The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 246-255

使用語者轉換技術於語音合成資料庫之音質改進

Speech Enhancement for TTS Speech Corpora by using Voice Conversion Technologies

林衍廷　Yan-ting Lin
國立臺北大學通訊工程學系
Department of Communication Engineering
National Taipei University
s8303232000@gmail.com

江振宇　Chen-yu Chiang
國立臺北大學通訊工程學系
Department of Communication Engineering
National Taipei University
cychiang@mail.ntpu.edu.tw

摘要

本論文將語者轉換技術用於修復語音資料庫的音質，其原由是在建立文字轉語音系統時，所使用的語音資料庫之部分受限於錄音器材與錄音環境而音質不佳，為了讓這些音質不佳的語料能夠被重新不浪費地使用於建立文字轉語音系統，本論文將利用語者轉換技術於語料庫的音質修復，利用同一語者的特性讓語者轉換的問題轉變成音質轉換的研究問題。在轉換技術上，聲學參數用 WORLD 聲碼器來分析語音訊號，轉換模型用傳統高斯混和模型以及深度學習模型，也嘗試了多種輸入及輸出參數組合，最後也探討以不同語速的語料進行音質修復的結果。客觀以及主觀測試結果顯示，轉換(修復)過的音質有明顯提升。另外也嘗試轉換同一位語者錄製的中英夾雜語料，該語料有一樣的問題，即使是跨語言轉換，實驗結果顯示有降低部分回音和雜訊。

一、緒論

（一）研究動機與方向

在語音研究上，好的語音資料庫對於語音相關的研究是很重要的。在建立文字轉語音系統時需要聲學模型來合成聲音，而在訓練聲學模型時，發現訓練用的語料庫有音質上的問題，使得訓練出來的聲學模型不好，進而影響輸出的音質不好。如果能修復這個

語料，就能減少語料產生的問題並繼續使用這些語料。本研究所使用的語音資料庫是由一位專業女性播音員(縮寫成語者 Tao)讀稿錄製之 4 種語速平行語料庫，總計 1,478 個音檔，語速分為：快、正常、中、以及慢，而有雜訊的語料為語速正常，本研究試著用語者轉換的方法，利用同一語者的特性使得語者轉換變成音質轉換，將語速正常的音質轉成其他語速的音質。現有的轉換技術從高斯混合模型(Gaussion mixture model,GMM)到深度類神經網路(deep neuron network,DNN)等等，能多方嘗試找出最好的轉換模型。另外，本研究也嘗試轉換另一個語者 Tao 所錄製的中英夾雜的語料，該語料有一樣的音質問題，嘗試能否也轉換非同一語言的情形。

（二）語者轉換相關研究文獻探討

語者轉換可以分為平行語料的轉換和非平行語料的轉換兩大區塊討論，平行語料的轉換最廣為人知的就是用 Gaussion mixture model(GMM)[1]來做轉換，將來源音框和目標音框用動態時間扭曲(dynamic time warping,DTW)做對齊後，來訓練高斯混合模型描述不同的發音，還有[2]GMM 加上 maximum likelihood parameter generation(MLPG)的作法，用機率來解這一問題到這邊算是有一個不錯的結果。

再來就是近年流行的機器學習，從[3]開始了用類神經網路(artificial neural networks, ANN)來轉換來源語者到目標語者，在[4]中實驗了很多 DNN 的變種，還有可以用一段訊號作為輸入的 DBLSTM [5]。除了用上述的方式做轉換，還有用 spectral differential[6]的作法，去學習來源語者和目標語者的差異；後來出現了用語音辨識加語音合成的作法[7]，先用語音辨識辨別來源語者的內容，再訓練目標語者的聲學模型，用聲學模型將內容合成聲音，這種作法讓非平行語料的語者轉換有更進一步的突破。

（三）語者轉換架構

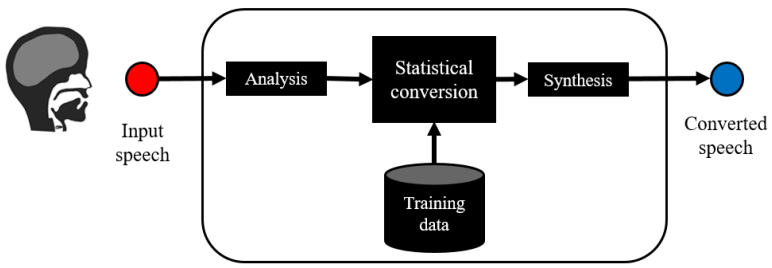

圖 一：語者轉換架構圖

語者轉換作法流程如圖一，輸入訊號經由分析成聲學參數，在訓練資料的部分，訓練的模型都是用來轉換聲學參數，統計式轉換(statistical conversion)所用的模型有 GMM 的機率模型也有神經網路(neuron network, NN)的機器學習模型，都是在將輸入的參數轉換成接近目標的參數，將聲學參數轉換完之後，進入合成器合成出聲音，這就是語者轉換的運作流程。

二、語音資料庫

(i). Treebank-Tao-SR-Corpus：是由一位專業女性播音員(縮寫成語者 Tao)讀稿錄製之 4 種語速平行語料庫，總計 1,478 個音檔，共有 203,746 個音節，語速分為：快、正常、中、以及慢，平均音節長度分別為 0.181 秒、0.198 秒、0.244 秒及 0.264 秒，音檔均為 20kHz 的取樣率及 16-bit 之 PCM 格式，主要內容大多摘錄自新聞、網路文章。

(ii). English-Tao-CEMix-Spell-Corpus：是由語者 Tao 所錄製而成的中英夾雜語料，以中文為主體並穿插英文字母於中文語句中，共 539 個語句，總音節數為 13,540 個音節，包含 11,688 個中文音節與 1,872 個英文字母。音檔為取樣頻率 20,000 赫茲(Hertz)及 16 位元數之 PCM 格式，平均語速為一秒 3.5 個音節。

(iii). English-Tao-CEMix-Word-Corpus：是由語者 Tao 所錄製而成的中英夾雜語料，以中文為主體並穿插英文詞(word)於中文語句中，共 843 個語句，總音節數為 18,103 個音節，包含 15,885 個中文音節與 2,218 個英文音節。音檔為取樣頻率 20,000 赫茲(Hertz)及 16 位元數之 PCM 格式，平均語速為一秒 4.85 個音節。

由於本實驗所需平行的語料，將(i)語料庫中的 4 個語速經文本比對之後，我們用其中 1048 個音檔作為實驗的語料且取樣率為 16kHz。

三、語者轉換技術簡介

（一）聲碼器

聲碼器就是將聲音訊號分解成有意義的參數，像 Mel-cepstral coefficient(MCC)就是將聲音訊號的頻譜取對數後按照人耳對頻率的聽覺敏銳做縮放再做傅立葉逆轉換得到的係數。這個依據人耳特性對頻率軸縮放的就是梅爾刻度(Mel-scale)，每一個刻度都是一維的 MCC。Mel-log spectral approximation filter (MLSA filter)就是將 MCC 轉回頻譜包絡(spectral envelope)，就可以把頻譜包絡和激發訊號卷積來得到聲音訊號。另外，

WORLD [8]聲碼器是近年被認為目前音質最好的聲碼器，它將訊號分成音調(pitch)、頻譜包絡和非週期性(aperiodicity)三個部分，WORLD 將非週期性從頻譜包絡分出來之後，做頻譜包絡的轉換會更好，整體音質會較穩定。

（二）映射函數

1、Gaussian Mixture Model (GMM)

高斯混合模型 GMM，可以很好的近似任意的機率分布，在這邊用來描述聲學參數的機率分布，用於語者轉換。令來源的 MCC 為x_t、目標為y_t，t為音框數，為了考慮音框之間的關聯性，需要一個有前後音框資訊的 delta $= \Delta data = -0.5 * data_{t-1} + 0.5 data_{t+1}$，令$X_t = [x_t \ \ \Delta x_t]$和$Y_t = [y_t \ \ \Delta y_t]$，而$Z_t = [X_t \ \ Y_t]$，假設$Z_t$可以由高斯混合模型表示成$P(Z_t|\lambda^{(Z)})$，$\lambda^{(Z)}$為機率參數。接下來將$P(Z_t|\lambda^{(Z)})$用貝式定理展開，$P(Y_t|X_t, \lambda^{(Z)})$就是將來源轉到目標的映射函數。

$$P(Z_t|\lambda^{(Z)}) = P(X_t, Y_t|\lambda^{(Z)}) = P(Y_t|X_t, \lambda^{(Z)})P(X_t|\lambda^{(Z)}) \tag{3.1}$$

在估計目標$\hat{y}$時，我們會希望 likelihood 最大，也就是$P(Y_t|X_t, \lambda^{(Z)})$最大，寫成

$$\hat{y} = argmax\ logP(Y|X, \lambda^{(Z)}) \tag{3.2}$$

目標函數$Q(Y, \hat{Y})$可寫成式 3.11，為了使目標函數最大，將$Q(Y, \hat{Y})$對y微分算出$\hat{y}$整個過程稱為 maximum likelihood parameter generation(MLPG)。

$$Q(Y, \hat{Y}) = \sum_{all\ m} P(m|X, Y, \lambda^{(Z)}) logP(\hat{Y}, m|X, \lambda^{(Z)}) \tag{3.3}$$

2、Deep Neural Networks (DNN)

類神經網路(Artificial Neural Networks, ANN 亦可以說是 NN)是由多個像神經元的感測器所組成，每個感測器的輸出是由輸入向量與權重向量的內積後，經過一個非線性函數所得的純量，其架構是一層輸入層、一層隱藏層和一層輸出層，而 DNN 就是多層的隱藏層。本研究用的非線性的函數是 sigmoid，訓練整個 DNN 的準則用的是 minimum mean squared error (MMSE)。在語者轉換中，DNN 是音框對音框的轉換，跟 GMM 一樣需要有 delta 作為輔助輸入來考慮前後音框的影響，轉換結果才會比較好。

3、Deep Bidirectional Long Short-Term Memory (DBLSTM)

上述兩種轉換都需要 delta 的資訊來輔助訓練，所以也出現用 Recurrent Neural Networks (RNN)可以考慮整段資料的模型。為了避免 RNN 在資料過長的時候有梯度爆

炸或梯度遺失的問題，而改用 Long Short-Term Memory (LSTM)。本研究使用的是 Deep Bidirectional LSTM (DBLSTM)，DBLSTM 是將多層 BLSTM 疊在一起來加強學習效果。BLSTM 隱藏層分為時序向前和時序向後兩種，這樣的架構可以使下一時刻的輸出考慮前後文的關係，就不需要 delta 作為輔助輸入，$\vec{h}_t$為時序向前，迭代計算$\mathrm{t} = 1$ 到 $\mathrm{t} = \mathrm{T}$；$\overleftarrow{h}_t$為時序向後，迭代計算$\mathrm{t} = \mathrm{T}$ 到 $\mathrm{t} = 1$，$H(*)$表示整個 LSTM 單元過程，BLSTM 單層運算如式(3.4~3.6)，整個 DBLSTM 可以用 back propagation though time(BPTT)算出權重。

$$\vec{h}_t = H(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}) \tag{3.4}$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{3.5}$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{3.6}$$

4、Spectral Differential

Spectral differential[6]目的是學x和y的差異，然後作為 MLSA filter 的輸入。令$d_t = y_t - x_t$。在 GMM 中，算出Δd_t，令$D_t = [d_t \;\; \Delta d_t]$，將目標函數寫成下列：

$$\hat{d} = argmax\, P(D|X, \lambda^{(Z)}) \tag{3.7}$$

這個$\hat{d}$就是轉換的x和y距離，最終要合成的 MCC 參數為$y' = x + d'$。在[6]提出的做法中，激發訊號是原始的音訊，然後 MLSA filter 的參數是$\hat{d}$，像是將原本的訊號通過一個 filter，所以音質不會卡在聲碼器的音質上限。

（三）Sprocket toolkit

Sprocket 是一個以 python 為基礎的語者轉換的工具，用的轉換方法是 GMM，而 Sprocket 中的 spectral differential 的作法和[6]中的作法有些出入，Sprocket 並未另外訓練出x對d的轉換模型，而是用原x對y的模型轉出y'後，令$d' = y' - x$來運算，本論文將以 sprocket 的作法來實驗 spectral differential 並和直接轉換的實驗做比較。

四、實驗結果和探討

（一）實驗設計

本節實驗主要為語速正常轉語速中，轉換的方法分為 GMM、DNN 和 DBLSTM，在 GMM 中轉換的目標分為 spectral differential(Diff)和目標語者，另外還有非週期性的轉換。

1、映射函數之比較

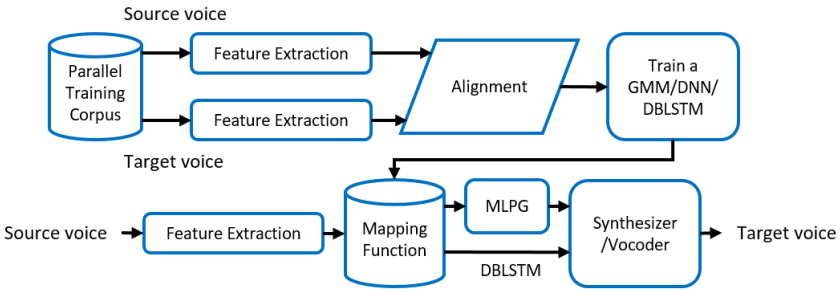

圖 二：GMM/DNN/ DBLSTM 轉換之流程圖

圖二是 GMM/DNN/DBLSTM 語者轉換的流程圖，MCC 的維度都是 40 維，GMM 和 DNN 都需要前後音框的資訊 delta 作為輔助，delta 也是 40 維，所以輸入是 80 維，輸入資料需要做正規化，轉換後的結果經過 MLPG 得到目標的 MCC。DNN 的架構是隱藏層 2 層，每層 2048 個節點。DBLSTM 用雙向訓練來考慮前後音框的影響，一筆資料需要一段的音框最為輸入，這裡用 512 個音框作為一筆資料的長度，整句音檔末端不足 512 的部分會往前湊齊 512 為該音檔的最後一筆資料，而且每個音框的資訊都需要正規化，其架構是隱藏層 2 層，每層 512 個節點，256 個節點向前、256 個節點向後。

2、增加輸入參數之比較

在映射函數比較的實驗中，DBLSTM 來轉換得到較好的結果，但是碰到了在發音邊界會有雜音的問題，為了改善這一問題在模型訓練的輸入端加入了語言參數(language parameter, 以下簡稱 lp)。這語言參數是 64 維 one-hot 的發音標籤(代表音節的聲母和韻母資訊)和 2 維的該音框在該發音中的正規化位置和整段發音的長度，這 66 維的資訊在輸入端時所對應的的目標之語言參數也要作為輸入，因此輸入維度是 172 維，輸出是 40 維的 MCC。另外實驗了加入了兩維的 voice/unvoice(以下簡稱 uv)的資訊，如果第一維是 voice 為 1，是 unvoice 為 0；第二維是 lf0，unvoice 的 lf0 是用前一個 voice 和後一個 voice 的端點做內插取得，也是需要加入對應的目標 uv 資訊，輸入維度是 176 維。

3、加入非週期性轉換之比較

前面都是針對頻譜包絡的轉換，這個實驗是將另一個參數非週期性(aperiodicity, 以下簡稱 ap)也轉換。在來源 MCC 和目標 MCC 動態時間扭曲之後，將其對齊好的索引值拿來對齊 ap(用 MCC 表示)。在轉換 ap 的模型我們也設計兩種，一種是 ap 轉成 ap，

輸入 40 維，輸出 40 維；另一種是加入語言參數的輸入 172 維，輸出 40 維，兩種的訓練除了輸入維度不同，其架構都同 DBLSTM 的實驗。

4、Spectral Differential 與 GMM 轉換之比較

Diff 的做法和 GMM 語者轉換在資料對齊之前都一致，對齊資訊還是依據 MCC 對齊後的索引值，再計算 MCC differential 的部分$d = y - x$，訓練出一個x轉到d的轉換模型，將轉出d'的作為 MLSA filter 的參數，輸入為來源音檔，得出聲音。但 sprocket 中，並未訓練一個x轉到d的轉換模型，而是用x轉到 y 的轉換模型，轉出y'之後，令$d' = y' - x$來得到差異的資訊，本實驗以 sprocket 的作法實踐，並和原 GMM 轉換的聲音做比較。

（二）實驗結果[1]

我們對外部音檔共 25 句來轉到語速中，每一個實驗有客觀評量或主觀評量。客觀評量中，我們並未有正確且長度一致的音檔可以來算 stoi 和 pesq，所以我們提出用 source、target 和 transformed 的 mel-cepstral distortion (MCD)來做相互比較，target 在算 MCD 前須和 source 做 DTW 對齊成 source 的長度。MCD 是描述兩個 MCC 距離的方法， y和x都是 MCC，N 是 MCC 的維度如公式 4.1

$$MCD = \frac{10}{ln\,10}\sqrt{2\sum_{n=1}^{N}(y[n] - x[n])^2} \tag{4.1}$$

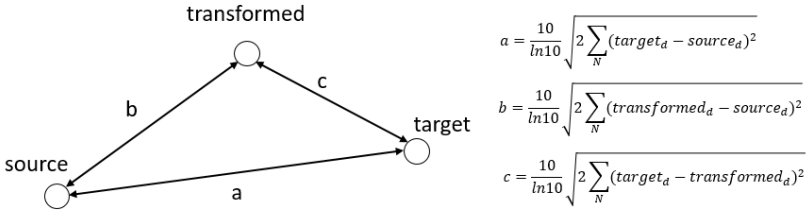

圖 三：MCD 比較示意圖

如圖三將 source、target 和 transformed 的 MCD 算出，只要能符合$a > b > c$，就能保證這個轉換的結果比原來的好，如果只比較$a > c$，那麼以目標為圓心，半徑為 c 的圓都會符合此條件而無法確定圓上的點之間的優劣；另外沒有比較$b > c$的話，也不能確定轉換後的結果是否有比較靠近目標。在做其他變因比較的時候，在符合$a > b > c$情形下比較c也可以看出哪個轉換的效果較優；在主觀測試，我們找了 6 位受測者，每一個實驗從 25 句裡隨機抽 5 句聽測，讓測試者投票哪個音質比較好，再比較票數算比例。

[1] https://drive.google.com/drive/folders/1YsCJNmhw6RFWzfI9DMyiX8ciNBTaAjwu?usp=sharing

1、映射函數之比較結果

表一是 GMM、DNN 和 DBLSTM 的 MCD 比較和票數比較。在客觀評量中，GMM 並未符合$a > b > c$，且b和c還比a長，在聽感上有點像語速中但又有點不像的樣子。DNN 轉換的結果符合$a > b > c$，在距離 c 的差異上 DNN 的表現比較好；在聽覺測試的投票中，73.3%認為 DNN 的音質比 GMM 好。DBLSTM 的轉換結果也符合$a > b > c$。在 DNN 跟 DBLSTM 的比較中，距離 c 的值是 DNN 略小，但在聽覺測試的投票中，僅有 30%認為 DNN 音質較好，70%認為 DBLSTM 音質比 DNN 好，在轉換模型的實驗比較中，DBLSTM 是我們覺得表現較好的。

表 一：映射函數之 MCD 比較和聽覺測試

	a	b	c	聽覺測試	
GMM	7.942	12.815	13.061	26.6%	
DNN	7.942	6.644	5.298	73.3%	30%
DBLSTM	7.942	7.433	5.433		70%

2、增加輸入參數之比較結果

在確認 DBLSTM 是較好的架構之後，我們開始增加輸入參數的實驗，為了使得訓練轉換的結果更接近目標，加入了語言參數- DBLSTM 172 和額外再加入 uv 資訊的- DBLSTM 176，由表二可以看出距離 c 的值有減少了一些；聽覺測試中是 46.6%比 53.3%，聽感上有點接近，必須用耳機聽才能聽出發音邊界雜音的差異。另外增加 voice/unvoice 參數的實驗，由於距離 b 和距離 c 跟 DBLSTM 172 是一樣的所以不做聽力測試。

表 二：增加輸入參數實驗之比較 MCD 和聽覺測試

	a	b	c	聽覺測試
DBLSTM	7.942	7.433	5.433	46.6%
DBLSTM 172	7.942	6.818	4.732	53.3%
DBLSTM 176	7.942	6.818	4.732	

3、加入非週期性轉換之比較結果

在加入 ap 轉換的實驗中，MCC 的轉換是用 DBLSTM 172，所以 MCD 的比較會一致，我們用 ap 加語言參數的實驗來做聽覺測試，聽覺測試的結果 53.3%的票數覺得有轉 ap 的結果聽起來比較好一點，聽感上有轉 ap 的結果比較沒有嗡嗡的聲音。

4、Spectral Differential 與 GMM 轉換之比較結果

因為 GMM Diff 沒有經過聲碼器，所以不做 MCD 比較，只做聽覺測試。聽覺測試的結果 86.6%的票數覺得 Diff 的聲音比較好聽，在沒有經過聲碼器的情況下音質聽起來就像是通過濾波器。

（三）轉到不同語速之音質比較

我們用 DBLSTM172 來實驗語速正常轉到哪種語速的音質比較好，在 MCD 的比較中，由於各語速和語速正常距離並不一致且分布不同，所以只能比較語速正常轉到哪種語速的結果比較像該語速，由表三可以看出語速慢距離語速正常是最遠的，語速快和語速中距離語速正常的長度差不多且語速快的距離 c 比語速中的略小，可以得知語速快和語速中各自的轉換效果差不多好；而在聽覺測試中 42.8%的語速中對上 50%語速快，在聽感上可能語速快會好一些，結論是轉到語速中或是語速快都是不錯的。

表 三：轉到不同語速實驗之 MCD 比較

	a	b	c	聽覺測試
語速中	7.942	6.818	4.732	42.8%
語速快	7.976	6.530	4.710	50%
語速慢	8.632	6.706	5.085	7.1%

（四）中文轉換模型對中英夾雜語料之嘗試

在確認中文對中文的轉換上音質確實有變好之後，我們對 CE_word 和 CE_spell 分別做語者轉換，轉換模型是 DBLSTM，輸入是 40 維。由於是轉成中英夾雜的句子，沒有內容一致且音質較好的音檔可供計算 MCD，只能做聽覺測試。CE_spell 投票結果顯示轉得並不好，聽感上回音的部分少了很多但有點破嗓，整體感覺還是沒有原本好；CE_word 投票結果各 50%，有改善到某些低頻雜音。雖然整體實驗效果沒有特別顯著，但我們發現即使原本中文轉換模型沒看過英文的發音，在轉換的過程還是會轉成類似的發音，使得音檔中英文的部分還是能的聽出來。

五、結論

本論文提出用語者轉換的技術用於修復語音資料庫，藉由同一位語者的特性使得語者轉換變得像是音質轉換，經過研究和分析得出以下結論：(1)在轉換模型的架構上，

DBLSTM 的表現比起 GMM 和 DNN 都來的好；(2)藉由文本的發音標記和發音位置做為輸入來輔助訓練模型，確實能讓結果更近似於目標聲音；(3) 非週期性的資料對齊應跟隨頻譜包絡，且轉換非週期性對於音質有一定的幫助；(4)以本論文所使用的 4 種語速語料庫而言，將語速正常轉到語速快或是語速中是較為恰當的，音質也比較好；(5)跨語言的轉換中，發音相近的會共用相同的轉換對。本研究嘗試用語者轉換技術在同一語者的語料庫進行音質修復，由主觀測試結果可得知，在 DBLSTM、DBLSTM 172 和加入 ap 轉換的實驗中皆有改善音質和減少雜訊。

參考文獻

[1] Y. Stylianou, O. Capp´e, and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion,” in IEEE Trans. on Audio, Speech and Language Processing, 1998

[2] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” in IEEE Trans. on Audio, Speech and Language Processing, 2007

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, “Spectral Mapping Using Artificial Neural Networks for Voice Conversion,” in IEEE Trans. on Audio, Speech and Language Processing, 2010

[4] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, “Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training,” in IEEE Trans. on Audio, Speech and Language Processing, 2014

[5] L. Sun, S. Kang, K. Li and H. Meng, “Voice Conversion Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks,” in ICASSP, Apr. 2015.

[6] K. Kobayashi, T. Toda, S. Nakamura, “F0 Transformation Techniques for Statistical Voice Conversion with Directwaveform Modification with Spectral Differential,” in IEEE Spoken Language Technology Workshop, Dec. 2016.

[7] L. Sun, K. Li, H. Wang, S. Kang and H. Meng, “Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training,” in ICME, July. 2016.

[8] M. MORISE, F. YOKOMORI, K. OZAWA, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” in IEICE, 2016