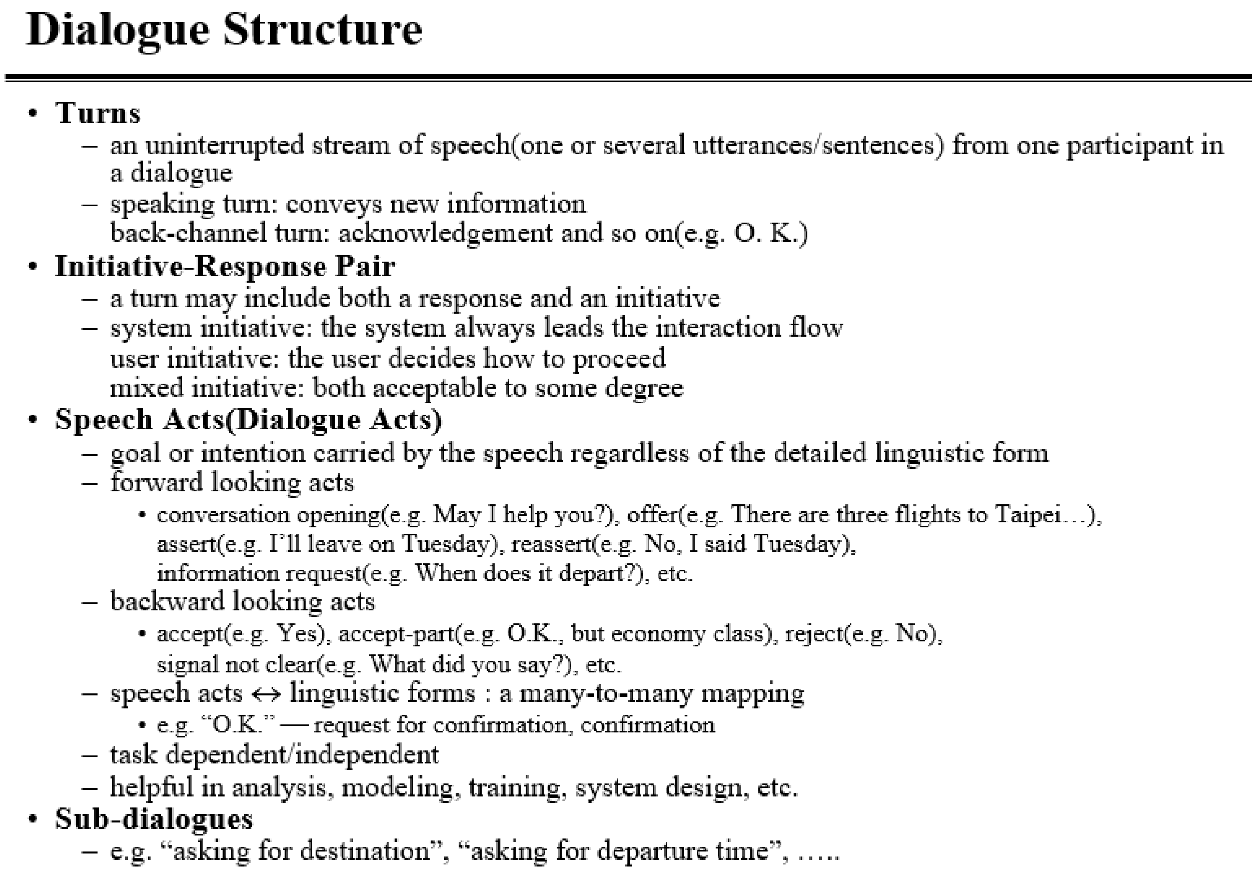
- base-approach 11-4 18-4
- CMS (Cepstral Mean Substraction) 15-2 16-4 note-4
- CRF (Conditional Random Field) 15-11 16-10 18-7 **18**-2 note-10
- CNN 16-7 18-9 note-7
- Dialogue Acts 18-12 *below-1*
- Distribute Speech Recognition (client-server/server-only) 7-8
- Eigenvoice 7-1 16-1 note-1
- EM algorism 7-5 11-6 12-6 15-10 16-2 18-6 7-note
- ETF、TF、IDF (Term Frequency) 11-5 12-5 15-4 16-3 18-7
- Extractive and Abtractive Summarization 18-2 *below-2*
- Gaussian Mixture Model **18**-3 *below-3*
- HEQ (Histogram EQunlization) 15-2 note-4
- lattiee 15-4 16-8
- LDA (Linear Discriminative Analysis) 7-3 18-5 7-note
- Likelihood Ratio Test 7-6
- LSA (Latent Semantie Analysis) 7-2 15-7
- MAP 15-1
- MDP (Markov Decision Process) 15-3 16-6 18-8 note-6
- MLLR (Maximun Likelihood Linear Regression) 11-2 12-2 15-1**18**-1
- mismatch in accoustic 11-4 12-4 18-5 **18**-10
- MMR (Maxinum Margin Relevance) 15-6
- PCA 7-note **18**-1
- PLSA 16-5 **18**-5
- PRF (Psuedo Relevance Feedback) 15-5 18-11
- Random Forest 15-9 18-10
- RNN 16-7 18-9 note-7
- Speach Robustness note-4
- Spoken dialogue system 7-7 11-1 12-1
- Spoken Document Understanding and Organization **18**-7 *below*-4
- SVD (Singular Value Decomposition) 15-7
- SVM **18**-8 *below*-5
- Vector Space Model 7-4 11-5 **18**-9 note-3
- WFST 15-8 16-9 **18**-6 note-9
- 中文 words/subwords 11-3 12-3 18-3

Dialogue Acts

- 

Dialogue Structure

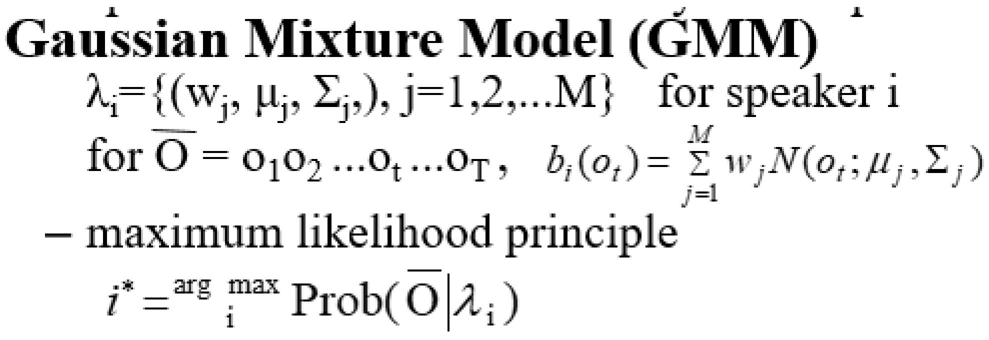
- **Turns**
  - an uninterrupted stream of speech(one or several utterances/sentences) from one participant in a dialogue
  - speaking turn: conveys new information
    back-channel turn: acknowledgement and so on(e.g. O. K.)
- **Initiative-Response Pair**
  - a turn may include both a response and an initiative
  - system initiative: the system always leads the interaction flow
    user initiative: the user decides how to proceed
    mixed initiative: both acceptable to some degree
- **Speech Acts(Dialogue Acts)**
  - goal or intention carried by the speech regardless of the detailed linguistic form
  - forward looking acts
    - conversation opening(e.g. May I help you?), offer(e.g. There are three flights to Taipei…), assert(e.g. I'll leave on Tuesday), reassert(e.g. No, I said Tuesday), information request(e.g. When does it depart?), etc.
  - backward looking acts
    - accept(e.g. Yes), accept-part(e.g. O.K., but economy class), reject(e.g. No), signal not clear(e.g. What did you say?), etc.
  - speech acts ↔ linguistic forms : a many-to-many mapping
    - e.g. "O.K." — request for confirmation, confirmation
  - task dependent/independent
  - helpful in analysis, modeling, training, system design, etc.
- **Sub-dialogues**
  - e.g. "asking for destination", "asking for departure time", …..

- 在語言學中，特別是在自然語言理解中，對話行為是在對話對話的上下文中的一種話語，其在對話中起作用。 對話行為的類型包括問題，陳述或行動請求。對話行為是一種言語行為。

Extractive and Abstractive Summarization

- 摘要總結: 選擇文檔中的句子
- 抽象概述: 生成描述文檔內容的句子

Gaussian Mixture Model

- 

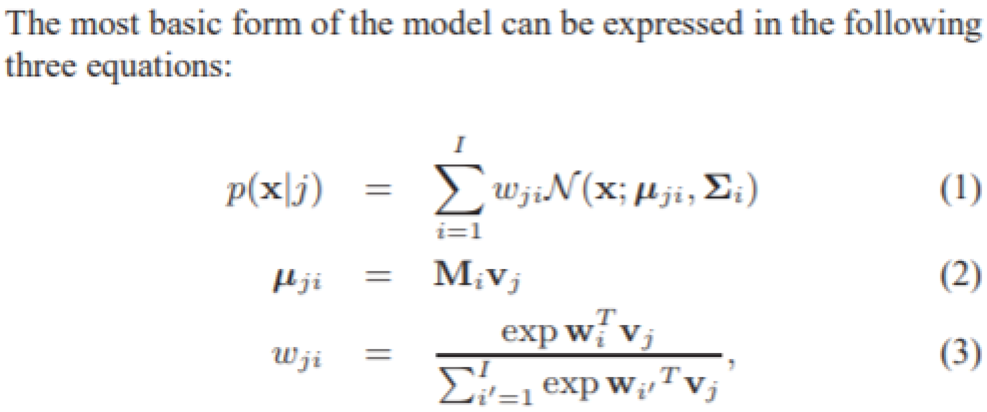
Gaussian Mixture Model (GMM)

$\lambda_i=\{(w_j, \mu_j, \Sigma_j,), j=1,2,...M\}$ for speaker i

for $\overline{O} = o_1 o_2 \ldots o_t \ldots o_T$, $b_i(o_t) = \sum_{j=1}^{M} w_j N(o_t; \mu_j, \Sigma_j)$

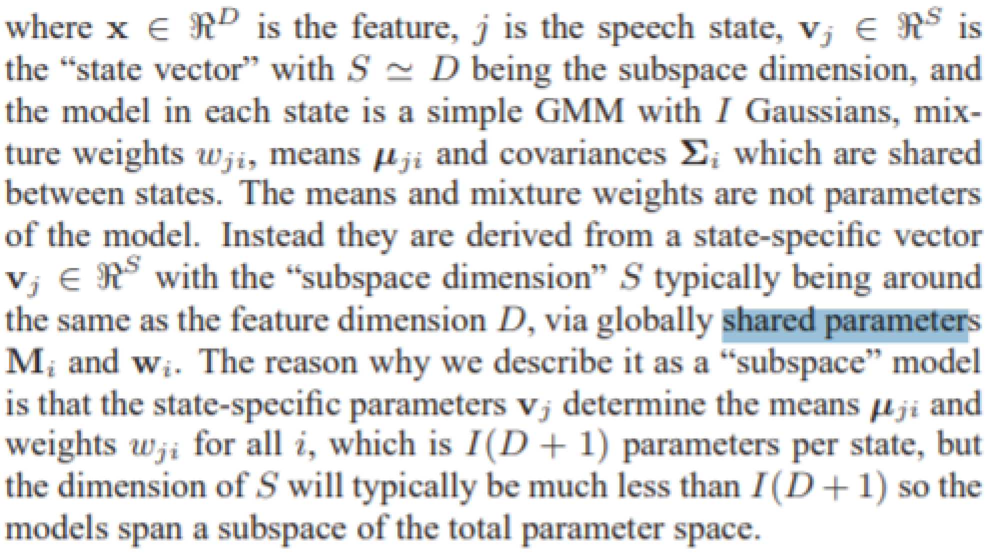
– maximum likelihood principle

$i^* = \arg\max_{i} \mathrm{Prob}(\overline{O} | \lambda_i)$



The most basic form of the model can be expressed in the following three equations:

$$\begin{aligned} p(\mathbf{x}|j) &= \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i) && (1) \\ \boldsymbol{\mu}_{ji} &= \mathbf{M}_i\mathbf{v}_j && (2) \\ w_{ji} &= \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}{}^T \mathbf{v}_j}, && (3) \end{aligned}$$

- 

  where $\mathbf{x} \in \Re^D$ is the feature, j is the speech state, $\mathbf{v}_j \in \Re^S$ is the "state vector" with $S \simeq D$ being the subspace dimension, and the model in each state is a simple GMM with I Gaussians, mixture weights w_{ji}, means $\boldsymbol{\mu}_{ji}$ and covariances $\boldsymbol{\Sigma}_i$ which are shared between states. The means and mixture weights are not parameters of the model. Instead they are derived from a state-specific vector $\mathbf{v}_j \in \Re^S$ with the "subspace dimension" S typically being around the same as the feature dimension D, via globally shared parameters $\mathbf{M}_i$ and $\mathbf{w}_i$. The reason why we describe it as a "subspace" model is that the state-specific parameters $\mathbf{v}_j$ determine the means $\boldsymbol{\mu}_{ji}$ and weights w_{ji} for all i, which is $I(D+1)$ parameters per state, but the dimension of S will typically be much less than $I(D+1)$ so the models span a subspace of the total parameter space.

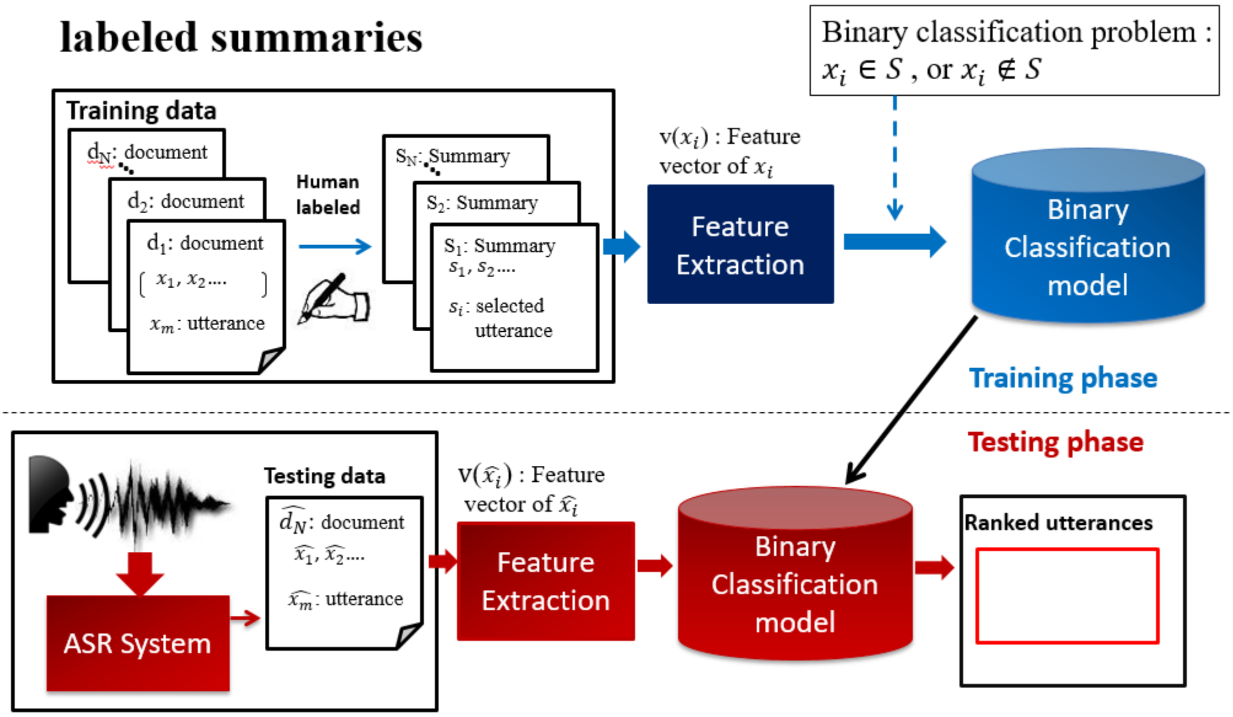
Spoken Document Understanding and Organization

- 與用標題和段落更好地構造並因此更容易檢索和瀏覽的書面文檔不同，多媒體/口述文檔僅僅是視頻/音頻信號，或者即使自動轉錄也包括很長的單詞序列，包括錯誤。 一般而言，它們不會被分段為段落，並且段落中沒有提及標題。 因此，它們更難以檢索和瀏覽，因為用戶根本無法從頭到尾瀏覽每個。 因此，為了更容易檢索/瀏覽，需要更好的方法來理解和組織口頭文檔（或相關的多媒體內容）。

SVM

Supervised Approach: SVM or Similar

- **Trained with documents with human labeled summaries**



Binary classification problem : $x_i \in S$, or $x_i \notin S$
Training data
d_N: document
d_2: document
d_1: document
x_1, x_2....
x_m: utterance
Human labeled
S_N: Summary
S_2: Summary
S_1: Summary
s_1, s_2....
s_i: selected utterance
$v(x_i)$: Feature vector of x_i
Feature Extraction
Binary Classification model
Training phase
Testing phase
Testing data
$\widehat{d_N}$: document
$\widehat{x_1}, \widehat{x_2}$....
$\widehat{x_m}$: utterance
ASR System
$v(\widehat{x_i})$: Feature vector of $\widehat{x_i}$
Feature Extraction
Binary Classification model
Ranked utterances