Robust Speech Detection Based on Phoneme Recognition Features

Author:France Mihelic and Janez Žibert
Professor:陳嘉平
Reporter :吳柏鋒

摘要

- 簡介
- 以音素為基礎的特徵辨識
- Speech/Non-speech 的音段分割
- 實驗

簡介

- 使用音素辨識器來分出speech/non-speech音段的差異

- 提出主要以子音-母音(CV)與有聲-無聲(VU)這兩組配對來作為語音辨識器的測量基準

簡介

- Speech 音段

  -在音訊某範圍內含有人聲

- Non-speech音段

  -由一個或多個不同發音源組成(例如:音樂、機械噪音等)

以音素為基礎的特徵辨識

- 連續語音訊號透過音素辨識器處理產生的輸出，可以找出決定性的發音特徵

- 在此擷取發音特徵是以基本單元的時間持續率 (time duration)和改變率(changing rate)為基準

以音素為基礎的特徵辨識

(1)CV的時間持續率(time duration)作正規化:

$$\frac{|t_C - t_V|}{t_{CVS}} + \alpha \cdot \frac{t_S}{t_{CVS}}$$

其中 t_{CVS} 為整個訊號的持續時間，t_C 為子音的持續時間，t_V 為母音的持續時間，t_S 為安靜部分的時間，α 為用來增強安靜部分訊號的參數，$0 \leq \alpha \leq 1$

以音素為基礎的特徵辨識

(2) CV的Speaking rate作正規化 ：

$$\frac{n_C + n_V}{t_{CVS}}$$

其中 n_C 為子音單元數，n_V 為母音單元數
在此比較強調說話的方式，不考慮 S 單元數

以音素為基礎的特徵辨識

(3) **正規化 CVS的改變 ：**

$$\frac{c(C, V, S)}{t_{CVS}}$$

其中 $c(C, V, S)$ 為在 t_{CVS} 時間內 C, V and S 單元之改變次數

以音素為基礎的特徵辨識

(4) **正規化平均CV的duration rate :**

$$\frac{|\bar{t}_C - \bar{t}_V|}{\bar{t}_{CV}}$$

其中 $\bar{t}_C$ 和 $\bar{t}_V$ 分別表示給定的音段中 C和V單元的平均持續時間

Speech/non-speech音段分割

- 使用EM演算法找出適當GMMs，並建立以N 個HMMs模型連結成的網路，其中N表示用來作分類的GMM個數

- 每個HMM由內部幾個相同PDF的state連結建構而成且每個HMM會附加minimum duration

- 以HMMs為基本概念來達到分類且使用Viterbi decoding作分割

實驗

- 使用Slovenian data建構的Si-recognizer與TIMIT database建構的En-recognizer的兩種以音素為基礎的特徵辨識器

- 使用語料庫:

  (1)SiBN（斯洛維尼亞語的廣播新聞）

  (2)COST278BN（9個歐洲語系的廣播新聞）

實驗

- 設定運算數據的門檻和參數值來達到最佳化的運算效能

- 由(1)~(4)計算出CVS特徵，並以frame-by-frame對特徵向量作分類

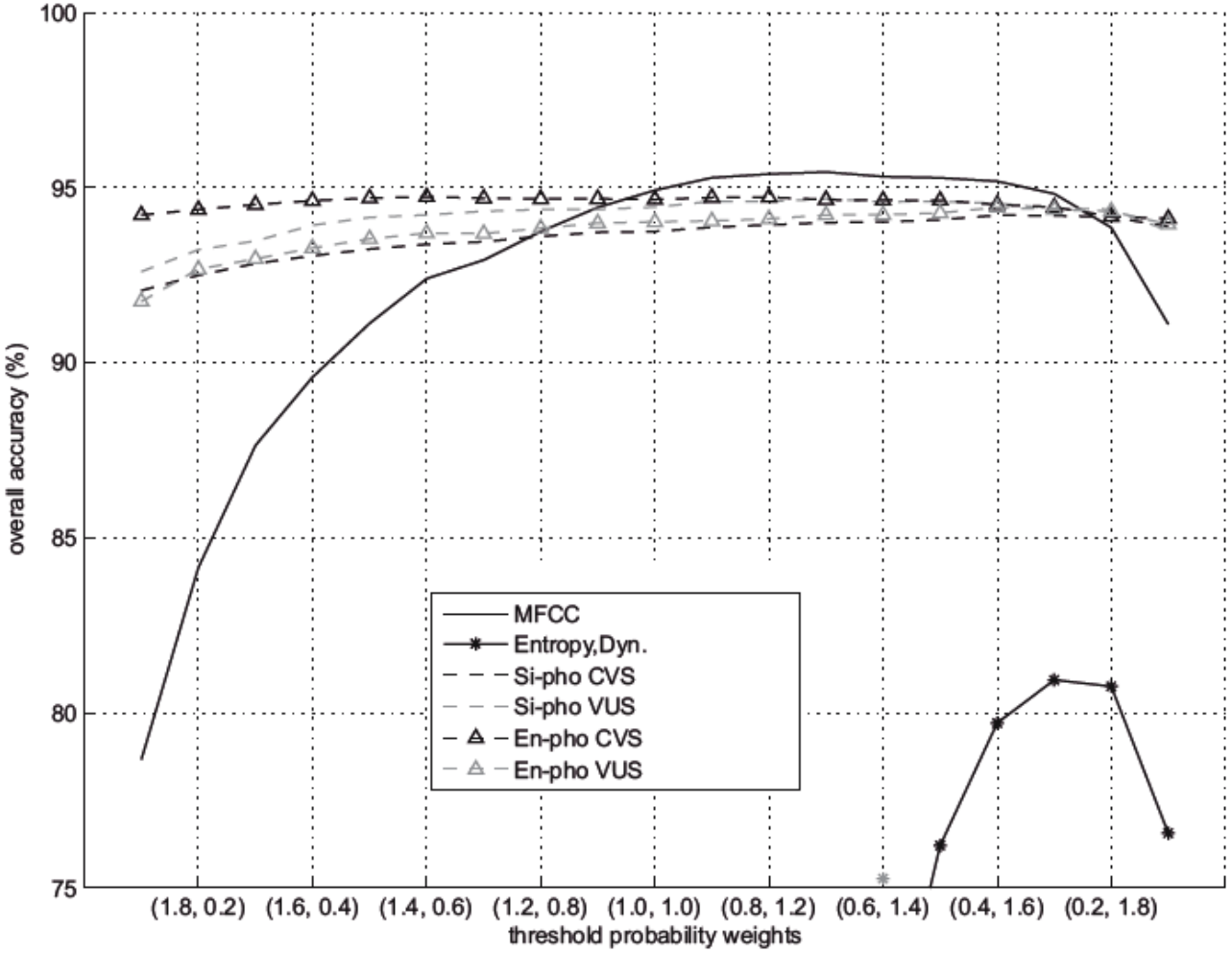

Fig. 2. Setting the optimal threshold weights of speech and non-speech models to maximize overall accuracy of different representations and approaches. (*MFCC*) – 12 MFCC features with energy and first delta coefficients modeled by 128 mixture GMMs. (*Entropy, Dyn.*) – entropy and dynamism features modeled by 4 mixture GMMs, (*Si-pho CVS, Si-pho VUS, En-pho CVS, En-pho VUS*) phonemes feature representations based on CVS and VUS phoneme groups obtained from Slovenian (Si) and English (En) phoneme recognizers.

實驗

Table 1. Speech/non-speech classification results on SiBN and COST278 dataset. Values in brackets denote results obtained from non-optimal models using equal threshold probability weights.

Features Type	SiBN dataset			COST278 dataset		
	Speech	*Non-Speech*	*Accuracy*	*Speech*	*Non-Speech*	*Accuracy*
MFCC	97.9 (96.4)	58.7 (72.3)	95.3 (94.8)	98.7 (97.8)	44.0 (54.2)	94.6 (94.6)
Entropy, Dyn.	99.3 (89.9)	55.8 (93.8)	96.5 (90.1)	99.6 (83.1)	37.4 (84.7)	95.0 (83.2)
Si-pho, CVS	98.2 (97.6)	91.1 (93.0)	97.8 (97.3)	96.6 (95.6)	76.9 (79.3)	95.1 (94.3)
Si-pho, VUS	98.1 (97.7)	88.7 (90.1)	97.5 (97.2)	97.2 (96.6)	72.2 (74.3)	95.3 (95.0)
En-pho, CVS	98.5 (98.4)	88.2 (88.8)	**97.8** (97.7)	97.9 (97.8)	71.1 (71.6)	**95.9** (95.8)
En-pho, VUS	97.5 (96.7)	90.0 (92.9)	97.0 (96.4)	96.8 (96.6)	72.4 (74.3)	95.0 (95.0)