SpokenDoc11

YLAB@RU at Spoken Term Detection Task in NTCIR-10 SpokenDoc-2

Iori Sakamoto, Kook Cho, Masanori Morise, Yoichi Yamashita (Ritsumeikan University)

Abstract:

This paper describes improvement of the STD method which is based on the vector quantization (VQ). Spoken documents are represented as sequences of VQ codes, and they are matched with a text query to be detected based on the V-P score which measures the relationship between a VQ code and a phoneme. The matching score between VQ codes and phonemes is calculated after normalization for each phoneme in a query term to avoid biased scoring particular phonemes.

1. Objectives

-To improve the detection performance of unknown words for the spoken term detection (STD).

2. STD Method Based on Vector Quantization

- To represent spoken documents as sequences of VQ codes .

> Conventional methods represent spoken documents as sequences of sub-words, such as phonemes, to detect unknown words.

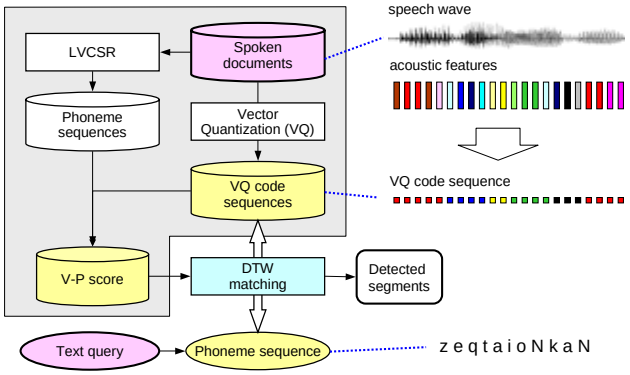

Figure 1: A flowchart of the STD method based on vector quantization.

2-1. V-P score

- The V-P score $s(v, p)$ measures the cooccurence of a VQ code v and a phoneme p.

$$s(v,p) = \log\left(\frac{C_v(p)}{N_v}\right) - \log\left(\frac{C_v(p_{best})}{N_v}\right)$$

$C_v(p)$: the number of frames which are labeled with a phoneme p and are quantized into a VQ code v
N_v : the total number of frames of v
p_{best} : the phoneme which appears most in v

- Acoustic features for VQ

> 60-dimensional parameters including 12 MFCCs of 2 preceding and 2 following frame as well as the current frame

2-2. Term detection

- To detect query term segments by continuous DTW.

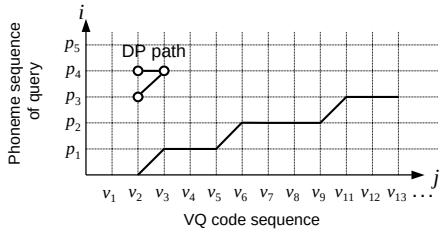

Figure 2: An example of DTW matching.

- To re-evaluate of candidates considering duration structure .

z	e	q	t	a	i	o	N	k	a	N

Figure 3: An example of false alarms for a query term 'zeqtaioNkaN', which is matched with "(watakushido)mowa'

- The distortion of phoneme duration $D(i)$ is defined based on the average frame length of a phoneme.

$$D(i) = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{d_l(p_j)}{L_l} - \frac{d_d(p_j)}{L_d}\right)^2$$

K : phoneme length of the query term
$d_l(p)$: the average frame length of a phoneme p
$d_d(p)$: the frame length of a phoneme p in the detected segment
L_l : estimated total duration of the detected term
L_d : total duration of the detected segment

- The unified score $P(i)$ is defined based on matching score and distortion of phoneme duration

$$P(i) = P_{\bar{S}(i)} - P_{D(i)}$$

$P_{\bar{S}}(i)$: matching score normalized based on statistical distribution of matching scores for all candidate segments
$P_D(i)$: the distortion of phoneme duration normalized based on statistical distribution

3. Proposed Method : Intra-phone normalization

- The matching score $\overline{N}(i)$ is normalized by averaging scores within a phoneme .

$$\overline{N}(i) = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{1}{N}\sum_{m=m_j}^{m_j+n_j-1} s(v_m, p_j)\right)$$

N : the length of candidate segment
n_j : the number of frames matching with j-th phoneme of the query term
m_j : the starting frame of the segment j-th phoneme

4. Evaluation

- Evaluation 1

Spoken documents: 177 spoken lectures in the CORE set of CSJ
Query term: 20 words (The average phoneme length is 11.0.)
VQ size: 1024

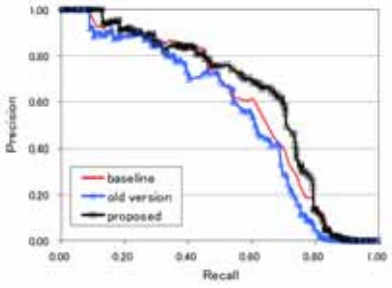

Figure 4: Recall and precision of STD methods.

Table 1: Comparison of STD methods.

method	F-measure[%]	MAP[%]
baseline	60.9	50.0
old version	59.3	61.1
proposed	65.9	67.5

- Evaluation 2 (formal run)

We used 1-best results of unmatched_syllable that were provided by the task organizer to train the V-P score definition.

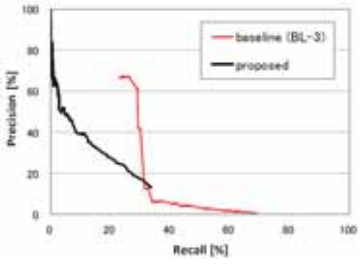

Figure 5: Recall and precision in the formal run.

Table 2: Performance of the proposed method in the formal run.

	F-measure[%] (max)	F-measure[%] (spec.)	MAP[%]
baseline(BL-3)	39.36	39.16	39.3
proposed	24.10	24.04	22.1

5. Conclusion

- The score normalization improves the STD performance by 6% of F-measure
- The proposed method shows the low performance for SDPWS data in the formal run
- To improve the performance
> To train V-P score using transcription of similar speaker