10.0 Utterance Verification and Keyword/Key Phrase Spotting

References: 1. "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score", IEEE Trans. Speech & Audio Processing, Nov 2001

2. "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", IEEE Trans. Acoustics, Speech & Signal Processing, Nov 1990

3. "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures", IEEE Trans. Speech & Audio Processing, March 2000

4. "Confidence Measures for Large Vocabulary Continuous Speech Recognition", IEEE Trans. Speech & Audio Processing, March 2001

5. "Key Phrase Detection and Verification for Flexible Speech Understanding", IEEE Trans. Speech & Audio Processing, Nov 1998

Likelihood Ratio Test and Utterance Verification

- **Detection Theory— Hypothesis Testing/Likelihood Ratio Test**
  - 2 Hypotheses: H_0, H_1 with prior probabilities: $P(H_0)$,$P(H_1)$
    observation: X with probabilistic law: $P(X|H_0)$, $P(X|H_1)$
  - MAP principle

    choose H_0 if $P(H_0|X) > P(H_1|X)$
    choose H_1 if $P(H_1|X) > P(H_0|X)$

    $$\Rightarrow \frac{P(H_0|X)}{P(H_1|X)} \underset{H_1}{\overset{H_0}{\gtrless}} 1$$

  - Likelihood Ratio Test

    $P(H_i|X) = P(X|H_i)P(H_i)/P(X)$, i=0,1

    $$\Rightarrow \frac{P(X|H_0)}{P(X|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \frac{P(H_1)}{P(H_0)} = Th$$

    likelihood ratio-Likelihood Ratio Test

- **Utterance Verification**

$$\rho(X; w_i, \overline{w_i}) = \frac{P(X|w_i)}{P(X|\overline{w_i})} > Th$$

w_i : HMM for a given word

$\overline{w_i}$: anti-model (background model) of w_i, or alterative hypothesis, trained with undesired phone units, cohort set, competing units, or similar

$\rho(X; w_i, \overline{w_i})$:confidence score, confidence measure

Type I error: missing (false rejection)
Type II error: false alarm (false detection)
false alarm rate, false rejection rate, detection rate, recall rate, precision rate
Th: a threshold value adjusted by balancing among different performance rates

Generalized Confidence Score for Utterance Verification

- **Frame-level Confidence Score**

$$\rho_i(o_t;\lambda_i,\overline{\lambda_i}) = \log[\frac{p(o_t|\lambda_i)}{p(o_t|\overline{\lambda_i})}]$$

o_t : observation vector at frame t

λ_i : state i of HMM for a phone unit p

$\overline{\lambda_i}$: anti - model (or background model) for state i, trained with the cohort set for the phone unit p

$$\ell[\rho_\mathrm{i}(o_t;\lambda_i,\overline{\lambda_i})] = \ell[\rho_\mathrm{i}] = \log[\frac{1}{1+\exp[-\gamma(\rho_\mathrm{i}-\theta)]}] \quad \text{log sigmoid function}$$

$\ell[\rho_\mathrm{i}] \to 0$ if ρ_i large, not affecting the local search decisions

$\ell[\rho_\mathrm{i}] \to$ very negative if ρ_i small, rejecting unlikely paths

- **Phone-level Confidence Score**

$$\rho_p(o_t) = (1/\tau)\sum_{u=t-\tau+1}^{t}\rho_j(o_u)$$

τ : length of the phone p

$\ell[\rho_p(\mathrm{o}_\mathrm{t})] = \ell[\rho_p]$, evaluated at the end of a phone p

- **Word-level Confidence Score**

$$\rho_w(o_t) = (1/N)\sum_p \ell[\rho_p]$$

p : a phone unit in the word w

N : total number of phone units in the word w

evaluated at the end of a word

- **Multi-level Confidence Score**

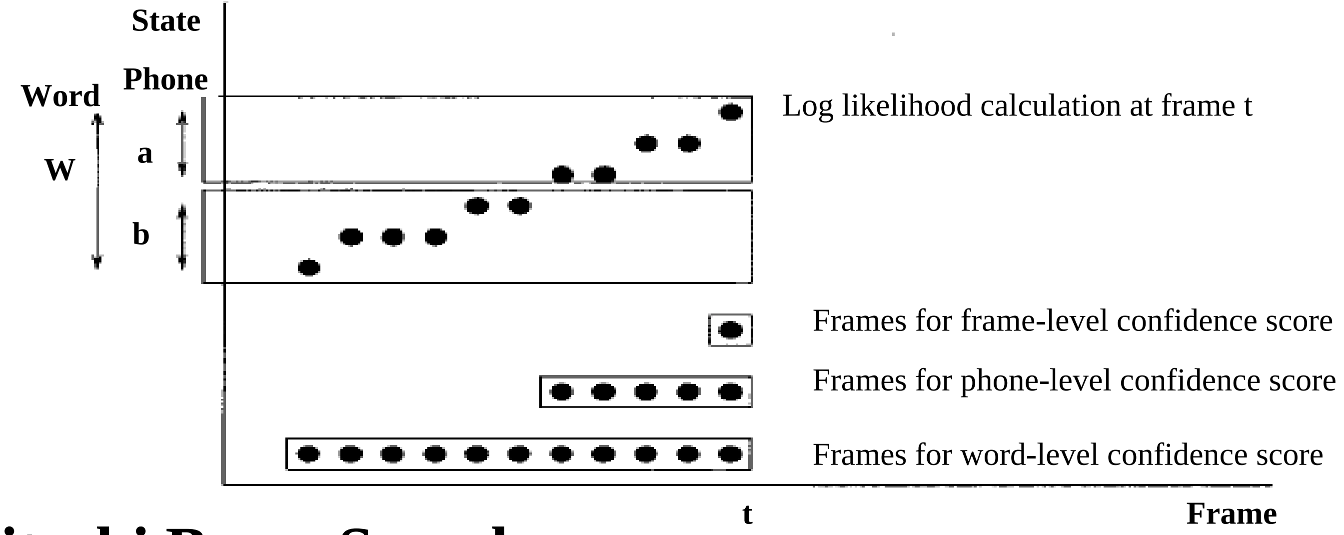
$$\rho_{M,i,t} = w_f \bullet \ell[\rho_i] + w_p \bullet \ell[\rho_\rho] + w_w \bullet \rho_\mathrm{w}(o_t)$$

frame-level score may not be stable enough, average over phone and word gives better results

w_f, w_p, w_W: weights, w_p=0 if not at the end of a phone, w_W=0 if not at the end of a word

Generalized Confidence Score in Continuous Speech Recognition

- **Evaluation of Multi-level Confidence Scores**



- **Viterbi Beam Search**

  D (t, q_t, w): objective function for the best path ending at time t in state q_t for word w

  – Intra-word Transition as an example

  $$D(t,q_t,w)=\max_{q_{t-1}}[D(t-1,q_{t-1},w)+d(o_t,q_t|q_{t-1},w)]$$

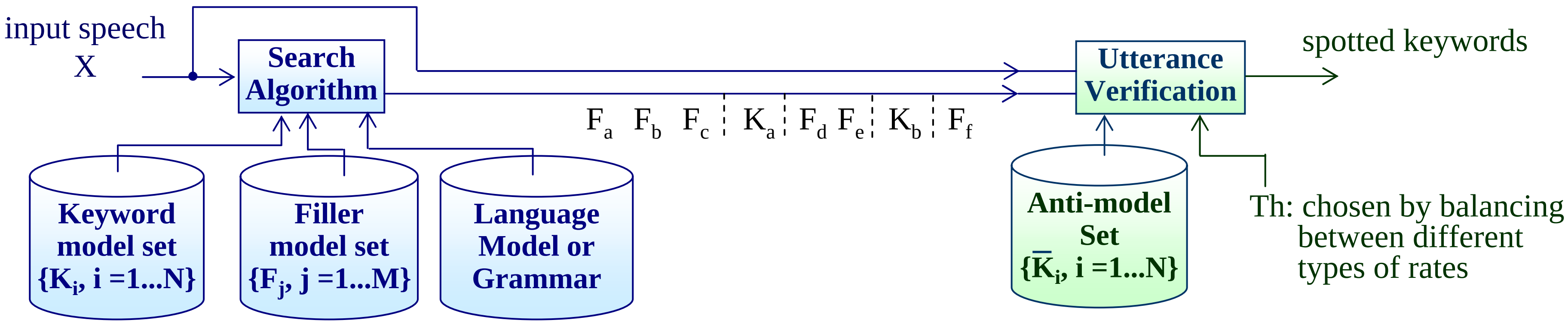
  with generalized confidence score for utterance verification

  $$D(t,q_t,w)=\max_{q_{t-1}}[D(t-1,q_{t-1},w)+d(o_t,q_t|q_{t-1},w)+\varepsilon\rho_{M,i,t}]$$

  – unlikely paths rejected while likely paths unchanged
  helpful in beam search

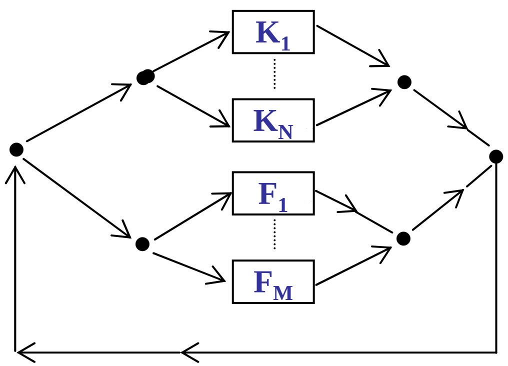
Keyword Spotting

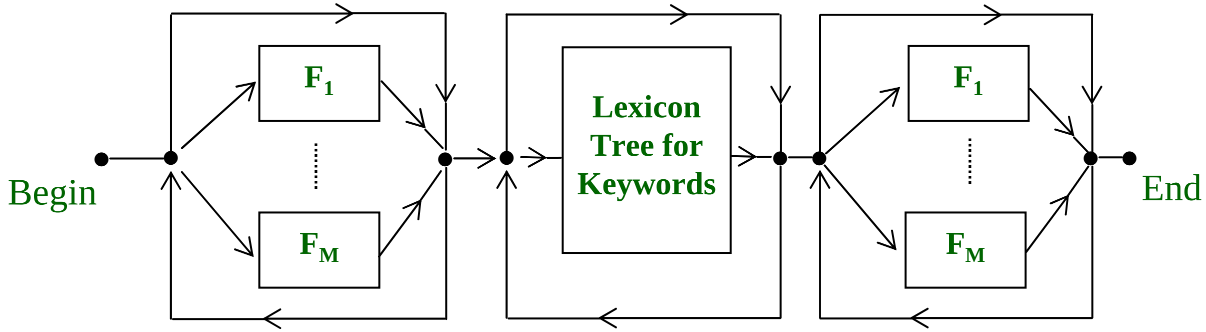
- **To Determine if a Keyword out of a Predefined Keyword Set was Spoken in an Utterance**
  - no need to recognize (or transcribe) all the words in the utterance
  - utterances under more unconstrained conditions
  - applications in speech understanding, spoken dialogues, human-network interaction
- **General Principle: Filler Models, Utterance Verification plus Search Algorithm**



– filler: models specially trained for non-keyword speech

- **Viterbi Search through Networks**

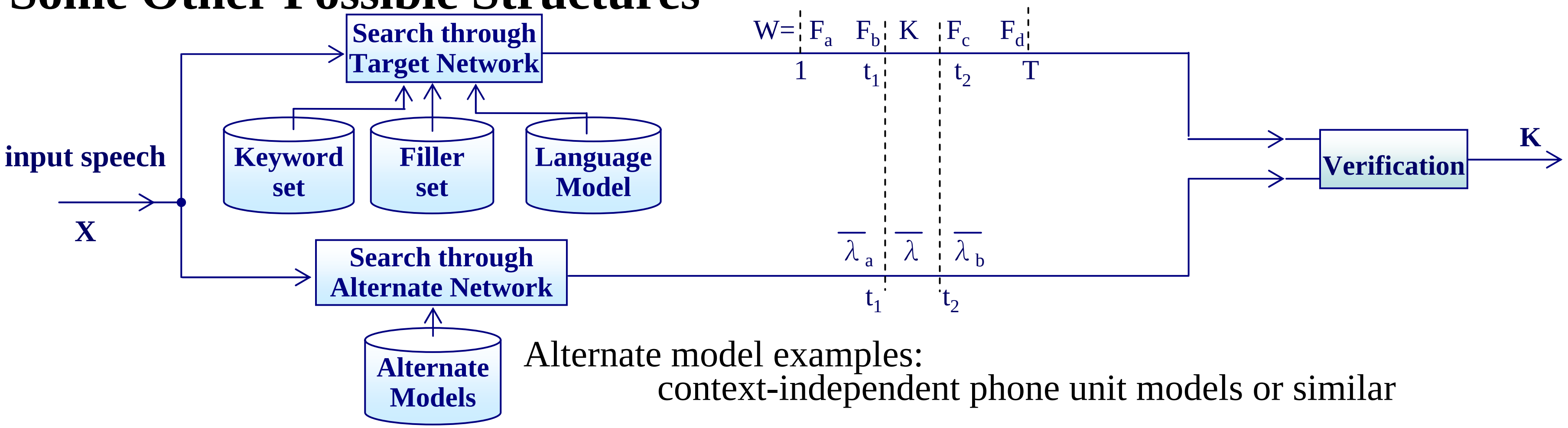




- **All Different Search Algorithms Possible: A*, Multi-pass, etc.**

Keyword Spotting

- **Some Other Possible Structures**



Alternate model examples:
context-independent phone unit models or similar

– several approaches for verification

$$\rho(X;K)=\frac{p[X(t_1,t_2)|K]}{\underset{q}{Max}\; p[X(t_1,t_2)|q,\bar{\lambda}]}$$

q, $\bar{\lambda}$: state sequence/model sequence in (t_1,t_2)

$X(t_1,t_2)$: X frames in (t_1,t_2)

$$\rho(X;K)=P(q_{t_2}=q_{e,k}|X(1,T),\Lambda)=\frac{\alpha_{t_2}(q_{e,k})\beta_{t_2}(q_{e,k})}{\sum_{q\neq q_{e,k}}\alpha_{t_2}(q)\beta_{t_2}(q)}$$

q_{t_2}: state at time t_2,

$q_{e,k}$: ending state of model k

q: any state

- **MCE Training of All Models (Keyword, Filler, Anti-model, Alternate, etc.)**

$$d(X,K)=\log p(X|K)-\log p(X|\bar{K})$$

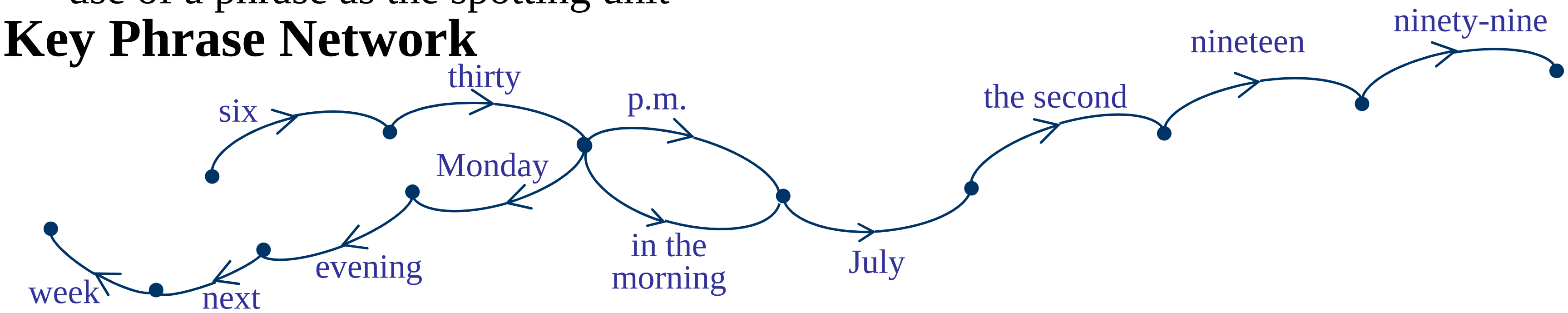
$$L(\Lambda)=\sum_X(\sum_K \ell[d(X,K)]\,\delta(X\in K)+\sum_K \ell[-d(X,K)]\,\delta(X\notin K))$$

$$\Lambda_{n+1}=\Lambda_n-\varepsilon_n\nabla L(\Lambda)$$

$X\in K$: X does include the keyword K, X : an utterance segment for a keyword hypothesis K

Key Phrase Spotting/Detection

- **Key Phrase: one or a few keywords connected, or connected with some function words**
  - e.g. on Sunday, from Taipei to Hong Kong, Six thirty p.m.
- **Spotting/Detection of Longer Phrase is More Reliable**
  - a single keyword may be triggered by local noise or confusing sounds
  - similar verification performed with longer phrase (on frame level, phone level, etc.)
  - use of a phrase as the spotting unit
- **Key Phrase Network**



  - every arc represents a group of possible key words
  - grammar for permitted connection defined manually or statistically
  - N-gram probabilities trained with a corpus
  - key phrases are easier mapped to semantic concepts for further understanding
- **Automatic Algorithms to Identify Key Phrases from a Corpus**
  - grouping keywords with semantic concepts, e.g. City Name (Taipei, New York,...)
  - starting with a core semantic concept, growing on both sides by some criteria, etc.
  - example criteria:

    "stickiness" $= P(c,c_0)/[P(c)\cdot P(c_0)] = P(c\,|\,c_0)/P(c) = I(c;c_0)$ — c: a semantic concept

    "forward-backward bigram" $= [P(c\,|\,c_0)\cdot \overline{P}(c_0\,|\,c)]^{1/2}$ — c_0: the core semantic concept