음성 특징에 대한 시간 지연 효과 분석

Analysis of the Time Delayed Effect for Speech Feature

안 영 목*
(Youngmok Ahn*)

※이 연구는 정보통신부의 지원으로 이루어진 결과물 입니다.

요 약

본 논문에서는 음성 특징의 시간 지연 효과에 대해서 분석한다. 여기에서 시간 지연 효과란 과거의 음성 특징 벡터가 현재의 음성 특징 벡터에 미치는 영향을 의미한다. 본 논문에서는 선형 예측 계수를 바탕으로 한 켑스트럼을 사용하였으며, 켑스트럼의 시간 지연 효과는 음성 인식 시스템의 성능을 바탕으로 평가하였다.

실험에 사용한 음성 데이터는 남성 화자 50명이 발성한 22단어 이며, 50명의 화자 중에서 25명은 음성 인식기의 훈련용으로 사용하였으며 나머지 25명은 평가용으로 사용하였다. 실험의 결과, 특징 벡터에서 시간 지연 효과는 저차원으로 갈수록 그 영향이 커지고, 고차원에서는 시간 지연 효과가 적었다.

ABSTRACT

In this paper, we analyze the time delayed effect of speech feature. Here, the time delayed effect means that the current feature vector of speech is under the influence of the previous feature vectors. In this paper, we use a set of LPC driven cepstral coefficients and evaluate the time delayed effect of cepstrum with the performance of the speech recognition system.

For the experiments, we used the speech database consisting of 22 words which uttered by 50 male speakers. The speech database uttered by 25 male speakers was used for training, and the other set was used for testing. The experimental results show that the time delayed effect is large in the lower orders of feature vector but small in the higher orders.

I. 서 론

음성 생성 모델에서 여기 신호(excitation)는 무성음일 경우 성대에서의 난기류에 의해 발생되는 백색 잡음(white noise)으로, 유성음의 경우 성대의 진동으로 모델링할 수 있다[1]. 또한 유성음의 경우 준주기적인 특성과 성도 특성에 따른 고유한 공명 특성이 존재한다. 따라서 이러한 모델을 바탕으로 한 단구간 퓨리에 변환(short time Fourier transform)은 음성의 여기원(excitation source)에 의한 성도 특성 및 성대의 진동 특성을 콘볼루션(convolution)한 것으로 볼 수 있다. 켑스트럼은 단구간 스펙트럼에 대한 로그 스케일의 크기를 역퓨리에 변환한 것으로 다음 식(1)과 같이 정의된다.

$$C_n=\frac{1}{N}\sum_{k=0}^{N-1}\log|X(k)|\,e^{j2\pi kn/N}\qquad(0\le n\le N-1)\qquad(1)$$

그러므로 켑스트럼은 주파수 영역의 파라미터를 역퓨리에 변환하였으므로 시간 영역의 파라미터이다. 그리고 주파수 영역에서의 단구간 스펙트럼 중에서 천천히 변하는 부분은 켑스트럼의 저차원 특성으로 반영이 되고, 급격히 변하는 부분은 켑스트럼의 고차원 특성으로 반영이 된다[2]. 한편 인간의 청각 특성은 저주파 성분에 대한 주파수 분해능은 우수하되 시간 분해능이 나쁘며, 고주파 성분에 대한 시간 분해능은 우수하되 주파수 분해능이 나쁜 것으로 알려져 있다. 또한 인간의 인지는 입력의 절대적인 값보다는 상대적인 값에 의해서 처리된다. 이러한 인간의 청각 특성을 바탕으로 한 음성 특징으로서 라스타(RASTA-relative spectra)가 제안되었다[3]. 따라서 음성 특징을 추출함에 있어서, 모든 주파수 대역에 대한 동일한 분석 구간의 적용보다는 주파수 대역별로 서로 다

*한국전자통신연구소 음성언어연구실
Spoken Language Processing Section, ETRI
접수일자 : 1996년 11월 28일

른 분석 구간을 적용하는 것과 이웃하는 음성 특징 벡터들 사이의 연관성을 반영하는 것은 의미가 있을 것이다.

본 연구는 음성 특징의 주파수 대역에 따른 서로 다른 분석 구간의 적용 및 음성 특징 벡터들 사이의 연관성을 알아내기 위해서, 켑스트럼에서 과거의 신호가 현재의 신호에 미치는 영향을 분석하여 음성 특징 벡터의 각 차원마다 현재 신호에 반영할 과거 신호의 가중치 및 서로 다른 분석 구간을 찾는 것이 목적이다.

Ⅱ. 음성 특징의 시간 지연 효과

잡음이 있는 곳에서 그 환경에 맞도록 화자가 목소리를 변화시킨다는 롬바드 효과는 잡음 환경에서의 음성 인식 성능 개선에 많은 정보를 제공한다[4]. 자신의 발성이 잡음 환경에서도 잘 들리게 하기 위해서 목소리를 변화시키는 것은 인간의 청각 특성에 바탕을 둔 것이므로 변화된 음성의 포만트 구조 등을 분석함으로써 잡음 환경에서의 인간의 청각 특성 변화를 밝힐 수 있는 것이다. 따라서 이러한 결과를 음성 특징 추출 단계에 반영함으로써 잡음 환경에서 보다 성능이 향상된 음성 인식기를 개발할 수 있다. 한편 좋은 음성 특징을 얻기 위한 도구로써 인식 시스템 자체의 분류 성능을 사용하기도 한다. 예를 들면 은닉 마코프 모델(HMM)을 바탕으로 한 화자 인식에 있어서 최소 분류 오류(Minimum Classification Error) 알고리즘을 사용하여 음성 특징의 성능 향상을 꾀하였다[5]. 또한 차분 켑스트럼(Differenced Cepstrum)은 시간에 따른 스펙트럼의 변화량이 인간의 인지에 중요한 역할을 담당한다는 사실을 반영한 것이다[6].

이웃 프레임 사이의 켑스트럼 변화량이 음성 인식기의 성능 향상에 기여했다는 사실은 프레임간의 음성 특징들이 상호 연관이 있다는 것을 의미한다. 그리고 음성 특징 프레임 사이의 상호 연관성은 음성 특징의 시간 지연 특성으로 나타난다고 볼 수 있다. 본 논문에서는 음성 특징 벡터의 각 차원에서의 시간 지연 효과를 분석하여 이웃 프레임간의 상호 연관성을 밝히려고 한다.

다음의 설명은 실험에 적용할 음성 특징 추출 방법에 관한 것이다. 아래 수식 (2)는 t번째 프레임에서의 음성 특징 벡터 $\overline{V}(t)$는 M개의 차원으로 이루어진 것을 표시한다.

$$\overline{V}(t)=\{V_1(t), V_2(t), ..., V_M(t)\} \tag{2}$$

그리고 벡터 $\overline{V}(t)$는 현재 프레임 및 과거의 프레임에서 구해진 음성 특징 벡터들에 의해서 구해진다. 이것은 식 (3)에 나타냈다. 식 (3)에서 $V_j(t)$는 t번째 프레임에서의 j번째 차원의 값을 의미한다. 그리고 $W_{j,k}(t)$는 $V_j(t)$에 반영할 k번째 프레임에서의 j번째 차원에 대한 가중치이다.

$$V_j(t)=\sum_{k=t-N}^{k<t+1} W_{j,k}V_j(t) \tag{3}$$

단 식 (3)에서 $\sum_{k=t-N}^{k<t+1} W_{j,k}=1.0$이며 N은 반영할 과거 프레임의 수이다. 그리고 식(2)와 식(3)을 통해서 $\overline{V}(t)$는 가중치 $W_{j,k}$에 의한 현재 및 과거 음성 특징 벡터의 결합임을 알 수 있다. 본 실험에서는 가중치 $W_{j,k}$의 변화에 따른 성능을 바탕으로 해당 차원에서 시간 지연 효과를 측정한다.

Ⅲ. 실험 결과

1. 시스템 개요

기본 시스템의 성능 평가에 사용된 음성 데이터는 1100개의 고립단어로 이루어졌다. 이것은 50명의 화자가 22개의 부서명을 발성한 것이다. 본 실험에 사용한 음성 특징의 추출 과정은 표 1에 나타냈다.

표 1. 음성 특징의 추출 과정
Table 1. The procedure of speech feature extraction

처리 변수	적용치
Sampling Rate	16 kHz
A/D Precision	16 bit
Preemphasis	$H(z)=1-0.95z^{-1}$
분석 구간	20 msec(320 points)
분석 주기	10 msec(160 points)
Hamming Window 사용	
20차 Autocorrelation 분석후 LPC 계수 추출	
Band Pass Lifter 적용한 12차 LPC-Cepstrum 변환	
12차 Delta Cepstrum	

본 실험에서 25명의 화자에 대한 음성은 음성 인식 시스템의 훈련에 사용하였고, 나머지 25명의 화자에 대한 음성은 평가에 사용하였다. 본 실험에 사용한 음성 인식기는 벡터 양자화를 기반으로 한 것으로 각 단어에 대한 코드워드의 분포, 발생 순서, 지속 시간의 정보 등을 이용한 것으로 아래 수식 (4)로 표시된다.

$$S_w=\sum_{t=1}^{n} V(y_t, W)+i_w \qquad (w=1, 2, ..., N) \tag{4}$$

식 (4)에서 $y_1, y_2, ..., y_n$은 벡터 양자화를 위한 코드워드 인덱스이다. S_w는 단어 w에 대한 발생값을 뜻하며, N은 전체 인식 대상 단어이다. 여기에서 i_w는 S_w에 대한 초기값이며, $V(y_t, W)$는 $y_1, y_2, ..., y_n$에 대한 단어 w의 모델에서 발생되는 값이다[7].

2. 실험 결과 및 분석

본 실험은 음성 특징의 시간 지연 효과를 알기 위한 것이다. 이것은 가중치 변화에 따른 음성 인식기의 성능을 바탕으로 추정한다. 가중치는 켑스트럼의 차원 및 시간에 따라서 다르게 적용된다. 또한 가중치는 해당 차원 및

시간에서 여러 가지 값을 적용하여 그 결과를 살펴봄으로써 최적의 가중치를 찾게 된다.

본 논문에서는 두 종류의 실험이 수행되었다. 첫째, 음성 특징으로 켑스트럼만을 이용한 경우이다. 즉, 훈련 및 실험에 켑스트럼만을 사용한 것이다. 둘째, 음성 특징으로 켑스트럼 및 델타 켑스트럼을 사용한 경우이다. 첫번째 실험에서는 각각의 차원에 대한 최적의 가중치를 얻기 위해서 하나의 차원에만 가중치를 적용하고 나머지 차원에 대해서는 원래의 값을 유지하였다. 즉, 가중치를 적용하지 않는 차원은 식 (3)에서 $W_{i,k}$의 값이 $k=t$인 경우 1.0이 되고 나머지는 0.0이 된다. 또한 임의의 차원에서 이전 프레임에 대한 가중치 증가로 인식 성능이 나빠질 경우에는 해당 차원에서의 실험을 중지하였다.

이와 같은 조건을 켑스트럼에 적용했을때 표 2의 결과를 얻었다. 표 2의 결과에서 시간 지연의 특성은 저차원의 켑스트럼에서 강하며, 고차원에서는 존재하지 않는 사실을 보여준다. 다시 말하면 켑스트럼에 시간 지연 특성을 반영할 경우 6차 이상에서는 성능 향상이 없거나 성능 저하가 발생되었다. 또한 가중치는 저차원에서 크게 나타나고, 고차원으로 갈수록 작아지는 경향을 보였다. 그리고 평균적으로 $k=t$의 가중치는 0.9가 되고, $k=t-1$의 가중치는 0.1이 되었다. 이것은 현재의 음성 특징에 대한 반영 비율은 0.9, 과거의 음성 특징에 대한 반영 비율은 0.1로 했다는 것이다. 따라서 이러한 결과를 기존의 음성 특징에 적용하게 되면 저차원의 경우 그 분석 구간이 20 msec보다 커지고 6차 이상에서는 동일해진다고 볼 수 있다. 또한 저차원의 켑스트럼에서 성능 향상이 뚜렷하며 가중치도 크다. 이것은 이전 프레임의 음성 특징을 현재 프레임의 음성 특징에 보다 많이 반영한 결과이다. 시간 지연 효과란 이전 프레임의 음성 특징이 현재 프레임에 미치는 영향을 뜻하므로 시간 지연 효과는 각 차원에 따라서 다르게 존재함을 알 수 있다. 한편, N이 2인 경우의 실험에서 1차원 및 2 차원을 제외한 다른 차원에서는 특별한 성능 향상이 없었다. 위의 실험 결과는 음성 특징의 시간 지연 특성이 오랜 시간 지속되지 않음을 뜻한다.

표 2. 켑스트럼에 대한 가중치 변화 실험
Table 2. Experiment of weighting coefficients for cpestrum

켑스트럼 차원	$N=1$에 대한 $k=t$에서의 가중치 및 인식 결과(%) 켑스트럼에 대한 기준 성능:93.98%		
	0.9	0.8	0.7
1	94.70	95.07	94.89
2	94.34	94.16	
3	94.71	94.16	
4	94.71	94.16	
5	94.16		
6	93.98		
7	93.61		
11	93.98		
12	93.80		

그리고 보다 세밀한 시간 지연 특성을 밝히기 위해서는 그 분석 단위를 보다 세밀하게 해야 하며, 표 1의 분석 구간도 다양하게 적용해야 할 것이다.

표 3은 첫번째 실험에서 얻어진 각 차원에 대한 최적의 가중치를 바탕으로 한 켑스트럼 및 델타 켑스트럼을 음성 인식기에 적용한 결과이다. 여기에서 말하는 최적의 가중치란 기본 시스템보다 인식 성능이 향상된 차원에서의 최대 성능을 보이는 가중치를 뜻한다. 따라서 N=1에서 현재 음성 특징에 반영하는 이전 프레임의 음성 특징은 1차원인 경우 0.2, 2차원에서 5차원까지는 0.1, 6차원 이상은 0.0이 된다. 그러므로 6차원 이상은 시간 지연 특성이 반영되지 않은 원래의 음성 특징이 사용된다. 실험에서 시간 지연 특성을 켑스트럼에만 적용한 경우에 델타 켑스트럼은 시간 지연 특성이 반영되기 이전의 켑스트럼을 이용하여 구하였다. 이 실험에서 시간 지연 특성을 켑스트럼과 델타 켑스트럼에 동시에 적용한 결과가 켑스트럼에만 적용한 결과보다 우수함을 알 수 있다. 본 실험의 결과, 시간 지연의 효과는 델타 켑스트럼에서도 유지됨을 볼 수 있었다.

표 3. 켑스트럼 및 델타 켑스트럼에 대한 가중치 변화 실험
Table 3. Experiment of weighting coefficients for cepstrum and delta-cepstrum

시간 지연 특성(O:적용, X:미적용)		인식률(%)
켑스트럼	델타 켑스트럼	
X	X	95.26
O	X	95.99
O	O	96.35

Ⅳ. 결 론

본 논문에서는 음성 특징의 시간 지연 특성을 살펴보았다. 특히 켑스트럼의 각 차원별로 각기 다른 가중치를 적용하여 얻은 인식 성능을 통해서 과거의 음성 특징 벡터가 현재의 음성 특징 벡터에 미치는 영향을 분석하였다. 실험의 결과, 켑스트럼의 5차 이하의 영역에서는 시간 지연의 효과가 존재하였으나 6차 이상에서는 없었다. 이러한 결과를 기존의 켑스트럼 추출 방식에 적용할 경우, 켑스트럼의 저차원에서 이전 프레임과 현재 프레임 사이의 급격한 변화를 막아주는 효과가 있다. 또한 기존의 켑스트럼보다 성능이 향상되었다.

그리고 앞으로 보강해야 할 내용은 다음과 같다. 첫째, 음성 특징 벡터 안에서 각 차원끼리 미치는 영향을 분석하고 이것의 결과를 시간 지연 효과 분석에 적용시켜야 한다. 둘째, 보다 세밀한 가중치를 적용할 필요가 있다. 셋째, 본 실험에서 얻어진 시간 지연 특성이 연속어 음성 및 대화체 음성에서도 똑같은 가중치로 적용되는지 살펴보아야 한다.

참 고 문 헌

1. L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.
2. Alex Waibel, Kai-Fu Lee, Reading in SPEECH RECOGNITION, Morgan Kaufmann Publishers, San Mateo, California, pp. 58-59, 1990.
3. Hynek Hermansky, Nelson Morgan, "RASTA Processing of Speech," IEEE Trans. Speech and Audio Processing, vol.2, NO.4, pp. 578-589, October 1994.
4. W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow & M.A. Stokes, "Effects of Noise on Speech Production: Acoustic and Perceptual Analyses," JASA, 34, pp. 936-941, 1988.
5. Chi-shi Liu, "A General Framework of Feature Extraction: Application to Speaker Recognition," Proc. ICASSP, pp. 669-672, 1996.
6. L. R. Rabiner, J. G. Wilpon, F.K. Soong, "High Performance Connected Digit Recognition, Using Hidden Markov Models," Proc. ICASSP, pp. 119-122, 1988.
7. Lalit R. Bahl, Raimo Bakis, Peter V.de Souza and Robert L. Mercer, "Obtaining candidate words by polling in a large vocabulary speech recognition system," Proc. ICASSP, pp. 489-493, 1988.

▲안 영 목(Youngmok Ahn)

1991년 1월:홍익대학교 전자공학과 졸업(학사)

1991년 1월~현재:한국전자통신연구소 음성언어연구실 연구원

※주관심분야:음성인식, 화자인식, 음성신호처리