Virtual dubber

Student : Sian-Yi Chen

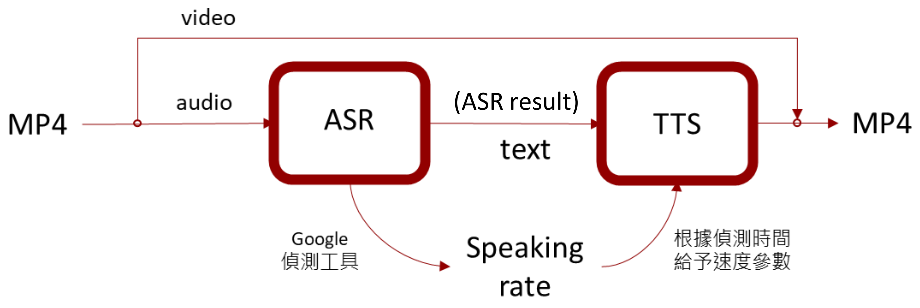
Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

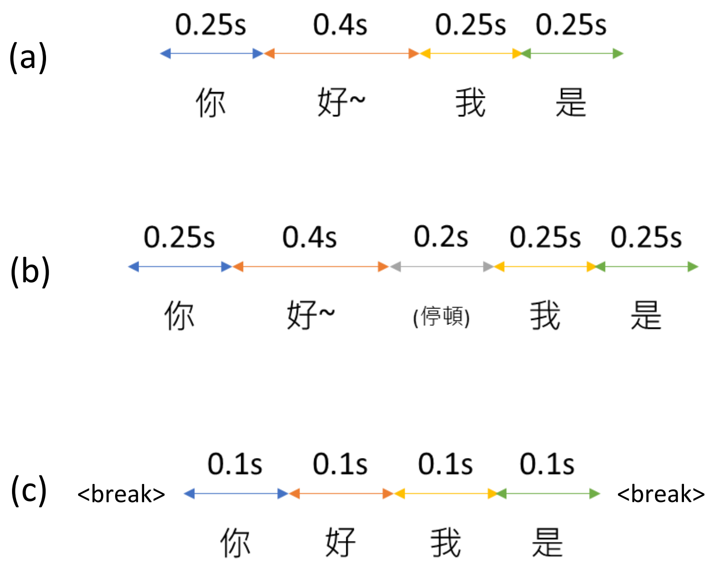
虛擬配音員

- Action item
  - □ Speaking rate control：去除語速控制後的不自然感

- Status report
  - □ PSOLA 是一種可以保持共振峰完整的語速控制方法，而 PSOLA 的使用方式為輸入音檔後將音檔變長 (語速變慢) 或變短 (語速變快)，因此須與 TTS 合併做使用，並透過 ASR 偵測語速工具判斷快慢
  - □ 然後發現 ASR 的語速偵測工具有一些瑕疵，以下是目前發現的三個問題：
    1. 在圖二 (a) 中，拖長音的後半段會因為聲音的減弱導致偵測不到，假設 “好” 被偵測持續時間長為 0.25s，那剩下的 0.15s 就會併入計算 “我” 的持續時間中，時長為 0.4s
    2. 在圖二 (b) 中，若拖長音後又有停頓，停頓的時間也會列入 “我” 的持續時間的計算，因此 “我” 的持續時間會特別長為 0.6s
    3. 在圖二 (c) 中，一句話每個字持續時間很短或是連音的部分，會導致某些字偵測後判斷持續時間為 0
  - □ 針對上述三個問題的解決方式：
    1. 透過主觀聽覺判斷，一個字持續時間若長於 0.4 秒聽起來就會不自然，因此若持續時間超過 0.4 秒就在此字前加入語氣減弱標籤：<break strength=“weak”/>
    2. 如果 “我” 持續時間異常的久，除了加入語氣減弱標籤外可以再加停頓標籤：<break time=“0.2s”/>，並且再扣除一些 “我” 的時間加回 “好”
    3. 在第一種情況出現時，就會加入 break 標籤，透過此標籤做為判斷依據，將兩標籤中間所有時間加總取平均，再平分給每一個字



(圖一) Virtual dubber 架構



(圖二) 音檔持續時間示意圖

Google 偵測工具問題改善 example

```
<prosody duration = “0.2”>          主</prosody>
<prosody duration = "0.199999">要</prosody>
<prosody duration = "0.199999">指</prosody>
<prosody duration = “0.1”>         的</prosody>
<prosody duration = "0.199999">是</prosody>
<prosody duration = “0.9”>         靜</prosody>
<prosody duration = "0.299999">態</prosody>
<prosody duration = “0.5”>         隨</prosody>
```

第二種：加入 break weak、strength 再將一部分時間 (0.2s) 往前一個字移動

第一種：大於 0.4s 因此加入 break weak

```
<prosody duration = "0.162499">要</prosody>
<prosody duration = "0.162499">指</prosody>
<prosody duration = "0.162499">的</prosody>
<prosody duration = “0.162499 + 0.2">是</prosody>
<break time="0.1s"/>
<break strength="weak"/>
<prosody duration = "0.4">         靜</prosody>

<prosody duration = "0.299999">態</prosody>
<break strength="weak"/>
<prosody duration = "0.3">         隨</prosody>
```

```
1  <prosody duration = "0.299999">機</prosody>
2  <prosody duration = “0.2”>         存</prosody>
3  <prosody duration = “0.2”>         取</prosody>
4  <prosody duration = “0.2”>         記</prosody>
5  <prosody duration = "0.299999">憶</prosody>
6  <prosody duration = “0.0”>          體</prosody>
7  <prosody duration = "0.399999">這</prosody>
8  <prosody duration = “0.2”>         些</prosody>
9  <prosody duration = “0.2”>         電</prosody>
10 <prosody duration = "0.099999">子</prosody>
11 <prosody duration = “0.1”>         的</prosody>
12 <prosody duration = "0.299999">產</prosody>
13 <prosody duration = "0.099999">品</prosody>
```

第三種：2.599994 / 13 = 0.19999

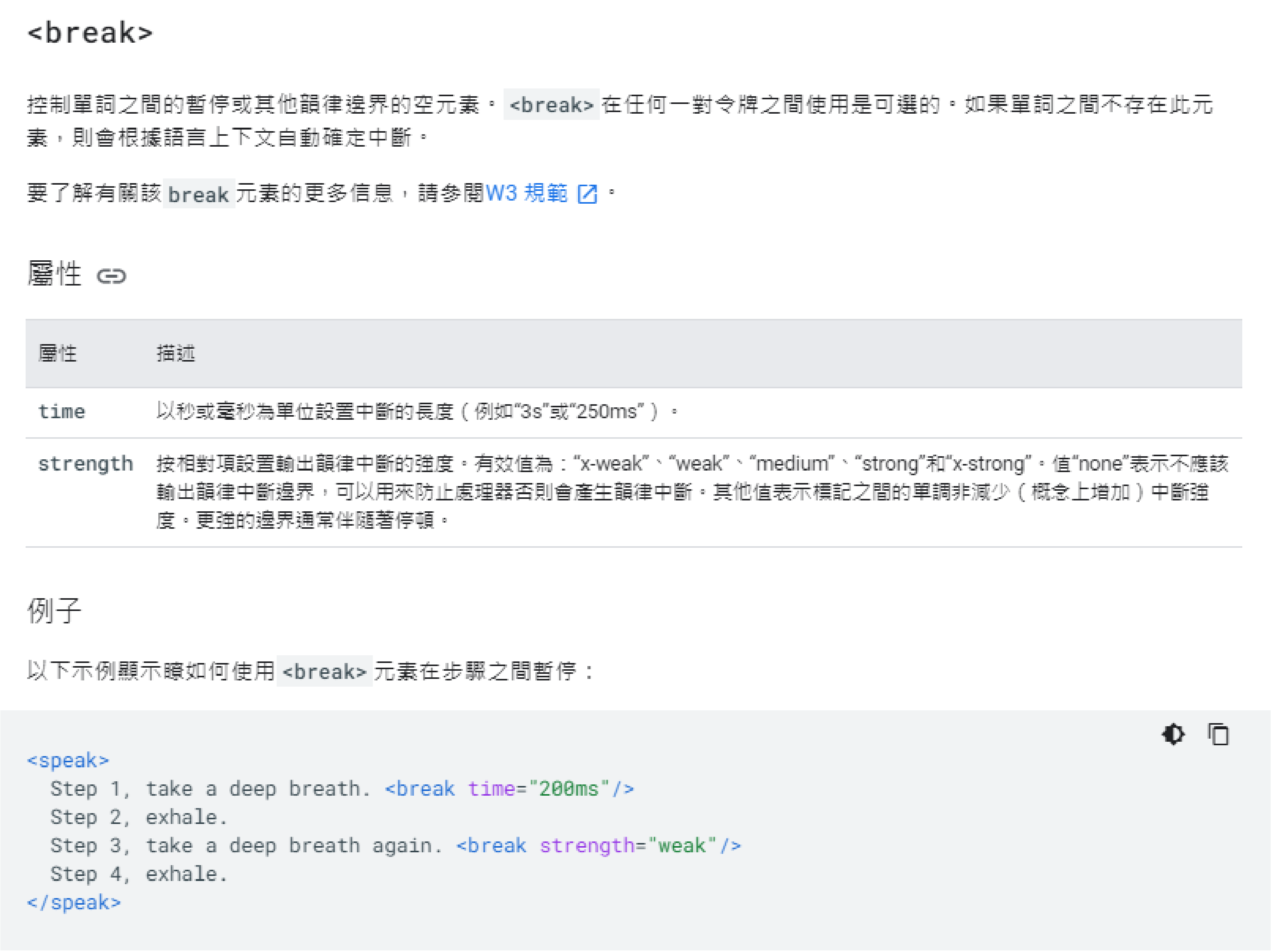
```
<prosody duration = “0.19999”>  機</prosody>
<prosody duration = "0.19999">  存</prosody>
<prosody duration = “0.19999”>  取</prosody>
<prosody duration = "0.19999">   記</prosody>
<prosody duration = “0.19999”>  憶</prosody>
<prosody duration = "0.19999">  體</prosody>
<prosody duration = “0.19999”>  這</prosody>
<prosody duration = "0.19999">  些</prosody>
<prosody duration = "0.19999">  電</prosody>
<prosody duration = “0.19999”>  子</prosody>
<prosody duration = "0.19999">  的</prosody>
<prosody duration = “0.19999”>  產</prosody>
<prosody duration = “0.19999”>  品</prosody>
```

```
<prosody duration = “0.5“>         他</prosody>
```

第一種：大於 0.4s 因此加入 break weak

```
<break strength="weak"/>
<prosody duration = “0.3”>         他</prosody>
```

附錄



<break>

控制單詞之間的暫停或其他韻律邊界的空元素。`<break>`在任何一對令牌之間使用是可選的。如果單詞之間不存在此元素，則會根據語言上下文自動確定中斷。

要了解有關該`break`元素的更多信息，請參閱W3 規範 。

屬性

屬性	描述
`time`	以秒或毫秒為單位設置中斷的長度（例如“3s”或“250ms”）。
`strength`	按相對項設置輸出韻律中斷的強度。有效值為：“x-weak”、“weak”、“medium”、“strong”和“x-strong”。值“none”表示不應該輸出韻律中斷邊界，可以用來防止處理器否則會產生韻律中斷。其他值表示標記之間的單調非減少（概念上增加）中斷強度。更強的邊界通常伴隨著停頓。

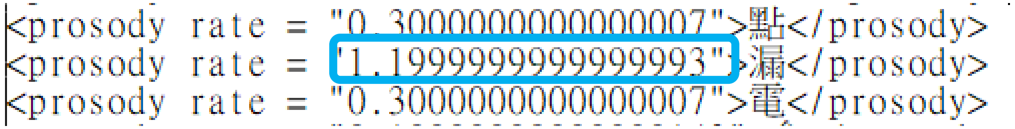
例子

以下示例顯示瞭如何使用`<break>`元素在步驟之間暫停：

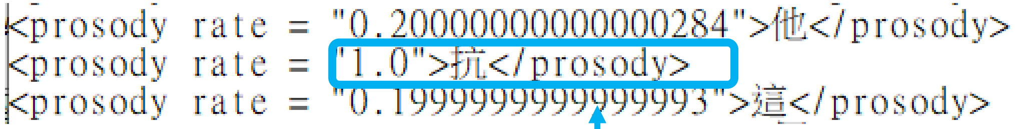
```
<speak>
  Step 1, take a deep breath. <break time="200ms"/>
  Step 2, exhale.
  Step 3, take a deep breath again. <break strength="weak"/>
  Step 4, exhale.
</speak>
```

感覺這個 0.5 秒也有點久，聽起來有點不太自然，
另外需要計算一下語氣強弱的 ssml 標籤，時間長度是多少

小於 0.1 秒的時間太短，聽起來也很不自然，這個也要處理一下像是做平均，至少每個字聽起來是均速，才不會一下 0.5 秒，然後接下來又念超快

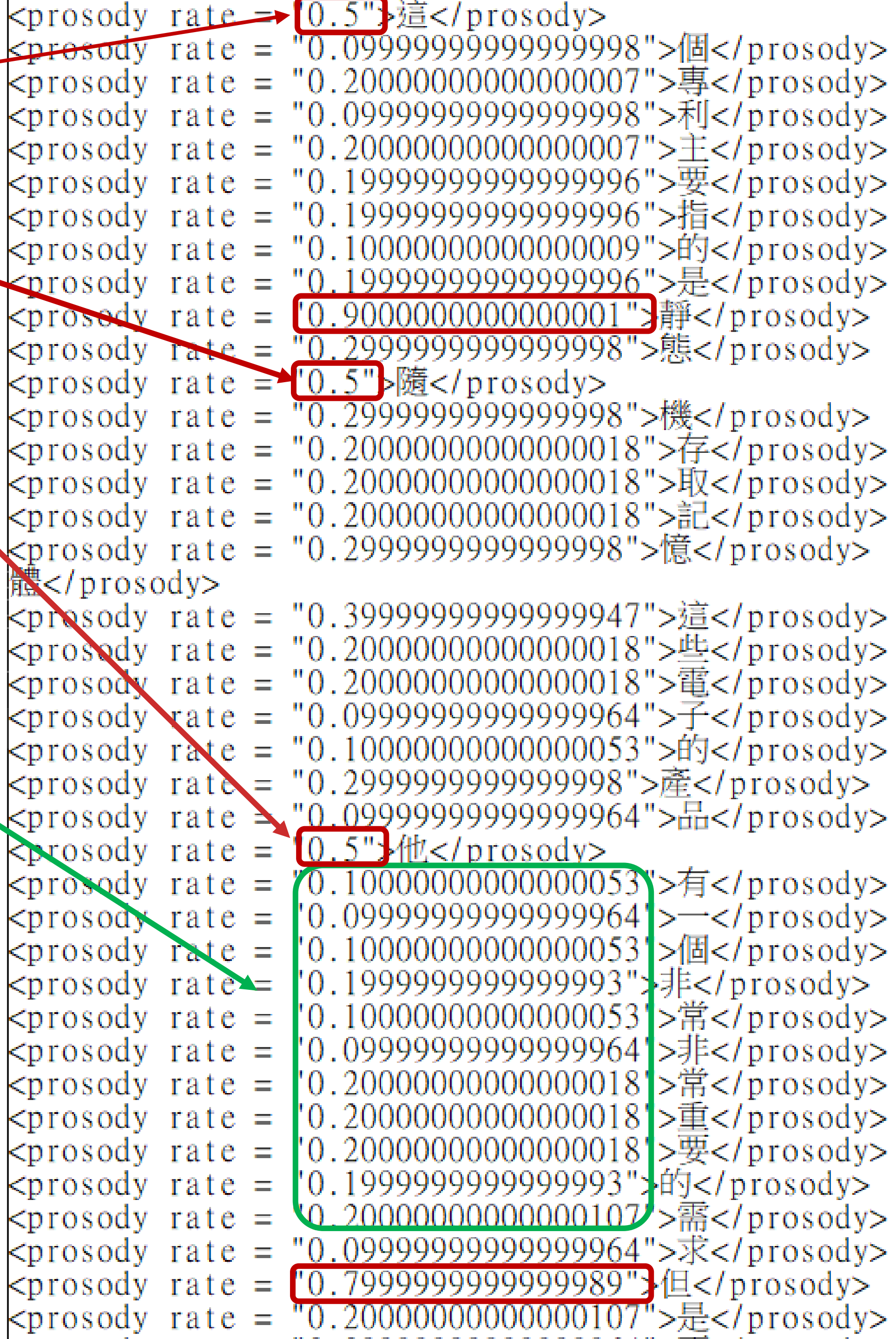


```
<prosody rate = "0.30000000000000007">點</prosody>
<prosody rate = "1.1999999999999993">漏</prosody>
<prosody rate = "0.30000000000000007">電</prosody>
```



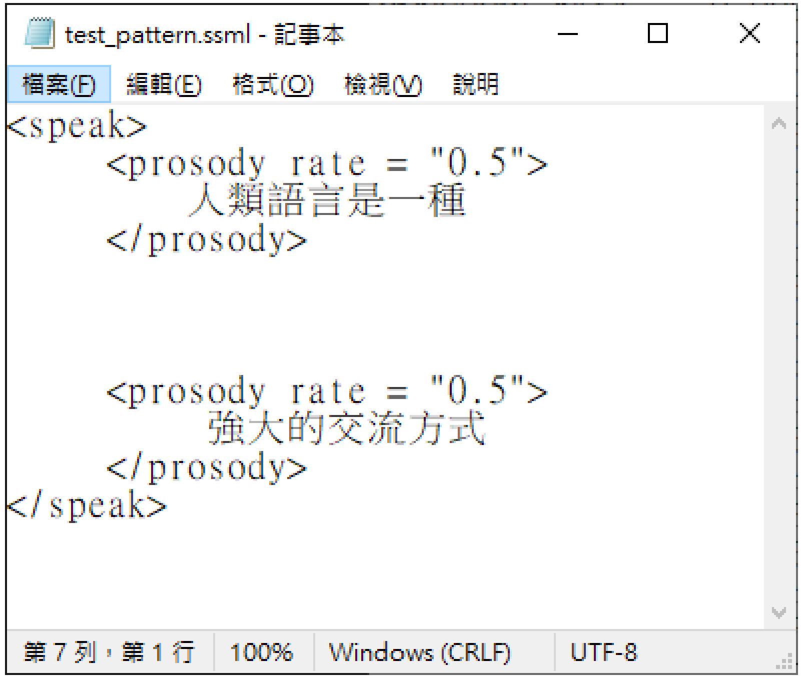
```
<prosody rate = "0.20000000000000284">他</prosody>
<prosody rate = "1.0">抗</prosody>
<prosody rate = "0.1999999999999993">這</prosody>
```

時間長達1秒，聽起來超怪



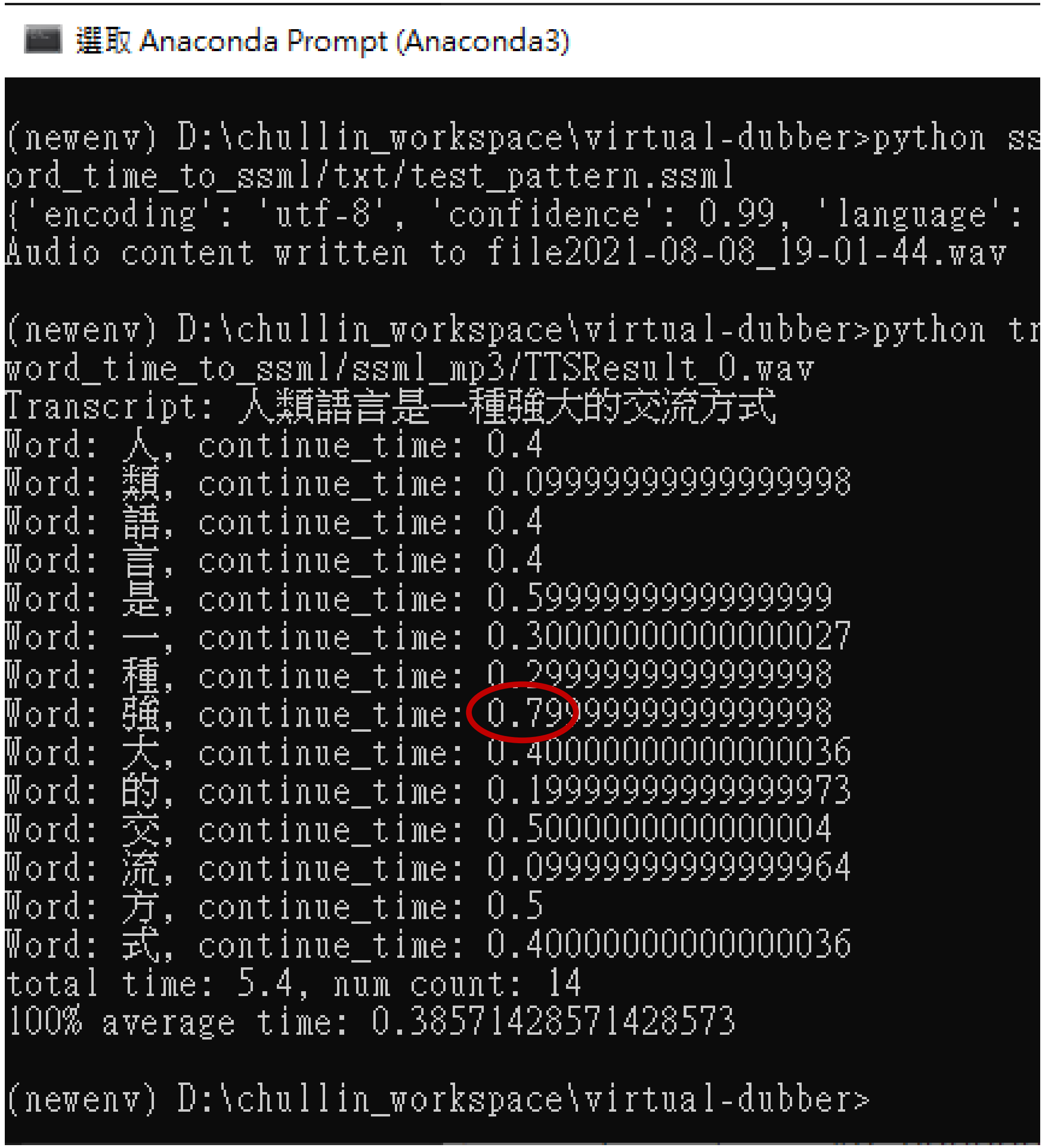
```
<prosody rate = "0.5">這</prosody>
<prosody rate = "0.09999999999999998">個</prosody>
<prosody rate = "0.20000000000000007">專</prosody>
<prosody rate = "0.09999999999999998">利</prosody>
<prosody rate = "0.20000000000000007">主</prosody>
<prosody rate = "0.19999999999999996">要</prosody>
<prosody rate = "0.19999999999999996">指</prosody>
<prosody rate = "0.10000000000000009">的</prosody>
<prosody rate = "0.19999999999999996">是</prosody>
<prosody rate = "0.9000000000000001">靜</prosody>
<prosody rate = "0.2999999999999998">態</prosody>
<prosody rate = "0.5">隨</prosody>
<prosody rate = "0.2999999999999998">機</prosody>
<prosody rate = "0.20000000000000018">存</prosody>
<prosody rate = "0.20000000000000018">取</prosody>
<prosody rate = "0.20000000000000018">記</prosody>
<prosody rate = "0.2999999999999998">憶</prosody>
體</prosody>
<prosody rate = "0.39999999999999947">這</prosody>
<prosody rate = "0.20000000000000018">些</prosody>
<prosody rate = "0.20000000000000018">電</prosody>
<prosody rate = "0.09999999999999964">子</prosody>
<prosody rate = "0.10000000000000053">的</prosody>
<prosody rate = "0.2999999999999998">產</prosody>
<prosody rate = "0.09999999999999964">品</prosody>
<prosody rate = "0.5">他</prosody>
<prosody rate = "0.10000000000000053">有</prosody>
<prosody rate = "0.09999999999999964">一</prosody>
<prosody rate = "0.10000000000000053">個</prosody>
<prosody rate = "0.1999999999999993">非</prosody>
<prosody rate = "0.10000000000000053">常</prosody>
<prosody rate = "0.09999999999999964">非</prosody>
<prosody rate = "0.20000000000000018">常</prosody>
<prosody rate = "0.20000000000000018">重</prosody>
<prosody rate = "0.20000000000000018">要</prosody>
<prosody rate = "0.1999999999999993">的</prosody>
<prosody rate = "0.20000000000000107">需</prosody>
<prosody rate = "0.09999999999999964">求</prosody>
<prosody rate = "0.7999999999999989">但</prosody>
<prosody rate = "0.20000000000000107">是</prosody>
```

<prosody rate = “0.5">
原始



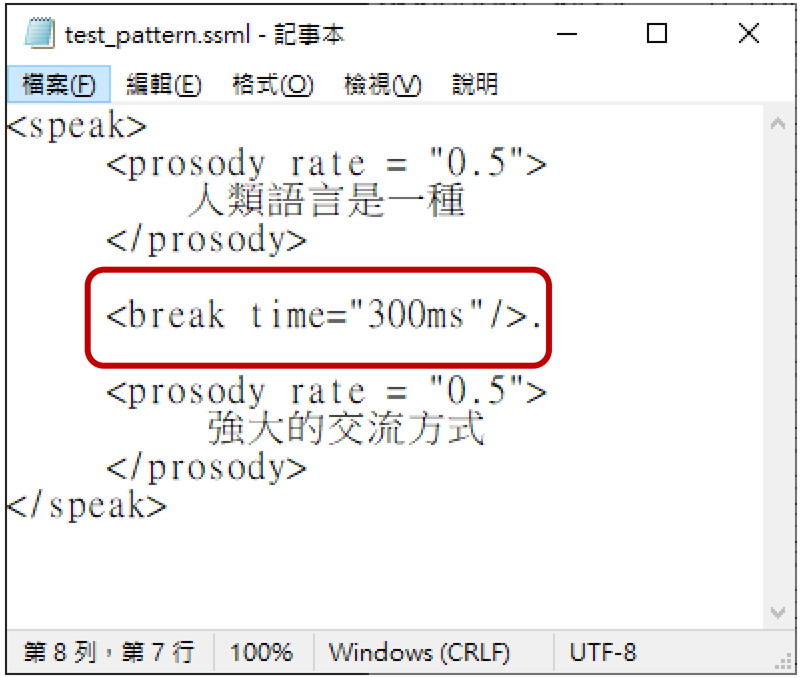
```
<speak>
     <prosody rate = "0.5">
         人類語言是一種
     </prosody>


     <prosody rate = "0.5">
          強大的交流方式
     </prosody>
</speak>
```

接著插入 <break time="300ms"/>

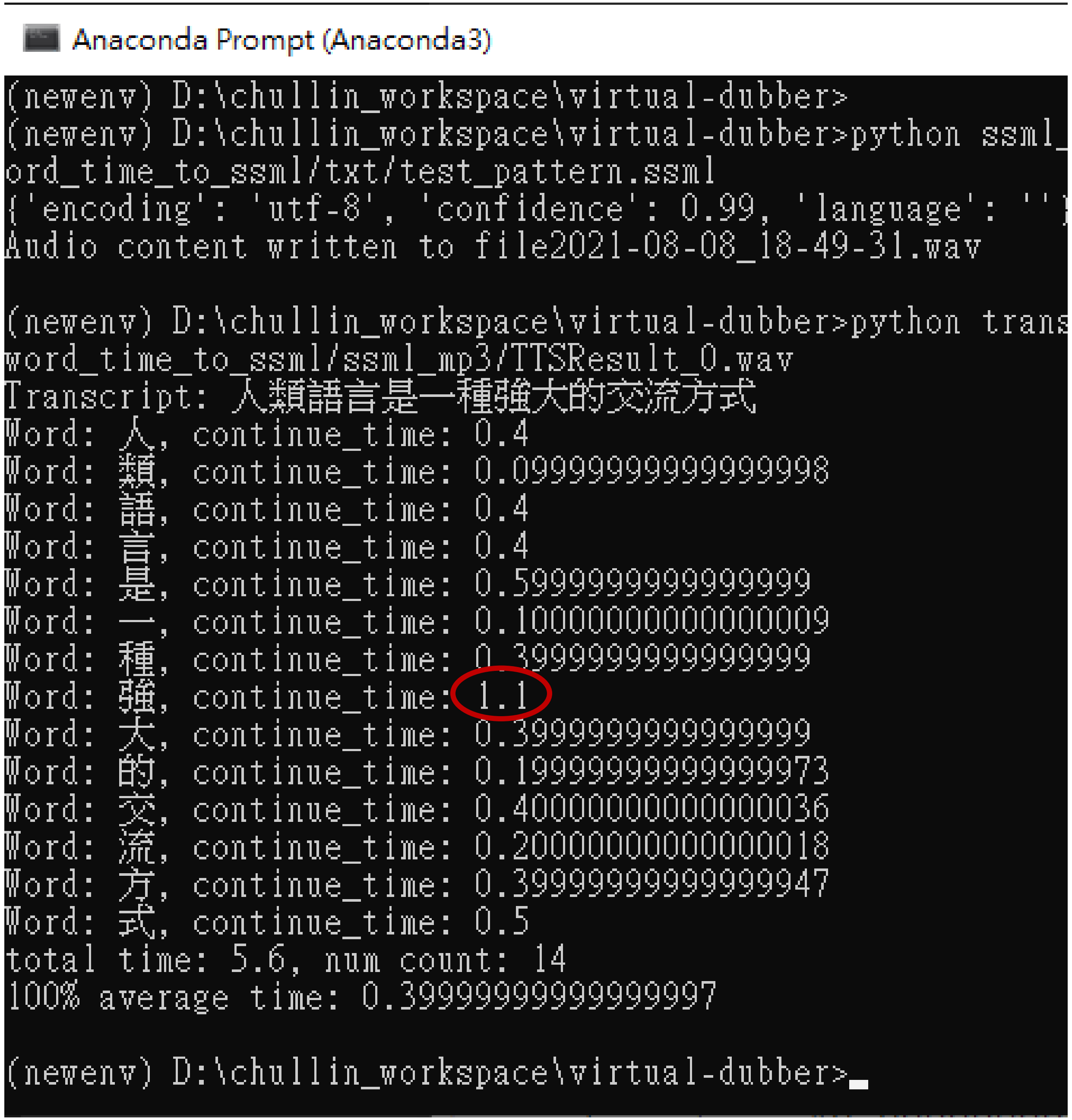


```
(newenv) D:\chullin_workspace\virtual-dubber>python ss
ord_time_to_ssml/txt/test_pattern.ssml
{'encoding': 'utf-8', 'confidence': 0.99, 'language':
Audio content written to file2021-08-08_19-01-44.wav

(newenv) D:\chullin_workspace\virtual-dubber>python tr
word_time_to_ssml/ssml_mp3/TTSResult_0.wav
Transcript: 人類語言是一種強大的交流方式
Word: 人, continue_time: 0.4
Word: 類, continue_time: 0.09999999999999998
Word: 語, continue_time: 0.4
Word: 言, continue_time: 0.4
Word: 是, continue_time: 0.5999999999999999
Word: 一, continue_time: 0.30000000000000027
Word: 種, continue_time: 0.2999999999999998
Word: 強, continue_time: 0.7999999999999998
Word: 大, continue_time: 0.40000000000000036
Word: 的, continue_time: 0.19999999999999973
Word: 交, continue_time: 0.5000000000000004
Word: 流, continue_time: 0.09999999999999964
Word: 方, continue_time: 0.5
Word: 式, continue_time: 0.40000000000000036
total time: 5.4, num count: 14
100% average time: 0.38571428571428573

(newenv) D:\chullin_workspace\virtual-dubber>
```

<prosody rate = “0.5">


test_pattern.ssml - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明
第 8 列，第 7 行
100%
Windows (CRLF)
UTF-8

```
<speak>
    <prosody rate = "0.5">
        人類語言是一種
    </prosody>

    <break time="300ms"/>.

    <prosody rate = "0.5">
         強大的交流方式
    </prosody>
</speak>
```

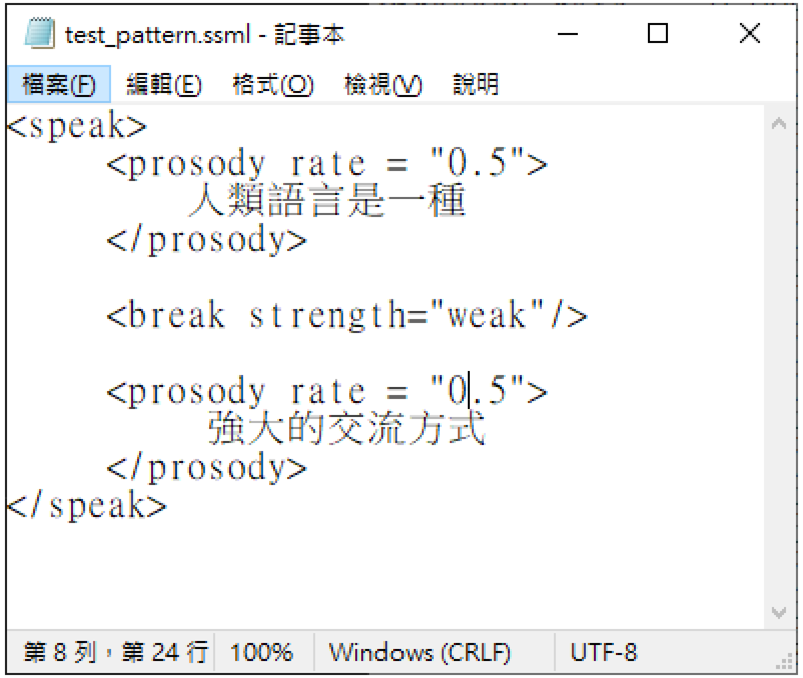


```
(newenv) D:\chullin_workspace\virtual-dubber>
(newenv) D:\chullin_workspace\virtual-dubber>python ssml_
ord_time_to_ssml/txt/test_pattern.ssml
{'encoding': 'utf-8', 'confidence': 0.99, 'language': ''}
Audio content written to file2021-08-08_18-49-31.wav

(newenv) D:\chullin_workspace\virtual-dubber>python trans
word_time_to_ssml/ssml_mp3/TTSResult_0.wav
Transcript: 人類語言是一種強大的交流方式
Word: 人, continue_time: 0.4
Word: 類, continue_time: 0.09999999999999998
Word: 語, continue_time: 0.4
Word: 言, continue_time: 0.4
Word: 是, continue_time: 0.5999999999999999
Word: 一, continue_time: 0.10000000000000009
Word: 種, continue_time: 0.3999999999999999
Word: 強, continue_time: 1.1
Word: 大, continue_time: 0.3999999999999999
Word: 的, continue_time: 0.19999999999999973
Word: 交, continue_time: 0.40000000000000036
Word: 流, continue_time: 0.20000000000000018
Word: 方, continue_time: 0.39999999999999947
Word: 式, continue_time: 0.5
total time: 5.6, num count: 14
100% average time: 0.39999999999999997

(newenv) D:\chullin_workspace\virtual-dubber>_
```

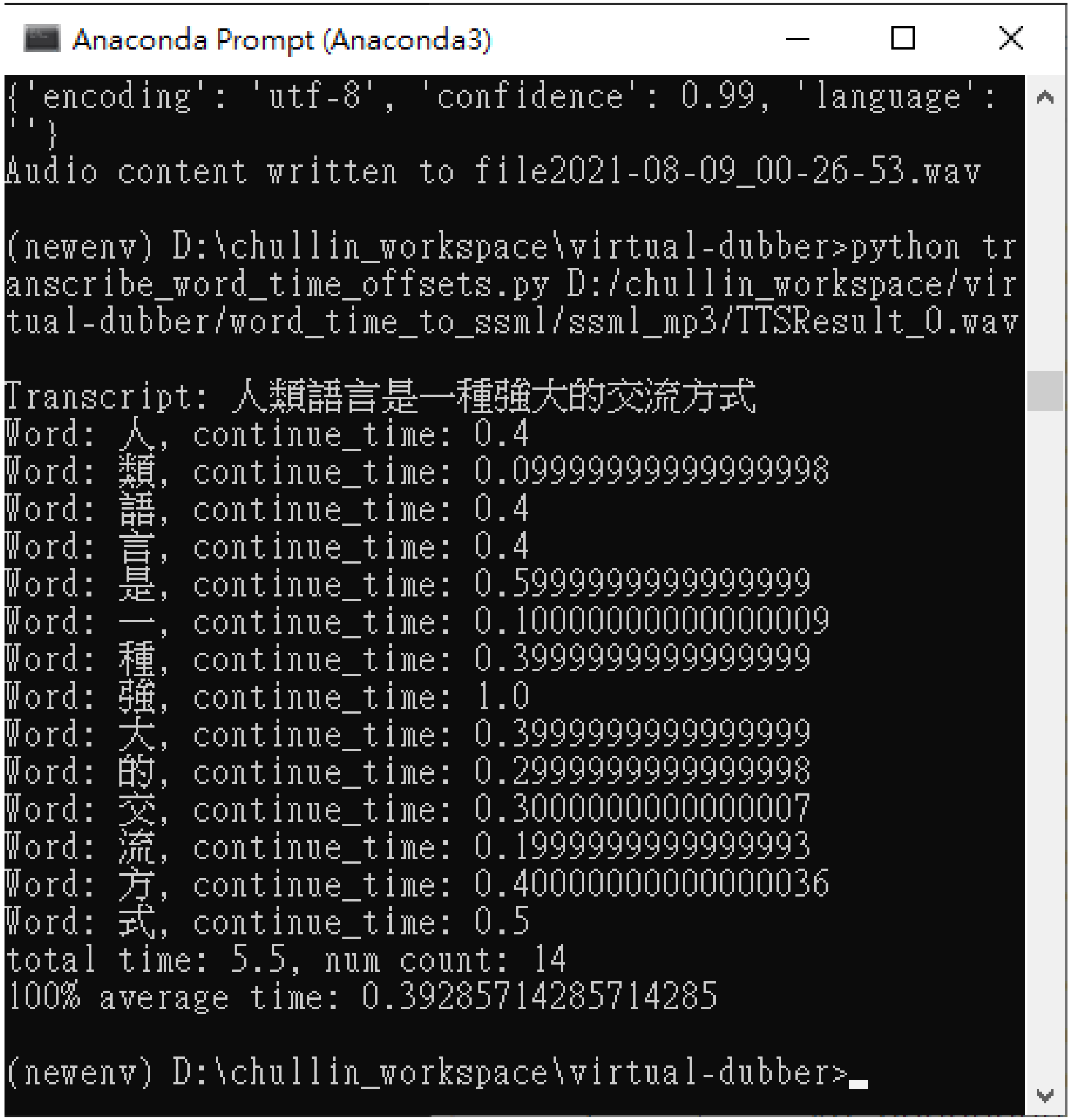
<break time="300ms"/>
就差這個 0.3 秒

0.8加 0.3 秒
變成1.1

<prosody rate = “0.5">


test_pattern.ssml - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明
第 8 列，第 24 行 100% Windows (CRLF) UTF-8

```
<speak>
    <prosody rate = "0.5">
        人類語言是一種
    </prosody>

    <break strength="weak"/>

    <prosody rate = "0.5">
         強大的交流方式
    </prosody>
</speak>
```



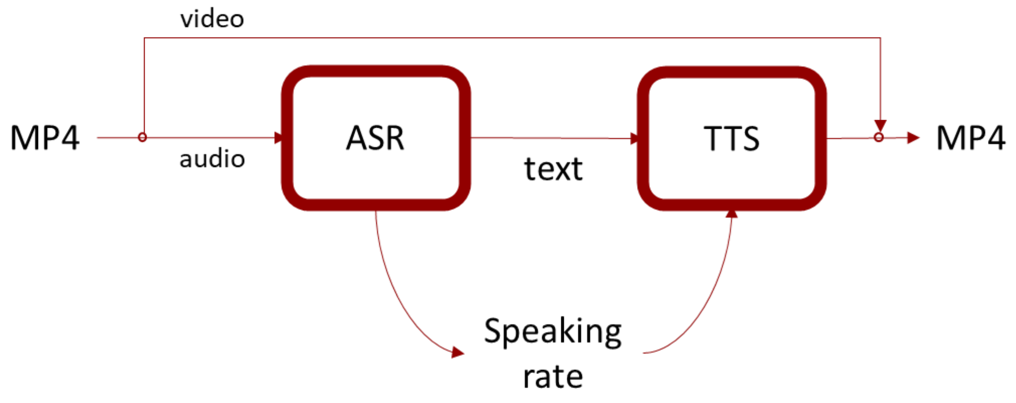
```
{'encoding': 'utf-8', 'confidence': 0.99, 'language':
''}
Audio content written to file2021-08-09_00-26-53.wav

(newenv) D:\chullin_workspace\virtual-dubber>python tr
anscribe_word_time_offsets.py D:/chullin_workspace/vir
tual-dubber/word_time_to_ssml/ssml_mp3/TTSResult_0.wav

Transcript: 人類語言是一種強大的交流方式
Word: 人, continue_time: 0.4
Word: 類, continue_time: 0.09999999999999998
Word: 語, continue_time: 0.4
Word: 言, continue_time: 0.4
Word: 是, continue_time: 0.5999999999999999
Word: 一, continue_time: 0.10000000000000009
Word: 種, continue_time: 0.3999999999999999
Word: 強, continue_time: 1.0
Word: 大, continue_time: 0.3999999999999999
Word: 的, continue_time: 0.2999999999999998
Word: 交, continue_time: 0.3000000000000007
Word: 流, continue_time: 0.1999999999999993
Word: 方, continue_time: 0.40000000000000036
Word: 式, continue_time: 0.5
total time: 5.5, num count: 14
100% average time: 0.39285714285714285

(newenv) D:\chullin_workspace\virtual-dubber>
```

原始大約為 0.8 秒
加了減弱語氣，時間被拖長了 0.2 秒
所以 0.8 秒 + 0.2 秒 = 1.0秒

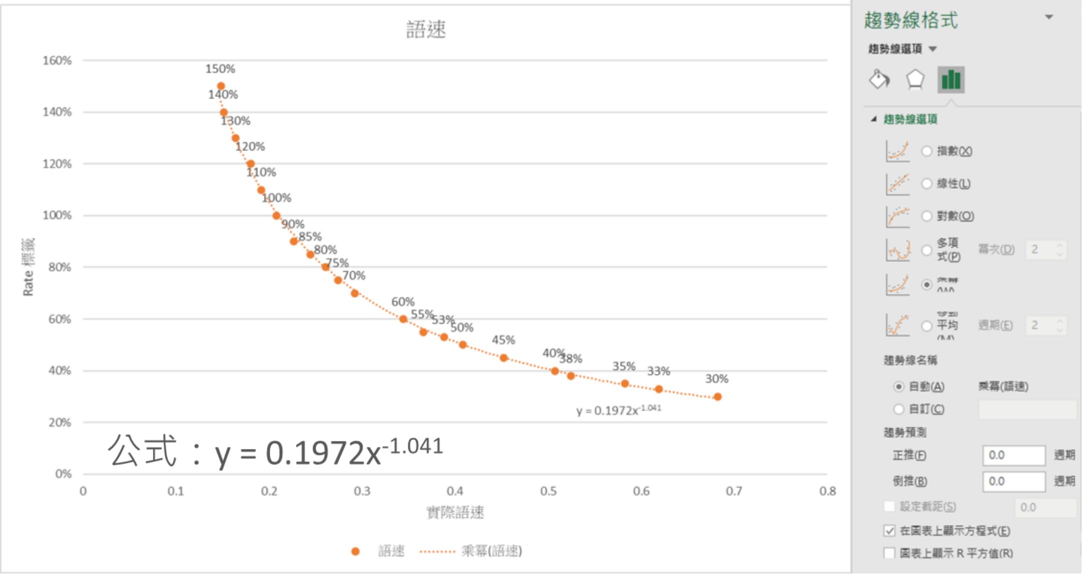
Outline

虛擬配音員

- ## Action item
  - □ Speaking rate control
- ## Demo link
  - □ Test sequence 王進賢教授：https://youtu.be/PR23ZwADHeQ
- ## Speaking rate control method
  1. 利用 TTS 提供的百分比語速 (e.g., 90%, 100%, 110%) 合成語音用於模擬實際音檔
  2. 利用 ASR 提供 “計算每個字聲音持續的時間” 的功能計算 TTS 合成出來的語音速度
  3. 將合成語音語速與提供的百分比製成表格
  4. 利用 excel 將表格內容繪製成散佈圖
  5. 選擇乘冪趨勢線並算出曲線方程式
     - 公式：$y = 0.1972x^{-1.041}$
  6. 利用此公式將文字標上速度標籤



(圖一) Virtual dubber 架構



公式：$y = 0.1972x^{-1.041}$

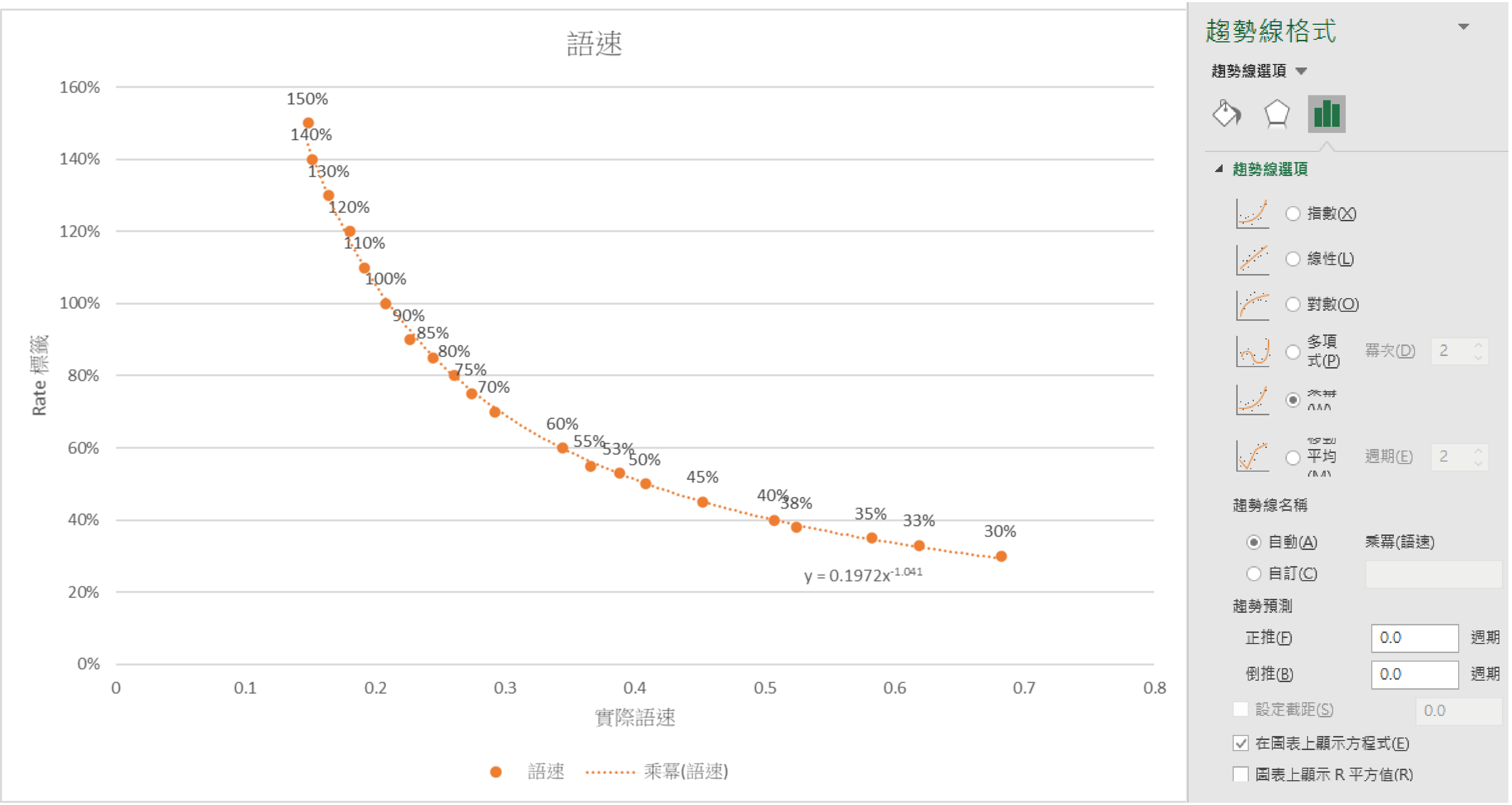
○ 指數(X)
○ 線性(L)
○ 對數(O)
○ 多項式(P) 冪次(D) 2
● 乘冪(W)
○ 移動平均(M) 週期(E) 2

趨勢線名稱
● 自動(A) 乘冪(語速)
○ 自訂(C)

趨勢預測
正推(F) 0.0 週期
倒推(B) 0.0 週期

□ 設定截距(S) 0.0
☑ 在圖表上顯示方程式(E)
□ 圖表上顯示 R 平方值(R)

(圖二) 實際語速與 TTS 提供語速的曲線



趨勢線格式

趨勢線選項

◢ 趨勢線選項

○ 指數(X)

○ 線性(L)

○ 對數(O)

○ 多項式(P) 冪次(D) 2

● 乘冪

○ 移動平均 週期(E) 2

趨勢線名稱

● 自動(A) 乘冪(語速)

○ 自訂(C)

趨勢預測

正推(F) 0.0 週期

倒推(B) 0.0 週期

□ 設定截距(S) 0.0

☑ 在圖表上顯示方程式(E)

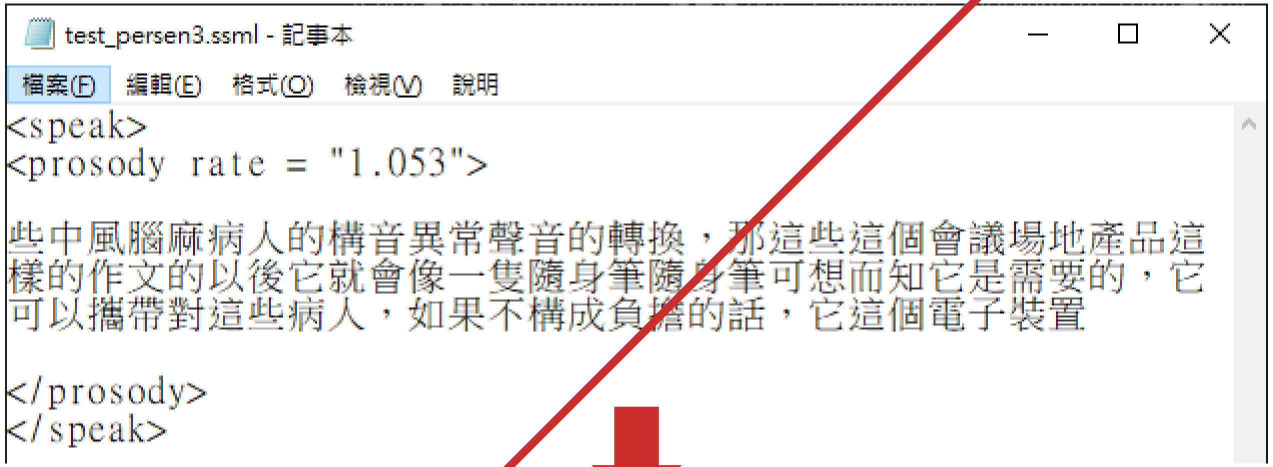
□ 圖表上顯示 R 平方值(R)

公式：$y = 0.1972x^{-1.041}$

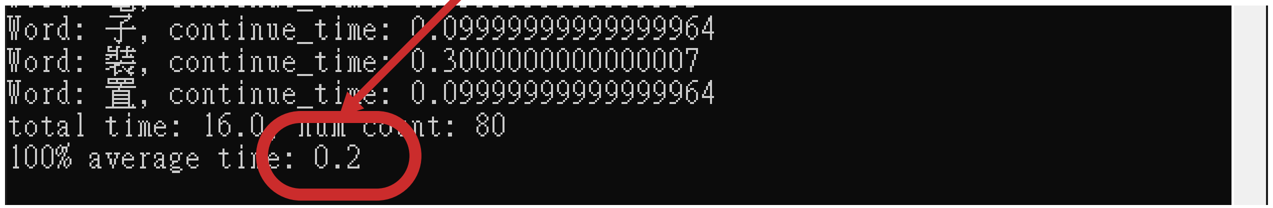
```
print ("0.1972 * math.pow(x, -1.041) : ", 0.1972 * math.pow(0.19999999, -1.041))
```

公式：$y = 0.1972x^{-1.041}$

```
0.1972 * math.pow(x, -1.041) :  1.0532578574884441
```


test_persen3.ssml - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

```
<speak>
<prosody rate = "1.053">

些中風腦麻病人的構音異常聲音的轉換，那這些這個會議場地產品這
樣的作文的以後它就會像一隻隨身筆隨身筆可想而知它是需要的，它
可以攜帶對這些病人，如果不構成負擔的話，它這個電子裝置

</prosody>
</speak>
```



```
Word: 子, continue_time: 0.09999999999999964
Word: 裝, continue_time: 0.3000000000000007
Word: 置, continue_time: 0.09999999999999964
total time: 16.0, num count: 80
100% average time: 0.2
```

得證，此公式合理且準確

(已完成) 30sec_In_addtag_optimize_syn.py