COMBINATION OF GMM AND SVM IN SPEAKER VERIFICATION

Lukáš MACHLICA [1]

1 INTRODUCTION

The task of speaker recognition may be viewed as a validation process, where a decision about the true identity of an unknown speaker represented by her/his speech recording has to be made. Several subtasks may be examined, however let us focus on the Text Independent Speaker Recognition (TISR), for list of all of the subtasks see Psutka (2007). Hence, none a-priori assumption is made about the presence of acoustic events (phones, syllables, words, etc.) occurring in the speech recording. Well-known techniques commonly utilized in automatic TISR are based on Cepstral Coefficients (CCs) and Gaussian Mixture Models (GMMs). At first CCs are extracted from the speech recording (an acoustic space is formed), and subsequently, GMMs are trained to represent the speaker specific regions in the acoustic space. In order to train a GMM for a given (target) speaker, Expectation-Maximization (EM) algorithm based on Maximum Likelihood (ML) is utilized. To cope with small amount of training data rather than train each GMM for each speaker via EM algorithm (may lead to ill-conditioned solutions), ML based adaptation of an Universal Background Model (UBM) was proposed by Reynolds (2000). UBM is trained on a huge amount of (impostor/non-target) data, and should reflect environment conditions of a given TISR task. Since ML estimation relies only on target data, it does not reflect the topology/location/characteristic of impostor data, it is quite handy to involve also discriminative techniques providing such an additional information. One of discriminative training methods, successfully implemented into TISR task (see Campbell (2006), Longworth (2008)), is the concept of Support Vector Machines (SVMs) introduced by Vapnik (1995). The objective of SVM training is to find a hyperplane separating two classes given by target speaker data and impostor data so that the margin between these two classes would be maximized. Approach combining GMMs and SVMs with additional improvements will be described in sequel.

2 COMBINATION OF GMM AND SVM TRAINING

In order to describe the training process, some notations have to be made. Let $\lambda_s = \{\omega_i, \mu_i, \mathbf{C}_i\}_{i=1}^{M}$ denote the set of parameters belonging to the s-th speakers' GMM, where $\omega_i, \mu_i, \mathbf{C}_i$ are the i-th mixture weight, mean, and covariance matrix, respectively, and M is the number of mixtures. Let $\psi(\lambda_s) = [\mu_{s1}^{\mathrm{T}}, \ldots, \mu_{sM}^{\mathrm{T}}]^{\mathrm{T}}$ denote a mapping of λ_s to a high dimensional SuperVector (SV) consisting of concatenated GMM means. Assume that the GMM parameters were obtained according to the Maximum A-Posteriory (MAP) adaptation - *ML stage*. Gaussian mixtures cover the speaker specific regions (location of which is given by the means $\mu_\mathbf{i}$) in the acoustic space. Now, speaker specific SV $\psi(\lambda_s)$ along with a set of impostor SVs (acquired from distinct speakers) is handed to the SVM training - *discriminative stage*. Output of the SVM estimation process is a normal vector $\mathbf{w}_s$ of a hyperplane (assuming involvement of a linear kernel, see Vapnik (1995)) separating target speaker SV and impostor SVs. Hence, elements in $\mathbf{w}_s$ may be

[1]Ing. Lukáš Machlica, University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 377632584, e-mail: machlica@kky.zcu.cz

interpreted as discriminative "weights" posed on the s-th speakers' GMM means. Thus, an additional information is supplied concerning location of regions in the acoustic space occupied by other speakers. Note that the MAP adaptation involved in the training process is crucial since the sequence of means μ_i in $\psi(\lambda)$ among distinct GMMs would be otherwise inconsistent. For more details (e.g. description of the verification process) see Campbell (2006).

3 IMPROVEMENTS

Generally, the whole speech recording (more precisely, all the extracted feature vectors) of one speaker may be utilized at once in order to train one GMM, thus only one SV per speaker is extracted. Hence, when speaker specific SVM model $\mathbf{w}_s$ is trained, the orientation of the separating hyperplane is determined only through impostor SVs (variation between speaker SVs does not exist since only one target speaker SV is present). This may cause poor generalization to unseen data in the verification process. Therefore, it is more suitable to divide the speakers' speech recording into several (e.g. uniform) parts, train a GMM for each of the parts and map each GMM to a distinct SV. The division of the speech recording can follow random selection or sequence based selection. Both approaches were studied. A simple sequence selection, where the input stream is stepwise partitioned into equally large groups, outperformed the random selection. The result is most likely caused by the fact that the random selection lowers the variation in final SVs. Hence, the orientation of separating hyperplane depends more on impostor SVs than in the sequence based case.

4 CONCLUSION

In this paper methods incorporating ML based estimation and discriminative techniques were presented. Focus was laid on combination of GMM and SVM. In addition, some improvements described in the previous section were proposed in order to improve the estimation process. Experiments were performed on the NIST SRE 2008 corpus [2]. NIST SRE 2008 evaluation contained 98776 trials (trial = one evaluation involving one speaker model and one test segment), where 20449 were true trials (speaker model and test segment correspond to the same speaker), and the rest were false trials. Experiments proved the evidence of improvements, a decrease in error rate of 1% absolutely was observed.

Acknowledgement: The work has been supported by "Studentská grantová soutěž: Inteligentní metody strojového vnímání a porozumění", project No. SGS-2010-054.

REFERENCES

Vapnik V., 1995. The Nature of Statistical Learning Theory. *Springer-Verlag*, New York.

Reynolds D.A., Quatieri T.F., and Dunn R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, Vol. 10. pp 19-41.

Campbell W.M., Sturim D.E., and Reynolds D.A., 2006. Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters, IEEE*, pp 308-311.

Psutka J. , Müller L., Matoušek J., and Radová V., 2007. Mluvíme s počítačem česky. *ACADEMIA Praha*.

Longworth C., and Gales M.J.F., 2008. Multiple kernel learning for speaker verification. *In Acoustics, Speech and Signal Processing, IEEE International Conference*, pp 1581-1584.

[2] http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf