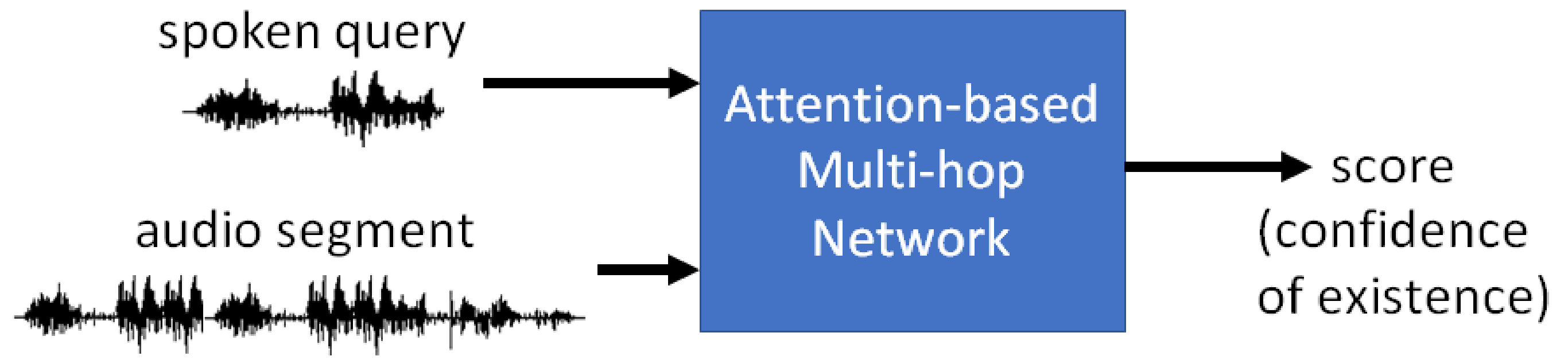
Query-by-Example Spoken Term Detection Using Attention-Based Multi-Hop Networks

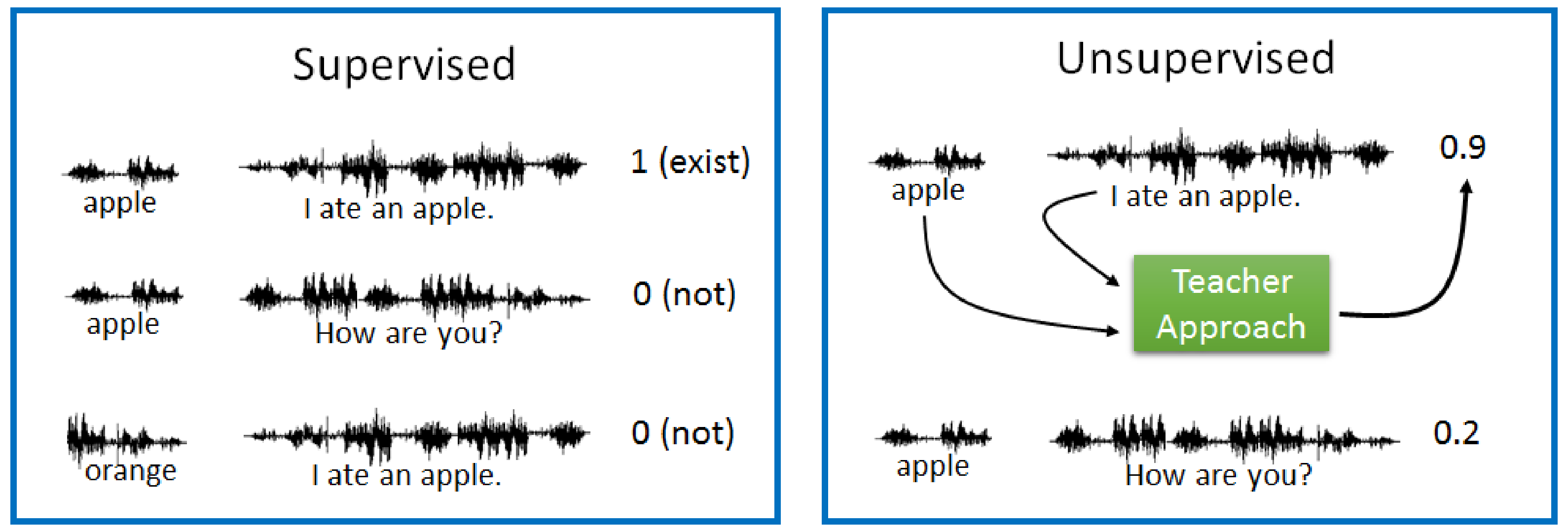
Chia-wei Ao, Hung-yi Lee,
National Taiwan University

1. Introduction

- Task: query-by-example spoken term detection
  - Given a spoken query, detecting whether an audio segment contains the spoken query
  - Matching of signals directly on the acoustic level without transcribing them into text.
  - We propose an end-to-end attention-based multi-hop


spoken query
audio segment
Attention-based Multi-hop Network
score (confidence of existence)

- Supervised Learning
  - Given the true labels to learn, become a classification problem
- Unsupervised Learning
  - Generating the labels by teacher approach (e.g. Dynamic Time Warping)



2. Proposed Approach

- Query Representation (A)
  - Input MFCC feature sequence: X_1 , X_2 , … , X_T
  - Using LSTM encode to V_Q
- Audio Segment Representation (B)
  - Input MFCC feature sequence: X_1', X_2' , … , X_T'
  - Using the same LSTM encode each frame S_1 , S_2 , … , $S_{T.}$

  Attention mechanism:

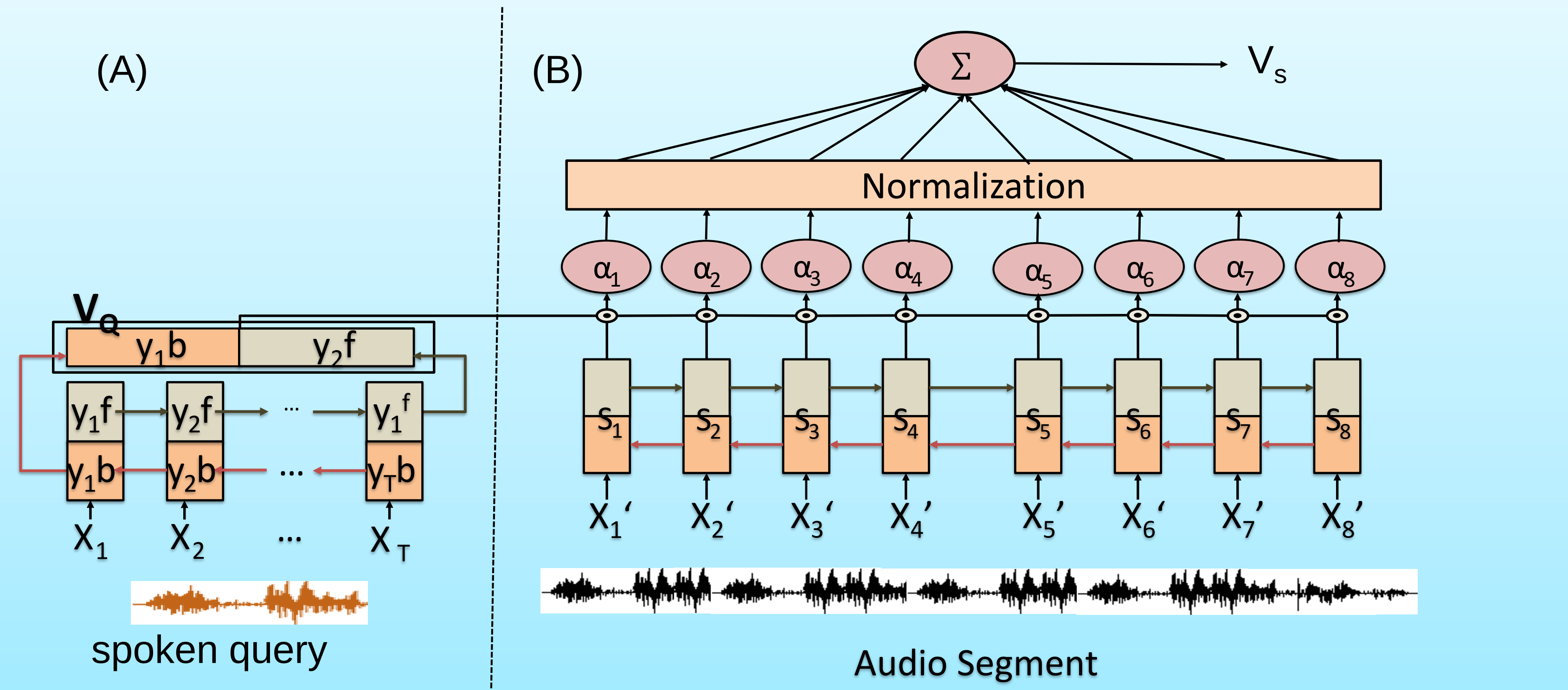
  $$\alpha_t = S_t \odot V_Q \qquad \odot\text{: Cosine simiarity}$$

  Normalization:

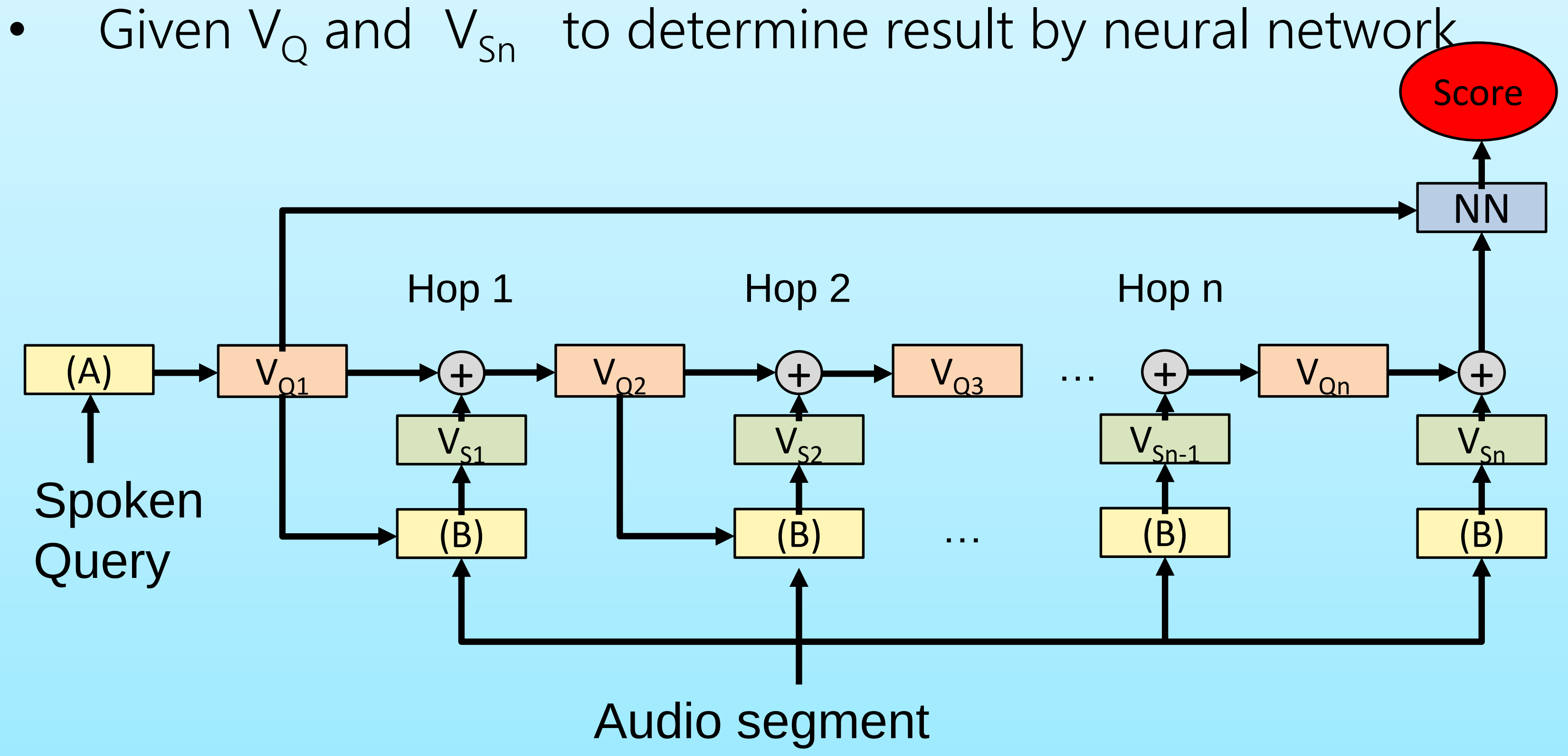
  $$\alpha_t' = \frac{\exp(\alpha_t)}{\sum_{t=1}^{T} \exp(\alpha_t)}$$

  Audio segment vector:

  $$V_S = \sum_{t=1}^{T} \alpha_t' S_t$$

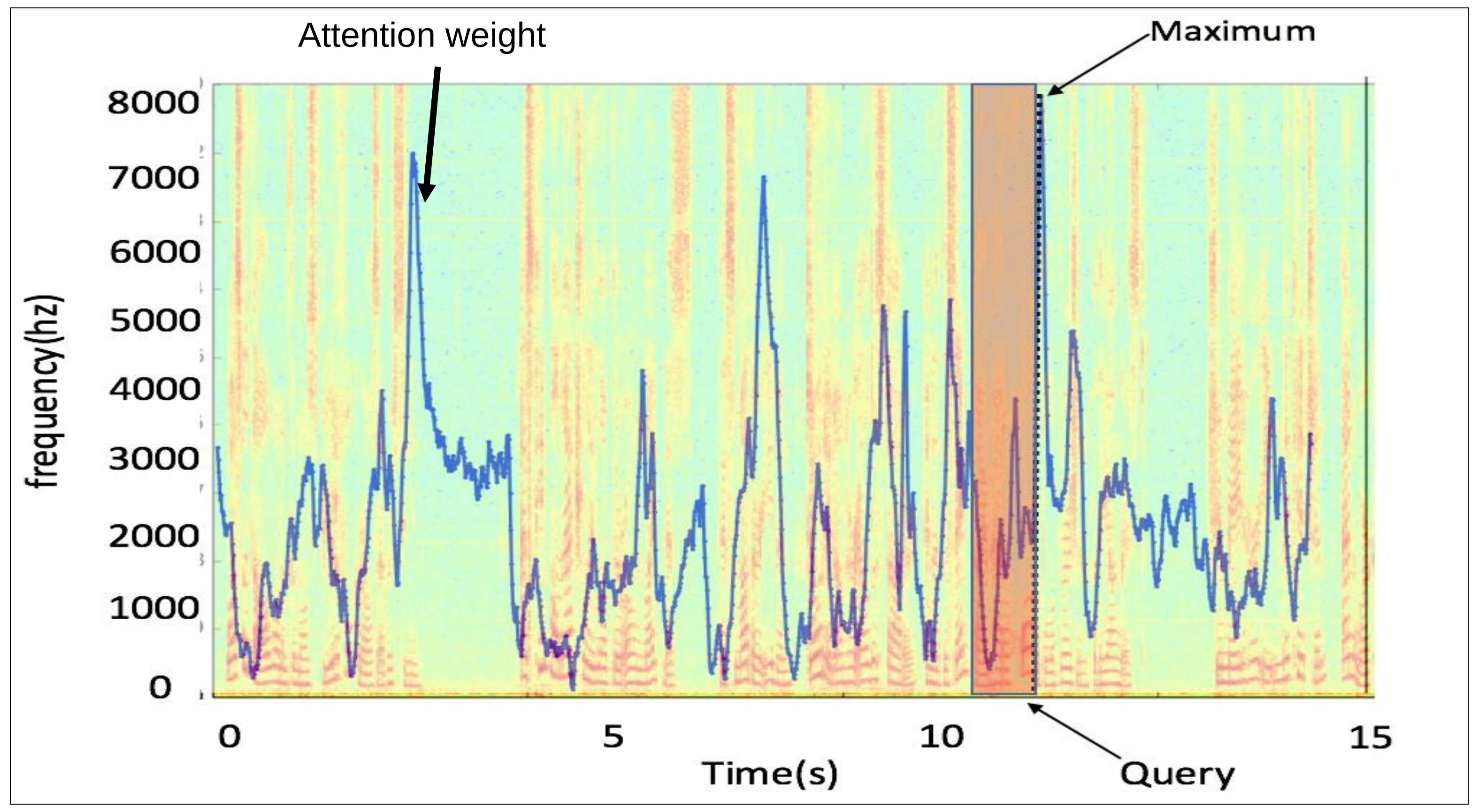

(A)
(B)
V_Q
y_1b
y_Tf
y_1f
y_2f
y_Tf
y_1b
y_2b
y_Tb
X_1
X_2
…
X_T
spoken query
Σ
V_s
Normalization
α_1 α_2 α_3 α_4 α_5 α_6 α_7 α_8
S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8
X_1' X_2' X_3' X_4' X_5' X_6' X_7' X_8'
Audio Segment

- Hopping
- Using attention mechanism repeatedly to extract more relative information from audio segment.
- Keyword Detection
- Given V_Q and V_{Sn} to determine result by neural network

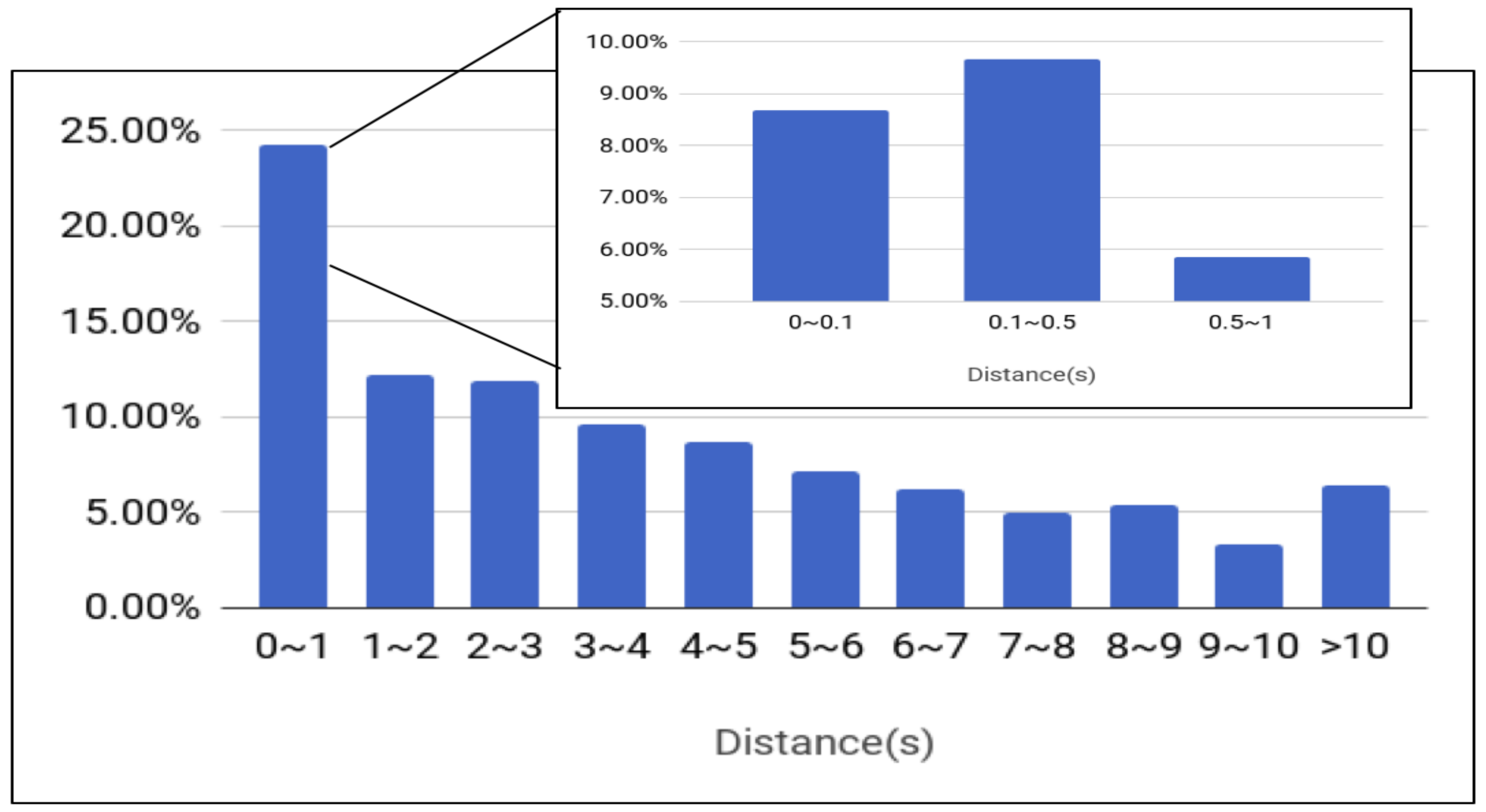


- Attention analysis

Blue line: attention value , Red box : query position.

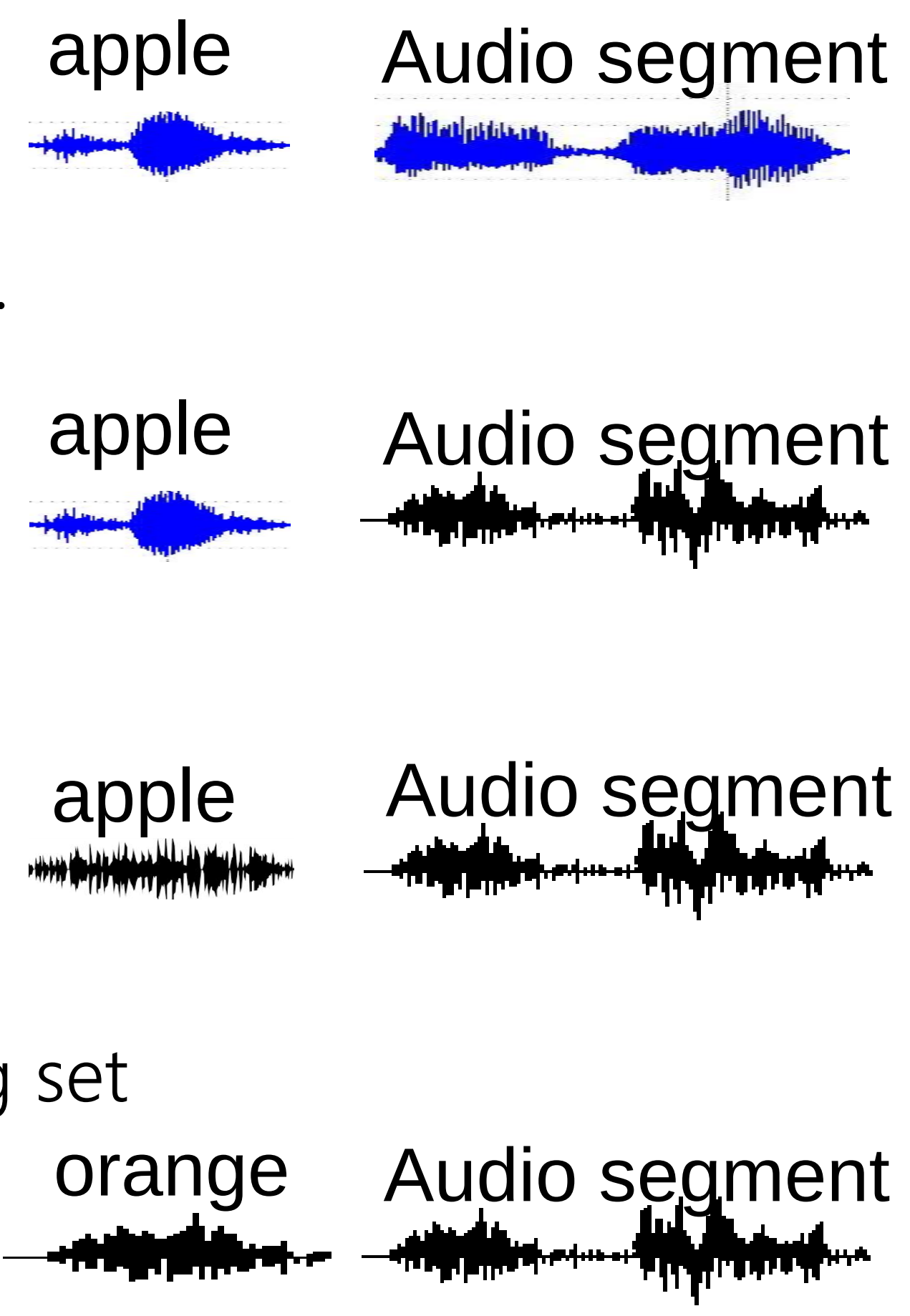


Distance: difference between the position with the highest attention weight and the end of the query



3. Experiments

- Data set : LibriSpeech
- Training set :
  - Query set : 500
  - Query and Audio segment pair : 70,000
- Testing set 1 :
  - Query acoustic feature from training set.
  - Query set : 30
  - Query and Audio segment pair : 1,500
- Testing set 2 :
  - Query Acoustic feature is different.
  - Query set : 30
  - Query and Audio segment pair : 1,500
- Testing set 3:
  - Query keyword didn't present in training set
  - Query set : 100
  - Query and Audio segment pair :10,000



		Test set1	Test set2	Test set 3
Supervised	(a): DTW	0.6173	0.5778	0.5678
	(b): 1-hop	0.6523	0.6246	0.5754
	(c): 2-hop	0.6472	0.6430	0.5842
	(d): 3-hop	0.6676	0.6404	0.5837
	(e): 4-hop	0.6417	0.6476	0.5792
	(f): (a) + (d)	0.6789	0.6430	0.5830
Unsupervised	(e): 1-hop	0.6128	0.5893	0.5548
	(g) 3-hop	0.6141	0.5964	0.5702