A study of variational method for text-independent speaker recognition

Liang He[1], Yao Tian[1], Yi Liu[1], Fang Dong[2], WeiQiang Zhang[1], Jia Liu[1]

[1]Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]School of Information and Electrical Engineering,
Zhejiang University City College, Hangzhou 310015, China
heliang@mail.tsinghua.edu.cn

Introduction

The i-vector firstly proposed by Najim Dehak has become the state-of-the-art algorithm for text-independent speaker recognition. In Dehak's classic work[1], he combines the eigenvoice and eigenchannel subspaces together to form a total variability subspace. The corresponding subspace loading factor is termed as the identity vector (i-vector for short). Unfortunately, the theoretic part is not fully addressed in that paper and mainly from the joint factor analysis (JFA) proposed by Patrick J. Kenny [2]. Kenny's JFA related works are classical but hard to read, which motivates us to re-consider the derivation.

In this paper, we propose a concise derivation based on the variational method. Our proposed variational method avoids solving the log likelihood directly and tries to maximize its lower bound by the Jensen's inequality. We also extend it to the maximum a posteriori (MAP) and maximum marginal likelihood (MML) criterion. The MAP criterion takes the prior distribution into account which may improve the recognition performance. The MML criterion takes the uncertainty of model's parameters into account by integrating them out.

An inequality for the *log sum exp*

Let p and q are mixture models for probability density. We examine log (p)

$$
\begin{aligned}
\log(p) &= \log\Big(\sum_{m=1}^{M}\alpha_m p_m\Big) \\
&= \log\Big(\sum_{m=1}^{M}(\alpha_m p_m \frac{\beta_m q_m}{\beta_m q_m})\Big) + \log(q) - \log(q) \\
&= \log\Big(\sum_{m=1}^{M}\frac{\beta_m q_m}{q}\frac{\alpha_m p_m}{\beta_m q_m}\Big) + \log(q) \\
&\geq \sum_{m=1}^{M}\frac{\beta_m q_m}{q}\log\big(\frac{\alpha_m p_m}{\beta_m q_m}\big) + \log(q) \\
&\geq \sum_{m=1}^{M}\frac{\beta_m q_m}{q}\log\big(\alpha_m p_m\big) + \log(q) - \sum_{m=1}^{M}\frac{\beta_m q_m}{q}\log(\beta_m q_m)
\end{aligned}
$$

From the perspective of variational method, we avoid solving the objective function directly and turn to maximizing the lower bound by selecting a proper q with a simple structure or known parameters. This is especially suitable for mixture models with exponential family distributions which are widely used in the field of machine learning. It transforms a complex *log sum exp* problem into a simple *sum* problem. Further more, we re-consider it from the view of information theory by re-arranging and integrating

$$
\begin{aligned}
D_{\mathrm{KL}}(q\|p) &= \int q\log(\frac{q}{p})dx \\
&\leq \int\Big(\sum_{m=1}^{M}\beta_m q_m\log(\frac{q_m}{p_m}) + \sum_{m=1}^{M}q_m\beta_m\log(\frac{\beta_m}{\alpha_m})\Big)dx \\
&\leq \sum_{m=1}^{M}\beta_m D_{\mathrm{KL}}(q_m\|p_m) + D_{\mathrm{KL}}(\beta\|\alpha)
\end{aligned}
$$

This inequality states that the KL divergence between two mixture models is upper bounded by two types of divergences: a weighted summation of mixture component divergences and weight divergence.

Variational methods

- ML criterion: Frequentist setting, no prior and take w as a fixed value. We use the above equality to maximize the lower bound of the objective function.
- MAP criterion: Bayesian setting, we can select prior from different considerations. In this paper, we study prior selection using the maximum entropy criterion and empirical MAP criterion.
- MML criterion : Bayesian setting but take prior as a variable. After complicated derivation, we found it has the same form to the ML criterion. After some variable manipulation, the estimation procedure of MML is the same to the ML.
- There are headache formula derivations which are not presented here. You can find more details in the paper.

Comparison of different criteria

Algorithm	ML	MAP, ME	MAP, EBM	MML
prior	-	Identity matrix	from data	variable

Experiments

used previous NIST SRE data corpus to train models. Speech/silence segmentation was performed by a G.723.1 VAD detector. A 13-d MFCC + Δ + ΔΔ was extracted. UBM, 1024, gender dependent; T, 400; length normalization; a simplified PLDA.

Experimental results on the NIST SRE08 tel-tel-eng female task

Algorithm	cosine		PLDA	
	EER	MinDCF	EER	MinDCF
ML	5.75	0.250	2.82	0.123
MAP, ME (i-vector)	5.76	0.255	2.84	0.125
MAP, EBM	6.17	0.279	3.03	0.132
MML	5.75	0.250	2.82	0.123

Analysis and Conclusions

We propose a concise method for the derivation of i-vector (also suitable for JFA) based on the variational method and compare several variational methods based on the ML, MAP and MML criteria. The ML and MML criteria lead to the same result in our setting. In the case of MAP criterion, the prior selection has an obvious influence. Although in our experimental result, MAP(EBM) is inferior to the other criteria, we believe that it may achieve a better performance by selecting training data carefully which is also our future work.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, *Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech Language Processing*, vol. 19, no. 4, pp. 788--798, 2011.

[2] P. Kenny, Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13, Montreal, CRIM, 2005. http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf, 2016.