G.723기반의 음성인식을 위한 변별적인 음성 특징 벡터 선정

이규환, 정민화
서강대학교 컴퓨터학과

Discriminative Feature Selection for G.723-based Speech Recognition

Kyuwhan Lee, Minhwa Chung
Department of Computer Science, Sogang Univercity

요 약

정보 통신 분야의 발달로 사람들의 전화 사용이 늘어나고 또한 전화기에 여러 가지 멀티미디어 기능들이 추가되면서 음성 인식의 필요성이 점차 증가하고 있다. 그러나 현재의 기술로는 음성 인식의 성능이 사람들의 기대치를 만족시키지 못하고 있다. 본 연구에서는 G.723을 이용한 네트워크 상에서 음성 인식 시간을 줄이고 같은 차수에서 더 좋은 음성 인식 성능을 얻을 수 있는 방법에 대해 연구하였다. 일반적인 보코더는 채널을 통과시킬 때 왜곡을 최소화 하기 위해 양자화할 때 안정적이라고 알려져 있는 LSP 파라메터를 양자화하여 전송한다. 전송된 양자화된 LSP 파라메터는 복호화기를 통과하게 되는데 본 연구에서는 양자화된 LSP 파라메터를 음성인식에 직접 이용하여 음성 합성한 후 음성 특징 파라메터를 추출하는 시간을 줄일 수 있고 음성 합성 시 왜곡을 미연에 방지할 수 있다. 본 연구에서는 변별적인 기준에 의해 특징 벡터 요소들을 순서화를 이용하여 음성 특징 벡터의 차수를 동적으로 조절할 수 있는 방법을 G.723에 적용시켜 보았다. 순서화 된 음성 특징 요소들 중에서 인식 목적에 적절한 차수를 선정하면 차수를 줄이면서도 음성인식 성능은 유지 또는 향상시킬 수 있음을 확인하였다. 특히 네트워크 통신망에서도 음성인식 성능을 향상시킬 수 있음을 확인하였고, 기존의 합성음에서 음성인식을 하는 방법보다 시간도 크게 단축할 수 있었다.

1. 개 요

정보 통신의 발달로 전화기의 사용이 늘어나고 또한 전화기에 여러 가지 멀티미디어 기능들이 추가 되면서 음성인식은 새로운 장애물을 넘어야만 하는 문제에 봉착하였다. 전화를 이용한 음성인식은 입력되는 음성이 일반 마이크와는 달리 전화기를 통하기 때문에 왜곡이 많고 인식률도 현저하게 감소하고, 전화선의 기본적인 대역폭(band width) 때문에 음성 정보가 상당부분 왜곡/손실이 되기 때문이다. 그러나 이러한 장벽은 정보통신의 발달로 인하여 음성인식의 필요성이 높아짐에 따라 반드시 넘어야 할 장벽으로 여겨지고 있다. 현재 우리의 음성인식 기술은 이제 제한된 고립단어 인식 수준을 뛰어넘어 더욱 편리한 형태로 우리 생활에 다가오고 있다. 선진 외국에서는 연속음성을 빠른 속도로 받아쓰는 dictation 프로그램에서부터, 전화를 통한 정보검색, 금융정보 이용 및 거래 등 그 응용범위가 날로 확산되고 있으며, 국내에서도 음성 다이얼링 기술을 이용한 휴대전화기가 등장하는 등 사람들의 관심이 고조되고 있다. 그러나 그 인식 성능이 사람들의 기대치를 만족시키지는 못하고 있다.

본 연구에서는 음성특징 벡터로 LSP 파라메터를 사용한다. 지금까지 음성인식에 LSP 를 이용하지 않은 것은 아니지만, 다른 특징 파라메터, LPC(Linear Predictive Coefficients), MFCC(Mel-frequency Cepstral Coefficients), 그리고 요즘 많이 사용하는 PLP(Perceptual Linear Prediction) 등에 비해서는 연구가 부족한 상황이다. 본 연구에서는 LSP 에서 PCEP(Pseudo-Cepstrum)을 구하여 인식에 사용하고 객관적인 기준에 의해서 특징 벡터를 순서화 한다. 순서화에 사용되는 기준은 음성인식에 기여도를 거리로 표현한다. CHMM(Continuous Hidden Markov Model)을 이용하는 음성인식기에서 음성의 한 프레임은 HMM 의 어느 한 state 에 정렬(align)된다. 이때 잘못 인식된 PLU 에 정렬된 프레임들과 그 때의 state 간에 서로 구성 요소별로 평균거리(Average Distance)를 구한다. 잘못 인식된 " incorrect" PLU 에 정열된 음성 프레임 요소들 중에 평균거리가 제일 큰 값을 가지는 요소는 인식하는데 기여도가 더 높다고 볼 수 있다. 즉, 잘못 인식된 PLU 에 대한 평균 거리이므로 값이 클수록 더 올바른 인식쪽에 가깝다는 것이다. 평균거리별로 음성특징 벡터의 각 요소를 값이 큰 순서대로 나열한 다음 원하는 차수를 정해 그 만큼만 인식에 이용한다면 인식 성능에 기여도가 큰 요소들로만 인식을 수행하므로 인식률은 당연히 증가할 것이다. 차수를 조절할 수 있다면 네트워크 통신상에서 교환기, 혹은 단말기 파워가 가능한 적절한 차수로 음성인식을 할 수 있을 것으로 기대한다.

2. 연구 배경

2.1 CHMM 기반의 음성 인식 시스템

본 연구에서는 CHMM 을 사용하였고, 사용한 HMM 모델은 3 state 인 단순 left-to-right 모델을 사용하였다. 처음과 끝은 dummy state 로 되어있다. 학습과정에서는 음성 특징 벡터의 연속 확률 밀도 함수에 대한 평균과 분산으로 이루어진 각 state 에 대한 관측확률과 state 간 천이확률을 Baum-Welch re-estimation에 의해 구한다. 인식과정에서는 Viterbi 탐색으로 구해진 확률값을 이용하여 입력된 음성과 가장 가까운 모델을 인식단어로 결정하였다.

본 연구의 관심사인 관측확률(Obsevation Probability) 구하는 과정을 자세히 살펴보면 다음과 같다. 주어진 상태 s 와 시간 t 에서 관측 심볼 ft 의 likelihood(ls(ft)) 를 대각 공분산(Diagonal covariance) 과 M 개의 정규분포(Normal distribution)로 표현하면 다음과 같다.

$$l_s(f_t) = \sum_{m=1}^{M} \varepsilon_{s,m} N(f_t, \mu_{s,m}, u_{s,m}) \qquad \text{식(2.1)}$$

$$N(f_t, \mu_{s,m}, u_{s,m}) = \frac{1}{\sqrt{(2\pi)^d |u_{s,m}|}} e^{-\frac{1}{2}(f_t-\mu_{s,m})' u_{s,m}^{-1} (f_t-\mu_{s,m})} \qquad \text{식(2.2)}$$

이때 d 는 벡터의 차수이고, εs,m은 Gaussian mixture 의 가중치, μs.m은 평균 벡터, 그리고 us.m 은 공분산의 대각요소의 벡터이다.

식 (2.2) 의 양변에 logarithm 을 취하면

$$\ln[N(f_t, \mu_{s,m}, u_{s,m})] = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\sum_{n=1}^{d}\ln(u_{s,m}(n)) - \frac{1}{2}\sum_{n=1}^{d}\frac{[f_t(n)-\mu_{s,m}(n)]^2}{u_{s,m}(n)} \qquad \text{식(2.3)}$$

그러나 본 연구의 특성상 Gaussian mixture 의 합으로 ls(ft)를 계산하는 대신에 이것을 다음과 같이 근사화 할 수 있다.

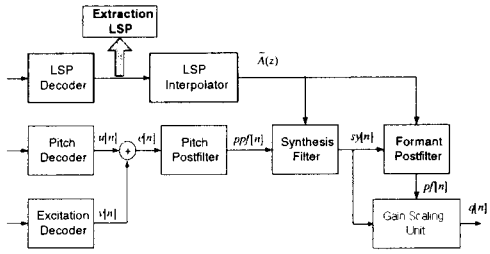
$$l_s(f_t) = \varepsilon_{s,m'} N(f_t, \mu_{s,m'}, u_{s,m'}), m' = \underset{m=1}{\overset{M}{\operatorname{argmax}}} \varepsilon_{s,m} N(f_t, \mu_{s,m} u_{s,m}) \qquad \text{식(2.4)}$$

본 연구에서 사용한 통계적 문법은 PLU 단위로 모델을 구성하여 phone bigram 을 사용하였다.

2.2. G.723[1]

ITU-T[1]의 표준의 하나인 G.723 은 현재 우리나라에서 한국통신의 소리샘 서비스의 표준이고 인터넷 전화의 표준으로 사용되고 있다.

본 연구에서 얻고자 하는 QLSP 신호는 Decoder [그림 1] 에서 큰 화살표 부분에서 QLSP 를 추출하고 PCEP[5] (Pseudo Cepstrum)을 구하여 음성인식에 사용하게 된다.



[그림 1] Decoder 의 블록 다이어 그램

2.3 변별적인 음성특징벡터 방법 [2]

Viterbi 알고리즘은 음성의 프레임 열을 ML(Maximum Likelihood)를 이용하여 확률값이 가장 높은 PLU 열로 바꾸어 준다.

$$f_1 f_2 f_3 \ldots f_T \xrightarrow{Viterbi} G\ AA\ U \ldots EY \qquad \text{식(3.1)}$$

역추적(backtracking)을 하면, 각각의 PLU 에서 최적의 state 열을 찾아낼 수 있다.

$\theta(t) = \theta(t, w(t))$ 라고 하자. 여기서 $w(t)$ 는 해당 PLU 의 t 번째 프레임 f_t 에 정열되는 state 이다. 식 (3.1) 프레임 열의 log-likelihood $\ln(L)$ 는 아래와 같다.

$$\ln(L) = \sum_{t=1}^{T} \ln[l_{\theta(t)}(f_t)] \qquad \text{식(3.2)}$$

식 (3.2) 에 2 장에서 구한 식 (2.2) 과 식 (2.3) 을 대입하면,

$$\ln(\ell) = -\frac{Td}{2}\ln(2\pi) + \sum_{t=1}^{T}\varepsilon_{\theta(t),m} - \frac{1}{2}\sum_{t=1}^{T}\sum_{n=1}^{d}\ln(u_{\theta(t),m}(n)) - \frac{1}{2}\sum_{n=1}^{d}\sum_{t=1}^{T}\frac{[f_t(n)-\mu_{\theta(t),m}(n)]^2}{u_{\theta(t),m}(n)} \qquad \text{식(3.3)}$$

우리는 위의 식 중에서 마지막 항에 초점을 맞춘다. 마지막 항은 평균거리(average distance), D(n) 이다. 식을 다시 쓰면 아래와 같다.

$$D(n) = \frac{1}{T}\sum_{t=1}^{T}\frac{[f_t(n)-\mu_{\theta(t),m'}(n)]^2}{u_{\theta(t),m'}(n)} = E[\frac{[f_t(n)-\mu_{\theta(t),m'}(n)]^2}{u_{\theta(t),m'}(n)}] \qquad \text{식(3.4)}$$

E[ ] 항은 산술적인 평균을 의미한다. 학습하는 동안에 supervised Viterbi 과정은 항상 정확한 PLU 열을 찾아낸다. 모델 파라메터인 $\mu_{\theta(t),m'}(n)$ 와 $u_{\theta(t),m'}(n)$ 는 ML 과정을 통해 구해진다. 그러므로, 위의 정의에 의하면

$$D(n) = 1.0\ ,\quad n = 1,2,\ldots,d$$

이와 같은 결과는 다음과 같은 특징 벡터 순서화의 방법적

접근이 가능하다.

모든 PLU HMMs 을 supervised maximum likelihood 방법으로 학습을 한 후, 훈련 데이터에서 인식 실험을 수행한다. DP는 참조 PLU 열과 인식한 결과를 정열(align)한다. 모든 대체(substitution)나 삽입(insertion) 에러를 찾기 위해 잘못 인식된 incorrect PLU 모델($\tilde{w}(t)$)에 정열된 프레임들($\tilde{f}_t$)을 찾아낼 수 있다. $\tilde{\theta}(t) = \tilde{\theta}(t, \tilde{w}(t))$ 를 정열된 프레임의 각 모델($\tilde{w}(t)$) state 라 하고, $\tilde{m}$ 을 state $\tilde{\theta}(t)$ 의 mixture 요소라 하자.

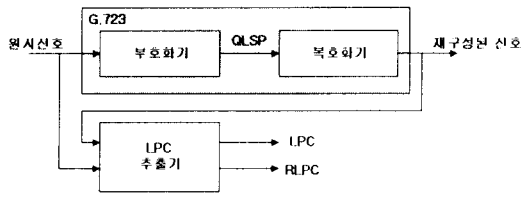
$$\tilde{m} = \arg\max_{m} \varepsilon_{\tilde{\theta}(t),m} N(\tilde{f}_t, \mu_{\tilde{\theta}(t),m}, u_{\tilde{\theta}(t),m}) \qquad \text{식 (3.5)}$$

그러면, 잘못 인식된 incorrect PLU 모델 $\tilde{\theta}(t)$ 에 정열된 모든 학습 데이터 $\tilde{f}_t$ 에 대하여 평균 거리를 계산할 수 있다.

$$\tilde{D}(n) = E[\frac{[\tilde{f}_t(n) - \mu_{\tilde{\theta}(t),m}(n)]^2}{u_{\tilde{\theta}(t),m}(n)}], \quad n = 1,2,...d \qquad \text{식 (3.6)}$$

계산된 $\tilde{D}(n)$ 중에서 가장 평균 거리가 큰 요소들 즉, $\tilde{D}(n)$ 의 값이 큰 요소들은 음성 프레임 $\tilde{f}_t$ 과 " incorrect" PLU 모델 간에 평균 거리가 크다는 의미이다. 반대로 생각해 보면 $\tilde{D}(n)$,평균거리가 큰 요소들은 올바른 인식에 더 공헌도가 크다는 것을 의미하기도 한다.

3. 실험 및 결과



3.1 실험 방법

위의 그림과 같이 3 종류의 음성특징벡터 집합을 선정하여 실험을 하였다. 실험방법은 2.3 에서 나온 변별적인 음성특징벡터 선정 방법을 적용한 결과($\tilde{D}(n)$)와 적용하지 않은 결과를 서로 비교하였고, 본 연구와 비슷한 연구논문과도 인식결과를 비교하여 보았다.

3.2 인식 결과

기존의 연구 결과[3]는 아래와 같은 인식률을 나타내었다.

Coder의 종류	Bit Rate(kbs)	Feature 종류	인식률 (26 차)
ADPCM G.723	40	MFCC	51.74
ADPCM G.723	24	MFCC	48.29
GSM	13	MFCC	49.58
CELP-1016	4.8	MFCC	45.23

본 연구의 연구 결과는 음성 인식 feature 종류도 LSP 에서 뽑은 PCEP(Pseudo Cepstrum)[4] 을 사용하였고 차수 26 차보다 적은 11 차와 22 차를 이용하였기 때문에 직접적인 비교는 할 수 없지만 좋은 인식결과를 얻을 수 있었다.

특징 종류	LPC		QLSP		RLPC	
재구성	NO	YES	NO	YES	NO	YES
FS(11차)	44.50	47.5	45.56	49.82	40.73	44.95
DFS(22차)	53.72	56.2	55.69	54.84	52.10	55.68

4. 결론 및 향후과제

본 연구는 G.723 에서 사용되고 있는 파라메터인 LSP 11차를 직접 인식에 이용하여 계산적으로도 빠르게 구할 수 있고 코덱 자체의 차수를 그대로 이용함으로 추가의 음성인식벡터 추출과정이 없이 인식실험이 성공적임을 보여주고 있다.

또한 적은 차수에서도 인식률을 높이기 위하여 변별적인 음성인식 추출방법[2]을 G.723 에 적용시켜 보았는데 11 차에서는 성공적이라 할 수 있지만 22 차에서는 성공적이지 못하였는데 이는 향후 과제로 남기기로 하겠다.

참고문헌

[1] " ITU-T RECOMMENDATION G.72 3" , Draft of 1995-October-17

[2] E.L. Bocchieri and J.G. Wilpon, "Discriminative Feature Selection for Speech Recognition", *Computer Speech and Language*, vol 7, pp229-246, 1993

[3] B.T. Lilly and K.K. Paliwal, "Effect of Speech Coders on Speech Recognition Performance", *ICSLP* 96

[4] K.K. Paliwal, "On the Use of Line Spectral Frequency Parameters for Speech Recognition", *Digital Signal Processing* 2, p80-87, 1992