Pattern Classification (I)

杜俊
jundu@ustc.edu.cn

Pattern Classification Problem

- Given a fixed set of finite number of classes: C_1, C_2, …, C_N.
- We have an unknown pattern/object *P* without its class identity.
- BUT we can measure (observe) some feature(s) *X* about *P:*
  - *X* is called feature or observation of the pattern *P.*
  - *X* can be scalar or vector or vector sequence
  - *X* can be continuous or discrete
- Pattern classification problem: $X \rightarrow C_i$
  - Determine the class identity for a pattern based on its observation.

Examples of Pattern Classification

- Speech Recognition:
  - Pattern: voice spoken by a human being
  - Classes: language words/sentences used by the speaker
  - Features: speech signal → a sequence of feature vectors
    - Continuous, high-dimensional vector

- Image Understanding:
  - Pattern: given images
  - Classes: all known object categories
  - Features: color or gray scales in all pixels
    - Continuous, multiple vectors/matrix
  - Examples: face recognition, OCR (optical character recognition).

Pattern Classification: Paradigm Shift

- Knowledge based
  - Reply on expert(s), small data samples
  - Simple toy problems

- Data-Driven
  - Large data samples
  - Statistical models, machine learning algorithms

- Big Data Era
  - Massive real-world data samples
  - Data intensive computing
  - Parallel/distributed platform: GPU, map-reduce

Big Data in Real Applications?

- How to Measure “Big”?

- Chinese Speech Recognition
  - 1K hours – 10K hours for training models
  - 0.2B- recording characters

- Handwritten Chinese Character Recognition
  - 1000 samples for each character class
  - The most commonly used 3755 characters
  - Totally 0.004B- character samples

- Population in China: 1.3B+

Statistical Pattern Classification

- Feature Extraction:
  - Domain knowledge
  - Varies from different objects
  - Speech, text, image, handwriting, video, gestures,…

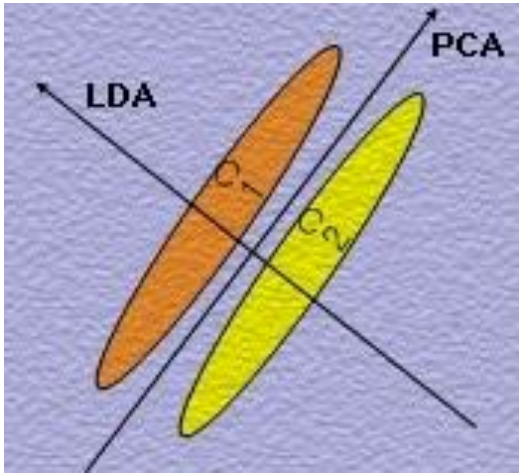
- Statistical Model Training/Learning
- Inference, Matching, Decision

} The Basic Theory

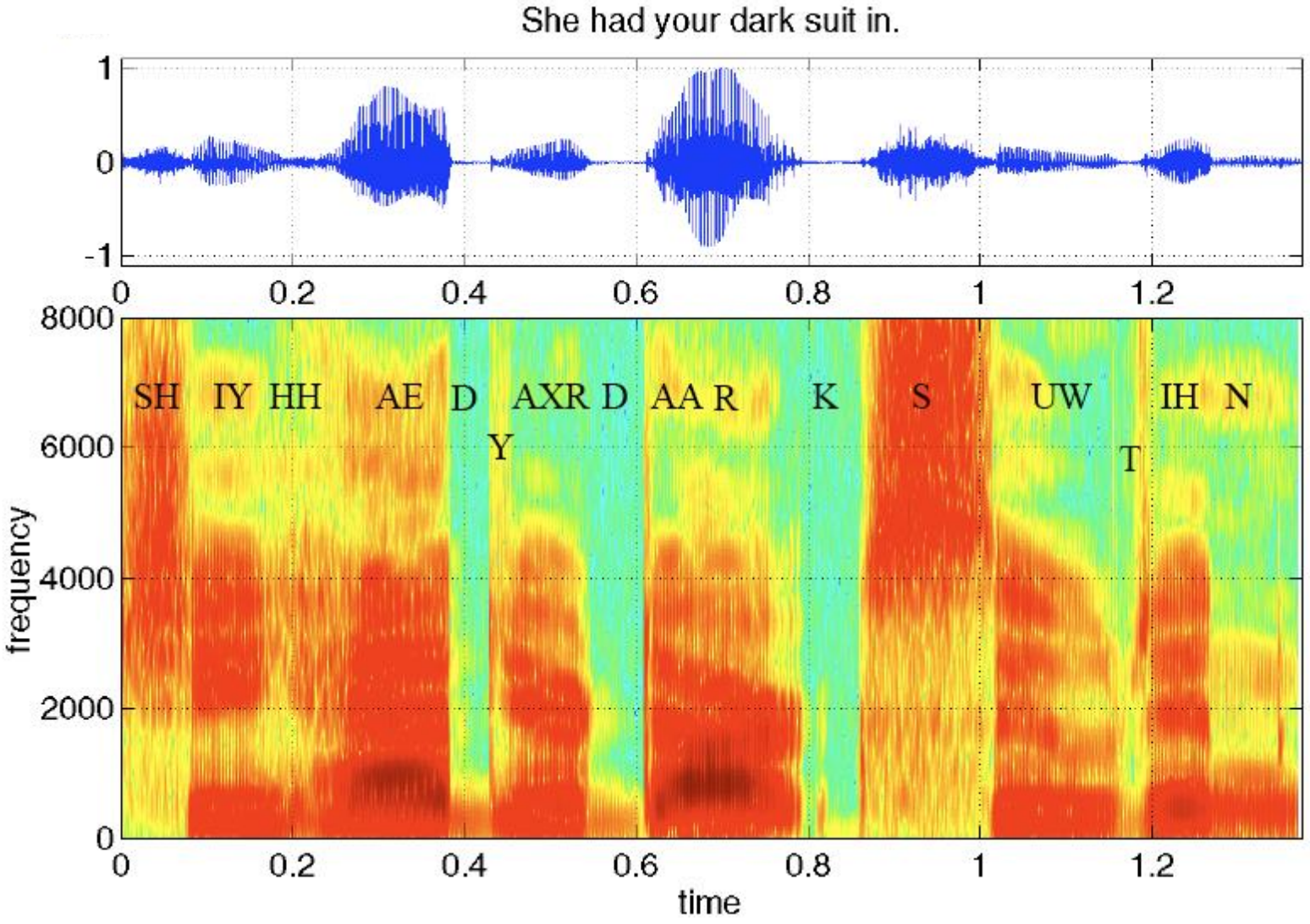
- Feature VS. Model （which is more important）

Feature Extraction

- Extraction of Feature Vector
  - Task dependent
  - Domain knowledge
  - Curse of dimensionality
  - Information redundancy and high correlation
- Dimensionality Reduction
  - Linear approaches
    - Principle Component Analysis (PCA)
    - Linear Discriminative Analysis (LDA)
  - Nonlinear approaches
    - Multi-Dimensional Scaling (MDS)
    - Stochastic Neighborhood Embedding (SNE)
    - Manifold Learning: IsoMap, Locally-Linear Embedding (LLE)
    - Neural Networks: Bottleneck



Speech Signal



Fundamental Speech Units

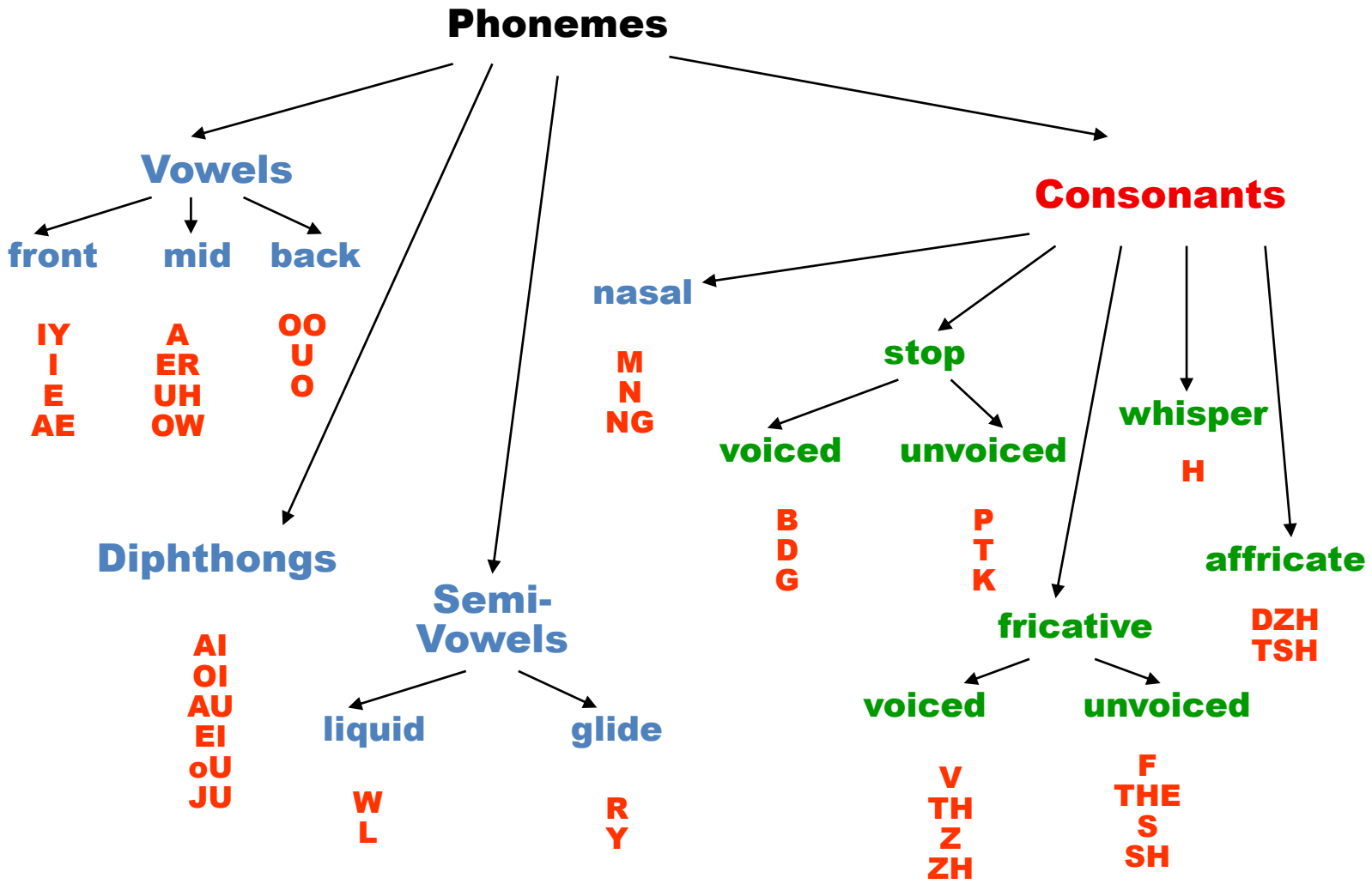
- Sentence/Utterance → Phrase → Word → Syllable → Phone
- Phone （音素）
  - Abstract name is called “phoneme” （音位）
  - Infinite number of acoustic realization
  - Monophone: context-independent phone
  - Allophone: context-dependent phone

- Other considerations:
  - Language dependency
  - Task or vocabulary dependency

- Example:

  Words: How do they turn out later
  Syllables: How do they turn out la-ter
  Phones: h aw d uh dh eh t er n aw t l ai t er

Phonemes in American English

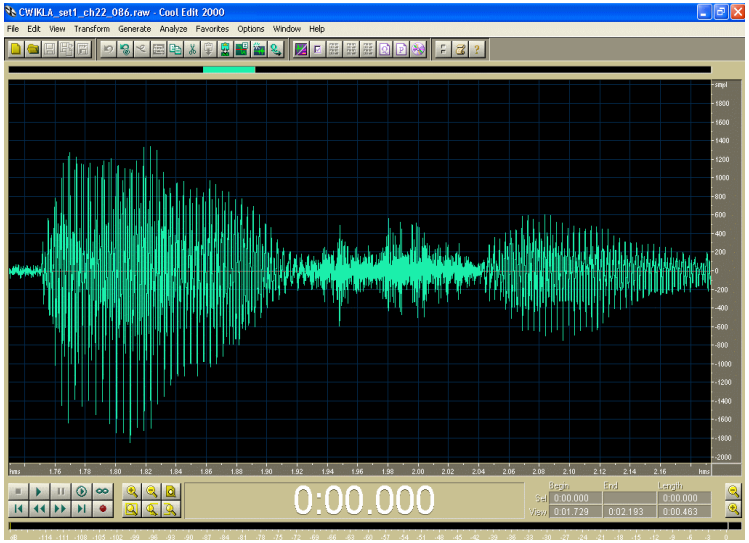


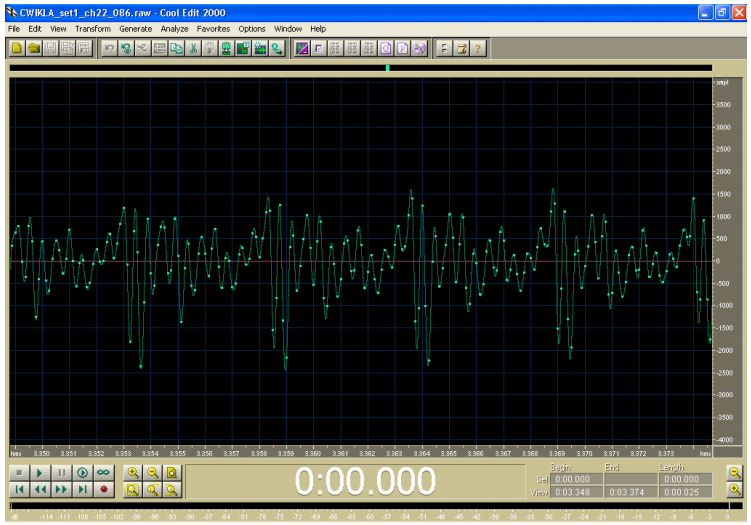
Coarticulation（协同发音）

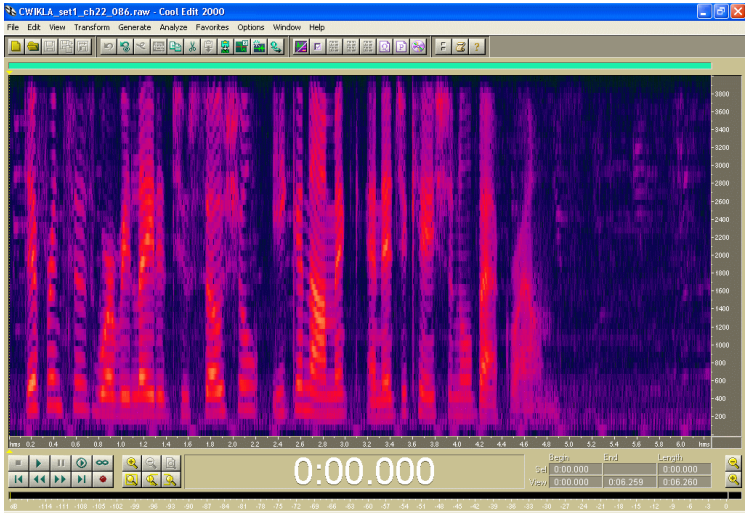
- Definition: acoustic realization of a phone is largely affected by its neighboring contexts.
- Reason: in speech production, articulatory gestures follow dynamics constrained by mechanical time constants associated with the articulator to keep the effort of muscles to a minimum.
- In speech recognition, how to model a phone:
  - Context-independent phone: monophone
  - Context-dependent phone: biphone (left and right) and triphone

Words:	*How*		*do*		*they*		*turn*		*out*		*later*			
Phones:	*h*	*aw*	*d*	*uh*	*dh*	*eh*	*t*	*er*	*n*	*aw*	*t*	*l*	*ai*	*t*	*er*
Monophone							*t*				*t*			*t*	
Biphone(left)						*eh-t*				*aw-t*			*ai-t*		
Biphone(right)							*t+er*				*t+l*			*t+er*	
Triphone						*eh-t+er*				*aw-t+l*			*ai-t+er*		

Acoustic Realization: Speech Waveform



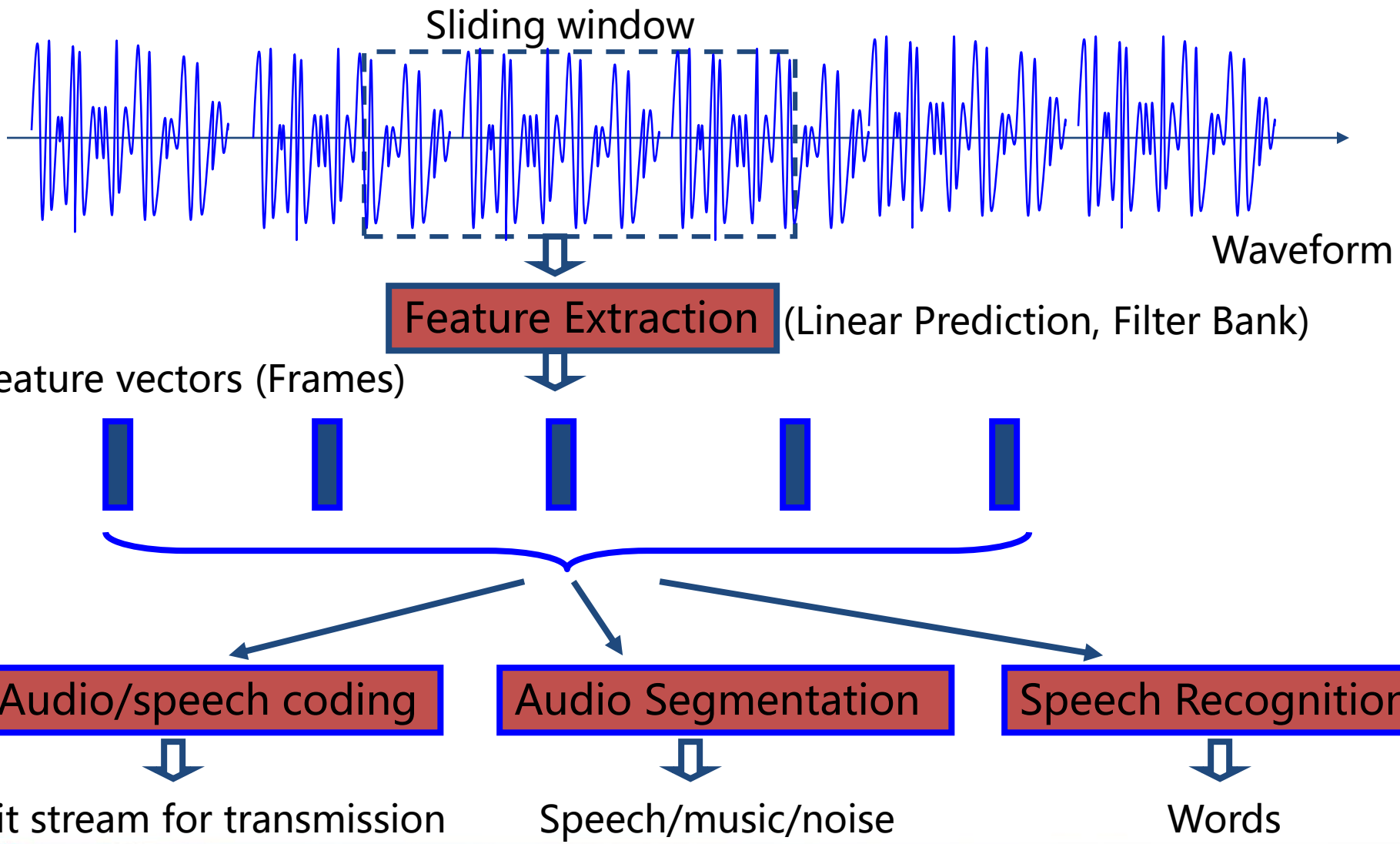




Speech Waveform: Digital Form

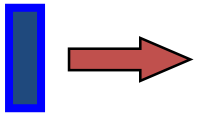
```
.....
1400:   -529  -405  -601 -1038    51   323  -324   115   698   465
1410:    485   251   166   433  -346  -908  -303  -414  -773  -475
1420:     65   406   672   566  1160  1000  -354   519   417  -702
1430:   -728  -487  -769  -511  -719  -811   227   149  -130   476
1440:    726   439   556   273   175    49  -718  -733  -363  -661
1450:   -754   -11   318   684   782  1088   999  -108   559   409
1460:   -704  -789  -509  -833  -735  -762  -712   205    80   -88
1470:    576   847   390   552   369   170  -193  -833  -719  -481
1480:   -739  -707   143   408   811   888  1321   685  -101   815
1490:     33  -963  -795  -498  -966  -741  -809  -456   399    66
1500:     -5   817   892   294   496   279    -9  -696  -820  -698
1510:   -534  -753  -254   392   757   985  1265  1187  -266   657
1520:    517  -887 -1134  -406  -830  -987  -568  -691   239   424
1530:     15   507  1212   474   325   435   -24  -784  -741  -812
1540:   -653  -532  -278   240   982   999  1221  1196  -463   630
1550:    500 -1023 -1331  -298  -819 -1110  -597  -520   344   443
1560:     49   526  1297   406   184   367  -438  -883  -589  -949
1570:   -704   -90   -74   261  1413  1188  1332   292  -234   895
1580:   -213 -1468 -1065  -191 -1017  -838  -640    20   688   379
1590:    157   941  1170   194    88  -313  -689  -674  -952  -938
1600:   -124   257   -30  1089  1539  1506   545  -636   687   269
1610:  -1439 -1751  -253  -534 -1033  -691   -74   862   709   156
1620:    555  1408   382  -249  -600  -476  -632 -1063  -938   -96
1630:    548    57   902  1527  1922   309  -874   618   315 -1606
1640:  -1961  -326  -416  -789  -673   118   970   917   256   494
1650:   1231   439  -591  -940  -278  -724 -1031  -728   223   613
1660:    420  1039  1578  2126  -260 -1047   674    16 -2048 -1771
.....
```

Feature Extraction: Short-Time Analysis



Feature Extraction: Feature Vector

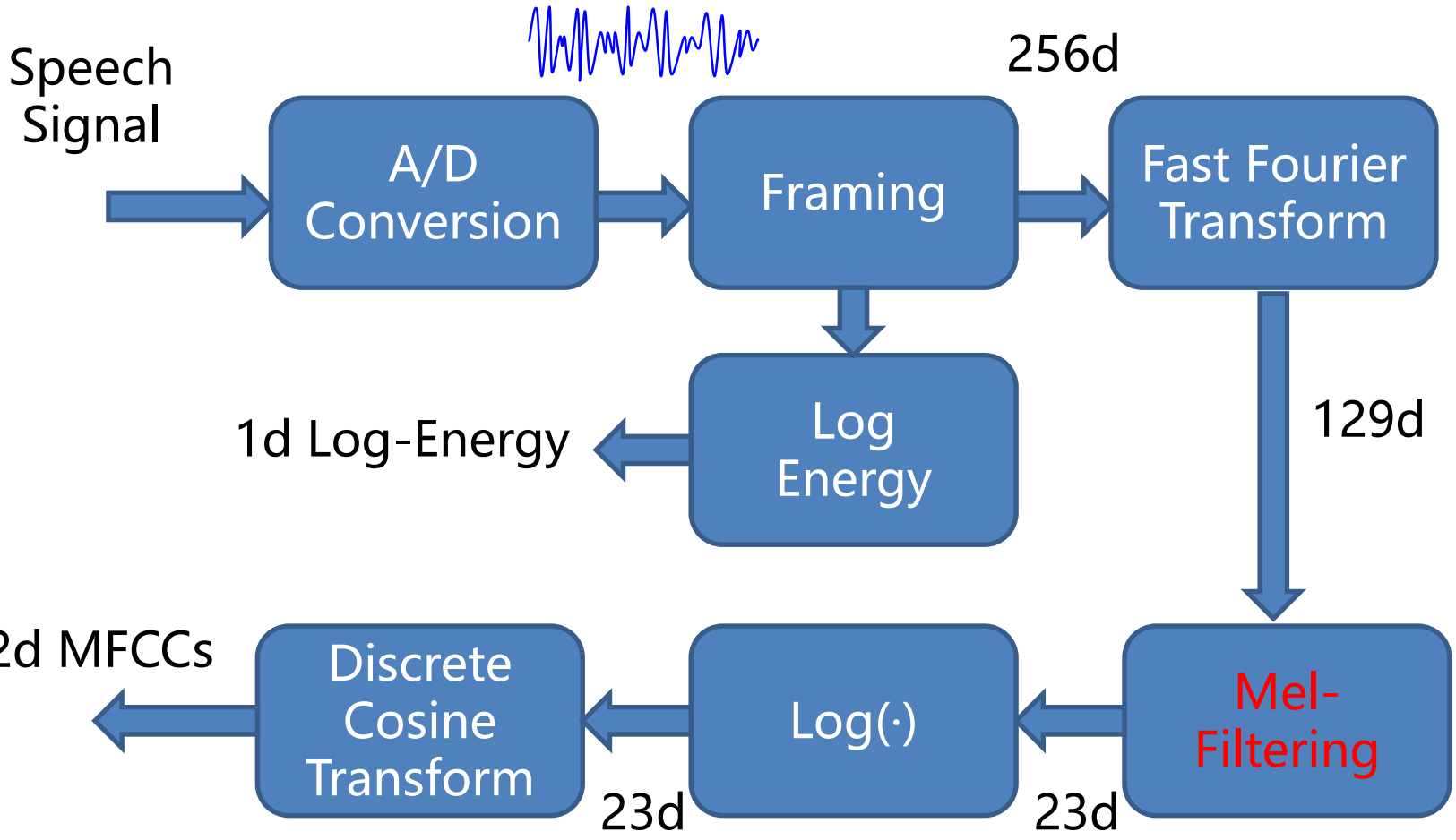
- A typical setting in most speech recognition system (a 39-dimensional vector)
  - Static feature
    - MFCCs(12d) + Log-Energy (1)
  - Dynamic feature
    - Delta(13d) +Delta-Delta(13d)



-12.520
-6.378
-9.335
-13.065
-13.997
-8.246
4.866
8.722
-4.418
0.149
-12.092
-0.341
0.814
0.434
3.185
2.058
-2.153
-1.276
1.346
-1.841
-3.689
0.826
-1.413
-0.378
2.650
-0.056
-0.550
0.352
0.023
1.102
1.315
0.649
-0.787
-1.324
-0.189
0.251
0.870
0.056
-0.016

MFCC: Mel-Frequency Cepstral Coefficients

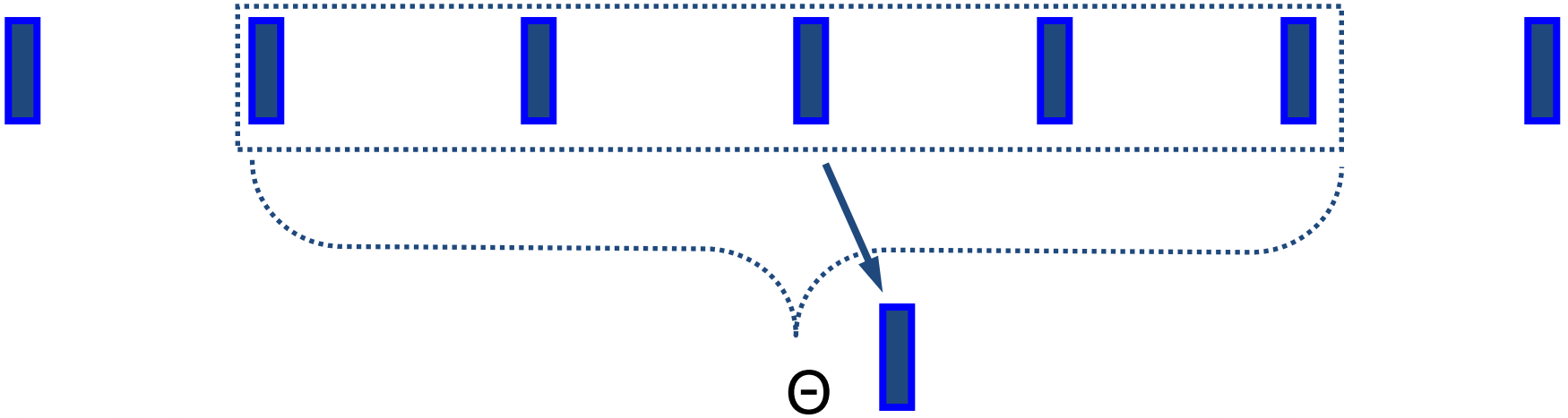
Static Feature Extraction



Dynamic Feature Calculation

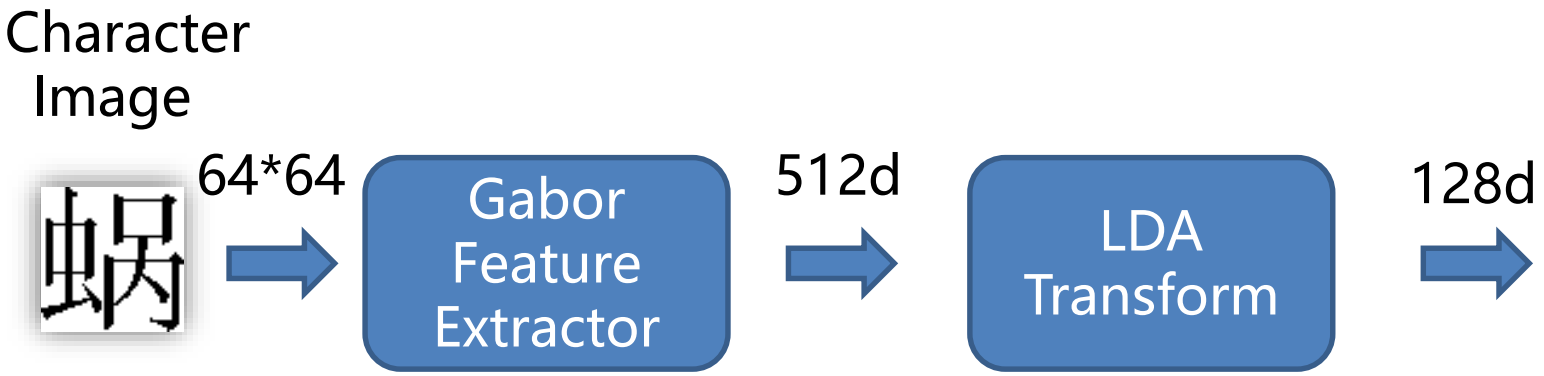
Delta coefficients : difference of static feature among consecutive frames
Delta-Delta coefficients : difference of delta among consecutive frames

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$



OCR: Feature Extraction

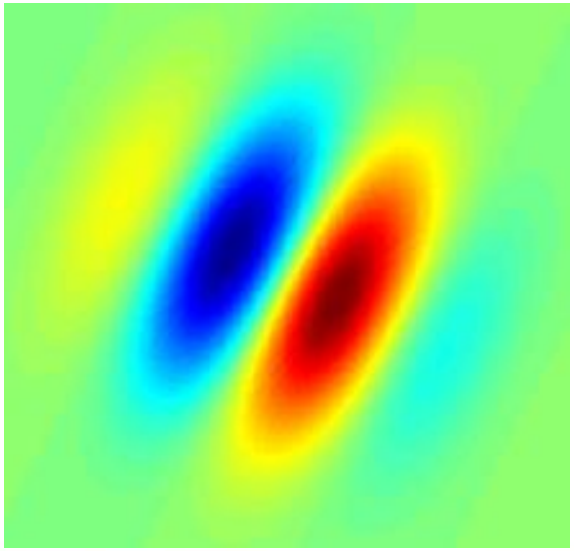
- An Example for Chinese OCR

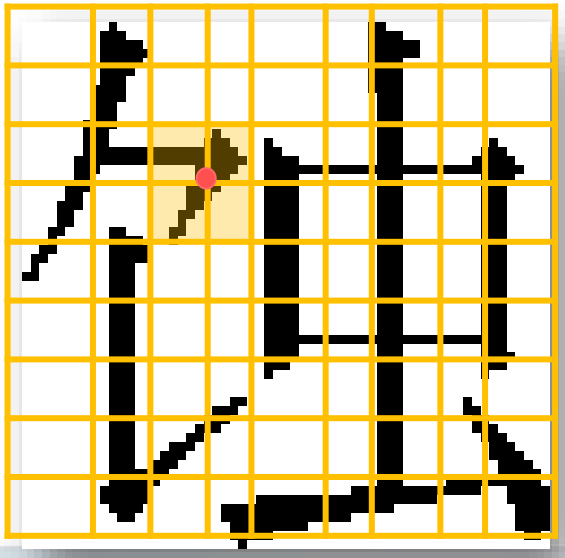


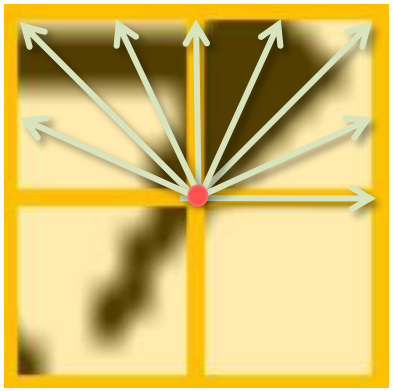
Gabor Features

- Uniformly choose 8 x 8 image sampling points
- At each point compute Gabor filter output in 8 directions
- Using 8 x 8 sampling points on 64x64 image we will have 8 x 8 x 8 = 512 dimensional feature vector

Frequency and orientation representations of Gabor filters are similar to those of the human visual system





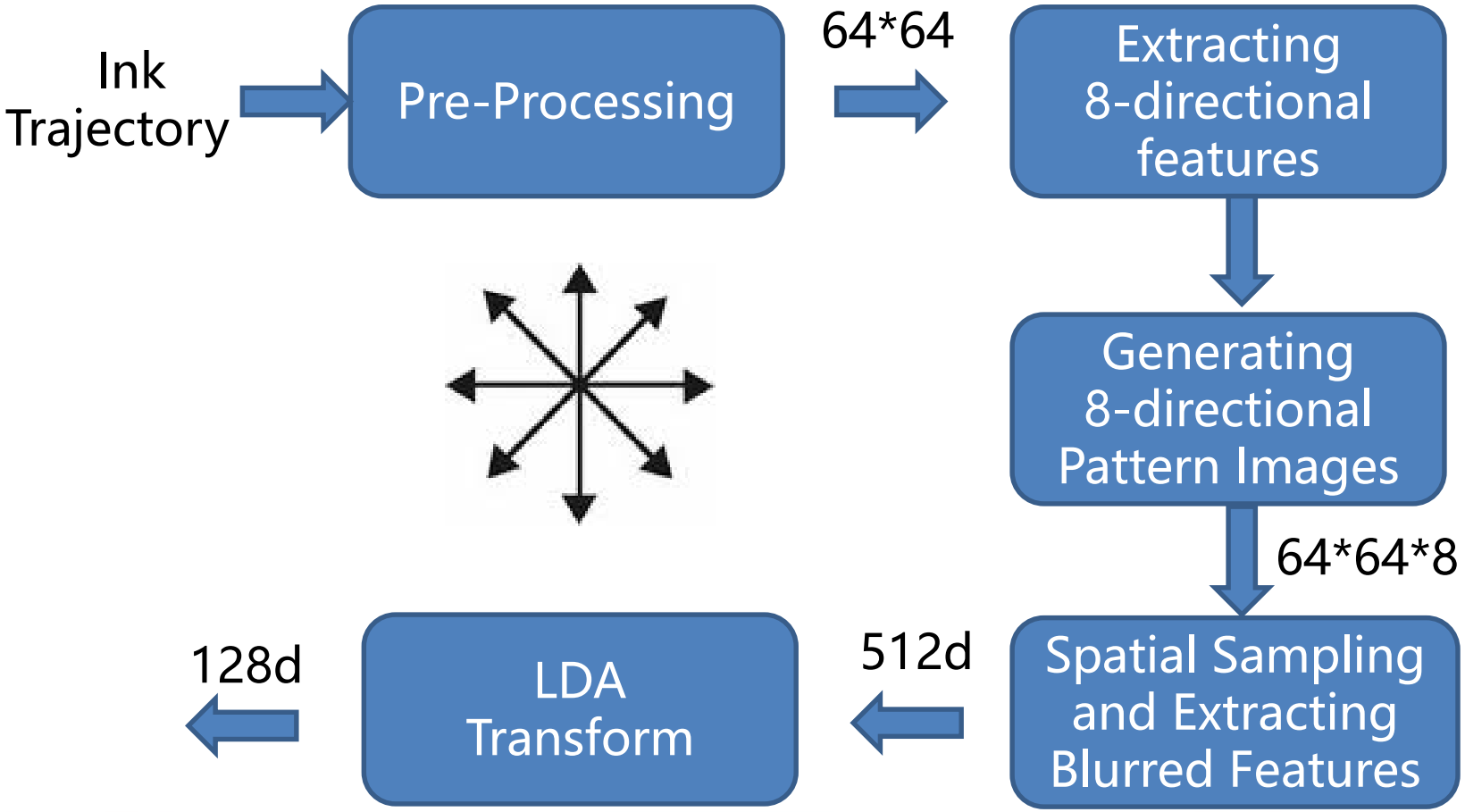


LDA (Linear Discriminant Analysis)

- Reduce the dimension of input feature vector
- LDA matrix is computed in a way to minimize within class feature variations and to maximize between class feature variations

Handwriting Recognition: Feature Extraction

- An Example for Handwritten Chinese Character Recognition



References for Feature Extraction

- “Speech Signal Representations”, Chapter 6 in Spoken Language Processing Book (for MFCC extraction)

- M. Lades et al., “Distortion invariant object recognition in the dynamic link architecture,” *IEEE Trans. on Computer*, Vol. 42, No. 3, pp.300-311, 1993. (for Gabor filter)

- Z.-L. Bai, Q. Huo, “A study on the use of 8-directional features for online handwritten Chinese character recognition,” *Proc. ICDAR*, 2005, pp.262-266. (for feature extraction of Chinese handwriting recognition)