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

적은 훈련 데이타를 이용한 LSP 파라메터 기반의 화자종속 음성인식에 관한 연구

곽수주, 김형순
부산대학교 전자공학과

A Speaker Dependent Speech Recognition Method Using LSP Parameters for Small Training Data

Suju Kwag, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
{ sjkwag, kimhs }@hyowon.cc.pusan.ac.kr

요 약

통신 수단의 발달로 휴대단말기의 사용이 증가하고 있으며, 이와 함께 휴대단말기에서의 음성인식에 대한 수요도 증가하고 있다. 휴대단말기의 경우 저 전송율을 가지는 음성 부호화기를 사용하게 되며, 이러한 저전송율의 음성 부호화기에서 음성인식을 수행할 경우 인식 성능이 저하되는 현상을 보이게 된다. 본 논문에서는 이러한 문제를 해결하기 위하여 LSP 파라메터 기반의 거리척도에 관하여 비교 검토하였으며, 적은 훈련 데이터에서 사용 가능한 화자 종속 음성인식 방법으로 Dynamic Time Warping(DTW)과 변형된 Hidden Markov Model(HMM)에 관하여 검토하였다. QCELP 음성 부호화기에서 인식 어휘 당 2번의 훈련 데이터만을 이용한 화자종속 인식방법을 사용한 결과 95% 이상의 인식 성능을 얻을 수 있었다.

1. 서 론

음성 부호화 기술이 발전함에 따라 저 전송율에서도 좋은 음질을 가지는 다수의 알고리즘이 개발되었으며, 이의 표준화 작업에도 많은 진전이 있었다. 이러한 발전은 통신기기의 발달과 더불어 실생활의 통신 분야에 저전송 음성 부호화(low-bit rate speech coder) 알고리즘을 사용할 수 있게 했다. 휴대단말기능의 통신기기에서 사용하는 음성 부호화 알고리즘의 경우 그 전송율을 낮추기 위하여 음성을 여러가지 방법으로 처리하게 되고, 그 과정에서 음성 정보의 일부분은 없어지거나 왜곡되게 된다. 그 결과, 음성 부호화기를 통과한 음성으로 음성인식을 수행하게 되면 인식 성능에 상당한 저하를 가져오게 된다[1].

음성 부호화기를 통과한 후 복원된 음성 신호로부터 인식용 파라메터를 별도로 추출하여 인식을 수행하는 방법은 음성 부호화기에 의한 왜곡이 인식성능의 저하로 직접 나타날 뿐만 아니라, 음성 신호의 복원이 필요하므로 그 과정에서 많은 계산량을 필요로 한다. 대부분의 저전송 음성 부호화기가 선형예측 부호화(LPC)에 기반한 Line Spectrum Pair (LSP)파라메터를 이용하여 음성의 스펙트럼 정보를 표현하기 때문에 본 논문에서는 LSP 파라메터를 기반으로한 음성인식 방법을 검토하였다. LSP를 음성인식용 파라메터로 사용하는 경우 음성 신호의 복원이 필요 없을 뿐만 아니라 적절한 거리척도를 사용하거나 다른 파라메터로의 전환[2,3]을 통하여 인식 성능을 향상 시킬 수 있다.

휴대단말기에서의 음성인식의 일차적인 응용분야는 음성 다이얼링(voice dialing)이다. 이 경우, 인식 대상 어휘를 제한하기 어렵고 그러한 이유로 많은 훈련 데이터를 얻는 것도 용이하지 않다. 연속 확률 분포를 가지는 HMM(Hidden Markov Model)방법은 일반적으로 높은 인식율을 나타내지만, 파라메터 추정을 위해 매우 많은 훈련용 음성을 필요로 하기 때문에 본 논문에서 다루는 화자종속 음성인식에는 적용되기 어렵다. 이를 해결하

기 위하여, 본 논문에서는 적은 데이터로도 상대적으로 높은 인식율을 얻을 수 있다고 알려져 있는 Dynamic Time Warping(DTW)기법과 적은 데이터 환경에 사용할 수 있도록 HMM을 변형한 화자종속 음성인식 방법에 대하여 검토하고 그 성능을 비교하였다.

서론에 이어 2장에서는 저 전송율의 음성 부호화기에서 사용하는 파라메타인 LSP에 관하여 설명하고, 3장에서는 적은 훈련 데이터에서 사용 가능한 음성인식 방법에 관하여 기술한다. 그리고, 4장에서 더 나은 인식 성능을 얻기 위한 LSP 거리척도에 관하여 언급하고, 5장에서 실험 결과를 다루며, 6장에서 결론을 맺도록 한다.

2. LSP 파라메터

LSP 파라메터는 Itakura에 의해 선형 예측 계수(LPC)의 다른 표현 형태로서 제안되었다[4]. LSP를 이용한 음성 부호화의 경우 다른 LPC 파라메터들에 비해 더 나은 양자화성능을 나타내며, 복원된 음질의 저하없이 기존에 사용되던 LPC 파라메터들에 비해 25~30% 정도의 전송율 감소를 얻을 수 있다고 알려져 있다. 이러한 이유로 현재 사용되는 많은 음성 부호화기에서 음성 스펙트럼을 나타내기 위하여 LSP를 사용한다.

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p} \tag{1}$$

식(1)에서 $\{a_k\}$는 LPC 파라메터이며, p는 LPC 차수이다. 식(1)을 이용하여 식(2)와 식(3)을 정의할 수 있으며, 이 두식의 근을 LSP라 한다.

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{2}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{3}$$

3. 인식 방법

3.1 Dynamic Time Warping(DTW)

DTW는 동적 프로그래밍(Dynamic Programming)방식을 통해 시간축에서 발생하는 차이를 보상하면서 두 음성 간의 유사도를 측정하는 기법이다. 화자종속 고립단어 인식에서 높은 성능을 얻을 수 있다고 알려져 있는 방법으로 특히 적은 훈련 데이터에서도 비교적 높은 인식율을 얻을 수 있는 장점이 있다.

본 논문에서는 인식 어휘 당 하나의 비교 패턴을 만드는 방법을 사용하였으며, 전역 경로 제한(global path constraint)은 비교 패턴의 1/2 ~ 2 배까지로 하였고, 국부 경로 제한(local path constraint)은 Itakura가 제안한 방식을 채택하였다[9].

3.2 변형된 Hidden Markov Model (HMM)

일반적인 HMM의 경우, 추정해야 할 파라메터가 많으므로 적은 훈련 데이터로 신뢰할 만한 추정치를 얻기 힘들다. 그러므로, 본 논문에서는 변형된 HMM을 사용하였다. 변형된 HMM은 VQ에 기반한 거리 측정 방법과 HMM을 결합한 방법이다[5].

$$\alpha_{t+1}(j) = \left[\sum_i \alpha_t(i) a_{ij}\right] b_j(\mathbf{o}_{t+1}) \tag{4}$$

식(4)는 일반적인 HMM에서 확률값을 구하는 수식이다. 일반적인 HMM의 경우 $b_j(\mathbf{o}_t)$의 확률값을 구하기 위하여 Gaussian 모델을 이용한 방법을 사용하는데 이는 많은 수의 파라메터를 추정해야 한다. 변형된 HMM의 경우 이 확률값 대신에 상태(state)에 따른 VQ 코드북을 구성하여 이 코드워드와의 거리를 사용하게 된다.

$$b_j(o_t) = \exp\left[\max_k \left[-d(\mathbf{x}, \mathbf{c}_k^j)\right]\right] \tag{5}$$

여기서 $d(\mathbf{x}, \mathbf{c}_k^j)$는 입력 패턴과 j번째 상태(state)의 코드북에서 k번째 코드워드와의 거리를 의미한다. 이 방법은 상태당 코드북만 구성하면 되므로, 기존의 HMM에 비해 적은 훈련 데이타만으로도 인식기를 구성할 수 있게 한다. 인식 실험에 사용된 변형된 HMM 방법은 8개의 상태(state)를 가지는 whole word 모델을 사용하였고, 인식 대상 어휘 당 단지 2회의 훈련 데이터를 사용하였다. 훈련 데이터가 적은 관계로 상태 당 코드워드(codeword)는 하나만 사용하였으며, 훈련을 위한 과정에서는 segmental K-means 알고리즘을 사용하였다. 본 논문에서 사용한 변형된 HMM의 경우 인식에 필요한 계산량은 DTW와 유사한 정도이다.

4. LSP 파라메터에 대한 거리척도 및 변형된 LSP 파라메터

위에서 제안한 두 가지 인식 방법 모두 입력 패턴과 기준 패턴과의 거리에 의하여 인식을 수행하는 방법이

므로 거리 척도(distance measure)의 선정이 인식 성능에 영향을 주게 된다. LSP를 이용한 거리 척도에 관하여서는 음성 부호화의 관점에서 여러 연구가 진행되었다[6,7,8]. 이들은 VQ에서의 spectral distorsion을 줄이기 위하여 제안된 방법들로서, 음성인식에서의 인식율 향상에도 도움을 줄 수 있다. 본 논문에서는 이러한 LSP를 이용한 거리척도 중 그 계산량이 비교적 작은 IHMW(Inverse Harmonic Mean Weighting)[8]에 대하여 살펴보고 또 하나의 파라메터로 PCEP(pseudo-cepstrum)에 대하여 살펴본다. 이는 높은 인식율을 얻기위하여 LSP를 인식에 가장 널리 사용되는 켑스트럼 파라메터로 바꾸는 방법이다[2,3].

4.1 유클리디언 거리척도

일반적으로 사용되는 거리 척도로써 식(6)과 같이 정의 된다.

$$D_{EUC}(\mathbf{x},\mathbf{c}_i) = (\mathbf{x}-\mathbf{c}_i)'(\mathbf{x}-\mathbf{c}_i) \tag{6}$$

여기서, $\mathbf{x}$는 입력 패턴을 $\mathbf{c}_i$는 기준 패턴 혹은 코드워드를 나타낸다.

4.2 Inverse Harmonic Mean Weighting (IHMW)[7]

IHMW은 가중(weighted) 유클리디언 거리 척도의 일종으로 식(7)과 같이 정의된다.

$$D_{IHMW}(\mathbf{x},\mathbf{c}_i) = (\mathbf{x}-\mathbf{c}_i)'\mathbf{W}(\mathbf{x}-\mathbf{c}_i) \tag{7}$$

이때, 사용되는 가중치는 식(8)과 같다.

$$w_i = s_i^2\left(\frac{1}{\omega_i-\omega_{i-1}}+\frac{1}{\omega_{i+1}-\omega_i}\right),\ 1\le i\le p \tag{8}$$

여기서 ω_i는 i번째 LSP 파라메터이며, p는 LSP 차수이다. 차수가 10인 경우 s_i는 다음과 같이 정의된다.

$$s_i=\begin{cases}1, & 1\le i\le 8\\ 0.8, & i=9\\ 0.4, & i=10\end{cases} \tag{9}$$

IHMW은 음성신호의 스펙트럼에서 피크(peak)가 존재하는 곳은 LSP 파라메터가 서로 가까이 존재하는 성질을 이용하여, 피크 부분에 큰 가중치를 주는 방법이다. 이는 음성인식에 있어 중요한 역할을 하는 음성의 포만트를 강조하는 효과가 있으므로 IHMW을 이용할 경우 인식성능의 향상을 기대할 수 있다.

4.3 Pseudo-Cepstrum (PCEP)[2,3]

PCEP은 LSP 파라메터와 켑스트럼 파라메터와의 관계를 근사화함으로써, LSP 파라메터를 LPC로 바꾸는 과정이 없이 직접 켑스트럼 파라메터로 바꾼 것이다[2].

$$c_n = \frac{1}{n}\sum_{1}^{p}\cos n\omega_i \tag{10}$$

PCEP은 식(10)과 같이 정의되는데, 여기서 ω_i는 i번째 LSP 파라메터이며, p는 LSP 차수이다.

4.4 Sine Liftered PCEP

켑스트럼 파라메터의 경우, 가중치를 줌으로써 인식 성능을 향상시킬 수 있다. 본 논문에서는 식(11)과 같은 Sine Lifter를 사용하여 PCEP에 가중치를 줌으로써 인식 성능을 향상시킬 수 있었다.

$$w_n = \left[1+\frac{Q}{2}\sin\left(\frac{\pi n}{Q}\right)\right],\quad 1\le n\le Q \tag{11}$$

여기서 Q는 켑스트럼 파라메터의 차수이다.

5. 실험 결과 및 검토

본 논문에서는 제안된 인식 방법들의 성능을 측정하기 위하여 저 전송 음성부호화기의 하나인 QCELP 음성 부호화기(Qualcomm Code-Excited Linear Predictive Coder)를 이용하였다. QCELP는 매 20ms마다 10차의 LSP를 추출하며 이를 양자화(Quantized LSP, QLSP)하여 음성의 스펙트럼을 표현한다.

실험을 위해 성인 남성 8명과 성인 여성 1명이 각각 28단어를 3회씩 발음한 음성 데이터를 구축하여 이용하였다. 이 음성 데이터는 일반적인 사무실 환경에서 PC를 통하여 11.025KHz 및 16bit로 녹음하였으며, 8KHz로 다운샘플(down sample)하여 사용하였다.

먼저 구현된 인식기의 기본적인 성능을 평가하기 위하여 수집된 데이터로부터 일반적으로 음성인식에 널리 사용하는 LPC 켑스트럼계수를 구하여 인식 실험을 수행하였다. 이때 QCELP 부호화기의 규격과 동일한 환경을 맞추기 위하여 매 20ms마다 10차의 LPC 켑스트럼을 추출하였다. 그 결과는 표 1과 같다.

표 1. LPCCEP의 인식율(%)

인식방법	DTW	변형된 HMM
LPCCEP	95.4	96.0

다음으로는 저전송 음성 부호화기에서 사용하는 파라메터인 LSP의 특성과 저전송 음성 부호화기를 통과한 후 나타나는 인식 성능의 저하를 살펴보기 위하여, 음성 데이터로부터 직접 LSP 파라메터를 추출한 일반적인 LSP와 QCELP를 거쳐 부호화기의 양자화 오차를 포함한 QLSP(Quantized LSP)를 이용하여 인식실험을 수행하였다. 이 실험에서는 일반적인 유클리디언 거리척도를 이용하였으며, 결과는 표 2와 같다.

표 2. LSP 파라메터의 인식율과 코디의 영향(%)

인식방법	DTW	변형된 HMM
LSP	94.6	90.1
QLSP	91.7	86.1

표 2에서 보는 바와 같이 음성 부호화기를 거치지 않은 LSP 파라메터의 경우에 비해, 부호화기를 통과하여 양자화 오차를 포함한 QLSP의 경우 인식 성능에 상당히 많은 저하가 있음을 볼 수 있다.

이러한 인식 성능의 저하를 보완하기 위하여 본 논문에서는 QLSP에 향상된 LSP 기반의 거리척도 및 변형된 LSP 파라메터를 적용하여 인식실험을 수행하였다. 그 결과는 표 3과 같다.

표 3. 제안된 거리 척도에 의한 인식율(%)

인식방법	DTW	변형된 HMM
QLSP	91.7	86.1
IHMW	92.5	92.1
PCEP	92.5	91.3
sine liftered PCEP	92.9	95.2

IHMW과 PCEP 모두 인식 성능의 향상에 효과가 있었으며, PCEP의 인식성능을 향상시키기 위하여 PCEP에 lifter를 사용한 결과가 가장 우수했다. 인식 방법면에 있어서 DTW의 경우는 거리 척도에 다소 덜 민감한 특성을 나타냈으나, 변형된 HMM을 사용한 경우는 거리 척도에 상당히 민감한 특성을 나타냈다.

특히 변형된 HMM의 경우, 유클리디언 거리척도를 적용한 QLSP의 경우 상당한 인식성능 저하를 나타냈으나, PCEP에 lifter를 적용한 향상된 거리 척도를 사용했을 경우, 부호화기를 통과하지 않았을 때의 인식성능에 가까운 인식율을 나타내어 많은 인식성능의 향상을 보였다.

6. 결 론

본 논문에서는 휴대단말기 등의 저전송 음성부호화기에서 사용할 수 있는 화자 종속 음성 인식 방법에 관하여 연구하였다. 이러한 분야의 경우, 저 전송 음성부호화기를 사용하므로 인식 성능의 저하를 가져올 수 있고, 많은 훈련데이타를 얻는 것도 용이하지 않다. 이를 해결하기 위하여, 본 논문에서는 적은 데이타에서 사용가능한 LSP 파라메터 기반의 화자종속 인식 방법을 연구하였다. 이의 검증을 위하여 QCELP를 사용하여 인식 실험을 수행하였으며, 그 결과 파라메터로는 LSP를 변형한 PCEP에 sine lifter를 적용한 것이, 인식 방법으로는 변형된 HMM을 사용하는 것이 가장 우수한 성능을 나타내었다. 이 경우 2번의 훈련 데이터만을 이용하여 95%이상의 비교적 높은 인식성능을 얻을 수 있었다.

앞으로 보다 나은 인식 성능을 얻기 위한 LSP 파라메터기반의 거리척도에 대한 지속적인 연구와 더불어 잡음 환경에서의 인식 성능 개선에 관한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in Proc. ICSLP, 1996, pp.2344 - 2347.

[2] H. C. Hong, H. K. Kim, H. S. Lee and R. M. Gray, "Speech recognition method using quantised LSP parameters in CELP-type coders," Electro. Lett., 1998, 34,(2), pp.156-157.

[3] H. K. Kim, K. C. Kim and H. S. Lee, "Enhanced distance measure for LSP-based speech recognition," Electro. Lett., 1993, 29, (16), pp. 1463 - 1465.

[4] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," J. Acoust. Soc. Am., (abstract) vol. 57, pp. 535, 1975.

[5] S. Nakagawa and H. Suzuki, "A new speech recognition method based on VQ-distortion measure and HMM," in Proc. ICASSP, 1993, pp. 676-679.

[6] H. L. Vu and L. Lois, "A new general distance measure for quantization of LSF and their transformed coefficients," in Proc. ICASSP, 1998, Vol.1, pp. 45-48

[7] R. Laroia, N. Phamdo and N. Farvardin, "Robust and Efficient Quantization of Speech LSP Parameters Using Structured Vector Quantizers," in Proc ICASSP, 1991, pp. 641 - 644.

[8] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," in Proc. ICASSP, 1991, pp. 661 - 664.

[9] L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.