改良式統計圖等化法於強健性語音辨識之研究

Improved Histogram Equalization Methods for Robust Speech Recognition

謝欣汝 Hsin-Ju Hsieh[1, 2], 洪志偉 Jeih-weih Hung[2], 陳柏琳 Berlin Chen[1]
[1] 國立臺灣師範大學資訊工程學系
[2] 國立暨南國際大學電機工程學系
hsinju@ntnu.edu.tw, berlin@ntnu.edu.tw, jwhung@ncnu.edu.tw

摘要

統計圖等化法(Histogram Equalization, HEQ)[1]是一種概念簡單且有效的語音特徵處理技術，近年來被廣泛地研究與應用於強健性語音辨識的領域。在本論文中，我們延續統計圖等化法的研究，提出一系列使用語音特徵的空間－時間之文脈統計資訊(Spatial-Temporal Contextual Statistics)的語音特徵強健方法，這些方法主要的架構是利用一個簡易的差分(Differencing)和平均(Averaging)的處理方式，對語音之倒頻譜特徵的空間域與時間域加以分割，以擷取出語音特徵在空間域與時間域上不同頻率成分之統計資訊後，將其分別作統計正規化處理並結合，來達到降低雜訊對語音特徵所造成的影響。其所用的差分和平均的公式如下所示:

$$x_{s-diff}(d,t)=\begin{cases}\dfrac{x(d,t)-x(d-1,t)}{2}, & 2\le d\le D\\ x(d,t), & d=1\end{cases}$$

$$x_{s-avg}(d,t)=\begin{cases}\dfrac{x(d,t)+x(d-1,t)}{2}, & 2\le d\le D\\ 0, & d=1\end{cases}$$

其中 $x_{s-diff}(d,t)$ 與 $x_{s-avg}(d,t)$ 分別表示從原始語音特徵之空間域上所擷取出的高頻和低頻的統計資訊。同樣地，將此處理方式作用於同一維度之任意兩個相鄰的音框，亦可得到原始語音特徵在時間域上之高頻 $x_{t-diff}(d,t)$ 和低頻 $x_{t-avg}(d,t)$ 的文脈統計資訊。此外，本論文另外提出一個變型的方法，將空間域和時間域上所求得的高頻特徵 $x_{diff}(d,t)$ 及低頻特徵 $x_{avg}(d,t)$ 以線性加權的方式結合，來觀察辨識率是否有進一步提升的空間。

有別於傳統運用於語音特徵時間序列上之各別維度獨立正規化(Dimension-Wise)的方法例如:倒頻譜平均值消去法(Cepstral Mean Subtraction, CMS)[2]、倒頻譜平均值與變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)[3]等，本論文所提出的一系列新方法能進一步地正規化不同空間與時間之間的特徵分布資訊，能更有效的降低不同聲學環境所產生的偏差並且嘗試消除傳統之統計圖等化法無法補償的問題，亦即隨機性雜訊對語音所產生的影響。值得注意的是，對於語音特徵之時間域或空間域上的正規化處理方式，過去已有學者提出概念類似的語音特徵之單一域的正規化處理技術[4-5]，

而本論文所提出之結合式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊的技術，於目前為止則是相對較少被研究與探討的議題。

本論文的辨識實驗是作用於國際通用的語音語料庫 Aurora-2[6]上，我們驗證了所提出之新方法能夠大幅提升各種雜訊環境下之語音辨識的精確度。其辨識效能都明顯高於許多傳統作用於語音特徵之時間序列上的正規化處理技術與只單獨正規化語音特徵之時間域或空間域的結果。此外以線性加權空間域與時間域上所求得的高頻特徵 $x_{diff}(d,t)$ 及低頻特徵 $x_{avg}(d,t)$ 的組合方式，使得辨識率從原始未加權的 83.33%進步至 85.05%，其絕對錯誤降低率為 1.72%。最終進一步地結合進階式前端標準(Advanced Front-End Standard, AFE)[7]強健性語音特徵，足足能使辨識率從原始的 87.17%提升至 88.22%，相對錯誤降低率約有 8%，足見這些新方法能有效提升語音特徵的強健性。

關鍵詞：自動語音辨識，雜訊強健性，統計圖等化法，特徵文脈的統計

Keywords: automatic speech recognition, noise robustness, histogram equalization, feature contextual statistics.

致謝：本論文之研究承蒙教育部－國立台灣師範大學邁向頂尖大學計畫（101J1A0900 和 101J1A0901）與行政院國家科學委員會研究計畫(NSC 101-2221-E-003 -024 -MY3 和 NSC 99 -2221-E-003 -017 -MY3)之經費支持，謹此致謝。

參考文獻

[1] Angel de la Torre, Antonio M. Peinado, Jose C. Segura, Jose L. Perez-Cordoba, Ma Carmen Benitez and Antonio J. Rubio, “Histogram equalization of speech representation for robust speech recognition”, IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 3, pp. 355-366, 2005.

[2] S. Furui, “Cepstral analysis technique for automatic speaker verification”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 29, No. 2, pp. 254-272, 1981.

[3] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition”, Speech Communication, Vol. 25, No. 1-3, pp. 133-147, 1998.

[4] J. W. Hung and H. T. Fan, “Subband feature statistics normalization techniques based on a discrete wavelet transform for robust speech recognition”, Signal Processing Letters, IEEE, Vol. 16, No. 9, 2009.

[5] V. Joshi, R. Bilgi, S. Umesh, L. Garcia and C. Benitez, “Sub-band level histogram equalization for robust speech recognition”, 12th Annual Conference of the International Speech Communication Association (Interspeech), 2011.

[6] H-G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Automatic Speech Recognition: Challenges for the Next Millennium, pp. 181-188, 2000.

[7] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce and F. Saadoun, “Evaluation of a noise-robust DSR front-end on Aurora databases”, 3th Annual Conference of the International Speech Communication Association (Interspeech), 2002.