차량 항법용 음성 인식 시스템 구현

김지성, 이태한, 신원호, 양태영,
김원구*, 이충용, 윤대희, 차일환

연세대학교 전자공학과, *군산대학교 전기공학과

Implementation of Speech Recognition System for Car Navigation

Ji-Sung Kim, Tae-Han Lee, Won-Ho Shin, Tae-Young Yang,
Weon-Goo Kim*, Chungyong Lee, Dae-Hee Youn, Il-Whan Cha

Dept. of Electronics Eng., Yonsei Univ, *Dept. of Electrical Eng., Kunsan National Univ.

E-mail : kjs@caas.yonsei.ac.kr

요 약

본 논문에서는 자동차 잡음 환경에서 녹음된 데이터 베이스를 이용하여 인식 시스템의 성능을 향상시키기 위한 효율적인 잡음 제거 방법을 연구하였다.

먼저, 잡음 및 주변 환경 변화에 강인한 것으로 알려져 있는 특징 벡터들의 인식 성능을 비교하고, 가중 켑스트랄 거리 측정 방법을 이용한 인식 실험을 통하여 시스템의 성능 향상을 확인하였다. 실험 결과, 본 논문에서 기준 시스템으로 사용한 LPC 켑스트럼의 경우에 비하여 MFCC나 root-cepstrum을 사용한 경우 인식률이 향상되었다. 켑스트럼간의 거리 측정에 있어서는 RPS와 BPL과 같은 가중 켑스트랄 거리 측정 함수들이 인식 성능 향상에 도움을 주었다.

또한, 켑스트럼 평균 차감법이라는 간단한 잡음 제거 기술을 적용하여 자동차 잡음 환경에서 인식 성능 향상을 보였다.

마지막으로, 차량 항법용 음성 인식 시스템의 실시간 구현을 위하여 여러 경우의 인식 성능을 비교하고, 메모리 량과 실행 시간 등을 고려하여 최적 시스템을 제시하였다.

1. 서 론

최근에는 음성 인식 기술에 대한 연구가 활발히 진행되고 있으며, 다양한 음성 인식 시스템이 개발되고 있다. 이러한 음성 인식 가능을 갖춘 시스템이 실용화되기 위해서는 잡음에 강인한 방법에 대한 연구가 필요하다. 그 이유는 잡음이 없는 조용한 환경에서 우수한 성능을 나타내는 음성 인식 시스템의 성능이 주위에 잡음이 존재하는 환경에서는 급격히 떨어지기 때문이다[1].

본 논문에서는 차량 항법용 음성 인식 시스템 구현을 위하여, 자동차 잡음 환경에서의 시스템 성능 향상을 위한 효율적인 기술의 개발을 목표로 하였다. 이를 위해 자동차 잡음 환경에서 녹음된 데이터를 가지고 잡음에 강인한 특징 벡터 및 가중 켑스트랄 거리 측정 방법, 켑스트럼 평균 차감법 등을 적용하여 인식 실험을 하였다. 그 결과 켑스트럼 평균 차감법을 적용한 경우가 우수한 인식 성능을 보였다. 또한, 메모리와 실행 시간 등을 고려하여 최적의 실시간 시스템을 제안하였다.

본 논문은 서론에 이어 2장에서 잡음 처리 기술에 대하여 소개하고, 3장에서 이에 따른 실험 및 결과를 고찰하였고, 4장에서는 실시간 시스템에 적용하기 위한 방법을 제시하였으며, 5장에서 결론을 맺었다.

2. 잡음 처리 기술

2.1. 잡음에 강인한 특징 벡터와 거리 측정 방법

잡음에 강인한 것으로 알려진 특징 벡터들 중에서 대표적인 것으로는 멜 켑스트럼 계수(Mel Cepstrum Coefficient: MFCC)[2], 루트 켑스트럼 계수[3] 등이 있다. 멜 켑스트럼 분석 방법은 인간의 청각 특성을 이용한 것으로, 실제 물리적인 주파수와 인지된 주파수 사이의 대응 관계를 이용하여 켑스트럼 계수로 표현한 것이다.

루트 켑스트럼은 로그 연산 대신 루트 연산을 사용하여 켑스트럼 계수를 얻는 방법이다. 일반적으로 사용되고 있는 로가리듬 켑스트럼 분석은 잡음에 매우 민감하기 때문에, 이러한 문제를 극복하기 위하여 루트 켑스트럼 도메인에서의 분석으로 확장하면 잡음에 민감하지 않은 파라메터를 얻을 수 있다[3].

잡음에 강인한 거리 측정 방법의 핵심은 선택적이고 자동적으로 잡음이 적게 들어간 스펙트럼 영역을 강조하는 것이다. 이러한 방법 중의 하나가 가중 켑스트랄

거리 측정 방법이다[4]. 이는 다음과 같이 정의되는데,

$$d(c, c') = \sum_{k=1}^{p} w_k^2 (c_k - c_k')^2$$

여기서 $w \equiv (w_1, \cdots, w_p)$는 켑스트랄 리프터(cepstral lifter)인 가중 함수이고 $c = (c_1, \cdots, c_p)$와 $c' = (c'_1, \cdots, c'_p)$은 켑스트럼 계수이다. 음성 인식에서 좋은 성능을 나타낸 가중 함수들로는, 스펙트럼 기울기에 기초를 둔 RPS(Root Power Sum), 가우시안(gaussian) 형태로 스무딩된 선형 리프터(smoothed linear lifter: SLL), 지수 함수 리프터(general exponential lifter: GEL), 켑스트럼 계수의 높은 차수와 낮은 차수의 바람직하지 못한 변화를 제거하기 위한 밴드 패스 리프터(band pass lifter: BPL), 켑스트럼 계수의 통계적인 분포에 따라 가중 함수를 결정한 것 등이 있다.

2.2. 켑스트랄 평균 차감법 (Cepstral Mean Subtraction: CMS)

켑스트랄 평균 차감법은 전체 구간에 대하여 켑스트럼의 평균을 구하고, 이를 차감하여 채널의 효과를 제거하는 방법이다[5].

음성 신호에 가해진 왜곡의 영향이 선형 필터로 표현된다면, 왜곡은 관찰된 신호를 역 필터링 함으로써 제거될 수 있다. 켑스트럼 영역에서 왜곡의 영향은 관찰된 신호의 켑스트럼에서 왜곡과 관련된 켑스트럼을 빼줌으로써 제거될 수 있다. 즉,

$$c_x = c_y - c_D$$

로 표현될 수 있는데, 여기서 c_x, c_y, 그리고 c_D는 각각 원 신호, 관찰된 신호, 채널 왜곡 성분에 해당하는 켑스트럼을 의미한다. 채널 왜곡 특성이 음성 신호의 관찰 구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 켑스트럼의 추정치는 다음 수식으로 표현될 수 있다.

$$m_y = \frac{1}{N(s)} \sum_{t=1}^{N(s)} c_y^t$$

$$c_{comp}^t = c_y^t - m_y$$

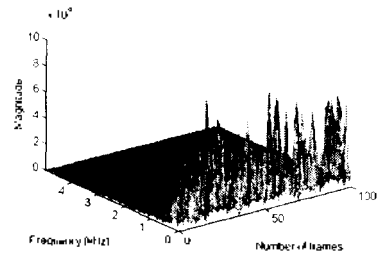
여기서, m_y는 음성의 모든 프레임에서 켑스트럼의 평균이고, $N(s)$는 입력 음성의 전체 프레임 수이며, c_{comp}^t는 t번째 프레임에서 CMS를 통해 보상된 켑스트럼을 의미한다.

3. 실험 및 결과

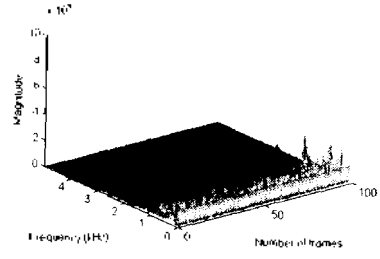
3.1. Database 및 잡음 분석

실험에 이용한 데이터는 자동차 환경에서 녹음된 69단어로 구성되어 있다. 기준 데이터는 50명(남 25명, 여 25명)이 각 단어를 정지 중 시동 끄고 3번, 켜고 3번 발음한 것으로, 시험 데이터는 19명(남 11명, 여 8명)이 각 단어를 정지 중 시동 켜고 1번, 100 Km/h로 주행 중 2번 발음한 것을 사용하였다. 이때, 헤드 셋(head-set) 마이크를 사용하였고, DAT(Digital Audio Tape)를 사용하여 녹음하였다.

인식 실험을 수행하기 전에 자동차 잡음의 특징을 살펴보면 그림 1과 같다. 그림 1(a)는 100 Km/h로 주행 중인 자동차에서 1초 동안 녹음한 자동차 잡음의 스펙트로그램(spectrogram)으로 대부분의 에너지가 저주파 영역에 집중된 것을 알 수 있다. 본 논문에서는 이러한 잡음을 제거하기 위하여 230Hz의 고역 통과 필터를 사용하였다. 그림 1(b)는 그림 1(a) 신호를 고역 통과 필터를 통과시켰을 때의 신호의 스펙트로그램이다. 이렇게 하여 자동차 잡음이 첨가된 음성 신호에 거의 영향을 끼치지 않으면서 상당 부분의 잡음을 제거할 수 있다.



(a)



(b)

그림 1. 100 km/h의 속도로 주행중인 자동차 잡음
(a) 잡음 신호의 스펙트로그램
(b) 고역통과 필터링된 잡음 신호의 스펙트로그램

3.2. 잡음 환경에서의 음성 구간 검출

음성 구간 검출을 사용하는 인식 시스템은 구해진 음성 구간에 대해 인식을 수행하므로 인식률에 미치는 영향이 크기 때문에 정확한 음성 구간 검출이 요구된다.

따라서, 본 논문에서는 자동차 잡음 환경에서 정확한

음성 구간을 검출하기 위하여 다음과 같은 음성 구간 검출 알고리즘을 제안하였다. 음성 구간 검출 알고리즘을 크게 두 부분으로 나누어, 한 부분은 자동차 잡음을 전극 필터로 모델링한 AR(AutoRegressive) 계수를 이용하여 입력 신호를 역 필터링(inverse filtering)한 후, 신호의 크기를 이용하여 음성 구간 검출을 수행하고, 다른 부분은 자동차 잡음의 대부분이 저주파 영역에 집중되어 있으므로 230Hz의 차단 주파수를 갖는 고역 통과 필터를 통과시켜서 잡음을 제거한 신호의 크기를 이용하여 음성 구간 검출을 수행하고 역 필터링을 이용하여 구한 음성 구간을 보정한다. 이 방법에 대한 블럭도가 그림 2에 주어져있다.

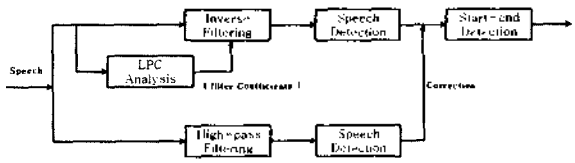


그림 2. 음성 구간 검출 방법

3.3. 특징 추출 및 인식 시스템 구성

반연속 HMM을 이용한 화자 독립 단독음 인식 시스템의 구성은 다음과 같다. 음성 신호는 10kHz, 16비트로 표본화되고, 이 음성 신호는 $1-0.95z^{-1}$의 전달 함수를 갖는 프리엠퍼시스(pre-emphasis) 필터를 통과하고 20ms(200sample)의 크기를 갖는 해밍 창을 사용하여 10ms씩 이동하면서 분석한다. 각 분석 구간으로부터 얻은 LPC 켑스트럼 또는 멜 켑스트럼 분석을 이용하여 12차의 켑스트럼 계수를 구하였다. 또한 현재 프레임을 기준으로 전후 20ms의 켑스트럼의 차를 이용하여 차등 켑스트럼을 구하였다. 에너지 파라메터는 전후 10ms의 에너지 차와 에너지 차의 전후 20ms의 차를 구하여 2차의 에너지 벡터를 구하였다. 이러한 특징 벡터는 256개의 코드를 갖는 3개의 코드북을 구하는데 사용되었고, 각 단어의 모델은 10개의 상태 개수를 가진다.

3.4. 처리 방법에 따른 실험

인식 시스템의 학습 및 인식은 LPC 켑스트럼 방법을 기준으로, 잡음에 강한 특징 벡터인 멜 켑스트럼(MFCC)과 루트 멜 켑스트럼을 비교하여 보았다. 이때, 1024샘플로 FFT를 수행하고, 켑스트럼 차수는 12차로 실험을 하였다.

표 1의 결과를 살펴보면, 두 가지 경우 모두 기준 시스템보다 인식률 향상을 보이고 있다. 특히, MFCC의 경우보다 Root MFCC인 경우 더 높은 인식률 향상을 보이고 있는데, 이는 루트 켑스트럼 도메인에서의 분석이 잡음에 덜 민감하다는 것을 나타낸다.

위의 실험에 이어서 각 특징 벡터에 대하여 가중 함수를 가하여 이에 따른 인식 성능의 변화를 표 2에 나타내었다. 이때, 가중 함수를 사용한 경우 인식률이 향상되며, 가중 함수끼리는 서로 비슷한 성능을 보였다.

다음은 멜 켑스트럼 및 루트 멜 켑스트럼에 대하여 CMS를 이용하여 인식 실험을 하였다. 이 경우 각 데이터의 전체 구간에 대하여 켑스트럼의 평균을 구한 후, 이것을 차감하여 구한 데이터들로부터 훈련 및 인식 실험을 하였다. 표 3을 보면, 전체적으로 인식률 향상을 보이고 있다.

또한, 코드워드 크기가 인식률에 미치는 영향을 알아보기 위해서 128개의 코드를 사용한 경우에 대해 실험을 하여 결과를 표 4에 나타내었다. 이를 앞의 256개의 코드를 사용한 경우와 비교해 보면 대체적으로 1% 정도의 인식률이 저하되었다. 그 이유는 벡터 양자화의 오차가 256개 코드인 경우보다 128개 코드인 경우가 더 크기 때문이다.

지금까지의 실험에서는 HMM의 상태 개수를 10개로 고정하여 사용하였는데, 이는 데이터 베이스 중의 "네비게이션", "페이지다운" 같은 긴 단어를 표현하는데는 적합하지 않다. 따라서, 이들을 표현하는데 보다 적합하도록 각 음절당 5개의 상태를 사용하여 실험을 하였다. 이의 결과는 표 5에 나타내었다. 이를 보면 상태 개수를 10개로 고정한 경우에 비해 약 2~3%의 인식률 향상을 보였다.

실험에 사용한 멜 켑스트럼이나 루트 멜 켑스트럼은 FFT에 근거한 특징 벡터이다. 따라서, FFT 크기가 인식률에 미치는 영향을 살펴보기 위하여 512샘플로 FFT를 수행하여 실험을 하였다. 이때, 가중 함수는 RPS를 사용하였다. 이에 대한 결과는 표 6에 나타내었다. 결과를 보면 512샘플인 경우 인식률이 약 1~5% 저하되었다. 이는 FFT의 크기가 줄어들면 주파수의 해상도(resolution)가 떨어지기 때문이다.

이제까지의 실험 결과를 바탕으로 하여 멜 켑스트럼을 사용하고, 평균 차감법을 적용하며, 가중 함수 RPS, 128개의 코드워드, 1음절 당 5개의 상태 개수, 1024 샘플 FFT를 사용한 경우를 차량 항법용 음성 인식 시스템으로 제시한다. 이 경우의 인식률은 94.64%이며, 두 번째 후보까지의 인식률이 97.38%이므로 인식 오류가 발생하였을 때 다음 후보를 사용하는 인식 시스템을 구성하면 더 효과적일 것이다.

4. 실시간 시스템 구현

본 논문에서는 실험 결과를 바탕으로 차량 항법용 음성 인식 시스템을 TI사의 TMS320C31을 이용하여 실시간으로 구현하였다. 그런데 실시간 시스템으로 구현할 때는 메모리 사용량과 계산 시간을 고려하여야 하므로 3장에서 제시된 구성은 적합하지 않다. 따라서 실시간 시스템은 멜 켑스트럼을 사용하고, 평균 차감법을 적용하며, 가중 함수 RPS, 128개의 코드워드, 10개의 상태,

표 1. 특징 벡터에 따른 인식률 (%)

feature vector	recognition rate(%)
reference (LPCC)	85.10
MFCC	87.41
Root_MFCC	90.49

표 2. 특징 벡터와 가중 함수에 따른 인식률 (%)

weight function	EUC	RPS	BPL
reference (LPCC)	85.10	91.18	89.63
MFCC	87.41	91.46	90.87
Root_MFCC	90.49	91.36	91.56

표 3. CMS방법에 따른 인식률 (%)

weight function	EUC	RPS	BPL
MFCC_CMS	91.79	92.65	92.83
Root_MFCC_CMS	91.56	92.60	92.68

표 4. 코드워드 크기에 따른 인식률(%)

codeword size	256			128		
weight function	EUC	RPS	BPL	EUC	RPS	BPL
LPCC	85.10	91.18	89.93	84.64	90.92	90.31
MFCC	87.41	91.46	90.87	84.01	88.89	87.59
Root_MFCC	90.49	91.36	91.56	90.16	91.25	90.77
MFCC_CMS	91.79	92.65	92.83	91.28	92.52	92.70
Root_MFCC_CMS	91.56	92.60	92.68	91.46	92.50	92.47

표 5. 음절 당 5개의 상태를 사용한 경우의 인식률(%)

weight function	EUC		RPS		BPL	
codeword size	256	128	256	128	256	128
Root_MFCC	93.69	93.85	94.10	94.43	94.58	94.20
MFCC_CMS	93.54	93.11	94.25	94.64	94.25	94.28
Root_MFCC_CMS	93.90	93.39	94.36	94.28	94.46	94.02

표 6. FFT 크기에 따른 인식률(%)

FFT size	1024		512	
codeword size	256	128	256	128
MFCC	91.46	88.89	85.81	85.74
Root_MFCC	91.36	91.25	90.03	90.29
MFCC_CMS	92.65	92.52	91.51	91.79

표 7. 인식 함수의 소요 클럭수

함수	소요 클럭수 (40MHz)	시간(s)	점유율(%)
끝점 검출 함수	8,000	0.0004	0.030
벡터 양자화	4,196,000	0.2098	15.884
확률 밀도 함수	1,084,000	0.0542	4.104
혼합 확률 밀도함수와 로그 비터비 함수	21,127,800	1.0564	79.982
합계	26,415,800	1.3208	100

512 샘플 FFT를 사용한 최적의 시스템으로 구현하였다. 여기서 상태 개수를 10개로 고정한 이유는 메모리 사용량의 문제와 프로그램화가 편리하기 때문이다.

위와 같이 구현된 실시간 시스템의 메모리 사용량은 약 290k word이며 인식 시간은 1.3208초가 소요되었다. 표 7에는 인식 함수의 소요 클럭수를 나타내었다.

5. 결 론

본 논문에서는 자동차 잡음 환경에서 음성 인식 시스템의 성능 향상을 위한 효율적인 특징 추출 및 전처리 기술을 비교, 연구하였다. 자동차 잡음 환경에서 녹음한 데이터 베이스를 이용하여, 반연속 HMM을 기반으로 한 인식 시스템을 구현하여 화자 독립 단독음 인식을 수행하였다.

여러 가지 잡음 처리 기술을 이용하여 실험한 결과 다음과 같은 결론을 얻을 수 있었다. 첫째, 멜 켑스트럼이나 루트 멜 켑스트럼을 특징 벡터로 사용한 결과 기존의 LPC 켑스트럼을 사용한 경우보다 우수한 인식 성능을 나타냈다. 둘째, 특징 벡터를 주변 잡음에 강인하게 하는 가중 함수(RPS, BPL)를 사용한 경우 인식 성능이 향상되었다. 셋째, 잡음 처리 기술인 CMS를 이용하여 인식 성능을 향상시켰다. 넷째, 멜 켑스트럼이나 루트 멜 켑스트럼을 특징 벡터로 사용하고 가중 함수를 사용하며 CMS를 적용한 경우에 가장 우수한 성능을 얻었다. 마지막으로, 인식 실험을 바탕으로 차량 항법용 음성 인식 시스템을 TI사의 TMS320C31을 이용하여 실시간으로 구현하였다.

참고 문헌

[1] H. W. Ruehl, S. Dobler, J. Weith and P. Meyer, "Speech Recognition in the Noisy Car Environment," Speech Commun., Vol. 10, No. 1, pp. 11-22, Feb. 1991.

[2] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, 1993.

[3] Patrice Alexandre and Philip Lockwood, "Root Cepstral analysis: a Unified View. Application to Speech Processing in Car Noise Environments," Speech Communication, Vol. 12, No. 3, pp. 277-288, Jul. 1993.

[4] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-35, No. 10, pp. 1414-1422, Oct. 1987.

[5] Richard J. Mammone, Xiaoyu Zhang, Ravi P. Ramachandran, "Robust Speaker Recognition - A Feature-based Approach", in IEEE Signal Processing Mag., pp. 58-87, Sep. 1996