Stereo-Based Stochastic Mapping for Robust Speech Recognition

Author : Mohamed Afify, Xiaodong Cui, and Yuqing Gao

Professor:陳嘉平

Reporter :吳柏鋒

We present a stochastic mapping technique for robust speech recognition that uses stereo data. The idea is based on building a GMM for the joint distribution of the clean and noisy channels during training and using an iterative compensation algorithm during testing. The proposed mapping was also interpreted as a mixture of linear transforms that are estimated in a special way using stereo data. The proposed method results in 28% relative improvement in string error rate (SER) for digit recognition in the car, and in about 10% relative improvement in word error rate (WER), when applied in conjunction with multi-style training (MST), for large vocabulary English speech recognition.

We present a stochastic mapping technique for robust speech recognition that uses stereo data. The idea is based on building a GMM for the joint distribution of the clean and noisy channels during training and using an iterative compensation algorithm during testing. The proposed mapping was also interpreted as a mixture of linear transforms that are estimated in a special way using stereo data. The proposed method results in 28% relative improvement in string error rate (SER) for digit recognition in the car, and in about 10% relative improvement in word error rate (WER), when applied in conjunction with multi-style training (MST), for large vocabulary English speech recognition.

We present a stochastic mapping technique for robust speech recognition that uses stereo data. The idea is based on building a GMM for the joint distribution of the clean and noisy channels during training and using an iterative compensation algorithm during testing. The proposed mapping was also interpreted as a mixture of linear transforms that are estimated in a special way using stereo data. The proposed method results in 28% relative improvement in string error rate (SER) for digit recognition in the car, and in about 10% relative improvement in word error rate (WER), when applied in conjunction with multi-style training (MST), for large vocabulary English speech recognition.

We present a stochastic mapping technique for robust speech recognition that uses stereo data. The idea is based on building a GMM for the joint distribution of the clean and noisy channels during training and using an iterative compensation algorithm during testing. The proposed mapping was also interpreted as a mixture of linear transforms that are estimated in a special way using stereo data. The proposed method results in 28% relative improvement in string error rate (SER) for digit recognition in the car, and in about 10% relative improvement in word error rate (WER), when applied in conjunction with multi-style training (MST), for large vocabulary English speech recognition.

簡介

- 在此提出以**立體音為基礎隨機對映(SSM)技術**來強健語音辨識

- 主要概念是著重在訓練階段GMM的訓練時的clean和noisy的聯合機率分佈上，在測試階段重複使用補償演算法

簡介

- 以立體音為基礎的對映技術可視為一個混合線性轉換

- **補償演算法**的概念是將clean和noisy堆疊起來形成大的增廣空間且在一個新的空間建立一個統計模型

簡介

- 在第一個實驗部分，我們將提出的對映技術在車內環境與SPLICE作比較
- 第二個實驗部分，將對映技術結合Multistyle training方法，來對大型英文字彙作語音辨識
- 使用混合線性特徵空間轉換(例FMLLR)證實在對映有不錯的結果

補償演算法

- Step1. 為了建立對映，必須訓練聯合機率 $p(z)$

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

其中 $\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix}$ 平均值

$\Sigma_{zz,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix}$ 共變數

最大期望值(EM)演算法

- 使用EM來估算(1)

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, p(x|y) = \underset{x}{\operatorname{argmax}} \sum_{k} p(x,k|y)$$

$$= \underset{x}{\operatorname{argmax}} \sum_{k} p(k|y)p(x|k,y) \tag{2}$$

- 重複執行EM演算法:

$$
\begin{aligned}
\hat{x} &= \operatorname*{argmax}_x \sum_k p(k|\bar{x}, y) \log p(k|y) p(x|k, y) \\
&= \operatorname*{argmax}_x \sum_k p(k|\bar{x}, y) [\log p(k|y) + \log p(x|k, y)] \\
&\equiv \operatorname*{argmax}_x \sum_k p(k|\bar{x}, y) \log p(x|k, y) \\
&\equiv \operatorname*{argmax}_x \frac{-1}{2} \sum_k p(k|\bar{x}, y) \left[\log |\Sigma_{x|y,k}| + \right. \\
&\quad \left. (x - \mu_{x|y,k})^T \Sigma_{x|y,k}^{-1} (x - \mu_{x|y,k}) \right] \qquad (3)
\end{aligned}
$$

- 針對 (3) 的 x 作微分，並將微分結果設為0

$$\sum_k p(k|\bar{x},y)\Sigma^{-1}_{x|y,k}\hat{x} = \sum_k p(k|\bar{x},y)\Sigma^{-1}_{x|y,k}\mu_{x|y,k} \tag{4}$$

其中

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

$$\Sigma_{x|y,k} = \Sigma_{xx,k} - \Sigma_{xy,k}\Sigma^{-1}_{yy,k}\Sigma_{yx,k}$$

簡化

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

其中

$$A_k = CD_k,\ b_k = Ce_k$$

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

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

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

實驗一

- 車內環境的數字辨識
- 使用語料庫CARVUI，此語料庫使用固定式(CT)和手持式 (HF)兩種麥克風在Bell實驗室錄製而成
- 訓練部分有7000句，測試部分有800句，且共有12HMMs(10digits+oh+sil)，且每個模型有6 states,每個state有8個高斯
- 特徵空間有39維，分別是13維倒頻譜係數(含C_0)+第一次微分+第二次微分

實驗一

- Baseline

Condition	SER
clean/clean	12.9
clean/noisy	31.7
noisy/noisy	16.8
clean/VTS	28.6

Table 1: Baseline sentence error rate (SER) results (in %) of the close-talking (CT) microphone data and Hands-Free (HF) data.

實驗一

- 在這提出兩個對映的方法:
  (1)對映是建構在clean和noisy間相同MFCC係數的對映
  (2)使用time window，其包含目前音框與左右內容，利用noisy MFCC係數來計算出對映的clean MFCC係數

實驗一

- 實驗在訓練階段是重複執行三次EM，在測試階段是只有執行一次補償演算法

- 實驗發現重複執行次數越多次，可能可以改善likelihood，但是WER仍然會增加

實驗一

- 在不同GMM個數下，SPLICE與SSM-1(無time window)的SER比較

	16	64	256
SPLICE	27.0	26.2	25.5
SSM-1	24.5	24.5	24.0

Table 2: Sentence error rate results (in %) of Hands-Free (HF) data using the proposed mapping (SSM-1) and SPLICE for different GMM sizes.

實驗一

- SMM(有time window)，GMM數為256

	SER
SSM-1	24.0
SSM-3	22.8
SSM-5	23.0

Table 3: Sentence error rate results (in %) of Hands-Free (HF) data using three different configurations of the proposed mapping (SSM) for 256 GMM size.

實驗二

- 提出的對映技術結合上mutistyle training(MST)，在大型英文字彙上作語音辨識

- 此語音辨識是建構在**IBM語音產生引擎**，MFCC係數24維(含energy)，且計算出MFCC平均值與能量正規化共有9個向量，所以參數空間共216維

實驗二

- 此特徵空間可以透過線性差分析(LDA)和最大近似線性轉換(MLLT)將維度降到40
- 聲學模型是使用決策樹之樹葉來產生相關的GMMs，透過音素內容(此有54種英文音素為依據)來對樹作分群
- 在分配特徵向量到樹葉後，GMMs的樹葉為第一次的初始，之後再執行四次FB演算法

實驗二

Condition	Clean	15 dB	10 dB
Clean	7.64	10.33	31.47
MST	4.07	5.96	14.06
MST+SSM(512)	3.71	5.92	13.88
MST+SSM(1024)	3.80	5.48	12.74
MST+SSM(2048)	3.96	5.39	13.45

Table 4: Word error rate (WER) for different testing scenarios including the Clean model, MST model, and SSM of different size applied in conjunction with MST.