An Investigation of Phonological Feature System Used in Detection-Based ASR

Author : I-Fan Chen
Hsin-Min Wang
Professor : 陳嘉平
Reporter : 許峰閤

大綱

- 介紹
- 語音特徵集
- 利用條件隨機域做後端處理
- 實驗

介紹

- 在此系統中利用特徵偵測器爲前端處理,條件隨機域(CRF)爲後端處理

- 在前端處理中比較三種語音特徵集對語音辨識系統準確率的影響

語音特徵集

- 在此我們提出三種語音特徵集,分別爲:

1. Sound Pattern of English feature set
2. Multi-valued feature set
3. Government phonology feature set

語音特徵集

- SPE 可將複雜的語音規則簡化成較簡單的形式,整個語音集包含了13個二元的語音特徵

- MV雖然語音特徵的分類較少,但是他對於每一個語音的分類,都可以有2~10個值

語音特徵集

Table 1. *The SPE feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
Anterior	90 %	Nasal	97 %
Back	88 %	Round	94 %
Consonantal	90 %	Silence	98 %
Continuant	93 %	Strident	97 %
Coronal	89 %	Tense	90 %
High	88 %	Vocalic	87 %
Low	93 %	Voice	92 %

語音特徵集

Table 2. *The MV feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
Centrality	84 %	Phonation	91 %
Front back	82 %	Place	71 %
Manner	85 %	Roundness	91 %

語音特徵集

- GP feature set 是利用一些較小的單元去組合成phone, 母音是取A U I 這三個單元去組合其它的母音, 例如: (A,I)那會將A視爲operator, I 視爲head來組成[e], 相反的(I,A)會變成 [æ]

- 其他子音也是取一些基本單元利用上述方式拼出子音

語音特徵集

Table 3. *The GP feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
A	85 %	H	93 %
I	90 %	N	97 %
U	86 %	a	96 %
E	86 %	i	94 %
S	91 %	u	95 %
h	95 %		

語音特徵集

- SPE 跟 GP 都是使用單一的類神經網路然後有多個輸出
- MV是針對每個特徵都使用個別的類神經網路然後也是多個輸出
- 類神經網路的輸入使用12維的MFCC加一個Energy

利用CRF做後端處理

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_i \left(\sum_j \lambda_j s_j(y_i, \mathbf{x}) + \sum_k \mu_k t_k(y_{i-1}, y_i, \mathbf{x}) \right) \right)$$

x and y are the observation and output sequence

Z(x) is the normalization term

i is the index of the current position of the output sequence

$s_j()$ is the state feature function

$t_k()$ is the transition feature function

利用CRF做後端處理

$$s(y_i, \mathbf{x}) = \begin{cases} 1, & \text{if } y_i = /\text{ix}/, \text{voice}(x_i) = true, \text{and vocal}(x_{i\text{-}1}) = false \\ 0, & \text{otherwise} \end{cases}.$$

在CRF中state feature function可針對整個observation來做考慮,並不侷限於當下的狀態,在較長的observation中CRF的這個特性是較HMM好的地方

利用CRF做後端處理

- 接著要將前端處理做出的frame-based的feature sequence利用CRF來對應到輸出的frame-based的phone sequence,最後再將這些frame-based的phone sequence合併成output
- 在此state feature function 爲

$$s(y_i, x_{i-1}, x_i, x_{i+1})$$

- Transition feature function 爲bi-grams

$$t(y_{i-1}, y_i)$$

實驗

- 利用TIMIT的語音資料庫

- 第一個實驗是假設所有的feature都可以被偵測到,藉此來觀察哪一種feature對detection-based的ASR有較好的潛力

實驗

Table 4. *The oracle phone recognition results derived by using different phonological feature sets.*

	Corr (%)	Acc (%)
SPE	93.28	93.20
MV	88.75	88.56
GP	98.39	98.36

實驗

- 在第二個實驗中使用HMM-based phone recognizer, 再用第一個實驗中經過人工作phone lable的訓練資料來訓練CRF, 將這個模型稱爲OT

- 但是發現Correction rate 雖然高但是Accuracy卻較低, 這個問題可能出在前端處理中的分類錯誤

實驗

- 欲解決此問題, 我們可以使用前端處理中偵測出來的結果來訓練CRF, 讓CRF可以學習到前端處理中的錯誤, 來降低訓練跟測試之間的mismatch, 稱此模型爲DT

實驗

Table 5. *The real phone recognition results derived by different recognizers, where OT means using oracle-data trained CRFs and DT means using detected-data trained CRFs.*

		Corr (%)	Acc (%)
HMM-based		69.02	63.45
OT detection-based	SPE	66.19	29.68
	MV	59.24	30.33
	GP	69.03	31.38
DT detection-based	SPE	56.56	55.27
	MV	51.84	50.68
	GP	55.74	54.53

實驗

- 接著可以從第一個實驗中發現, 在每個feature set中, 都會有一些容易混淆的phone, 但是每個feature set中容易被混淆的phone pair都不太一樣

- 接著可以用CRF來將這些不同的feature set的output做合併

實驗

Table 6. *Confusion pairs identified from the oracle phone recognition results by using different feature sets.*

Feature Set	#pair	Top 5 most confused pairs and their frequency counts
SPE	38	(iy,dh):1809 (z,aw):1236 (p,ey):956 (m,en):939 (f,v):911
MV	59	(iy,ih):995 (s,sh):395 (er,ah):394 (ey,iy):371 (ae,ah):315
GP	14	(el,sil):163 (uh,ah):126 (w,uw):64 (y,ih):39 (ah,sil):6

實驗

Table 7. *The real phone recognition results of the combined recognizers.*

Method	#sys	Corr (%)	Acc (%)
Baseline HMM	1	69.02	63.45
OT: SPE+MV+GP	3	61.97	**60.65**
DT: SPE+MV+GP	3	52.90	52.06
OT+DT: SPE+MV+GP	6	60.81	59.20
OT: SPE+MV+GP plus HMM	4	65.53	**64.31**
DT: SPE+MV+GP plus HMM	4	59.57	58.64
OT+DT: SPE+MV+GP plus HMM	7	64.22	62.59