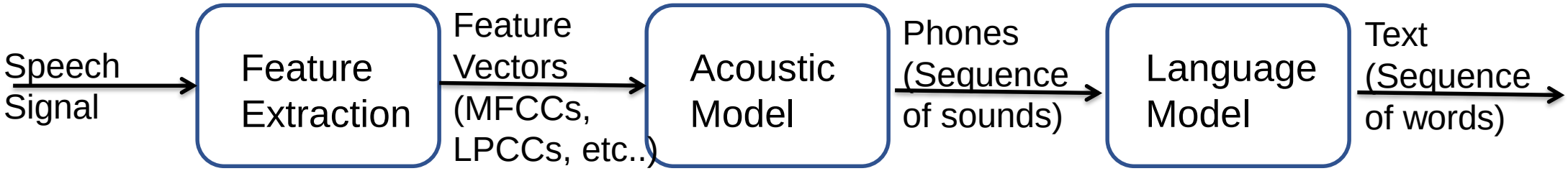
Automatic Speech Recognition (ASR)

ASR : Framework Formulation



Speech signal → Text (Sequence of words)

$w^* = \arg\max P(w/y)$

w = word sequence

y = observation sequence (sequence of feature vectors)

w* = optimal word sequence (Best word sequence)

P(w/y) = {P(y/w)P(w)}/P(y)

P(w/y) → Aposterior probability

P(y/w) → likelihood (pdf of specific word model)

P(w) → Language model (prior probability)

Types of ASR

- Isolated word recognition
- Connected word recognition
- Continuous speech recognition

- Isolated word recognition
  - ✓ End-point detection (begin-end detection)
  - ✓ Large number of words (100-1000 words)
  - ✓ Template matching (Recognition step)
  - ✓ Time alignment (dynamic time warping : DTW)
  - ✓ No contextual information (no grammar, no syntax) : Difficulty
  - ✓ Less variability (includes co-articulation)
  - ✓ All words are equally important

Types of ASR (Cont..)

- Connected word recognition
  - ✓Small vocabulary (approx.. 20 words)
  - ✓Begin-end & word-boundary detection
  - ✓How many words & what are they?
  - ✓2-level DTW
  - ✓Complexity in recognition (Ex: 5 words → 20^5 word sequences
  - ✓Best word sequence
  - ✓No language model (No contextual & No grammar)
  - ✓All words are equally important

Types of ASR (Cont..)

- Continuous Speech Recognition
  - ✓ Large vocabulary (size of dictionary)
  - ✓ Suitable unit for recognition
    - In view of speech : Phone or Syllable
    - In view of language : word, phrase, sentence
  - ✓ Contextual information (Language Model)
  - ✓ Begin-end & word-boundary detection
  - ✓ All words are not important (WER ←→ Message understanding)
  - ✓ Read / Extempore / Conversational / Spontaneous modes of speech

Basic Issues in Speech Recognition

- Speech unit for recognition (smaller vs larger units)
- Task syntax : Association with grammar
- Task perplexibility : Average word branching factor
- Speaking mode : isolated / connected-word / continuous / spontaneous
- Speaker-specific / speaker independent /speaker adaptive
- Speaker interaction : machine vs human
- Speaker environment : indoor, outdoor, crowded places, …
- Effect of transducer
- Effect of transmission channel
- Inherent speech variability

Language Model

Word Sequence $w = w_1 w_2 \ldots w_q$

$P(w) = P(w_1)P(w_2/w_1)P(w_3/w_1 w_2) \ldots\ldots P(w_q/w_1 w_2 \ldots\ldots w_{q-1})$

N-gram model : $P(w_j/w_1 w_2 \ldots\ldots w_{j-1}) = P(w_j/w_{j-N+1} \ldots\ldots w_{j-1})$

N = 1 : Unigram model : $P(w_j)$ → No context

N = 2 : Bigram model : $P(w_j/w_{j-1})$

N = 3 : Trigram model : $P(w_j/w_{j-2} w_{j-1})$

Word-pair models : $P(w_j/w_k) = 1$ ($w_k w_j$ is present in the language)

= 0 (($w_k w_j$ is not present in the language

No-grammar model : $P(w_j/w_k) = 1$ (for all j & k)

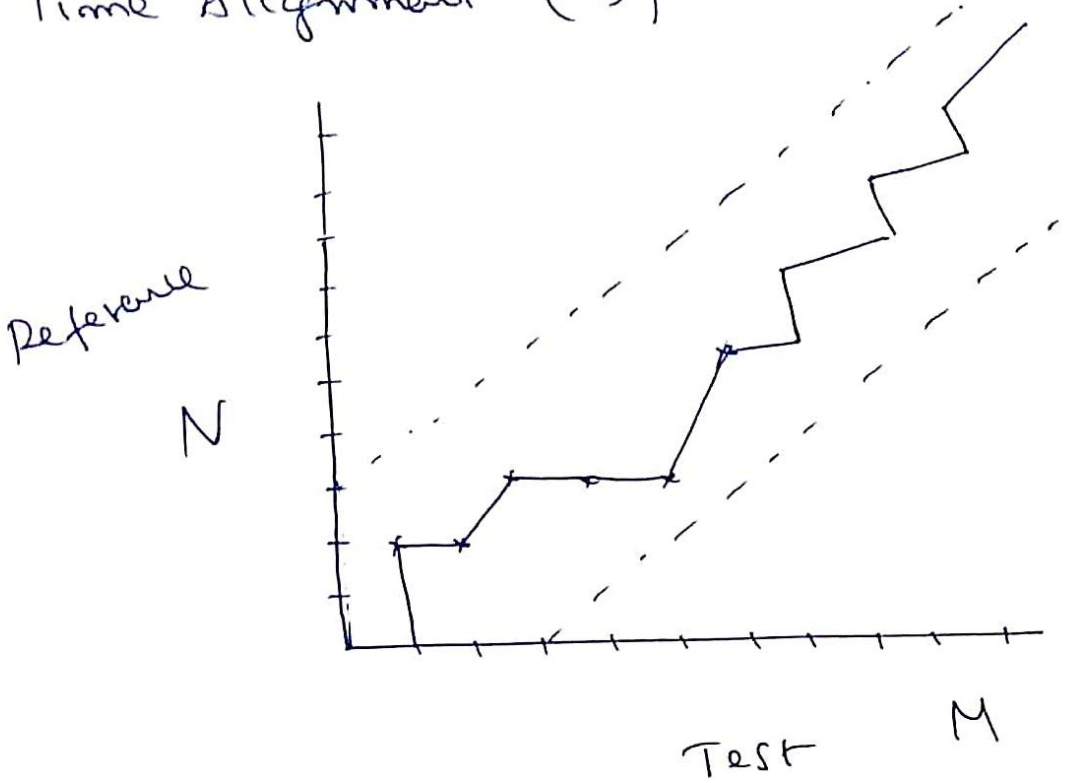
Statistical language model : $\mathrm{P}_N(w) = \prod_{i=1}^{q} P(w_i/w_{i-1} w_{i-2} \ldots\ldots w_{i-N+1})$

$$P\hat{}(w_i/w_{i-1} w_{i-2} \ldots\ldots w_{i-N+1}) = \frac{F(w_i w_{i-1} w_{i-2} \ldots\ldots w_{i-N+1})}{F(w_{i-1} w_{i-2} \ldots\ldots w_{i-N+1})}$$

Template Matching (Dynamic Time Warping : DTW)

Template Matching

Time Alignment (Dynamic Time warping : DTW)



computation = M × ω

ω → window length

Development of ASR systems

- Using acoustic models followed by language models
  - ✓ HMMs, GMMs, NNs, DNNs, SVM, …
- End-to-end models
  - ✓ DNNs (LSTMs, BLSTM, CNN, …

Issues in developing acoustic models

- ✓ Choice of symbols
  - Phonemes : 40-50
  - Syllables (C^mVC^n) : 10^3 - 10^4
  - Words : 10^5 - 10^6
  - Phrases : > 10^6

Feature Extraction

- ✓ Spectral features
- ✓ Auto-encoder based features

Detection of symbol boundaries in speech signal

Applications of ASR Systems

- Voice interface to machines
- Speech to speech translation
- Automatic dictation systems
- Personalized applications (home appliances, mobile phones, … etc.
- Automatic form-filling
- Indexing and retrieval of speech documents
- Social data analysis (twitter, whatsup, facebook, …)
- Healthcare
- Speech coding, speech compression, speech transmission, ….

Thank You