 90%). This departs from another situation already reported for the online adaptation of a speaker verification system, where fewer impostors than true speakers are expected in normal exploitation [4, 5].

In the next section we describe the experimental conditions and the baseline system without adaptation. In Section 3 we review the supervised and unsupervised adaptation protocols that were tested. The experimental results are presented and discussed in Section 4.

2. Experimental setup

In this section, we describe the NIST one-speaker detection task, the corpora used to carried out the experiments, and the baseline speaker verification system.

2.1. Corpus and task

The speaker recognition experiments were conducted on cellular telephone conversational speech from the Switchboard corpus. This data was selected by NIST for the 2002 one-speaker detection task [3]. Given a speech segment of about 30 seconds, the goal is to to decide whether this segment was spoken by a specific target speaker or not. For each of 330 target speakers (139 males and 191 females), two minutes of untranscribed, concatenated speech is available as enrollment data for training the target model. Overall 2679 test segments (1085 males and 1594 females), lasting between 15 and 45 seconds, as defined by NIST for the primary test condition, were selected for these experiments. For each of the 330 target speakers, between 74 and 110 tests are conducted, with a mean of about 89 trials per target, and a total of about 30.000 trials. There are up to 17 true speaker trials per target speaker, with a mean of about 7 true speaker trials, and for 15 targets there are no true speaker trials at all. In the standard, unsupervised protocol, each trial has to be performed independently of the other, ignoring the score of all other trials. The gender of the target speaker is known and only gender-matching trials are considered. The proportion of impostor trials is 92.3%.

We made use of the cellular data from the NIST 2001 evaluation in order to train background models or impostor models for score normalization, and estimate a priori distribution of impostor and true target scores. This data includes files from 60 development speakers (2 minutes of speech for each of 38 males and 22 females) which are used to train the background models, and files from 174 target speakers (2 minutes of speech for each of 74 males and 100 females) used as enrollment for impostor target models.

2.2. Baseline system

Acoustic features are extracted from the speech signal every 10ms using a 30ms window. The feature vector estimated on the 0-3.8kHz bandwidth is comprised of 15 MEL-PLP cepstrum coefficients, 15 delta coefficients plus the delta energy, for a total of 31 features. Acoustic features are normalized using feature warping [7] over a 3 seconds sliding window before computing the delta coefficients. Feature warping consists in mapping the observed cepstral feature distribution to a normal distribution. It was shown to outperform the classical Cepstral Mean Substraction approach for speaker recognition tasks.

For each target speaker, a speaker-specific GMM with diag-

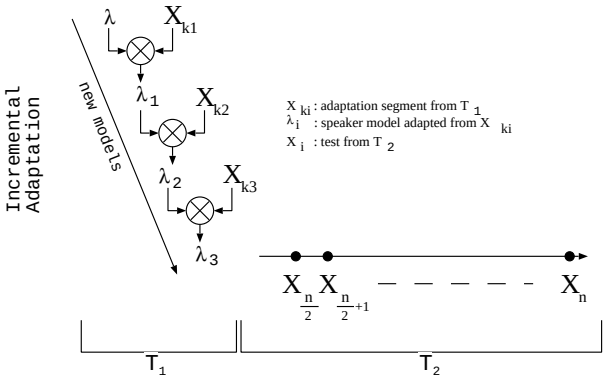


Figure 1: *Protocol used for the offline adaptation of the system. Some test segments from T_1 were used for incremental adaptation of the target model λ, which was then used for scoring T_2 trials.*

onal covariance matrices was trained via maximum a posteriori (MAP) adaptation [8] of the Gaussian means of the matching gender background model using 5 iterations of the EM algorithm, with a prior weight $\tau = 10$. Each of the two gender-dependent background models includes 1024 Gaussians. These two models were trained on a total of about 2 hours of data from the 60 development speakers.

Each verification trial is comprised of a test segment and a target speaker. The test segment is scored against the target model and a cohort of gender-matching impostor models, ignoring low energy frames (about 10%). According to T-norm score distribution scaling [9], for a given test segment X and a target model λ, the decision score is

$$S(X,\lambda) = \frac{\log f'(X|\lambda) - \mu_X}{\sigma_X}$$

where $f'(X|\lambda)$ is the normalized likelihood of the speech segment (of length $L(X)$) for a given model λ, i.e., $f'(X|\lambda) = f(X|\lambda)^{1/L(X)}$, and is scaled according to the mean μ_X and standard deviation σ_X of the likelihoods of the test segment given the gender-matching impostor cohort models.

2.3. Performance measure

The primary performance measure for the NIST speaker detection task is the detection cost function (DCF) defined as a weighted sum of missed detection and false alarm probabilities (see [3]) $DCF = P_{Miss} + 9.9 \times P_{FalseAlarm}$. For our experiments, we report the minimal DCF value obtained a posteriori for the best possible detection threshold, and the equal-error rate (EER). For unsupervised online adaptation, we also give the ratio of both types of adaptation errors: false adaptation using an impostor test segment (FAd) and missing the adaptation for a true speaker test segment (MAd).

3. Adaptation protocols

The evaluation of the performance of an online unsupervised adaptation system is complicated by its inherent non-stationarity, the goal being to improve the system performance along its use. Before testing the online unsupervised situation, it was decided to calibrate the performance of such a system by testing it under more controlled conditions: supervised offline and supervised online mode.

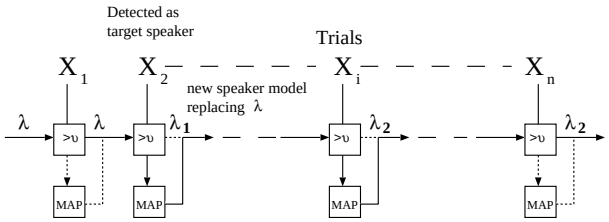


Figure 2: *Incremental protocol used for the online adaptation of the system.*

3.1. Offline adaptation

For the offline adaptation experiments, all trials involving a given target were shuffled in random order, and split in two halfs, T_1 and T_2. T_1 was used for the adaptation of the target models, T_2 was used for assessing the performance of the model without further modification. Two protocols for supervised offline adaptation were used:

- Oracle: One to four true speaker tests were selected from T_1 for supervised adaptation of the target model. Due to the limited and varied amount of true target tests per target model, targets for which the count of true test segments was not available were discarded from the experiments[1].
- Nearest impostors: for each target, the impostor test segments in T_1 best matching the target model were used for adaptation of the target model. One to four impostor test segments were selected. The goal of this protocol was to assess the corruption of the target models brought by the most confusable impostor trials.

In all situations, the adaptation consisted in the MAP adaptation of the Gaussian means only with a fixed prior weight τ. For adaptation using several test segments, the segments were presented in sequence, and the adapted target replaced the current one after each adaptation (cf. Figure 1).

3.2. Online adaptation

For the online adaptation experiments, all trials involving a given target λ were shuffled in random order, and processed in sequence. For each segment X, the decision score $S(X,\lambda)$ is computed. According to this score, a decision is made to adapt or not the target model using the test segment. The adapted model then replaces the initial model for the subsequent trials (cf. Figure 2). Obviously, the adaptation threshold ν may be different from the decision threshold, in order to control the corruption of the target models. For the simplicity of the application, we chose to test here incremental adaptation, instead of re-adaption from the background model using all speech data attributed to the speaker.

Supervised and unsupervised online adaptation were compared. In the supervised mode, the adaptation is done for all and only for the true target trials. This is the online counterpart of the oracle offline adaptation described in the previous section, and provides an upper bound for the performances of an unsupervised online adaptation system.

Unsupervised online adaptation was first tested using a fixed decision threshold ν and a fixed adaptation weight τ. However, this binary decision mode may be too rigid, and a

[1] Contrastive experiments without adaptation were made with the same restricted target set to insure that the results remained comparable to the default condition.

Nb. adaptation	*min. DCF*	*EER (%)*
Baseline (on T_2)	0.333	8.9
1	0.234	6.1
2	0.187	4.6
3	0.164	3.9
4	0.145	3.6

Table 1: *System performance of the oracle system with an increased count of true segment adaptations, using a MAP adaptation weight $\tau = 10$.*

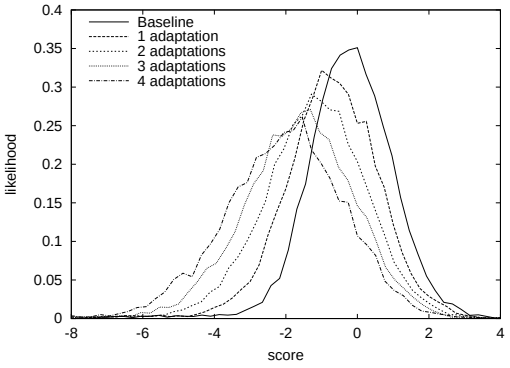


Figure 3: *Shift of the impostor scores distribution for the oracle system, with an increased count of true segment adaptations.*

probabilized adaptation mode was also tested. Given the a priori probability of the true speaker and using the conditional distribution of true speaker and impostor scores estimated on the development database, it is possible to estimate the a posteriori probability of the true target $P(\lambda|X)$ given the score $S(X, \lambda)$. Given a test segment $X = \{x_t\}$, the EM MAP adaptation of the mean μ_k of the k^{th} Gaussian of λ is performed by weighting the contribution of the new segment with the a posteriori probability as follows:

$$\mu'_k = \frac{\tau\mu_k + P(\lambda|X)\sum_t \gamma_{kt} x_t}{\tau + P(\lambda|X)\sum_t \gamma_{kt}}$$

where γ_{kt} is the posterior probability of the Gaussian k for the frame x_t.

4. Experimental results

The baseline system has a minimal DCF of 0.330 and an EER of 8.3% using all primary condition trials of the NIST 2002 one-speaker limited data evaluation. This is similar to the performances we already reported on this task [10].

4.1. Offline adaptation results

For testing offline adaptation, the data set was randomly split in two parts T_1 and T_2. On the T_2 part, the baseline system has a minimal DCF of 0.333 and an EER of 8.9%, showing only a slight bias from the baseline system due to the random split.

We tested the oracle, supervised offline protocol, with an increased count of true speaker test segment for adaptation ranging from 1 to 4, and using a MAP adaptation weight $\tau = 10$. Results in Table 1 show that a reduction above 50% of both minimal DCF and EER can be reached with at least 3 adaptations. Performances seem to further improve beyond 3 segments, but experiments were limited by the count of true speaker segments per target speaker in the corpus. The adaptation affects both impostor and true speaker scores. The distribution of impostor scores is shifted towards lower scores (cf. Figure 3), and the distribution of true speaker scores is only slightly shifted towards higher scores. The shift of impostor scores may be related to an increased distance between the adapted models and the unchanged cohort models used for the T-norm.

Nb. adaptation	*min. DCF*	*EER (%)*
Baseline (on T_2)	0.333	8.9
1	0.407	10.5
2	0.479	11.9
3	0.547	12.5
4	0.585	13.9

Table 2: *System performance degradation with an increased number of nearest impostor segment adaptations, using a MAP adaptation weight $\tau = 10$.*

τ	*min. DCF*	*EER (%)*
Baseline	0.330	8.3
8	0.223	5.6
10	0.189	4.9
12	**0.177**	**4.7**
14	0.177	4.7

Table 3: *Performances of the online supervised adaptation of the system, as a funtion of the MAP adaptation weight τ.*

As a contrastive experiment, offline adaptation using 1 to 4 nearest impostor segments was also performed, keeping the same MAP adaptation weight $\tau = 10$. We observe in Table 2 a performance degradation of about 20% for a single misadaptation, quickly exceeding 50% of relative degradation with several misadaptations.

4.2. Online adaptation results

Online, supervised adaptation was evaluated for different values of τ. More than 40% relative reduction of the minimal DCF and EER can be reached for $\tau = 12$ (cf Table 3). This score covers the global performance of the system for all the trials, using increasingly adapted target models. This must be seen as an upper bound for the unsupervised adaptation protocol.

Varying the decision threshold and the adaptation weight for a binary unsupervised adaptation, the best result observed was a minimal DCF of 0.250 and an EER of 6.6, i.e. a relative improvement of 20-25% over the baseline system (cf Table 4). Looking at the adaptation errors, we see in this situation that the MAd rate is about 25%, with a very low FAd rate of 0.35%. Taking into account the a priori probability of true targets under 8%, it follows that only 5% of model adaptations were per-

τ	ν	*min. DCF*	*EER (%)*	*MAd (%)*	*FAd (%)*
8	3	0.279	7.2	30.6	0.18
10	3	0.258	6.5	25.9	0.26
12	3	**0.250**	**6.6**	**23.4**	**0.35**
14	3	0.255	6.7	21.9	0.48
12	2	0.258	6.8	11.9	2.11
12	2.5	0.251	6.8	16.9	0.96
12	3	0.250	6.6	23.4	0.35
12	3.5	0.261	6.9	31.1	0.14

Table 4: *Performances of the online unsupervised adaptation of the system along with the missed adaptation (MAd) and false adaptation (FAd) errors, as a function of the MAP adaptation weight τ and the adaptation threshold ν.*

max. τ	*min. DCF*	*EER (%)*
Baseline	0.330	8.3
8	0.271	6.8
10	0.252	6.7
12	0.252	6.8
14	0.261	6.6

Table 5: *Performances of the online unsupervised adaptation of the system, for a soft adaptation, as a function of the MAP adaptation weight* τ.

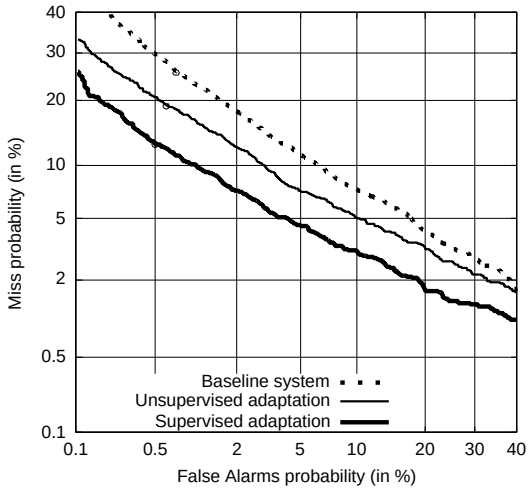


Figure 4: *DET curves for the baseline system and the sytems with online supervised and online unsupervised adaptation. Circles are drawn at minimal DCF operating point.*

formed using an impostor segment. Target model corruption is thus very limited; more precisely, 10% of target models were corrupted with one impostor segment, 2.5% with two segments, and none with more than two. Missed adaptation were less balanced: all true speaker instances were used for adaptation for 30% of target speakers, but more than half of the true trials were missed for adaptation for also about 30% of target speakers. All speakers do not benefit equally from the unsupervised adaptation. In average, about 160 seconds of true speaker speech were added to the 2 minutes of initial enrollment, i.e. more than doubling the training data.

The unsupervised online soft adaptation, taking into account the a posteriori probability of the target for weighting the adaptation, provides very similar results (cf Table 5). Distributions of impostor scores and of true speaker scores were estimated by histograms on the NIST 2001 cellular data, leading to the estimation of the probability $P(\lambda|X)$ given the score $S(X,\lambda)$. For efficiency, a minimal adaptation threshold was still used for low values of $P(\lambda|X)$: the audio segments for which $P(\lambda|X) < 0.1$ were not used for adaptation. However the system is much less sensitive to this threshold than to the hard decision threshold ν, this is an advantage for practical use.

The Detection Error Tradeoff (DET) curves for both the online supervised and unsupervised systems are shown along with the DET curve of the baseline system (cf Table 4). The unsupervised online adaptation of the system seems to provide performance balanced along the DET curve between the baseline and the supervised adaptation setup, except for low miss detections where it gets closer of the baseline system: the unsupervised adaptation did not improve performances for the target speakers with the lowest detection scores.

5. Conclusion

Online unsupervised adaptation of a speaker verification system was evaluated in the context of a speaker detection task, where impostor trials are the majority. It was shown that, despite the low proportion of true speaker trials, the performances of the system can be significantly increased, with a reduction of the minimal DCF from 0.33 to 0.25 on cellular speech data taken from the NIST 2002 one speaker detection task. This result is quite encouraging. There is in fact very few misadaptations, thus limiting the corruption of the target models. By contrast, a minimal DCF of about 0.18 could be reached by using an oracle supervised adaptation.

Two incremental adaptation modes were tested, either by using a fixed decision threshold for adaptation, or by using the a posteriori probability of the true target for weighting the adaptation. Both provide similar results in the best configuration, but the latter takes into account the whole distribution of the impostor and true speaker trial scores. It is thus less sensitive to the actual threshold value.

Experiments were limited by the amount of trials per target speaker in the database, which was not initially designed for an adaptation task. Further improvements should probably be observed when increasing the number of trials. Since a shift was observed in the distribution of the scores when adapting the models, reduction of this phenomenon may further increase the performance. At last, some target models showing lower true speaker trial scores did not benefit of the unsupervised adaptation; speaker-specific score normalization [5] may be needed for these cases.

6. References

[1] L. Lamel and J.L. Gauvain, "Speaker verification over the telephone," *Speech Communication*, vol. 31, pp. 141–154, 2000.

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

[3] "The NIST year 2002 speaker recognition evaluation plan," 2002, `http://www.nist.gov/speech/tests/spk/2002/doc/`.

[4] C. Fredouille, J. Mariethoz, C. Jaboulet, J. Hennebert, J.-F. Bonnastre, C. Mokbel and F. Bimbot, "Behaviour of a bayesian adaptation method for incremental enrollment in speaker verification," in *Proc. ICASSP*, 2000.

[5] N. Mirghafori and L. Heck, "An adaptive speaker verification system with speaker dependent a priori decision thresholds," in *Proc. ICSLP*, 2002.

[6] G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, June 2001.

[8] J.-L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[10] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP*, 2003.