1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

다중 대역기반 우도 측정을 이용한 잡음 환경에서의 음성 인식

신원호, 양태영, 이충용, 윤대희, 차일환

연세대학교 전자공학과

Speech Recognition in the Noisy Environment Using Multi-Band-Based Likelihood Measure

Won-Ho Shin, Tae-Young Yang, Chungyong Lee, Dae-Hee Youn, Il-Whan Cha

Dept. of Electronic Eng. Yonsei Univ.

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

요약

본 논문에서는 서브밴드 및 전 대역(full band)으로부터 얻은 특징 벡터를 함께 사용하여 잡음 환경에서 음성 인식 시스템의 성능을 향상시키는 방법을 제안하였다. 이는 인식시 잡음에 오염된 대역에서 얻은 특징 벡터를 제거하는데 따른 정보 손실을 막기위해 전 대역으로부터 얻은 특징 벡터를 함께 이용하며 신호 대 잡음비가 높은 대역을 강조하여 각 모델에 대한 확률 값을 계산한다. 전화망에서 수집된 데이터베이스를 이용하여 인식 실험을 수행한 결과 비교적 넓은 주파수 대역에 걸쳐 분포된 잡음의 경우에도 인식 성능을 향상시킬 수 있었다.

I. 서론

일반적으로 음성 인식에 사용되는 특징 벡터를 얻기 위해서는 전체 대역 스펙트럼을 사용하는데, 주파수 대역의 일부분만 잡음에 의해 영향을 받아도 특징 벡터 전체가 영향을 받게 된다. 따라서 최근에는 서브밴드 특징 벡터를 독립적으로 얻은 후 인식 단에서 결합하는 연구가 많이 이루어지고 있다. 서브밴드 음성 인식의 아이디어는 Bourlard와 Hermansky로부터 출발하였다[1][2]. Lippmann과 Carlson[3]은 멜 필터 뱅크의 스펙트럼을 특징 벡터로 이용하여 잡음에 손상된 필터 뱅크를 인식에서 제외시키는 방법을 제안하였고 Bourlard와 Dupont[4]는 다중 시간 비율(time scale)을 갖는 인식 방법을 제안하였는데, 이는 명료 HMM에 대하여 다른 시간 비율로 얻어진 음성 신호의 개별적인 동적 특성을 강조하도록 하였다. Okawa[5]는 서브밴드로부터 얻은 특징 벡터를 결합하여 한 개의 벡터로 이용하는 방법을 제안하였다.

그러나 이와 같은 서브밴드 구조를 갖는 음성 인식 시스템은 주로 부분적인 주파수 대역을 가린 잡음 환경에서 사용하기에 적합한 특성을 가지고 있어 따라서 실제로 주변에 존재하는 자동차 잡음이나 기타 여러 가지 배경 잡음들은 보다 넓은 주파수 대역을 차지하므로 서브밴드 구조의 결점을 보완해야 한다.

본 논문에서는 잡음에 오염된 대역을 제거하는 기존의 방법 대신 전체 대역으로부터 얻은 특징 벡터를 서브밴드 특징 벡터와 함께 이용하는 방법을 제안하였다. 이는 인식시 잡음에 의해 오염된 대역의 특징 벡터를 제거하는 데 따른 정보 손실을 막으며 상대적으로 SNR이 높은 대역을 확률 계산시 강조하게 된다. 2장에서는 다중 대역기반 우도 측정에 대하여 설명하였고 3장에서는 이를 이용한 실험 결과를 제시하였고 4장에서 결론을 맺었다.

II. 다중 대역기반 우도 측정

2.1 다중 대역기반 음성 인식

다중 대역기반 음성 인식은 기본적으로 주파수 대역을 분할하여 각 대역별로 특징 벡터를 추출하여 인식하는 방법이다. 이는 서브밴드 음성 인식이라고도 한다.

이와 같이 다중 대역을 이용한 연구가 주목을 받게 된 이유는 다음과 같다. Fletcher와 이의 연구를 바탕으로 한 Allen[11]에 의하면 언어적 정보는 각각 다른 주파수 영역에서 독립적으로 복호화 되며 최종 결정은 각 서브밴드의 결정을 종합함으로써 이루어진다고 제안했다. 또한 어느 서브밴드의 조잡어진적 이것으로부터 충분한 정보를 얻을 수 있다면 다른 서브밴드를 정보의 복호화에 이용하지 않아도 된다는 것이다. 즉 주변 잡음이 유색이거나 몇 개의 대역만 심하게 오염되었을 경우, 오염된 대역이 인식 결정에 영향을 미치지 않도록 서브밴드의 확률 값 및 결합 방법을 구상할 수 있다. 이 밖에도 안정된 구간(segment) 사이의 전환이 대역별로 동시에 일어나지 않을 수 있으므로 각 대역간의 동기화에 대한 제한을 완화할 수 있다. 그리고 각 대역별로 다른 인식 방법을 적용할 수 있는 이점을 가지고 있다.

입력 음성에 대한 확률 값을 구하기 위해서는 각 서브밴드 특징 벡터에 대한 우도 계산한 방법이 필요한데, 각 서브밴드는 대역마다 다른 정도의 정보량을 갖을 수 있다. 또한 잡음이 심한 대역이 영향을 줄일 필요가 있으므로 각 대역 별로 가중치를 고려하여 계산하여야 한다. 상태 s_j에 대한 t번째 프레임의 우도를 계산하면 다음과 같다.

$$f(o_t \mid s_j) = \prod_b f(o_t^b \mid s_j)^{w_b}, \qquad (1)$$

여기서 각 서브밴드에 대한 가중치 w_b를 결정하는 방법은 각 대역별로 정규화된 음소 단위 인식률을 가중치로 이용하거나 각 대역별로 SNR을 추정하여 대역별로 상대적인 신뢰도를 측정한다. 특별히 대역별로 다른 시간 비율의 특징 벡터를 사용하는 경우, 전체 확률 값 및 인식 결정을 위해서는 결합 기준(merging level)을 정할 필요가 있다. 이는 프레임보다 상위의 단위인 음소나 음절, 단어 등의 단위로 결합할 수 있는데 어느 단

위로 동기화 하여 결합하는 것이 가장 좋은 것인지에 대해서는 현재까지 아직 뚜렷한 연구 결과가 없으므로 이의 지속적인 연구가 필요하다[1][6].

2.2. 제안된 다중 대역기반 우도 측정

잡음이 비교적 광대역에 걸쳐 분포하는 경우 잡음에 오염된 대역을 모두 제거하기가 힘들게 된다. 실제로 어느 대역이 잡음에 오염되어 있다고 하더라도 어느 정도의 정보를 가지고 있기 때문이다. 따라서 전 대역으로부터 얻은 특징 벡터를 이용하여 각 대역 별로 신호대 잡음비가 좋은 대역을 강조하면 잡음에 강인한 성능을 얻을 수 있다. 이를 수식으로 나타내면 다음과 같다.

$$f(o_t|s_j) = f(o_t^f|s_j)^{w_f} \cdot \prod_b f(o_t^b|s_j)^{w_b}, \quad (2)$$

여기에서 o^f는 전 대역으로부터 얻은 특징 벡터를 나타내며, w_f는 전 대역 특징 벡터에 대한 가중치를 의미한다. 그림 1은 이를 이용한 우도 측정 방법을 간단히 나타낸 것이다.

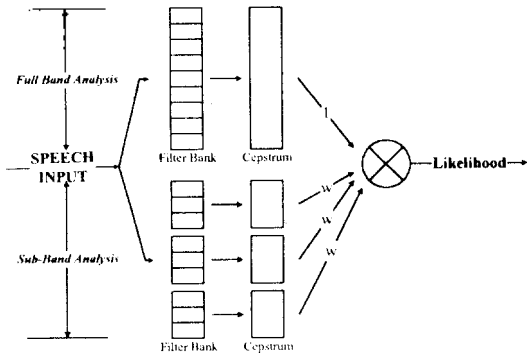


그림 1. 제안된 다중 대역기반 우도 측정

제안된 방법은 다음과 같은 이점을 가지고 있다. 첫째 잡음에 오염된 대역의 특징 벡터는 상대적으로 전 대역으로부터 얻은 특징 벡터보다 심하게 손상되어 있다. 따라서 잡음에 오염된 대역의 특징 벡터를 사용하지 않거나 전 대역 특징 벡터를 이용하여 잡음의 영향을 줄일 수 있다. 둘째 서브밴드 특징 벡터를 이용하여 SNR이 상대적으로 높은 부분을 강조할 수 있다. 세째 전 대역 특징 벡터는 서브밴드 특징 벡터 사이의 상관(correlation) 정보를 포함하고 있다.

2.3 다중 대역기반 인식 시스템의 구현

서브밴드 구조를 갖는 인식 시스템에서 문제가 되는 것 중의 하나는 실험에 이용할 서브밴드의 구조 및 수이다. 현재까지는 주로 2개에서 7개 정도의 서브밴드가 많이 이용되고 있다. 본 논문에서는 1kHz이상의 대역을 갖는 주변 잡음을 이용하므로 2개 및 4개의 서브밴드 구조를 갖는 인식 시스템을 구성하기로 하였다. 8kHz의 샘플링 주파수를 갖는 경우, 다음과 같은 범위를 갖게 되는데 [6] 각 대역별로 약간의 중첩을 허용하도록 한다.

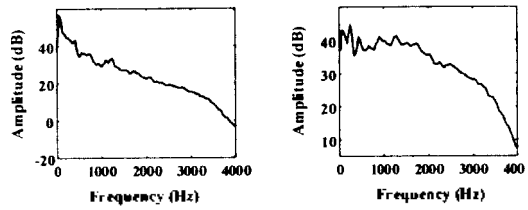
2 대역: 0 ~ 1140Hz, 1046 ~ 4000Hz

4 대역: 0 ~ 765Hz, 700 ~ 1640Hz, 1515 ~ 2700Hz, 2100 ~ 4000Hz

서브밴드의 인식 결과에 이하여 서브밴드 이 대역이 좁아질수록 그 안에 포함된 정보가 줄어들므로 인식 성능이 다소 저하되는 것으로 알려져 있다[6]. 그러나 전 대역을 이용한 인식 시스템의 성능과 유사하거나 다소 향상된 성능을 보이고 있다[6]. 다음으로 고려해야 할 것은 각 대역에 이용할 특징 벡터이다. 필터 뱅크를 이용하거나 FFT를 이용하여 각 대역의 특징 벡터를 구할 수 있다. 본 논문에서는 전 대역으로부터 얻은 특징 벡터를 함께 이용하므로 FFT를 이용하여 전 대역 및 서브밴드의 멜 켑스트럼을 생성한다. 이때 상위 서브밴드 멜 켑스트럼의 경우 각 임계 대역 스펙트럼을 기저 대역(base band)으로 낮추어 계산한다.

III. 실험 및 결과

3.1 데이터베이스 및 인식 시스템 구성



(a)자동차 잡음 (b)도로변 잡음

그림 2. 잡음 신호의 평균 스펙트럼

음성 인식 실험에 사용된 데이터 베이스는 전화망을 통하여 50명의 남녀 화자로부터 얻은 50개의 단어로 구성되어 있다. 학습에 이용된 단어는 3회씩 발음되었고 테스트용 단어는 다른 14명의 화자에 의해 3회씩 발음된 것을 이용하였다. 잡음에 오염된 단어를 얻기 위하여 자동차, 도로변 잡음 그리고 백색 잡음을 실어 20,10,5dB의 SNR에 대하여 실험을 수행하였다. 그림 2는 앞의 두 가지 잡음 신호의 평균적인 스펙트럼 분포를 나타내고 있다. 그림에서 보는 바와 같이 도로변 잡음의 경우 자동차 잡음보다 넓은 주파수 대역을 가지고 있다.

각 단어 모델은 3개의 상태를 가진 분백 종속 음소 모델(context dependent phoneme)들을 결합하여 구성되었다. 이들은 대각 행렬의 공분산(diagonal covariance matrix)과 여러 개의 혼합 밀도 계수(multiple mixture)를 가진 좌우(left-to-right) 형태 반연속(semi-continuous) HMM에 의하여 모델링되었다. 또한 각 단어 모델의 전후에는 묵음이 존재한다고 간주하여 2개의 상태를 가진 묵음 모델을 구성하였다.

실험에 이용된 음성 신호는 8kHz로 샘플링되어 매 10ms마다 20ms의 구간을 가진 해밍 윈도우에 의해 분석되었다. 12차의 전 대역으로부터 얻은 멜 켑스트럼(MFCC: Mel Frequency Cepstral Coefficient)과 6차의 서브밴드 멜 켑스트럼을 구하였다. 본 실험에서는 자동 켑스트럼을 추가로 이용하였으며 전 대역 및 서브밴드 특징 벡터에 대하여 전후 2프레임에 서어 값을 구하였다. 실험에 사용된 데이터는 전화망에서 수집되었으므로 채널의 특성을 보상하기 위해 켑스트럼 평균 차감법(CMS)을 적용하여 채널 바이어스를 제거하였다. 또한 각 켑스트럼의 차이 계수에는 음성에 강인한 것으로 알려진 RPS 가중치[12]를 이용하였다. 학습과 인식 과정은 Baum-

Welch 알고리즘 및 Viterbi 디코딩 방법을 이용하였다. 각 대역별 가중치는 1로 고정하여 학습하였고 인식시에 잡음에 따라 가중치의 값을 조정하였다.

3.2 인식 결과

먼저 각 대역에 대한 잡음의 변화 정도를 살펴 보기 위하여 다음과 같은 평균 거리 측정 값 D를 정의하였다.

$$D = E\left\{\frac{\sum_{d=1}^{d=P}\{w(d)(c_c(d)-c_n(d)\}^2}{\sum_{d=1}^{d=P}w(d)^2}\right\} \quad (3)$$

여기서 $w(d)$는 실험에 이용된 가중 함수이며 $c_c(d)$ 및 $c_c(d)$는 잡음이 포함되지 않은 캡스트럼 및 잡음이 섞인 캡스트럼이다. P 는 캡스트럼의 차수이다. 다음의 표 1은 잡음이 포함되지 않은 신호와 실험에 쓰인 자동차 잡음을 혼합한 10dB 신호와의 평균 거리를 나타낸 것이다. 하위 대역의 경우 상위 대역에 비하여 잡음에 대한 오염도가 심하며 또한 전 대역에 비하여 손상 정도가 크다. 따라서 하위 대역을 사용하는 것보다 전 대역 특징 벡터를 이용하는 것이 유리함을 알 수 있다.

표. 1. 각 대역별 잡음이 포함된 신호와의 캡스트럼 평균 거리

	L	H	F
D	0.01642	0.00540	0.00530

표 2는 전 대역과 2개의 서브밴드 특징 벡터를 이용한 인식 결과이다. 다중 특징 벡터의 효과를 비교하기 위해 자동 캡스트럼이나 에너지등은 사용하지 않았다. 표에서 'F'는 기존 전 대역 특징 벡터만을 이용한 경우이며, 'LH'는 2개의 서브밴드를 이용한 결과이다. 전 대역 특징 벡터를 함께 이용한 경우는 'FLH'로 나타내었고, 괄호안은 우도 측정시 각 대역간의 가중치이다. 인식 실험에 사용된 두 잡음이 주로 저주파 대역에 분포되어 있으므로 하위 대역을 제거하였으나 많은 정보의 손실로 인하여 '$LH(0:1)$'은 전 대역 특징 벡터만을 이용한 기존의 방법인 'F'에 미치지 못하였다. 두 가지 잡음에 대하여 전 대역과 상위 대역을 이용한 'FLH (1:0:1)'가 가장 좋은 성능을 나타내었다. 이는 'F'및 '$LH(0:1)$'에 비하여 각각 12.6%, 19.5%정도의 인식률을 향상시켰다. 전 대역 특징 벡터를 사용하지 않고 여와 유사한 가중 효과를 갖기 위하여 저지 및 두 특징 벡터만을 이용한 '$LH(0.5:1)$'은 '$LH(1:1)$'에 비하여 다소간의 향상을 보였으나 '$FLH(1:0:1)$'의 성능에는 미치지 못하였다. 따라서 제안된 전 대역을 이용한 방법이 유용함을 볼 수 있다. 그러나 백색 잡음의 경우에는 잡음이 전 대역에 걸쳐 골고루 분포되어 있으므로 잡음이 증가함에 따라 전대역으로 SNR이 유리한 'FLH (1:1:0)'이 성능에 미치지 못하였다.

다음으로 기존의 전 대역 특징 벡터와 서브밴드 특징 벡터의 비교에 앞서지는 서브밴드 특징 벡터를 이용한 'LH (1:1)'과 전 대역 특징 벡터를 이용한 'F'에 미치지 못하지만 다중 대역을 기반으로 한 인식 방법이 유용함을 보여주고 있다.

표 2. 전 대역 및 2개의 서브밴드를 이용한 인식 결과

		F	LH (0:1)	LH (1:1)	LH (0.5:1)	FLH (1:1:0)	FLH (1:0:1)
Clean		86.57	72.10	88.52	88.43	85.57	89.52
자동차 (dB)	20	85.24	70.33	85.95	86.43	82.95	89.14
	10	69.43	63.10	74.67	76.90	63.52	78.24
	5	52.43	52.00	60.62	63.57	48.14	64.57
도로변 (dB)	20	81.24	58.05	84.05	82.67	80.00	84.19
	10	55.00	26.90	55.14	52.00	54.48	56.33
	5	29.38	12.67	27.38	25.14	31.14	30.00
백색 (dB)	20	83.71	60.05	84.71	83.14	84.00	85.71
	10	62.19	26.10	61.71	56.10	64.00	60.76
	5	36.48	10.10	31.62	25.95	41.48	33.90

다음 표 3은 전 대역과 4개의 서브밴드 특징 벡터를 이용한 인식 결과이다. 우도 가중치로 나타낸 수는 전 대역과 하위 대역에서 상위 대역으로 가중치를 나타낸 것이다. 자동차 잡음의 경우에는 '10111', 도로변 잡음의 경우에는 '10110', 낮은 SNR의 백색 잡음에서는 '11100'이 향상된 성능을 나타내었다.

표 3. 전 대역 및 4 개의 서브밴드를 이용한 인식 결과

		F	LH_a	LH_b	LH_c	LH_d	
		10000	01111	10011	10111	11100	10110
Clean		86.95	85.43	87.90	89.76	84.52	88.76
자동차 (dB)	20	84.19	83.33	84.57	87.71	82.86	86.38
	10	68.14	72.62	75.95	79.67	65.19	77.19
	5	54.38	59.10	61.90	67.76	49.76	63.95
도로변 (dB)	20	80.76	78.48	79.76	83.38	78.86	84.29
	10	54.38	48.76	52.62	57.48	53.38	57.86
	5	29.48	26.38	25.19	30.43	29.52	31.10
백색 (dB)	20	84.24	81.33	81.10	85.19	84.05	85.62
	10	62.38	56.57	51.95	59.86	64.43	64.43
	5	38.57	30.81	22.05	32.90	40.62	38.38

표 2의 2개의 서브밴드를 이용한 인식 결과에 비하여 가장 높은 인식율을 기준으로 약 0.06% 가량 인식률이 향상되었다. 그러나 증가된 특징 벡터 수를 고려하면 인식 성능이 개선되었다고 보기 어렵다. 본 실험에 사용된 잡음의 특성을 고려하면 2개 정도의 서브밴드가 적합한 것으로 보인다. 다음으로 고려할 사항은 우도 측정시 사용한 가중치의 결정이다.

표 4는 전 대역 특징 벡터의 가중치를 1로 고정한 후 상위 대역의 가중치를 변화시켜 가며 인식한 결과이다. 표에 나타낸 'X'의 숫자는 상위 대역에 적용된 가중치 값이다. 두 가지 잡음의 경우에 있어서는 잡음의 양이 증가함에 따라 상위 대역의 가중치를 증가 시킴에 따라 인식 성능이 향상되었다.

표. 5는 다중 대역의 차등 캡스트럼을 함께 이용한 인식 결과이다. 표에 나타낸 숫자는 각 대역별 가중치의 값이다. 차등 캡스트럼을 함께 이용하였으므로 전반적인 인식 성능이 향상 되었고 자동차 잡음과 같이 비교적

하위 대역에 잡음이 많이 집중된 경우에는 차등 켑스트럼을 전 대역 및 상위 대역에 가중치를 둔 '101/101'의 경우에 인식률의 향상이 있었다. 5dB의 자동차 잡음 환경에서는 상위 대역을 강조하지않은 '101/100'의 경우에 비하여 3.5% 정도 인식률이 향상되었다. 반면 도로변 잡음과 같이 잡음이 보다 넓게 분포하거나 백색 잡음과 같이 전 대역에 분포한 경우에는 서브밴드 차등 켑스트럼이 큰 효과를 나타내지 못하였다. 따라서 잡음의 특성에 따라서 서브밴드를 이용한 차등 켑스트럼이 유용하게 쓰일 수 있으나 특징 벡터의 추가에 따라 계산량 등이 함께 고려되어야 할 것이다.

표 4 상위 대역의 가중치 변화에 따른 인식률 비교

X		F.L:H=1.0:X				
		0.5	0.75	1	1.25	1.5
Clean		89.81	89.38	89.52	88.86	88.62
자동차 (dB)	20	87.81	88.14	89.14	87.86	87.38
	10	77.33	78.19	78.24	78.43	78.00
	5	62.29	63.52	64.57	65.38	65.48
도로변 (dB)	20	84.52	84.62	84.19	82.62	82.05
	10	58.10	57.38	56.33	57.62	57.00
	5	31.48	30.62	30.00	32.43	31.67
백색 (dB)	20	86.52	86.19	85.71	84.90	84.24
	10	63.38	62.38	60.76	63.10	61.95
	5	36.71	36.00	33.90	36.48	35.48

표 5 다중 대역의 차등 켑스트럼을 이용한 인식 결과

X		F.L:H(cep)/F.L:H(dcp)				
		101/101	101/100	110/110	110/100	100/100
Clean		97.71	97.81	96.81	96.33	97.10
자동차 (dB)	20	96.19	96.24	95.10	94.76	95.71
	10	88.38	87.52	83.29	82.14	84.81
	5	77.10	73.62	62.38	62.86	66.86
도로변 (dB)	20	94.67	94.38	94.29	94.24	94.71
	10	71.29	70.43	72.52	72.10	72.14
	5	32.52	32.76	36.10	37.81	33.52
백색 (dB)	20	95.71	95.62	95.38	95.05	95.67
	10	75.05	75.57	77.62	78.14	76.90
	5	37.19	38.19	47.29	47.90	40.10

결론

본 논문에서는 주변에 존재하는 비교적 넓은 주파수 대역을 갖는 잡음 환경에서 서브밴드 특징 벡터를 이용한 인식 시스템의 성능 향상을 위해 전 대역 특징 벡터를 함께 이용하였다. 이러한 가중 우도 측정 방법은 상위 대역을 강조함에따라 조용한 환경에서도 향상된 인식 성능을 나타내었고 가중치의 조성에 따라 잡음 환경에서도 개선된 인식 성능을 나타내었다. 다중 대역을 이용한 차등 켑스트럼의 경우에도 잡음이 하위 대역에 많이 집중된 자동차 잡음의 경우 인식 성능의 향상되었다.

참고 문헌

[1] H. Bourlard and S. Dupont, "A New ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP*, pp. 426-429, Oct. 1996.

[2] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on partially corrupted speech," in *ICSLP*, pp. 1579-1582, Oct. 1996.

[3] R. P. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise," in *Eurospeech*, pp. KN37-KN40, Sep. 1997.

[4] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proc. ICASSP*, pp. 1251-1254, Apr. 1997.

[5] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. ICASSP*, vol. 2, pp. 641-645, May 1998.

[6] S. Tibrewala and H. Hermansky, "subband based recognition of noisy speech," in *Proc ICASSP*, pp. 1225-1258, Apr. 1997.

[7] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Eurospeech*, pp 2619-2622, Sep. 1997.

[8] Y. Normandin, R. Cardin, and R. DeMori, "High-performance connected digit recognition using maximun mutual information estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 299-311, Apr. 1994.

[9] X. D. Huang, Y. Ariki, and M. A. Jack, *edits: Hidden Markov Models for Speech Recognition*, Edinburgh Univ. Press, 1990.

[10] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition - a feature-based approach," *IEEE Signal Processing Mag.*, pp. 58-71, Sep. 1996.

[11] Allen, J.B., "How do humans process and recognize speech?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4., pp. 567-577, 1994.

[12] B.A. Hanson, H. Wakita, "Spectral slope distance measure with linear prediction analysis for word recognition in noise," *IEEE Trans On Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 7, Jul. 1987.