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

一、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。

二、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。

三、動機

- 我們講話時所要表達得**訊息**自然是非常重要的，但有時候我們更在意**說話者是誰(Speaker Identity)**。每個人說話的**聲音特徵都不相同**，因此我們能夠辨識出是誰在說話。
- 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**聲音轉換(Voice Conversion)**來達到目的，或者是因為某種意外讓人無法再說話了，我們可以**讓機器去聽他以前的聲音片段**，來讓機器學習他的說話聲音。

四、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 類似。不同的是在 Voice Conversion 中，最後所輸出的結果是要給人類聽到的語音訊號。
- 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。

五、相關技術介紹

聲音轉換(Voice Conversion)

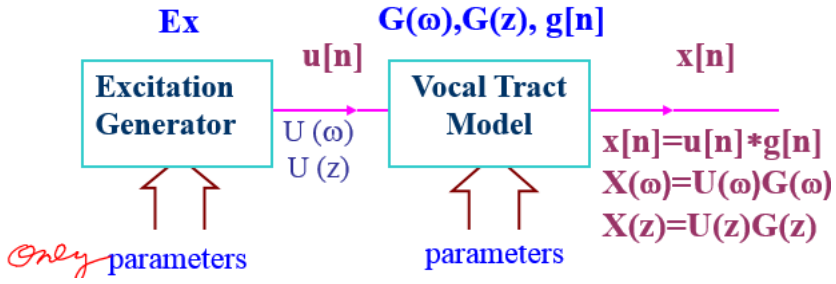
1. 傳統的問題定義：盡可能將說話者甲(Source speaker)的聲音特徵轉換並重組成近似說話者乙(Target speaker)的聲音特徵

2. 解決方法：想辦法找到一個轉換函式(conversion/mapping function)

3. 聲音特徵(feature)的種類：spectral envelope、vocal tract、prosody

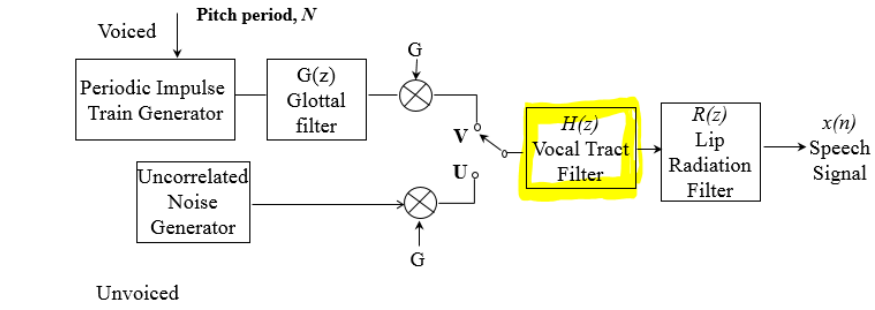
4. 聲音特徵參數化(parameterization)表示：formant structure、MFCCs、LSFs、MGCs、fundamental frequency F_0

5. 以聲道(vocal tract)模型如何將聲音特徵參數化為例
   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特徵參數

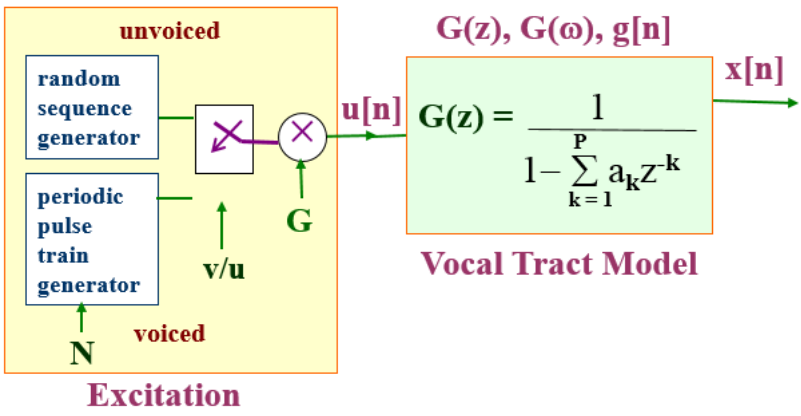


講義第七章 page 15

- **Sophisticated model for speech production**

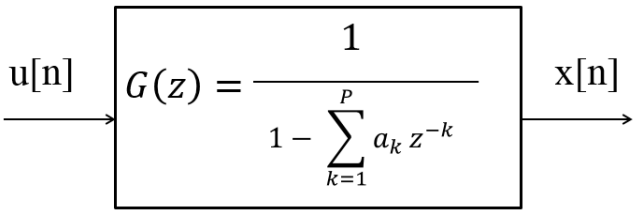


講義第七章 page 17



講義第七章 page 18

Speech Source Model



講義第七章 page 19

6. 表達轉換函式的方法：
   - Codebook Mapping[3][4] ：使用 Vector Quantization 流程在 Source and Target speakers 上。缺點：Prone to error due to discontinuity, need large training data
   - Spectral interpolation approach[5][6]
   - Acoustic space modeling with GMM: 優點：提供一個 performance 不錯的選擇。缺點：oversmoothing, overfitting, time-independent→Solution: MLE(maximum likelihood estimation), frequency warping with a GMM
   - HMM-based: 優點：不需要太多的 training data 即可做到 voice adaptation。缺點：需要大量預存的 speakers，以及 parallel training data，實行上有困難
   - Eigenvoice method: 優點：方法很簡單(PCA)。缺點：同 HMM-based
   - PLS + GMM(Partial Least Square Regression + Gaussian Mixture Model)[12]: 最重要的優點：have good performance on new data with only a small amount of training observations
   - NHM (Harmonic + Noise Model)

7. 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

8. 試舉 NMH 為例
   - NHM 流程圖[1]：
   - Conversion function：當 Source speaker 跟 Target speaker 先經由 Harmonic + Noise model system(NHM)分析進到系統中，再轉換成 Spectral envelope(可以表達聲音的特徵[2])。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：

   甲、Analysis/Synthesis Model：
   - 整個 Voice conversion 的系統是 based on 在 Harmonic + Noise model(HNM)[7]，提供了 high-quality 的語音描述。

   乙、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。

   丙、Learning Procedure：完整的 learning procedure 如 Fig1。細節會在第 9 點提到。

   丁、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

訊號，產生 Converted speech(跟 Target speaker 相當接近)。完整的 Voice transformation 如 Fig2。

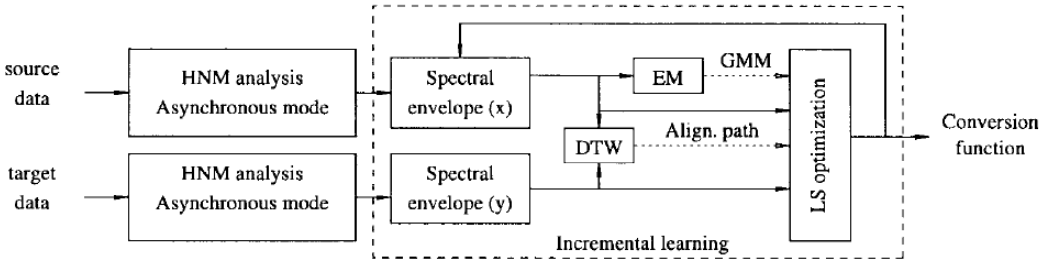


Fig. 1. Block diagram of the learning procedure.

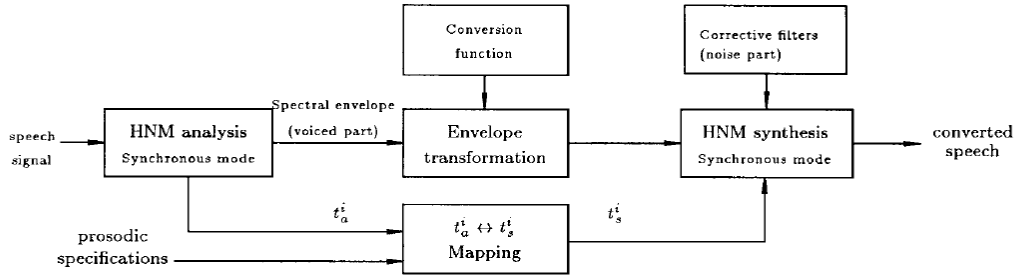


Fig. 2. Block diagram of the voice conversion system (not including the training of the spectral conversion function). t_a^i: analysis time-instants, t_s^i: synthesis time-instants.

9. 訓練轉換函式的方法：In general，如何透過 data，train 出轉換函式[1]：
   - 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 Xt 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。

   甲、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[8] [9]，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 $N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，在 GMM 的 model，每個 acoustic class 就是用 Gaussian 的 mean vector 跟 covariance matrix 所組成，個別的 component 有個別的 weights。如{1}{2}。

   {1}:

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$$

{2}:

$$N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

因為條件機率，當我們看到 X 時，P(CilX)可以從{1}推論出來:

{3}:

$$P(\mathcal{C}_i \mid \mathbf{x}) = \frac{\alpha_i N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{j=1}^{m}\alpha_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}.$$

而{2}{3}合起來可以推導出:

{4}:

$$P(\mathcal{C}_i \mid \mathbf{x}) = \frac{\alpha_i|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]}{\sum_{j=1}^{m}\alpha_j|\boldsymbol{\Sigma}_j|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]}.$$

- Training GMM 的 parameter，使用 EM algorithm[10]。如果要詳細看 EM 用在 Gaussian mixtures 可以在[8]找到。

i. EM Algorithm：可以參考老師講義[14]

- 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。[1]

乙、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}:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i \mid \mathbf{x}_t)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]. \quad (5)$$

The conversion function $\mathcal{F}$ is entirely defined by the p-dimensional vectors $\boldsymbol{\nu}_i$ and the $p \times p$ matrices $\boldsymbol{\Gamma}_i$, for $i = 1,\cdots,m$ (where m is the number of mixture components).

- Minimum mean square error(MMSE)計算 Target vector[1][11]。

$$E[\mathbf{y} \mid \mathbf{x} = \mathbf{x}_t] = \boldsymbol{\nu} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}) \quad (6)$$

where $E[\,]$ denotes expectation, and $\boldsymbol{\nu}$ and $\boldsymbol{\Gamma}$ are, respectively, the mean target vector

$$\boldsymbol{\nu} = E[\mathbf{y}]$$

and the cross-covariance matrix of the source and target vectors

$$\boldsymbol{\Gamma} = E[(\mathbf{y} - \boldsymbol{\nu})(\mathbf{x} - \boldsymbol{\mu})^T]$$

- Training conversion function 的 parameters，藉由 minimize the total squared conversion error 用 least squares optimization。

{7}:

$$\epsilon = \sum_{t=1}^{n} \left\| \mathbf{y}_t - \mathcal{F}(\mathbf{x}_t) \right\|^2.$$

- 在[1]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。

i. Full Conversion：第{5}的大部分 case，GMM 的參數跟 Conversion function 的參數是沒有限制的，計算量大。

ii. 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。

iii. VQ-Type Conversion：
- 假如我們省略訂正(Correction)的 term 可以使

{5}=>{9}:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i \mid \mathbf{x}_t)\boldsymbol{\nu}_i.$$

丙、Optimization of the Conversion Function：先把 $P(\mathcal{C}_i \mid \mathbf{x}_t)$ 簡化為 $p_t(i)$。

i. Full Conversion：計算量很大。
- 上述{5}可以推導成
{10}:如下

$$\mathbf{y}_t = \sum_{i=1}^{m} p_t(i)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]$$

- 上述{10}可以推導成 matrix 形式{11}
{11}: 如下

$$\begin{aligned} \mathbf{y} &= \mathbf{P} \cdot \boldsymbol{\nu} + \boldsymbol{\Delta} \cdot \boldsymbol{\Gamma} \\ &= \begin{bmatrix} \mathbf{P} & \vdots & \boldsymbol{\Delta} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\nu} \\ \cdots \\ \boldsymbol{\Gamma} \end{bmatrix} \end{aligned}$$

- Y 是 n*p 的 matrix 表示 Target spectral。
- P 是 n*m 的 matrix 表示 conditional probabilities。
- $\boldsymbol{\Delta}$是 n*pm 的 matrix

$$\boldsymbol{\Delta} = \begin{bmatrix} p_1(1)(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1^T} & p_1(2)(\mathbf{x}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1^T} & \cdots & p_1(m)(\mathbf{x}_1-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1^T} \\ p_2(1)(\mathbf{x}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1^T} & p_2(2)(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1^T} & \cdots & p_2(m)(\mathbf{x}_2-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1^T} \\ \vdots & \vdots & & \vdots \\ p_n(1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1^T} & p_n(2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1^T} & \cdots & p_n(m)(\mathbf{x}_n-\boldsymbol{\mu}_m)^T\boldsymbol{\Sigma}_m^{-1^T} \end{bmatrix}_{(n\times mp)}$$

- $$\boldsymbol{\nu} = \left[\boldsymbol{\nu}_1 \vdots \boldsymbol{\nu}_2 \vdots \cdots \vdots \boldsymbol{\nu}_m\right]^T_{(m\times p)}$$

- $$\boldsymbol{\Gamma} = \left[\boldsymbol{\Gamma}_1 \vdots \boldsymbol{\Gamma}_2 \vdots \cdots \vdots \boldsymbol{\Gamma}_m\right]^T_{((m\times p)\times p)}$$

- 上述{11}是 Standard Least-squares problem 他的解可以由線性代數的 normal equation [11] 解變成

{14}:如下

$$\left(\begin{bmatrix}\mathbf{P}^T \\ \cdots \\ \boldsymbol{\Delta}^T\end{bmatrix}\cdot\begin{bmatrix}\mathbf{P} & \vdots & \boldsymbol{\Delta}\end{bmatrix}\right)\cdot\begin{bmatrix}\boldsymbol{\nu} \\ \cdots \\ \boldsymbol{\Gamma}\end{bmatrix} = \begin{bmatrix}\mathbf{P}^T \\ \cdots \\ \boldsymbol{\Delta}^T\end{bmatrix}\cdot\mathbf{y}$$

$$\begin{bmatrix}\mathbf{P}^T\mathbf{P} & \vdots & \mathbf{P}^T\boldsymbol{\Delta} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Delta}^T\mathbf{P} & \vdots & \boldsymbol{\Delta}^T\boldsymbol{\Delta}\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{\nu} \\ \cdots \\ \boldsymbol{\Gamma}\end{bmatrix} = \begin{bmatrix}\mathbf{P}^T\mathbf{y} \\ \cdots \\ \boldsymbol{\Delta}^T\mathbf{y}\end{bmatrix}$$

ii. Diagonal Conversion：計算量會是 full conversion 除以 p/4 倍[1]。

- Diagonal case 的 Conversion function 最佳化是簡化的，因為 covariance matrices of the GMM $\boldsymbol{\Sigma}_i$ 跟 conversion matrices $\boldsymbol{\Gamma}_i$ are diagonal。
- 上述的{10}可以簡寫成

{16}:如下

$$y_t^{(k)} = \sum_{i=1}^{m} p_t(i)\left[\gamma_i^{(k)}\left(x_t^{(k)} - \mu_i^{(k)}\right)/\sigma_i^{(k)} + \nu_i^{(k)}\right]$$

- 上標 K 表示第 K 個 coordinate(例如$y_t^{(k)}$就相當於 Yt)
- $\sigma_i^{(k)}$跟$\gamma_i^{(k)}$相當於$\boldsymbol{\Sigma}_i$跟$\boldsymbol{\Gamma}_i$的第 K 個 diagonal elements。
- 上述的{14}可以簡寫成

{17}:如下

$$\begin{bmatrix}\mathbf{P}^T\mathbf{P} & \vdots & \mathbf{P}^T\boldsymbol{\Delta}^{(k)} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{P} & \vdots & \boldsymbol{\Delta}^{(k)^T}\boldsymbol{\Delta}^{(k)}\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{\nu}^{(k)} \\ \cdots \\ \boldsymbol{\gamma}^{(k)}\end{bmatrix} = \begin{bmatrix}\mathbf{P}^T\mathbf{y}^{(k)} \\ \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{y}^{(k)}\end{bmatrix}$$

iii. VQ-Type Conversion：

- 是{17}的 special case 省略 the diagonal matrix elements $\boldsymbol{\Gamma}^{(k)}$，因此 Conversion vector

變成

{20}:如右 $$\nu^{(k)} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}^{(k)}.$$

10. 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

六、結論

- 2012 年時 PLS + GMM 宣稱他們是聲音轉換的 state-of-the-art
- 根據以上的探討，想要做出阿笠博士領結型變聲器技術上應該不是問題，但也不容易，其中難點是 1. 要如何更精細、明確地表達一個特定的人的聲音特徵　2. 如何訓練出更 robust 的轉換函式/模型，最極端的目標是：只需要聽到某人講幾句話，即可任意模仿他的聲音講所有的話

七、相關軟體

- 劍橋大學工程學院的 Steve Young 和 Hui YE 在 2004 做的一個簡易版聲音轉換軟體，可以透過調整一些聲音特徵參數達到不錯的變調效果，連結的網站可以免費下載 http://svr-www.eng.cam.ac.uk/~hy216/VoiceMorphingPrj

八、分工

- 成毅：問題發想、問題定義、Survey 相關技術和應用
- 冠鈞：NHM 模型細節、訓練轉換函式的方法

九、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]Voice Conversion through vector quantization-M.Abe and S.Nakamura

[4]Voice Conversion through vector quantization-J.Acoust.Soc.Jpn

[5]Speech spectrum transformation by speaker interpolation-N.Iwahashi and Y.Sagisaka

[6]Perceptually Weighted Linear Transformations for Voice Conversion -Hui Ye and Steve Young

[7]HNS : Speech modification based on a harmonic + noise model-J.Laroche

[8]Robust text-independent speaker identification using Gaussian mixture speaker models-D.A.Reynolds and R.C.Rose

[9]Continuous probabilistic acoustic map for speaker recognition-B.L.Tseng and F.K.Soong and A.E.Rosemberg

[10]Maximum likelihood from incomplete data via the EM algorithm-A.P.Dempster and N.M.Laird and D.B.Rubin

[11]Solving Least-Squares Problems-C.L.Lawson and R.J.Hanson

[12]2012-ASLP-Voice Conversion Using Dynamic Kernel Partial Least Squares Regression-E Helander

[13]維基百科

[14]李琳山教授數位語音處理上課講義