

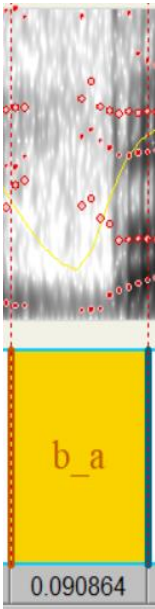
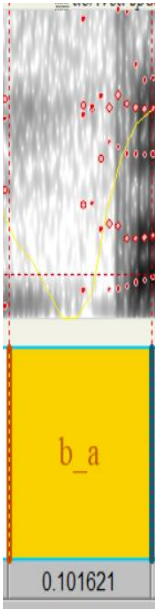
DSP HW2

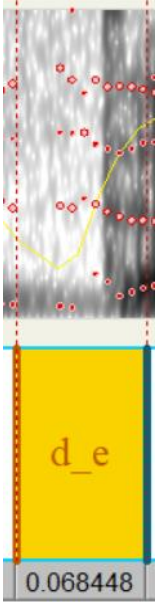
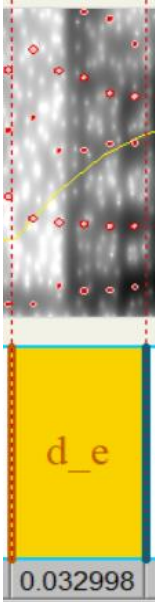
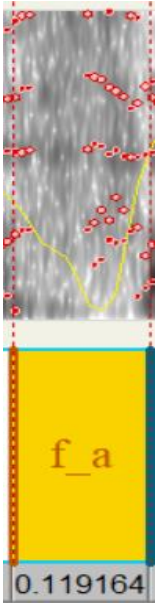
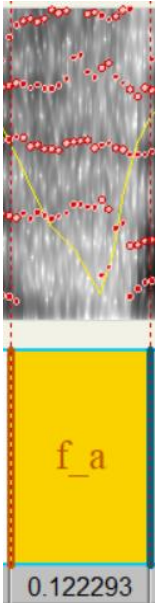
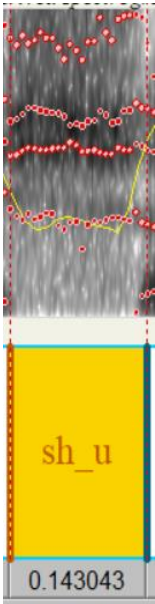
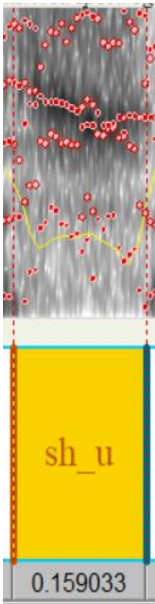
B09901080 電機三 吳宣逸

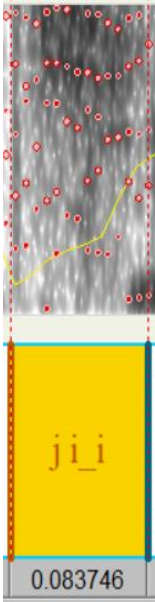
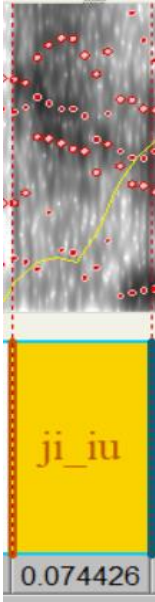
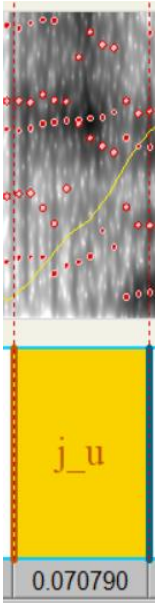
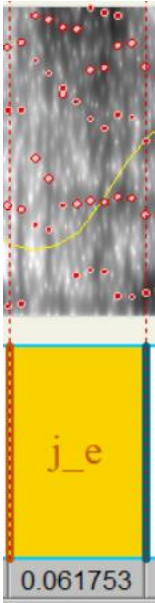
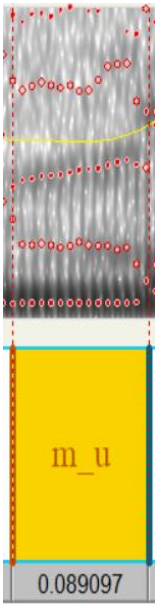
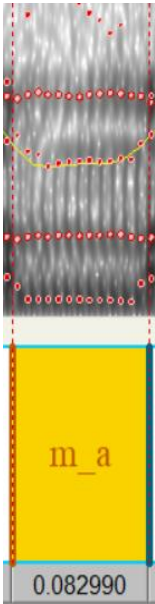
Part1

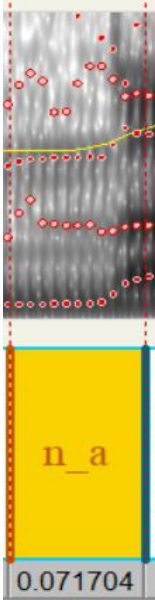
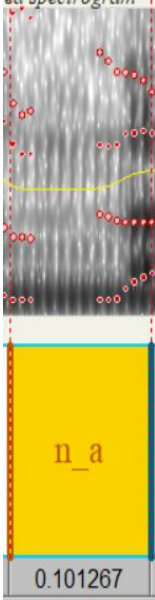
- 1. NTU\_m112901\_0
- 2. NTU\_m112902\_0
- 3. NTU\_m112903\_0
- 4. NTU\_m112905\_0
- 5. NTU\_m112908\_0
- 6. NTU\_m1129010\_0
- 7. NTU\_m1129012\_0
- 8. NTU\_m1129015\_0
- 9. NTU\_m1129025\_0
- 10. NTU\_m1129026\_0

Part 2

Phonetic Class			
Plosive	b [ㄅ]		

Plosive	d [ㇰ]	 <p>0.068448</p>	 <p>0.032998</p>
Fricatives	f [ㇱ]	 <p>0.119164</p>	 <p>0.122293</p>
Fricatives	sh [㇯]	 <p>0.143043</p>	 <p>0.159033</p>

Affricates	ji [ʨ]	 <p>ji_i</p> <p>0.083746</p>	 <p>ji_iu</p> <p>0.074426</p>
Affricates	j [ɕ]	 <p>j_u</p> <p>0.070790</p>	 <p>j_e</p> <p>0.061753</p>
Nasals	m [ɱ]	 <p>m_u</p> <p>0.089097</p>	 <p>m_a</p> <p>0.082990</p>

Nasals	n [ㄋ]		
--------	-------	---	---

### Part3

#### 1. (20%) What are the consistencies of the spectrogram in each phonetic class? (Plosive, Fricative, Affricate, Nasal)

以下的 intensity 可以從頻譜的深淺看出，大致是越深者 intensity 越大，越淺者 intensity 越小。

Plosive (塞音)：

頻率較集中於低頻。在前段 intensity 可能會先下降，但在中後段會快速上升到超過一開始的大小。

Fricative (擦音)：

頻率較集中於中高频。在發音中段 intensity 變小，結束時回復到與一開始差不多的大小。

Affricates (塞擦音)：

頻率較集中於中高频。intensity 會上升到超過一開始的大小，上升時間較 plosive 長一些。

Nasals (鼻音)：

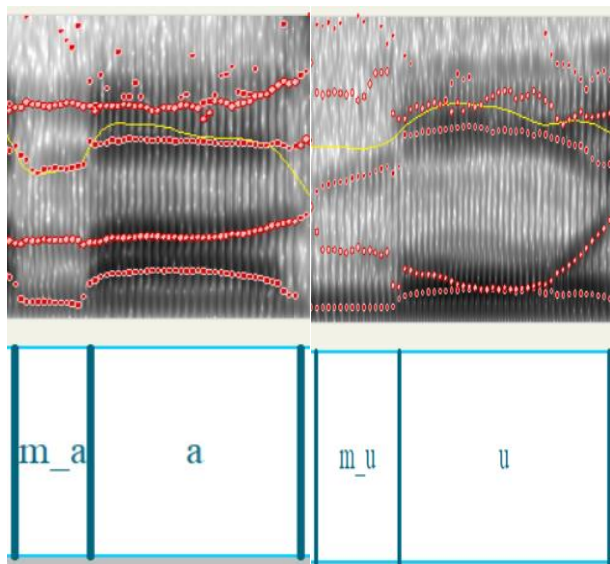
鼻音的 first formant 頻率相較其他三個 phonetic class 低，而且 intensity 幾乎不變化。

#### 2. (10%) Is the boundary between neighboring initial and final clear? What is the benefit of using "right-context dependent" initial model (ex: sh\_a) instead of pure initial model (ex: sh) to model initials? Please explain with examples.

相鄰的 initial 和 final 通常不會有明顯邊界，因為會受到 context 影響同個 phoneme 的發音。

相較於直接使用 initial 本身訓練 initial 模型，考慮到前述發音會受到 context 影響，使用 right-context dependent 能夠更明確地區分同一個 phoneme 可能產生的不同發音。

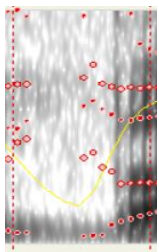
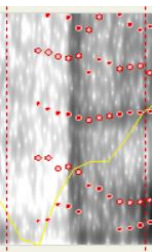
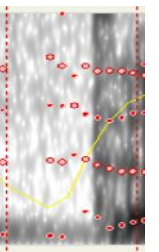
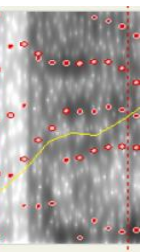
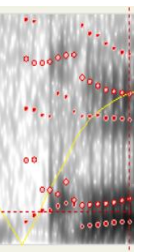
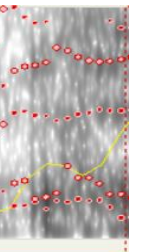
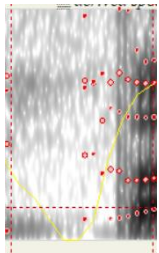
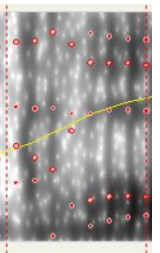
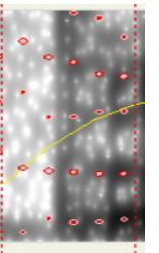
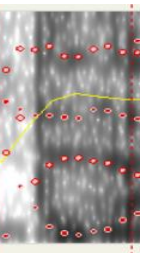
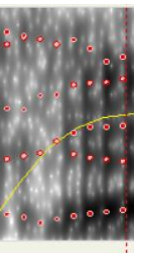
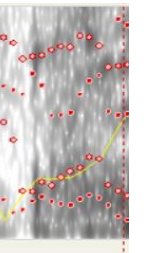
以「ㄋㄚˇ」和「ㄋㄨˇ」與為例，以下是對應的頻譜與讀音。其中的藍線是我用耳朵聽所能切出能最清晰分割 initial 和 final 的邊界，不論如何調整多少都還是會聽到 initial 或 final 其中一個聲音被對方混到。仔細觀察下方兩個「ㄋ」的四個 formant，除了 first formant 之外(因為 m 是 nasals)，另外三個 formant 受到 final 的影響而差異滿大的，可以預見兩個「ㄋ」的 MFCC 應該也有一定差異，分成兩個類別的音會更適當。因此根據 right-context dependent 將「ㄋㄚˇ」和「ㄋㄨˇ」的「ㄋ」根據 final 分成「m\_a」和「m\_u」兩種音將有助於提取各自更精確的特徵，提高模型的準確率。



3. (10%) What are the differences when pronouncing ㄇ & ㄇˊ? How can you tell the differences in spectrogram for ㄇ & ㄇˊ? (You may also want to compare ㄌ & ㄌˊ, ㄍ & ㄍˊ respectively)

因為「ㄇㄌㄍ」是 voiced 而「ㄇˊㄌˊㄍˊ」是 unvoiced，發 unvoiced 的時候聲門常開，因此可以明顯感受到比發「ㄇㄌㄍ」時更多的空氣自嘴巴噴出。

對照下方「ㄇㄌㄍ」、「ㄇˊㄌˊㄍˊ」的頻譜，可以用 voiced 和 unvoiced 的 frequency domain 中 excitation 在低頻造成的週期性進行比較，大致可以看出「ㄇㄌㄍ」在低頻的波型相對整齊、週期性較明顯，而「ㄇˊㄌˊㄍˊ」低頻的波型則較為紊亂無章。(voiced 在高頻的周期性不明顯故觀察低頻週期性)

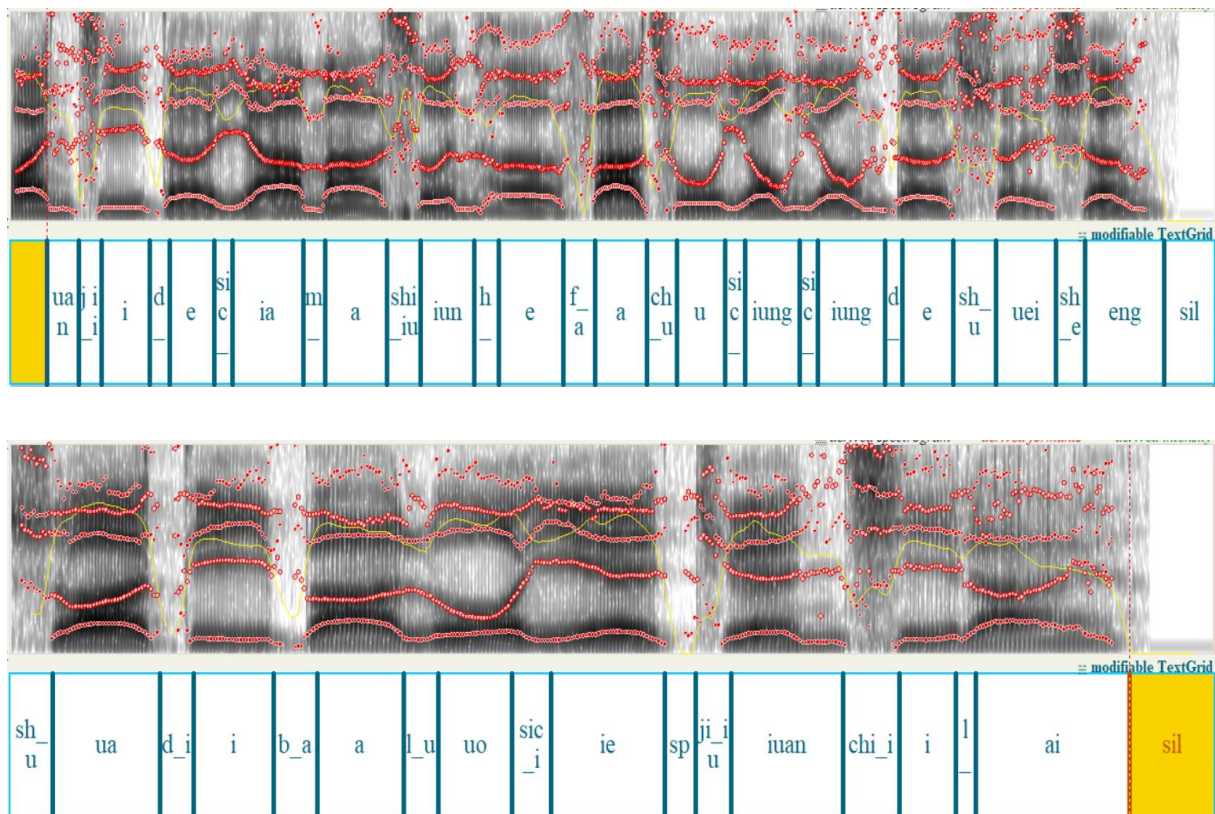
b [ㄇ]	p [ㄇˊ]	d [ㄌ]	t [ㄌˊ]	g [ㄍ]	k [ㄍˊ]
					
b_a	p_e	d_e	t_i	g_u	k_u
0.090864	0.084506	0.068448	0.060503	0.111952	0.084192
					
b_a	p_e	d_e	t_a	g_a	k_u
0.101621	0.046773	0.032998	0.057946	0.050384	0.090095



4. (10%) Take a look at the spectrogram of finals. Is there any simple rules to discriminate initials from finals provided only spectrogram?

若並沒有需要找出清楚的邊界，可以先觀察頻譜切出一些邊界，切出來的單位大致就是中文的音節。觀察自己標記的頻譜(下方為其中兩個)以及作業附的 phonetic\_class.pdf，會發現 initial 通常都是短音節，而 final 通常是長音節，因此從每段音節的時間長度可以大致分辨出頻譜中大部分 initial 和 final 的位置。

另外，若再加上 intensity 判斷(大致看頻譜顏色深淺)會發現，initial 大概都是在音訊號 attack 和 release 的部分，而 final 幾乎都處於音訊號 hold 的部分，若音節時距較難分辨 initial 和 final 時(短短相連或長長相連)時，可以用 intensity 加以判斷，例如下方第二張頻譜的「chi\_i i」即可以從 intensity 看出 initial 和 final 而不至於無法判斷。



## Bonus

答案應該是「純粹的聲」。

第一個線索是助教給的「川端康成的作品」以及「4 字詞」，於是首先排查出《伊豆舞孃》、《淺草紅團》、《水晶幻想》、《純粹的聲》、《少女的心》、《女性開眼》、《愛的人們》、《駒鳥溫泉》、《晚霞少女》、《一草一花》、《我的伊豆》、《伊豆之旅》、《東京的人》、《燕之童女》、《生為女人》、《愛與哀愁》、《落花流水》、《月下之門》、《竹聲桃花》共 19 個候選詞。

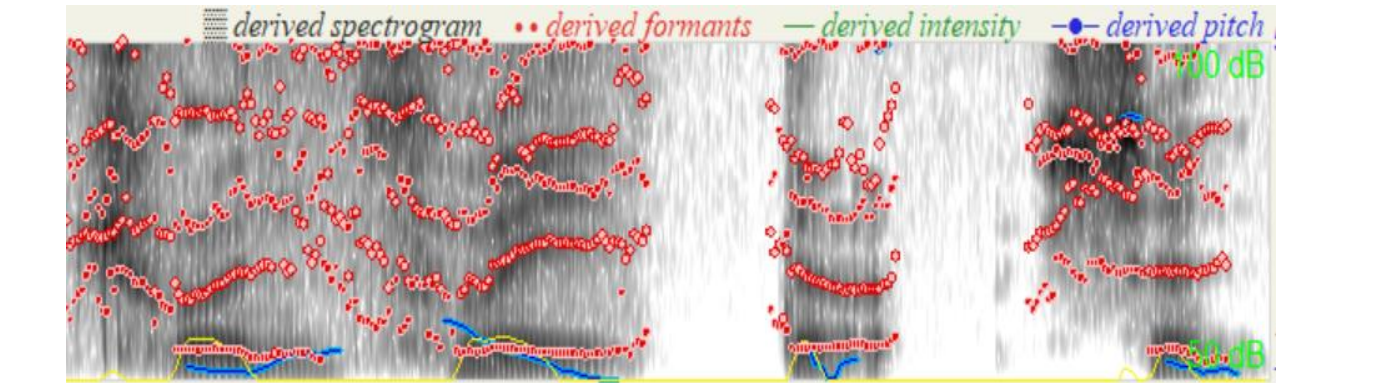
第二個線索是頻譜的第 1、4 個字的 final 結尾能觀察到 first formant 頻率到韻尾有稍微降低，因此從 part3 第一題對 part2 頻譜的分析，得知 nasal 有此一性質，因此先嘗試排查出第 1、4 字的 final 是 nasal ending(-n 或-ng)的詞，結果剩下《淺草紅團》、《純粹的聲》、《東京的人》、《生為女人》。

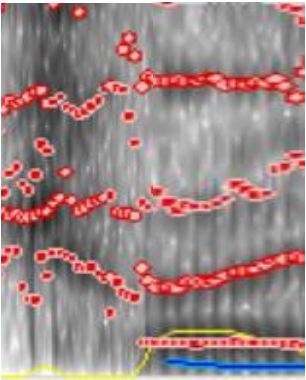
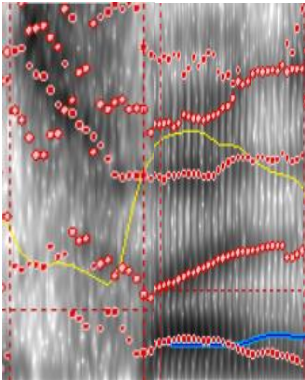
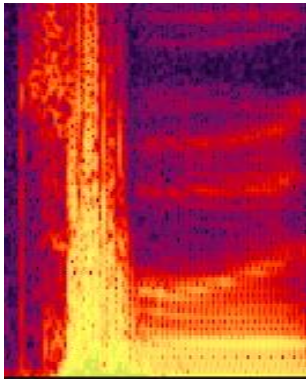
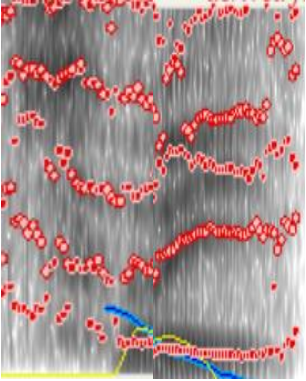
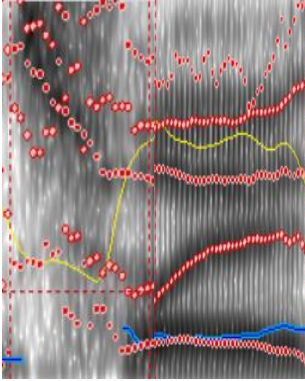
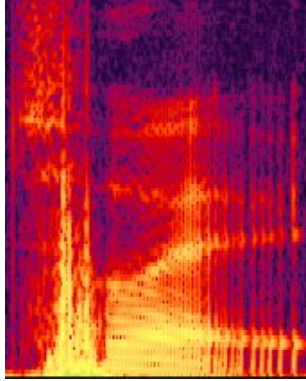
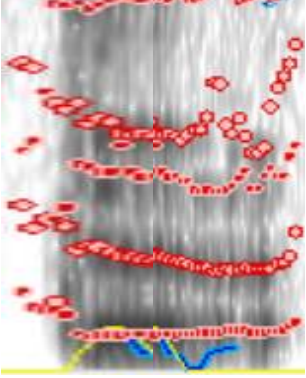
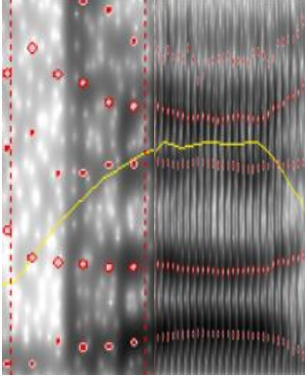
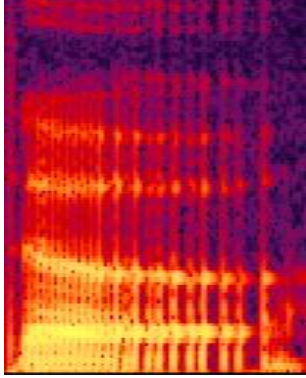
第三個線索是從頻譜可看出第 1、2、4 個字的 initial 集中於中高頻，所以很有可能是 fricative 或 affricates，唯一符合條件的只有《純粹的聲》，加上這門課是 DSP，因此更增加答案是《純粹的聲》的可能性。

最後，我有自己錄音製成頻譜，以及從本次作業中的素材擷取相同的 syllable 進行拼接，與助教的第一份頻譜逐

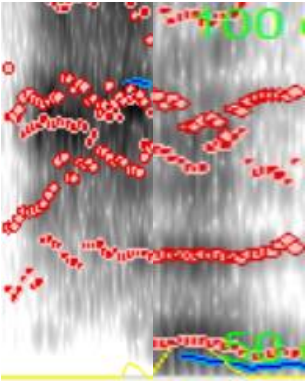
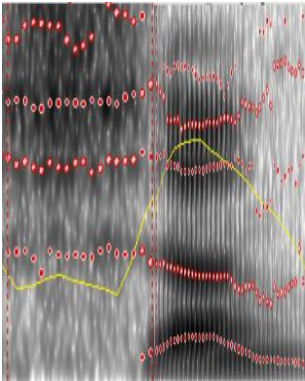
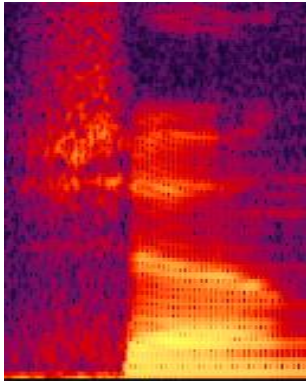
字比對，三者的四個 formant 的位置大致都可以對得上，因此答案應該是《純粹的聲》無誤。

ps. 助教給的頻譜：



		Given	我標記的	我自己的錄音
純	ch_u uen			
粹	ts_u uei			
的	d_e e			



聲	sh_e eng			
---	----------	--	---	--