改善孩童語音辨識之策略
Strategies to Improve Children's Speech Recognition

組別：A22　組員姓名：鄭詠、李丞恩　指導教授：劉奕汶

- 以深度學習進行孩童語音辨識遭遇的困難：
  1. 缺乏足夠資料訓練模型。
  2. 孩童常有不正確發音。
  3. 孩童語言架構常有不合邏輯或顛倒語法的情況。
  4. 孩童語音有較大的音高變化

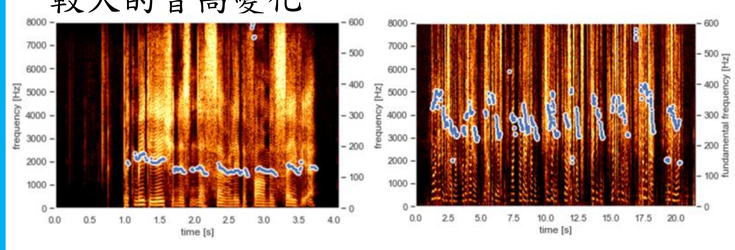
我們實作語音辨識的軟體為Kaldi。我們使用PF-star作為孩童語料庫。PF-star語料庫蒐集了158位年齡介於4到14歲的英國孩童語音，每位孩童約有8-10個utterance（音檔），training data由92位說話者錄製而成。我們使用WSJCAM0作為成人語料庫。WSJCAM0語料庫的語者為不分年齡、男女皆有的英國成人，每位成人約有90個utterance。

Word Error Rate（WER）用以判斷語音辨識系統表現，其定義為辨識正確的單字佔所有單字的百分比。常見的錯誤如下圖所示。我們訓練的Baseline系統對孩童的WER為16.36%；成人的WER為3.93%。

substitution	TWO	TO	33
substitution	THE	A	22
substitution	RED	READ	16
substitution	GREY	GRAY	14
substitution	WEAR	WHERE	13
substitution	YOU'RE	YOUR	13
substitution	NOSE	KNOWS	12
substitution	THE	FOR	12
substitution	MUM	MOM	11
substitution	THREE	FREE	11

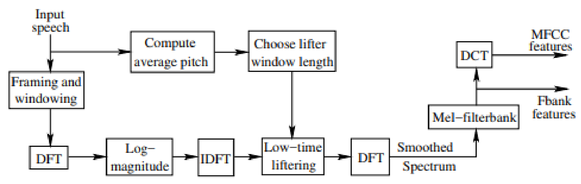
I. Pitch-adaptation

從時頻圖（Spectrogram）來看，孩童有較大的音高變化

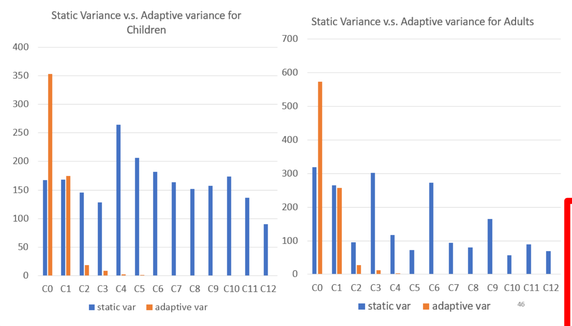


▲成人　▲孩童

因此，我們嘗試改變梅爾頻率倒譜系數（MFCC）的演算法如下[1]



但高頻的MFCC被消除後（如下圖），Kaldi難以推斷每個frame是否包含人聲。也因此，Kaldi無法進行alignment，導致這個方法失敗。

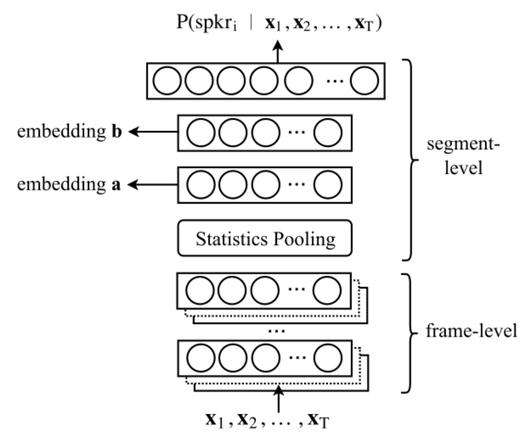


Oops!

II. Speaker Embedding

我們使用的Speaker Embedding有i-vector[2]與xvector[3]兩種。i-vector是一種speaker embedding，將每個frame中語者的MFCC隱藏的資訊，比如：音高、年齡、性別、情緒等額外特徵，利用統計的手法壓縮成一個向量。

Xvector是神經網路的中間產物，如下圖[3]。計算xvector的神經網路本質上是一個分類器。Pooling layer之後第一層的輸出向量即為所需的embeddings



我們將xvector和i-vector串接形成一個228維的speaker embedding，使它的功能更為強大。接著，再把這個embedding與MFCC串接，作為神經網路的輸入

Success!

III. Language model interpolation

當我們研究WER的成分時，發現許多辨識錯誤的單字具有相似的發音。這是語言模型不夠完善所造成的。我們採用IRSTLM[4]這套軟體並用孩童的文本來產生語言模型。

```
\4-grams:
-0.07493048     WILL ALLOW A RARE
-0.6221532      I AM A BLACKBIRD
-0.6203443      I AM A WHITE
-1.124449       I AM A YOUNG
-0.03970881     YOU ARE A CHILD
-0.5368701      <s> AS A ROMANTIC
-0.8224477      <s> AS A TREAT
-0.1289204      FAIL AS A ROMANTIC
-0.08698717     FLOSS AS A TREAT
-0.1495632      <s> ATE A LOT
-0.06837481     BUTTERFLY ATE A LOT
```

▲IRSTLM中關於4-gram的統計資料

接著將baseline的語言模型$\widehat{P_2}$，和孩童文本建立的語言模型$\widehat{P_1}$以插值法（interpolation）混和。

$$\hat{P}(\omega_k|\omega_1 \cdots \omega_{k-1}) = \lambda\widehat{P_1}(\omega_k|\omega_1 \cdots \omega_{k-1}) + (1-\lambda)\widehat{P_2}(\omega_k|\omega_1 \cdots \omega_{k-1})$$

當混合的語言模型對孩童dev-set文本測式，在所得到的機率最高時，λ為0.991425。

Success!

	Baseline	Pitch Adaptation	Speaker embedding	Language model interpolation
adult	3.61%	X	3.45%	7.03%
children	16.36%	X	15.72%	8.97%

結論

1. 在pitch adaptation的方法中，由於某些phone在高頻率上才有存在的痕跡。因此higher pitch被挪去可能導致本方法失敗。
2. Speaker embedding可以顯露出MFCC所隱藏的資訊，特徵較為多樣。聲學模型因為有多樣的選擇，更能做到模型泛化的能力。
3. 語言模型的interpolation使效果大幅進步。雖然speaker dependent的語言模型表現較佳，但成人的語音辨識模型也會因此犧牲。

參考資料

[1] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. INTERSPEECH*, 2016, pp. 3459–3463.
[2] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE Speech Recognition and Understanding Workshop*, pp. 55–59, 2013.
[3] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.
[4] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models," in *Interspeech*, 2008, pp. 1618–1621.