대역폭 변화에 따른 음성 인식률 비교연구

손 일현, 도 삼주, 구 명완

한국통신 연구개발단 기초기술1연구실

A Comparative Study of Recognition Rate According to the Variance of Speech Bandwidth

Il-Hyun Sohn, Sam-Joo Doh and Myoung-Wan Koo

Basic Research Section 1, Korea Telecom Research Center

요 약

이 논문에서는 123개 단어의 한국어 음성에 대하여 음성의 대역폭 변화에 따른 인식률을 비교하였다. 인식률 비교실험을 위해 hidden Markov model과 음소와 유사한 131개의 한국어 subword 유니트를 사용한 화자독립 격리단어 인식 시스템을 사용하였다. 이 실험은 대역폭이 각각 0 – 4.5kHz 및 0.3 – 3.3kHz인 두가지 종류의 음성 데이타베이스를 사용하였다. 훈련과정에서 corrective training의 반복회수를 2로 하고 state transition duration 정보를 사용하였을 때, 0 – 4.5kHz 와 0.3 – 3.3kHz 대역폭에 대해 각각 98.8 % 및 98.2 % 의 최고 인식률을 얻었다. 이로부터 전화대역폭에서도 음성인식률은 크게 저하되지 않음을 알 수 있다.

I. 서 론

음성인식은 인간과 기계간의 자연스러운 인터페이스의 한 방법으로서 연구되어 왔으며, 정보검색 시스템에 효율적으로 사용될 수 있다. 특히 전화선을 통한 음성을 인식할 수 있는 기술은 음성을 이용한 공중 정보검색 시스템에 매우 유용하게 사용될 수 있을 것이다.

전화선을 통한 음성의 인식에 영향을 미치는 요소는 여러가지가 있지만 이 논문에서는 이들 요소 중 대역폭의 제한에 초점을 맞추었으며 대역폭의 변화에 따른 음성의 인식률을 비교하였다. 인식률 비교를 위해 대역폭이 각각 0 - 4.5 kHz와 0.3 - 3.3 kHz인 두가지 종류의 음성 데이타베이스를 사용하였다. 실험에는 화자독립 격리단어 인식방식으로 123개의 단어를 인식하는 부서안내 시스템을 사용하였다.

제2절에는 실험에 사용한 음성인식 시스템의 개관을 기술하였고, 3절에는 실험에 사용한 데이타베이스를 설명하였다. 그리고 4절에서 실험 및 그 결과를 제시하였고 5절에서 결론을 맺었다.

II. 실험 시스템 개관

이 논문에서 실험을 위해 사용한 부서안내 시스템은 hidden Markov model(HMM)을 기반으로 하는 화자독립 격리단어 인식 시스템이다[1]. 그림1은 훈련과정의 시스템 개요을 나타낸 것이다.

입력된 음성은 4.5kHz 저역통과 필터를 거쳐 12bit 디지탈 신호로 변환되고, 음성의 끝점이 자동검출된다. 훈련시에는 끝점 검출과정에서의 약간의 오류는 수동으로 수정한다. 다음으로 이 음성신호는 전달함수가 $1 - 0.95z^{-1}$인 필터를 통해 pre-emphasis된다. 이 신호

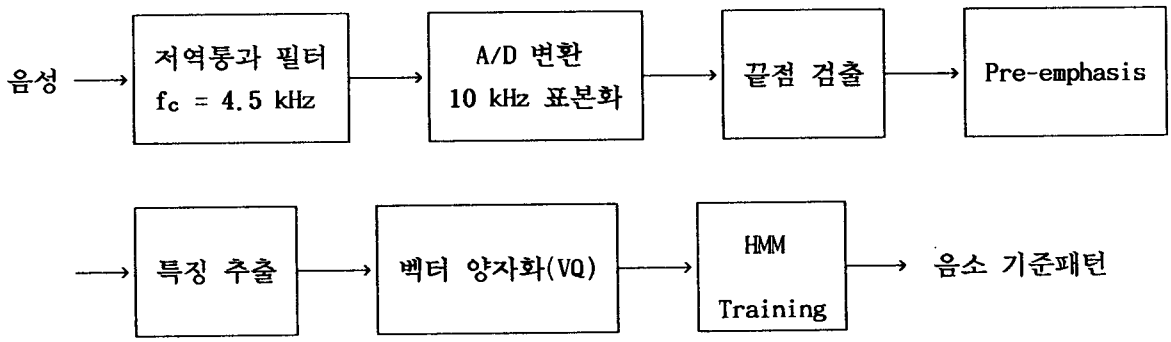


그림 1: HMM 훈련 과정

로부터 매 10msec 단위로 특징 파라미터를 계산한다. 다음에 이 값은 벡터 양자화(vector quantization) 된다. 벡터 양자화에는 256 codeword를 갖는 3개의 VQ codebook을 사용하였다. 3개의 codebook은 첫째, 12개의 mel-scale된 LPC cepstral 계수, 둘째, cepstral 계수의 차, 세째, 정규화된 log power와 그 값의 차이다. VQ 알고리즘은 Linde-Buzo-Gray 알고리즘을 사용하였다. 이 논문에서는 3개의 출력 확률밀도함수가 독립이라 가정하였다. 이 경우 출력 확률은 3개의 출력확률의 곱이된다.

실험 시스팀은 음소와 유사한, context dependent 모델의 131개 subword 유니트를 사용하였다. 먼저 대상어휘에 필요한 한국어 음소를 context independent 모델에 의해 44개의 유니트로 나타내고, 이것을 다시 131개의 context dependent 유니트로 확장한 것이다. 이 시스팀에서는 5개의 연속적인 context에 의해 결정되는 pentaphone을 사용하였다[2]. 이 시스팀의 HMM모델 topology는 K. -F. Lee가 사용한 모델과 유사한것으로 7개의 state와 12개의 transition으로 구성된다[3].

III. 음성 데이타베이스

이 논문에서는 두 종류의 음성 데이타를 사용하여 실험을 하였다. 먼저 대역폭이 0 – 4.5kHz 이고 표본화율이 10kHz인 음성 데이타를 구성하였다. 다음으로 이 데이타를 대역폭이 0.3 – 3.3kHz이고 표본화율이 8kHz인 또하나의 음성 데이타로 변환하였다. 2번째 음성데이타의 대역폭은 전화음성의 대역폭과 같다. 그림2에 표본화율 변환방법을 나타내었다[4].

여기서 사용한 필터는 98개의 계수를 갖는 FIR 디지탈 필터이며, 필터계수는 Remez ex-

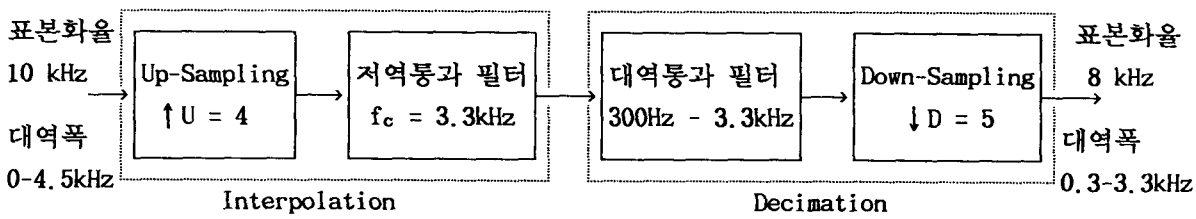


그림 2: 표본화율 변환

change 알고리즘을 사용하여 구하였다. 그림3은 이 필터의 주파수 특성을 나타낸 것이다.

IV. 실험 및 결과

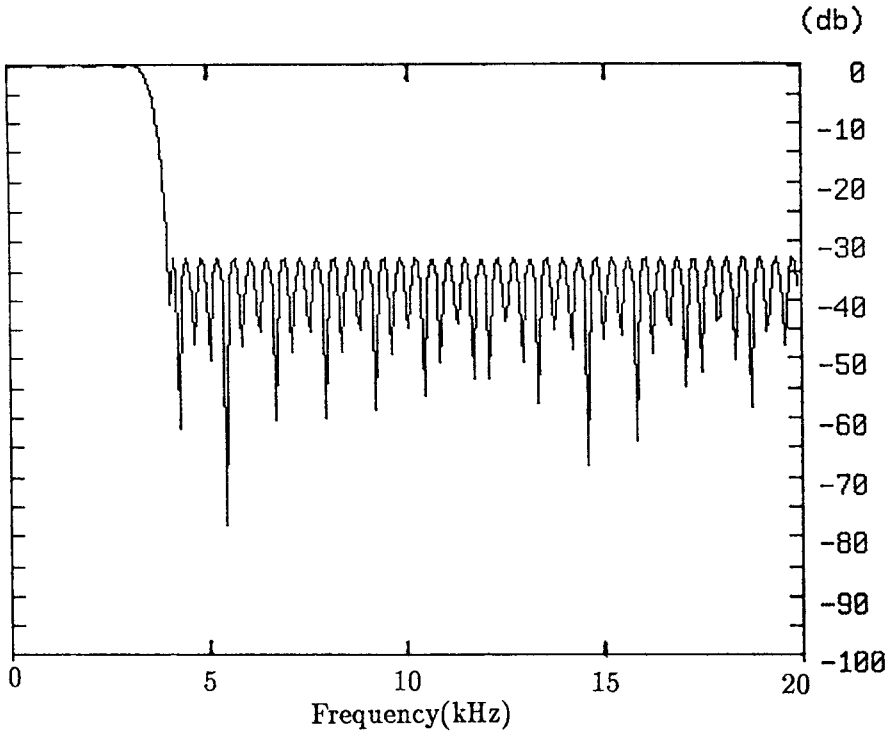
실험에서 사용한 음성인식 시스팀은 평균 lexicon의 길이가 16.0인 123개 단어를 인식하는 시스팀이다. 훈련 및 인식을 위한 음성 데이타는 연령이 27 – 35세인 10명의 남성화자가 조용한 사무실 환경에서 각 단어를 2회씩 발음한 것을 사용하였다. 10명의 화자의 음성 데이타 중 8명분의 데이타는 HMM훈련을 위해 사용되었으며 나머지 2명분의 데이타는 인식률 시험을 위해 사용되었다. HMM 초기 훈련을 위해 492개 단어(123단어×4명×1회)에 대한 데이타는 수동으로 segmentation을 하여 사용하였다. 음성인식에서 대역폭의 영향을 알기 위하여 대역폭이 0 – 4.5kHz와 0.3 – 3.3 kHz인 2개 그룹의 음성 데이타를 사용하였다. 훈련과 시험은 그룹별로 각각 실시하였다.

검색시간 감축을 위해 Viterbi beam 검색 알고리즘을 사용하여 beam 폭의 변화에 따른 인식률을 구하였다. 다음으로 state transition duration 정보를 사용하고 훈련시의 iteration 회수를 2로하여 corrective training(CT) 알고리즘을 사용하는 인식 시스팀에서 동일한 실험을 수행하여, 이것을 사용하지 않은 경우와 비교하였다. 인식실험 결과는 표 1 및 그림 4 와 같다.

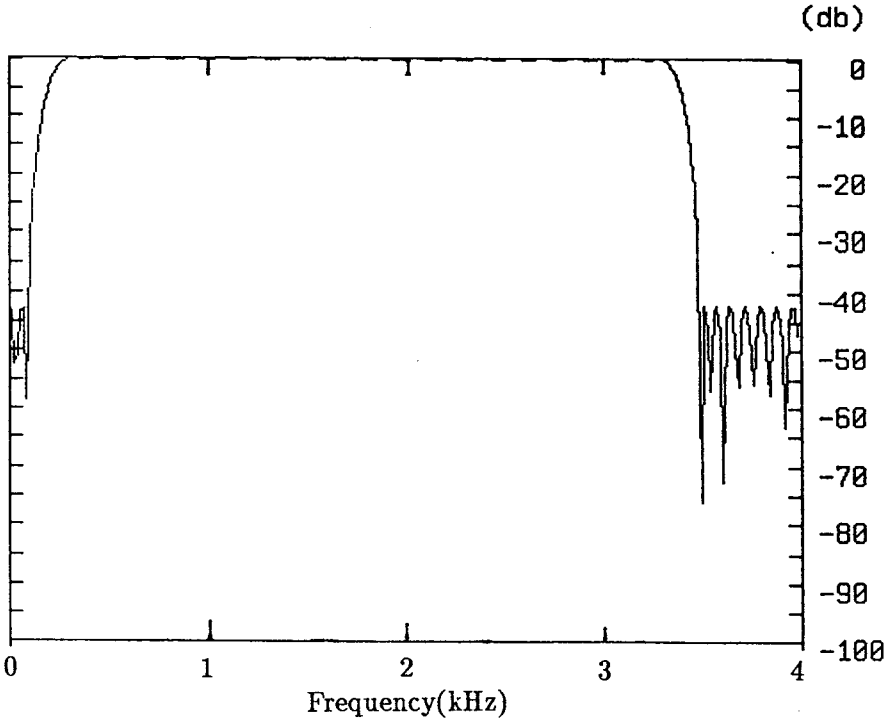
두 그룹 모두 beam 폭이 증가할 수록 오인식률이 감소하였는데 beam 폭이 10이상일 때는 인식률에 크게 영향을 미치지 않았다. 또 CT를 사용하였을 경우에는 이를 사용하지 않은 경우보다 오인식률이 약 반으로 줄어듬을 알 수 있다. 이 때 beam 폭이 증가하면 차이가 줄어드는 경향을 보였다. 대역폭이 다른 두 그룹의 경우 전화대역폭 음성의 오인식률이 더 컸으나 전체적으로 그 차이가 적음을 알 수 있다.

V. 결 론

이 논문에서는 대역폭이 각각 0 – 4.5kHz와 0.3 – 3.3kHz인 두가지 종류의 음성 데이타에 대해 인식률을 비교하였다. 실험용 음성인식 시스팀은 HMM을 사용하고 음소와 유사한 131개 subword 유니트를 갖는 화자독립 격리단어 인식 시스팀을 사용하였다. 실험 결과 두 종류의 음성에 대한 인식률은 큰 차이가 없음을 알 수 있었다. 이것은 전화선 대역폭 음성에 대해서도



(가)



(나)

그림 3: 필터의 주파수 특성

(가) 저역통과 필터 (b) 대역통과 필터

표 1. 인식 결과
(가) CT 와 state duration 정보를 사용하지 않았을 때
(나) CT 와 state duration 정보를 사용하였을 때

(가)

		인식률(%)					
Beam 폭		5	7	8	10	20	30
대역폭	4.5 kHz	94.5	97.8	98.0	97.6	97.8	97.8
	0.3-3.3 kHz	92.1	95.9	96.1	96.3	97.8	97.8

(나)

		인식률(%)					
Beam 폭		5	7	8	10	20	30
대역폭	4.5 kHz	95.7	97.6	97.6	98.6	98.8	98.8
	0.3-3.3 kHz	95.7	97.6	97.8	98.2	98.2	98.2

높은 음성인식률을 얻을 수 있음을 나타낸 것으로, 앞으로 공중전화망을 통한 음성인식 정보서비스의 실용화를 위해서는 실제의 전화음성에 대한 인식실험을 하여야 할 것이다.

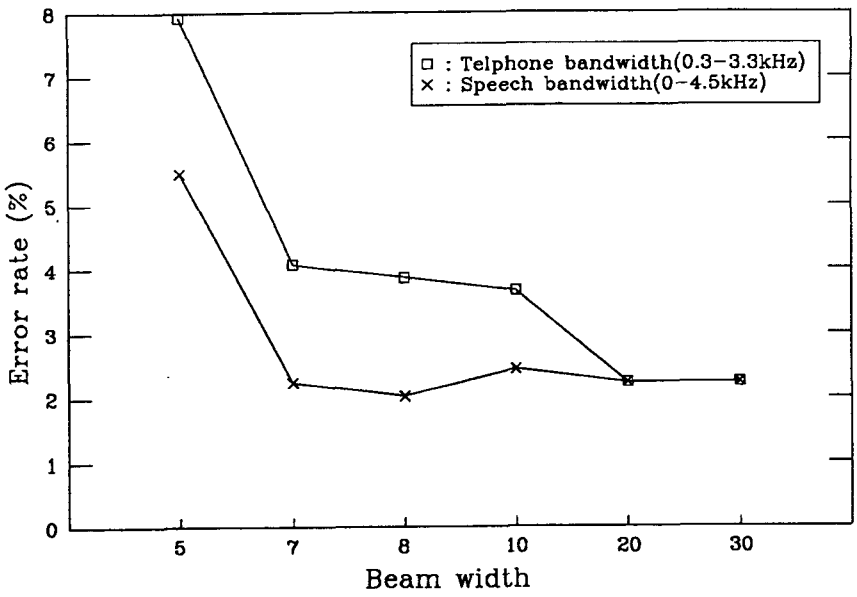
참고 문헌

[1] 구명완 외, "음성인식기술을 이용한 정보검색 시스템," 음성통신 및 신호처리 워크샵, pp. 251-256, 1992. 8.

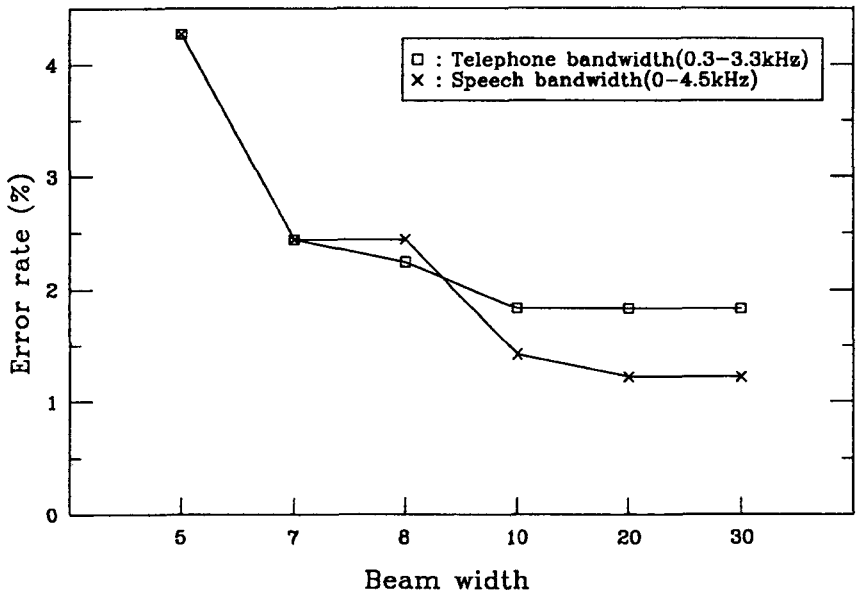
[2] C. H. Lee, et al., "Acoustic modeling of subword units for speech recognition," in *Proc. 1990 IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 721-724, April 1990.

[3] K. -F. Lee, *Automatic speech recognition: the developement of the SPHINX system,* Kluwer Academic Publisher, Norwell, Mass., 1989.

[4] J. G. Proakis, et al., *Introduction to digital signal processing,* Macmillan Publishing Company, New York, NY., 1988.



(가)



(나)

그림 4: 인식 결과

(가) CT 와 state duration 정보를 사용하지 않았을 때

(나) CT 와 state duration 정보를 사용하였을 때