Phoneme Boundary Refinement Using Ranking Methods

Hung-Yi Lo and Hsin-Ming Wang

ISCSLP 2010

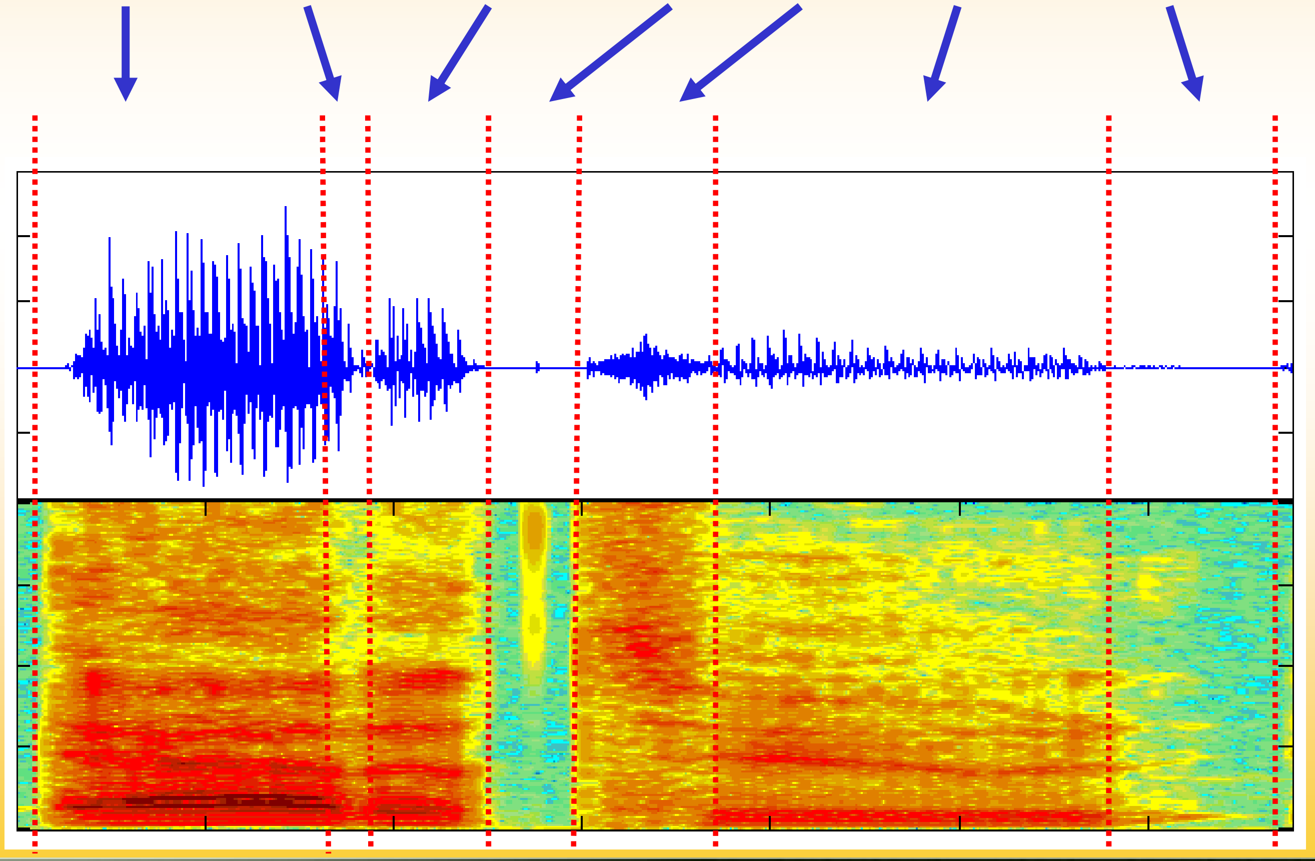
Dec. 01, 2010

Spoken Language Group
Institute of Information Science
Academia Sinica, Taipei, Taiwan
http://sovideo.iis.sinica.edu.tw/SLG/index.htm

Automatic Phoneme Alignment Problem

Context attitude In "A good attitude is unbeatable"

Phonetic Transcription ae dx ax tcl t ux dcl



Acoustic Signal

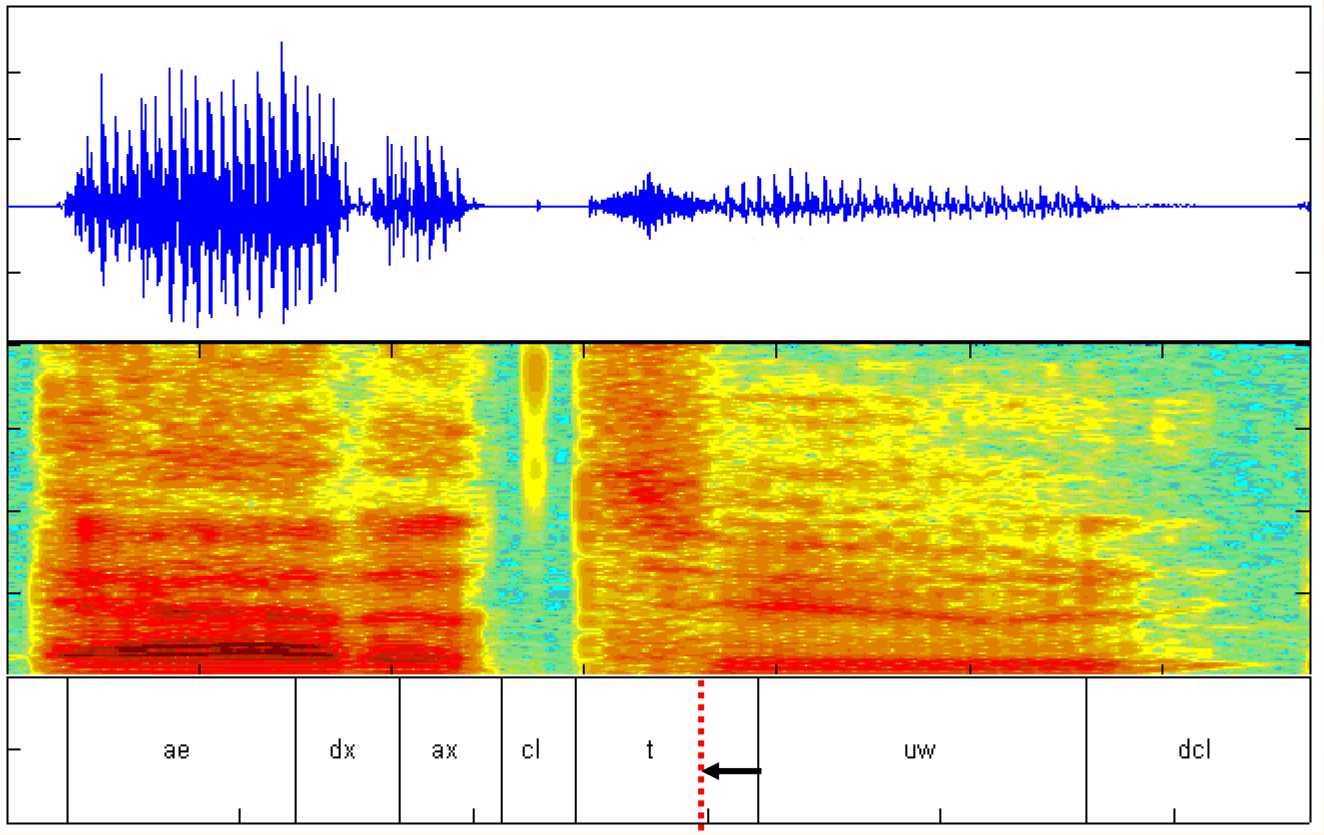
Spectrogram

HMM/SVM-based Two-stage Framework

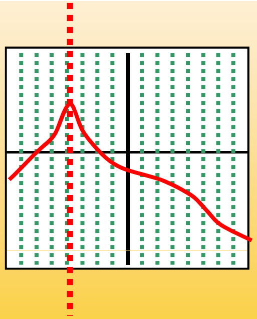
Acoustic Signal

Spectrogram

Boundaries detected by HMM-based Forced Alignment



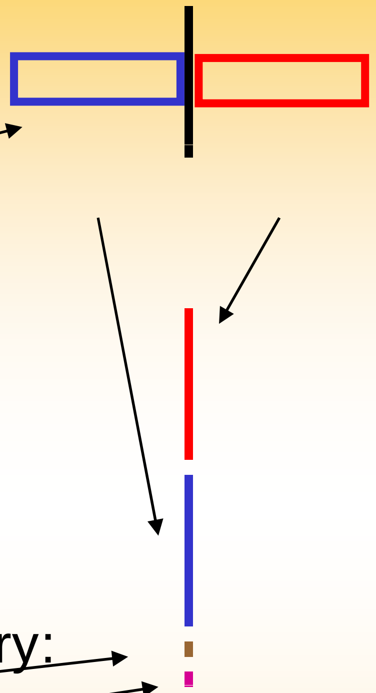
Hypothesized Boundaries as Input to SVM



Prediction Score of Boundary Classifier

Feature Extraction

- Each frame represented by a 45-dim vector:
  - 39 MFCC-based coefficients
  - Zero crossing rate
  - Bisector frequency
  - Burst degree
  - Spectral entropy
  - General weighted entropy
  - Subband energy

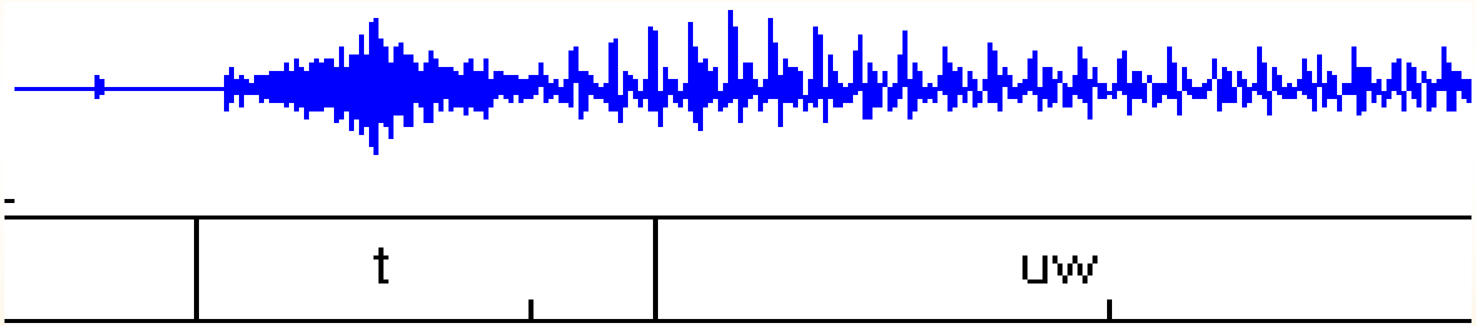


- Two features are extracted for each hypothesis boundary:
  - Symmetrical Kullback-Leibler distance
  - Spectral feature transition rate

- Each hypothesized boundary is represented by a 92-dim

Training Data for Boundary Classifier

- Positive samples: the feature vectors associated with the true phone boundaries
- Negative samples: the randomly selected feature vectors at least 20ms away from the true boundaries

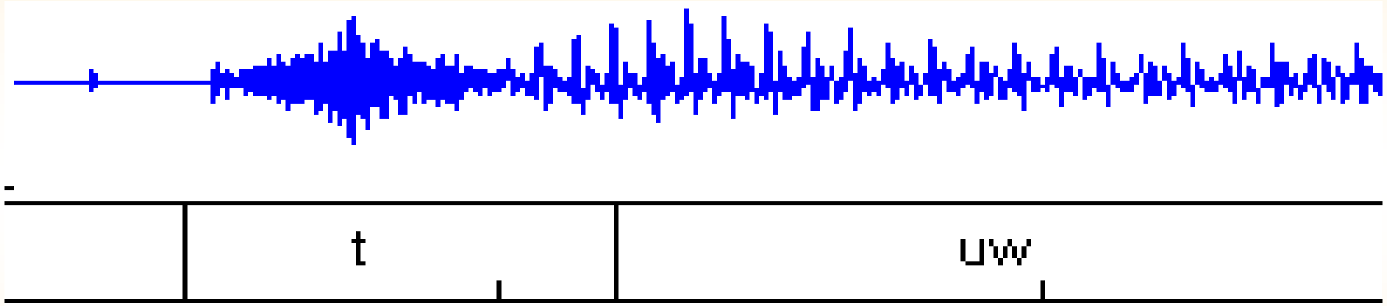


Cla

However, this classification-based method has two drawbacks

First Drawback: Losing Information (1/2)

- Only information about the boundary and far away non-boundary signal characteristics is used
  - What about the information nearby the boundary?



Classification Model

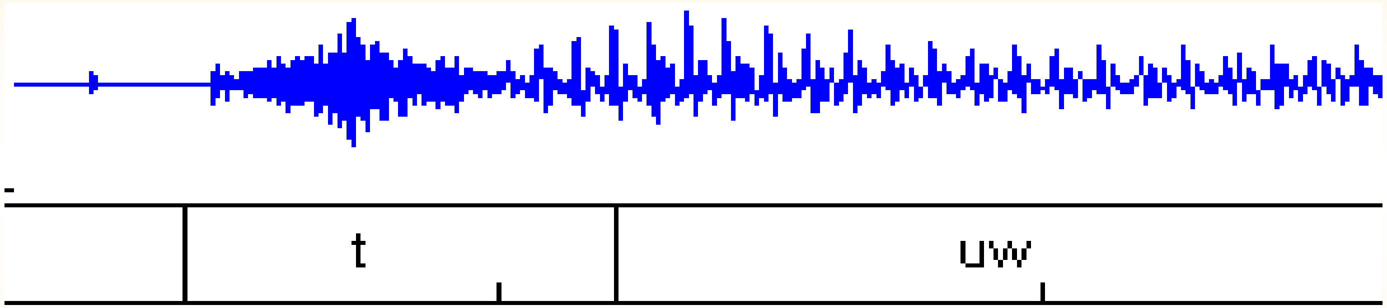
Negative Instance

Positive Instance

Negative Instance

First Drawback: Losing Information (2/2)

- Preference ranking
  - Instances extracted from the true boundaries: high preference
  - Nearby instances: medium preference
  - Far away instances: low preference



Classification Model

Negative Instance

Positive Instance

Negative Instance

Preference Ranking

Low Mid High Mid Low

Second Drawback : Imbalanced Training

- A lot of negative instances but only a limited amount of positive instances

- General classification algorithms will be biased to predict all instances to be negative
  - Since they are learned to minimize the number of incorrectly classified instances

Boundary Refinement as a Ranking Problem

- Learn a function H: X → R where $H(x_i) > H(x_j)$ means that instance x_i is preferred to x_j

- The hypothesized boundary closed to the true boundary should have higher score

- We only care about relative order
  - Correct order: A-B-C-D
  - OK: {A:-100, B:-10, C: 0, D:1000}
  - OK: {A: 0.1, B: 0.3, C: 0.4, D: 0.41}

- We exploit two learning-to-rank methods:
  - Ranking SVM
  - RankBoost

Ranking SVM

- Optimization problem:

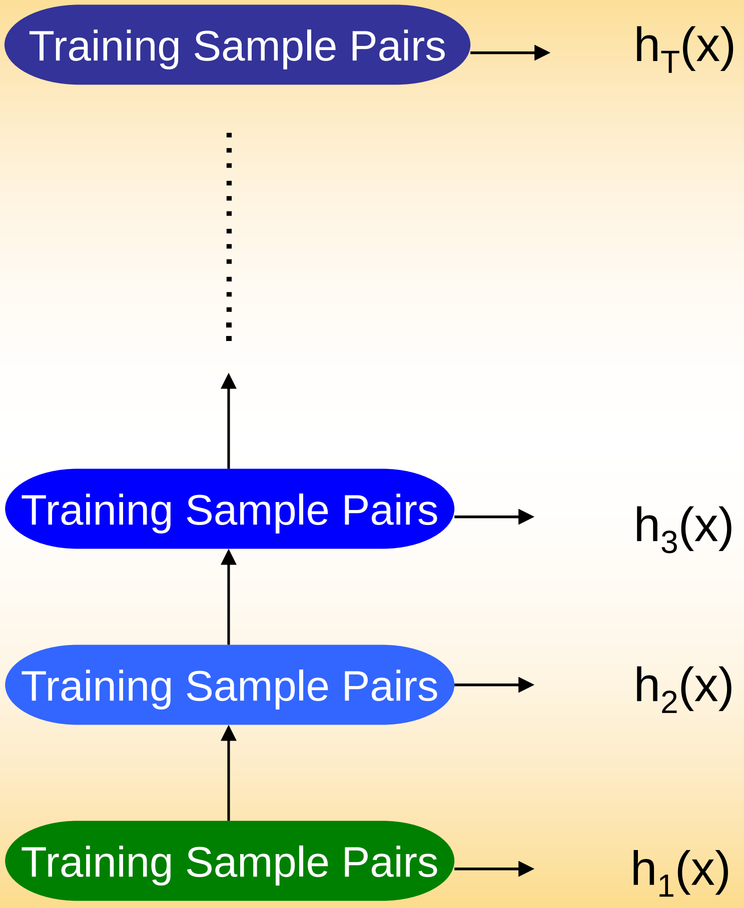
$$\min_{(w,\xi)\in R^{n+l}} \quad \frac{1}{2}w'w + C\sum_{i,j}\xi_{ij}$$

$$\textbf{s. t.} \quad w'\phi(x_i) \geq w'\phi(x_j) + 1 - \xi_{ij} \;,$$

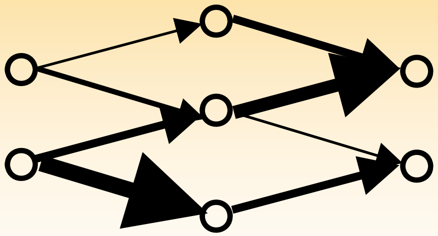
$$\xi_{ij} \geq 0$$

- The training instances are given in ordered pairs
  - $x_i \succ x_j$ means that x_i should be ranked higher than x_j

RankBoost



Weight on Instance Pairs *D_t*



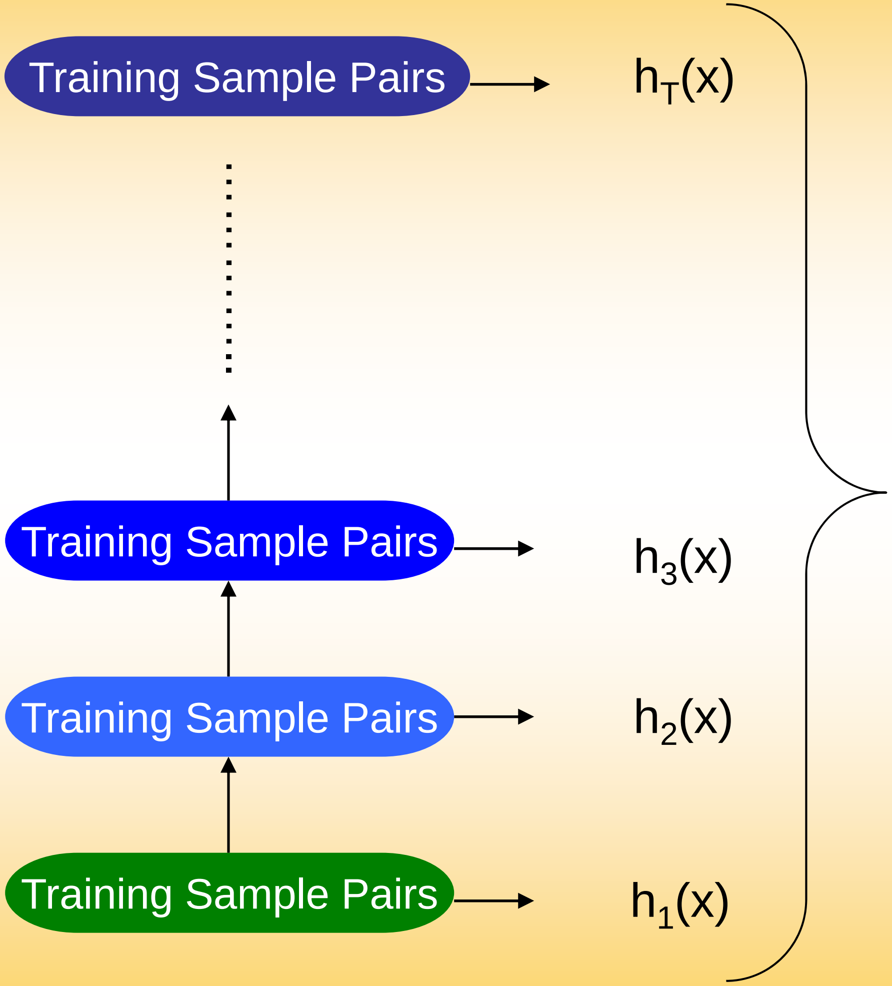
Minimize Ranking Loss:

$$\sum_{x_i, x_j} D_t(x_i, x_j) e^{\alpha_t (h_t(x_j) - h_t(x_i))}$$

Data Weight Update Rule:

$$D_{t+1}(x_i, x_j) = \frac{D_t(x_i, x_j) e^{\alpha_t (h_t(x_j) - h_t(x_i))}}{Z_t}$$

RankBoost

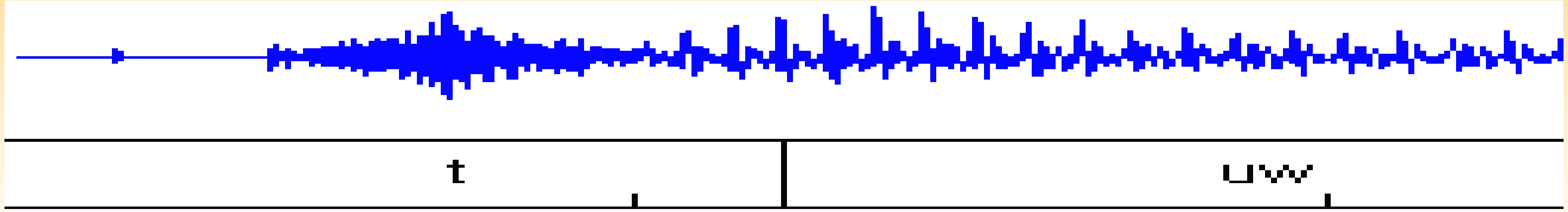


Final Classifier:

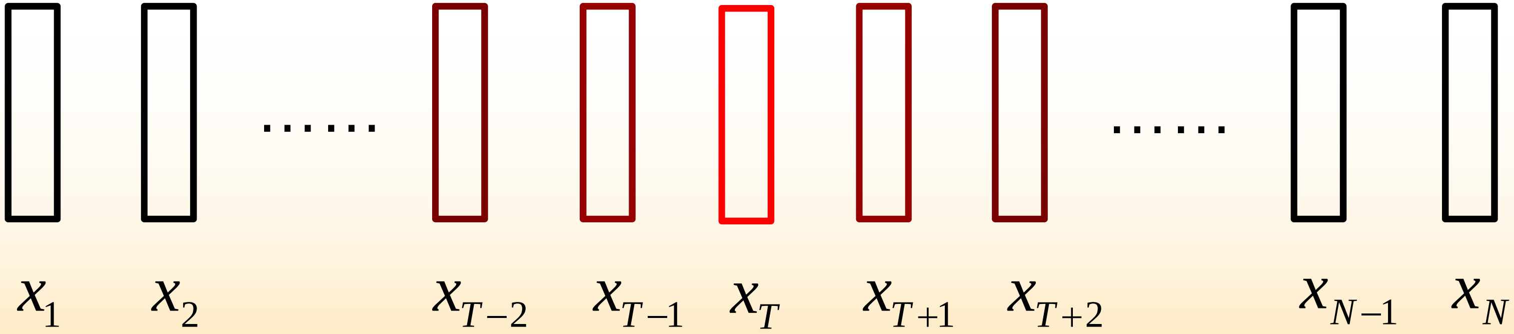
$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

Generation of Training Pairs

- Four ordered ranking lists generated from each true boundary

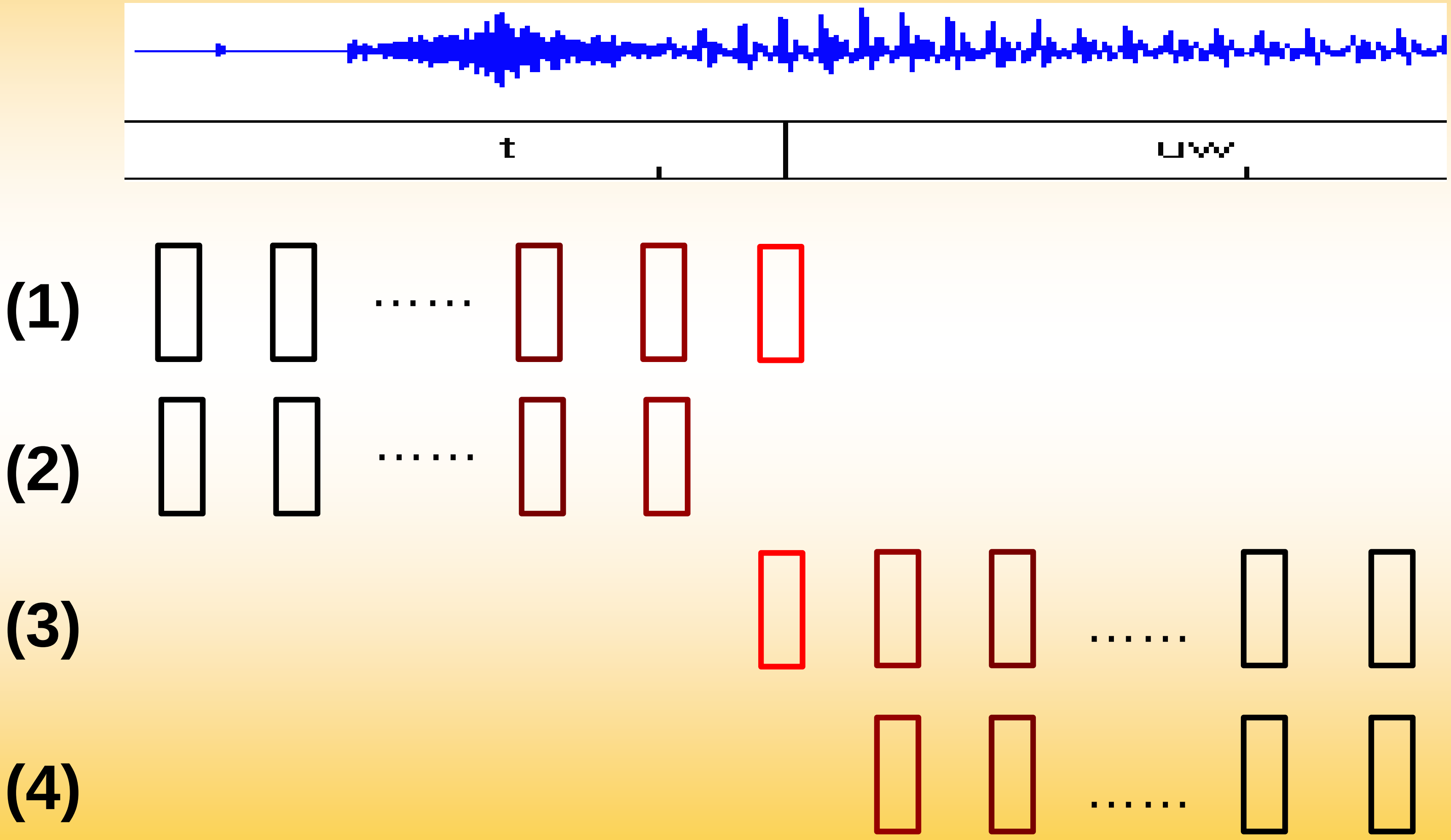


True Boundary



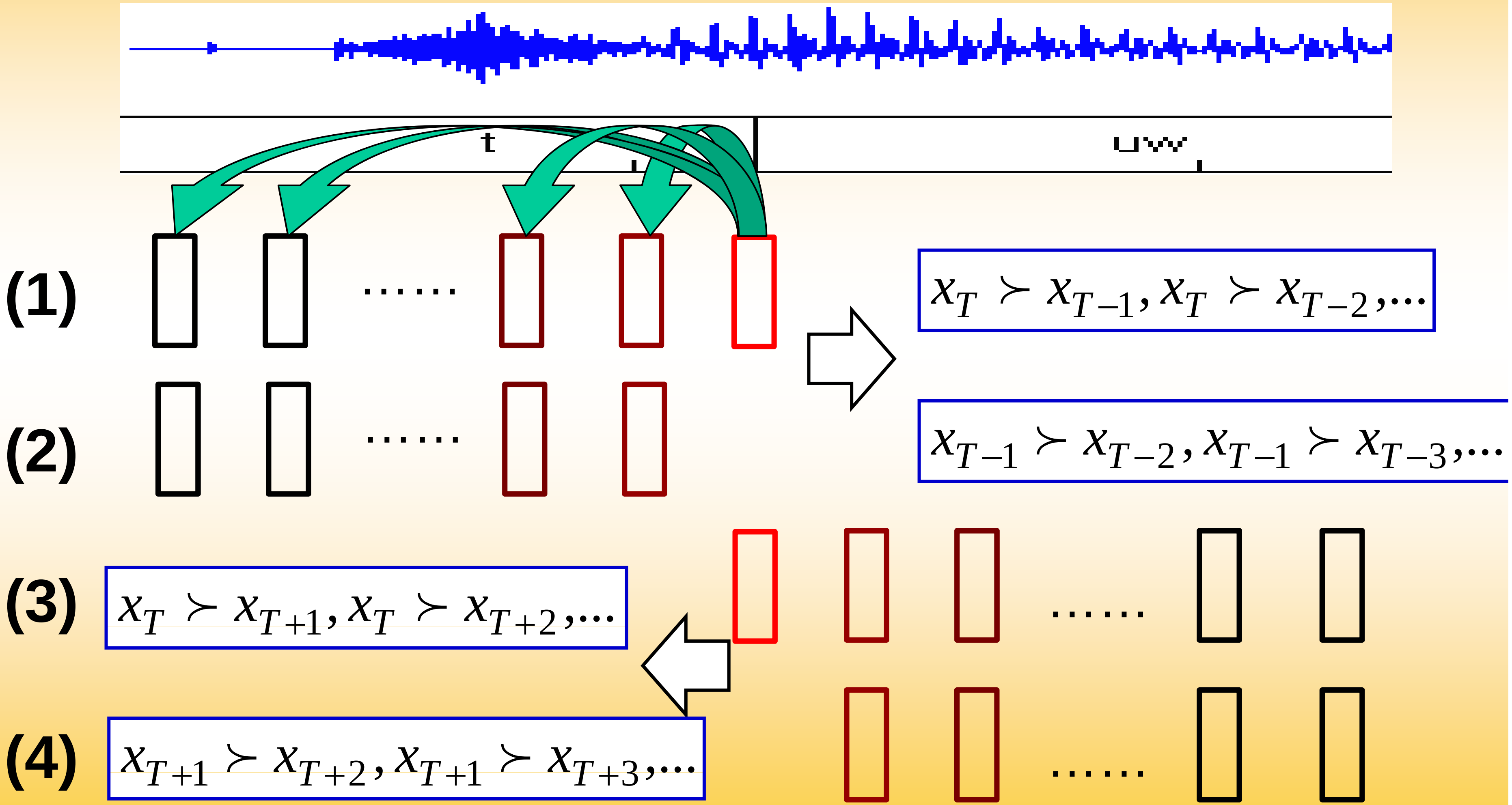
Generation of Training Pairs

- Four ordered ranking lists generated from each true boundary



Generation of Training Pairs

- Couple the preferred instances with each remaining instance



(1) $x_T \succ x_{T-1}, x_T \succ x_{T-2}, \ldots$

(2) $x_{T-1} \succ x_{T-2}, x_{T-1} \succ x_{T-3}, \ldots$

(3) $x_T \succ x_{T+1}, x_T \succ x_{T+2}, \ldots$

(4) $x_{T+1} \succ x_{T+2}, x_{T+1} \succ x_{T+3}, \ldots$

Phone-Transition-Dependent Ranker

- Training data is always limited
  - Cannot train a ranker or classifier for each type of phone transition
- Many phone transitions have similar acoustic characteristics, we can partition them into clusters
- The phone transitions with little training data can be covered by the rankers or classifiers of the categories they belong to
- Two methods for phone transition clustering:
  - K-means-based Clustering (KM)
  - Decision-tree-based Clustering (DT)

Experimental Setup

- TIMIT corpus (dialect sentences are excluded)
  - Training set: 3696 utterances
  - Testing set: 1312 utterances
- Initial segmentation by HMM-based forced alignment
- In the refinement phase, 5 hypothesized boundaries extracted every 5 ms around the initial boundary within ±10 ms will be examined by Ranking SVM and RankBoost

Experiment Results

Method	Mean Boundary Distance (ms)	% Correctness	
		<10ms	<20ms
HMM	7.14	81.57	93.73
Linear SVM$_{KM}$	6.84	83.51	93.85
Linear SVM$_{DT}$	6.89	83.44	93.79
RBF SVM$_{KM}$	6.75	84.00	94.33
RBF SVM$_{DT}$	6.83	83.70	94.12
Linear RankSVM$_{KM}$	6.62	83.89	94.17
Linear RankSVM$_{KM}$	6.76	83.90	94.01
RankBoost$_{KM}$	6.66	84.20	94.14
RankBoost$_{KM}$	6.66	84.13	94.11

Conclusion

- We have presented a ranking-based boundary refinement approach to refine the hypothesized phone boundaries given by the HMM-based Viterbi forced alignment
- We have described how to generate the training instance pairs for training the ranking SVM and RankBoost
- The experiment results on the TIMIT corpus show that the proposed ranking-based approach outperforms the conventional classification-based approach

Thank you!