i-vector and GMM-UBM

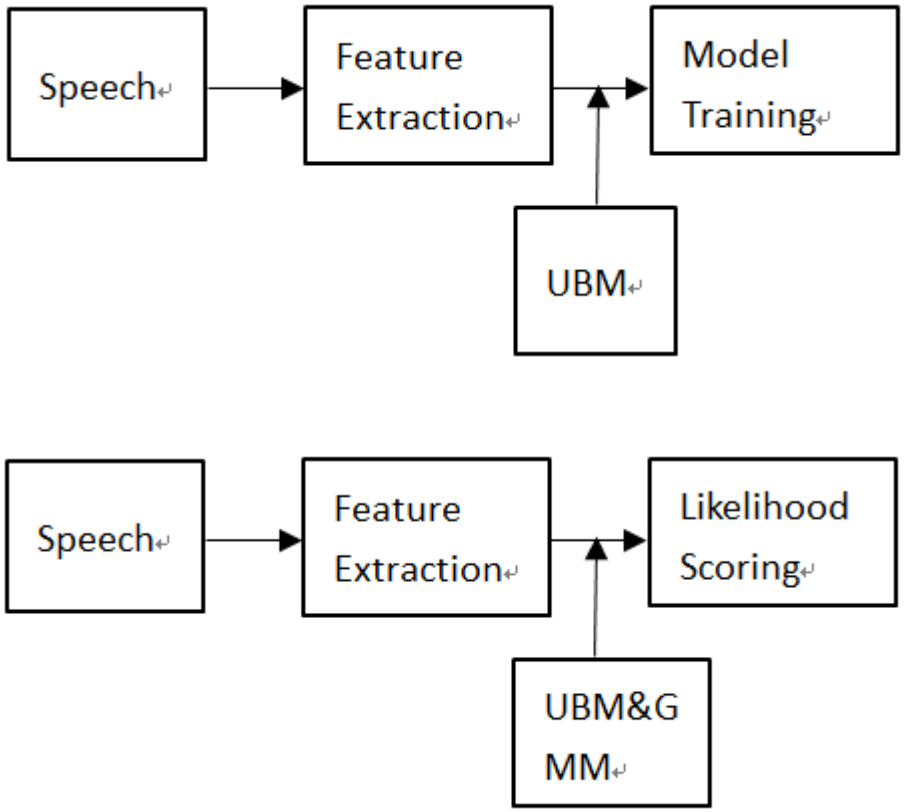
Bie Fanhu

CSLT, RIIT, THU

2013-11-18

Framework

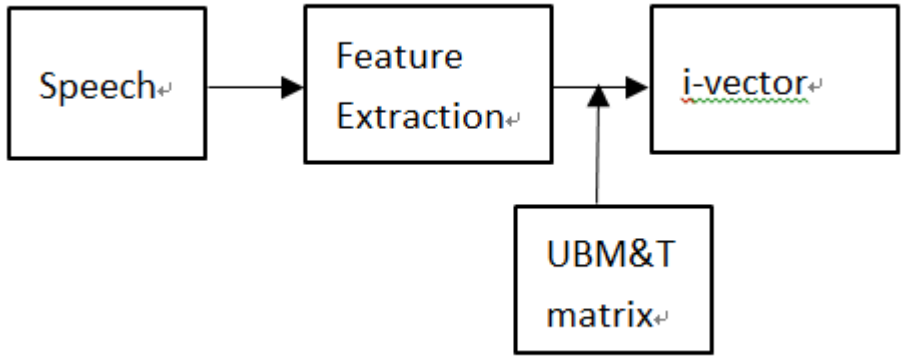
1. GMM-UBM

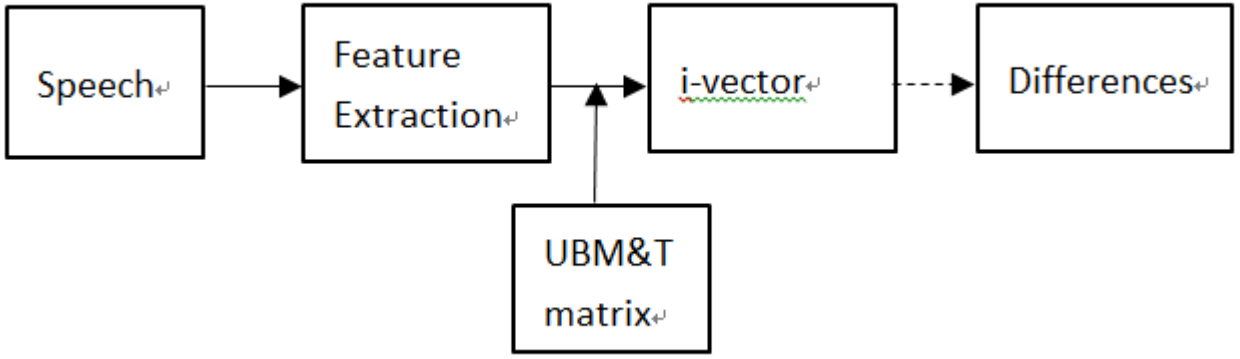


- Feature is extracted by frame. Number of features are unfixed.
- Gaussian Mixtures are used to fit all the features. The mixtures are fixed so the size of the model is fixed.
- The difference of the likelihood ratio from the GMM to the UBM is used to describe the result.
- Calculate the differences between 2 distributions.

Framework

2. i-vector





- I-vector can be seen as both features and models.
- I-vector is calculated with all the features at one time.
- I-vector is a vector which can be seen as compressed from the supervector of the mean vectors from the GMM.

i-vector theory

- I-vector is carried out in a defined subspace
  - Define a low dimension subspace, which is referred as the "total variability space".
  - The subspace contains the speaker and session variability simultaneously.
  - In the subspace, the speaker and session variability can be easier to separate.
  - Defined with the eigenvectors with the largest eigenvalues.

i-vector theory

- $M = m + Tw$
  - M stands for the speaker, assumed to be normally distributed with mean vector m and covarience matrix TT^t
  - m stands for the mean (supervector) of the UBM, which is considered as session independent, speaker independent, so it should use the UBM means trained from all the data.
  - T is rectangular matrix of low rank, which is used to map the supervector to a low demission space called total variability space.
  - W is a random vector having a normal distribution N(0, I), which is extracted from each utterance.

Calculate w

- <<Pattern recognition>>

Calculate p(x|y) in the condition of p(x) and p(y|x) assumed the Gaussian distribution

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

Posterior probability:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \quad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

Calculate w

- <<Pattern recognition>>

Assume the prior and the likelihood function to be Gaussian distribution as follows:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \tag{3.10}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{3.48}$$

Posterior probability:

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) \tag{3.50}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}. \tag{3.51}$$

Calculate w

- <<Front-End Factor Analysis for Speaker Verification>>

w is calculated as follows:

$$w = \left(I + T^t \Sigma^{-1} N(u) T\right)^{-1} . T^t \Sigma^{-1} \tilde{F}(u).$$

where:

$$N_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega)$$

$$F_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega) y_t$$

$$\tilde{F}_c = \sum_{t=1}^{L} \mathrm{P}(c|y_t, \Omega)(y_t - m_c)$$

N(u) is a diagonal matrix of dimension CF*CF whose diagonal blocks are $N_c I$. $\tilde{F}(u)$ Is a supervector of dimension CF*1 obtained by concatenating all $\tilde{F}_c$

Calculate w

- Calculation and conclusion

Compare the equations above, we can get the following equations:

$$S_0 = I$$
$$\Phi = T$$
$$\beta = \Sigma^{-1} N(u)$$

So the matrix T is considered as a linear transformation for the original feature(point), question: Φ is treated as a function activated on x(original point), but T is a fixed formation.

Under the above conclusion(if it`s right), we need to discuss the following equation:

$$N(u)\boldsymbol{t} = \tilde{F}(u)$$

Question: $\boldsymbol{t}$ stands for the target value for the train set. What does it mean here?

Some discussions:

- Discussions:
  - w is assumed to be N(0, $\boldsymbol{I}$), is it right according to the formula? Is the assumption right?

    This assumption equals to M obeys Gaussian distribution. The matrix TT^t is used to normalize the w distribution. According to the formula, the mean for m is 0.

  - What`s the target for the matrix T?

    1. Decrease the dimension, and compressed the information, which used for WCCN, PLDA, because of the data limitation.
    2. Separate the speaker and session variability (not perfect), because the T calculation is based on the eigenvalues, so the transformation will compressed the information to the mail eigenvectors.

Some discussions:

- Comparing with the GMM-UBM, the w(mean) calculation with a prior distribution N(0,$\boldsymbol{I}$). If getting away with the prior , will it get a same or familiar result?

Both systems use the Gaussian distribution to fit the data, not considering the dimension difference.

In GMM mapping, the feature is calculated by frame with MAP. In i-vector, the features of all the frame from one speech are calculated first and then using ML in the given model. So it can be guessed, they will give a familiar result.

Thanks