Noisy Speech Recognition by using Output Combination of Discrete-Mixture HMMs and Continuous-Mixture HMMs

Authors:Tetsuo Kosaka, You Saito
and Masaharu Kato
Professor:陳嘉平
Reporter:吳柏鋒

摘要

- 簡介
- 使用離散混合HMMs(DMHMMs)作強健語音辨識
- 使用ROVER作系統整合
- 實驗

簡介

- 主要改善在不利情況下之固定(stationary)與非固定(non-stationary)噪音的辨識效果
- 使用離散混合HMMs(DMHMMs)與連續混合HMMs(CMHMMs)作為聲學模型，並使用MAP估算DMHMM參數
- 提出將兩種模型的系統輸出做整合，更進一步改善在不同噪音環境下語音的辨識效果

使用離散混合HMMs強健語音辨識

- 使用兩種方法來減少量化大小:

（1） <u>subvector-based</u>

-將特徵向量分成幾個子向量並將它們各別用codebooks作量化

（2） <u>scalar-based</u>

-將每個特徵向量的維度作常數化

使用離散混合HMMs強健語音辨識

- 針對DMHMMs提出MAP估算來更進一步減少training data量

$$\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \ldots, \boldsymbol{o}_{st}, \ldots, \boldsymbol{o}_{St}]$$:特徵向量之分割向量

$$q(\boldsymbol{o}_t) = [q_1(\boldsymbol{o}_{1t}), \ldots, q_s(\boldsymbol{o}_{st}), \ldots, q_S(\boldsymbol{o}_{St})]$$

:使用VQ codebook

使用離散混合HMMs強健語音辨識

- DMHMM的分散式輸出：

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st}))$$

where w_{im} is the mixture coefficient for the mth mixture in state i, and $\hat{p}_{sim}$ is the probability of the discrete symbol for the sth subvector.

使用離散混合HMMs強健語音辨識

- 離散機率的Maximum likelihood(ML)估算:

$$p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{imt}\, \delta(q_s(\boldsymbol{o}_{st}), k)}{\sum_{t=1}^{T} \gamma_{imt}}$$

$$\delta(q_s(\boldsymbol{o}_{st}), k) = \begin{cases} 1 & q_s(\boldsymbol{o}_{st}) = k \\ 0 & \text{otherwise} \end{cases}$$

where k is the index of the subvector codebook and γ_{imt} is the probability of the mth mixture component being in state i at time t.

使用離散混合HMMs強健語音辨識

- DMHMM的MAP估算:

$$\hat{p}_{sim}(k) = \frac{\tau \cdot p^0_{sim}(k) + n_{im} \cdot p_{sim}(k)}{\tau + n_{im}}$$

$$n_{im} = \sum_{t=1}^{T} \gamma_{imt}$$

where $p^0_{sim}(k)$ is the constrained prior parameter and τ indicates the relative balance between the corresponding prior parameter and the observed data. In our experiments, τ was set to 10.0

使用離散混合HMMs強健語音辨識

- Prior distribution參數
  -models轉換CMHMMs成DMHMMs

$$p_{sim}^{0}(k) = \frac{b'_{sim}(\boldsymbol{\nu}_s(k))}{\sum_k b'_{sim}(\boldsymbol{\nu}_s(k))}$$

where $b'_{sim}()$ is the probability density of the CMHMM, and $\boldsymbol{\nu}_s(k)$ is the centroid for each subvector s.

使用histogram equation(HEQ)作正規化

- 使用此方法主要應用在將特徵空間正規化，可以補償訓練與測試環境不匹配之情況

$$\boldsymbol{o}'_{st} = HEQ_f(\boldsymbol{o}_{st}) = C_T^{-1}(C_E(\boldsymbol{o}_{st}))$$

where C_E is the CDF estimated from test data and C_T is the CDF from training data.

使用ROVER作系統整合

- 使用ROVER(Recognizer Output Voting Error Reduction)辨識系統結果投票結合法，整合DMHMM與CMHMM兩個發聲模型所產生的輸出

- 當兩個系統有相對差異很大時，使用ROVER會有很大的改進效果，ROVER是簡單的表決(vote)機制，用來作出最適當選擇

實驗

- 使用JNAS語料庫
  (Japanese Newspaper Article Sentences)

- 共15732句，由102個男生錄音

- 分別在多條件環境下作training
  汽車、展覽館、人群、火車

- 分別在兩種多條件環境下作testing
  A. 汽車、展覽館、人群、火車(與train同)
  B. 車站、工廠、交叉路口、電梯(與train異)

實驗

- 使用HEQ對特徵作正規化(normalization)

  主要分為:

  (1)<u>utterance</u>

  – 針對要做辨識的單一句子計算

  (2)<u>noise</u>

  – 針對每個noise型態中的所有句子計算

實驗

Table 2: Results of output combination for tetstset A (WER(%)). Bold font shows the best performance among three methods.

SNR(dB)	CMHMM	DMHMM	combination
w/o normalization			
∞	6.83	**6.42**	6.63
20	7.96	8.85	**7.79**
15	10.72	10.66	**9.97**
10	15.55	14.88	**14.65**
5	25.93	25.31	**24.69**
ave.	16.75	16.53	**15.93**
normalization by HEQ (utterance)			
∞	6.00	6.31	**5.80**
20	8.03	8.28	**7.63**
15	10.64	10.20	**9.55**
10	14.67	14.57	**13.72**
5	21.74	21.51	**20.03**
ave.	15.27	15.22	**14.18**
normalization by HEQ (noise)			
∞	5.80	6.52	**5.69**
20	7.92	7.97	**7.43**
15	10.07	9.97	**9.45**
10	13.98	13.90	**13.15**
5	21.92	23.27	**21.74**
ave.	14.92	15.41	**14.37**

實驗

Table 3: Results of output combination for tetstset B (WER(%)). Bold font shows the best performance among the three methods.

w/o normalization			
SNR(dB)	CMHMM	DMHMM	combination
∞	6.83	**6.42**	6.63
20	8.31	8.28	**8.05**
15	16.75	**14.26**	15.19
10	37.09	**32.17**	33.96
5	67.80	**61.96**	64.34
ave.	34.20	**30.77**	32.04
normalization by HEQ (utterance)			
∞	6.00	6.31	**5.80**
20	9.47	**8.85**	8.93
15	14.57	13.87	**13.46**
10	25.62	25.83	**24.15**
5	53.65	50.75	**50.21**
ave.	27.33	26.40	**25.64**
normalization by HEQ (noise)			
∞	5.80	6.52	**5.69**
20	8.60	8.54	**8.28**
15	13.05	13.54	**12.81**
10	26.48	26.71	**25.11**
5	52.95	53.08	**51.22**
ave.	26.72	27.10	**25.78**