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

연속숫자 음성인식에서 화자 적응에 관한 연구

최광표°, 윤재선, 홍광석
성균관대학교 전기전자컴퓨타공학부 HCI연구실

A Study on Speaker Adaptation in Continuous Digits Speech Recognition

Kwang-Pyo Choi, Jeh-Seon Youn, Kwang-Seok Hong
HCI Lab, Electronic Engineering, Sung Kyun Kwan University
sunhci@chollian.net, kshong@yurim.skku.ac.kr

요 약

본 논문에서는 반음절 단위 HMM을 이용한 연속 숫자 음성인식 시스템의 2단계로 이루어지는 화자 적응 알고리즘을 수행하였다.

음성인식 시스템에서 사용되는 훈련데이터의 양이 많더라도 발성속도, 발성크기 등의 화자 발성 습관에 따라 화자독립 음성인식 시스템에서는 많은 문제점들이 발생하게 된다. 불특정 화자를 대상으로 한 음성 인식에 있어서 개인차에 의한 변동을 대처하는 방법으로 유효한 음향적 특성을 추출하기 위해 스펙트럼의 동적인(Dynamic) 특성을 주로 이용하고 있다.

따라서 본 논문에서는 화자 적응 기법의 하나인 frequency warped spectral matching방법을 연속숫자 음성인식시스템에 적용하였으며, 이때 인식에 의한 적절한 화자별 스케일링 계수 선정 방법을 수행하여 오인식률이 감소함을 확인하였다.

I. 서 론

음성 인식분야에서 최종적으로 필요로 하는 기술 중의 하나는 불특정 화자를 대상으로 한 음성인식이다. 특정화자의 음성인식에 있어서는 고립단어 뿐만 아니라, 연속음성에 있어서도 높은 인식률을 보이고 있으나, 불특정화자의 음성인식에 대해서는 그다지 좋은 성능을 보이고 못하고 있는 실정이다. 이러한 불특정화자 음성인식의 문제는 단어를 발성하는 화자가 바뀌에 따라 성도길이, 구강 크기 등의 해부학적 차이와 액센트, 발성 속도, 발음 크기 등의 화자 발성 습관에 따른 차이 등에 의해 발생하게 된다. 이러한 단점을 해결하기 위해서는 화자적응이 필요하다.[1]

따라서 본 논문에서는 반음절HMM을 이용한 연속 숫자음성 인식시스템에서 특징파라미터를 추출할 때 scale factor를 추가함으로써 성대파 스펙트럼과 성도의 길이를 정규화하는 방법을 제안하여 그 성능을 확인한다.

II. 반음절 단위 HMM시스템에 적용한 화자 적응

2.1 반음절 단위 HMM 시스템

본 논문에서 입력된 숫자음의 인식방법은 입력음성의 모음 구간 분할을 통해 반음절, 반음절 + 반음절, …, 반음절 + 반음절, 반음절 단위로 구분한 후 각각의 분할된 반음절 단위들을 반음절, 반음절+반음절 HMM과 비교하며 인식하는 것이다. 이 방법은 그림 1에 나타내었다.

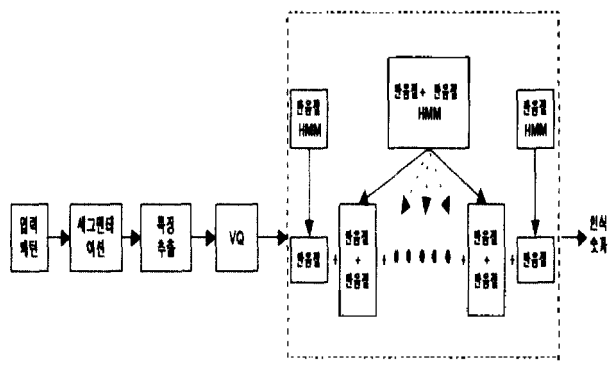


그림 1. 반음절 단위 HMM 인식 시스템

또한 인식된 반음절, 반음절+반음절 패턴에는 다음에 올 반음절 패턴 정보를 포함되어 있기 때문에 규칙을 적용하여 인식후보를 줄일 수 있다.[2]

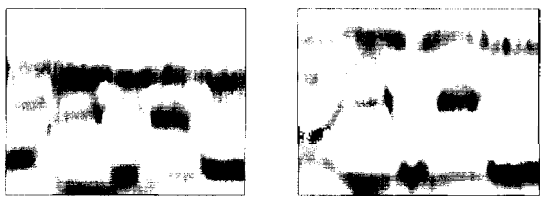
2.2 Spectral Warping

Vocal Tract 길이에 따른 Spectrum상의 차이점은 파이프 내에서 음파를 발생시키는 것과 유사하다. 그 공명 주파수는 식(1)과 같다.

$$f = \frac{v}{aL} \tag{1}$$

a : 진동수에 따라 결정되는 상수, L : 파이프의 길이

각 진동 주파수는 파이프의 길이에 따라 값이 달라지는 것을 알 수 있다. 이것을 Vocal Tract에 적용을 한다면 각 화자마다 Vocal Tract의 길이가 다르므로 같은 단어를 발성하더라도 Peak성분의 위치가 주파수에 따라 조금씩 다르게 된다. 극단적인 비교로 여자의 경우는 남자에 비해 일반적으로 Vocal Tract의 길이가 짧으므로 주파수 특성은 남성에 비해 높은 특징을 가지고 있다. 연속으로 발성한 "아 아 우 에 오"를 성별에 따라 분석한 스펙트로그램을 그림 2에 나타내었다.



(a) 남자 음성 (b) 여자 음성

그림 2. 성별에 따른 스펙트로그램

따라서 본 논문에서는 각 화자의 vocal tract의 길이의 차이를 spectral warping을 통해 보상함으로써 화자독립의 인식성능을 향상시키고자 한다.

적용된 spectral warping기법은 다음과 같은 식을 사용한다.[3]

$$\omega = \arctan\left[\frac{(1-\alpha^2)\sin\theta}{2\alpha+(1+\alpha^2)\cos\theta}\right] \tag{2}$$

α : 화자마다 결정되는 scale factor

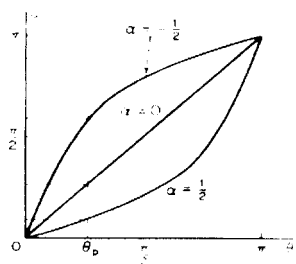


그림 3. Frequency Warping

α에 따른 spectral warping을 그림 3에 나타내었다. 식(2)를 이용함으로써 warping 기법의 불연속점의 문제점을 해결하였으며 α값을 각 화자마다 설정을 하고 그 값에 따라 각 화자의 음성으로부터 추출된 특징 벡터를 Warping 함으로써 개인차를 줄일 수 있다.

기존의 반음절 HMM 시스템의 음성특징 추출부분에 spectral warping을 추가하면 그림 4와 같다.

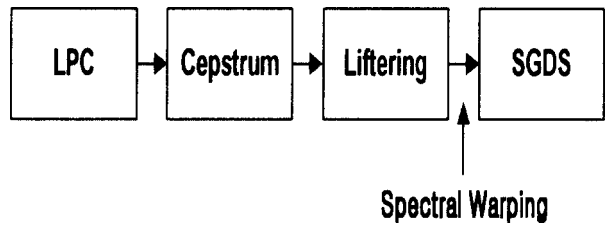


그림 4. 음성특징벡터 추출

Smoothed Group Delay Spectrm(SGDS)의 식은 다음과 같다.[4]

$$S(\theta) = \sum_{i=0}^{M-1} c_i \cos\left(\frac{\pi \cdot i \cdot \theta}{M}\right) \tag{3}$$

M : 특징벡터 차수, c_i : 캡스트럼 계수

식(3)의 θ를 식(2)의 ω로 변환함으로써 spectral warping을 취하면 식(4)와 같다.

$$S(\theta) = \sum_{i=0}^{M-1} c_i \cos\left(\frac{\pi \cdot i}{M} \cdot \arctan\left[\frac{(1-\alpha^2)\sin\theta}{2\alpha+(1+\alpha^2)\cos\theta}\right]\right) \tag{4}$$

α에 따라 spectral warping을 통해 변화된 SGDS를 그림 5에 나타내었다.

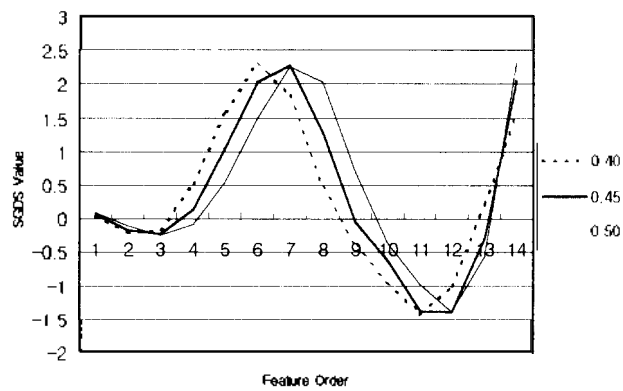


그림 5. α 값에 따른 SGDS

그림 5에서는 α 값에 따라 SGDS의 peak값이 변화됨을 알 수 있다

III. 실험 및 결과

본 논문에서 사용한 음성 데이터는 소음이 45~50dB 정도 되는 사무실 환경에서 남성 화자 20명이 0~9의 단독 발성음 1회씩 발성한 것과 00~99까지 두자리 숫자음 1회씩 자연스럽게 발성한 음성 신호를 16bit, 11.025kHz로 샘플링하여 사용하고, 이로부터 반음절 20종 각각 20개와 반음절 + 반음절 100종 각각 20개를 훈련 데이터로 사용하였다. 반음절 표준패턴의 상태수는 2개를 두었고, 반음절 + 반음절 표준패턴은 3개의 상태수로 구성하였다.

음성신호는 20ms의 프레임 단위로 Hamming window를 사용하였고, 10ms씩 중첩하여 14차 SGDS계수를 구하였다. K-means알고리즘을 이용한 벡터 양자화에서 코드북은 128레벨로 하였다.[5]

인식 실험 데이터는 남성 독립 화자 4명이 한 자릿수(0~9) 10개, 두 자릿수(00~99) 100개의 숫자음을 1번씩 자연스럽게 발성한 음성은 자동으로 끝점 추출 및 반음절 분할을 하여 Scale factor(α)에 따른 한자리 숫자음에 대한 인식 결과는 표 1, 두자리 숫자음에 대한 인식 결과는 표 2에 나타내었다. α의 범위는 -0.05~0.05, step size는 0.01로 하였으며, α값이 0인 경우는 spectral warping을 적용하지 않았을 때의 인식률이다.

α에 따른 한자리 숫자음의 인식률은 동일하게 나타났으며, 두자리의 숫자음의 인식률은 화자 1을 제외하고 화자 2는 86%에서 89%로 화자 3은 80%에서 90%로, 화자 4는 71%에서 76%로 향상되었다. 두자리 숫자음에서 spectral warping을 적용하지 않았을 때의 화자 4명의 평균은 77%이며, 가장 높은 인식률을 나타내는 값을 기준으로 화자적응기법을 적용했을 때의 인식률은 81.5%로 4.5%의 인식성능이 향상되었다.

표 1. 한자리 숫자음 인식결과

(단위 : %)

α \ 화자	1	2	3	4
-0.05	80	100	100	90
-0.04	80	100	100	90
-0.02	80	100	100	90
0	**80**	**100**	**100**	**90**
0.02	80	100	100	90
0.04	80	100	100	90
0.05	80	100	100	90

표 2. 두자리 숫자음 인식결과

(단위 : %)

α \ 화자	1	2	3	4
-0.05	71	85	84	73
-0.04	69	85	83	75
-0.03	66	86	83	70
-0.02	65	84	82	72
-0.01	67	85	81	69
0	**71**	**86**	**80**	**71**
0.01	70	85	82	73
0.02	67	86	83	74
0.03	66	88	85	70
0.04	71	89	85	75
0.05	70	83	90	76

표 2에서는 화자에 따라 가장 높은 인식률을 나타내는 scale factor가 화자 1은 -0.05, 0.05와 0, 화자 2는 0.04, 화자 3은 0.05, 화자 4는 0.05이다. 즉 화자에 따라 성도의 길이를 정규화하는 α값이 각각 다름을 알 수 있다

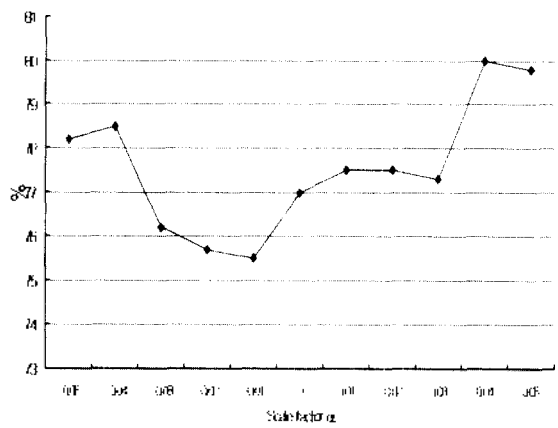


그림 6. α에 따른 4명 화자의 평균

두자리 숫자음에 대한 α에 따른 화자의 평균을 그림 6에 나타내었다.

그림 6을 보면 대체적으로 화자 적응 기법을 적용하지 않은 평균값보다는 spectral warping을 적용했을 때의 평균값이 높은 값을 가지는 것을 볼 수 있으며, α가 0.04인 때 가장 높은 인식률을 나타내었다.

Ⅳ. 결론

본 논문에서는 반음절 단위 HMM을 이용한 연속 숫자 음성인식 시스템에 spectral warping을 적용한 화자 적응을 수행하였다. 최초의 인식을 수행하기 전에 소수의 단어 혹은 짧은 문장을 발성하여 그 사람의 음성에 자동적으로 적응할 수 있는 화자 적응 시스템을 구현[6]하기 앞서 한자리 숫자음과 두자리 숫자음에 대해 spectral warping을 적용하여 인식률을 확인하였다.

한자리 숫자음에서는 scale factor에 따른 인식률의 차이가 없었으나 두자리 숫자음에서 spectral warping을 적용하지 않았을 때의 평균인식률은 77%이었으며, 가장 높은 인식률을 나타내는 값을 기준으로 화자적응기법을 적용했을 때의 인식률은 81.5%로 4.5%의 인식성능이 향상되었다. 또한 화자에 따라 정규화하는 scale factor가 각각 다름을 확인하였다.

따라서 앞으로는 최초의 인식을 수행하기 전에 소수의 단어나 짧은 문장을 발성하여 scale factor를 찾는 연구를 진행할 계획이다.

참고 문헌

[1] Li Lee, Richard Rose, "A Frequency Warping Approach to Speaker Normalization," IEEE Trans. on Speech and Audio Processing, Vol. 6, No. 1, pp. 49-60, January 1998.

[2] 윤재선, 홍광석, "반음절 단위HMM을 이용한 연속 숫자 음성인식," 한국음향학회지 제17권 제5호, pp.73-78, 1998.

[3] Alan V. Oppenheim, Donald H. Johnson, "Discrete Representation of Signals," IEEE Vol.60, No. 6, pp.681-691, 1972.

[4] Itakura. J, Umezaki.T, "Distance measure for speech recognition based on the smoothed group delay spectrum," Proc. ICASSP, April, pp. 1257-1260, 1987.

[5] J.G. Wilpon, L.R. Rabiner, "A Modified K-Means Clustering Algorithms for use in Isolated Word Recognition," IEEE Trans. on Acoust. Speech and Signal Proc., Vol. 33, No. 3, pp. 587-594, 1985.

[6] K.Shikano er. al, "Speaker Adaptation through Vector Quantization," Proc. ICASSP 86, 49.5, 1986.