數位語音處理概論

期末專題

教授:李琳山

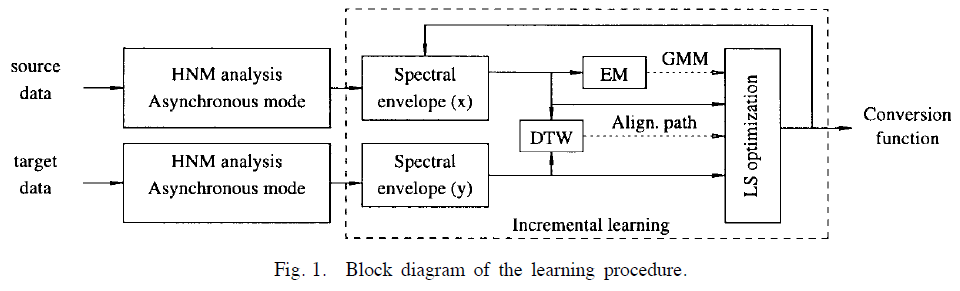
學生: 莊成毅、傅冠鈞

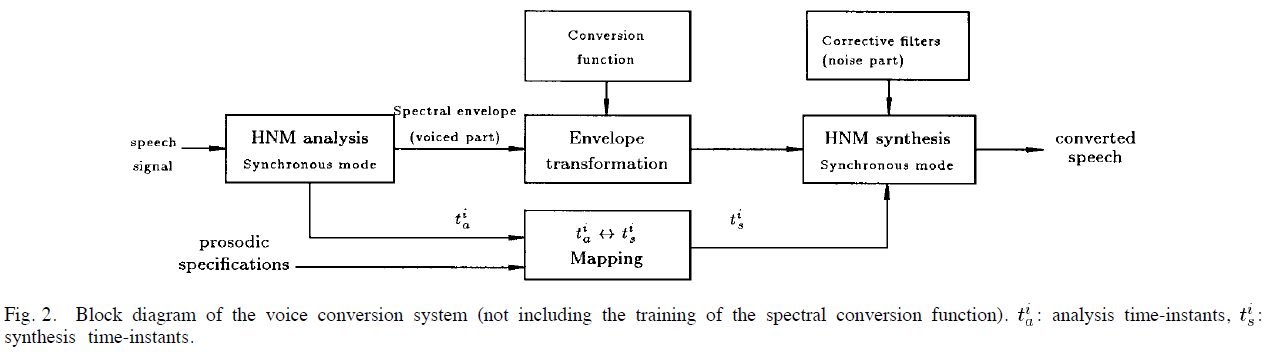
學號：b99901008、r03920184

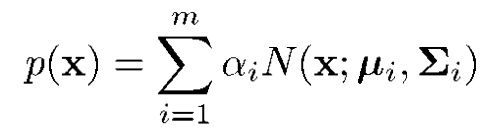
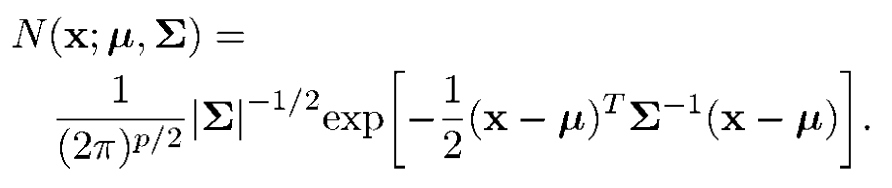
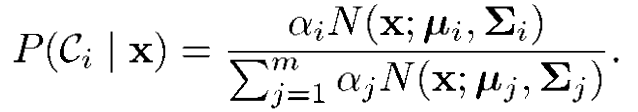
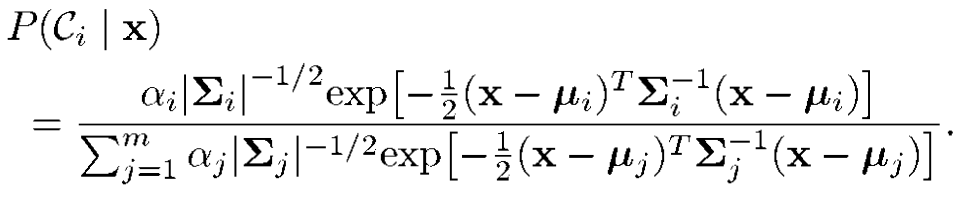
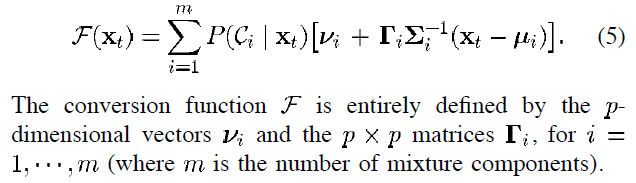
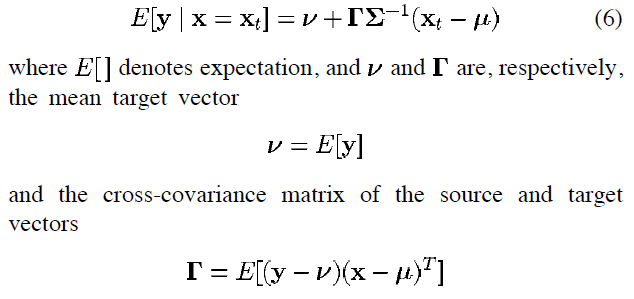
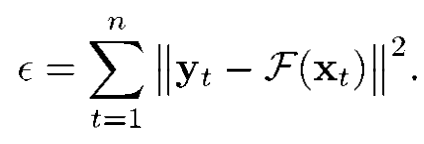
1. 領結型變聲器  
    

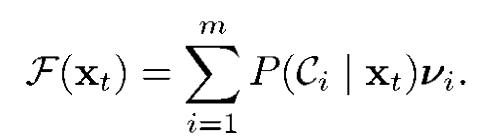
阿笠博士發明的第一個道具，**內含兩個旋鈕，一個調聲調，一個調音量，調整至適當的聲音後，可以讓柯南發出所有他曾經聽過的人聲。**一般用於使毛利小五郎（被麻醉的小五郎總是低著頭，以嚴肅的表情進行推理，「沉睡的小五郎」這個綽號即由此而來）（號碼設定在轉盤上的59號）或鈴木園子（通常是因小五郎不在現場才選她，後來自稱是高中女偵探），還有一位是山村刑警（在園子和小五郎不在場時或者一些情況時才用的），使用時搭配手錶型麻醉槍使他們沉睡後，**模仿其聲音進行案件推理**，而阿笠博士則是模仿柯南說話的口形推理。**缺點是沒聽過的人聲就不能模仿**，曾在動畫第116-117話《推理小說家失蹤事件》中，不小心射中別人，正想將計就計時卻發現沒聽過他的聲音而不能利用他來推理。

1. 相關應用  
   - 現在許多KTV的麥克風系統可以設定輸出為男聲、女聲、唐老鴨聲等等，達到real-time將說話者的聲音變調的效果。  
   - 有些新聞採訪為了保護受訪者會將其聲音進行變調處理。
2. 動機  
   - 我們講話時所要表達得**訊息**是非常重要的，而有時候我們更在意**說話者是誰（Speaker Identity）。**每個人的**說話聲音特徵都不相同**，因此我們可以輕易辨識是誰在說話。  
   - 如果啞巴人士（speech impaired）想要藉由機器發出自己製造出來獨一無二的說話聲音的話，我們就可以藉由**Voice Conversion** 來達到目的，或者是因為某種意外讓人無法在說話了，我們可以**讓機器去聽他以前的聲音（video）**，來讓機器學習他的說話聲音。
3. 探討重點  
   - 我們要研究的方向是聲音轉換(Voice Conversion)，Voice Conversion的研究問題跟Speaker adaptation跟Speaker recognition非常相似。不同的是在Voice Conversion中，最後所輸出的語音訊號是要給人類聽到的。  
   - 既然即時變調系統現已存在（如相關應用第一項），是否可能specifically訓練出特定的人聲模型？以求達即時將輸入聲音變調後輸出成特定人聲的效果。
4. 相關技術介紹  
   聲音轉換(Voice Conversion)
5. 傳統的問題定義：盡可能將說話者甲的聲音(source speaker)特徵轉換並重組成近似說話者乙的聲音(target speaker)特徵。
6. 解決方法：想辦法找到一個轉換函式(conversion/mapping function)。
7. 流程圖[1]：  
   - Conversion function：當Source speaker跟 Target speaker先經由Harmonic + Noise model system(HNM)分析進到系統中，再轉換成Spectral envelope(可以表達聲音的特徵[2][3][4])。  
   在Incremental Learning box(可以重複做好幾個Iterations)中做了DTW的Alignment 和 EM algorithm，最後經過Least Squares(LS) Optimization，產生Conversion function。如Fig1。  
   - 如何實作：
   1. Analysis/Synthesis Model：  
      - 整個Voice conversion的系統是based on在Harmonic + Noise model(HNM)[9][10][11]，提供了high-quality的語音描述。
   2. Spectral Parameters：分成voiced part 跟 unvoiced part的轉換。Voiced 的部分被拿來training Conversion function，而Unvoiced 的部分被拿來training noise。
   3. Learning Procedure：完整的learning procedure如Fig1。細節會在第8點提到。
   4. Voice Conversion System：當Conversion function得到了之後，就可以把Speech signal經由HNM分析，在通過Conversion Function 跟 Envelope transformation 最後處理Noise產生Converted speech(跟Target speaker相當接近)。完整的Voice transformation如Fig2。

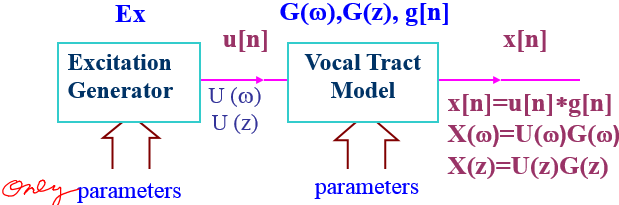
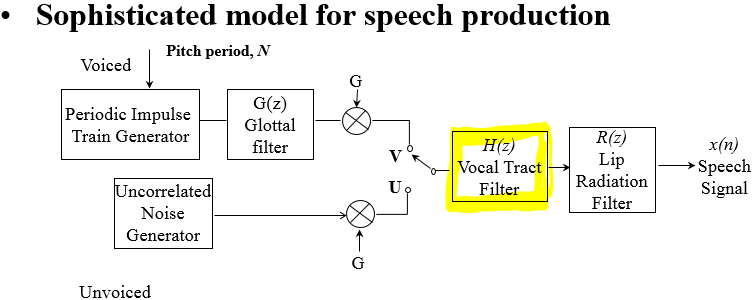
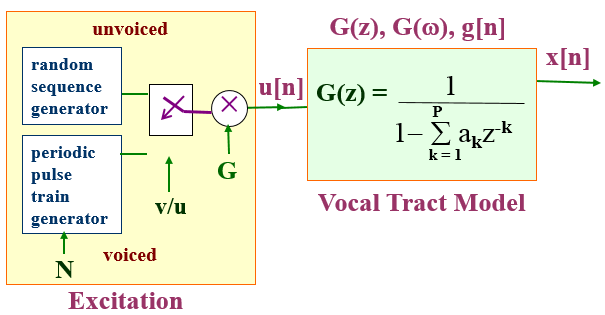
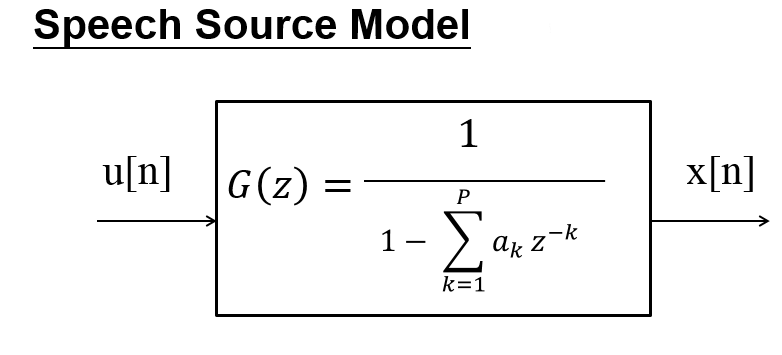




1. 決定聲音特徵： Average pitch、Formant structure、Vocal tract、Fundamental frequency *F0*、Phoneme durations、Prosody、Spectral envelope。
2. 聲音特徵參數化表示(Feature)：MFCCs。
3. Spectral conversion 的方法：  
   - Codebook mapping[5][6] ：使用Vector quantization流程在Source and Target speakers上。缺點：Poor quality 跟缺乏Robustness。  
   - Spectral interpolation approach[7][8] 。  
   - Acoustic space modeling with GMM。  
   - Harmonic + Noise Mode System [1]。  
   - PLS  
   - PLS + GMM
4. 聲音轉換的基本問題：如何選擇使用的模型(Model)，以在Oversmoothing和Overfitting之間取得平衡。
5. Training Conversion Function：怎麼從data中學習到spectral conversion function。  
   - Xt：spectral envelopes of the source which is p-dimensional vector of MFCC's   
   - Yt：spectral envelopes of the target which is p-dimensional vector of MFCC's  
   - Xt：t = 1,……n   
   - Yt：t = 1,……n  
   - F(Xt)：轉換Source envelope 到 Target envelope Yt where t = 1,….n  
   - 在這裡是使用連續機率模型GMM。  
   - 在Mapping Codebook的方法中，可以把這個問題化簡成low dimensional problem。
   1. Gaussian Mixture Model(GMM) ：  
      - GMM是一個很普遍很經典使用在很多pattern recognition techniques，可以有效率的運用在Speaker recognition[12][13][14]，GMM可以用機率分布去描述所觀察到的parameter。  
      - GMM可以想像成是HMM with Gaussian state-conditional distribution，如同老師課堂所解釋的，用一把一把的Gaussian去描述與音訊號。  
      - 為什麼用GMM?  
       因為我們在意的是segmental conversion functions 在時間t的converted envelope只有依據在相同時間下的Source envelope Xt。   
       又因為GMM提供了soft classification，用許多的component組合，每個component就是一個unimodal Gaussian distributions ，在GMM的model，每個acoustic class就是用Gaussian的mean vector（Ui）跟它的covariance matrix（SIGMAi）所組成，而ALPHAi就是一個mixture weights。如{1}{2}。  
      {1}:  
        
      {2}:  
        
      因為條件機率，當我們看到X時，P(Ci|X)可以從{1}推論出來:  
      {3}:  
         
      而{2}{3}合起來可以推導出:  
      {4}:   
        
      Training GMM的parameter，使用EM algorithm [15]。如果要詳細看EM用在Gaussian mixtures 可以在[16][12]找到。
      1. EM Algorithm：可以參考老師講義[]  
         - 在實作EM algorithm時最重要的是它的initialization，EM algorithm只保證converge toward a stationary point of  the likelihood function，EM algorithm的initialization不只影響convergence rate也影響final estimate。
   2. Conversion Function：  
      - 我們現在要找一個conversion function F( )，這個function可以轉換source data set {Xt}的每個vector到相對應的target data set {Yt}，Conversion Function定義[1]。  
      {5}:  
        
      - Minimum mean square error(MMSE)計算Target vector[1][17][18]。  
        
        
      - Training conversion function的parameters，藉由minimize the total squared conversion error用least squares optimization。  
      {7}:  
        
      - 在這篇paper中使用spectral parameters基本上是cepstral coefficients，而Total squared error 是 minimize 所有的 GMM 下的acoustic space。從算式{5}比較三個不同種類的conversion function。
      1. Full Conversion：general case of {5}，parameter of the GMM and parameter conversion function are  unconstrained.
      2. Diagonal Conversion：  
         - 用Diagonal covariance matrix 是非常常見的一種實作，可以有效減少計算量。  
         - 在cepstral parameters的case中，這種modification是非常適合的，因為correlation between distinct cepstral coefficients 是非常小的。
      3. VQ-Type Conversion：  
         - 假如我們省略訂正的term可以使

{5} => {9}:  


* 1. Optimization of the Conversion Function：

1. 聲音轉換的難點：  
   - 品質問題(Quality)  
   - 實際上training data通常很有限  
   - 缺乏客觀衡量performance好壞的判準  
   - 一個人講同一句話也可能有多種方式
2. 結論
3. 聲道(Vocal Tract)模型和其他人類聲音特徵參數  
     
   　　 講義第七章page 15  
     
   　講義第七章page 17  
    講義第七章 page 18  
     
    講義第七章 page 19

努力點：希望能從每個人的聲音data中找出個人化的聲道模型參數和其他特徵參數

1. 實作  
   - 方法一：語音辨識🡪選擇聲調🡪語音合成  
   - 方法二：不經辨識，直接修改輸入波型，輸出時即達特定變調的效果
2. 參考資料  
   [1]Continuous Probabilistic Transform for Voice Conversion-Yannis Stylanou and Eric Moulines  
   [2]Speaker-identifying features based on formant tracks-J.Acoust  
   [3]The Acoustics of Crime-H.Hollien  
   [4]Acoustic characteristics of speaker individuality:Control and Conversion-H.Kuwavara and Y.Sagisaka  
   [5]Voice Conversion through vector quantization-M.Abe and S.Nakamura   
   [6]Voice Conversion through vector quantization-J.Acoust.Soc.Jpn  
   [7]Speech spectrum transformation by speaker interpolation-N.Iwahashi and Y.Sagisaka  
   [8]Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks -N.Iwahashi and Y.Sagisaka  
   [9]HNS : Speech modification based on a harmonic + noise model-J.Laroche  
   [10]Harmonic plus noise models for speech combined with statistical methods for speech and speaker modification-Y.Stylianou  
   [11]High-quality speech modification based on a harmonic + noise model-Y.Stylianou  
   [12]Robust text-independent speaker identification using Gaussian mixture speaker models-D.A.Reynolds and R.C.Rose  
   [13]Text independent speaker identification using automatic acoustic segmentation- D.A.Reynolds and R.C.Rose  
   [14]Continuous probabilistic acoustic map for speaker recognition-B.L.Tseng and F.K.Soong and A.E.Rosemberg  
   [15]Maximum likelihood from incomplete data via the EM algorithm-A.P.Dempster and N.M.Laird and D.B.Rubin  
   [16]Finite Mixture Distributions-B.S.Everitt and D.J.Hand  
   [17]Fundamentals of Statistical Signal Processing-S.M.Kay  
   [18]Introduction to Multivariate Analysis-C.Chatfield and A.J.Collins  
   []維基百科  
   []李琳山教授數位語音處理上課講義