음성인식을 위한 주파수 부대역별 효과적인 특징추출

지상문*

Effective Feature Extraction in the Individual Frequency Sub-bands for Speech Recognition

Sang-Mun Chi

이 논문은 2003학년도 경성대학교 특별과제연구비에 의하여 연구되었음

요 약

본 논문에서는 주파수 부대역마다 최적의 특징추출을 위해서, 음성인식률을 기준으로 최적의 방법을 선택한다. 다중대역 음성인식 접근을 사용하여 각기 다른 주파수 영역에서 특징벡터를 독립적으로 추출함으로써 부대역별로 다른 특징추출 방법을 적용할 수 있었다. 저주파 대역의 음성은 비교적 스펙트럼의 구조가 명확하므로 전극모델을 사용하는 것이 효과적이었고, 고주파 대역에서는 비모수적인 변환방법인 이산 코사인 변환을 사용한 켑스트럼이 효과적이었다. 부대역별로 효과적인 특징추출 방법을 사용함으로써, 각 주파수 부대역에 포함된 음성인식을 위한 언어정보를 보다 효과적으로 추출할 수 있었다. 음성인식 실험결과, 제안한 방법은 전대역 특징추출보다 우수한 성능을 나타내었다.

ABSTRACT

This paper presents a sub-band feature extraction approach in which the feature extraction method in the individual frequency sub-bands is determined in terms of speech recognition accuracy. As in the multi-band paradigm, features are extracted independently in frequency sub-regions of the speech signal. Since the spectral shape is well structured in the low frequency region, the all pole model is effective for feature extraction. But, in the high frequency region, the nonparametric transform, discrete cosine transform is effective for the extraction of cepstrum. Using the sub-band specific feature extraction method, the linguistic information in the individual frequency sub-bands can be extracted effectively for automatic speech recognition. The validity of the proposed method is shown by comparing the results of speech recognition experiments for our method with those obtained using a full-band feature extraction method.

키워드

부대역 특징추출, 전대역 특징추출, 다중대역 음성인식

I. 서 론

최근에 인간의 음성인식에 대한 Fletcher의 연구결과를 음성인식시스템에 적용하려는 시도로서, 다중대역 음성인식 방법이 연구되고 있다. 다중대역 음성인식은 인간이 음성을 인식할 때, 전체 주파수 대역을 대상으로 하지 않고 다수의 부대역으로 나누고, 독립적으로 인식한다는 연구결과에 기반을 두고 있다[1]. 다중대역 음성인식의 주요 연구 분야는 부대역의 정의 및 구성 방법, 부대역별 효과적인 특징 추출 및 인

*경성대학교 컴퓨터과학과

접수일자 : 2003. 4. 23

식 단위 선정, 부대역간의 비동기적 인식, 부대역의 인식모델, 각기 다른 부대역 인식결과의 통합방법 등으로 알려져 있다[2,3].

부대역간의 비동기적 인식에 관한 연구는 각기 다른 부대역에서의 음운의 천이는 동기적이지 않다는 관찰로부터, 주파수 부대역의 인식결과를 비동기적으로 결합함으로써 보다 음성의 특성을 효과적으로 이용하려는 방법이다. 그러나, 비동기적 결합을 위해 고려하여야 할 결합의 위치가 많아서 연산량이 크게 증가하는 단점과 더불어 동기적인 음성의 정보인 부대간의 특징적 상관관계를 상실하여 오히려 인식률이 저하되기도 한다[3].

가장 활발히 연구되는 부대역 인식결과를 통합하는 방법은 부대역 인식 결과에 동일한 가중을 주거나 부대역의 인식 정확도로 가중을 하는 방식 또는 신경망에 의한 통합 등이 연구되고 있으며[2,4], MCE(minimum classification error) 학습에 의해 각 부대역의 가중치가 인식 집단간의 변별력을 최대화하도록 가중치를 부여하는 방법[5], 부대역 상호 정보 또는 최대우도에 기반한 부대역 신뢰도 측정방법[6], 입력음성으로부터 부대역 가중치를 최적화하기 위해, 대역의 신호대 잡음비에 의한 가중방법과 부대역 상호 정보를 이용한 가중[7] 등이 연구되었다. 이밖에도 부대역별로 학습된 은닉 마르코프 모델을 이용하여 부대역의 가중치를 추정하는 방법과[8] 주파수 부대역의 켑스트럼 벡터의 차원을 조정하여 주파수 부대역의 언어정보가 최적으로 가중되게 하는 방법[9]이 있다.

본 논문의 연구대상인 부대역별 효과적인 특징 추출 및 인식 단위 선정방법에 대한 연구는 보고되지 않고 있다. 사람의 말소리를 이루는 음운은 각기 음향물리학적인 특성이 주파수 대역별로 다르다. 예를 들어, 모음의 경우에는 포만트(formant)와 배음(harmonics)이 나타나는 반면에, 대부분의 자음들은 이러한 특성이 없고 고주파 대역에 에너지가 분포한다. 따라서, 주파수 부대역별로 효과적인 음성인식을 위해서는 각 부대역에 특징적으로 발생하는 음성인식 정보를 효과적으로 추출하여야 한다. 하지만, 부대역별 최적화된 특징추출을 위한 물리음향학적 분석은 보고되지 않고 있다. 본 논문에서는 주파수 부대역별로 나타나는 음향학적 특성을 분석하기에 앞서, 기존의 특징추출 방법을 부대역별로 적용하여 효과적인 특징추출 방법을 조사해 보고자 한다.

2장에서는 부대역별 특징추출을 위한 다중대역 음성인식 접근의 이용에 대해서 설명하고, 3장에서는 부대역별로 적용할 수 있는 특징추출 방법들을 알아본다. 4장에서는 음성인식 실험 및 결과를 기술하고, 마지막으로 5장에서 결론을 맺는다.

II. 다중대역 음성인식을 이용한 부대역 특징 추출

주파수 전대역에서 특징을 추출하게 되면 특정 주파수 부대역에서 각기 다른 특징추출 방법을 적용할 수 없다. 따라서, 주파수 부대역마다 효과적인 특징추출 방법을 사용하기 위해서는 부대역별로 독립적인 특징추출이 가능하여야 한다. 본 논문에서는 이를 위하여 다중대역 음성인식 접근을 사용한다.

다중대역 음성인식은 그림 1에서 보듯이, 주파수 전대역을 여러 개의 부대역으로 나누고, 부대역별로 인식을 수행한 후에 최종 결과를 통합한다. 즉, B개의 부대역에서 추출된 각각의 특징벡터를 $x_1, x_2, \cdots, x_B$ 라 하고, 부대역 특징들이 확률적으로 서로 독립이라고 가정하면.

$$\Pr(x|\lambda) = \prod_{b=0}^{B} \Pr(x_b|\lambda) = \prod_{b=0}^{B} \Pr(x_b|\lambda_b). \quad (1)$$

단, λ는 전대역 특징을 사용한 음성인식 모델이고 λ_b는 부대역 특징을 사용한 음성인식 모델이다. 식 (1)의 양변을 로그 변환하면

$$\log \Pr(x|\lambda) = \sum_{b=0}^{B} \log \Pr(x_b|\lambda_b) \quad (2)$$

다중대역 음성인식에서는 식 (2)를 사용하는 대신에 부대역마다 최종 음성인식에 기여하는 정도 w_b를 조정할 수 있는 식 (3)을 사용한다.

$$\log \Pr(x|\lambda) = \sum_{b=0}^{B} w_b \circ \log \Pr(x_b|\lambda_b) \quad (3)$$

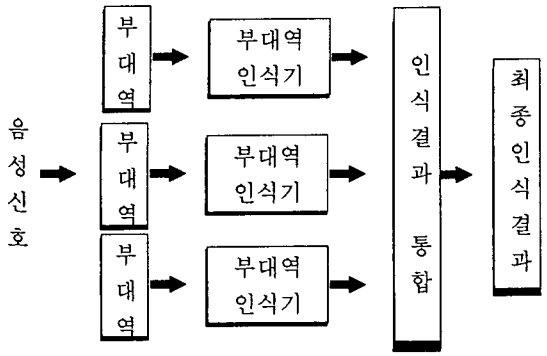

그림 1. 다중대역 음성인식의 개략도

Fig. 1 A simplified overview of multi-band speech recognition

일반적인 다중대역 음성인식은 주파수 부대역을 독립적으로 처리하므로, 주파수 대역의 일부분을 오염시키는 잡음음성의 인식에는 효과적이나, 부대역간의 상관관계나 부대역간의 특징적인 스펙트럼 천이 패턴를 이용하지 않으므로 조용한 환경의 음성인식에는 오히려 인식률이 감소하는 경향이 있고[10], 부대역 특징과 전대역 특징은 서로 상호 보완할 수 있는 특성 때문에 전대역 특징벡터를 보조적으로 사용할 경우 성능이 향상된다는 보고가 있다[10].

본 논문의 특징추출에서도 다중대역 음성인식 방법과 동일하게 각각의 주파수 부대역에서 특징벡터를 독립적으로 추출한다. 하지만, 일반적인 다중대역 음성인식방법처럼 부대역별로 추출된 특징벡터에 대해 인식기를 독립적으로 구성하지 않고, 부대역간의 동기적 인식과 학습을 수행한다. 이러한 방법은 음성인식의 중요정보인 부대역간의 특징적인 스펙트럼 천이 패턴과 부대역간의 상관관계를 유지할 수 있게 하여 조용한 환경의 음성의 인식에도 성능을 유지할 수 있게 한다. 또한 부대역마다 독립적으로 특징을 추출하므로, 주파수 일부분의 오염은 해당되는 주파수 부대역외의 부대역에서 추출한 특징벡터에는 영향을 미치지 않는 장점이 있다.

III. 부대역 특징추출

주파수 부대역별로 특징추출을 위해서는 음성을 여러 개의 부대역으로 분리하여야 한다. 본 논문에서는 음성신호에 대해 필터뱅크 분석을 수행하고, 각 필터뱅크의 에너지를 부대역별로 나누는 방법을 사용한다.

먼저 주파수 전대역에 대해서 필터뱅크 분석을 수행한다. 본 논문에서는 바아크 스케일 필터뱅크 분석을 사용한다. 바아크 단위는 인간의 청각 처리 단위를 공학적으로 근사한 주파수 스케일이다[11]. 바아크 스케일 필터뱅크를 통해서 얻은 필터뱅크의 에너지를 E 라고 할 때,

$$E=[e_1, e_2, \cdots, e_d]^T \tag{4}$$

여기서 e_k는 k번째 필터뱅크의 에너지이다.

주파수 부대역별로 독립적인 특징추출을 위해서, 필터뱅크 에너지 E는 B개의 주파수 부대역들로 분할한다.

$$[[e_{s(1)}, \cdots e_{e(1)}], \cdots [e_{s(B)}, \cdots e_{e(B)}]]^T \tag{5}$$

여기서, $s(b)$와 $e(b)$는 각각 b번째 주파수 부대역의 시작과 마지막의 필터뱅크 번호이다.

본 논문에서는 b번째 주파수 부대역을 이루는 필터뱅크 에너지 $[e_{s(b)}, \cdots e_{e(b)}]]^T$로부터 특징벡터를 추출할 때 두 가지 방법을 사용하였다. 첫 번째 방법은 PLP 분석[11]과 유사하고, 이를 MPLP (Modified Perceptual Linear Predictive) 방법이라 하자. MPLP 방법에서는 주파수 부대역을 이루는 필터뱅크 에너지를 IDFT(inverse discrete Fourier transform)을 통하여 자기상관 함수를 얻고, 이를 다시 Yule-Walker 방정식을 푸는데 사용하여 전극(all pole)모델의 자기회귀계수 (autoregressive coefficients)를 얻는다. 이 계수를 켑스트럼 계수로 변환하여 특징벡터로 사용하였다. 그러나, PLP 분석에서 사용하는 equal loudness preemphasis와 intensity loudness power law는 적용하지 않았다. 이는 잡음 음성을 인식할 경우에는 효과적이나 조용한 환경의 음성인식에는 오히려 성능을 약간 저하시키는 예비 실험 결과를 보였기 때문이다.

본 논문에서 주파수 부대역을 이루는 필터뱅크 에너지로부터 특징벡터를 추출하는 두 번째 방법은 DCT (discrete cosine transform)를 사용하여 필터뱅크 에너지를 켑스트럼 계수로 변환하는 방법이다. 이를 BCEP (Bark Cepstrum)이라 하자.

$$C_{b,n}=\sum_{k=s(b)}^{e(b)}\frac{\log(e_k)\cos(k+0.5-s(b))*n*\pi}{e(b)-s(b)+1} \quad (6)$$

여기서 b는 부대역 번호이고, n은 n번째 켑스트럼 계수임을 나타낸다.

IV. 실험 및 결과

4.1 실험 환경

음성인식 실험에는 TIDIGITS 자료를 사용하였다[12]. 남성 55명, 여성 57명이 발성한 음성을 HMM(hidden Markov model)의 학습을 위하여 사용하였고, 학습에 사용되지 않은 남성 20명, 여성 20명이 발성한 음성은 부대역별 최적 특징추출 방법을 탐색하는데 사용하였고, 이외의 남성 36명 여성 37명이 발성한 음성을 평가에 사용하였다. 각 화자는 11개의 영어 숫자(zero, oh, one, two, three, four, five, six, seven, eight, nine)로 이루어진 숫자열을 발성한다. 각 화자는 11개의 한자리 숫자들을 두 번씩, 그리고 두 자리부터 일곱 자리까지 연결된 숫자를 55번 발성하였다.

16개의 Gaussian mixture와 10개의 상태를 가지는 연속분포 HMM을 사용하여 음성을 모델링 하였다. 분석프레임의 길이는 32ms이고 16ms씩 이동하였고, 19개의 바아크 스케일 필터뱅크를 사용하였다. 특징추출 방법의 비교를 위하여 정적인 특징만을 사용하였다.

4.2 실험 결과

본 논문의 비교 대상이 되는 주파수 전대역에서 추출한 특징벡터의 음성인식 성능을 알아본다. 그림 2는 학습에 참여하지 않은 평가자료인 남성 36명 여성 37명이 발성한 음성을 대상으로 한 음성인식실험에서 숫자열의 문장 인식률을 보여준다. 문장 인식률은 숫자열에 포함된 모든 숫자가 정확히 인식된 비율이다. 가장 높은 인식률은 MPLP와 BCEP이 각각 12차원일 경우 94.47%이었다. 여기서, MPLP는 3장에서 설명한 조용한 환경의 음성을 인식하기 위해 PLP 분석과정의 일부를 적용하지 않은 특징추출 방법이고, BCEP은 바아크 스케일 필터뱅크 분석과 DCT를 사용하여 얻은 켑스트럼이며, LCEP은 선형예측계수에서 유도한 켑스트럼이다. 그림 2에서 보듯이 MPLP와 BCEP이 LCEP보다 우수한 성능을 보였다. 이밖에도 가산 또는 채널잡음의 존재하에 우수한 성능을 보이는 것으로 알려진 특징추출 방법인 RASTA와 J-RASTA[13]를 사용하여 실험하였으나, 조용한 환경의 음성이므로 인식성능이 MPLP, BCEP, LCEP보다도 저하되어 이후의 실험에는 사용하지 않았다.

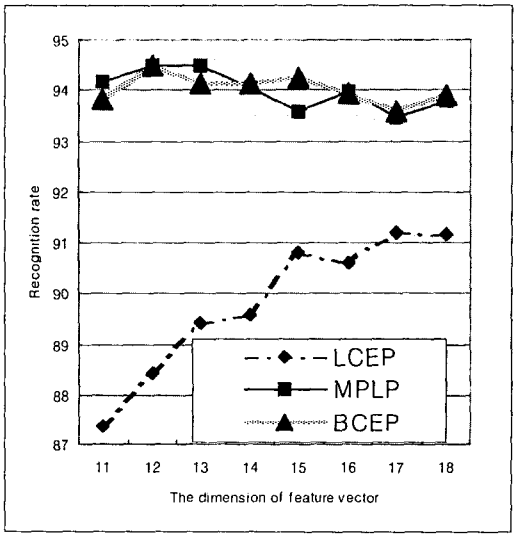

그림 2. 전대역 특징들을 사용한 문장 인식률 (%)
Fig. 2 Sentence accuracy using full-band features

그림 2의 결과를 바탕으로, 전대역에서 인식률이 높은 두 개의 특징추출 방법만을 사용하여 부대역별로 효과적인 특징추출 방법이 어떤 것인가를 조사하였다. 전대역 주파수를 바아크 스케일에 따라 균등하게 삼등분하여 각각의 부대역에 두 개의 특징추출 방법을 적용하였다. 표 1에서는 학습에 사용되지 않은 남성 20명, 여성 20명이 발성한 음성을 사용하여, 부대역별로 효과적인 특징추출 방법을 탐색하는데 사용하였는데, 이는 최종 평가자료와는 다른 별도의 자료이다. 표 1에는 각 부대역의 특징벡터의 차원이 4와 5일 때의 평균 인식률을 나타내었다. 인식률의 차이는 크지는 않았으나, 저대역에서는 음성의 배음(harmonics)이 잘 나타나고, 명확한 스펙트럼 구조를 가지므로, 극(pole)의 위치에 중점을 두어 특징을 추출하는 전극(all-pole) 모델을 사용하는 MPLP가 BCEP보다 인식률이 높았고, 중대역에서는 거의 비슷하지만

BCEP이 조금 성능이 높았고, 고대역에서는 비모수적인 특징추출인 BCEP이 우수하였다. 이러한 결과는 음성신호의 특성으로부터 어느 정도 예측할 수 있었던 결과이나, 보다 효과적인 처리를 위해서는 기존의 특징추출 방법을 부대역별로 적용하는 것보다는 부대역에 특화된 특징추출을 개발하는 것이 필요하다.

표 1. 세 개의 주파수 대역별 문장인식률
Table. 1 Sentence accuracies for three sub-bands

대역 / 특징	저대역	중대역	고대역
MPLP	69.725	82.37	48.23
BCEP	69.205	82.385	48.34

표 2에서는 표 1에서 각 부대역에서 효과적인 특징추출 방법을 탐색할 때 사용한 음성자료와는 별개인 자료로서, 학습에 참여하지 않은 남성 36명과 여성 37명이 발성한 음성을 사용하였고, 각 부대역 벡터의 차원은 5로 고정하였다. 표 1의 결과로부터 저대역에서는 MPLP가 중대역과 고대역에서는 BCEP이 효과적인 특징추출 방법으로 선택되었고, 이를 적용한 방법의 성능을 조사한 결과 95.34%로 가장 성능이 좋았고, 전대역 특징의 최적 성능인 94.47%보다 인식률이 향상되었다. 표 2를 살펴보면, 효과적인 특징추출과는 다른 특징추출 방법의 조합을 사용하였을 경우에도 전대역 특징보다는 조금 성능이 높았으나, 부대역별로 효과적인 특징추출 방법을 적용한 경우보다는 성능이 좋지 않았다.

표 2. 대역별 상이한 특징을 사용한 문장인식률
Table. 2 Sentence accuracies using different feature for each sub-band

저대역 특징	중대역 특징	고대역 특징	인식률
BCEP	BCEP	BCEP	95.21
BCEP	BCEP	MPLP	95.02
BCEP	MPLP	BCEP	95.05
BCEP	MPLP	MPLP	94.98
MPLP	BCEP	BCEP	95.34
MPLP	BCEP	MPLP	95.14
MPLP	MPLP	BCEP	95.07
MPLP	MPLP	MPLP	94.79

V. 결 론

본 논문에서는 각각의 주파수 부대역에 효과적인 특징추출 방법을 사용함으로써, 주파수 부대역에 포함된 음성인식을 위한 언어정보를 보다 효과적으로 추출하려는 시도를 하였다. 이를 위하여 기존의 특징추출 방법을 주파수 부대역마다 독립적으로 적용하여 보았다. 하지만, 일반적인 다중대역 음성인식방법처럼 부대역별로 추출된 특징벡터에 대해 인식기를 독립적으로 구성하지 않고, 부대역간의 동기적 인식과 학습을 수행하여 음성인식의 중요정보인 부대역간의 특징적인 스펙트럼 천이 패턴과 부대역간의 상관관계를 유지할 수 있게 하여 조용한 환경의 음성의 인식에도 성능을 유지할 수 있게 하였다.

본 논문의 실험결과에 따르면, 명확한 스펙트럼 구조를 가지는 저주파 대역의 음성신호는 극(pole)의 위치에 중점을 두어 특징을 추출하는 전극(all-pole)모델을 사용하는 방법이 인식률이 높았고, 고주파 대역에서는 비모수적인 특징추출 방법이 우수하였다. 이러한 결과는 음성신호의 특성으로부터 예측할 수 있는 결과와도 일치함으로서, 향후 부대역별 최적 특징추출을 위한 더 많은 연구를 수행하면 보다 향상된 결과를 얻을 수 있을 것이라 예측된다.

본 논문에서는 부대역의 특징추출을 위해서 기존의 특징추출 방법을 사용하여 전대역 특징보다 약간 향상된 결과를 얻었지만, 음성인식을 위한 부대역의 물리음향학적 특성에 대한 심도 있는 분석이 병행된다면 더욱 향상된 결과를 얻을 수 있으리라 판단된다. 또한, 제안한 방법의 성능을 검증하기 위해서는, 많은 인식단위의 변이를 포함하고 있는 대용량 연속음성인식에도 적용해 볼 필요가 있다.

참고문헌

[1] J. B. Allen, "How do humans process and recognize speech?," IEEE Trans. On Speech and Audio Processing, 2 (4), 567-577, October 1994.

[2] H. Bourland and S. Dupont. "ASR based on independent processing and recombination of partial frequency bands," Proc. Int. Conf. on Spoken Language Processing, 1. 422-425, 1996.
[3] H. N. Mirghafori, "A multi-band approach to automatic speech recognition," ICI TR-99-04, 1999.
[4] H. Hermansky, S. Tibrewala and M. Pavel, "Towards ASR on partially corrupted speech," Proc. Int. Conf. on Spoken Language Processing, 1. 462-465, 1996.
[5] C. Christophe, H. J. Paul and F. Dominique, "Towards a global optimization scheme for multi-band speech recognition," Proc. EUROSPEECH, 2, 587-590, 1999.
[6] Y. C. Tam and B. Mak, "Optimization of sub-band weights using simulated noisy speech in multi-band speech recognition," Proc, Int. Conf. on Spoken Language Processing, 2000.
[7] S. Okawa, T. Nakajima and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," Proc. EUROSPEECH, 2, 603-606, 1999.
[8] 조훈영, 지상문, 오영환, "다중대역 음성인식을 위한 부대역 신뢰도의 추정 및 가중," 한국음향학회지, 제 21권 제 6호, 2002.
[9] 지상문, 조훈영, 오영환, "주파수 부대역의 켑스트럼 해상도 최적화에 의한 특징추출," 한국음향학회지, 제 22권 제 1호, 2003.
[10] C. Cerisara and D. Fohr, "Multi-band automatic speech recogniton," Computer Speech and Language, 15, 151-174, 2001.
[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. 87 (4), 1738-1752, April 1990.
[12] R. G. Reonard, "A database for speaker-independent digit recognition," Proc. ICASSP, 3, 42.11/1-4, 1984.
[13] H. Hermansky and N. Morgan, "RASTA Processing of speech," IEEE Trans. On Speech and Audio Processing, 2 (4), 578-589, October 1994.

저자 소개

지상문(Sang-Mun Chi)

1991년: 서울대학교 수학교육과(학사)
1993년: 과학기술원 수학과 (석사)
1998년: 과학기술원 전산학과 (박사)
1993년~2000년: 삼성전자 (선임연구원)
2000년~2001년: L&H (책임연구원)
2001~현재: 경성대학교 컴퓨터과학과 전임강사
※관심분야 : 음성처리, 패턴인식