Speech Communication Lab (SCL) SRE10 system description

Daniel Garcia-Romero, Xinhui Zhou and Carol Y. Espy-Wilson

{dgromero, zxinhui, espy}@umd.edu

Overview

The SCL submission comprises a single system based on the novel paradigm of signal coding via overcomplete dictionaries [1]. Under this paradigm, the data corresponding to the train and test segments is summarized into a fixed-length vector (GMM supervector) and subsequently coded using an overcomplete dictionary via ridge regression. After the encoding stage, a pair of supervectors $\boldsymbol{\eta}_{model}$ and $\boldsymbol{\eta}_{test}$ are obtained and a similarity measure between them is computed by means of a normalized inner product (cosine of the angle) followed by ZTnorm. The overcomplete dictionary $\boldsymbol{\Phi}$ used to encode the data is constructed by appending a set of eigensession supervectors $\mathbf{U}$ to a diagonal matrix $\mathbf{D}$ resulting in $\boldsymbol{\Phi} = [\mathbf{U}\ \mathbf{D}]$. The columns of $\mathbf{U}$ are learned from a development data set via ML. The matrix $\mathbf{D}$ is fixed and obtained from a UBM to implement relevance MAP [2].

In the following we present a more detailed description of the fundamental stages of the system.

Feature extraction[1]

The ASR transcripts along with a set of heuristics based on energy were used to perform VAD. Subsequently, 38 MFCCs (c1-c19 + Delta) were computed using a 20ms Hamming window shifted by 10ms. The MFB energies were processed with RASTA and the 38MFCCs were normalized to zero mean and unit variance.

UBM

A gender-independent 2048 mixture GMM was trained from 10 hours of speech from telephone and microphone speech from SRE04, SRE05 , SRE06 as well as microphone speech from the SRE08-follow-up evaluation. 15 EM iterations were used to train the UBM.

Dictionary learning (Eigen-session subspace)

Two eigensession subspaces $\mathbf{U}_{mic}$ and $\mathbf{U}_{tel}$ were trained from the same data used for the UBM. In particular, 17676 files of 937 speakers from SRE04, SRE05 , SRE06 were used to learn $\mathbf{U}_{tel}$ via ML estimation. Four iterations of EM were used and initialization was performed through PCA. The same procedure was used for $\mathbf{U}_{mic}$ but using the microphone data from SRE08-follow-up with 5579 files from 147 speakers. Each of the subspaces was of dimension 50. The final subspace $\mathbf{U} = [\mathbf{U}_{mic}\,\mathbf{U}_{tel}]$ of dimension 100 was constructed by appending together the individual subspaces.

[11] The authors would like to thank MIT-LL for providing the feature extractor as well as the GMM training binaries. We used them to parameterize the data as well as to train the UBM.

Train and test segments coding

For each train and test segment a supervector of dimension $p = 2048 \times 38$ is obtained by computing the zero and first order sufficient statistics of the data with respect to the UBM. The resulting supervector $\boldsymbol{\eta}$ is subsequently encoded as a linear combination of the columns of $\boldsymbol{\Phi}$ by minimizing the ridge regression objective [1]:

$$\min_{\boldsymbol{\beta}} \Psi(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \mathbf{W}^{\frac{1}{2}} (\boldsymbol{\eta} - \boldsymbol{\Phi}\boldsymbol{\beta}) \right\|_2^2 + \frac{1}{2} \|\boldsymbol{\beta}\|_2^2. \quad (1)$$

This operation results in an encoded $\widehat{\boldsymbol{\eta}} = \boldsymbol{\Phi}\boldsymbol{\beta}$ that can be session-compensated by setting the coefficients of the vector $\boldsymbol{\beta}$ that correspond to the columns of $\mathbf{U}$ to zero. Thus, each train and test segment is represented by a session-compensated supervector $\widehat{\boldsymbol{\eta}}_c$.

Scoring

Once all the test and training segments have been encoded, a similarity measure between them is obtained by a normalized inner product (cosine of the angle):

$$norm\_score = \frac{\langle \widehat{\boldsymbol{\eta}}_{A|c}, \widehat{\boldsymbol{\eta}}_{B|c} \rangle_{\mathbf{W}}}{\langle \widehat{\boldsymbol{\eta}}_{A|c}, \widehat{\boldsymbol{\eta}}_{A|c} \rangle_{\mathbf{W}}^{1/2} \langle \widehat{\boldsymbol{\eta}}_{B|c}, \widehat{\boldsymbol{\eta}}_{B|c} \rangle_{\mathbf{W}}^{1/2}} \quad (2)$$

where $\widehat{\boldsymbol{\eta}}_{A|c}$ and $\widehat{\boldsymbol{\eta}}_{B|c}$ represent the model and test session-compensated supervectors and $\mathbf{W}$ is a positive-definite matrix defining the inner product. In particular, $\mathbf{W} = \boldsymbol{\Sigma}_{UBM}^{-1} \boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma} = \mathrm{diag}\{\mathrm{weights}_{\mathrm{ubm}}\}$. Finally, the scores were normalized by ZTnorm.

Computation time

The system was run on a node with 16 Intel(R) Xeon(R) E5520 @ 2.27GHz cores and 24 GB of RAM. The times reported here are for a single core.

Task	Time
UBM training	6 hours
Model encoding (5460 models)	0.027sec/per model (147.5 sec for all models)
Test segment (13,344 segments)	0.027sec (360.3 sec for all data)
ZTnorm + Scoring	3,764.8 seconds for all 610,748 trials

References

[1] D. Garcia-Romero and C. Espy-Wilson, “Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete Dictionaries”, admitted to Odyssey 2010.

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