使用模型適應法解碼可變音框率語音之分散式語音辨識

李立民[a]、洪英愷[b]、簡福榮*[b]、譚旦旭[b]
大葉大學電機工程學系[a]
國立臺北科技大學電機工程學系[b]

摘要 — 在分散式語音辨識架構中，伺服端之語音辨識模型是由全音框率之訓練語料訓練而成。當客戶端採用可變音框率傳送語音特徵串流至伺服端時，將因音框率不同，導致辨識模型與待測語料間之不匹配。本論文提出一語音辨識模型之適應法以補償兩者間之不匹配。實驗結果顯示：我們所提出之模型適應法搭配音框降取樣或降取樣結合最小距離之音框選取演算法在(1/2)、甚至(1/3)音框率下都有非常趨近於全音框率之語音辨識率，並具運算量大幅降低優點。

一、 簡介

在分散式語音辨識(distributed speech recognition, DSR)[1][2]架構中，客戶端(client end)使用者利用手持行動裝置將擷取到的語音特徵串流傳送至伺服端(server end)進行語音辨識，伺服端再將響應結果回傳至客戶端。為減少客戶端傳送語音特徵串流的頻寬，並維持伺服端之語音辨識率，可變音框率[3][4][5]及半音框率((1/2)音框率)[6][7][8]等研究方法由此因應而生。可變音框率選取音框的要旨是：進行語音辨識時，變化大的音框提供更多有關語音感知的訊息所以應該被保留，緩慢改變的音框所提供的訊息很有限所以可以被捨棄。半音框率則是選取奇數索引或偶數索引之音框直接傳送至伺服端，所以可以節省一半的傳輸量。半音框率之音框選取方法可以視為可變音框率選取音框的一個特例。由於伺服端之語音辨識模型是由全音框率(full frame rate, FFR)之訓練語料訓練而成，使用可變音框率或半音框率之測試語料將因為音框率的不同造成訓練模型與測試語料不匹配的狀況發生，導致語音辨識率的下降。為減輕由於音框率不匹配所引起的性能退化，目前主要有兩種補償方法。第一種方法是利用特徵參數間插法(feature interpolation)[9][10][11]將可變音框率或半音框率語音特徵串流在伺服端重建為全音框率語音特徵串流。第二種方法則是模型適應法(model adaptation)，在伺服端適應性地調整語音辨識模型用以匹配所接收到的可變音框率或半音框率語音特徵串流[6][8]。本論文將提出一屬於後者之模型適應法補償方法，在伺服端對可變音框率語音特徵串流進行語音辨識。

本論文將在第二節說明所提出之模型適應法的數學理論，第三節說明欲採用的三種可變音框率音框選取演算法，第四節是實驗結果與討論，最後則是本論文的結論。

二、 HMM模型適應法

令全音框率(FFR)之語音觀測序列為 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_T$ ，對應於全音框率之隱藏式馬可夫模型(hiden Markov model, HMM)其參數集記為 λ ，而由左到右的 HMM 模型設有 N 個狀態，其狀態索引(state index)由 1 到 N。我們也令狀態 0 和狀態 N+1 分別代表在第一個觀測向量 $\mathbf{o}_1$ 之前和最後一個觀測向量 $\mathbf{o}_T$ 之後的兩個特殊邊界狀態。可變音框率(variable frame rate, VFR)的語音觀測序列 $\mathbf{o}_{t_1},\mathbf{o}_{t_2},\cdots,\mathbf{o}_{t_k},\cdots,\mathbf{o}_{t_K},1\le t_k\le T,t_k<t_{k+1}$ 可以視為全音框率觀測序列 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_T$ 的子序列(subsequence)。為計算在此 HMM 模型 λ 下，產生此觀測子序列的機率 $P\left(\mathbf{o}_{t_1},\cdots,\mathbf{o}_{t_K}\,|\lambda\right)$，我們先定義局部觀測子序列的聯合機率密度如下。

$$\alpha_{VFR}\left(q_{t_k},t_k\right)=P\left(\mathbf{o}_{t_1}\cdots\mathbf{o}_{t_k},Q_{t_k}=q_{t_k}\,|\lambda\right), \tag{1}$$

其中 $Q_{t_k}=q_{t_k}$ 意指在 t_k 時，最後狀態是停駐在 q_{t_k} 。

類似於全音框率之推導過程[12]，我們可以疊代的方式計算出截至時間點 t_{k+1} 的局部觀測子序列聯合機率密度函數。

$$\alpha_{VFR}\left(q_{t_{k+1}},t_{k+1}\right)=\sum_{q_{t_k}=1}^{N}\alpha_{VFR}(q_{t_k},t_k)a_{q_{t_k}q_{t_{k+1}}}^{(t_{k+1}-t_k)}b_{q_{t_{k+1}}}(\mathbf{o}_{t_{k+1}}), \tag{2}$$

其中

$$a_{q_{t_k}q_{t_{k+1}}}^{(t_{k+1}-t_k)}=\left(\sum_{q_{t_k+1}=1}^{N}\cdots\sum_{q_{t_{k+1}-1}=1}^{N}a_{q_{t_k}q_{t_k+1}}\cdots a_{q_{t_{k+1}-1}q_{t_{k+1}}}\right) \tag{3}$$

代表在時間點由 t_k 至 t_{k+1} 期間，狀態由 q_{t_k} 轉移至 $q_{t_{k+1}}$ 的狀態轉移機率。請特別注意此轉移機率的上標是當成符號標記並非執行指數運算。(3)式內所用的 $a_{q_{t_k}q_{t_k+1}}$ 則是原先全音框率 HMM 模型 λ 下，下一時間點會由狀態 q_{t_k} 轉移至 $q_{t_{k+1}}$ 的狀態轉移機率。而 $b_{q_{t_{k+1}}}(\mathbf{o}_{t_{k+1}})$ 是指在狀態 $q_{t_{k+1}}$ 下會觀察到特徵向量 $\mathbf{o}_{t_{k+1}}$ 的機率。我們定義(2)式為廣義式向前似然函數(generalized forward likelihood function)，其特例是當 $t_k=k,\ \ k\in\{1,2,\cdots,T\}$ 時，(2)式即收斂至全音框率之一般順向似然函數的定義。

Corresponding author:frjean@ntut.edu.tw
DOI:10.6159/IJSE.2014.(4-2).69

當計算時間點 t_{k+1} 的向前機率時，我們是使用和全音框率 HMM 相同的觀察(輸出)機率分布 $b_{q_{t_{k+1}}}(\mathbf{o}_{t_{k+1}})$，但使用(3)式的狀態轉移機率 $a_{q_{t_k} q_{t_{k+1}}}^{(t_{k+1}-t_k)}$。最後可以整理出在使用此 HMM 模型 λ 下，產生此觀測子序列的機率。

$$P\left(\mathbf{o}_{t_1}\cdots\mathbf{o}_{t_K}\,|\lambda\right)=\sum_{i=1}^{N}\alpha_{VFR}(i,t_K)a_{i,N+1}^{(T+1-t_K)}\,. \qquad (4)$$

實務上為減少運算複雜度，我們使用維特比(Viterbi)演算法[12]解碼可變音框率的觀測序列 $\mathbf{o}_{t_1},\mathbf{o}_{t_2},\cdots,\mathbf{o}_{t_K}$，以得到最大相似路徑及其機率。

三、 可變音框率之音框選取演算法

全音框率(FFR)之語音觀測序列，如前所述令為 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_n,\cdots,\mathbf{o}_T$，且每一個觀測特徵向量是由前 12 個梅爾頻率倒頻譜係數(mel-frequency cepstral coefficients, MFCC)加上對數能量(log energy)所構成，以組成一 13 維 MFCC_E 參數如下。

$$\mathbf{o}_n=\{o_n(1)\quad o_n(2)\quad\cdots\quad o_n(13)\},\quad 1\le n\le T. \qquad (5)$$

我們採用三種演算法進行可變音框率語音之音框選取(或捨棄)，包括有最小距離方法(minimum distance method, MD)、音框降取樣法(frame decimation method, FD)[6][8]、以及降取樣結合最小距離方法(combined decimation and minimum distance method, CDAMD)，分述如下。

3.1 最小距離方法

最小距離方法(MD)是先計算出語音觀測序列中相鄰音框間的距離，選其中距離前一音框為最小的音框逐一捨去，直到保留的音框數目達到設定的保留比例 k %為止。

現說明以最小距離方法選取音框的流程如下：

1. 初始化。
   令全音框率之語音觀測序列為 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_T$，音框索引集合為 $n=\{1,2,\cdots,T\}$，且設定音框保留比例為 k %。
2. 計算相鄰音框間距離。
   對所有屬於 n 的音框索引 t，$t\in n$，除設 $D_1=\infty$ 外，計算其與前一個音框距離。
   $$D_t=\|\mathbf{o}_t-\mathbf{o}_{t-1}\|=\sqrt{\sum_{i=1}^{13}\left(o_t(i)-o_{t-1}(i)\right)^2}\ ,\ t\in n \qquad (6)$$
3. 選取欲捨棄音框。
   $$t^*=\arg\min_t D_t \qquad (7)$$
   捨棄音框 $\mathbf{o}_{t^*}$，並在音框索引集合 n 排除索引 t^*。
   如果 (剩餘音框數 $\le T\times k\%$)，跳至 5，結束;
   否則，跳至 4，繼續。
4. 更新 t^* 前後音框距離。
   令 t_{next}^* 為 t^* 之後，目前還未遭排除的最近音框索引，$t_{previous}^*$ 為 t^* 之前，目前還未遭排除的最近音框索引，更新計算兩音框間距離。
   $$D_{t_{next}^*}=\left\|\mathbf{o}_{t_{next}^*}-\mathbf{o}_{t_{previous}^*}\right\|,\ t_{next}^*\in n\,. \qquad (8)$$
   跳至 3，繼續選取下一個欲捨棄的音框。
5. 結束。

3.2 音框降取樣法

在音框降取樣法(FD)中，音框之選取與捨棄其實跟語音特徵參數無關，而是跟它的索引值相關。讓保留音框間之索引擁有相同的距離，直接降低語音觀測序列的取樣率，使得傳送所需的頻寬減少。其演算法流程說明如下：

1. 初始化。
   令全音框率之語音觀測序列為 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_T$，設降取樣因子(decimation factor) M 為大於一的整數，且令音框索引初始值 $t=1$。
2. 選取欲保留音框。
   如果 t/M 等於整數，則選取音框 $\mathbf{o}_t$ 及音框索引 t；
   否則，捨棄音框 $\mathbf{o}_t$ 及音框索引 t。
   更新音框索引值，$t\leftarrow t+1$。
3. 迴圈判斷。
   如果 $t\le T$，跳至 2，繼續選取音框;
   否則，跳至 4，結束。
4. 結束。

不同的 M 值代表不同的降取樣率，如 $M=2$，所選取之 MFCC_E 串流為(1/2)音框率(half frame rate, HFR)的語音觀測序列 $\mathbf{o}_2,\mathbf{o}_4,\mathbf{o}_6,\cdots,\mathbf{o}_{\lfloor T/2\rfloor}$，其中 $\lfloor x\rfloor$ 代表小於等於 x 的最大整數。若 $M=3$，所選取之 MFCC_E 串流為(1/3)音框率的語音觀測序列 $\mathbf{o}_3,\mathbf{o}_6,\mathbf{o}_9,\cdots,\mathbf{o}_{\lfloor T/3\rfloor}$，依此類推。

3.3 降取樣結合最小距離方法

由於音框降取樣法的方法簡單，而最小距離方法又有緩慢改變的音框會被挑出捨棄的特性，所以結合兩者優點發展為降取樣結合最小距離方法(CDAMD)。其演算法流程說明如下：

1. 初始化。
   令全音框率之語音觀測序列為 $\mathbf{o}_1,\mathbf{o}_2,\cdots,\mathbf{o}_T$，降取樣因子 M 為大於一的整數。
   設定 t_{Last} 為最後保留音框的索引，t_{Ptr} 為下 M 個要比較距離之相鄰音框的音框指標。
   設定初始值 $t_{Last}=M$，$t_{Ptr}=M+1$。
2. 計算相鄰音框間距離。
   計算下 M 個相鄰音框與最後保留音框間距離。
   $$D_t=\left\|\mathbf{o}_t-\mathbf{o}_{t_{Last}}\right\|\ ,\ t_{Ptr}\le t\le t_{Ptr}+M-1 \qquad (9)$$
3. 選取欲保留音框。

$$t^* = \arg\max_t D_t \quad (10)$$

選取音框 $\mathbf{o}_{t^*}$ 及其音框索引 t^* 。

更新 t_{Last} 及 t_{Ptr} ，即 $t_{Last} \leftarrow t^*$ ， $t_{Ptr} \leftarrow t_{Ptr} + M$ 。

4. 迴圈判斷。

如果 $t_{Ptr} \le T$ ，跳至 2，繼續迴圈;

否則，跳至 5，結束。

5. 結束。

四、 實驗結果

4.1 語音資料庫

本論文訓練與測試語料使用中華電信研究所語音資料庫中的獨立數字如表一所示，其錄音格式為 10 kHz，16-bit PCM。錄音人數為 50 位男性及 50 位女性，共錄製 6000 個音檔，其中一半的語料用來訓練模型，另一半語料用來做性能測試。雜訊部分使用 AURORA2 語音資料庫中所附之雜訊，分別有機場(airport)、吵雜人聲(babble)、汽車(car)、展覽會(exhibition)、餐廳(restaurant)、車站(station)、街道(street)以及火車(train)，並將雜訊分別以 20 dB、15 dB、10 dB、5 dB、0 dB 的輸入訊雜比，加入到測試語音。求取 MFCC_E 參數之音框長度及音框位移則各是 25 ms 及 10 ms。

表 I
實驗語料規格

語音資料庫	中華電信研究所之數字語音資料庫
錄音者	50 位男性、50 位女性
音訊檔	6000 wav 檔
取樣頻率	10 kHz
訓練用語料	3000 個獨立數字
測試用語料	3000 個獨立數字
音框長度	25 ms
音框位移	10 ms
輸入語料訊雜比	20 dB、15 dB、10 dB、5 dB、0 dB
雜訊語音資料庫	AURORA2
訊雜類型	機場、吵雜人聲、汽車、展覽會、餐廳、車站、街道、火車

當伺服端(server end)接收到由客戶端(client end)傳送至的 MFCC_E 參數串流將進行語音辨識。但如果有使用可變音框率音框選取將因音框率的不同而造成伺服端全音框率 HMM 模型與可變音框率待測語料不匹配的狀況發生，造成辨識率的下降，因此需透過語音匹配之補償方法以隱藏所遺失音框。在本論文中我們將使用所提出的模型適應法(model adaptation method, MA)並和常被使用的特徵參數間插法(feature interpolation method, FE)做比較。在語音進行辨識之前，我們先執行將 13 維 MFCC_E 參數串流展延成 13 維速度(velocity)成份和 13 維加速度(acceleration)成份，共 39 維 MFCC_EDA 參數串流之前處理。要特別注意的是：對於可變音框率 MFCC_E 參數串流不可以直接做速度和加速度展延運算，而是必須先間插為全音框率 MFCC_E 參數串流，再做速度和加速度展延運算。若使用模型適應法，再將間插入的音框捨棄。

4.2 實驗結果

表 II 為客戶端採用最小距離方法(MD)、音框降取樣法(FD)、或降取樣結合最小距離方法(CDAMD)選取(1/2)音框率 MFCC_E 參數串流，伺服端採用模型適應(MA)或特徵參數間插(FE)補償方法，或不採用任何補償方法之語音辨識率。其中就平均辨識率而言，FD 配合 MA 的辨識率 88.09 %以及 CDAMD 配合 MA 的辨識率 88.45 %皆比全音框率的語音辨識率 88.05 %略高。意味 FD 或 CDAMD 只用 50 %的音框，搭配 MA 補償方法，其辨識率與全音框率語音的辨識率幾乎相同。不採用任何補償方法之語音辨識率明顯較低。

表 II
選取(1/2)音框率語音之語音辨識率

音框選取法	FFR	MD method			FD method			CDAMD method		
補償方法	none	none	MA	FE	none	MA	FE	none	MA	FE
20 dB	98.10	92.00	97.50	97.37	94.83	98.00	97.77	96.13	98.20	97.97
15 dB	96.87	90.63	96.30	95.93	93.40	96.83	96.53	95.20	97.03	96.53
10 dB	94.33	86.33	93.17	92.97	90.73	94.37	93.70	92.17	94.67	93.83
5 dB	86.97	74.43	83.83	82.57	81.13	86.80	85.50	82.43	86.77	85.63
0 dB	64.00	48.77	60.13	60.10	57.23	64.43	63.53	57.30	65.57	62.90
AVERGE	**88.05**	78.43	86.19	85.79	83.46	**88.09**	87.41	84.65	**88.45**	87.37

表 III 為選取(1/3)音框率之語音辨識率。其中就平均辨識率而言，FD 配合 MA 的辨識率 88.14 %也比全音框率的語音辨識率 88.05 %略高，為(1/3)音框率下的最高語音辨識率。CDAMD 配合 MA 可達到辨識率 87.82 %。不採用任何補償方法之語音辨識率則急遽降低。

表 III
選取(1/3)音框率語音之語音辨識率

音框選取	FFR	MD method			FD method			CDAMD method		
補償方法	none	none	MA	FE	none	MA	FE	none	MA	FE
20 dB	98.1	56.67	96.53	95.13	64.5	97.73	97.67	70.37	97.73	97.57
15 dB	96.87	56.03	94.87	93.03	62.67	96.6	96.47	68.57	96.23	96.27
10 dB	94.33	52.43	90.3	87.67	59.43	94.2	93.4	65	93.9	92.97
5 dB	86.97	41.5	78.23	74.37	50.7	86.23	84.57	54.6	85.8	83.63
0 dB	64	24.97	52.17	50.63	31.4	65.93	62.57	35.07	65.43	62.43
AVERGE	**88.05**	46.32	82.42	80.17	53.74	**88.14**	86.94	58.72	**87.82**	86.57

表 IV 為選取(1/4)音框率之語音辨識率。其中就平均辨識率而言，FD 配合 MA 的辨識率 85.98 %比全音框率的語音辨識率 88.05 %略低，為(1/4)音框率下的最高語音辨識率。表 V 為選取(1/5)音框率之語音辨識率。其中 CDAMD 配合 MA 的平均辨識率 83.57 %為(1/5)音框率下的最高語音辨識率。

從表 II 至表 V 我們可以得知音框率降低至(1/3)，其最佳語音辨識率還是與全音框率的語音辨識率很接近，當音框率降低至(1/4)以下則開始有明顯差異。另外也可以得知採用模型適應(MA)補償方法的語音辨識率幾乎皆優於採用特徵參數間插(FE)補償方法的語音辨識率。

表 IV
選取(1/4)音框率語音之語音辨識率

音框選取	FFR	MD method			FD method			CDAMD method		
補償方法	none	none	MA	FE	none	MA	FE	none	MA	FE
20 dB	98.10	23.70	94.93	90.03	27.13	96.93	96.50	31.20	96.97	96.53
15 dB	96.87	23.33	91.47	87.27	26.50	95.80	94.90	30.30	95.43	94.57
10 dB	94.33	22.07	85.50	78.37	24.73	91.70	91.27	28.20	91.70	90.80
5 dB	86.97	17.60	70.87	63.07	20.97	83.23	82.03	20.97	83.23	82.03
0 dB	64.00	12.50	46.20	39.67	14.13	62.23	60.33	14.13	62.23	60.33
AVERGE	**88.05**	19.84	77.79	71.68	22.69	**85.98**	85.01	24.96	**85.91**	84.85

表 V
選取(1/5)音框率語音之語音辨識率

音框選取	FFR	MD method			FD method			CDAMD method		
補償方法	none	none	MA	FE	none	MA	FE	none	MA	FE
20 dB	98.10	12.20	91.77	83.80	12.17	95.83	95.93	13.20	95.83	95.63
15 dB	96.87	12.13	88.20	80.60	12.07	93.83	93.83	13.07	93.93	93.97
10 dB	94.33	11.80	80.70	70.63	11.57	89.80	89.07	12.67	89.87	88.37
5 dB	86.97	11.10	64.57	54.33	10.93	78.93	77.07	11.63	80.10	76.13
0 dB	64.00	10.33	41.13	32.87	10.23	55.70	52.27	10.20	58.10	53.00
AVERGE	**88.05**	11.51	73.27	64.45	11.39	**82.82**	81.63	12.15	**83.57**	81.42

表 VI
於不同音框率下，使用不同補償方法的辨識時間
(將全音框率(FFR)的辨識時間正規化到 100 %)

補償方法	FFR	none	MA	FE
(1/2)音框率	100.00 %	49.71 %	**52.76 %**	99.33 %
(1/3)音框率	100.00 %	33.23 %	**34.97 %**	99.39 %
(1/4)音框率	100.00 %	24.97 %	**26.40 %**	99.56 %
(1/5)音框率	100.00 %	20.16 %	**21.02 %**	99.12 %

表 VI 是伺服端語音辨識器於不同音框率下，使用不同補償方法的辨識時間。表中我們將全音框率(FFR)的辨識時間正規化到 100 %。FE 是將音框間插重建為全音框率，因此辨識時間於不同音框率下大致都會和全音框率相同。MA 由於音框數的減少，從表中可以觀察到辨識時間是隨著音框數刪減多寡呈現出近似於等比例減少。和全音框率相比，使用 MA 補償方法在(1/2)、(1/3)、(1/4)、及(1/5)音框率下，其語音辨識時間(運算量)只有全音框率的 52.76 %、34.97 %、26.40 %、及 21.02 %。

結論

本論文提出一在伺服端可以直接對可變音框率語音特徵串流進行語音辨識之 HMM 模型適應法補償方法。客戶端則使用三種音框選取演算法選取音框，分別是最小距離方法(MD)、音框降取樣法(FD)、及降取樣結合最小距離方法(CDAMD)。除全音框率外，實驗所選取的可變音框率分別是(1/2)、(1/3)、(1/4)、(1/5)音框率。實驗結果顯示本論文提出的模型適應法配合 FD 或 CDAMD 在(1/2)，甚至(1/3)音框率下有幾乎等同於全音框率的語音辨識率，也就是說在刪除 66.67 %音框後，其語音辨識率也非常趨近全音框率的辨識率。本論文所提出模型適應法的另一大優點便是進行語音辨識時，所需運算量的大幅降低。以客戶端傳送(1/3)音框率語音特徵串流為例，採用本論文模型適應法的伺服端可以幾乎增加兩倍用戶容量而不需採購裝置新設備。

參考文獻

[1] Z.-H. Tan and I. Varga, "Network, distributed and embedded speech recognition: an overview," in *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, Z.-H. Tan and B. Lindberg Ed. Springer-Verlag, 2008, pp. 1-26.

[2] Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, *ETSI Std. ES 201 108 V1.1.1*, Feb. 2000.

[3] K. M. Pointing and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech and Language*, vol. 5, no. 2, Apr. 1991, pp. 169-179.

[4] P. Le Cerf and D. Van Compernolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Process. Lett.*, vol. 1, no. 12, Dec. 1994, pp. 185-187.

[5] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 4, no. 5, 2010, pp. 798-807.

[6] L.-M. Lee, "Adaptation of hidden Markov models for half frame rate observations," *IET Electronics. Lett.*, vol. 46, no. 10, 2010, pp. 723-724.

[7] Z.-H. Tan, P. Dalsgaard, and B. Lindberg, "Exploiting temporal correlation of speech for error-robust and bandwidth-flexible distributed speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, 2007, pp. 1391-1403.

[8] L.-M. Lee and F.-R. Jean, "Adaptation of hidden Markov models for recognizing speech of reduced frame rate," *IEEE Transactions on Cybernetics*, to be published.

[9] H. Deng, D. O'Shaughnessy, J. Dahan, and W.F. Ganong, "Interpolative variable frame rate transmission of speech features for distributed speech recognition," *IEEE Workshop on Automatic Speech Recognition & Understanding*, Kyoto, Japan, 2007, pp. 591-595.

[10] Z. Tan, P. Dalsgaard, and B. Lindberg, "Exploiting temporal correlation of speech for error robust and bandwidth flexible distributed speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, pp. 1391–1403, 2007.

[11] W. Kim, and J. H. L. Hansen, "Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, pp. 2111-2120, 2010.

[12] L.R. Rabiner and B. Juang, *Fundamental of Speech Recognition*, Prentice Hall, 1993.