음성 질의 기반 디지털 사진 검색 기법*

김태성(ICU), 서영주(ICU), 김희린(ICU), 이용주(원광대)

<차 례>

<Abstract>

A Query-by-Speech Scheme for Photo Albuming

Taesung Kim, Youngjoo Suh, Hoirin Kim, Yong-Ju Lee

In this paper, we introduce two retrieval methods for photos with speech documents. We compare the pattern of speech query with those of speech documents recorded in digital cameras, and measure the similarities, and retrieve photos corresponding to the speech documents which have high similarity scores. As the first approach, a phoneme recognition scheme is used as the pre-processor for the pattern matching, and in the second one, the vector quantization (VQ) and the dynamic time warping (DTW) are applied to match the speech query with the documents in signal domain itself.

Experimental results show that the performance of the first approach is highly dependent on that of phoneme recognition while the processing time is short. The second method provides a great improvement of performance. While the processing time is longer than that of the first method due to DTW, but we can reduce it by taking approximated methods.

* Keywords: Query-by-speech, Spoken document retrieval, Contents-based retrieval.

* 이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업 헬스케어기술개발사업단의 지원에 의하여 연구되었음.

1. 서 론

디지털 카메라의 보급으로 사진을 찍고 저장하기가 쉬워졌다. 반면 저장된 사진의 수가 늘어날수록, 원하는 사진을 찾기가 어려워졌는데, 이를 위해 디지털 카메라의 음성메모기능을 이용하는 방법을 소개한다. 즉 사진을 찾기 위해 어떤 음성 질의어를 입력했을 때, 음성메모기능에 의해 녹음된 파일들을 검색하여, 해당 질의어를 포함하고 있다고 간주되는 음성파일들이 가리키는 사진을 검색하는 방식이다. 이를 위해서 본 논문에서 다루는 방법은 크게 두 가지이다. 첫 번째는 음소인식에 기반한 방법이고, 두 번째는 vector quantization(VQ)와 dynamic time warping(DTW)를 사용하는 방법이다.

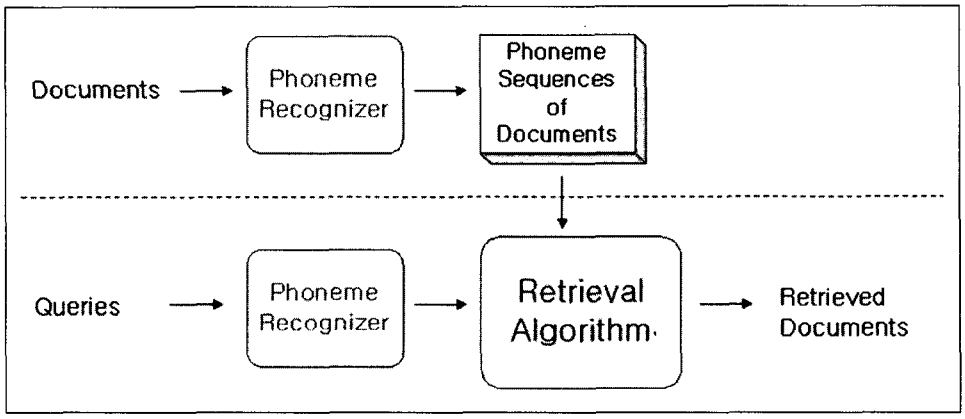
첫 번째 방법은 spoken document retrieval에서 많이 사용되는 음소인식에 기반한 방법으로 이 가운데 음성발생정보를 사용하는 방법은 [1]에서 소개되고 있다. 음소의 발생정보를 이용하여 vector space를 구성한 뒤 질의어와 음성 document 사이의 유사도를 측정한다[1][2]. 한편, 음성순서정보를 이용하는 방법은 음소열들을 dynamic programing을 통하여 유사도를 측정한다[1][2]. 두 번째 방법은 음성 질의어와 음성 docuemnt들을 vector quantization을 통해 codeword index의 sequence로 변환한 뒤 dynamic time warping을 이용하여 pattern을 비교하는 방법이다. 이는 음성 질의어나 음성 document가 어떤 내용을 가지는 지 상관하지 않고, 그 둘 사이의 유사도를 측정해서 음성 document가 음성질의어를 포함하고 있는 지의 여부를 나타낸다. 두 번째 방법에서는 dynamic time warping을 사용하기 때문에 처리시간이 길어지는 단점이 있다. 본 논문에서는 이를 극복하기 위해 Dual DTW라는 방식을 제안한다. 이는 일반적인 DTW를 사용한 경우보다 약간의 성능 감소를 가져오지만 처리시간을 반으로 줄인다.

본 논문은 2장에서 음소인식을 이용한 방법에 대해 소개하고, 3장에서 VQ와 DTW를 이용한 방법에 대해 제안한다. 4장에서는 실험방법 및 성능측정과 실험결과 등을 다루며, 5장에서는 결론을 제시한다.

2. 음소인식을 이용한 정보검색 방법

본 논문에서는 음소 인식을 이용한 방법으로 음소발생정보를 이용한 방법과 음소순서정보를 이용한 방법을 소개한다. <그림1>에 음소인식을 이용한 정보검색 시스템의 개요도가 나타나 있다. 음소인식을 이용하는 방법은 음성질의어와 음성 document들을 음소인식기를 통하여 음소열로 변환한 뒤, 둘 사이의 유사도를 측정한다. 본 논문에서 사용한 음소인식기는 46개의 음소를 인식하며, 각 음소모델의 훈련을 위해 39차의 MFCC를 사용하였고, 5states, 12mixture의 HMM으로 모델링하

였다. 여기에 음소에 대한 bigram language model을 적용하였는데, 이때 사용된 코퍼스는 한국전자통신연구원에서 만든 phonetically balanced sentence(PBS) DB이다. 음소 인식기를 훈련시키는 데에는 국어공학 연구소의 낭독음성DB 보급판 가운데 PBW(phonetic balanced word) 452어절 DB의 63명(남: 35명, 여: 28명)분의 데이터를 사용하였고, 평가에는 훈련에 사용되지 않은 7명(남: 4명, 여: 3명)분을 사용했다. 음소 인식기는 54.82%의 correction rate와 44.97%의 accuracy rate의 인식률을 가진다[3].



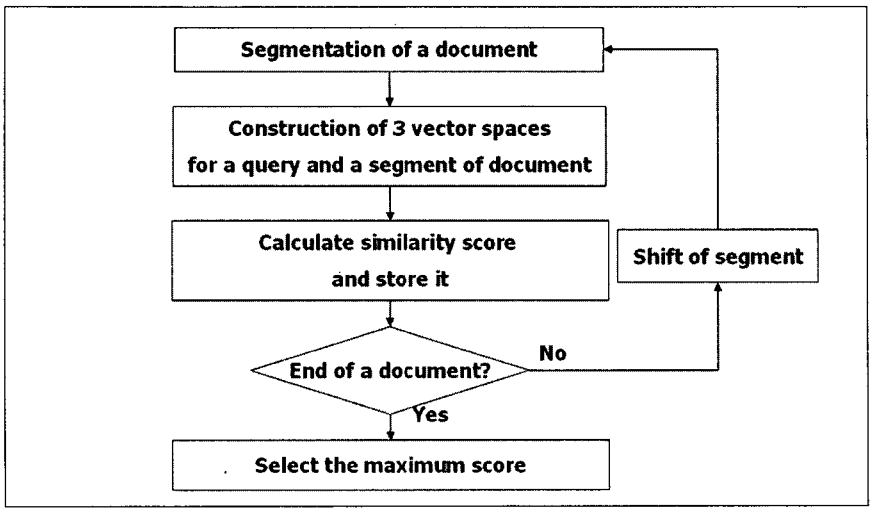
<그림 1> 음소인식을 이용한 정보검색 방법

2.1. 음소 발생정보를 이용한 방법

음소발생정보를 이용하는 방법은 <그림 2>에서 나타난 순서로 행해진다. 음성질의어와 비교 대상이 되는 음성 document의 일부분에 대해 각각 3개의 vector를 구성한다. 이 때 비교대상이 되는 음성 document의 일부분은 음성질의어의 음소열과 같은 길이의 음소열이고, segment라 한다. 음성질의어와 segment로부터 각각 3개의 vector가 만들어지는데, 첫 번째 vector는 46개의 원소를 가지며, 이는 46개의 monophone에 대한 identity 정보를 나타낸다. 즉 음성질의어나 segment가 가지고 있는 monophone들에 대해서는 1의 값을 가지고, 그렇지 않은 것들에 대해서는 0의 값을 가진다. 두 번째 vector는 46×46개의 원소를 가지며, 이는 46개의 monophone들의 2개의 순서적 조합에 대한 identity를 나타낸다. 세 번째 vector는 46×46×46개의 원소를 가지며 3개의 monophone의 순서적 조합에 대한 identity를 나타낸다[1][2]. 이를 수식 (1)을 통하여 나타내면[1][2],

$$q(c)=\begin{cases}1 & \text{if } c \in Q \\ 0 & otherwise\end{cases} \quad \text{and} \quad d(c)=\begin{cases}1 & \text{if } c \in D \\ 0 & otherwise\end{cases} \tag{1}$$

여기서 c는 vector의 한 원소 q(c)가 나타내는 어떤 음소열이다. c는 길이가 1, 2, 또는 3인 음소열이고, c가 음성질의어 Q내에 들어 있으면 q(c)는 1, 그렇지 않으면 0이 된다. 같은 방법으로 c가 음성 document의 segment D 내에 들어 있으면 vector의 한 원소 d(c)는 1이 되고 그렇지 않으면 0이 된다.



<그림 2> 음소발생정보를 이용하는 방법

이렇게 만들어진 vector들(음성질의어로부터 3가지, segment로부터 3가지)로부터 유사도 점수를 계산하는데, 같은 차원의 vector들 간의 내적을 구한 뒤, 그 값들을 더한다. 자세한 수식은 식 (2)와 같다[1][2].

$$S_N(q,\ d) = \sum_{c \in Q} P_N(u_c |\ c) \cdot q(c) \cdot d(u_c) \qquad (2)$$

식 (2)에서 N은 vector의 한 원소가 나타내는 음소열 c의 길이이다. 즉, N=1, 2, 3일 때 각각의 score가 식 (2)를 통해 각각 구해진다. u_c는 c가 음소열 Q에 포함되어 있을 때 segment에도 포함되어 있으면 c로 해석되고, segment에 포함되어 있지 않으면, segment가 포함하고 있는 음소열 중 $P_N(u_c \mid c)$의 값을 크게 하는 음소열로 해석된다[1][2]. 이는 식 (3)으로 나타낼 수 있다. 한편, 여기서 $P_N(u_c \mid c)$는 음소에 대한 confusion matrix로부터 계산되고, 식 (4)에서 자세히 나타나 있다. 음소에 대한 confusion matrix는 음소인식기를 구성할 때 구해진 것을 사용한다[1][2].

$$u_c = \begin{cases} c & \text{if } c \in Q \\ \arg[\max_{c' \in D} P(c' \mid c)] & otherwise \end{cases} \quad (3)$$

여기서 $P(c' \mid c)$는 식 (2)에서의 $P_N(c' \mid c)$과 같은 식이다. 즉, 음소열 c가 c'로 나타날 확률이다. 식 (4)에서 이 값에 대해 자세히 설명하면,

$$P_N(c_d \mid c_q) = \left\{ \prod_{k=1}^{N} P(\beta_k \mid \alpha_k) \right\}^{\frac{1}{N}} \quad (4)$$

α_1, α_2, α_3, β_1, β_2, β_3가 각각 하나의 음소를 나타낼 때, N=1인 경우에 c_q는 하나의 음소 α_1으로, c_d는 하나의 음소 β_1로 나타낼 수 있고, N=2인 경우에는 c_q는 α_1과 α_2의 음소열로, c_d는 β_1과 β_2의 음소열로 나타낼 수 있다. N=3인 경우는 c_q와 c_d를 각각 $\alpha_1, \alpha_2, \alpha_3$의 음소열과 $\beta_1, \beta_2, \beta_3$의 음소열로 나타낼 수 있다. 이 때 식 (4)에 의해, $P_N(c_d \mid c_q)$는 N값에 따른 음소에 대한 confusion probability 값의 기하평균으로 나타낼 수 있다[1][2].

한편, 식 (2)로부터 구해진 N에 따른 3개의 유사도 score는 식 (5)와 같이 더해져서 음성질의어와 segment의 유사도 score로 사용된다[1][2].

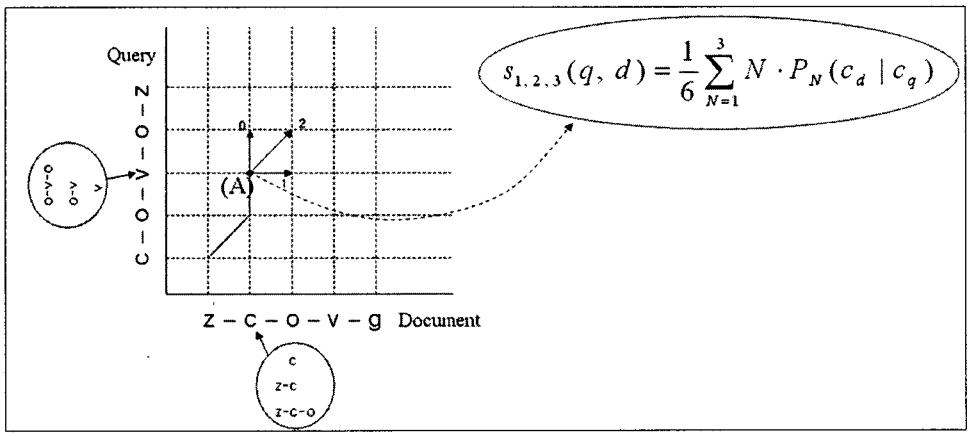
$$S_{1,2,3}(q, d) = \frac{1}{6} \sum_{N=1}^{3} N \cdot S_N(q, d) \quad (5)$$

이렇게 구해진 유사도 score는 음성 document가 끝날 때까지 segment를 shift 시키면서 계속 계산된 뒤, 가장 큰 값을 음성질의어에 대한 그 음성 document의 유사도 score로 삼는다. 즉 유사도 score의 값이 클수록 음성질의어를 포함하고 있을 가능성이 크다.

2.2. 음소 순서정보를 이용한 방법

음소발생정보를 이용한 방법은 음성질의어와 segment를 비교할 때, 시간정보를 사용하지 않았다. 여기서는 앞서 소개한 음소발생정보를 이용하는 방법의 성능을 개선시키기 위해 음소들 간의 순서정보를 이용하여 음성질의어와 음성 document를 비교하였다. 즉, 길이가 같은 음성질의어의 음소열과 segment의 음소열들을 dynamic programing을 이용하여 유사도를 측정한다[2]. 보통의 dynamic programing

은 두 pattern들 간의 distance를 측정하여 가장 적은 distance를 구하는 데 반해, 여기서는 두 pattern들 간의 유사도를 측정하여 가장 큰 값을 구한다[2][4]. <그림 3>에서 dynamic programing을 이용하여 유사도를 구하는 방법을 보여주고 있다.



<그림 3> 음소순서정보를 이용하여 계산되는 유사도 score

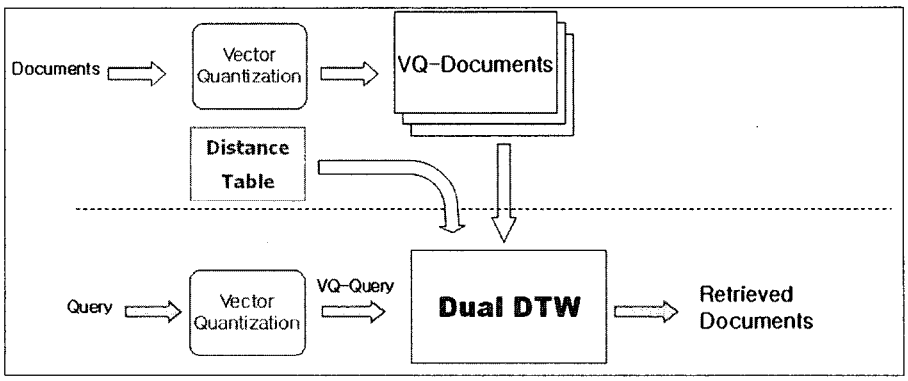
위의 그림에서 보는 바와 같이 점(A)에서의 지역 유사도 score는 식 (6)을 통해 구해진다. 즉 (A)에서는 음성질의어의 v, o-v, o-v-o와 segment의 c, z-c, z-c-o에 대한 유사도가 고려된다. 이로부터 지역 유사도 score가 구해지면 앞서 구해진 전역 유사도 score에 더해진 뒤 0, 1, 2의 세 방향 중 지역 유사도 score가 큰 방향으로 진행한다. 이렇게 구해진 전역 유사도 score는 저장되고, segment를 이동시키면서 같은 과정이 되풀이된다. 저장된 전역 유사도 score 가운데 가장 큰 값을 문장의 대표 전역 유사도 score 값으로 사용하고. 값이 클수록 해당되는 음성질의어가 포함될 가능성이 크다[2].

$$s_{1,2,3}(q,\ d)=\frac{1}{6}\sum_{N=1}^{3} N\cdot P_N(c_d \mid c_q) \tag{6}$$

음소의 발생정보나 순서정보를 사용하는 경우에는 음소 인식기의 인식성능이 시스템의 전체 성능에 많은 영향을 끼친다. 다음 장에서는 음소인식기를 사용하지 않는 방법에 대해 소개하겠다. 이는 vector quantization과 dynamic time warping을 사용하여 pattern을 비교한다.

3. Vector Quantization과 Dynamic Time Warping을 이용한 방법

Vector quantization과 dynamic time warping을 사용하여 음성질의어와 음성 document를 비교하는 방법은 <그림 4>에 잘 나타나 있다. 그림에서 제일 왼편의 Query 와 Document는 39차 MFCC 특성벡터를 의미한다. 이들은 vector quantization 과정을 거친 뒤 본 논문에서 제안하는 dual-DTW를 통해 비교된다.



<그림 4> Vector Quantization과 Dynamic Time Warping을 이용하는 정보검색 방법

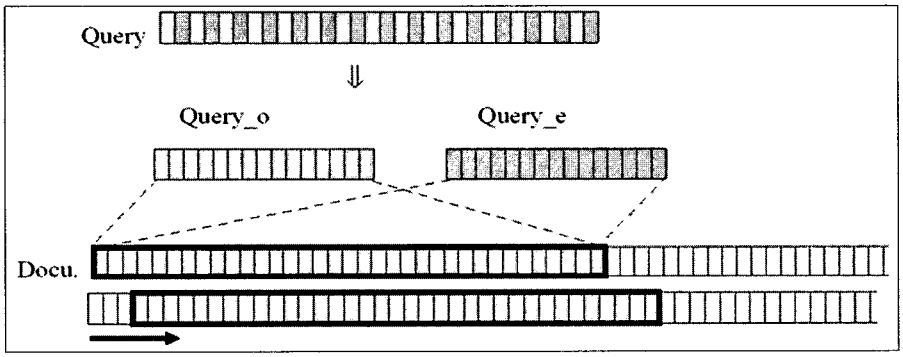
3.1. Codebook 구성 및 Distance Table 구성

LBG algorithm을 통하여 512, 1024, 2048개의 codeword를 가지는 3가지의 codebook이 만들어진다[5]. 이 때 국어공학 연구소의 낭독음성DB 보급판 가운데 PBW(phonetic balanced word) 452어절 DB의 70명(남: 39명, 여: 31명)분의 데이터를 사용하였다. 한편 distance table은 Euclidean distance를 사용하며, 512 size의 codebook에 대해서는 512×512 size의 table을 1024, 2048 size의 codebook들에 대해서는 1024×1024, 2048×2048 size의 table을 각각 만든다. Distance table은 symmetric matrix이므로 각 원소들은 주대각선 상에 위치한 원소들에 대해 대칭이 된다. 따라서 실제 저장되는 양은 이 table들의 size의 절반정도이다.

3.2. Dual DTW

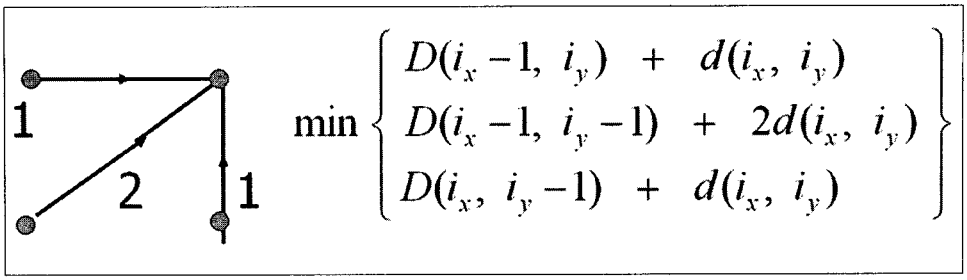
Dynamic time warping은 두 pattern들 간의 정확한 비교를 제공해 주지만, 계산량이 많아 pattern 비교를 수행하는 데 많은 시간을 소비한다. Dual DTW는 DTW의 이러한 단점을 극복하고자 제안되었다. 기본적으로는 DTW를 이용하지만, 비교 대상이 되는 frame들의 수를 줄임으로써 수행시간을 줄이는 방식이다. 즉, 음성은 짧은 시간동안 급격하게 변하지 않으므로[6] 10msec 간격으로 추출된 특성벡터들

에서 인접한 frame들은 거의 비슷한 값을 가진다. 따라서 DTW 수행 시 모든 frame을 사용하지 않고 하나씩을 건너 뛰어 사용해도 성능의 저하를 많이 야기하지 않을 것이다. 이는 <그림 5>에 잘 나타나 있다. 음성질의어의 frame들 중 홀수 번째 frame들만 따로 모아서 Query_o를 만들고 짝수 번째 것만 따로 모아서 Query_e를 만든 다음, Document의 한 segment 내의 홀수 번째 frame들(<그림 5>에서 회색으로 표시된 frame들)과 DTW를 수행한다. 그런 다음segment를 이동시킬 때 홀수개의 frame 단위로 이동시키면 (segment의 size는 그대로이다) 현재 segment에서 비교대상이 아니었던 짝수 번째 frame들이 segment의 이동 후에는 홀수 번째 frame이 되어 비교대상이 된다. 따라서 Dual DTW 수행 중, 음성 docuement 내의 frame들 가운데 비교대상에서 제외되는 것은 하나도 없거나(segment의 이동단위가 1개의 frame일 때), 첫 번째 segment의 첫 몇 개의 frame에 불과하다.



<그림 5> Dual DTW에서의 음성질의어의 분리 및 segment와의 비교

<그림 6>에서는 본 논문에서 사용된 DTW에서의 local continuity constraint와 slope weight를 나타내고 있는데, 이 경우 global constraint에 의한 DTW의 수행면적은 직사각형을 이룬다[4]. 한편, DTW 수행 시 음성질의어의 길이가 반으로 줄고 segment의 길이가 반으로 줄어들면 global constraint에 의한 DTW의 수행 면적은 기존의 면적의 4분의 1로 줄어든다. 그리고 Dual-DTW에서는 한 segment에 대하여 DTW 수행이 2번이기 때문에(Query_o에 대해서 1번, Query_e에 대해서 1번) 전체 계산양은 4분의 1로 줄어드는 것이 아니라, 반으로 줄어들게 된다($\frac{1}{4}\times 2=\frac{1}{2}$).



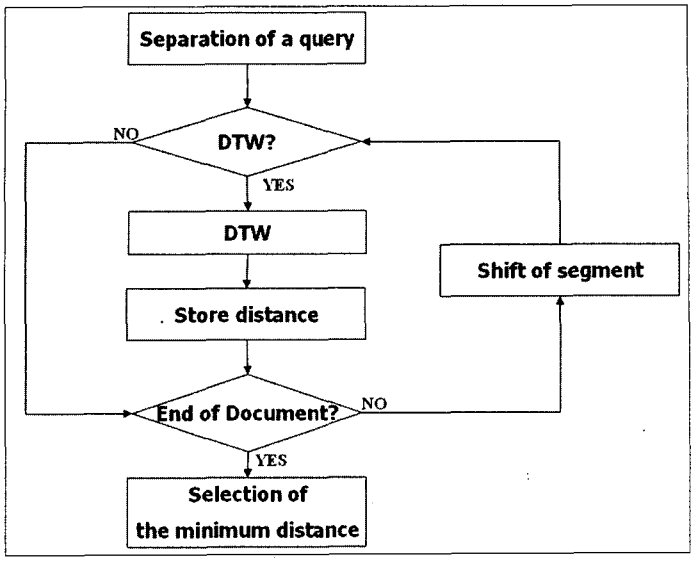
<그림 6> Local continuity constraint with slope weighting and DP recursion formula

이와 같이 Dual-DTW를 사용하면, 계산량이 줄어들어 수행시간을 줄일 수 있다. 하지만, DTW의 수행횟수가 많을수록 전체수행시간은 길어지므로, 두 패턴 간의 비교에서 불필요한 DTW의 수행횟수를 줄이는 것이 전체수행시간을 줄이는 방법이다. 그래서 DTW의 수행이 필요한 지를 결정하기 위해 candidate score를 정의하여, 그 값이 특정값보다 작으면 DTW를 수행하지 않고 segment를 이동시키도록 고안하였다. 이는 식 (7)에 잘 나타나 있다.

$$\text{Candidate score} = \frac{1}{N} \sum_{c \in Q} n_c \tag{7}$$

Candidate score는 segment의 비교대상이 되는 frame들 가운데 음성질의어의 frame과 같은 VQ index들을 가지는 frame들의 수의 비이다. 여기서 N은 segment의 frame들 중에서 비교대상이 되는 frame들의 수이고, c는 음성질의어Q에 포함된 VQ index(각 frame들은 VQ codeword의 index로 표현되어 있다)이며, n_c는 segment 내의 비교대상이 되는 frame들 가운데 특정 VQ index c를 가지는 frame들의 수이다.

<그림 7>은 Dual-DTW를 수행하는 전체 과정을 나타낸다. 음성질의어를 Query_o와 Query_e로 분리한 후, 분리된 각 음성질의어와 비교대상이 되는 segment를 비교하여 식 (7)에서 나타낸 Candidate score 값에 따라 DTW 수행여부를 판단한다. 이렇게 수행된 DTW 후에는 음성질의어와 segment 간의 전역거리가 저장되며, 최종적으로 저장된 전역거리 가운데 가장 짧은 거리를 음성 document의 대표 전역거리로 정한다. 거리가 짧을수록 해당 음성 질의어가 포함되었을 확률이 높다.



<그림 7> Dual DTW

4. 실험 및 결과

4.1. 실험에 사용된 DB와 실험환경

13명의 화자(남자 10명, 여자 3명)로부터 질의어 10개, document 100개를 각각 녹음하였다. Document는 카메라의 음성메모처럼 3초 내외로 발성된 단일 문장으로 구성되었다. 질의어 1개는 10개의 document에 각각 포함되어 있으며, 질의어들의 목록은 <표 1>과 같다. 각 질의어는 <표 1>에 나타난 바와 같이 docuemnt의 앞, 가운데, 뒤에 고정되어 위치된 것들과, document의 여러 부분에 위치된 것들이 있다. 실험은 음성질의어를 발성한 화자와 음성 document를 발성한 화자가 같은 경우의 동일화자실험과 음성질의어를 발성한 화자와 음성 document를 발생한 화자가 반드시 같지 않은 비동일화자실험으로 나누어 행해졌다. 동일화자실험에서는 1개의 음성질의어는 100개의 음성 document와 비교되고, 비동일화자 실험에서는 1300개(100문장×13명)의 음성 document와 비교된다.

실험은 펜티엄4 Dual Core 프로세서 2.8GHz, 메모리 1GB, 운영체제 Windows XP (service pack 2)가 설치된 컴퓨터에서 수행되었다.

<표 1> DB에 사용된 질의어

질의어	문장내의 위치
졸업식	앞
생일날	앞
지리산	앞
롯데월드	가운데
아웃백	가운데
수목원	끝
에버랜드	끝
계룡산	앞, 가운데, 끝
제주도	앞, 가운데, 끝
경주	앞, 가운데, 끝

4.2. 성능측정 방법

본 실험에서는 정보 검색에서 일반적으로 사용하는 recall 값과 precision 값을 이용하여 성능을 평가하였다. Recall rate과 precision rate은 다음과 같다[7].

$$\text{recall rate} = \frac{\text{현재 불려온 음성문장들 중 바르게 불려온 수}}{\text{불려와야 될 음성문장들의 수}} \tag{8}$$

$$\text{precision rate} = \frac{\text{현재 불려온 음성문장들 중 바르게 불려온 수}}{\text{현재 불려온 음성문장들의 수}} \tag{9}$$

한편, 정보검색 시스템에서 recall 값에 따라 precision이 변하므로, 시스템의 성능을 나타낼 단일 척도가 필요하다. 이를 위해 11개의 recall 값(0, 0.1, 0.2, …, 1.0)에 대하여 precision을 측정한 뒤, 이 값들에 의한 recall-precision 곡선이 나타내는 면적을 구하여 mean average precision(mAP)라고 불리는 값을 계산한다. 이 값은 0과 1사이의 값을 가지며 이상적인 경우 1이 된다[7].

4.3. 실험 결과

실험은 음성질의어를 발성한 화자와 음성 document를 발성한 화자가 같은 경우

와, 그렇지 않은 경우에 대해 수행하였다. <표 2>에서는 같은 조건에서의 일반 DTW과 Dual-DTW의 성능을 나타내었다. 이 때 segment의 shift size는 두 경우 다 1 frame이고, Dual-DTW에서 사용한 candidate score 값은 0이다. 수행시간은 1개의 음성질의어가 100개의 document를 검색하는 시간을 나타내고 단위는 sec이다.

<표 2> 일반 DTW와 Dual-DTW의 성능비교

DTW종류	일반 DTW						Dual-DTW					
codebook size	512		1024		2048		512		1024		2048	
에 따른 성능	mAP	시간	mAP	시간	mAP	시간	mAP	시간	mAP	시간	mAP	시간
동일화자	0.83	2.4	0.85	2.4	0.85	2.6	0.83	1.4	0.83	1.4	0.84	1.5
비동일화자	0.52	2.4	0.52	2.4	0.53	2.6	0.51	1.4	0.51	1.4	0.51	1.4

<표 2>에서 알 수 있듯이 시스템의 성능은 Dual-DTW를 사용한 경우가 일반 DTW를 사용의 경우보다 비슷하거나 약간 낮아졌다. 하지만, 수행시간이 40%이상 줄어들어 Dual-DTW를 사용한 경우가 성능과 수행시간 면에서 우수함을 알 수 있다.

한편, 수행시간을 더 줄이기 위해 Dual-DTW의 candidate score를 0.2로 두면, 성능은 <표 2>에서 나타난 것보다 다소 낮아진다. Candidate score가 커질수록 DTW의 수행대상이 되는 frame들이 줄어들므로 그 수행횟수가 적어진다. 따라서 성능은 낮아지고 수행속도는 빨라진다. <표 3>에서는 candidate score가 0.2인 Daul-DTW의 성능과 2장에서 소개된 음소정보를 사용한 시스템의 성능을 나타내었다.

<표 3> 음소정보를 이용한 경우와 candidate score=0.2인 Dual-DTW의 성능비교

알고리즘 종류	음소인식 이용				Dual-DTW					
	음소발생정보		음소순서정보		512		1024		2048	
	mAP	시간	mAP	시간	mAP	시간	mAP	시간	mAP	시간
동일화자	0.55	0.6	0.58	0.5	0.83	1.1	0.82	0.9	0.81	0.8
비동일화자	0.36	0.6	0.39	0.5	0.50	1.0	0.48	0.8	0.45	0.7

음소인식을 사용하는 경우는 Dual-DTW에 비해 낮은 성능을 나타내는 대신 수행시간이 짧음을 알 수 있다. 그리고 음소순서정보를 이용하는 알고리즘이 음소발생정보를 이용하는 경우보다 성능이 더 높음을 알 수 있다. 음소발생정보를 이용하는 경우가 음소순서정보를 이용하는 경우보다 수행시간이 더 걸리는 이유는 음소발생정보에서의 vector space의 규모가 최대 46×46×46이기 때문에 이들 개개의 음소열들에 대해 음성질의어나 segment에 포함이 되어있는지를 확인하는데 시간이

많이 소모되기 때문이다. 한편, Dual-DTW에 candidate score를 적용한 결과 <표 2>에서 나타난 결과보다 성능이 낮아졌으나 수행시간은 짧아졌다.

5. 결 론

본 논문에서는 디지털 카메라에서 찍은 사진을 검색하기 위해, 디지털 카메라에서 사진과 같이 녹음한 음성 메모를 사진의 인덱스로 삼아, 사용자가 음성으로 질의어를 입력하였을 때 사진을 검색하는 시스템을 구현하였다. 구현에 사용된 알고리즘은 음소인식에 기반한 알고리즘 2가지와 음소인식을 사용하지 않고 VQ와 DTW를 사용한 알고리즘이다. 음소인식에 기반한 알고리즘들은 음소인식기의 성능에 종속되기 때문에 본 논문에서 제시한 낮은 성능의 인식기에서는 좋은 성능을 나타내지 못했다. 하지만, VQ와 DTW를 사용하는 알고리즘은 음소인식에 기반한 방법들에 비해 좋은 성능을 보여주었다. 본 논문에서 제안한 Dual-DTW를 이용한 방법은 DTW의 장점인 두 pattern 간의 정확한 비교를 취하는 동시에 DTW의 단점인 많은 계산량을 줄임으로써 성능향상을 도모하였다. 그리고 동일화자실험의 성능이 비동일화자의 경우보다 우수했는데, 카메라가 개인적인 용도로 사용되는 현실에 비추어볼 때, 동일화자실험의 성능이 더 의미가 있는 것으로 판단된다.

앞으로의 실험계획은 성능향상에 영향을 주는 local continuity constraint를 최적화 시키는 것과, 실험 DB의 양을 늘려 실험결과에 보다 신빙성을 높이는 것이다. 아울러 음성 document에 포함되지 않은 음성질의어에 대한 거부기능을 첨가할 계획이다.

참 고 문 헌

[1] N. Moreau, H. G. Kim, and T. Sikora, "Phone-based spoken document retrieval in conformance with the MPEG-7 standard", *Proc. AES 25th International Conference, 2004.*

[2] 김태성, 서영주, 김회린, "음소 인식기를 이용한 음성정보 검색시스템의 구현", *제22회 음성 통신 및 신호처리 학술대회*, 숭실대학교, pp. 21-24, 2005.

[3] HTK (Hidden Markov Model ToolKit), http://htk.eng.cam.ca.uk/.

[4] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

[5] X. Huang, A. Acero, H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001.

[6] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[7] B. Yates, R. Neto, *Modern Information Retrieval*, ACM Press, 1999.

접수일자: 2006년 2월 1일
게재결정: 2006년 3월 13일

▶ 김태성(Taesung Kim)
주소: 305-714 대전광역시 유성구 문지동 103-6 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6221
E-mail: taesung@icu.ac.kr

▶ 서영주(Youngjoo Suh)
주소: 305-714 대전광역시 유성구 문지동 103-6 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6221
E-mail: yjsuh@icu.ac.kr

▶ 김회린(Hoirin Kim) : 교신저자
주소: 305-714 대전광역시 유성구 문지동 103-6 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술 연구실
전화: 042) 866-6139
E-mail: hrkim@icu.ac.kr

▶ 이용주(Yong-Ju Lee)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 원광대학교 전기전자 및 정보공학부
전화: 063) 850-7451
E-mail: yjlee@wonkwang.ac.kr