Robust Endpoint Detection for Speech Recognition Based On Discriminative Feature Extraction

Author: Koichi Yamamoto ,
Klau Reinhard
Professor: 陳嘉平
Reporter : 許峰閤

大綱

- 介紹
- 聲音偵測(Voice Activty Detection)
- 端點偵測(Endpoint Detection)
- 實驗

介紹

- 端點偵測在語音辨識中可以增進效能，降低計算量，因為如果端點偵測都正確，只會有有用的speech frame進入後端處理

- 在以往都使用energy-based但是該方法在low SNR時準確度並不高，所以在這裡結合了log-likelihood的方法

Voice Activity Detection

- 在以能量為基礎的聲音偵測中,當log-energy超過門檻值時,變被歸類為speech否則為non-speech

- 該門檻值用以下算式界定

$$E_{noise}(t) = \lambda E_{noise}(t-1) + (1-\lambda)E(t)$$

E(t) 是frame t的log-energy λ 是 forgetting factor

$$T_e(t) = E_{noise}(t) + \gamma$$

Voice Activity Detection

- Likelihood-based使用GMM為分類器，首先要訓練speech及non-speech的GMM，non-speech及speech的log-likelihood ratio如下

$$L(t) = g_1(\mathbf{y}(t); \Lambda) - g_0(\mathbf{y}(t); \Lambda)$$

- g_0 與 g_1 分別代表non-speech及speech GMM的log-likelihood
- y(t) 代表feature vector for frame t

Endpoint Detection

- 爲了增加噪音環境時的強建性,在QBNF (quantile based noise estimation)之前,先用spectral subtraction(SS)爲前置處理

$$\hat{S}(k,t) = \max\{X(k,t) - \alpha\hat{N}(k,t), \beta X(k,t)\}$$

- X(k,t)代表 kth-PSD of noisy signal at frame t
- $\widehat{N}$ (k,t)代表 kth-PSD of noise estimation by QBNE
- $\widehat{S}$ (k,t)代表 kth-PSD of enhanced input signal

Endpoint Detection

- 每個frame的log-energy用下式得到

$$E(t) = \log \sum_{k=K_L}^{K_H} \hat{S}(k,t)$$

K_H 及 K_L 分別代表最高及最低的frequency component

Endpoint Detection

- 用log mel-filterbank來取得GMM的feature vector如下:

$$\mathbf{x}(t) = \left[x_1(t), \cdots, x_N(t), \Delta_1(t) \cdots, \Delta_N(t)\right]^T$$

N代表mel-filterbank的數量

$x_n(t)$ 爲n-th log mel-filterbank的energy

Endpoint Detection

- 接著將feature vector減掉每個frame的平均來作normalized得到下式

$$\bar{\mathbf{x}}(t) = \left[\bar{x}_1(t), \cdots, \bar{x}_N(t), \Delta_1(t) \cdots, \Delta_N(t)\right]^T$$

- 再將 $\bar{X}(t)$ 投影到lower feature vector y(t)來降低計算量

$$\mathbf{y}(t) = \mathbf{P}\bar{\mathbf{x}}(t)$$

- P為一個M * 2N的投影矩陣利用Principle component analysis(PCA) 求出

Endpoint Detection

- 接著就可以判斷該frame是否爲speech,只要符合下式該frame即爲speech

$$E(t) > T_e(t) \quad \& \quad L(t) > T_l(t)$$

- 作完VAD以後利用finite-state automaton決定start-of-speech和end-of-speech

Endpoint Detection

- 在之前我們知道投影矩陣使用PCA來求得,而GMM則使用EM alg.來訓練,但是這些方法並不是基於可以將speech及non-speech的分類得到最小的錯誤率,所以在這裡提出discriminative feature extraction
- DFE是based on minimum classification error/generalized probabilistic descent (MEC/GPD)

Endpoint Detection

- The frame-based misclassification measure of the likilihood ratio :

$$d = -g_j(\mathbf{y}(t);\Lambda) + g_{i \neq j}(\mathbf{y}(t);\Lambda)$$

$$\mathbf{y}(t) \in C_j \quad \text{and} \quad i, j \in [0,1]$$

C_j 是兩個分類分別為speech and non-speech
如果該frame分類正確則d會是負的

Endpoint Detection

- 由上式可以得到DFE的loss function

$$l = \frac{1}{1+\exp(-\tau d)}$$

τ 是控制sigmoid function的斜率

所有投影矩陣及GMM中的參數都設為 ϕ

ϕ is updated base on MCE/GPD traning rule:

$$\Phi[t+1] = \Phi[t] - \varepsilon_t \nabla_{\Phi} l(\overline{\mathbf{x}}(t); \Phi[t])$$

實驗

- 用來訓練投影矩陣及GMM的資料,分別有speech及noise的data
- Speech data 爲clean環境的三千句句子
- Noise data使用JEIDA noise database
- 在混和成noisy data

實驗

- Input signal 的 sample rate 11025Hz
- K_L 及 K_H 分別爲 130 Hz 和 4900Hz
- 有24個mel-filterbank
- Feature vector的dimension 爲16

實驗

Table 1. The statistical information of the histograms, where each value represents the rate (%) of the distribution.

Conditions	Clean				Car 5dB				Babble 5dB			
The difference of the number of frames	SOS		EOS		SOS		EOS		SOS		EOS	
	≤10	≤30	≤10	≤30	≤10	≤30	≤10	≤30	≤10	≤30	≤10	≤30
Energy	96.7	99.7	91.7	99.1	59.5	79.7	60.3	78.4	57.1	77.0	56.9	76.3
Proposed without DFE	94.0	98.9	92.7	98.2	67.5	82.5	60.0	79.6	63.3	78.0	60.2	78.1
Proposed with DFE	95.9	99.1	92.5	98.0	79.6	92.2	73.8	90.6	79.5	91.6	74.3	91.6

實驗

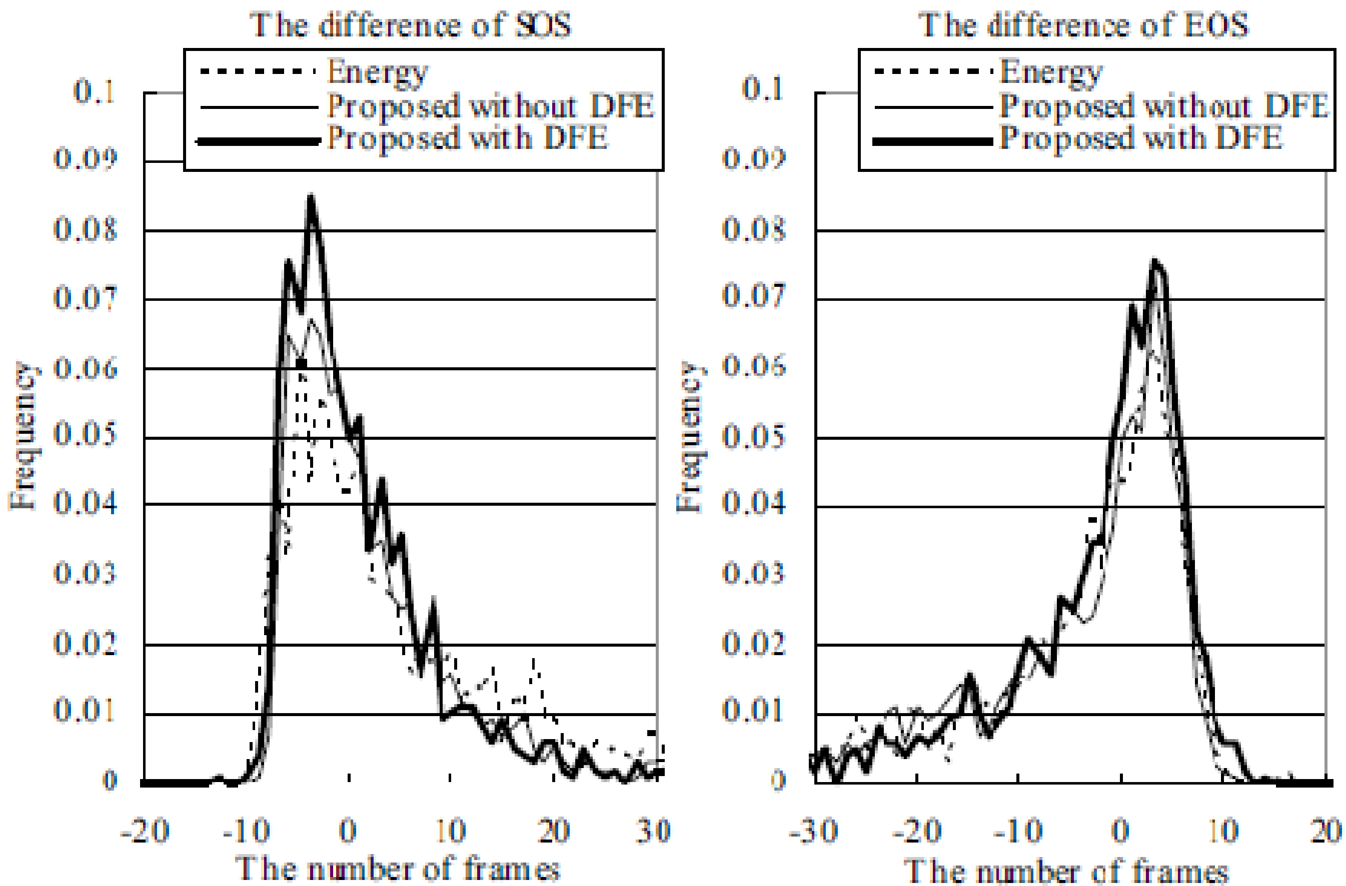
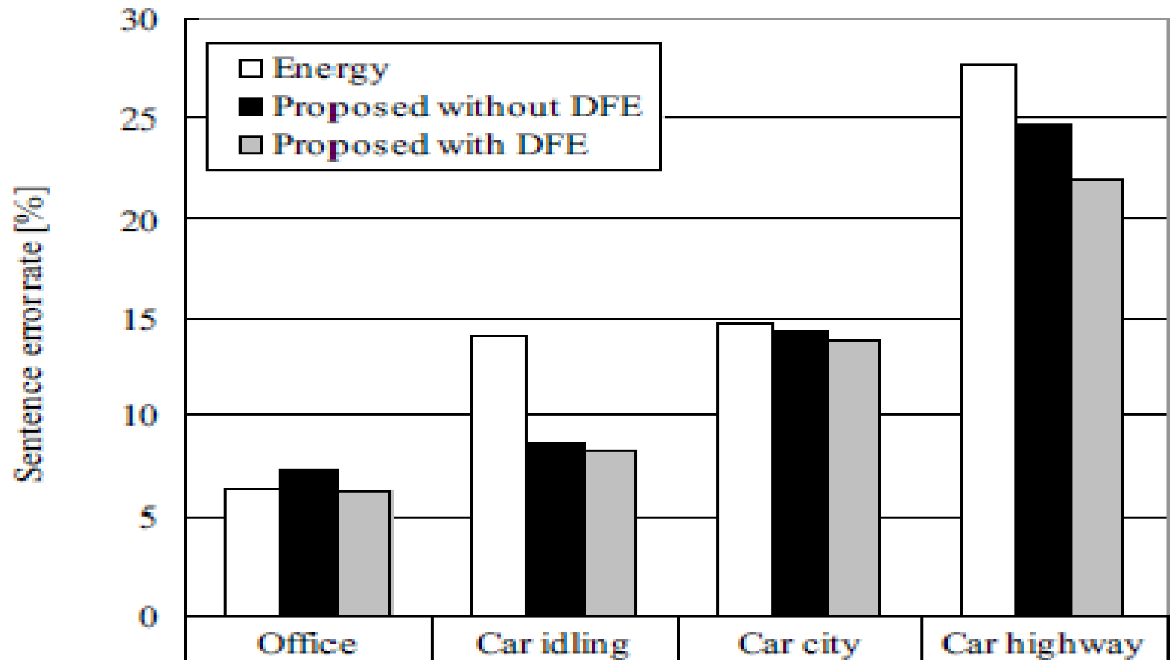


Figure 1. The histograms of the differences (# of frames) between manually labeled and detected endpoints: SOS (left) and EOS (right) points for 5dB SNR car noise.

實驗



	Office	Car idling	Car city	Car highway
□ Energy	6.31	14.08	14.66	27.7
■ Proposed without DFE	7.31	8.71	14.29	24.62
▒ Proposed with DFE	6.21	8.27	13.85	21.86

Recording environments

Figure 2. The sentence error rate of the ASR for the four recording environments.