Vol.2010-SLP-84 No.12
2010/12/21

Inter-speaker weighted MAP adaptation for GMM-supervector speaker recognition

Marc Ferràs ,†1 Koichi Shinoda †1
and Sadaoki Furui†1

Gaussian Mixture Models (GMM) are ubiquitously used in state-of-the-art speaker recognition systems. The popular GMM-SVM paradigm uses Maximum A Posteriori (MAP) speaker-adapted GMM models by stacking the mean vectors into a supervector that is fed into a Support Vector Machine classifier. In this paper, we modify the standard relevance MAP algorithm to better fit the speaker recognition task. We propose to emphasize the adaptation of the Gaussian mixtures according to the inter-speaker variability exhibited on a training set, thus accounting for both the occupation count and the speaker discrimination ability during adaptation. We evaluate our proposal on a relevance MAP based GMM-SVM system using a large telephone speech corpus such as the one provided in the 2006 NIST Speaker Recognition Evaluation. We show that despite its simplicity this technique is effective.

1. Introduction

Currently, adaptation of Gaussian Mixture Models (GMM) to speech data from a speaker is probably the most popular framework used for modeling speech in text-independent speaker recognition. Many of the adaptation techniques in the literature are founded on the Maximum a Posteriori (MAP) criterion which optimally combines prior knowledge with new data under the Bayesian framework. Relevance MAP[4] was first introduced in speaker recognition in[9], where a Universal Background Model (UBM) representing the acoustic space for a large number of speakers was used as prior knowledge for the adaptation process. Some techniques constrain adaptation onto a feature subspace, thus reducing the number of estimated parameters while adapting observed and non-observed Gaussian components. In the last years, several new techniques based on this framework have been proposed to disentangle different sources of variation in acoustic modeling. In this line, Joint Factor Analysis (JFA)[5] is a notable technique that allows to adapt speaker and session components separately, leading to significant improvements of speaker model robustness against inter-session variation.

Relevance MAP does not implement any session compensation scheme but it performs direct adaptation to the data. In this paper, we aim at improving the robustness against session variation of relevance MAP by weighting the relevance factor used for adaptation. For each Gaussian, we use a measure of inter-speaker variability to derive the weights that speed up the adaptation of the Gaussians with large inter-session variability and slow down adaptation of the Gaussians with small inter-session variability. A similar idea has been successfully used in[7] for VAD-based hypothesis decoding in speech recognition.

This paper is structured as follows: Section 2 describes relevance MAP adaptation. Section 3 presents the proposed relevance factor weighting technique used to improve robustness against inter-session variation. In Section 4, we detail the GMM-SVM speaker verification system used in the experiments, whose experimental protocol is described in Section 5. The experiments and the results are shown and discussed in Section 6. Conclusions are given in Section 7.

2. MAP adaptation

Relevance Maximum A Posteriori (MAP)[4] is a technique used to adapt the parameters of an Gaussian Mixture Model (GMM) to some speech data. It aims at finding direct parameter estimates that optimize the posteriori probability after adaptation, given a model representing prior knowledge. It exhibits the desirable property of asymptotically converging to Maximum Likelihood training estimates as more adaptation data is provided. Prior knowledge is captured by the prior distributions. i.e. the Gaussian mixture of the non-adapted model.

Given some adaptation data with observation vectors $\mathbf{x}_t$, with $1 \leq t \leq T$ and a GMM with M Gaussians with $1 \leq m \leq M$, the expected vector $\tilde{\boldsymbol{\mu}}_m$ given the data only is computed as

$$\tilde{\boldsymbol{\mu}}_m = \frac{\sum_{t=1}^{T} \gamma_{mt} \mathbf{x}_{mt}}{\sum_{t=1}^{T} \gamma_{mt}}, \tag{1}$$

where γ_{mt} is the occupation probability for Gaussian m and time t, that is,

†1 Tokyo Institute of Technology

$$\gamma_{mt} = \frac{\lambda_m \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_m; \boldsymbol{\Sigma}_m)}{\sum_{j=1}^{M} \lambda_j \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_j; \boldsymbol{\Sigma}_j)} \tag{2}$$

The new mean vector MAP estimates for each Gaussian m are then obtained as a linear interpolation of the prior and new estimates, as

$$\hat{\boldsymbol{\mu}}_m = \frac{N_m \tilde{\boldsymbol{\mu}}_m + \tau \boldsymbol{\mu}_m}{N_m + \tau}, \tag{3}$$

where $N_m = \sum_{t=1}^{T} \gamma_{mt}$, i.e. the average number of frames assigned to Gaussian m and τ is the MAP relevance factor that balances the priors and new estimates.

From (3) we can see how the updated mean vectors $\hat{\boldsymbol{\mu}}_m$ result from a linear interpolation between old, $\boldsymbol{\mu}_m$, and new, $\tilde{\boldsymbol{\mu}}_m$, estimates.

3. Relevance factor weighting

In this paper, we propose a technique to improve the robustness of the features to session mismatch between training and testing. In MAP adaptation, the sensitivity depends only on the occupation count, which is necessary to capture acoustic information but does not involve any speaker-related information. We believe adaptation can be improved by including a correction factor in the interpolation formula (3) used during adaptation. The proposed modification corrects the relevance factor τ according to a measure of inter-speaker distance, namely the between-speaker to within-speaker variance ratio of the occupation counts, i.e. $\sigma^{BS}_{\gamma_m}/\sigma^{WS}_{\gamma_m}$. Therefore, for Gaussian m the new relevance factor τ_m becomes

$$\tau_m = \frac{\tau}{\beta_m} \quad \text{with } \beta_m = \frac{\sigma^{BS}_{\gamma_m}}{\sigma^{WS}_{\gamma_m}} \tag{4}$$

where

$$\sigma^{WS}_{\gamma_m} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{N_s} \sum_{n=1}^{N_s} (\gamma_m^{n,s} - \overline{\gamma}_m^s)^2 \tag{5}$$

and

$$\sigma^{BS}_{\gamma_m} = \frac{1}{S} \sum_{s=1}^{S} (\overline{\gamma}_m^s - \overline{\gamma}_m)^2 \tag{6}$$

with $\gamma_m^{n,s}$ being the occupation count for session n of speaker s in the training database, $\overline{\gamma}_m^s$ the average count for speaker s and $\overline{\gamma}_m$ the average count for all speakers and sessions.

We further process β_m before applying the correction. A gain factor C is included to tune the amount of correction around the average of β_m as

$$\beta'_m = C(\beta_m - \overline{\beta}) + \overline{\beta} \tag{7}$$

where $\overline{\beta}$ is the average of β_m over all models in the training data. Finally, we apply the normalization $\beta''_m = \beta'_m/\overline{\beta}$ so that a correction factor of 1 is obtained when $\beta'_m = \overline{\beta}$.

Compensating the relevance factor results in a slight correction of the adapted model only. Note, for instance, that the actual statistics used for adaptation are still relevance MAP Baum-Welch statistics which, in principle, do not take any advantage of multiple sessions. More sophisticated techniques such as JFA can give explicit estimates for the speaker and session contributions of the adapted model.

4. GMM-SVM system description

In this paper, we experiment with a GMM-SVM speaker verification system using relevance MAP parameter estimates. A Universal Background Model (UBM) is trained off-line using data from many speakers, thus modeling the speaker-independent acoustic space. The UBM is adapted to the speech segments of interest and the mean-adapted vector parameters are stacked into supervector that are classified with Support Vector Machines (SVM).

We use Nuisance Attribute Projection (NAP) for inter-session variability compensation, which has shown to be effective for GMM-SVM systems[3]. NAP is a data-driven technique that projects out the supervector subspace with maximum inter-session variability. In the training phase, we perform a Principal Component Analysis (PCA) of the inter-session scatter matrix estimated on a database of GMM supervectors with multiple sessions per speaker. Its k eigenvectors with largest eigenvalues, $\mathbf{E} = (\mathbf{e_1}, \ldots, \mathbf{e_k})$, representing the subspace with maximum inter-session variability, are retained. In the test phase, the projection matrix $(\mathbf{I} - \mathbf{E}\mathbf{E}^{\mathbf{T}})$ is applied on a non-compensated GMM supervector to remove those components with maximum session variability. We apply the projection matrix

after adaptation and before classification.

We use the GMM supervector linear kernel[2] to compute the similarity between two supervectors in the SVM classifier. This kernel is an upper bound of the Kullback-Leibler divergence. For two speech segments s_a and s_b with the corresponding adapted models, it can be written by

$$k(s^a, s^b) = \sum_{m=1}^{M} \left(\sqrt{\lambda_m}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}\boldsymbol{\mu}_m^a\right)^{\mathbf{T}} \left(\sqrt{\lambda_m}\boldsymbol{\Sigma}_m^{-\frac{1}{2}}\boldsymbol{\mu}_m^b\right) \quad (8)$$

which can be computed as a linear kernel if we let GMM supervectors be normalized as $\mathbf{m} = (\sqrt{\lambda_1}\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\boldsymbol{\mu}_\mathbf{1}^\mathbf{T}, \ldots, \sqrt{\lambda_M}\boldsymbol{\Sigma}_M^{-\frac{1}{2}}\boldsymbol{\mu}_\mathbf{M}^\mathbf{T})^\mathbf{T}$.

4.1 System setup

The front-end extracts 15 Perceptual Linear Prediciton (PLP) features with normalized energy, plus their Δ and $\Delta\Delta$ coefficients, every 10ms using a window of 30ms. Feature warping[8] using a 3s window is later applied. The start and end timestamps in the provided transcriptions are used for speech/non-speech segmentation.

The UBM has 512 Gaussians and it is trained using 2900 segments taken from the 2004 NIST SRE training data. We use a Gaussian splitting strategy with 5 iterations of maximum likelihood estimation per step as well as one iteration of relevance MAP adaptation with a typical relevance factor of 10.

The NAP transform was trained using the same 2900 segments used for UBM training. 50 dimensions were removed from the GMM supervector space.

We use a soft-margin SVM (LIBSVM package) with linear kernel as the classifier. We use the same 2900 segments used for UBM training as the impostor speaker data. No score normalization is used.

5. Experimental protocol

The performance of the systems under study is assessed using the conversational telephone speech data of the 2006 NIST Speaker Recognition Evaluation (SRE) [⋆1], involving a large number of speakers as well as strong acoustic channel mismatch. The speaker verification system is asked to decide whether speech from a given target speaker is present in another speech segment. Evaluation was carried out on the English trials of the core condition, consisting of speech segments of 5 minutes with an average of 2 minutes of effective speech per conversation side. 816 and 3735 segments are available for training and test respectively for a total of 22316 scored trials. Trials involve same gender segments with an overall ratio of impostor to true trials of 10.

The primary performance measure for the NIST speaker detection task is the Detection Cost Function (DCF) defined as a cost function weighting the false alarm and miss error probabilities $\mathrm{DCF}_{\mathrm{Norm}} = P_{\mathrm{Miss}} + 9.9 \times P_{\mathrm{FalseAlarm}}$ according to the defined decision costs. We report the Minimal DCF (MDC) obtained a posteriori for the best possible detection threshold. Since this operating point favors false alarms, we also provide the Equal Error Rate (EER) as an alternative performance measure. The Detection Error Tradeoff (DET) curves[6] assess system behavior at all operating points.

6. Experiments and results

To assess the impact of the proposed technique we ran experiments for the MAP GMM-SVM systems using different values of the scaling factor C of equation (7). The scaled-and-normalized between-speaker to within-speaker variance ratio of the occupation counts, i.e. β''_m, exhibited maximum values of 1.6, 2.3, 3.0 and 3.6 for $C = 1, 2, 3$ and 4 respectively. Therefore, these maximum correction factors reduce the relevance factor by a factor of 3.6 at the most. Minimum values were as low as 0.02, resulting in a 50 times larger τ_m for $C = 4$.

We used NAP compensation in these experiments since the GMM-SVM performance is much better than without using it. However, note that NAP transforms the GMM supervectors according to a inter-session based criterion. According to the speaker labeling in our database, inter-session variability is equivalent to within-speaker variability, which is also being exploited in the proposed weighting technique. Therefore, some non-negligible interaction between both technique might be present, although not accounted for or evaluated in this study.

Table 1 shows results for systems using relevance MAP adaptation with/without relevance factor weighting. Absolute performance of the baseline MAP system is high, at the state-of-the-art compared to similar systems in[1]. Our proposed technique improves performance for almost all values of C, obtaining a maximum relative gain of 7% EER for $C = 3$. The relative gains in MDC

⋆1 The NIST 2006 SRE evaluation plan, `http://www.nist.gov/speech/tests/spk/`

Table 1 MDC and EER for GMM-SVM systems using MAP adaptation and inter-speaker weighting on the 2006 NIST SRE evaluation data. The minimum and maximum value for the corrected τ_m are shown in the second column. The lowest MDC and EER are shown in boldface.

System	τ_m (min-max)	MDC	EER (%)
MAP	1	0.0168	3.13
MAP $C = 1$	0.5-1.6	0.0166	3.04
MAP $C = 2$	0.3-2.3	0.0166	2.99
MAP $C = 3$	0.2-3.0	**0.0166**	**2.90**
MAP $C = 4$	0.02-3.6	0.0169	2.99

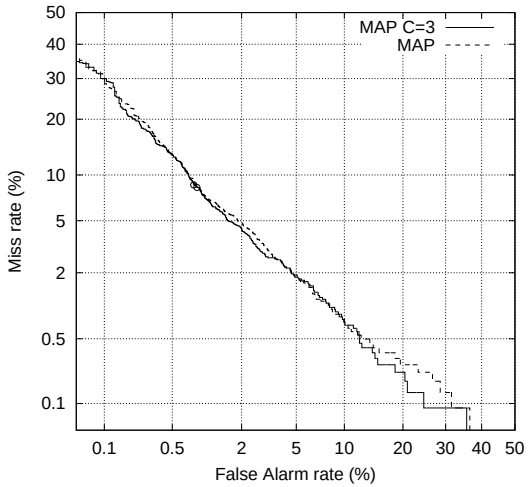

Fig. 1 DET curve for GMM-SVM systems using MAP adaptation with no relevance factor compensation and relevance factor compensation with $C = 3$. MDC operating points are shown by circles.

are smaller and more stable across different values of C. The DET curves of Figure 1 reveal that the improvement of the weighting technique concentrate in very low false alarm rate, very high false alarm and near EER areas. Therefore, although an improvement is obtained, it is still dependent on the application operating point.

7. Conclusions

We proposed a simple technique for improving the robustness of MAP adaptation of Gaussian Mixture Models to inter-session variability. Based on a state-of-the-art performing GMM-SVM system, a relative improvement of up to 7% EER was obtained on the 2006 NIST SRE data. Nonetheless, the improvement was dependent on the application operation point. Although this is a rather adhoc technique that is not derived directly from the Bayesian framework, the results show that it can still be effective for some applications.

References

1) N.Brummer, L.Burget, J.H. Cernocky, O.Glembek, F.Grezl, M.Karafiat, D.A. van Leeuwen, P.Matejka, P.Schwarz, and A.Strasheim. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Trans. on Audio, Speech and Language Processing*, 15:2072 – 2084, September 2007.
2) W. M. Campbell, D.E. Sturim, and D. A. Reynolds. Support Vector Machines using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.
3) W. M. Campbell, D.E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In *Proc. IEEE ICASSP*, pages 97–100, 2006.
4) 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*, 2:291–298, April 1994.
5) P.Kenny, G.Boulianne, P.Ouellet, and P.Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 15(4):1435–1447, 2007.
6) A.Martin, G.Doddington, T.Kamm, M.Ordowski, and M.Przybocki. The det curve in assessment of detection task performance. In *Proc. EUROSPEECH*, volume4, pages 1895–1898, 1997.
7) T.Oonishi, K.Iwano, and S.Furui. VAD-measure-embedded Decoder with Online Model Adaptation. In *Proc. INTERSPEECH*, pages 3122–3125, September 2010.
8) J.Pelecanos and S.Sridharan. Feature Warping for Speaker Verification. In *Proceedings of the IEEE Speaker Odyssey Workshop*, 2001.
9) D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.