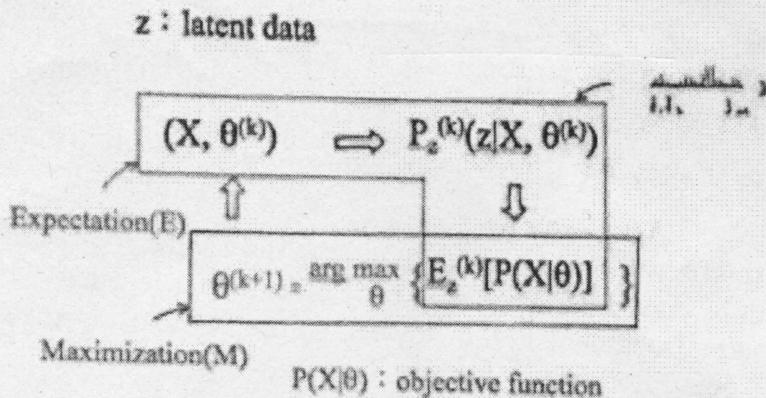


1. eigenspace 是將每個user的所有triphone的所有參數串成一條supervector後 將他們排成 matrix 然後做PCA，選出由較大eigenvalue的vector組成的子空間(ex, 200) 後在這個空間找出每個user所對應的點。它代表的意義是在supervector約480萬維中，挑出代表性高的latent feature，於是只要由這些少數的重要維度就可以代表一個user，而每個新的user都可以用低維度的eigenvector組合而成，代表他們聲音中的某些特性，然後可以再將低維空間的點轉到高維。因為若是200維的子空間，只有200個參數要調(相對於原來的480萬)，所以不需要太多的training data。

2. a)



b) M-step:先用第k輪的參數去算出每個latent data (ex, state seq in HMM)出現的機率，得到Z的機率分佈之後，跟計這個分佈算出 $P(X|\theta)$ 的期望值。E-step: theta是未知的參數，所以要算出能夠讓期望值最大化的參數，就是對未知的theta微分。

3.

不同的feature像是syllable, bi-syllable(feature是任兩個syllable組合),character,word,還可以將一連串的syllable每N個分為一個feature(ex s1s2s3, s2s3s4)。TF是每個term出現的頻率，可以用count或取log，越高代表term越常出現;IDF為 $\log(N/N_w)$ ，越高代表term出現在特定某些文章，可能帶有比較多文章相關資訊而不是function word。relevance score的計算可以用query vector 和 document vectors去算cosine similarity，較高代表這兩個所帶有得吃訓越像。

4.CMS是將在cepstral domain的input data直接減雜訊(若原始音訊X是zero mean則雜訊就是被雜訊干擾的音訊Y的mean; 也可以用EM算出產生Y的機率最高的h)，因為在cepstral domain  $Y = X + h$  故此方法可以避免任何種類的h。CMVS是在用CMS算出X後除以X的variance，讓分佈更靠近更接近原始的X。HEQ是 $Y = CDF_y^{-1}(CDF_x(X))$ ，代表將每個X值所在的CDF的位置對應到Y的CDF位置，然後再看Y等於多少會得到該位置的值，也就是兩個CDF直接轉換。

5. PLSA就是用一個latent topic的set，再用EM去train，所以給一個文件，用這方法可以用這些latent topic算出每個字出現的機率。

$$L_T = \sum_{i=1}^N \sum_{j=1}^n c(t_j, D_i) \log P(t_j | D_i)$$

$$P(t_j | D_i) = \sum_{t=1}^T P(t_j | T_t) P(T_t | D_i)$$

6.

a. Markov Decision Process (MAP) is mathematical framework for decision making

it is defined by a 5-tuple  $(S, A, T, R, \pi)$  where  
S is the set of states, current system status  
A is the set of actions the system can take at each state  
T transition probabilities between states where a certain action is taken  
R reward received when taking an action  
 $\pi$  is the policy or choice of action given the state  
the objective of MDP is to find a policy that maximizes the expected total reward

b.

在每個state，system可以有各種action像是回傳相關文件給user，然後user response可以是feedback，然後system可以根據這些response跳到下一個state然後再回傳更新過後的document給user。

7.a.DNN為有很多層hidden layer的類神經網路。像是hybrid system，可以將HMM裡面每個state的gaussian換成DNN。

b. RNN每次都會將上一輪的output當成新的input，所以我們每次將一句話裡面的一個字丟進去，RNN會根據句子之前的字來預測下一個字出現某個字的機率。

8.lattice 是在所有字之中，由某些較有可能組成該句子的字所組成的graph，每一條從左到右的字都是可能的句子。expected term frequency  $E(t,x)$ 代表在x這句話中，term t 出現在某一條可能的path中的count乘上這條path出現的機率，將所有的可能path加起來。代表某個term出現的期望值。

9.WFST是一種自動機，只是每個transition都會有score，然後會根據input不同，output不同的東西，而多個WFST可以combine在一起，weight也可以流動，所以可以把HMM, triphone phoneme sequence和word都看成WFST，train完這四個WFST後combine起來可以讓model更精簡快速，而要預測某一段聲音時只要將那段聲音也做成WFST後和原來的model座combination就好了。

10.CRF是一種ML的方法，每個input會output某個target Y(ex, POS)，然後每個tag都和對應的input還有前一個tag相關。應用上，將每個字使用CRF的方式training，可以依據之前句子所給的資料，決定這個字是否有重要資訊，像是在電影推薦系統，可以找出GENRE, PLOT, ACTOR等等。

- 1) (20) Explain the three key elements in a spoken dialogue system: what they are, how they operate, and how they are linked together.

a.speech recognition/understanding 接收音訊後理解句子的意思，可以用conditional random field的方法，如果有夠多的train data(ex,有標示POS tag)，則可以根據前後字之間的關係找到資訊；b.Discourse analysis：分析對話前後句之間的關係，將相對時間改為絕對或檢測前後文不連貫的錯誤等等，可以用填表的方式紀錄前文提到的資訊來分析後面得到的資訊；c.Dialog management：根據user說的話決定相對應的行為，可以用MDP，每個response都會對應一個action和分數，用reinforcement learning學習到能夠最大化得分的actions。他們之間得關聯是a.先接收音訊轉成可用的資訊ex句子，然後b會根據a得到的資訊填表，然後c可以根據使用者的問句和表的內容決定要問user什麼問題。

- 2) (20) What is Maximum Likelihood Linear Regression (MLLR) approach for speaker adaptation? Why the acoustic models for phonemes not observed in the adaptation data can also adapted?

先將相似的data分成一個個群(可以使用tree structure的方式，如果某一類的data太少則和另一類合併)，之後每一群用EM train出一個共用的 $\text{mean}' = A * \text{mean} + B$  讓 $\text{mean}'$ 產生機率最大given A B model. 可以調整沒被選到的data因為只要在同一群之中有任何點被選到，整個群共用的mean就會改變。

- 3) (20) Explain and discuss why words, subword units and keywords are useful in voice-based information retrieval for Mandarin Chinese, but each with respective limitations?

- 4) (20)

- (a) What is the mismatch in acoustic environment between training/testing conditions for speech recognition?

取得train data時環境的背景雜音或傳遞時生成的noise和取得test data的有可能不一樣，所以model的辨識能力會下降。

- (b) Explain what the model-based approaches, feature-based approaches and speech enhancement are, including mentioning the names of two examples for each of them.

model-based為直接real-time根據test data的noise調整model的參數(PMC,VTS)；feature-based為對cepstral domain的feature做調整，像是CMS為將訊號減掉它的mean(假設原始信號zero-mean)，就可以對所有noise免疫；CMVS則是算出 $X_{\text{CMS}}$ 後再除以它的標準差；speech enhancement 是利用noise在freq domain變化較慢，直接將聲音訊號轉成frequency domain，如果和訊號相差程超過原始信號的某個倍數就相減(spectral subtraction)，signal subspace方法概念是訊號是由噪音和我們要的音

訊組成的高維空間，所以將聲音投影到代表我們要的音訊的m維空間(用GSVD等方法)可以消除雜音。

5) (20) In vector space model of information retrieval.

(a) Explain how the vector representations of query q and document d can be constructed for each type of indexing feature?

不同的feature像是用word, subword, syllable來index，feature space分別是 vocabulary裡的字、兩個相連的字和所有可能的音節，然後用TF-IDF算出每個 feature的值，每個d和q就會有sparse的array。

(b) What are the Term Frequency (TF) and Inverse Document Frequency (IDF) and what they mean?

TF是每個單位(word,subword...)出現的頻率，可以用count或是取log讓它增加的比較平緩，TF越高代表某個字越常出現;IDF是 $\log(N/N_w)$ ,  $N_w$  是出現該單位的document 數，IDF越高代表該字越可能有某領域相關重要資訊，不是function word.

(c) How the relevance score can be computed using these vectors?  
每個document和query都會有一個vector，可以算d,q的vector之間的cosine similarity 找出和query比較接近的文件。

6) (20) Write down the two steps in each iteration of EM algorithm and explain how they operate and what they mean. (20%)

每輪先用上一輪的model和觀察到的data去算出每個隱藏資訊(ex, HMM裡的state seq)出現的機率，算出latent data的機率分佈之後，用這個機率分佈去求 $P(X|\theta)$ 的期望值。M step則是maximize這個期望值，選出最可能的 $\theta$ 作為下一輪的參數。

- OPEN Lecture Slides (Printed Version) and Personal Notes
- You have to use CHINESE sentences to answer all the questions, but you can use English terminologies
- Total points: 160

- 
1. (20) In eigenvoice approach, explain how the eigenvoice space is constructed, what that means, and why rapid speaker adaptation can be achieved with very limited quantity of adaptation data?
  2. (20) Write down the two steps in each iteration of EM algorithm and explain how they operate and what they mean.
  3. (20) In vector space model of information retrieval, explain how the vector representations of query  $q$  and document  $d$  can be constructed for each type of indexing feature, what the Term Frequency (TF) and Inverse Document Frequency (IDF) are, what they mean, and how the relevance score can be computed using these vectors.
  4. (20) In feature based approach of robust speech recognition, explain the Cepstral Mean Subtraction (CMS), Cepstral Mean and Variance Normalization (CMVN), and Histogram Equalization (HEQ); what they are and why they work.
  5. (10) What is Probabilistic Latent Semantic Analysis (PLSA)?
  6. (20)
    - (10) Describe the mathematical framework of a Markov Decision Process(MDP)?
    - (10) Describe how MDP can help in Multi-modal Interactive Dialogue for Spoken Content Retrieval?
  7. (20)
    - (10) What is a Deep Neural Network (DNN)? How can it be used in acoustic modeling?
    - (10) Explain how a Recurrent Neural Network (RNN) can be used in language modeling.
  8. (10) Please explain what is the "lattice" of an utterance, and what is the "expected term frequency" for a term in the lattice.
  9. (10) What is WFST? How is it useful in speech recognition?
  10. (10) Explain what is the Conditional Random Field (CRF) and how it can be used for filling in spoken dialogues.