논문 2011-4-26

잡음환경에서의 Noise Cancel DTW를 이용한 음성인식에 관한 연구

A Study on Voice Recognition using Noise Cancel DTW for Noise Environment

안종영*, 김성수**, 김수훈***, 고시영****, 허강인*****

Jong-Young Ahn, Sung-Su Kim, Su-Hoon Kim, Si-Young Koh, Kang-In Hur

요 약 본 논문에서는 잡음 환경에서의 음성인식 개선에 관한 내용으로 기존의 DTW에서 일종의 특징보상기법을 적용한 방식으로 예측잡음이 아닌 실생활에서의 음성잡음 데이터를 적용하여 인식모델을 잡음상황에 맞도록 적응시키는 방법으로 제안하는 Noise Cancel DTW를 사용하였다. 음성인식 시 주변노이즈를 고려한 참조패턴을 생성하여 특징 보상으로 인식률을 향상 시키는 방법으로 잡음 환경에서 음성 인식률을 향상 시켰다.

Abstract In this paper, we propose the Noise Cancel DTW that to use a kind of feature compensation. This method is not to use estimated noise but we use real life environment noise data for Voice Recognition. And we applied this contaminated data for recognition reference model that suitable for noise environment. NCDTW is combined with surround noise when generating reference patten. We improved voice recognition rate at mobile environment to use NCDTW.

Key Words : Voice Recognition, DTW, Noise Cancel

Ⅰ. 서 론

현재 시스템에 의한 연속음성인식에는 많은 어려움과 연구해야 될 부분이 많이 남아 있다. 최근에는 고립단어 기반의 상용제품도 등장하고 있어 향후 실용화 단계까지는 얼마 남지 않아서 상당히 고무적이다.

음성인식에 있어서의 가장 큰 영향을 미치는 요소 중의 하나가 바로 음성인식 시 환경적으로 발생하는 잡음이다.

특히, 자동차 환경에서는 그 잡음의 강도가 심해 음성인식을 수행하는데 어려움을 초래 한다.[1]

음성인식에는 특정화자, 불특정화자를 구별하여 화자종속, 화자독립으로 나눌 수 있으며 인식단어에 따른 고립단어인식과 연속단어인식으로 나누어진다. 화자종속에는 화자인증, 핵심어인증으로 나눌 수 있다.

현재 음성인식에서의 잡음처리 기술은 크게 음성향상(speech enhancement), 특징보상(feature compensation), 모델적응(model adaptation)과 같이 세 가지로 구분된다.

음성의 특징을 추출하여 참조 패턴을 만드는 것이 기본이 되는 데 비교 패턴이 주변잡음으로 인해 영향을 받아 인식률 저하를 발생 시킨다. 음성인식 알고리듬은 크게 확률론적인 접근방법인 HMM(Hidden Marcov Model)

*정회원, 동아대학교 전자공학과
**한국폴리텍2대 컴퓨터정보과
***부천대학교 모바일통신과
****경일대학교 전자공학과
*****동아대학교 전자공학과(교신저자)
접수일자 2011.5.20, 수정완료 2011.7.8
게재확정일자 2011.8.12

이 있고 신경세포를 모델링한 NN(Neural Network)이 있다.

화자 독립의 경우 특성상 HMM을 사용하여 참조패턴을 구성하나 화자종속의 경우 신경망을 사용하여 참조패턴을 만들 수 있으나 이 부분에 대해서는 아직도 다변적인 연구가 이루어지고 있다. 그리고, 데이터가 가지는 그 특징대표벡터를 추출하여 참조패턴을 만드는 VQ(vector Quantization), DTW(Dynamic Time Warping) 방법 등이 있다.[2]

특히, DTW는 시간 축 상에서의 비선형 신축을 허용하는 패턴매칭 알고리듬으로 정의를 하는데 수행을 통하여 참조패턴을 생성할 수가 있다. 그리고 생성된 참조패턴과 입력패턴을 비교하여 인식여부를 결정한다. 여기서 참조패턴에 대한 입력패턴 비교 시 유사도 기준을 분류하여 적용이 가능하다.[3]

본 연구에서는 상기 기법 중 참조패턴에 잡음섞인 음성데이터를 같이 사용하는 특징보상방법을 적용하여 잡음환경에서 인식률을 높이는 방법으로 DTW을 사용하였다. 음성인식 시 주변노이즈를 고려한 방법으로 기존 DTW방식에서 잡음이 부가된 데이터를 함께 사용 하는 NC(Noise Cancel)DTW를 제안 하고자 한다.

II. 본 론

1. DTW(Dynamic Time Warping)

패턴인식에서 인식의 대상이 되는 패턴은 정적 패턴과 동적 패턴으로 나눌 수 있는데 정적 패턴은 지문, 숫자, 문자와 같이 고정된 영상의 경우 이고 동적패턴은 음성과 같이 시간에 따라서 변하는 패턴에 해당한다.

동적 계획법이 다른 분할-정복(divide & conquer) 알고리즘 등과 구별되는 특징은 메모리에 해당하는 테이블 값과 점화식이 이루는 순환적인 성질을 이용한다는 것이다. 그런데, 모든 문제가 모두 동적 계획법으로 해결될 수 있는 것은 아니다. 어떠한 문제가 동적 계획법을 이용하여 해결 가능하기 위해서는 해당 문제가 최적화의 원리(principal of optimality)가 성립하여야 한다. 최적화의 원리가 적용되는 문제란 「한 문제에 대한 해가 최적이면 그 문제를 이루는 부분 문제들의 해도 최적이다」라는 명제가 성립하는 문제이다. 길이가 다른 두 열에서 어느 한 열을 기준으로 두 열을 비교하기 위해서는 어느 한 열이 신장(늘어남)되거나 축소(줄어듦) 되어야만 한다. 그림 6은 길이가 긴 A열을 길이가 작은 B열을 기준으로 비교하는 경우이다. 이 때 매핑 함수를 통하여 비교가 이루어지는데 이러한 매핑 함수가 직선과 같은 선형적인 경우를 「선형 신축 비교」라고 하고, 곡선과 같은 비선형적인 경우를 「비선형 신축 비교」라고 한다.

DTW 알고리즘을 이용하면 이러한 비선형 매핑 함수를 최적으로 찾아가면서 동시에 비교가 이루어진다. DTW 알고리즘은 일단 두 열의 각 성분에 대한 거리척도 값을 비용으로 설정 한다.

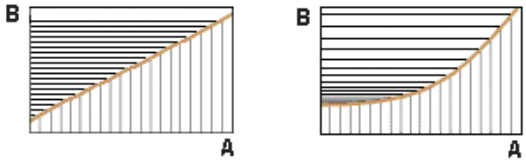

그림 1. 선형과 비선형 매핑
Fig. 1. Liner and Non-Liner Mapping

그리고 두 열이 이루는 격자(lattice)상에서 각 열의 시작 성분에서 시작하여 끝 성분에 이르기까지 비용 테이블에 최소 비용을 순환적으로 택하여 저장하는 점화식을 이용하는 동적 계획법으로 매핑 함수를 찾아가면서 두 열을 비교하는 알고리즘이다. 최종적으로 끝 성분에서 비용 테이블에 저장되는 비용 값이 두 열에 대한 유사도가 된다. 한편, 매핑 함수의 궤적은 앞의 동적 계획법의 최적 탐색패스를 찾는 것과 같이 탐색 과정에서 최소 비용을 택하는 경로를 별도의 경로 테이블에 매 단계마다 저장하고 끝 성분에서 최종 최소 비용을 구한 후에 역추적(backtracking)하여 찾게 된다.

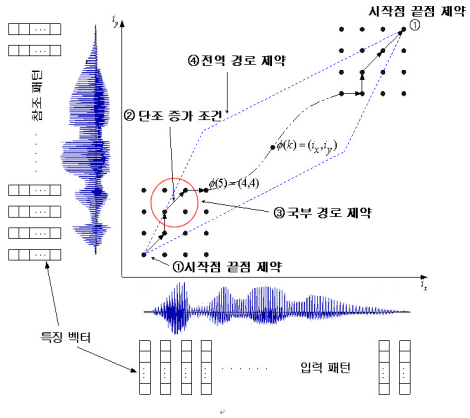

그림 2. DTW를 이용한 음성인식
Fig. 2. Voice recognition using DTW

그러므로 DTW 알고리즘은 열의 길이가 일치하지 않는 두 열의 유사도를 측정하는 매칭 알고리즘으로 안성맞춤이라고 할 수 있다. DTW 알고리즘은 주로 음성 인식에서 많이 이용한다. 비교 루틴이 아주 간결하고 단순하여 단어 단위의 간단한 음성 인식기에 적용 가능한 알고리즘이다.[4]

DTW를 이용한 음성 인식은 PCM 디지털 데이터를 그대로 사용하는 것이 아니라, 음성이 10-20ms 지속 시간 동안은 정상적(stationary)인 구간이라고 가정하고 행하는 단 구간 분석에 의하여 프레임 단위로 음성 특징벡터를 추출하는 전처리를 거친 후에 이루어진다. 그리고 인식 과정에서는 음성 인식 후보 단어 각각을 이러한 특징 추출 과정을 거쳐 기준 벡터 열로 미리 준비하여 두고, 인식할 단어에 대한 특징을 추출하여 시험 벡터 열을 각 후보 단어와 DTW 알고리즘을 이용하여 비교하여 최소가 되는 후보 단어 카테고리를 인식 결과로 결정하는 비교적 간단한 음성 인식 알고리즘이다.[5]

2. Noise Cancel DTW

인간이 발성하는 음성은 동일음 일지라도 발성자, 발성의 방법, 전후의 음운환경 등에 따라 시간적으로 불규칙하게 많은 차이가 있으며 스펙트럼의 형태도 다르게 나타난다. 이러한 시간축의 변동은 정상적인 발성에 비하여 30% 정도의 비선형 신축이 생기며 이것은 특히 패턴 매칭을 이용하는 음성인식에 있어서는 오인식의 원인이 된다. 변동의 최소화를 위해 비선형 시간 정규화가 요구되며 그 방법으로 DTW 알고리즘이 주로 사용 된다.

매칭을 실시하려고 하는 두 개의 패턴은 음성인식하기 위한 시험패턴(Test pattern) A와 이와 비교를 위한 구성되어지는 참조패턴(Reference pattern) B를 식(1)와 같은 특징의 시계열로 표현 할 수 있다.

$$A = a_1, a_2, a_3, \ldots, a_i \qquad (1)$$
$$B = a_1, a_2, a_3, \ldots, a_j$$

여기서, 패턴 A, B 사이의 시간적 대응은 i - j 평면상의 격자점 W로 표현할 수 있으며 식(2)와 같이 표현되는 열을 Warping 함수라고 한다.

$$S = W_1, W_2, W_3, \ldots, W_L \qquad (2)$$

이 함수가 매칭 경로가 되며 일반적으로 이 점열의 변화가 A와 B의 시간 대응의 변화에 해당된다. 특징 벡터 ai와 bj간의 차이를 구하는 척도로서 거리에 대한 개념을 도입하여 국소거리(Local distance) 즉 오차를 다음과 같이 표현한다.

$$d(W) = d(i,j) = |a_i - b_j|^2 \qquad (3)$$

여기서의 S 에 대한 누적 거리를 E(S)는 아래 식 (4)와 같이 나타낼 수 있다.

$$E(S) = \frac{\sum_{m=1}^{L} w_m d(W_m)}{\sum_{m=1}^{L} w_m} \qquad (4)$$

식(4)에서 점열 S를 변화시킬 때 E(S)의 최소치를 A와 B간의 거리로 정의해서 D(A,B)로 나타낸다.

$$D(A,B) = \min E(S) \qquad (5)$$

D(A,B)는 패턴 A와 B를 대응시킬 때 1 단계씩 부분적으로 최적인 경로만을 선택해 나가는 동적계획법을 (Dynamic programming, DP)을 이용하면 효율적으로 구할 수 있다.[6]

비선형 시간 정규화에 따른 방법에서의 DTW는 훌륭한 알고리즘이다. 하지만 모바일환경 즉, 잡음환경에서의 음성인식에 있어서는 인식률 저하를 막을 수 없다. 실제 생활환경에서의 잡음레벨은 통상 60dB ~ 80dB이며 이 정도의 잡음레벨에서의 음성에 대한 인식률은 현저히 떨어진다.

제안하는 방법은 깨끗한 음성데이터와 오염된 데이터를 동시에 사용하여 특징변환에 필요한 파리미터를 추정하여 인식하는 방식이다.

기존의 방식은 조용한 상태의 참조데이터를 생성하여 인식하는 방법으로 실제 주변잡음이 있다면 인식률은 현저히 떨어진다. 기존의 깨끗한 데이터와 잡음이 섞인 데이터를 참조데이터로 생성하여 인식하면 주변 잡음이 있는 상황일지라도 기존의 방법에 비해 인식률이 향상됨을 알 수 있다.

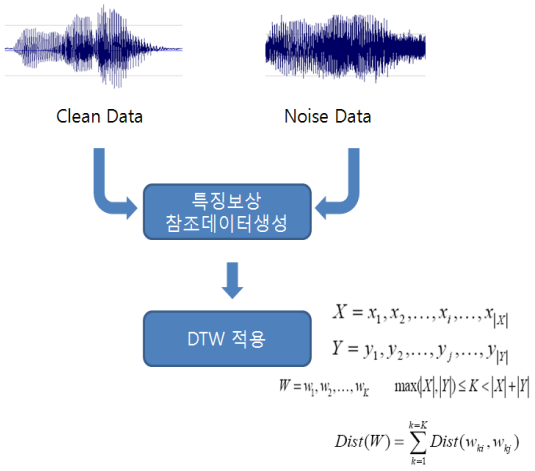

그림 3. Noise Cancel DTW 구조
Fig. 3. The Structure of Noise Cancel DTW

Ⅲ. 실험 및 인식 결과

본 논문에서의 특징보상은 두 가지의 방법으로 실험을 하였는데 우선 특징보상을 추정되는 White Gaussian Noise를 임의로 만들어서 원음에 부가시켜 참조패턴을 생성하여 실험하였다.

그리고 Data driven 기법으로 깨끗한 데이터와 오염된 데이터인 Real Life noise을 동시에 사용하여 인식실험을 하였다. 따라서 참조 모델은 3가지의 군으로 분류하여 실험 하였으며 Table 1과 같이 모바일 환경에서 사용될 수 있는 20개의 고립단어를 사용하였으며 사용된 음성데이터는 5명 화자에 대한 데이터를 사용하였다.

조건에 따라서 인식실험을 하였는데 인식결과는 평균 인식률은 그림51 그리고, Table 11, 12와 같이 나타내었으며 이후 각 Table별로 화자별 인식률을 나타 내었다. 화가간의 인식률이 다소 차이가 보이는데 이는 화자의 각 발성 음성에 대한 명료도 및 성량과 관계되어진 결과로 사료된다.

잡음환경(잡음 60~80dB) 에서의 실험을 위하여 Table 1와 같이 모바일 기기에 사용 가능한 20개의 음절데이터를 사용 하였고 신호 대 잡음비 각 20dB, 10dB, 5dB 로 실험 하였다.

표 1. 인식단어 리스트
Table 1. Recogntion Word List

NO	음절
1	전화
2	예
3	아니오
4	연결
5	다음
6	이전
7	취소
8	음악재생
9	멈춤
10	정지
11	전곡재생
12	통화
13	검색
14	메뉴
15	영상통화
16	메시지
17	카메라
18	메모
19	사진
20	일정

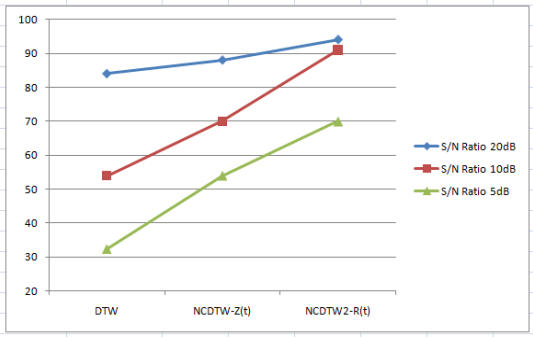

그림 4. NCDTW 인식률(%)
Fig. 4. Recognition Rate for NCDTW(%)

표 2. DTW 인식률(%)
Table 2. DTW Recognition rate(%)

	참조데이터	S/N Ratio	평균 인식률(%)
DTW	Clean voice data	20 dB	84
		10 dB	54
		5 dB	32

표 3. NCDTW 인식률(%)
Table 3. NCDTW Recognition rate(%)

	참조데이터	S/N Ratio	평균 인식률(%)
NCDTW	Clean voice data + Gaussian noise	20 dB	88
		10 dB	70
		5 dB	54
	Clean voice data + Real life voice data	20 dB	94
		10 dB	91
		5 dB	70

표 2,3에서 알 수 있듯이 기존의 방법에 비해서 약 10%이상 향상된 결과를 보였고 특히 신호 대 잡음비가 10dB이하에서는 인식률의 편차가 큰데 이는 참조패턴이 신호 대 잡음비가 10dB 전후 이라고 판단되는 결과로 사료된다.

Ⅳ. 결 론

본 논문에서 제안한 방법인 NCDTW의 경우 기존의 방법인 DTW에 비해 잡음환경에서 약 10% 향상된 94% 이상의 인식률을 얻을 수 있었다. 특히, 주변 잡음이 심한 자동차 도로 상황에서의 비교 데이터에 대해서도 비교적 양호한 결과를 나타내었는데 이는 잡음환경에서의 데이터를 함께 사용하여 주변 잡음의 특성을 많이 반영시킨 결과이라고 사료되어진다. 그러나 주변잡음이 상대적으로 많은 지역에서는 실험화자가 다소 크게 발성해야 인식 가능한 레벨에 도달 할 수 있을 것으로 예측되어진다. 그리고 NCDTW의 경우 데이터량이 많아 질 경우 인식 속도가 많이 걸린다는 단점은 여전히 연구과제로 남아있지만 고립단어 인식 시 단어수가 한정될 경우에는 적용 가능한 알고리즘으로 판단되어진다.

참 고 문 헌

[1] 안종영, 김영섭, 김수훈, 허강인, “자동차 ECU제어를 위한 음성인식 패턴매칭레벨에 관한 연구,” 한국인터넷방송통신학회 논문지 제10권 제1호 pp.75-80. 2010.
[2] L.Rabiner and B.H.Jung, "Fundamentals of Speech Recognition", PTR Prentice Hall, 1993.
[3] H.Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 263-271, Apr. 1984.
[4] 한학용저 “패턴인식개론” ,한빛미디어
[5] 안종영, 김영섭, 허강인, “차량용 BCM을 위한 음성인식 시스템 설계” 한국 인터넷 방송통신학회 추계학술 대회 논문집 pp. 169-171. 2009.
[6] 이종진, “한국어 연속음성 인식시스템의 구현”, 博士學位 論文, 1994.

※ 본 논문은 동아대학교 학술연구비 지원에 의하여 연구되었음.

저자 소개

안 종 영(정회원)

- 1993년 : 동아대학교 전자공학과 공학사
- 1996년 : 동아대학교 전자공학과 공학석사
- 2011년 : 동아대학교 전자공학과 공학박사
- 1996-2000 ;현대모비스(현) 전임연구원
- 2001-2003:한국폴리텍 아산캠퍼스 영상매체과 교수
- 2004-2006 : (주)대성전기 선임연구원

<주관심분야 : 음성신호처리, 임베디드 시스템, DSP, 전장 ECU>

김 성 수(정회원)

- 1989 : 건국대학교 전자공학과 공학사
- 1992 : 건국대학교 전자공학과 공학석사
- 1998 : 건국대학교 전자공학과 공학박사
- 1992-1996 : 대우전자(주) 중앙연구소 선임연구원
- 1996-현 : 한국폴리텍Ⅱ 컴퓨터정보과 교수

<주관심분야 : 임베디드 시스템, 영상신호처리, 무선통신>

김 수 훈 (정회원)

- 1990년 : 동아대학교 전자공학과 공학사
- 1992년 : 동아대학교 전자공학과 공학석사
- 1999년 : 동아대학교 전자공학과 공학박사

2001년~현: 부천대학 모바일통신과 부교수
<주관심분야 : DSP, 음성인식, 모바일콘텐츠>

허 강 인(정회원) :교신저자

- 1980년 : 동아대학교 전자공학과 공학사
- 1982년 : 동아대학교 전자공학과 공학석사
- 1990년 : 경희대학교 전자공학과 공학박사
- 1998년 9월~1989년 8월 일본 쓰쿠바 대학 객원연구원
- 1992년 9월~1993년 8월 일본 도요하시대학 객원연구원
- 1984년-현: 동아대학교 전자공학과 교수

<주관심분야 : DSP, 음성인식, 음성합성, 신경회로망>

고 시 영

- 1979년 : 영남대학교 전자공학과 공학사
- 1983년 : 영남대학교 전자공학과 공학석사
- 1992년 : 동아대학교 전자공학과 공학박사
- 1972년~1979년 한국전자 연구소
- 1986년-현: 경일대학교 전자공학과 교수

<주관심분야 : DSP, 음성신호처리, 회로이론, 신경회로망>