1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

유성/무성/묵음 정보를 이용한 동적 시간 정합 알고리즘 개선

최 민 석, 한 현 배, 한 민 수
한국정보통신대학원대학교

Improvement of Dynamic Time Warping Algorithm by Using Voiced/Unvoiced/Silence Information

Min Seok Choi , Hyun Bae Han, Min Soo Hahn
Information and Communications University
E-mail : mschoi@icu.ac.kr

요 약

본 연구에서는 고립단어 인식시스템에 사용되고 있는 DTW (Dynamic Time Warping) 알고리즘의 계산량을 줄일 수 있는 방법을 제안한다. 일반적으로 고립단어 인식 시 가장 인식률이 좋은 알고리즘은 DTW 라고 알려져 있으나, 인식대상어휘가 늘어나면 계산량이 비례해서 늘어나고 인식률이 저하되는 단점이 있으므로 일반적으로 200 단어 이하의 어휘에만 사용되고 있다. 따라서 대상어휘를 감소시켜 계산량을 줄이기 위해 본 논문에서는 유성/무성/묵음 (V/U/S) 정보를 이용하여 코드워드를 구성하고 같은 코드워드에 해당되는 단어들을 추출해 이들 만을 비교대상 어휘로 제한하므로서 DTW 알고리즘을 적용할 대상 어휘수를 줄이는 방법을 사용하여 계산 속도를 향상시켰다. 또한 입력 단어와 대상 단어와의 누적거리 계산 시 끝점 정보 뿐만 아니라 유성/무성/묵음 경계 정보를 이용하여 piecewise DTW 를 구현하므로서 탐색 영역을 축소함으로써 추가적인 계산량 감소가 가능하다. 따라서 상기 기법들을 이용하면 PC 상에서도 DTW 를 이용한 대어휘 고립단어 음성 인식기의 구현이 가능할 것이다.

I. 서 론

일반적으로 음성 인식 시스템에는 벡터 양자화 (Vector Quantization)를 이용하는 방법과 동적 시간 정합 (DTW)을 이용하는 방법, 신경 회로망(Neural Network)을 이용하는 방법, 은닉 마코프 모델(Hidden Markov Model, HMM)을 이용하는 방법 등이 사용되고 있다. [2] [3][4]

동적 시간 정합을 이용한 음성 인식 시스템은 일반적으로 인식어휘수가 100 단어 이하인 소규모 음성 인식 응용 시스템에 적용되어 좋은 성능을 보여주고 있으며 대부분의 고립단어 인식 기능을 갖는 상용시스템에 사용되고 있다. 하지만, 대규모 음성인식시스템을 구현할 경우 데이터베이스의 기준 패턴수가 증가함에 따라 계산량이 비례해서 증가하게 된다.[1][2] 이러한 계산량 부담은 대어휘 음성 인식 시스템 구현에 있어서 동적 시간 정합 알고리즘을 도입하는 것을 어렵게 만들었으며 소어휘의 경우에도 빠른 응답 시간을 보장하기 위해선 계산량을 감축하는 것이 필요하다. 동적 시간 정합의 계산량 부담을 줄일 수 있는 방법으로 본 논문에서 제안하고 있는 방법은 유성/무성/묵음 분류기의 분류 결과를 이용하는 방법이다. 즉, 발성된 고립단어를 유성/무성/묵음 분류기를 이용하여 유성음, 무성음, 묵음 구간으로 분류한 후 V/U/S 코드워드를 생성하여 V/U/S 코드워드 별 데이터베이스를 구축한다. 이렇게 한 후 비교 대상 어휘를 동일한 V/U/S 코드워드를 갖는 어휘로 제한함으로써 동적 시간 정합의 비교 탐색 범위를 줄일 수 있으므로 인식시간의 단축이 가능하다. 또한 누적거리 계산 시에도 유성.무성.묵음 분류기의 구간별 경계정보를 이용하여 한 단어 내의 유성음 구간은 유성음끼리, 무성음은 무성음끼리 순차적으로 piecewise 하게 탐색하므로써 탐색 범위를 줄일 수 있으므로 추가적인 계산시간 단축이 가능하다.

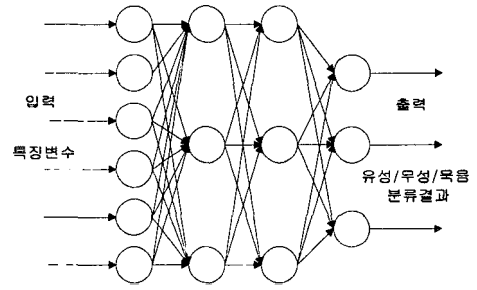
II. 신경망을 이용한 유성/무성/묵음 분류

최근 다양한 구조의 신경망이 패턴인식, 시스템 모델링, 비선형 예측 등과 같은 분야에서 널리 사용되고 있다. 신경망은 인간두뇌의 생리학적 구조와 기능을 모사한 인지적 정보처리 장치로서 학습, 패턴 인식, 연상기억능력이 있다. 이러한 신경망의 기능 때문에 음성 인식, 합성 및 분석 분야에서 많이 사용되고 있다. 본 논문에서는 신경망을 이용하여 유성/무성/묵음 분류기를 구현하였다. 신경망을 이용한 유성/무성/묵음 분류기의 입력으로 사용된 특징 변수는 원래 음성 신호와 전처리 (pre-emphasis)된 음성 신호에 대해 각각 에너지, 영교차율, 레벨 교차율의 6 가지를 사용하였다. 6 가지 특징 변수를 사용하여 신경망을 훈련 시킨 뒤 후처리 (post processing) 를 거쳐 입력 음성 신호를 유성음, 무성음, 묵음 구간으로 분류하였다. 신경망 훈련에 사용된 데이터는 남, 여 각각 3 명씩 30 단어를 사용하였다. 본 논문에서 사용된 신경망은 2 개의 은닉층 (Hidden layer)을 가지며 Generalized delta rule 을 사용하였고 사용된 파라미터는 표. 1 과 같다.

[표. 1] 신경망에 사용된 특징변수 및 상수

Parameter	Value
Number of input feature	6
Number of output unit	3
Momentum rate	0.5
Learning rate	0.3
Max total error	0.001
Max individual error	0.0001
Max number of iteration	20000
Number of hidden layers	2
Number of first hidden layer	12
Number of second hidden layer	8

신경망을 이용하여 구현한 유성/무성/묵음 분류기의 블록도는 그림. 1과 같다. 신경망의 입력으로는 앞에서 설명한 6가지의 특징 변수가 입력되고 출력으로는 각 음성 프레임 별로 유성음, 무성음, 묵음 정보가 얻어진다.



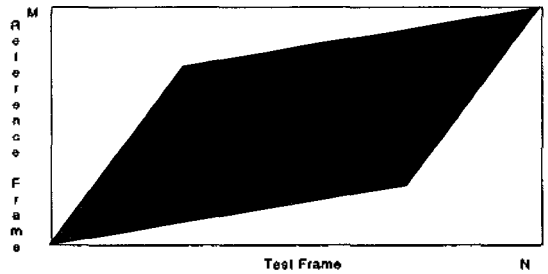
[그림. 1] 유성/무성/묵음 분류에 사용된 신경망

III. 동적 시간 정합

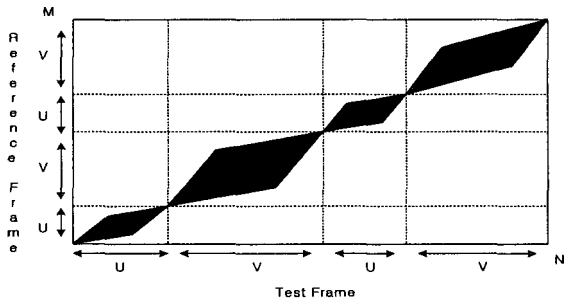
동일인이 같은 단어를 발성하는 경우에도 특별히 훈련받은 사람이 아니면 발성할 때마다 단어의 시간적 길이 뿐만 아니라 구성 음소별 지속시간이 변화한다. 이를 미리 준비된 표준 패턴과 단순히 비교하면 시간 축이 서로 다르기 때문에 고립단어 인식기의 인식률이 무척 저하된다. 이 영향을 줄이기 위한 방법에는 음소열을 이용한 인식, 시간축의 정규화 방법을 이용한 인식 등이 있다. 가장 초보적인 시간축의 정규화 방법은 선형 신축(linear scaling)에 의한 두 패턴 길이의 정규화 방법이다. 그러나 음성은 시간에 따른 구성요소의 신축 정도가 비선형적이고, 모음과 자음의 신축 정도가 상이하기 때문에 이 방법은 적절하지 않다. 이 문제를 비교적 고속이면서 효과적으로 처리할 수 있는 방법이 Vinsyuk, Chiba와 Sakoe에 의해 제안된 동적 프로그래밍 (dynamic programming; DP) 에 기반한 시간축의 비선형 신축에 의한 정합법이다. 동적시간 정합은 서로 다른 두 개의 자료에서 비선형의 최적의 정합 경로를 찾아 서로 다른 길이의 특징 벡터를 비교하는 방법이다. 현재까지 고립단어 인식에서 가장 우수한 인식률을 보이고 있으나, 인식 대상 어휘가 증가하면 계산량이 비례해서 증가할 뿐만 아니라 인식률도 상대적으로 감소한다는 단점이 있다. 본 논문에서는 유성/무성/묵음 분류기를 이용한 VUS 코드워드 생성 및 구간 별 경계 정보를 이용하여 동적 시간 정합에서의 탐색 범위를 축소하고 인식 대상어휘를 줄이므로서 고립단어 인식에 필요한 전체 계산 시간을 줄일 수 있는 방법을 제안한다.

III. V/U/S 인식 시스템

일반적으로 DTW를 이용하여 특징 벡터 간의 누적 거리값을 측정하기 위해서는 많은 계산량이 필요하다. 기존의 DTW 방법에서는 그림. 2 (a)와 같은 영역에서 거리값을 계산하게 되나 본 논문에서는 유성/무성/묵음 경계정보를 이용하여 순차적으로 구간별 탐색을 실현하므로서 그림. 2 (b)의 경우와 같이 탐색영역이 감소함을 알 수 있다. 이러한 방법를 이용하여 특징 벡터열 간의 누적거리 계산시의 계산량을 줄일 수 있는 것이다.



(a) 기존의 DTW에서의 계산영역

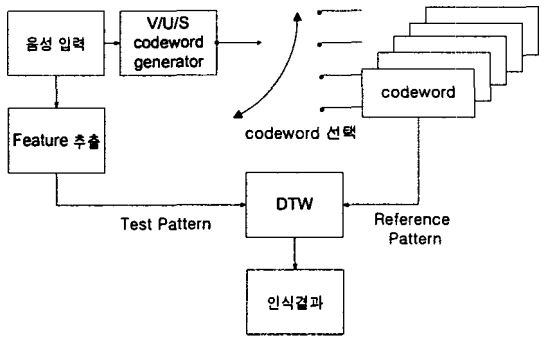


(b) 제안된 방법에서의 계산영역

[그림. 2] 계산영역 비교

본 논문에서 제안하는 고립단어 인식 시스템의 전체

블록도는 다음 그림. 3 과 같다. 입력된 발성단어에 대하여 V/U/S 코드워드를 생성시킨 후, 그 코드워드에 해당하는 대상어휘 군을 선택한다. 선택된 대상어휘 군에 속하는 기준 패턴 (template) 과 입력된 음성신호에서 추출한 테스트 패턴과의 유사도를 DTW 를 이용한 누적 거리값을 이용하여 측정하며 가장 유사한 기준 패턴 단어, 즉 누적 거리값이 최소가 되는 단어를 출력시키므로서 인식을 수행한다.



[그림. 3] 제안된 방법의 블록도

IV. 실험 및 결과

실험에 사용한 데이터는 한국전자통신연구원 (ETRI) 에서 구축한 452 음소 균등 단어(Phoneme balanced word) 데이터베이스 중 남녀 각 3 인, 즉 전체 6 명 화자가 발성한 30 단어 (전체 180 단어, 표. 2 참조)를 수작업으로 유성/무성/묵음 구간으로 분류한 후 유성/무성/묵음 분류기에 이용된 신경망의 학습에 사용하였으며, 남녀 각 5 인, 즉 전체 10 명의 화자가 발성한 50 단어 (전체 500 단어)를 유성/무성/묵음 분류기의 성능 평가에 사용하였다.

상기 10 인의 화자 중 남녀 각 2 인 즉, 4 명의 화자에 의해 발성된 음성데이터를 음성 인식에 사용하였다. 4 명의 화자에 의한 발성 중 남.녀 각 한명의 화자에 의한 발성을 수작업으로 유성/무성/묵음 구간을 분류한 후 기준 패턴으로 작성하였으며, 나머지 2 명의 화자의 발성은 인식기의 성능 평가를 위한 테스트 패턴으로 사용하였다. 유성/무성/묵음 분류기에서 사용한 음성 특징 변수는 원래의 음성 신호와 전처리(pre-emphasis)된 음성 신호에서 추출된 에너지, 영교차율, 레벨교차율의 6 가지를 사용하였으며, DTW 인식기에서 사용된 특징 벡터는 12 차 LPC (Linear Predictive Coding) 켑스트럼 (Cepstrum) 계수를 이용하였다.

수작업으로 V/U/S 분류된 테스트 패턴

수작업으로 V/U/S 분류된 테스트 패턴에 대한 기존의 DTW 인식기와 본 논문에서 제안한 V/U/S 정보를 이용한 DTW 인식기의 인식률 및 단어 당 평균 처리시간의 비교는 표. 3 과 같다.

[표. 2] 인식 단어 및 해당 V/U/S 코드워드

단어	V/U/S 코드	단어	V/U/S 코드
자유와	UV	취약한	UVSUV
주위의	UV	규범이	UVUV
되풀이	UVSUV	자외선	UVUV
뇌물을	V	세율을	UV
거예요	UV	요소와	VUV
누워서	VUV	민방위	VUV
이집트	VUVSUV	재활용	UVUV
왔지만	VSUV	에너지	VUV
경우와	UV	발맞춰	UVSUV
스위스	UVUV	키워야	UV
최우선	UVUV	위치에	VSUV
귀여운	UV	영어의	V
원만한	VUV	지위의	UV
괜찮은	UVUV	내뿜는	VSUV
여의치	VSUV	두번째	UVUVSUV

[표. 3] 기존의 DTW 와 제안된 DTW 의 성능 비교

	제안된 DTW		기존의 DTW	
	인식률	처리시간	인식률	처리시간
남	80.0 %	40.4 msec	86.7 %	231 msec
여	93.3 %	58.3 msec	100 %	359 msec

표.3 을 보면 본 논문에서 제안한 알고리즘이 기존의 DTW 인식기에 비해 비록 인식률에서는 약 6.7%의 저하를 가져오나 단어 당 인식 시간은 약 20% 정도로 줄어듬을 알 수 있다. 을 이용한 DTW 인식기의 계산량이 기존의 DTW 인식기에 비해 현저하게 감소함을 볼 수 있다. 상기 표에서 인식률의 저하는 주로 유성-무성 및 무성-유성 변이 구간에서의 부정확한 경계 정보 검출 때문에 발생하였다고 판단된다.

V/U/S 분류기로 분류된 테스트 패턴

V/U/S 분류기를 이용하여 자동으로 분류된 테스트 패턴에 대한 제안된 DTW 알고리즘을 이용한 인식기의 인식률 및 단어 당 평균 처리시간의 비교는 표. 4 와 같다. 기준 패턴에 대해 각각 하나씩의 V/U/S 코드워드를 생성하였을 때의 성능은 수작업으로 V/U/S 분류해 준 것에 비해 성능이 상당히 떨어진다. 이는 주로 유성음 사이의 'ㅎ' 음소가 발성되지 않는 경우나 순음 'ㅁ', 'ㅂ'이 혼합음(mixed-sound) 형태로 나타나는 데서 기인한 오류로 해석된다. 이 경우를 고려하여 한 단어에 대해 두개의 코드워드, 즉 이중 코드워드를 생성해 줌으로써 V/U/S 분류에 의해 검색 대상 코드북이 제한됨에 따른 오류를 감소시킬 수 있다. 표. 4 에서 보면 전체 처리 시간이 기존의 DTW 에 비해 큰 차이가 없음을 알

수 있으나 이는 본 논문에서 사용한 신경망이 유성/무성/묵음 분류 시에 이용되는 계산시간이 주이므로 유성/무성/묵음 분류기를 tree-structure 패턴 분류기로 바꿔주면 개선이 가능하다.[6][7]

[표. 4] V/U/S 분류기를 이용한 DTW 인식기의 성능

	단일 코드워드		이중 코드워드	
	인식률	처리시간	인식률	처리시간
남	63.5 %	256 msec	73.3 %	268 msec
여	76.7 %	330 msec	90.0 %	337 msec

V. 결론

일반적으로 고립단어 인식시스템에서는 동적 시간 정합이 가장 우수한 인식률을 나타낸다. 하지만 동적 시간 정합 알고리즘을 이용한 고립단어 인식 시스템의 단점은 인식대상어휘가 증가되면 계산량이 비례하여 증가된다는 단점을 가지고 있다. 본 논문에서는 이러한 계산량을 감소시키기 위하여 V/U/S 코드워드를 이용한다.

본 논문에서 제안하는 고립단어 인식기에서 기준패턴 생성시 V/U/S 코드워드 별로 분류를 하여 저장하고, 인식 시에는 입력음성에 대한 V/U/S 코드워드를 추출한 후, 그 코드워드에 해당하는 기준패턴들 만을 인식대상 후보로 포함시킨다. 따라서, 본 논문에서 제안한 인식 알고리즘을 이용하면, 인식대상 어휘를 상대적으로 줄이는 효과와 더불어 패턴 간의 누적거리 측정 시 계산량이 상대적으로 감소한다는 두가지 이득이 있다. 시뮬레이션한 결과 기존의 방법에 비하여 처리시간은 5배 이상 줄어들 수 있음을 보였다. 현재 제안된 알고리즘을 이용한 인식기의 성능은 기존의 DTW를 이용하는 경우보다 인식률이 저하되었다. 이는 주로 유성/무성/묵음 분류기의 에러에 기인하므로 향후 유성/무성/묵음 분류기의 성능 개선 연구를 수행할 예정이며 인식시간 개선에 장애가 된 신경망 구조도 1개의 은닉층을 갖는 신경망 구조나 tree-structure로 대치할 예정이다.

참고 문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentsals of speech recognition", Prentice-Hall, 1993.

[2] H. Sakoe and S.Chiba, "Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 43-49, February 1978.

[3] A. Weibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme Recognition Using Time Delay Neural Networks", IEEE Trans. Acoustics, Speech, Signal Proc,., ASSP-37: 328-339, 1989.

[4] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, January 1986.

[5] R.M. Gray, A. Buzo, A.H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing", IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-28 (4): 367-376, August 1980.

[6] M. Hahn, "Silence and Voiced-Unvoiced-Mixed Excitation Classification of Speech with Applications A Two-Channel and A One-Channel", Ph.D. Dissertation. University of Florida, 1989.

[7] Robert Schalkoff, "Pattern Recognition", WILEY, 1992.