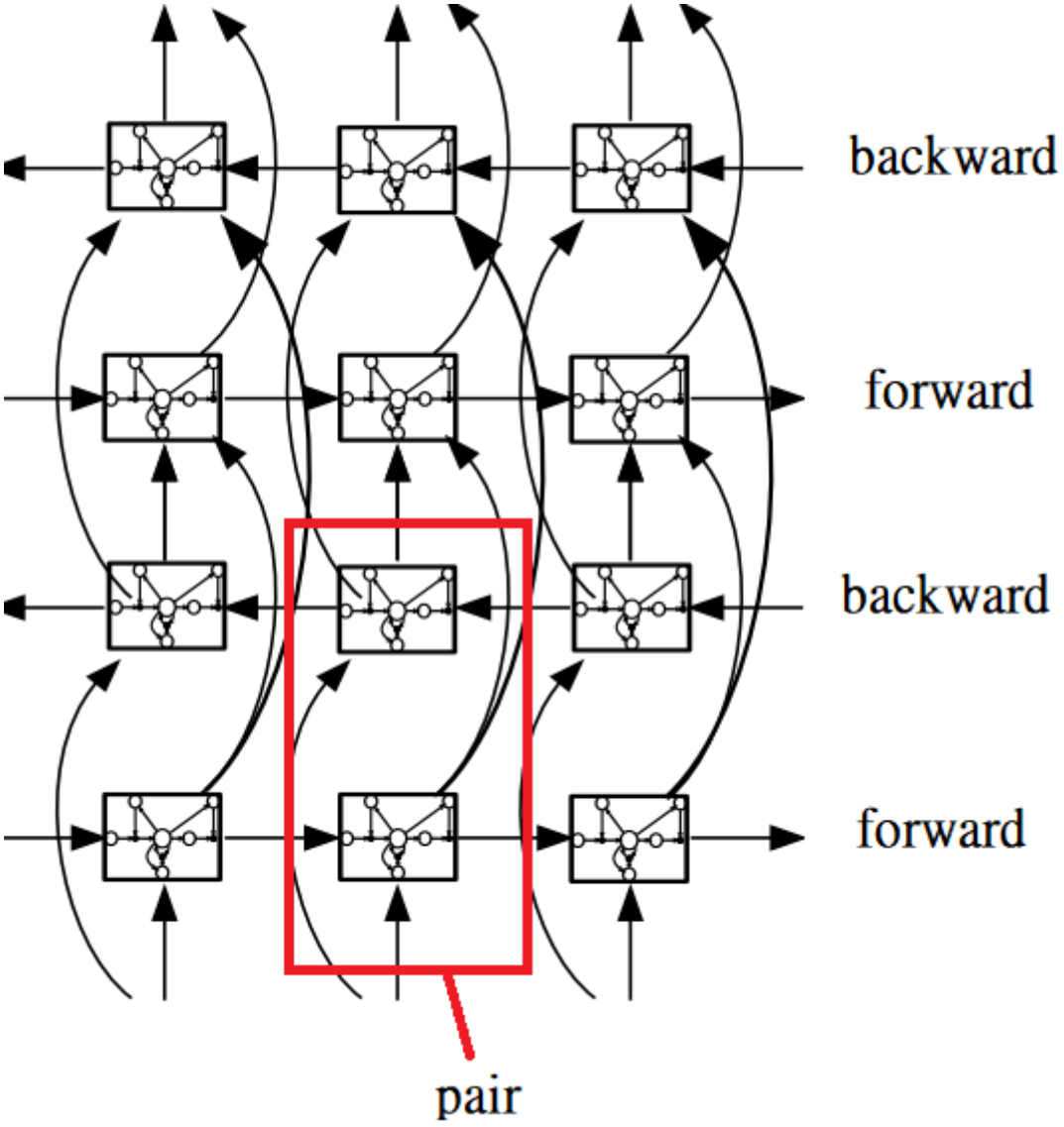
DBLSTM在數位語音上的應用

B07901169 電機三 楊宗桓 B07902042 資工三 葉璟諄

摘要:

1. Deep Bidirectional LSTM



LSTM

LSTM是具有**記憶功能**的cell，可以用LSTM取代NN中的node，讓NN訓練時會受前面資訊的影響，應用在語音辨識上可以利用過去的音訊和現在的input來判定現在的output

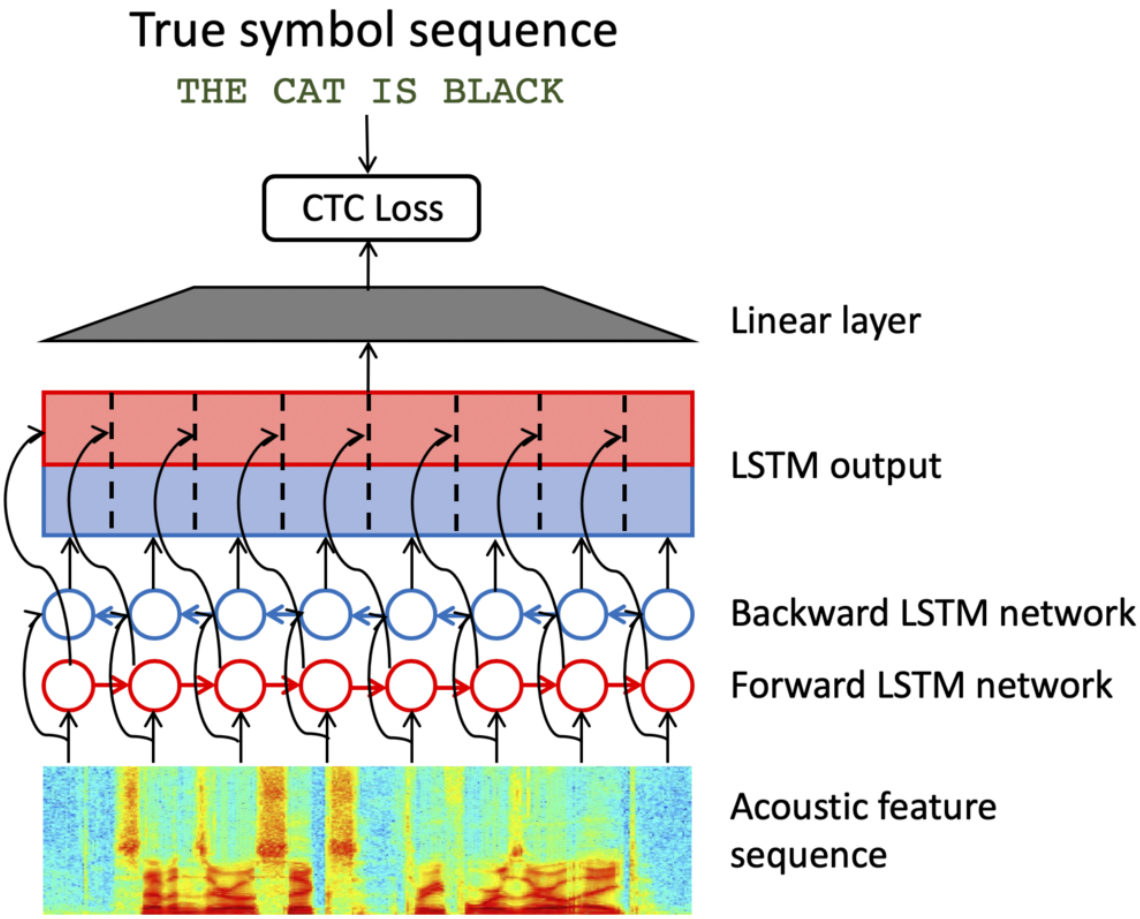
DBLSTM

而LSTM的優化版本叫做DBLSTM，把語音input從頭往後，從尾往前分別擷取出來，對於每一層layer，單一時間的LSTM變為1個pair(2倍)，一個負責forward，一個負責backward，並且把同時間的LSTM pair一起傳入下一層layer，這樣可以**用上下文來幫助語音辨識**，有更充分的資訊會比單向的LSTM好

2. End-to-end automatic speech recognition

簡介

直接用音訊作為輸入，經過DNN後直接輸出成文字，訓練流程如下圖



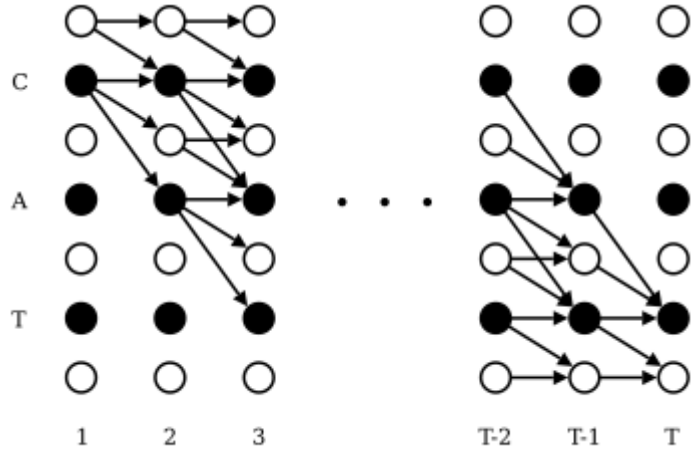
1.把語音訊號encode成feature sequence
2.以feature作為輸入，經過DBLSTM輸出
3.經過decoder，和正確文本做CTC loss，如此可以用back propagation做梯度下降

- 優點
  1.End-to-end 直接從聲音輸出成文字，不需要language model
  2.不需要對input sequence與label sequence 做 forced allignment(註1)
- 缺點
  1.無法使用以HMM為基礎的辨識系統，因此training data少很多，比較適合用在特定領域的語音辨識(可能出現的詞比較侷限)
  2.需要把data做label，成本可能很高

End-to-end training mehods:

(1) Connectionist Temporal Classification

CTC運作方式是把輸入decode成字母，並且在語音停頓處用blank填充，用符號Φ表示，把Φ之間相鄰的重複字合併成一個，如此一來疊字的部分就不會有問題，最後把Φ刪除就會得到一個輸出，好處是對於各種不同的輸入，在**順序正確的情況下，可以輸出同樣的結果**，就不需要先做**forced allignment**



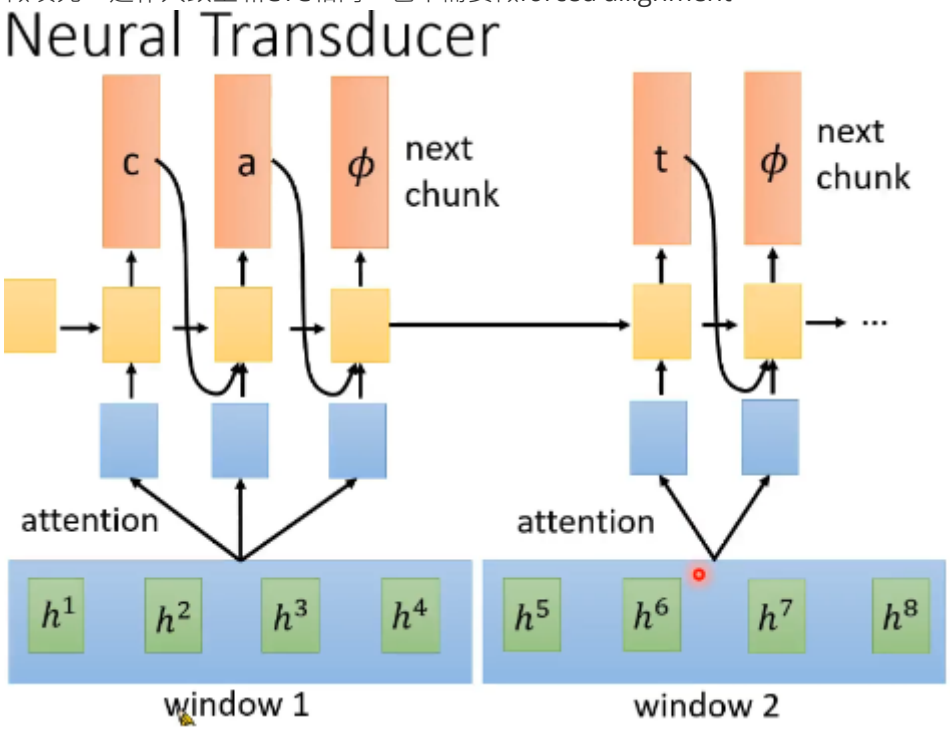
上圖可能有
ΦccΦΦaaat 或 ΦccΦΦaaaΦtt 等等路徑，但結果都會是cat
一般CTC會用來當DNN的loss function，例如判斷DNN輸出的詞和groundtruth有多少edit distance，DNN-CTC是**對於sequence**(不須對齊)去minimize loss ，和傳統的HMM**對於frame**(須對齊)去minimize loss不同
訓練流程:
1.把語音訊號input **X**={x1,...,xT} encode成feature sequence **F**={f1,...,fT}
2.把 **F**={f1,...,fT}用softmax轉成機率分布 **Y**={y1,...,yT}
3.給定**X**的條件下，任意文本**π**的條件機率為

$$p(\pi|X) = \Pi_{t=1}^{T} y_t^{\pi t}$$

4.以CTC做為loss做back propagation

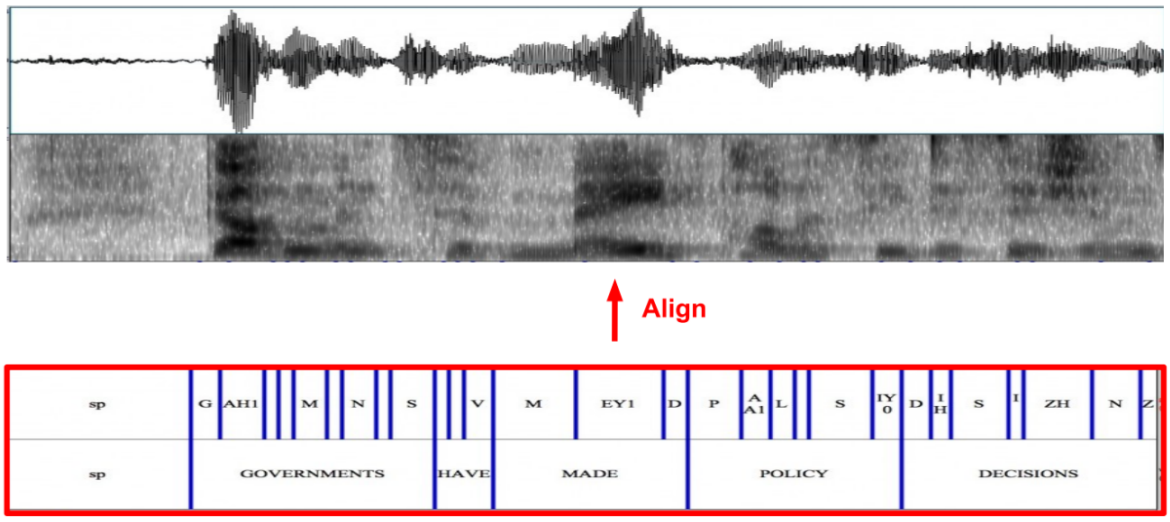
(2) Sequence Transduction:

一般和RNN連用，和CTC不同處是可以由**多個input綁成window**，**輸出多個output**，一樣有使用Φ去做填充，運作大致上和CTC相同，也不需要做forced allignment



註1: Forced allignment

將音訊與文本對齊，並在音訊中抓出各個phoneme的時間點，一般使用是先抽出MFCC features，並且結合HMM中使用過的viterbi algorithm找出可能性最大的切點，結果如下圖。

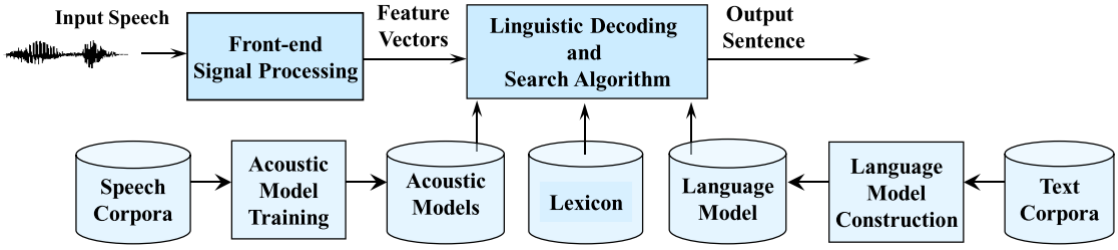


在Neural network訓練時，此技術可以直接使用 cross-entropy 做為loss function，比較直觀，但成本較高。

3. DBLSTM Hybrid with HMM

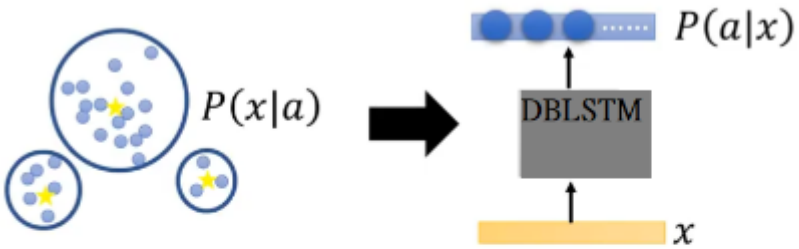
為何需要DBLSTM Hybrid model?

CTC， ST等End-to-End的model當初設計的優點在於**不用forced allignment** 事先Pre-segment訓練的data，就可以將input sequence和target sequence輸入model，訓練出implicit的語言模型，但這種模型有一個缺點是: 它沒辦法和大型字彙的語音辨識系統做結合，因為大型字彙的語音辨識系統是建立在HMM-GMM model上(也就是有經過**forced allignment**)，所以因此影響了end-to-end在大型字彙語音辨識系統下的表現。有一個統合大型字彙語音辨識系統的方法是用LSTM(或RNN)結合HMM-GMM的model得到一個新的hybrid model。



上圖為以HMM-GMM為基礎建立的語音辨識系統

如何把DBLSTM與HMM混用



$$P(x|a) = \frac{P(x,a)}{P(a)} = \frac{P(a|x)P(x)}{P(a)}$$

DBLSTM output

Count from training data

HMM是求給定state a，出現feature x的條件機率P(x|a)
DBLSTM的network得出的是給定feature x ，在各個state的條件機率P(a|x)
利用貝式定理可以把DBLSTM的結果轉為HMM的
此hybrid model屬於**discriminative model**，而傳統的HMM屬於**generative model**，判別模型不需要考慮X與a的聯合機率分布，可能是表現較傳統HMM優秀的一個原因
另一個原因，HMM各個state是由不同GMM混合而成，而hybrid是由同一個Network的出的機率分布，純粹受**data driven**，因此表現會比較好

4.實驗比較Hybrid DBLSTM與End-to-End的表現:

(名詞 WER:Phoneme Error Rate， FER:Frame Error Rate， CE:Cross Entropy)

對TIMIT語料庫

目的

在TIMIT語料庫下比較DBLSTM在**hybrid training**下和**end-to-end methods**的performance

實驗變數

- table1是以DBLSTM為模型做不同的End-to-End的training，這裡使用CTC跟ST
- table2是以DBLSTM或DBRNN作為模型，不過訓練方式為Hybrid training，也就是用HMM-GMM的架構，因為這種架構不是End-to-End所以訓練時需要語言模型的協助
- table1和table2有noise的model都是在先用no noise的訓練資料先訓練過後才用有noise的訓練資料再做retrain

Table 1. TIMIT Results with End-To-End Training.

TRAINING METHOD	DEV PER	TEST PER
CTC	19.05 ± 0.11	21.57 ± 0.25
CTC (NOISE)	16.34 ± 0.07	18.63 ± 0.16
TRANSDUCER	15.97 ± 0.28	$\mathbf{18.07 \pm 0.24}$

Table 2. TIMIT Results with Hybrid Training.

NETWORK	DEV PER TEST PER	DEV FER TEST FER	DEV CE TEST CE
DBRNN	19.91 ± 0.22 21.92 ± 0.35	30.82 ± 0.31 31.91 ± 0.47	1.07 ± 0.010 1.12 ± 0.014
DBLSTM	17.44 ± 0.156 19.34 ± 0.15	28.43 ± 0.14 29.55 ± 0.31	0.93 ± 0.011 0.98 ± 0.019
DBLSTM (NOISE)	16.11 ± 0.15 $\mathbf{17.99 \pm 0.13}$	26.64 ± 0.08 $\mathbf{27.88 \pm 0.16}$	0.88 ± 0.008 $\mathbf{0.93 \pm 0.004}$

結果:

可以看出經過hybrid training後的DBLSTM和單純的End-to-End training有相當的PER， 而且當有Noise時前者甚至有更好的表現， 另外值得注意的是**DBLSTM相較於DBRNN有較好的表現**，因為它在訓練時每個LSTMcell都會對過去或者未來的input額外做過濾，而不是像RNN一樣直接餵進所有有連接的input。

對WSJ語料庫(Wall Street Journal)

目的

在WSJ語料庫下(large vocabulary corpus)測試hybrid DBLSTM對大型字彙的語音辨識下是否依然合適

實驗變數

以不同的Model(hybrid-DBLSTM， DNN， sGMM)對大型字彙語料庫做預測

Table 3. WSJ Results. All results recorded on the dev93 evaluation set. 'WER' is word error rate, 'FER' is frame error rate and 'CE' is cross entropy error in nats per frame.

SYSTEM	WER	FER	CE
DBLSTM	**11.7**	30.0	1.15
DBLSTM (NOISE)	12.0	**28.2**	**1.12**
DNN	12.3	44.6	1.68
sGMM [20]	13.1	–	–

結果

DBLSTM MODEL正如預期的相對於單純的DNN或是GMM都有較好的WER、FER、CE，證明**透過混合模型能夠更精準地去預測大型字彙系統**，不過值得注意的是雖然DBLSTM降低了WER，但在CE的部分卻沒有成正比的下降， 另外**DBLSTM在有noise下卻反而增加了WER**， 這可能造成之後在做預測時失去generalization而出錯。

實驗討論:

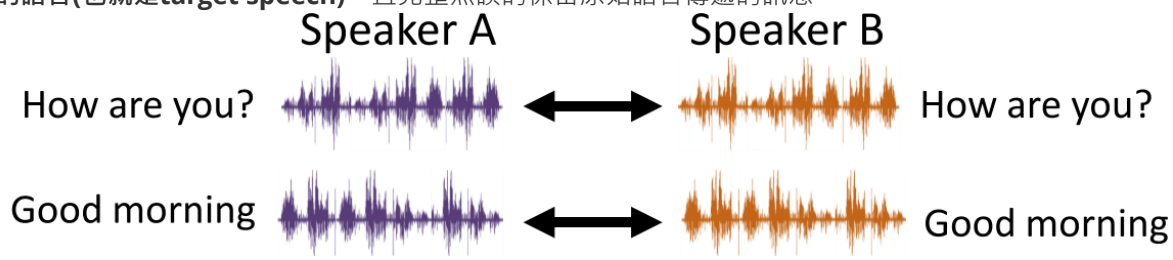
DBLSTM表現會比DNN比較好的原因在於因為採取雙向的NN，比較容易用上下文來預測，但在WSJ(大量詞彙)下，表現卻不比沒有經過pretrain的DNN好多少，Hybrid DBLSTM有主要兩個問題:

1. 因為Model的目標在於減少經過Forced allignment後的sequence與targe sequence之間的CE，因此**Forced Allignment的精準度也會間接影響到model的準確率**，所以當Forced Allignment是suboptimal時， model的準確率也會下降
2. 在WSJ實驗的結果中，發現CE和WER的關係，和想像中的正相關不太一樣，原因可能是CE並沒有考慮到lm(language model)中的機率，而作者懷疑這是因為DBLSTM因為考慮了前後文，因此**間接的學習到了word-level的lm**，而這樣的結果將會與之後接的**decoding時所用的lm互相干擾**而使得CE並不如預期的下降。

5. 應用: 用DBLSTM實作Voice conversion

為何要使用DBLSTM來Model Voice conversion?

VC(Voice Conversion)就是中文所說的變聲，目標就是**將一段來自A(也就是source)的語音轉換成B所講的語音(也就是target speech)**，且完整無誤的保留原始語音傳遞的訊息。



目前主要有兩種方式實現VC:

Rule-based approach

依據特定規則來調整聲音訊號中所帶的資訊，這樣做的好處是原始的資訊可以被完整地保留，但是會因為轉換的目標聲音不同而需要不同的規則而失去一般性。

Statistical approach

它是透過模型去預估Source跟target speech在頻率特徵上的對應函數(Mapping function)，目前已有GMM、DNN、RNN等模型，但這些模型都有些缺陷:

1. 正如前面所講的，DNN會把source視為各個獨立的speech frame去一一對應到target而**忽略到語音會與前後文有關係**
2. 而RNN的缺點在於雖然它可以捕捉到前後文字的語音，但遇到較長的Sequence時back propagation會有**gradient vanishing或exploding**的問題而導致train不起來。

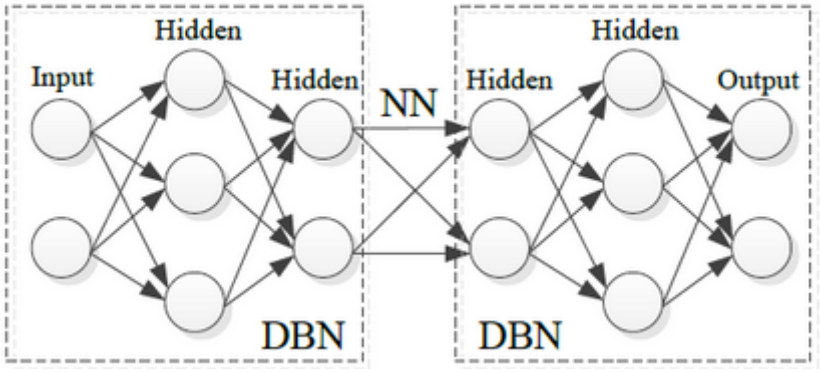
綜合以上，DBLSTM就很適合來做VC的model，因為DBLSTM運用了LSTM cell記憶來自過去與未來的sequence，同時LSTM cell只需要linear的記憶體空間就可以記下很長單位的語音訊號(因為每個LSTM cell只有常數個小單元)。

模型架構

為了**比較DBLSTM和以往模型**對VC的performance的差異，作者一共建立了4種模型，4種模型都是先對語音訊號做STRAIGHT analysis得到語音訊號的特徵(包括MCEPS(Mel cepstral coefficients = 頻譜的Envelop)、Fundamental frequency(F_0)、aperiodic component)，其中MCEPs會當作training的input餵給模型再輸出轉換後的的MCEPS，最後再把F_0跟aperiodic component經過轉換後與MCEPS合成，得到最終的語音，4種模型如下:

1. DNN:

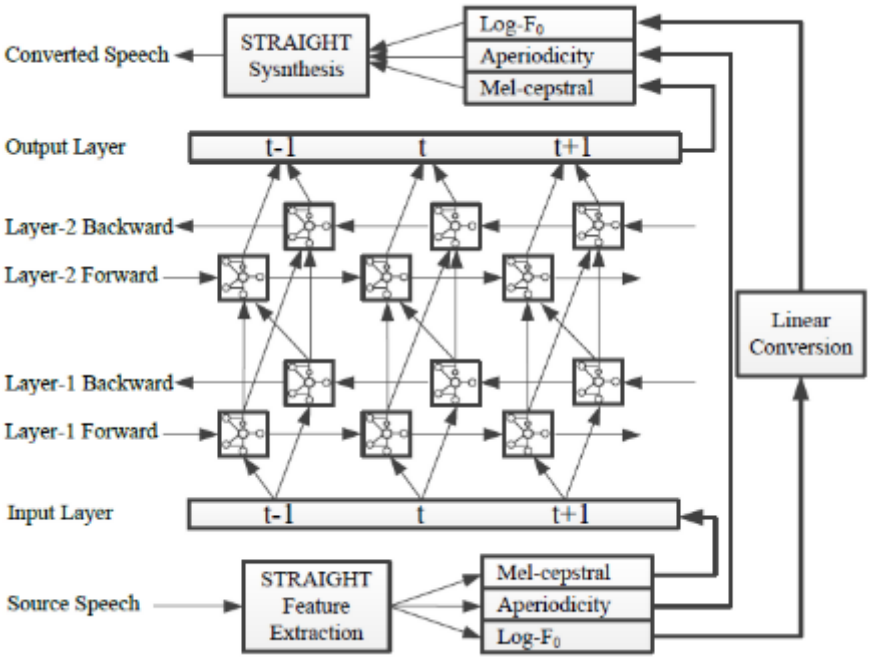
由兩個DBN(deep believe network)及連接的NN所建構而成。



2. DNN-DYN:

將DNN的input端額外加上Dynamic features，dynamic featuresu由語音訊號的特徵做一階和二階的微分得來，這樣做的好處在於可以幫助補足DNN無法捕捉語音訊號前後資訊的缺失。

3. DBLSTM:



模型前跟模型後的處理與DNN的前後處理相同，都是對語音訊號先做straight analysis之後再重新合成，只不過中間換成兩層的Bi-directional LSTM。

4. DBLSTM-DYN:

同樣是為DBLSTM加入Dynamic features作法是在最後的語音合成前多加入smoothing的函數。

模型訓練:

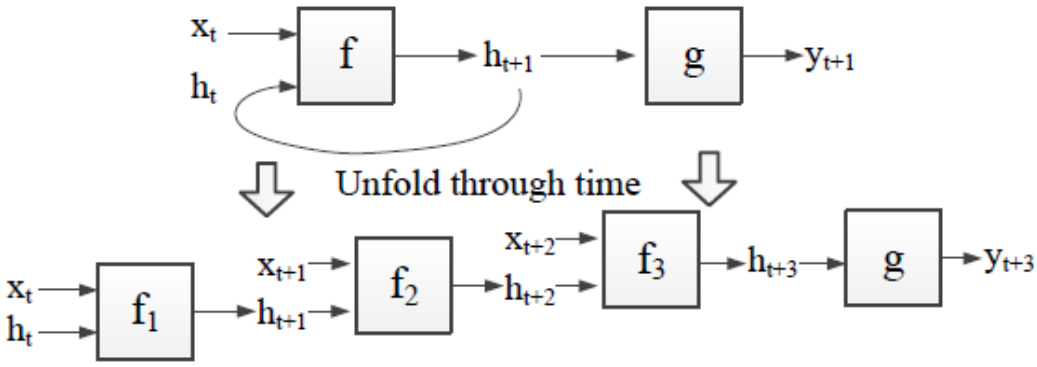
loss function

定義為Mel-cepstral Distortion(MCD)也就是衡量模型產生的MCEPS和Target speech的MCEPS的歐式距離(C_d代表第d個Mel_cepstrum的係數)

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{N}(c_d - c_d^{converted})^2}$$

Training

1. 因為DNN和DNN-DYN都是feed-forward的network，只需要做back propagation來迭代每次訓練的參數即可
2. DBLSTM因為輸出與過去及未來有關，需運用**Back-propagation through time(BPTT)**，做法是將原本的network展開(unfold)成一個共k層的feed-forward的network(係數k依model而定，下圖以k=3為例)，再用back-propagation來對參數做更新

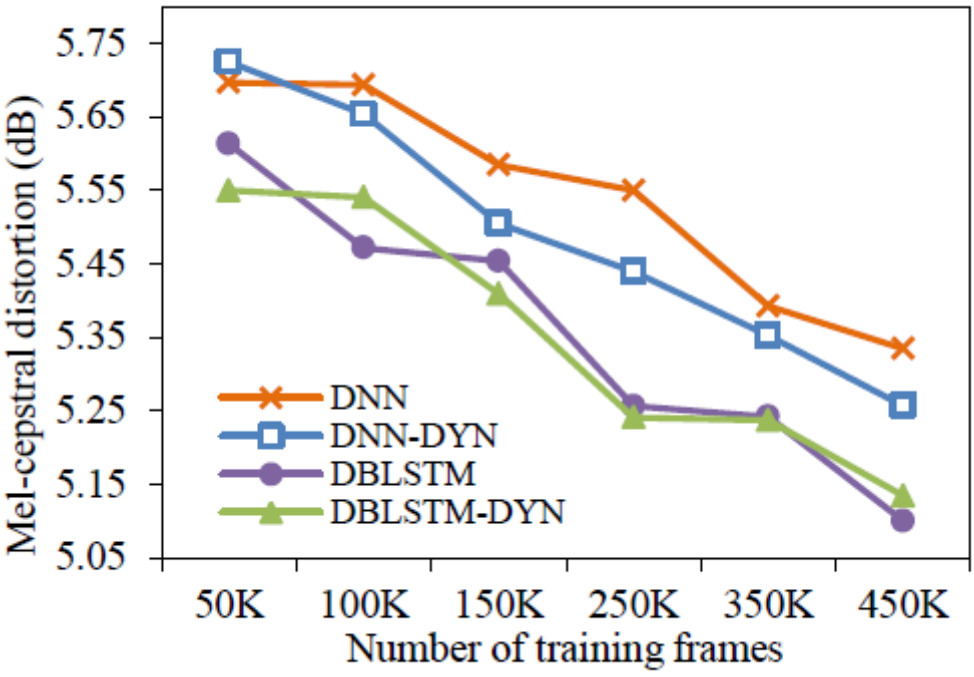


值得注意的是，因為DBLSTM與前後input有關，訓練時weight的gradient是**每input整個sentence**才更新一次，而DNN時則是語音訊號的**每一幀**就更新一次

比較DBLSTM與DNN對VC的表現:

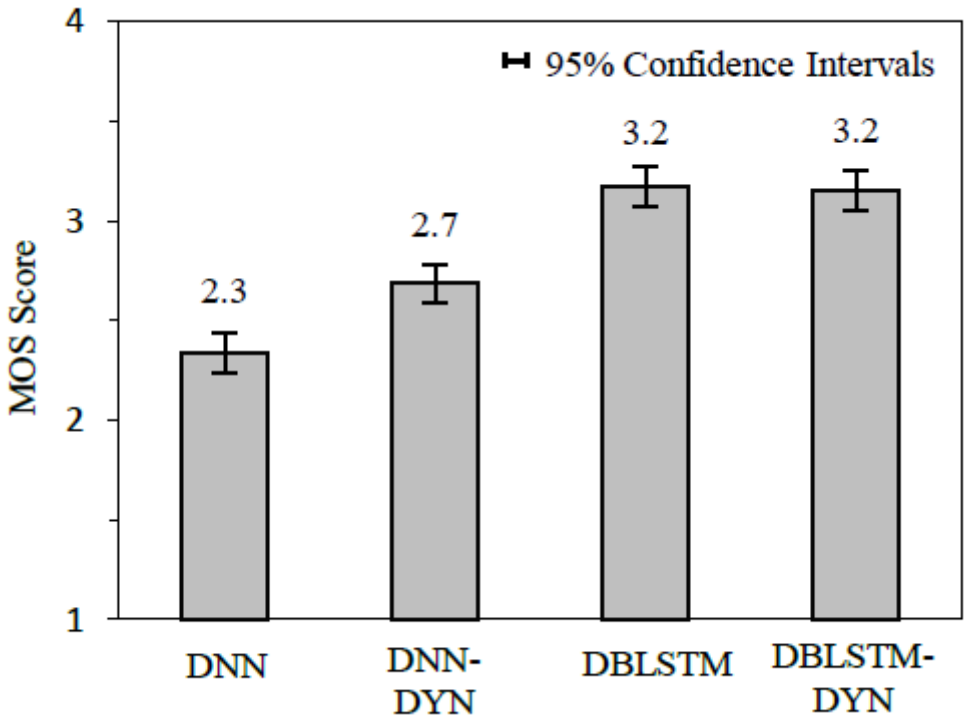
客觀分析:MCD

下圖是對不同的Model做訓練後，test他們與target speech的MCD，MCD越大代表Model對MCEPs的預測越差，由下圖可知**DBLSTM因為捕捉了前後文的語音訊號，確實比起DNN有更好的表現**，另外DNN-DYN也因為加入了Dynamic features而比DNN有較低的MCD，不過對於DBLSTM，加入Dynamic features並沒有降低DBLSTM的誤差，可能是因為**DBLSTM本身就有捕捉前後文的能力，加入Dynamic features反而有點多此一舉**。



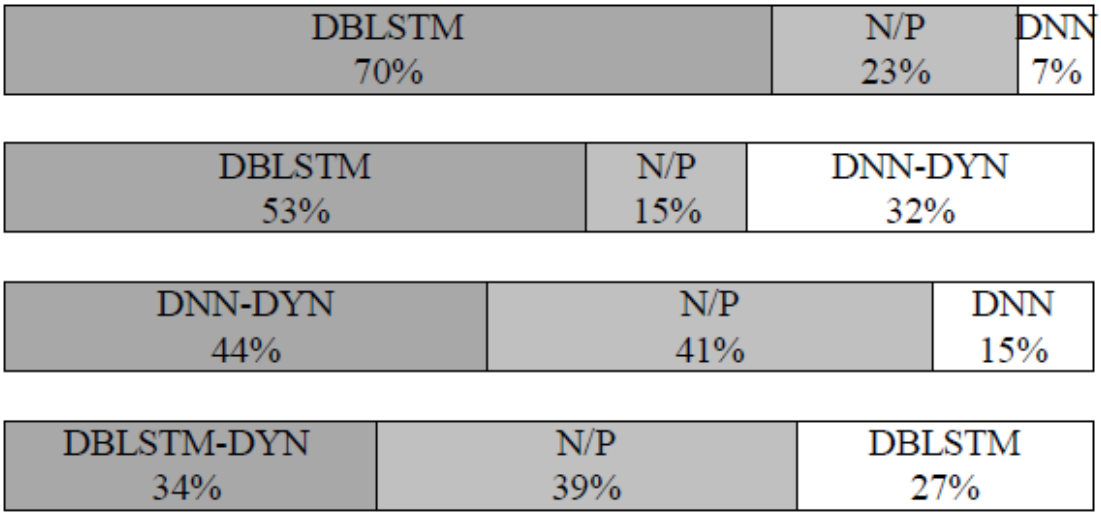
主觀分析:MOS

MOS是指Mean opinion score，也就是請受試者比較四種模型產生的語音訊號哪一種較為自然，下圖是結果，可以發現主觀分析上，還是**DBLSTM-DYN產生的語音訊號比較自然而得到較高的分數**



主觀分析:ABX

ABX是指將4種模型產生的訊號隨機打散於A、B，並請受試者比較A、B哪個聽其來比較接近X(或者是N/P，代表no preference)，結果如下圖，可以看出倆倆比較下，**有dynamic feature的優於沒有dynamic feature的，DBLSTM則優於DNN**



結論:

DBLSTM因為Model**掌握了前後的語音訊號**，所以能夠做出不管是客觀或主觀分析上都較好的VC，另外可以發現DNN產生的語音訊號斷斷續續且背景雜音較多，而**DBLSTM產生的訊號不只較乾淨，語音也較為連續**，DNN之所以斷斷續續原因就來自於DNN產生的語音訊號是根據各自獨立、分開的以幀為單位的Input訊號。

6. Reference:

- **Graves， N. Jaitly， A. Mohamed. “Hybrid Speech Recognition with Deep Bidirectional LSTM”， ASRU 2013.**
- **Graves， Alex， and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks." Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.**
- **Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks**
- **Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks. Author:L.S.K.H 2015 IEEE**
- **DLHLP 2020 HUNG-YI LEE 上課影片**
- 自動語音辨識 ASR 的前世今生

7. 分工

葉璟諄:
讀上述四篇論文
撰寫
1.甚麼是DBLSTM模型
2.將DBLSTM運用於End-to-End的語音辨識
3.結合DBLSTM與HMM(Hybrid DBLSTM)
楊宗桓:
讀上述四篇論文
撰寫
4.實驗比較Hybrid DBLSTM與End-to-End的表現
5.應用: 用DBLSTM實作Voice conversion