MMSE-based Stereo Feature Stochastic Mapping for Noise Robust Speech Recognition

Author : Xiaodong Cui , Mohamed Afify , Yuqing Gao

Professor : 陳嘉平

Reporter : 許峰閤

ABSTRACT

A stochastic mapping approach under the MMSE criterion based on stereo features is investigated in this paper for noise robust speech recognition. By learning the mapping from a joint GMM distribution of clean and noisy features, the MMSE estimate of the clean feature is shown to be a piece-wise linear transformation of the noisy feature. The mathematical relationship between the proposed MMSE mapping and other piece-wise linear estimates for noise robustness (i.e. MAP mapping and SPLICE) is also analyzed and discussed. Experimental results show that the proposed MMSE-based stochastic mapping yields superior performance over the MAP mapping on DARPA Transtac large vocabulary spontaneous speech test sets when using clean and multi-style acoustic models.

Mathematical Formulation

- 首先先設定stereo feature為$\{(x_i, y_i)\}$ 其中X為clean feauter而Y為其對應的noisy feature

- 接著將Z設定為$z \equiv (x, y)$也就是Z為一個將clean feature及noisy feature串起來的向量

Mathematical Formulation

- 一個聯合的GMM分佈如下

$$p(z) = \sum_{k=1}^{K} c_k \mathcal{N}(z; \mu_{z,k}, \Sigma_{zz,k})$$

K　為第K個GMM Component

c_k 為第K個Component的WEIGHT

$\mu_{z,k}$ 為第K個Component 的 Mean

$\Sigma_{zz,k}$ 為第K個Component的covariance

Mathematical Formulation

- 將其中的Mean及Covariance以下式表示

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \qquad \Sigma_{zz,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix}$$

- 當有一個noisy feature y時可用下式將其逼近clean feature x

$$\hat{x} = E[x|y]$$

Mathematical Formulation

- p(x, y)是一個GMM, 可進一步改寫成下式

$$
\begin{aligned}
\hat{x} &= \int_x p(x|y)x\,dx \\
&= \sum_k p(k|y)\int_x p(x|k,y)x\,dx \\
&= \sum_k p(k|y)E[x|k,y]
\end{aligned}
$$

Mathematical Formulation

- 而事後機率 $p(k|y)$ 以下式表示

$$p(k|y) = \frac{p(y|k)p(k)}{\sum_k p(y|k)p(k)}$$

- 而期望值 $E[x|k,y]$以下式表示

$$E[x|k,y] = \mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y - \mu_{y,k})$$

Mathematical Formulation

- 基於上面的推導可將欲補償的y以下式表達

$$\hat{x} = \sum_{k} p(k|y)(A_k y + b_k)$$

$$\begin{aligned} A_k &= \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \\ b_k &= \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \end{aligned}$$

SPLICE

- SPLICE為雙聲源為基礎的的分段線性補償
- 利用高斯混合模型來表示受雜訊干擾的語音特徵參數分佈,每個高斯分佈代表語音特徵參數在一種特定雜訊環境下的分佈情形
- 而每個高斯分佈都有一個對應的校正向量來對noisy做補償用來逼近clean feature

SPLICE

- SPLICE對noisy feature的補償可以下式表示

$$\hat{x} = \sum_k p(k|y)(y + r_k)$$

r_k 為其校正向量

$$r_k = \frac{\sum_n p(k|y_n)(x_n - y_n)}{\sum_n p(k|y_n)}$$

SPLICE

- 在SPLICE中的事後機率 $p(k|y)$ 是由noisy的高斯分佈中算得,而MMSE是由聯合的高斯分佈所算得

- SPLICE又假設轉換矩陣 A_k 為一個單位矩陣,也就是MMSE中的一個特殊例子當 $\Sigma_{xy,k} = \Sigma_{yy,k}$ 時,便可由前MMSE的式子得

$$r_k = \mu_{x,k} - \mu_{y,k}$$

MAP

- MAP為一個疊代的分段線性補償技術

$$\hat{x}^{(l)} = \sum_k p(k|\hat{x}^{(l-1)}, y)(A_k y + b_k)$$

$$A_k = \left(\sum_k p(k|\hat{x}^{(l-1)}, y) \Sigma_{x|y,k}^{-1} \right)^{-1}$$

$$\Sigma_{x|y,k}^{-1} \Sigma_{xy,k} \Sigma_{yy,k}^{-1}$$

$$b_k = \left(\sum_k p(k|\hat{x}^{(l-1)}, y) \Sigma_{x|y,k}^{-1} \right)^{-1} \cdot$$

$$\Sigma_{x|y,k}^{-1} \left(\mu_{x,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \mu_{y,k} \right)$$

MAP

- 在MAP中假設GMM中的每個component共用一個條件共變異矩陣 $\Sigma_{x|y}$

$$
\begin{aligned}
& \left(\sum_k p(k|\hat{x}^{(l-1)}, y)\Sigma_{x|y,k}^{-1}\right)^{-1} \\
= & \left(\Sigma_{x|y}^{-1}\sum_k p(k|\hat{x}^{(l-1)}, y)\right)^{-\mathbf{1}} \\
= & \left(\Sigma_{x|y}^{-1}\cdot 1\right)^{-1} = \Sigma_{x|y}
\end{aligned}
$$

MAP

- 所以我們可以對轉換矩陣 A_k 重寫

$$
\begin{aligned}
A_k &= \Sigma_{x|y}\Sigma_{x|y}^{-1}\Sigma_{xy,k}\Sigma_{yy,k}^{-1} = \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \\
b_k &= \Sigma_{x|y}\Sigma_{x|y}^{-1}\left(\mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k}\right) \\
&= \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k}
\end{aligned}
$$

實驗

- 使用DARPA Transtac project
- Test set A 包括11個男性語者共2070句

Condition	Clean	15 dB	10 dB
no compensation	15.96	31.97	40.72
MMSE-SSM24	14.84	31.21	40.58
MMSE-SSM40	14.70	28.74	35.47

Table 1. Word error rate (WER) with clean acoustic model on Set A when applying MMSE mapping to different domains.

實驗

Condition	Clean	15 dB	10 dB
no compensation	15.96	31.97	40.72
MAP-SSM40-1iter	14.77	30.63	39.23
MAP-SSM40-3iter	14.77	30.54	39.12
MMSE-SSM40	14.70	28.74	35.47

Table 2. Word error rate (WER) with clean acoustic model on Set A using MAP and MMSE mappings.

Condition	Clean	15 dB	10 dB
no compensation	10.48	20.16	27.15
MAP-SSM40-1iter	11.31	16.63	20.09
MAP-SSM40-3iter	10.96	17.10	20.58
MMSE-SSM40	11.25	16.94	20.24

Table 3. Word error rate (WER) with MST model on Set A using MAP and MMSE mappings.