車載系統應用之強健性語音辨識研究

王文俊*、陳保清、黃英峰
中華電信研究院 匯流服務研究所

摘要 — 本篇報告是探討應用於車載系統之強健性語音辨識技術的幾項相關議題，包括適合的語音模型訓練方式、不同麥克風陣列收音裝置之效能比較以及進一步結合雜訊消除處理之效益等。相關實驗是在本單位所開發之車載系統語音點歌應用服務進行，藉由外加汽車雜訊之語料與實際車內錄製之語料觀察語音辨識效能以最佳化此應用服務之處理流程。

研究背景

對於眼忙(eyes-busied)與手忙(hands-busied)的駕駛行為而言，語音辨識功能是各類車載應用系統必備之人機界面，而如何在高速行駛的車內雜訊環境有效地降低雜訊以提升語音辨識正確率則為最迫切的課題。本篇報告嘗試就三個有關強健性語音辨識技術之議題進行測試，以調整並提升本單位所開發之車載系統語音點歌應用服務之辨識正確率。此三個議題包括選擇適合的語音模型訓練方式、比較不同麥克風陣列收音裝置之效能以及觀察進一步結合雜訊消除處理之效益等。

提出方法

本篇報告的第一個重點是觀察多重狀況訓練法(multi-condition training)[1]之效能，此方法是希望降低語音辨識工作在訓練與測試階段所存在的不匹配(mismatch)現象。對於車內環境應用而言，就是希望在語音模型訓練階段加入不同 SNR 值的汽車雜訊於訓練語料，以提升訓練所得之語音模型的強健性，進而期望在實際車內雜訊環境獲得令人滿意的辨識率。

第二個重點是探討車內環境之收音問題，對於一般環境之語音辨識而言，近距離收音方式應能優於遠距離收音方式，因此頭戴式免持麥克風應是駕駛可採用之近距離收音方法。另外遠距離收音方式之相關研究近年來也持續受到關注，本篇報告之實驗就包括此兩類收音裝置之效能比較，同時為能同時有效降低雜訊，此兩類收音裝置均採用多個麥克風之架構以建立聲源定位與語音強化之功能。

第三個重點則是針對受雜訊汙染之信號或是麥克風陣列處理後之信號加入有效的頻譜處理方法，也就是 Wiener Filter 處理，以進一步消除殘餘雜訊(residual noise)[2]並觀察其效能。

實驗結果

實驗語料區分為訓練語料與測試語料兩部份，訓練語料是本單位歷年透過市話與行動電話所蒐集的多語者國語常用詞與短句語料。測試語料則是為測試開發中的車載系統語音點歌應用服務之效能所建立之語料，區分成如下 3 個測試集：

Set A：外加不同 SNR 值之汽車雜訊於辦公室環境錄製。

Set B：使用遠距麥克風陣列(Andrea DA-350)於車內錄製。

Set C：使用頭戴藍牙麥克風(SC VMX-100)於車內錄製。

相關實驗所比較的兩個語音模型分別為利用混合不同頻道特性之訓練語料所建立的基本語音模型(MIX)，以及採用外加不同 SNR 值之 ETSI AURORA 汽車雜訊[1]於上述訓練語料的 multi-condition training 方式所建立的語音模型(CAR)。表 I 為比較此兩種語音模型針對 Set A 之效能，同時也加入是否結合 Wiener Filter 處理之比較。表 II 與表 III 則分別為針對 Set B 與 Set C 所作之效能比較。

實驗結果顯示外加特定雜訊之語音模型訓練方式並未獲得較佳之結果，其可能理由之一為實際車內雜訊與外加於訓練語料之 ETSI AURORA 汽車雜訊仍存在顯著差異。另一理由為外加特定雜訊之語音模型訓練方式為將訓練語料分割再加入不同 SNR 值之特定雜訊，此方式雖能降低訓練與測試間的 mismatch 現象，但也有可能造成語音模型之污染，充分的訓練語料以提高語音模型精確度才應為關鍵所在。

在 Set A 之高雜訊環境下，Wiener Filter 有明顯的效能提升，但在 Set B 與 C 的實驗則顯示，由於麥克風陣列與藍牙麥克風已大幅有效的降低車內雜訊，致使 Wiener Filter 所能提供的效能再提升的幅度就不明顯。

另外根據 Set B 與 C 的 Top 1 與 Top 10 的實驗結果，頭戴式藍牙麥克風所獲得之效能確實優於遠距收音之麥克風陣列。

表 I：Set A 之實驗結果(%)

SNR	Filter	Top 1		Top 5		Top 10	
		MIX	CAR	MIX	CAR	MIX	CAR
Clean	NA	92	90	97	95	98	97
	Wiener	93	89	97	94	98	96
10 dB	NA	57	31	72	54	79	61
	Wiener	72	66	86	80	89	82
0 dB	NA	5	1	12	6	17	8
	Wiener	24	14	41	31	46	37

表 II：Set B 之實驗結果(%)，依 Set A 之實驗結果採用 MIX model

Filter	Top 1	Top 5	Top 10
NA	86	94	95
Wiener	87	92	94

表 III：Set C 之實驗結果(%)，依 Set A 之實驗結果採用 MIX model

Filter	Top 1	Top 5	Top 10
NA	92	94	97
Wiener	90	94	97

參考資料

[1] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", Proceedings of ISCA IIWR ASR2000, Paris, France, pp.181-188, 2000

[2] 洪維廷、陳弘啓、陳軍廷、廖宜斌，『麥克風陣列技術應用於語音加強及語音辨識之研究』，2008 年全國電信研討會。

Corresponding author:wernjun@cht.com.tw
DOI:10.6159/IJSE.2014.(4-1).29