피춰 추출 관점에서 기준 화자 수 증가에 따른 음성 인식 성능 분석

Performance Analysis of Speech Recognition by Increasing the Number of Reference Speaker

현 동 훈, 이 철 희
연세대학교 전자공학과
120-749 서울시 서대문구 신촌동 134
(02) 361-2779

Donghoon Hyun and Chulhee Lee
Dept. of Electronic Engineering, Yonsei University
E-Mail: hyungari@champ.yonsei.ac.kr

요 약

음성을 인식하기 위해서는 주어진 음성을 미리 정한 기준 음성과 비교하여 가장 유사한 것을 찾는 과정을 거치게 된다. 같은 단어라도 화자에 따라서 발음 속도, 음의 강약이 틀리므로 화자 독립 음성 인식을 위해서는 여러 화자가 발음한 음성을 기준 음성으로 사용하여 인식 성능을 향상시킬 수 있다. 그러나 화자 수를 증가시켜도 인식 성능의 향상에는 한계를 보이고 있다. 이러한 문제점은 현재 음성에서 추출되는 피춰가 인식에 필요한 정보를 충분히 포함하지 않는 것과 인식 알고리즘의 효율성 등에서 원인을 찾을 수 있다. 본 논문에서는 남자 10명과 여자 10명이 발음한 한국어 숫자음을 인식 대상으로 하여 멜켑스트럼을 추출하고 DTW에 의해 인식을 수행하여 피춰 추출의 관점에서 화자 수 증가에 따른 인식률의 변화와 그 한계에 대해서 분석한다.

Ⅰ. 서 론

음성 인식 시스템은 입력 음성에서 특징을 추출하는 피춰 추출 부분과 입력 피춰를 기준 피춰와 비교하는 패턴 비교 부분으로 구성되어 있다. 음성의 피춰로 LPC 켑스트럼과 멜켑스트럼을 널리 사용하고 있고 패턴 비교 단계에서는 DTW(dynamic time warping)와 HMM(hidden Markov model) 등을 사용하고 있다. DTW는 기준 음성과 입력 음성의 시간적 차이를 보상하면서 유사도를 측정하는 패턴 매칭 방법이므로 기준 음성의 수가 증가할수록 인식 시간도 증가한다. 따라서 어휘 규모가 큰 대규모 인식 시스템에서 DTW를 사용하는 것은 인식 시간을 고려할 때 부적절하다. 그러나 어휘의 수가 제한된 상황에서 DTW를 이용한 화자 독립 인식 시스템의 인식 성능을 향상시키기 위해서는 여러 명이 발음한 음성을 기준 음성으로 사용하여야 하는데, 앞에서 언급한 인식 시간의 문제로 인해 기준 화자 수를 무조건 증가시킬 수 없게 된다. 따라서 기준 화자 수에 따른 인식 성능의 변화에 대한 구체적인 분석과 인식 성능의 한계에 대한 고찰이 요구되며 이를 통해서 현재 사용되는 피춰의 한계점을 찾는 계기를 마련할 수 있다.

이를 위해 본 논문에서는 화자 독립 인식에서 기준 화자 수 증가와 피춰의 차원에 따른 인식 성능을 분석하고 그 한계에 대해서 고찰한다. Ⅱ장에서 음성 신호로부터 멜켑스트럼을 구하는 과정을 간략히 설명하고 Ⅲ장에서는 기준 화자 수와 멜켑스트럼의 차수 변화에 따른 인식 성능의 변화를 분석하고 Ⅳ장에서 결론을 맺는다.

Ⅱ. 멜켑스트럼 추출

현재 음성 인식에서 널리 사용되는 피춰로는 LPC 계수로부터 유도되는 LPC 켑스트럼과 인간의 청각 특성을 이용한 멜켑스트럼 등이 있는데, 본 논문에서는 멜켑스트럼을 피춰로 사용한다. 멜켑스트럼은 critical band 필터를 사용하여 그림 1과 같은 과정으로 구한다 [1].

이때 critical band 필터의 중심 간격(C, mel 단위)

과 대역폭(B, mel 단위)을 일정하게 하여 멜켑스트럼을 구하는데[2] (C, B)의 값을 변화시킴으로써 같은 음성에 대해서 다른 멜켑스트럼을 구할 수 있다[3]. 본 논문에서는 여러 가지의 (C, B)를 사용하여 멜켑스트럼을 구하고 각각의 경우에 대해서 인식 성능을 관찰한다.

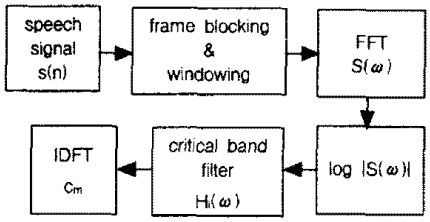


그림 1. 멜켑스트럼을 구하는 과정.

Ⅲ. 실험 및 결과

본 논문에서는 한국어 숫자음을 실험 대상으로 하여 인식 성능을 분석하였다. 실험을 위한 음성 데이터는 표 1에 나타나 있고, 주어진 음성 데이터로부터 표 2의 방법을 사용하여 피춰를 추출하였다. 또한 DTW에서 사용한 local path constraint와 path weighting은 그림 2와 같다.

표 1. 실험에 사용된 음성 데이터.

실험용 음성 데이터	한국어 숫자음 0(영)~9(구) 남자 10명, 여자 10명이 10번씩 발음 (총 2000개의 음성)
샘플링 주파수	11.025 kHz (16 bits 양자화)
끝점 검출	수동
기준 음성	화자 2명, 4명, 6명, 8명, 10명, 12명, 14명을 각각 임의로 30회 선택 (남자와 여자의 수는 동일)
테스트 음성	기준 음성을 제외한 실험용 음성 데이터

표 2. 피춰 추출 방법.

분석 프레임	300 samples (27.2ms)
프레임 간격	100 samples (9.1ms)
window 함수	Hamming window
FFT	1024-point FFT
피춰 벡터	멜켑스트럼 (10, 12, 14, 16차)

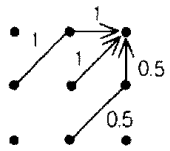
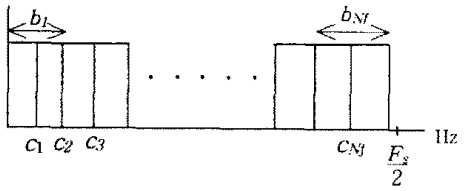


그림 2. Local path constraint and path weighting.



$c_i = c_{i-1} + C$, $b_i = B$ $(C, B$: *mel unit*$)$

그림 3. Critical band 필터.

멜켑스트럼을 구하기 위해서 그림 3의 critical band 필터를 사용하였는데 (C, B)를 각각 (75, 75), (100, 100), (125, 125)로 설정하여 10, 12, 14, 16차의 멜켑스트럼을 구하고 기준 화자 수를 증가시키면서 인식 성능을 관찰하였다. 그림 4는 각각의 기준 화자 수에 대해서 기준 음성을 임의로 30개 선택하여 인식률의 평균과 표준편차를 나타낸 것이다. 기준 화자 수가 증가할수록 인식 성능이 대체적으로 향상되었지만 멜켑스트럼 차수의 증가에 따른 인식 성능의 변화는 크게 나타나지 않았다. 특히, 12, 14, 16차 멜켑스트럼은 기준 화자 수가 증가할수록 비슷한 인식 성능을 보여주었다.

그림 5는 멜켑스트럼의 차수의 변화와 기준 화자 수의 변화에 따른 숫자음별 인식 성능을 보여준다. 멜켑스트럼의 차수에 따른 인식 성능의 변화를 살펴보면, 숫자음 5와 9의 경우에 멜켑스트럼의 차수가 증가함에 따라 인식 성능이 크게 향상되었으며 숫자음 1과 7은 오히려 인식 성능이 감소되었다. 그 외의 경우에는 멜켑스트럼의 차수에 따른 인식 성능의 변화가 크게 나타나지 않았다. 또한 기준 화자 수의 변화에 따른 인식 성능의 변화를 관찰하면, 숫자음 3에서는 화자 수가 증가함에 따라 성능이 감소하는 경우가 발생하였고 대부분의 숫자음에서 기준 화자 수가 증가할수록 인식률의 증가량이 감소하는 것을 나타내었다.

Ⅳ. 결 론

본 논문에서는 한국어 숫자음에 대해서 멜켑스트럼의 차수와 기준 화자 수의 증가에 따른 인식 성능

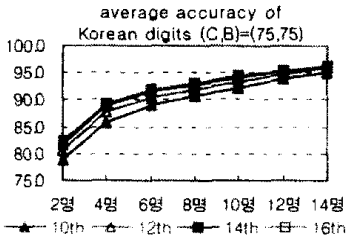
의 변화를 관찰하였다. 숫자음별 인식률을 고려할 때 멜켑스트럼의 차수가 증가할수록 대부분 인식 성능이 향상되었지만 몇몇 숫자음에서는 감소하였다. 또한 기준 화자 수가 증가함에 따라 대부분의 숫자음에서 인식률이 증가하였으나 증가량은 점차 감소하였다. 특히 숫자음 3에서는 인식률이 감소하는 경우도 발생하였다. 전체적인 인식 성능을 고려할 때 기준 화자 수의 증가에 따라 인식 성능이 향상되는 것이 관찰되었지만 몇몇 숫자음에서는 성능의 한계를 나타내었다.

참고문헌

[1] J. R. Deller J. R. J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice Hall, 1987.

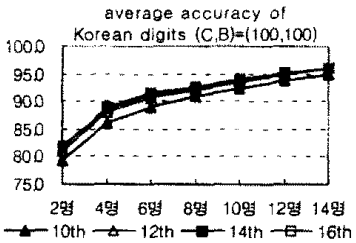
[2] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 357-366, Aug. 1980.

[3] 현동훈, 이철희, "멜켑스트럼의 성능 향상을 위한 critical band 필터의 최적화," 한국음향학회 학술발표대회 논문집, 제17권, 2호, pp. 403-406, 1998년 11월.



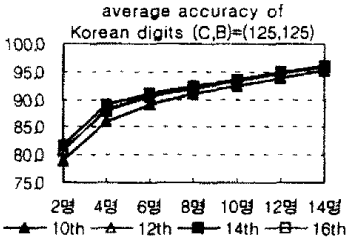
(a)

	10th	12th	14th	16th
2명	4.8	4.8	4.7	4.7
4명	2.3	2.2	1.8	1.9
6명	1.8	1.4	1.2	1.3
8명	1.3	1.3	1.2	1.3
10명	0.8	0.7	0.8	0.9
12명	0.6	0.6	0.6	0.7
14명	0.7	0.7	0.8	0.8



(b)

	10th	12th	14th	16th
2명	4.8	4.7	4.6	4.7
4명	2.2	2.1	1.8	1.7
6명	1.7	1.5	1.2	1.2
8명	1.3	1.3	1.2	1.2
10명	0.7	0.8	1.0	1.0
12명	0.7	0.6	0.6	0.7
14명	0.7	0.8	0.8	0.8



(c)

	10th	12th	14th	16th
2명	4.8	4.8	4.8	4.8
4명	2.3	2.2	2.1	2.3
6명	1.7	1.6	1.7	1.9
8명	1.4	1.4	1.4	1.4
10명	0.8	0.7	0.7	0.7
12명	0.6	0.6	0.6	0.7
14명	0.7	0.6	0.7	0.6

그림 4. 기준 화자 수 증가에 따른 인식률의 변화와 인식률의 표준편차. (% 단위)

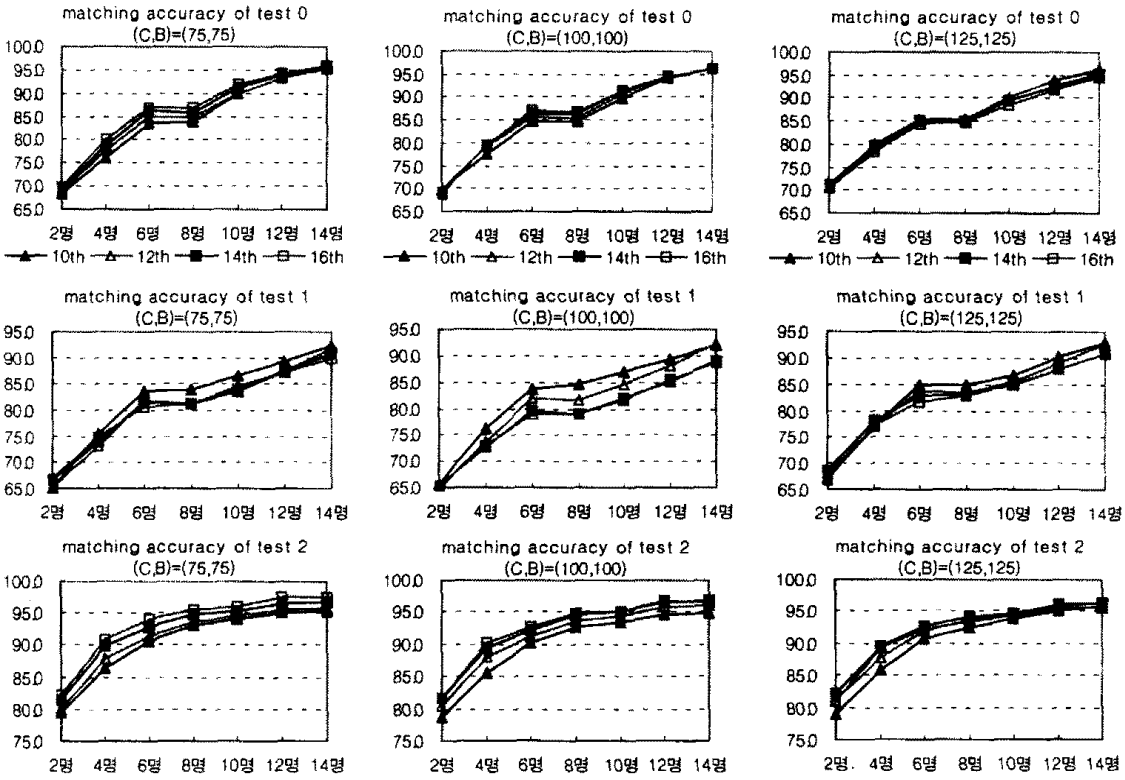


그림 5. 기준 화자 수 증가에 따른 각 숫자음별 인식률의 변화.

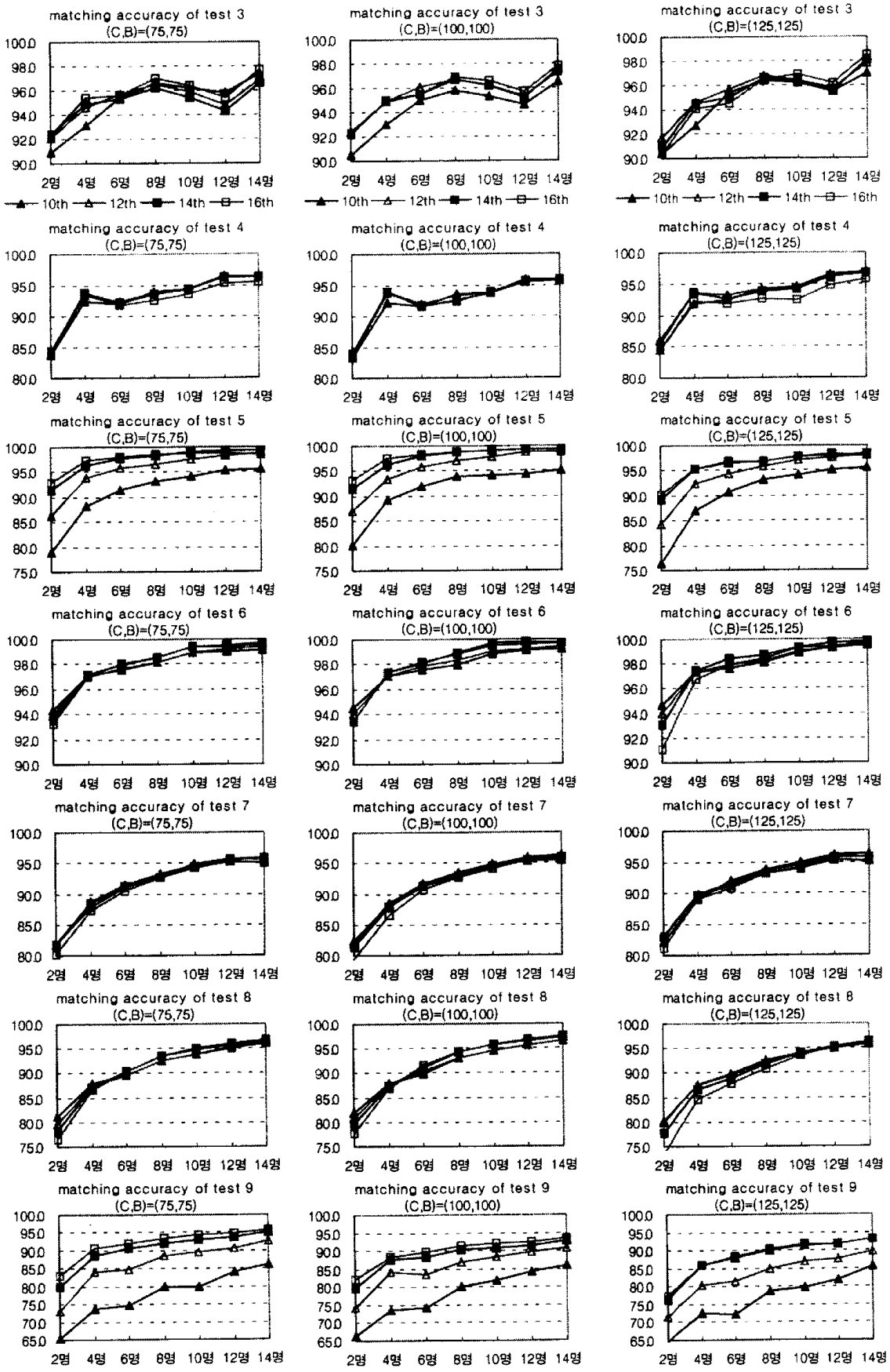


그림 5. (continued)