2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

선형워핑함수의 화자정규화에 의한 음성인식시스템의 성능향상

최석용*, 정경용*, 이정현*
*인하대학교 전자계산공학과
e-mail:syong@nlsun.inha.ac.kr

Performance Improvement of Speech Recognition System Based on Speaker Normalization Through Linear Warping Function

Seok-Yong Choi*, Kyoung-Yong Chung*, Jung-Hyun Lee*
*Dept of Computer Science and Engineering, In-Ha University

요 약

화자종속 음성인식 시스템은 훈련 데이터가 화자들 사이의 음향적 변이를 충분히 모델링할 수 있을 때, 화자독립 시스템보다 더 성능이 좋은 것으로 알려져 있다. 화자 정규화 기술은 입력음성의 스펙트럼을 수정하여 화자들 사이의 변이를 줄인다. 최근 성공적인 화자 정규화 알고리즘은 신호처리 단계에 화자 특유 주파수 워핑을 통합했다. 이런 알고리즘은 입력음성에 담겨있는 음향적 특징을 다 사용하지 않는다.

본 논문에서는 화자의 음향적 특징으로 세 개의 포만트 주파수를 이용하였고, 수집된 포만트 주파수들로부터 워핑함수를 정의하는데 선형회귀를 사용한 화자 정규화 방법을 제안한다. 이 방법을 사용하여 인식 성능을 향상할 수 있었다.

1. 서론

본 논문은 화자 특유 음향적 특징으로 주파수 축의 변환에 의해 달성되는 화자 정규화 방법을 제안한다. 화자 정규화는 화자들 사이의 변이를 줄여 음성인식 정확도를 향상시키는 것이다. 화자종속 시스템은 충분한 훈련 데이터가 가용할 때 화자독립 시스템보다 성능이 더 좋은 것으로 알려져 있다. 따라서 화자들 사이의 변이를 줄이는 것은 화자독립 시스템과 화자종속 시스템 사이의 차이를 줄여 성능을 향상하는 것이다.

화자 변이에는 여러 가지 원인들이 있다. 화자의 문화적 배경, 감정상태 등과 같은 외부적 요인과 성도의 형태와 크기의 차이에서 나타나는 내부적 요인이 그것이다. 성도의 차이는 명백하게 포만트의 위치로 나타난다. 포만트는 스펙트럼 포락선의 봉우리이다. 물리적으로, 포만트는 성도의 공명과 일치한다[1]. 몇 가지 음소에서 포만트와 성도 길이 사이에 종속성이 선형적으로 보인다면, 주파수 축의 크기 변환이 정규화 기술로서 좋은 해결책이 된다는 것을 의미한다.

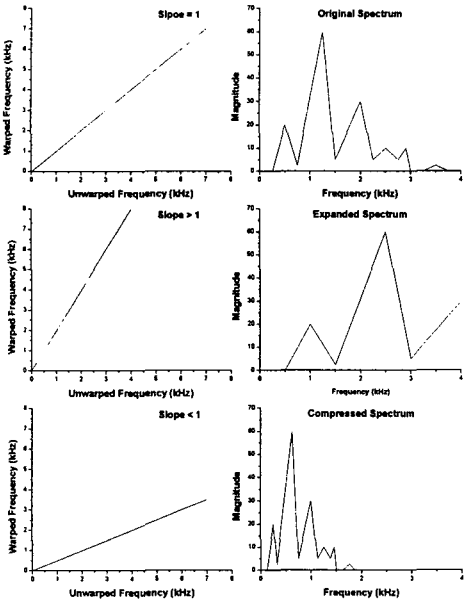
본 논문은 화자의 음향적 특징으로 세 개의 포만트 주파수를 이용하였고, 수집된 포만트 주파수의 중위수를 계산한 후, 특징점들을 보간하는 방법으로 선형회귀 기법을 이용하여 워핑함수의 파라미터인 기울기를 구하였다. 워핑함수를 사용하여 새로운 화자의 스펙트럼을 워핑한 후, 특징추출을 하여 인식 성능을 향상할 수 있었다.

2. 정규화 방법

워핑함수는 시스템적인 변이를 줄이기 위해 두 화자의 스펙트럼 사이의 매핑을 한다. 그러므로 화자 정규화의 맥락에서 워핑함수는 두 스펙트럼 사이의 매핑함수이다. 만약 스펙트럼 $X(w)$를 스펙트럼 $Y(w)$로 매핑하려 한다면, 함수 $f(w)$를 식(1)과 같이 사용할 수 있다.

$$Y(w) = X(f(w)) \qquad (1)$$

함수 $f(w)$의 결과는 f의 1차 도함수인 $f'(w)$가 단위보다 크거나 작거나에 따라 스펙트럼 $X(w)$의 확장이거나 축소가 결정되는데 [그림1]은 이 결과를 보여준다.



[그림1] 워핑함구 기울기에 따른 스펙트럼의 변화

2.1 포만트

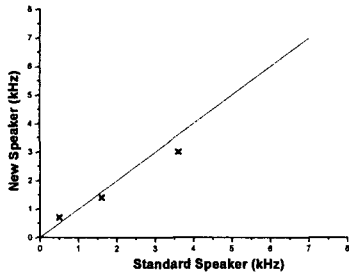
포만트는 모음의 확인에 아주 중요한 단서이다[1]. 본 논문에서는 포만트들로부터 추출된 특징을 사용하여 정규화를 이루는데 필요한 워핑함수를 구한다. 포만트의 분포를 구하는데 있어서 평균은 가장 단순하고 점들의 분포를 생각할 수 있는 가장 직접적인 측정이다. 그러나 중위수는 평균보다 특이점의 영향을 적게 받는다. 더욱이 본 논문에서는 포만트 주파수들의 두 분포들 사이의 매핑을 하려한다. 따라서 중위수의 사용은 매핑에서 가장 좋은 근사값이라 할 수 있다[4][5].

본 논문에서는 세 포만트 주파수 분포들의 중위수를 이용하여 워핑함수를 정의하고 이를 매핑한다.

2.2 선형워핑함수

본 논문에서는 포만트 주파수의 중위수에서 추출된 점들에 의해 워핑함수를 선택한다. 워핑함수는 훈련 셋에서 모든 화자들의 평균화된 특징값들과 특정한 화자에서 얻어진 특징들의 값을 비교하여 얻어진다.

[그림2]는 포만트기반 특징들로부터 얻어진 선형 워핑함수의 예를 보여준다. 세 점들의 세로좌표는 특정 화자로부터 계산된 포만트 중위수이고 가로좌표는 모든 화자에서 계산된 포만트의 중위수이다. 워핑함수는 이런 세 점들을 맞추는 것이다. 본 논문에서는 16kHz의 샘플링 주파수를 사용하기 때문에 최대 주파수는 나이키스트 주파수인 8kHz이다.



[그림2] 선형워핑함수의 예

선형회귀는 데이터점들에 가장 잘 맞는 직선 계수의 추정을 가능하게 한다. 워핑함수는 평면상의 특징들과 일치하는 점들을 맞추는 직선이다[6]. x_k는 기준 화자의 k-th 특징이고, y_k는 기준화자로 정규화하려고 하는 화자의 k-th 특징이다. 만약 모델 $y=ax+b$를 따르는 추정 $\widehat{y_k}$와 점 (x_k, y_k) 사이의 거리를 최소화하는 것을 선택하려면 다음의 식(2)를 최소화하여 파라미터 a와 b를 찾을 수 있다.

$$\sum_k (y_k - \widehat{y_k})^2 \qquad (2)$$

즉,

$$\sum_k [y_k - (ax_k + b)]^2 \qquad (3)$$

이다. 식(3)을 a와 b에 대해 편미분하여 식(4)와 (5)를 얻을 수 있고,

$$a\sum_k x_k^2 + b\sum_k x_k = \sum_k x_k y_k \quad (4)$$

$$a\sum_k x_k + b\sum_k = \sum_k y_k \qquad (5)$$

이를 정리하면 파라미터 a와 b를 얻을 수 있다.

여기에서 우리가 사용할 모델은 b를 0으로 두는 $y=ax$이므로, a는 식(6)과 같이 결정할 수 있다.

$$a = \frac{\sum_k x_k y_k}{\sum_k x_k^2} \qquad (6)$$

본 논문에서 식(6)을 사용하여 기울기 a를 구하여 점 (x_k, y_k)를 지나는 워핑함수를 얻는다.

2.3 정규화과정

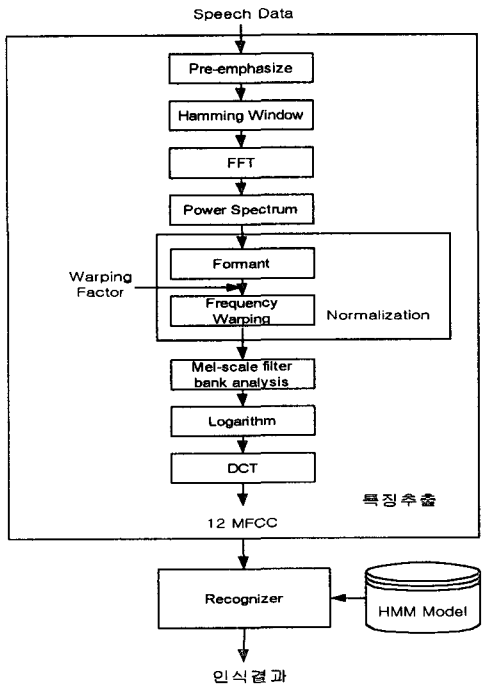
워핑함수에서 주요문제는 워핑함수 파라미터를 결정하는 것이다. 본 논문에서는 [알고리즘1]과 같은 과정으로 세 포만트의 중위수를 이용하여 워핑함수 파라미터를 구한다.

[알고리즘1] 워핑함수 파라미터 추출 알고리즘

1. 훈련데이터의 모든 프레임에서 포만트를 구한다.
2. 각 포만트(F1, F2, F3)의 중위수를 계산한다.
3. 훈련 데이터의 중위수와 워핑하려고 하는 화자의 중위수의 비율을 식(6)으로 구하고, 이에 따라 주파수축 워핑을 한다.
4. 워핑된 스펙트럼으로 음성 파라미터를 계산한다.

3. 인식시스템 기본 구성

정규화에 의한 음성인식 시스템의 구조는 [그림3]과 같으며 정규화 과정은 특징추출 내에 포함된다. 입력음성에 대하여 포만트를 미리 계산하여 정규화 인자를 먼저 구한 후, 정규화를 거친 데이터로 특징



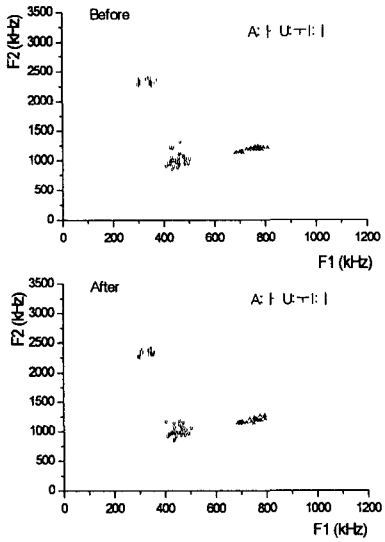
[그림3] 전체 시스템

벡터인 MFCC를 구하고 모노폰으로 구성된 HMM의 입력벡터로 사용한다[2].

4. 실험 및 평가

실험은 450MHz 펜터엄 PC와 Ultra 10 SPARC-II에서 16bit 사운드카드로 수행하였다. 음성신호는 표본화주파수 16kHz, 양자화해상도 16bit로 샘플링하였다. 제안된 시스템의 성능을 평가하기 위한 음성 데이터베이스는 20대 남자 10명의 화자가 47개의 음소를 포함한 20개의 단어를 2번씩 발음한 데이터로 구축하였고, 이 데이터베이스로부터 HMM을 학습시켰다. 실험은 잡음이 고려되지 않은 실험실에서 학습에 참여한 5명의 남자가 발음한 단어 20개로 구축하였다.

[그림4]는 /ㅏ//ㅜ//ㅣ/ 세 모음의 정규화 전과 후의 포만트 분포도를 나타내고 [표1]은 /ㅏ//ㅜ//ㅣ/ 세 모음의 정규화 전과 후의 중위수를 나타낸다.



[그림4] 정규화 전과 후의 포만트 분포도

[표1] 모음의 포만트 중앙값

	F1		F2		F3	
	Before	After	Before	After	Before	After
ㅏ	745.4	753.3	1199	1201.3	2707.2	2690.5
ㅜ	440.8	449.4	981.8	984	2758.6	2797.4
ㅣ	339.9	342.7	2309.6	2315.3	3136.5	3165.9

실험 결과를 평가하기 위해서 인식결과와 정규화의 정도를 계산하는 척도인 F-ratio를 이용하였다. F-ratio는 식(7)로 정의되고 전체 화자들의 평균치의 변이 값을 화자 각각의 변이의 평균으로 나눈 값이다. 이는 화자내 편차로 화자간의 편차를 나눈 것인데 화자내 편차는 작을수록 좋고 화자간의 편차는 클수록 좋다. 그러므로 F-ratio 값이 크면 좋은 특징이라 말할 수 있고 값이 작으면 성능이 좋지 못한 특징으로 생각할 수 있다[3].

$$F-ratio = \frac{variance\ of\ speaker\ means}{mean\ intraspeaker\ variance} \quad (7)$$

[표2]는 실험 결과를 나타낸다.

[표2] 실험 결과

Method	F-ratio	Recognition Rate	Improvement
No Normalization	22.3	89.2%	-
Formant-Based Frequency Warping	28.2	93.5%	4.3%

[그림4]와 [표2]의 결과에서 정규화 비율인 F-ratio를 더욱 향상시키기 위해서 워핑함수 형태 및 파라미터의 선택에 있어 다른 형태가 요구된다.

5. 결론 및 향후 연구과제

본 논문은 워핑함수를 정의하는데 화자 특유 특징인 포만트를 사용한 화자 정규화에 관한 연구이다. 여기에서는 포만트의 분포로부터 통계치를 계산하였고 이런 통계치는 워핑함수를 정의하는 점들의 좌표가 된다.

워핑함수는 기준화자의 주파수 축을 새로운 화자의 주파수 축으로의 매핑하는 함수이고, 본 논문에서 사용한 선형함수와 같은 워핑함수의 형태는 어떠한 제약조건이 없기 때문에 성능향상을 위해 바크나 멜과 같은 다른 형태로 바뀔 수도 있다.

본 논문의 시스템을 이용하여 실험한 결과, 인식률의 향상이 약 4.3%있었고, 정규화 척도인 F-ratio의 비율도 약 7정도 증가하였다. 만약 화자정규화가 잘된다면 새로운 화자를 더 정확히 인식할 것이고 이상적으로, 화자 독립 시스템은 실제로 기준 화자에 대해 완전하게 훈련된 화자 종속 시스템이 될 것이다.

참고문헌

[1] L.R.Rabiner, R.W.Schafer, *Digital Processing of Speech Signals,* Prentice Hall, Englewood cliffs, N.J., pp.378-385, 1978.

[2] Davis, S., Mermelstein, O., "Comparison of parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-28, no.4, pp.357-366, Aug. 1980.

[3] 이황수, "화자인식 기술," 제12회 음성통신 및 신호처리 워크샵 논문집, pp.42-46, 1995.

[4] Eide, E., Gish, H., "A Parametric Approach To Vocal Tract Length Normalization," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol.1, pp.346-348, May 1996.

[5] Zhan, P., Westphal, M., "Speaker Normalization Based On Frequency Warping," 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol.1, pp.1039-1042, May 1997.

[6] Gouvěa, E. B., *Acoustic-Feature-Based Frequency Warping for Speaker Normalization,* Ph.D. Dissertation, Carnegie Mellon University, pp.111-113, December 1998