An Investigation on Robust Confidence Measure and Model Compensations for Smartphone-based Speech Recognition

Wei-Tyng Hong
Department of Communications Engineering, Yuan Ze University, Taiwan
E-MAIL: wthong@saturn.yzu.edu.tw

Abstract

This paper presents a structural design for robust HMM-based speech recognition system under PDA/smart-phone platforms. Two main issues were investigated: (1) how to implement the CM (confidence measure) under noisy environment with resource-constrained platform? (2) how to perform model compensation for improving recognition rate under noisy environment with resource-constrained platform? We have gone through algorithm modification, fixed-point analysis and porting procedures to PDA/smart-phone. The final testing is performed with the Dopod 818pro machine. As the proposed design of robust modules need not consume many computing resources, it is conceivably suitable for speech recognition of resource-constrained platforms.

Keywords: robust speech recognition, confidence measure, model compensation.

中文摘要

本文針對 PDA/智慧型手機運算資源受限的平台，提出以 HMM 為基礎之語音辨認核心的強健式技術。我們主要探討的問題如下: (1) 如何在資源受限的移動式裝置與雜訊環境下進行信賴度分析 (2) 如何在資源受限的移動式裝置下進行模型補償，進而達到效能提升的目的。整個設計流程經歷過演算法修改、定點數分析、移植到 PDA/智慧型手機驗證等程序，最終將相關程式模組移植到 Dopod 818Pro 進行實機測試。由於本文提出的設計能不需耗用過多的資源下達到一定程度的抗雜訊能力，咸信非常適合作為運算資源受限的平台之語音辨認核心。

關鍵詞：強健式語音辨認、信賴度分析、模型調適。

1. 緒論

近年來標榜智慧型「移動式裝置」(mobile devices)產品，例如 PDA/Smart-Phone，隨著 3G/4G 時代的來臨，以及嵌入式微處理器與其他晶片在性能日漸強大且整合晶片大小卻日漸縮小的趨勢下開始蓬勃發展，其中尤以智慧型手機(Smart-Phone or PDA Phone)的前景看俏。根據資策會產業情報研究所（MIC）的調查顯示，預計 2014 年智慧型行動電話出貨量將接近 6.22 億支的規模。這類裝置輕薄短小且強調移動便利性，語音輸入法是此類裝置最適宜的人機界面。因此，應用於移動式智慧型裝置 (e.g., 車機/PDA/smart phone 等) 上之語音介面系統，在未來幾年內將是急迫且重要的研究重點，也將對移動裝置的應用造成重大的影響。

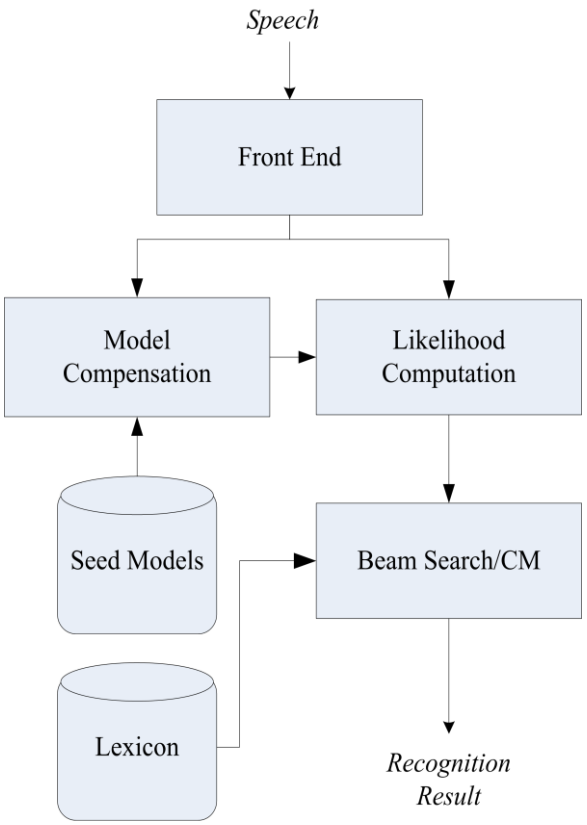
運算資源受限的抗雜訊強健式技術，是關係到移動式裝置語音辨認核心是否能突破的關鍵。本論文以智慧型手機為運算環境，以國語(Mandarin)語音辨認為標的，開發語音辨認之抗雜訊強健式技術。因為移動式裝置的無所

不在(anywhere)的特性，使用者可能在車內或街上使用語音辨認器，故系統需對環境雜訊，通道變化(例如 Hands-free 式麥克風 far-field 效應/低品質麥克風裝置)要有強健能力[1]。文獻[2]對移動式裝置語音辨認器的三種架構(分別是 embedded, network 和 distributed-based 架構)提出優缺點。本論文要發展的語音辨認核心是屬於 embedded 架構，也是對系統資源最敏感的架構。文獻[3-6]探討語音辨認系統定點化、前處理求參數時的簡化運算和辨認搜尋網路大小的縮減等問題。為了降低語音核心佔用的 storage 資源，文獻[7]採用 state-clustered tied-mixture 方法來製作密實模型(compact HMM)，在不破壞效能的前提下，減少語音辨認系統的 HMM 參數大小。在抗雜訊能力方面，文獻[8]提出使用多樣式訓練(multistyle training)來增加語音模型抗雜訊能力，並用雙麥克風式的 adaptive decorrelation filtering 技術，在 cepstral domain 上配合 multi-channel codebook dependent normalization 將雜訊訊號濾除，用以增加 PDA-based 語音辨認系統抗雜訊效能。文獻[9]提出一個外掛計算資源之 PDA 麥克風陣列的架構配合分散式語音辨認系統。前端的麥克風陣列用來做多通道(multi-channel)收音，這些多通道資料經由網路傳到後端伺服器進行 adaptive beamforming 與 noise suppression 而得到加強後語音(enhanced speech)再進入語音辨認器，但此文獻所提出的方法所需計算資源仍太大。文獻[10]提出一個結合 instantaneous MLLR[11]和 TDCM-based delay-and-sum beamformer 的語音辨認技術來達成抗車雜訊能力，但 instantaneous MLLR 需要 2 個程序，首先做語音切割並需有額外空間儲存切割資訊，然後進行模型轉換，模型轉換時牽涉到反矩陣等計算，這對資源受限裝置是相當耗資源的計算，不適合於 embedded-based 語音辨認架構。根據以上文獻探討，本論文聚焦到二個問題: (1) 如何在資源受限的移動式裝置與雜訊環境下進行信賴度分析? (2)如何在資源受限的移動式裝置下進行模型補償，進而達到效能提升的目的?

2. 系統架構

如圖 1 所示，為一個標準的以「模型補償」(Model Compensation)為抗雜訊技術的語音辨認流程圖。語音輸入 *Front End*(前處理)模組後，得到一連串的語音辨認參數。在模型補償的程序中，此參數將進入 *Model Compensation* (模型補償)模組，將 *Seed Models* (種子模型)轉換成符合目前環境的語音模型; *Likelihood Computation* (相似值計算)模組依此補償後模型對語音辨認參數求取每個音框(frame)的相似值。接著，此一連串的音框式(frame-leve)的相似值進入 *Beam Search* 模組配合 *CM*(Confidence Measure)模組過濾不可靠之辨認結果。

圖 1 結合模型補償與信賴度模組的語音辨認流程圖



在 HMM 觀測機率分佈和其 log-likelihood 計算方面，為了簡化後續計算計算量，我們採用 Laplacian Distribution 作為 HMM 的觀測機率分佈，其機率密度函數定義如下:

$$Lap(x;u,v)=\frac{1}{2v}\exp\left(-\frac{|x-u|}{v}\right) \qquad (1)$$

其中 u 為 location parameter，v 為 scale parameter。假設一個 D-dimensional 特徵值 x

的 covariance matrice 為 diagonal，我們以此狀態(state)中所有 mixture 最大的機率值來代表 HMM 第 j 個狀態的觀測機率:

$$b_j(\vec{x}) = \max_m \left\{ \prod_{d=1}^{D} Lap\left(x_d; u_{j,m,d}, v_{j,m,d}\right) \right\} = \max_m \left\{ \prod_{d=1}^{D} \frac{1}{2v_{j,m,d}} \exp\left(-\frac{\left|x_d - u_{j,m,d}\right|}{v_{j,m,d}} \right) \right\} \quad (2)$$

其中 $u_{j,m,d}$ 和 $v_{j,m,d}$ 分別為 HMM 第 j 狀態上第 m 個 mixture 之第 d 維度的 location parameter 和 scale parameter。假設 k 滿足下列式子:

$$k = \arg\max_m \left\{ \prod_{d=1}^{D} Lap\left(x_d; u_{j,m,d}, v_{j,m,d}\right) \right\} \quad (3)$$

則可得到如下的式子:

$$\log(b_j(\vec{x})) = C_{j,k} - \sum_{d=1}^{D} \frac{\left|x_d - u_{j,k,d}\right|}{v_{j,k,d}} \quad (4)$$

其中 $C_{j,k}$ 為和 HMM 第 j 狀態第 k 個 mixture 的參數相關的數值(和特徵值 x 無關)，可預先算出作為另一個模型參數。除法運算在嵌入式系統下是相當耗費運算資源，因此藉由下列方法把觀測相似值的計算中的除法以 bit-shift right 運算元取代。令整數 $\beta_{j,k,d} = 2^s / v'_{j,k,d}$ ，經過適當式子重整，得到下列的全整數運算逼近式:

$$\log(b_j(\vec{x})) \approx C'_{j,k} - \left[\sum_{d=1}^{D} \beta_{j,k,d} \times \left| x'_d - u'_{j,k,d} \right| \right] >> s \quad (5)$$

其中>>為 bit-shift right 運算元。

在信賴度(CM)分析模組方面，我們採用 Single Layer Perceptron (SLP) [12]為基礎的分類器來偵測 OOV(Out-of-Vocabulary)語句。選取 5 種 CM 特徵參數，分別來自辨認搜尋後得到的相似度分數(Likelihood Score, 簡稱 LS)和詞分數(Word-level Likelihood Score, 簡稱 WLS)，且分別進行音匡正規化(frame normalization)。為了符合資源限制下的運算環境，這些特徵參數都直接從搜尋過程中計算得到，不需進行額外的後追蹤(back trace)過程。CM 模組的 5 個特徵參數如下:

1. Difference of LS (Top1-Top2)
2. Difference of LS (Top1-avg(Top3…N))
3. Difference of WLS (Top1-Top2)
4. Difference of WLS(Top1-avg(Top3…N))
5. Variance of Top1's WLS

CM 模組的訓練階段，首先對訓練用的測試庫進行一次識別過程，得到各語句的 5 個 CM 特徵參數 $\boldsymbol{X}$ 和識別結果。利用其中正確的識別結果，每個結果有 IV 和 OOV 兩組 feature 參數，送入 SLP 的訓練過程進行訓練，經過 iteration 得到 SLP 的參數矩陣 $\boldsymbol{W}$。在 CM 辨認階段，將 W 應用到識別過程中，即可對每個識別結果進行句子驗證(Utterance Verification)或 IV/OOV 的判決，即計算 $\mathbf{Y}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$，根據預先確定的臨界值(Th) 進行如下之決策: if Y>=Th, Accept (i.e., decided as IV); else Reject (i.e., decided as OOV)。

從計算量上，SLP 只增加了 5 次乘法運算。另外，為了增加效能，我們將搜尋中的前 10 個競爭模型的相似度分數當作線上垃圾分數(online garbage score)，並與目前的模型分數相減，形成額外第 6 個 CM 特徵參數。此方法中的線上垃圾分數，是在搜索的過程中利用候選模型的競爭模型組成備選假設的方法進行 CM。在應用時需考慮系統實現的計算量等問題，因此選擇次音節單位(sub-syllable)為線上垃圾分數計算單元。具體實現方法是：對識別候選中的每一個次音節，在所有候選詞裡尋找其競爭單元(與目前次音節不同且非靜音模型)。每找到一個競爭單元，將其加入競爭集。最後對所有競爭候選按相似度排名，取前 10 個候選的平均值或中位值作為線上垃圾分數的得分，再以當前候選次音節的得分減去其線上垃圾分數作為一個新的 CM 參數，加入到原來的 5 個參數中，然後通過 SLP 進行決策。

在模型補償模組上，我們採用 MLLR[11]方法，並將之修改符合資源限制下的智慧型手機運算環境。MLLR 是一種對模型參數進行線性變換，使之適應新的說話人或者環境的方法。它使用一定數量的調適資料，根據 maximum likelihood 準則，得到對模型參數進行變換的轉換矩陣(Transformation Matrix)。關鍵的一點是，它將原有模型劃分成一定數量的 regression 類別，每個模型歸入其中的一個類別，每個類別 的所有成員共用同一個變換矩陣，這樣就克服了調適資料數量不足帶來的問題。因此 MLLR 具有自適應速度快，只有少量資料也能得到較好性能的特點。類別的劃分是 MLLR 中的一個重點。如果類別劃分為一類，稱之為 Global Transformation ，如果類別數目等於 model 數目，即每個 model 單獨

一個轉換矩陣，則 MLLR 就相當於 MAP。實際劃分類數介於這兩種方案之間，合理的類別數與調適資料數量有關。

3. 實驗分析

我們以 MAT2000 [13]為訓練語料，採用 100 個 *Initial* 及 38 個 *Final* 模型，並以 Laplacian Distribution 作為 HMM 的觀測機率分佈，在 CIVIC 6 dB 及 ROVER 0 dB 車雜訊環境[14]下，針對 1000 詞人名辨認做實驗，結果如表 1 所示，其操作點設定在平均拒絕綠在 10%以下，CM1 表示 SLP 的作法，CM2 表示 SLP 整合線上垃圾分數的方法。我們發現，由此方法建構的 CM 模組具備抗雜訊能力，CM 的結果錯誤比辨認系統本身錯誤要少得多，也就是說可以糾正原系統的一些錯誤，說明採用的方法是合理有效的，並且加入線上垃圾分數參數要比原方法對 CM 總體性能有明顯的提升。

表 1 在車雜訊下 CM1 及 CM2 的效能，其中數字代表系統整體辨認率 (%)

	Baseline	CM1	CM2
CIVIC 6 dB	87.8	90.6	92.1
ROVER 0 dB	84.5	88.6	90.6

本論文實現了用 MLLR 方法對 Laplacian Distribution 作為 HMM 觀測機率分佈的模型參數進行轉換。在具體實現中的重點如下: (1) 只對 mean 參數進行了變換。這一方面是因為 mean 參數在 HMM 參數中有主導作用，只需變換 mean 參數就能得到很好的性能；另一方面也是因為我們在定點模型中的 Laplacian 分布之 bias 參數為了避開除法運算，已經過正規化，存在一定的量化誤差，若再調整此 bias 參數由勢必帶來更大的誤差。(2) 在實現過程中，一律視 bias 參數為對角線矩陣來增進運算速度，避開複雜的矩陣運算。

對 MLLR 的性能測試的實驗語句一樣是由 iPAQ PDA 上進行錄製。此語料庫包含了 10 個語者(8 男 2 女)，每個人有 61 句話，從中最多抽取 20 句語句作為調適語句，其餘 41 句話則作為測試語句。實驗針對調適語句數量 10、15、20 下分別進行，考慮以對聲學分類和語音學分類方式來進行類別決定，結果如表 2，表中列的數字為相對於 baseline 系統的辨認正確率改進幅度。從結果來看，修改後的 MLLR 能有效地提升系統的性能，調適語句越多，性能越好，聲學分類方法比語音學分類結果要好一些，且此演算法符合在資源受限的移動式裝置下進行模型補償的需求。

表 2 不同的調適語句與分類方式的 MLLR 模組效能（表內數字為相對於 baseline 系統的辨認正確率改進幅度）

調適語句數	聲學分類	語音學分類
10	15.8 %	13.6%`
15	30.0 %	27.3%
20	42.2 %	36.4%

4. 結論

資源受限系統上之語音辨識系統，在未來幾年內將是所有移動式裝置(包括車機、智慧型手機等)極為重要的人機介面技術，此技術也將對移動式裝置的應用有重大的影響，一如圖形介面介面對 PC 產業造成重大助益一樣。本論文以智慧型手機為平台，建立起資源受限且具備抗雜訊強健式的語音辨認技術，重點在找出符合智慧型手機運算環境的強健式技術，雖然演算法的實現平台為智慧型手機，但核心技術可以延伸並應用到其他的移動式智慧型裝置，例如 PDA、 GPS、車機等等，其技術應用領域具備高度實用性。

參考文獻

[1] R. C. Rose et al., "***On the implementation of ASR algorithms for hand-held wireless mobile devices***", ICASSP 2001, 17-20, 2001.

[2] Dmitry Zaykovskiy, "***Survey of the Speech Recognition Techniques for Mobile Devices***",11-th International Conference Speech and Computer (SPECOM-2006).

[3] Y. M. Lam et al., "***Fixed-point***

implementations of speech recognition systems", Proceedings of the International Signal Processing Conference, DALLAS, 2003.

[4] Y. H. Kao et al., "***A low cost dynamic vocabulary speech recognizer on a GPP-DSP system***", ICASSP 2000, Vol. 6, 3215-3218, 2000.

[5] T. W. Kohler et al., "***Rapid Porting of ASR-Systems to Mobile Devices***", INTERSPEECH-2005, 2005.

[6] David Huggins-Daines et al., "***PocketSphinx: A Free, Real-Time Continuous Speech Recognition for Hand-held Devices***", ICASSP 2006, 2006.

[7] Junho Park et al., "***Achieving a reliable compact acoustic model for embedded speech recognition with high confusion frequency model handling***", Speech Communication, Vol. 48, 737-745, 2006.

[8] S. Deligne et al., "***A robust high accuracy speech recognition system for mobile applications***", IEEE Trans. Speech and Audio Processing, Vol. 10, 551-561, Nov. 2002.

[9] W. Herbordt et al., "***Hands-free speech recognition and communication on PDAs using microphone array technology***", ASRU-2005, 302-307, 2005.

[10] Jen-tzungChien and Jain-ray Lai, "***Use of Microphone Array and Model Adaptation for Hands-Free SpeechAcquisition and Recognition***", Journal of VLSI Signal Processing 36, 141–151, 2004.

[11] P.C. Woodland, D. Pye and M.J.F. Gales,"***Iterative Unsupervised Adaptation using Maximum Likelihood Linear Regression***", Fourth International Conference on Spoken Language, Vol. 2, pp. 1133—1136.

[12] Gian Luca Marcialis and Fabio Roli "***Fusion of multiple fingerprint matchers by single-layer perceptron with class-separation loss function***", Pattern Recognition Letters, Volume 26, Issue 12, September 2005, Pages 1830-1839.

[13] H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, "***MAT2000 – Design, Collection, and Validation on a Mandarin 2000-speaker Telephone Speech Database***", ICSLP, pp. 460-463, Beijing, China, 2000.

[14] "***Ambient Noise Database for Telephonometry 1996***", NTT Advanced Technology Corporation.