Final project
Voice conversion from child to adult

107011133 王靖淳 107061249 黃凱琳

動機:我們會想要將聲音轉換用在將小孩的聲音轉為大人的聲音是因為看到許多案子的受害者是小孩，若是在家中沒有大人時，有人按門鈴或是打電話，聽到小孩的聲音知道目前只有小孩在家，而小孩的反抗能力不高，作案的成功機率提高，希望能降低小孩受害的風險所以想將小孩的聲音轉為大人的聲音。

觀察:小孩與大人聲音除了音高、音色不同以外，還存在許多差異，小孩的口腔肌肉不足，不太會控制發音氣流，使他們講話的內容不太清楚，另外也有語速的問題，小孩容易在一句話中某幾個單子發音特別長，或是中間有較長的時間不講話。

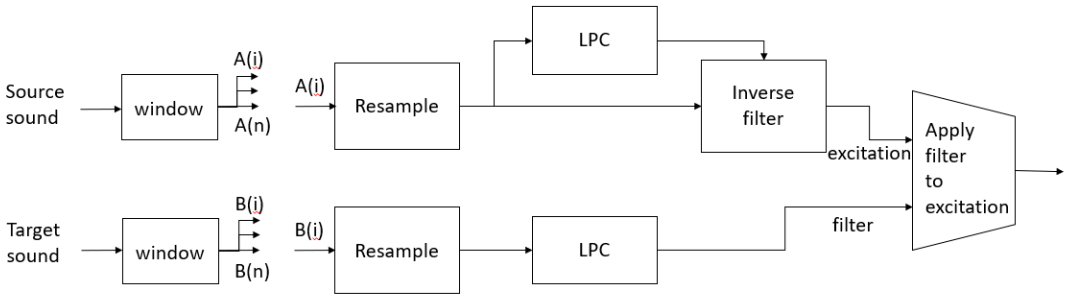
藉由以上的觀察分為幾點目標:

1. **聲音:音高、音色**
2. **每個音的速度**
3. **講話清晰度**

我們會針對第 1、2 點做改變，第 3 點講話清晰度的部份我們這裡沒有作探討，但認為可能的方向會是結合母音辨識，將每個音先做辨識再根據對應的發音更改共振峰的資訊或許能讓發音更準確，但這同時存在一個問題是辨識的部分是否能夠準確，因為在辨識時誤判為另一個音的話後面所做的一切都會是錯的，可能會造成語音轉換後的理解性很差。

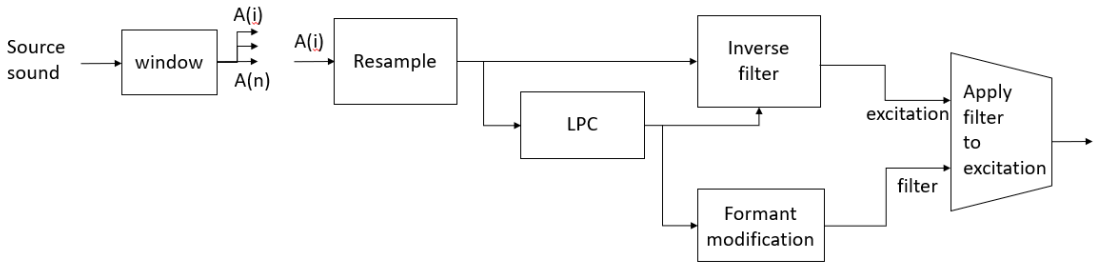
改變聲音:

LPC test



這是我們起初的設計架構，將源音檔的利用 LPC 抽出的語音特徵過濾掉後送入目標音檔利用 LPC 抽出的語音特徵參數構成的濾波器，以獲得另一種聲音特徵，這裡的測試結果是源音檔與目標音檔的內容混雜不清。

對此我們做了幾個測試是調整兩音檔的內容為相同的一個發音且速度相同，只有聲音不同，依照我們的想法是 LPC 可以取得共振峰的資訊，那我們源音檔取出的殘餘訊號代表的是基頻資訊，如果將基頻資訊移動到目標的基頻，搭配上目標發音的共振峰資訊是否能利用這個方法更改到共振峰，讓聲音聽起來比單純改基頻更具有目標音檔的特徵，但這裡的測試結果仍然是兩個很分離的聲音。



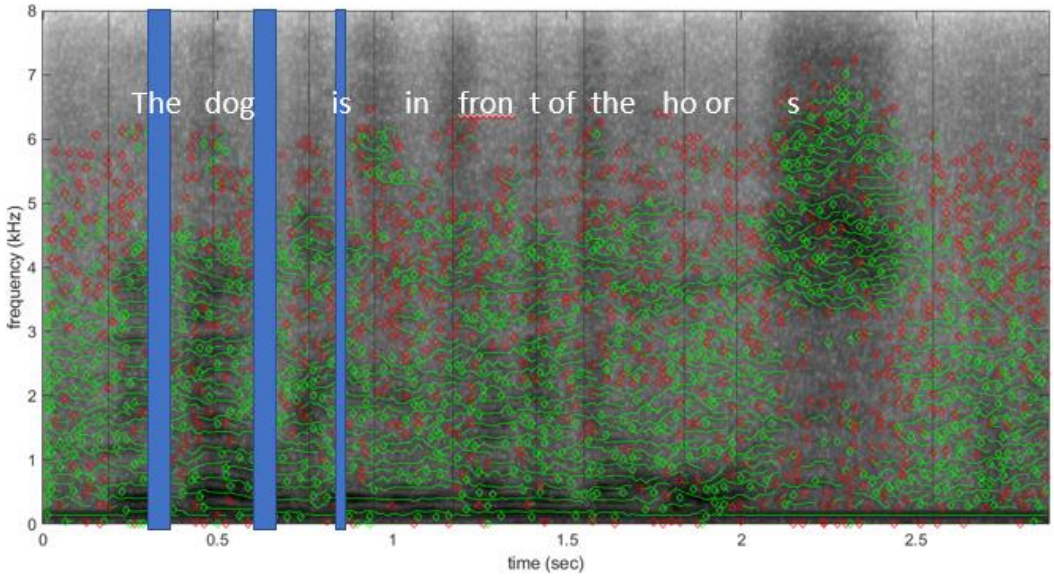
基於上面的結果不理想，我們又改動了架構，將目標語音拿掉，改用以參數調整的語音變換系統，會拿掉目標語音的原因是認為或許是目標與源音檔相差太大，不管是基頻或是共振峰，我們希望能以參數調整來記錄中間較小的改動聽起來的效果。

因為調整參數會變成我們需要蒐集的是小孩與大人之間的共振峰的差別，這裡我們有在一些文章中找到關於更改共振峰的資訊，可能造成的問題需要用其他額外的方法解決這個問題，以及小孩與大人與老人的共振峰的比例，但我們未能實做出來所以還沒辦法測試，將會把這些找到的資訊放在文獻參考。

Pitch shift

在改變聲音這部分我們選擇改變音高，以參數調整每個小孩的聲音能轉換到的聽起來最順耳且成熟的聲音。

改變每個音的速度



因為在作業中利用每時刻找可連接的頻率峰值形成基頻諧波，上圖綠色線的部分是連接好形成的基頻諧波，而用方塊標記的是找到卻未能連線的峰值，綠色方塊是此時刻的某些頻率峰值都指向前一時刻的相同頻率峰值，除了被選擇連線的峰值外剩餘的峰值都會被標記，而紅色方塊是此時刻的某一頻率的一固定距離內沒有找到上一時刻的連接頻率而被標記，從時域與頻域交互觀察能發現若是在不同

的字音轉換時會讓訊號不穩定而造成頻域沒有明顯的峰值，且容易抓到高頻峰值，所以正好利用此資訊作為切語句每個音的判斷依據。

在切線(黑線)切好後分出來的會是”the”、”dog”，可能某些音的前後會連接著前或後一個字的音，例如"the"那段聽到的可能是”the d”，隱約聽的到”d”的音，但這不影響聲音縮短，因為分好切割線後切掉的是線與線中間的一段時間的訊號(藍色區塊)，所以切到的範圍仍是”the”的發音，不會切到”d”而讓不同字與字之間的連接因為少了一部份”d”的音而變得不好辨識。

這裡會設定一些參數來調整切線位置，一是計算一幀需達多少個紅色方塊，二是因為多分布在高頻，需要達到多少頻率的紅色方塊才可被累計，三是雖然找的到紅色方塊的分布位置，但可以看到在上圖切線位置前後的一小段時間範圍內都有蠻多紅色方塊，為了讓切線能避免在一個切線位置卻設有超過一個切線所以設定切線之間的距離，四是藍色切掉的區塊與兩條切線之間的比例，因為黑色切線之間距離越大表示該發音越長，因此要調整整體的每個字發音長度接近就要將該發音切掉的部分加大。

聽測

第一部分

請受試者單純聽轉換後的音檔-大人說話的音檔來判斷音檔中的說話者年齡與性別以及音檔中的內容。

第二部分

請受試者聽完源音檔-小孩說話的音檔以及轉換後音檔-大人說話的音檔來評論可理解性、失真程度、自然性，評分為 1 到 5 分，1 為極差，5 為極佳。

可理解性:指的是語句中每個字的發音正確性，例如，在源音檔可能是母音 A，轉換後聽起來卻像母音 E，又或是轉換後某些子音變的不明顯，讓整句話聽起來不夠清楚或是誤認單字造成語意不同。

分數/等級		語音品質評估的定義
5	極佳	語句內容非常清楚
4	佳	語句稍微不清楚，但影響程度不大
3	普通	語句有點不清楚，但仍在可接受的程度內
2	差	語句不清楚，已經稍微影響到語句內容的判斷
1	極差	無法判斷語句內容

失真程度:語音轉換後可能產生的雜音造成合成語音品質下降。

分數/等級		語音品質評估的定義
5	極佳	與原始語音相比，沒有任何雜音
4	佳	有些為雜音，但大致還算悅耳
3	普通	有些雜音，但不至於到刺耳的程度
2	差	有雜音，讓人覺得有些刺耳
1	極差	與原始語音相比，雜音非常嚴重

自然性:語音轉換後聽起來是否像一般正常人所發出的聲音。

分數/等級		語音品質評估的定義
5	極佳	聽起來非常自然，完全像正常人聲
4	佳	聽起來還算自然，像正常人聲
3	普通	聽起來還算自然，稍微不像正常人聲
2	差	聽起來不是很自然，不是很像正常人聲
1	極差	聽起來一點都不像正常人聲

受試者資訊

	A	B	C	D	E	G
性別	男	男	女	女	男	男
耳機	Airpods 2	鐵三角耳塞系列	Airpods 2	Airpods 1	Airpods 2	Airpods pro
英聽成績(分級: A/B/C/F)	A	A	A	A	A	A
多益英聽(滿分495)	445	470	300	400	385	390

受試結果

語音 1

The dog is in front of the horse

	A	B	C	D	E	G	平均
語者性別	女	女	女	女	男	女	
語者年齡	45	25	中年30~40	26	18	40	31.5
可理解性	4	4	3	2	2	2	2.8
失真程度	4	4	4	2	3	4	3.5
自然性	4	5	4	4	4	4	4.2

語音 2

The fish is in the pond

	A	B	C	D	E	G	平均
語者性別	女	女	女	女	女	女	
語者年齡	35	32	中年 30~40	40	45	30	36
可理解性	4	3	3	4	2	4	3.3
失真程度	2	4	5	3	4	4	3.6
自然性	5	4	4	4	4	4	4.2

語音 3

The dog is on top of the shed

	A	B	C	D	E	G	平均
語者性別	女	女	女	女	女	女	
語者年齡	29	28	中年 30~40	45	30	30	33
可理解性	3	2	3	3	4	3	3
失真程度	4	3	4	2	3	2	3
自然性	4	4	5	3	5	4	4.2

結論

聲音的部分沒能實踐最後的架構，但對於共振峰與調整音高的原因更了解，在單純調整音高的部分小孩的聲音調為女生的聲音較為適合，因為原先小孩的音高就很高，調為男生的聲音會變得太低、聲音過於圓滑、內容模糊，我們取用的音檔小孩的年齡聽起來都太小，不管是男生或女生的小孩聲音都很容易被認成都是女生的聲音，在該年紀小孩的聲音不容易區分性別，所以通常都是調低音高，讓聲音聽起來像是我們認為的男聲，但我們認為的男聲通常是已經到了比原音檔評估的年紀還要大的年紀，還有，我認為音檔本身的小孩的聲音更具有奶音，不同小孩給的感覺不同，輕重不一，所以遇到奶音較重的音檔這裡是將音高再調低，比較可以消弭這個狀況。

語速的部分需要調整參數至能夠切出每個音，以避免以下狀況，”the dog”被分在一起，而切掉的是中間部分使得”d”的音正好被切掉。

聽測的部分，第一部分是在不知道專題題目與音檔內容的情況下請大家聽，有請同學記錄下聽到的內容或是有認出的單字，有的人是聽得出幾乎全對的內容，有的人只聽得出兩三個字，這裡先做測試聽聽看在不知道這原本是小孩音檔的情況下會不會覺得這個聲音是大人的聲音，避免在知道這是小孩轉換的音檔後怎麼聽都會覺得像小孩，並且小孩的音檔較慢、容易理解內容，所以避免大家在知道內容之後聽轉換後的音檔會覺得內容很好理解，就先做一次第一部

分的測試。

第二部分是在第一部分完成後，接著請大家聽小孩的音檔，在大家知道內容以及原本音檔的品質後聽轉換過的音檔，評分理解度、失真程度、自然性，在第一部份回答過內容後這裡再度評分理解度的原因是第一部分的測試有可能是因為有部分的人不習慣英聽，加上速度更快，所以無法辨認，才在第二部分設有在聽出內容後來評分轉換過的音檔是否能聽出相同的內容。而這裡的結果自然性是蠻高的，像是一般人的發音，理解度的部分或許是因為音高調低，聲音本身聽起來會較模糊，再把速度調快就更不容易理解，最後失真程度背景雜音較多，一部份可能的原因是源音檔背景就已經有許多雜音加上其他人聲，在縮短時間時雜音的感受更明顯。

文獻參考

1. Gina Upperman, Matthew Hutchinson, Brian Van Osdol, Justin Chen, Methods for Voice Conversion
2. 蘇培智,基於藉語音再取樣萃取共振峰變化之聲調調整技術, 2004/7
3. 孟猛,張樹武,基於與音分析與合成的高品質實時變聲方法,中華人民共和國國家知識產權局, 2006/1/11
4. Akash I. Mecwan, Vijay G. Savani, Shah Rajvi, voice conversion algorithm,ICAC3 '09: Proceedings of the International Conference on Advances in Computing, Communication and Control, Pages 615–619, 2009/1