2002년도 한국음향학회 학술발표대회 논문집 제21권 제1(s)호

전화망 환경에서의 연속숫자음 인식 성능평가

김성탁*, 김상진*, 정호영**, 김회린*, 한민수*

* 한국정보통신대학교(ICU) 공학부
** 한국전자통신연구원(ETRI) 음성정보연구센터

Performance Evaluation of Telephone Continuous Digit Recognition

SungTak Kim*, SangJin Kim*, Hoyoung Jung**, Hoirin Kim*, Minsoo Hahn*

* School of Engineering, ICU
** Speech Information Technology Center, ETRI
E-mail : stkim@icu.ac.kr

요 약

한국어 숫자는 단음절로 이루어져 있고, 연속적으로 발음할 때 조음현상에 의해 발음이 심하게 변하고, 숫자간의 경계를 규정하기가 어려워 진다. 특히 잡음환경에서는 한국어의 무성음인 자음구간의 주파수 특징이 많이 왜곡되어 성능이 저하된다. 본 논문에서는 전화망에서의 고성능 연속숫자음 인식기 개발을 위하여 그 첫 단계로서 다양한 조건에서 MFCC 특징계수를 구하는 방법들과 문맥독립 및 문맥종속 HMM의 상태수 및 각 상태에서의 mixture 수 변화에 대한 성능을 분석해본다. 음향모델로는 문맥독립 모델인 음소와 문맥종속 모델인 triphone 모델을 모두 평가하였다.

1. 서 론

현재 연속음성인식(ASR) 시스템의 가장 큰 문제는 잡음환경에서의 인식율 저하에 있다. 이 인식율 저하의 이유는 훈련환경과 테스트 환경의 불일치, 그리고 보편적으로 사용되어지고 있는 특징벡터, MFCC (Mel-Frequency Cepstrum Coefficient)가 잡음환경에서 성능이 저하된다는 사실에 있다[1]. 특히 이 논문에서 다루고자 하는 한국어 숫자음의 경우 숫자음 자체가 단음절로 되어있어 인식이 어려운데다가, 잡음환경에서 자음 모델링 시, 중요한 요소인 고주파영역의 왜곡으로 인해 성능이 떨어진다. 이런 사실에 입각하여 MFCC 특징벡터를 이용한 전화망환경에서의 한국어숫자음 인식실험을 통해 그 특징벡터와 HMM의 구성에 따른 성능의 변화를 관찰함으로써 전화망에서의 숫자음 인식에 대한 기존 알고리즘의 한계를 다각

도로 분석해서 향후 어떤 측면에서 성능을 개선해야 하는지를 검토하는 기본자료를 제공하고자 한다. 전화망 환경에서의 한국어 숫자음인식에 대한 성능평가는 다음 항목을 기준으로 평가하였다.

(1) 특징벡터 추출 시

- 20ms Hamming Window의 shifting rate
- Mel-Scale Filter bank의 수
- 주파수 영역에서의 에너지(C0) 사용유무
- CMN (Cepstrum Mean Normalization) 수행 유무

(2) HMM 구성 시

- Model의 state 수
- Model의 mixture 수

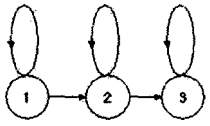
2. 음성 DB 및 기본 인식시스템

2.1 음성 DB

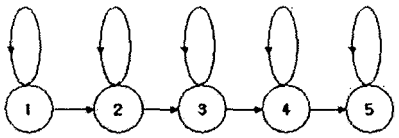
본 논문에서 사용한 음성 DB는 전화망환경에서 구축한 것으로, 8kHz로 샘플링, 16bit PCM 형식으로 저장되어 있다. 한국어 단연숫자음 "공", "영", "일", "이", "삼", "사", "오", "육", "칠", "팔", 그리고 "구"의 조합으로 사연숫자를 발성하였다. 실험에 사용된 음성샘플은 총 60,064개로 남성화자가 183명으로 37,554개의 발성을 여성화자가 111명으로 22,510개의 발성을 하였다. 실험 시, 훈련용으로 남성화자 153명, 여성화자 93명으로 구성된 49,999개의 음성샘플을 사용하였고, 테스트용으로 남성화자 30명, 여성화자 18명인 10,065개의 음성샘플을 사용하였다.

2.2 기본 인식시스템

실험에 사용된 인식시스템은 그림 1과 같은 left-to-right 연속밀도 HMM[2]을 사용하였고, 검색 네트워크 구조는 자릿수 제약을 둔 그림 2와 같은 구조를 사용하였다. Triphone 단위 인식실험에서 훈련 데이터에서 나타나지 않는 unseen triphone 문제를 해결하기 위해, decision tree-based state tying[3]을 수행하여 각각의 leaf에 있는 HMM state를 tying 하였다. 방법은 그림 3과 같다.



(가) State가 3개인 HMM



(나) State가 5개인 HMM

그림 1. HMM 구조

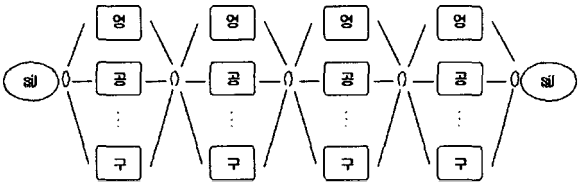


그림 2. 검색 네트워크 구조

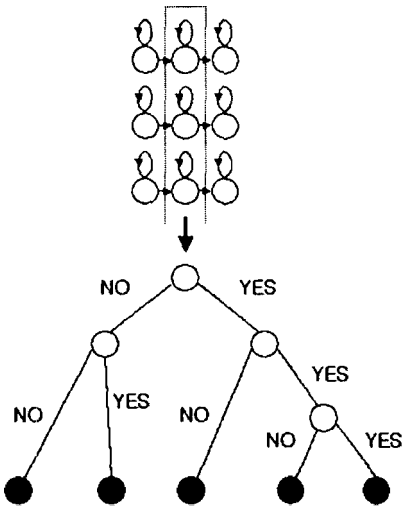


그림 3. Decision tree-based state tying.

3. 성능평가 결과

3.1 특징벡터 추출 조건에 따른 비교 평가

이 실험에서는 MFCC 특징벡터 추출 시, 윈도우 shifting rate, 필터뱅크의 수, 주파수 영역에서의 에너지(MFCC에서의 C0)의 사용유무, 그리고 CMN (Cepstrum Mean Normalization)의 사용유무에 따른 전화망환경에서의 인식성능을 평가하였다. 표 1은 39차 MFCC(with C0), state가 3개인 HMM을 사용하여 윈도우 shifting rate가 8ms와 10ms인 경우를 비교하였다.

표 1. 윈도우 shifting rate에 따른 결과

비교조건 / 인식율	8ms shift		10ms shift	
	Monophone	Triphone	Monophone	Triphone
WORD	81.29	95.59	88.30	97.40
SENT	43.93	84.33	46.34	86.02

표 2와 표 3은 39차 MFCC(with C0), state가 3개인 HMM, 윈도우 shifting rate가 10ms와 8ms일 때, filter bank 수에 따른 결과를 나타내었다.

표 2. Filterbank 수에 따른 결과 (shift:10ms)

비교조건 / 인식율	Filter bank 13		Filter bank 16	
	Monophone	Triphone	Monophone	Triphone
WORD	88.30	97.40	88.16	97.29
SENT	46.34	86.02	45.71	85.59

표 3. Filter bank 수에 따른 결과 (shift:8ms)

비교조건 / 인식율	Filter bank 13		Filter bank 16	
	Monophone	Triphone	Monophone	Triphone
WORD	81.29	95.95	81.19	95.21
SENT	43.93	84.33	43.67	83.09

표 4와 표 5는 state가 3개인 HMM, 13개의 filter bank, 윈도우 shifting rate가 10ms와 8ms일 때, 주파수영역에서의 에너지 사용유무에 따른 결과를 나타내었다.

표 4. C0 사용유무에 따른 결과(window shift:10ms)

비교조건 / 인식율	With C0		Without C0	
	Monophone	Triphone	Monophone	Triphone
WORD	88.30	97.40	87.55	97.34
SENT	46.34	86.02	43.08	85.61

표 5. C0 사용유무에 따른 결과(window shift:8ms)

비교조건 / 인식율	With C0		Without C0	
	Monophone	Triphone	Monophone	Triphone
WORD	81.29	95.59	80.51	94.99
SENT	43.93	84.33	42.51	82.35

표 6은 39차 MFCC(with C0), 16개의 filter bank, 그리고 10ms윈도우 shifting rate를 사용했을 때 CMN 사용유무에 따른 결과를 보여주고 있다.

표 6. CMN 사용유무에 따른 결과(window shift:10ms)

비교조건 / 인식율		With CMN		Without CMN	
		Mono PH	Tri PH	Mono PH	Tri PH
MIX 1	WORD	88.16	97.29	88.50	97.31
	SENT	45.71	85.59	46.43	85.35
MIX 3	WORD	92.07	98.54	92.84	98.36
	SENT	61.08	91.86	63.72	90.91
MIX 5	WORD	93.13	98.66	96.62	98.47
	SENT	65.62	92.52	67.53	91.43
MIX 7	WORD	93.06	98.63	93.71	98.51
	SENT	65.46	92.30	67.87	91.67

표 6의 결과를 보면 Gaussian mixture가 1일 경우는 CMN를 사용하지 않는 경우가 성능이 좋지만, mixture수를 높일수록 성능이 좋아지므로 CMN를 사용하는 경우가 성능이 좋아짐을 볼 수 있다.

3.2 HMM state수와 mixture수에 따른 비교 평가

본 실험은 실험 3.1의 결과를 바탕으로 가장 성능이 좋은 조건에서의 특징벡터(20ms Hamming window를 10ms씩 이동, 39차 MFCC(with C0), 13개의 filter bank, CMN를 적용한 경우)에 대해 state 수와 Gaussian mixture 수에 따른 성능평가이다. 결과는 표 7과 표 8과 같다. 표 7과 표 8을 보면 HMM state 수에 따른 성능의 차이가 없다. 평가 시, 음소단위로 인식을 수행했으므로 state 수 3개가 적당한 것 같다.

표 7. Shifting rate 10ms, filter bank 13, and state 3 with C0 and CMN.

		Monophone	Triphone
MIX-1	WORD	88.30	97.40
	SENT	46.43	86.02
MIX-3	WORD	92.50	98.50
	SENT	63.24	91.59
MIX-5	WORD	92.94	98.66
	SENT	64.63	92.51
MIX-7	WORD	92.93	98.67
	SENT	64.69	92.46

표 8. Shifting rate 10ms, filter bank 13, and state 5 with C0 and CMN.

		Monophone	Triphone
MIX-1	WORD	82.45	96.09
	SENT	46.34	86.02
MIX-3	WORD	88.75	97.75
	SENT	63.24	91.59
MIX-5	WORD	89.40	98.00
	SENT	64.63	92.51
MIX-7	WORD	89.39	98.00
	SENT	64.69	92.46

4. 결 론

본 실험을 통해서, 전화망에서의 한국어 연속숫자음 인식성능이 Hamming window의 shifting rate, 주파수 영역에서의 에너지(C0) 사용유무, filter bank의 수, 그리고 CMN의 사용유무에 따라 성능이 다름을 알 수 있었다. 가장 좋은 성능을 보인 경우는 39차 MFCC, 20ms Hamming window에 10ms shift, 13개의 filter bank, 그리고 CMN를 사용한 경우이다. 한국어 연속숫자음 인식의 경우 무잡음환경에서 통상 95% 이상의 문장인식율을 보이는 점에 미루어 볼 때, 전화망 환경의 경우 약 92%정도가 나오므로 성능이 떨어짐을 알 수 있다. 이러한 결과가 나온 이유는 전화망 환경이 한국어 자음 모델링에 중요한 역할을 하는 고주파성분을 왜곡시키기 때문이다. 그리고 평가에서 훈련용 데이터와 테스트용 데이터가 같은 전화망환경에서 녹음되어진 음성샘플을 이용하였으므로 훈련과 테스트의 불일치 문제는 해결했지만, 그렇지 못할 경우는 인식결과가 많이 저하될 것으로 예상되어진다.

참고문헌

1. Ramalingman Hariharan, Imre Kiss, and Olli Viikki, "Noise Robust Speech Parameterization Using Multiresolution Feature Extraction," *IEEE Trans. on Speech & Audio Processing*, vol.9, pp.856-864, November 2001.
2. Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc. 1993.
3. Steve Young, *The HTK BOOK(for HTK Version 3.0)*, 2000.