유성음의 정보를 이용한 화자식별에 관한 연구

오창환, 박대성, 최홍섭
대진대학교 전자공학과

On the speaker identification using the informations contained in the voiced intervals

Chang-Hwan Oh, Dae-Sung Park, Hong-Sub Choi
Dept. of Electronics Engineering, Daejin University
E-mail: mirage@road.daejin.ac.kr

요 약

GMM을 기반으로 하는 화자식별 시스템은 입력음성의 길이의 장단에 의해서 인식률에 차이가 생긴다. 이는 가우시안 모델의 파라미터를 추정할 때, 많은 데이터를 사용할수록 추정이 정확해지기 때문이다. 따라서 화자식별에 사용하는 입력데이터는 화자가 발성한 모든 음성신호에서 잡음구간만을 제거한 유,무성음을 이용하게 된다. 그러나 이 경우 데이터의 양이 많아져서 실시간처리에 어려움이 있겠다. 본 논문에서는 전체 음성구간을 이용하는 대신 유성음 구간만을 추출하여 이 구간의 켑스트럼과 피치값들을 특징파라미터로 이용하여 화자식별에 이용하였다. 특히 피치성분은 일반적으로 통신채널과 핸드셋의 영향에 상대적으로 강한 장점이 있다. 실험을 위하여 20대의 남성 및 여성화자 40명으로부터 얻은 음성데이터에서 유성음구간을 추출하여 GMM을 이용한 문장독립 화자식별 실험을 하였으며, 실험결과 스펙트럼정보와 함께 피치정보가 화자식별에 유용하게 사용될 수 있음을 알 수 있었다.

1. 서론

일반적으로 음성신호에서 추출한 피치와 에너지의 궤적에는 그 음성을 발화한 화자의 고유의 특징들이 있다고 알려져 있으며, 이를 이용한 화자인식의 연구가 과거 70년대와 80년대에 있었다[1,2]. 그러나 피치와 에너지와 같은 음성의 운율정보를 화자인식에 적용하는 경향이 근래에 많이 줄었는데, 이는 운율특징만으로는 문장종속 화자인식 스템에서도 사용하기에 만족할 만한 인식률을 얻을 수 없을 뿐만 아니라 문장독립 시스템으로 발전되기에 어려움이 있으며, 또한 피치 추출과정에서 오차가 발생하기 쉬우며 계산량도 적지않은 점이 단점으로 생각되었기 때문이다. 따라서 근래에는 주로 음성의 스펙트럼 정보를 표현하는 켑스트럼 계수를 이용하여 화자인식을 수행하고 있으며 좋은 결과를 보여주고 있다[3,4,5].

스펙트럼을 구하기 위하여 주로 사용하는 LPC분석은 계산이 간단하고 음성의 스펙트럼성분을 정확히 추정하기 때문에 보편적으로 음성분석에 많이 이용하고 있다. LPC분석은 음성발성을 전극(all-pole)필터로 모델화하여 이 필터의 계수인 LPC계수로서 스펙트럼 정보를 나타내는데, 모델이 정확할수록 LPC분석 오차신호인 잔차신호의 값이 작아지게 된다. 그러나 비음과 마찰음과 같이 반포만트 성분을 갖고 있는 음소에 대해서는 LPC분석이 적당한 모델이 아니므로 LPC계수는 여기에 포함되어 있는 화자의 특징을 표현하지 못한다고 하겠다. 반면에 LPC분석에서 생기는 오차인 잔차신호는 결국 LPC계수가 갖지 못하는 화자의 특징들을 내포하고 있다는 의미가 되며, 여기에 속하는 화자의 특징들로 기본주파수인 피치와 성문파의 모양 그리고 전극필터에 의해 모델링되지 않는 스펙트럼성분 등이 있으며 이를 이용한 화자인식 연구가 발표되었다[6,7].

일반적으로 화자인식의 성능에 많은 영향을 주는 요소로 전화채널의 왜곡과 잡음의 영향을 들고 있다. 그러나 피치는 스펙트럼정보를 갖고 있는 켑스트럼계수들에 비해 상대적으로 이러한 잡음에 대해서 영향을 덜 받는 것으로 알려져 있어[8], 피치와 같은 운율정보를 현재 많이 이용하고 있는 켑스트럼계수와 결합하여 화

자인식에 사용하려는 노력이 있다[9].

본 논문에서는 유성음에서 추출한 피치정보와 LPC켑스트럼계수를 결합한 형태의 문장독립화자식별 시스템의 성능에 대해서 알아보았다. 특히 화자가 발성한 전체 음성데이타를 사용한 경우와 유성음구간만을 검출하여 인식에 사용한 경우에 인식률을 비교하였고, 잡음환경에서의 피치정보의 유용성을 알아보기 위하여 입력음성의 SNR에 변화를 주어 인식률을 비교하였다.

논문의 구성은 서론에 이어 2장에서 본 논문의 화자인식 시스템으로 이용하는 GMM(Gaussian Mixture Model)에 대한 간단한 설명을 하였고, 3장에서는 사용하는 음성특징 파라미터로 피치와 LPC켑스트럼의 검출방법에 대한 내용을, 그리고 4장에서는 실험의 방법과 그 결과를 보여주고, 마지막으로 결론을 맺었다.

2. GMM(Gaussian Mixture Model)

GMM은 여러 개의 가우시안 확률밀도(Gaussian probability density)함수들에 각각의 가중치를 준 다음, 이를 선형 결합함으로써 임의의 모양을 갖는 확률밀도 함수를 표현할 수 있다. 그리고 음성의 특징 파라미터 벡터의 확률분포는 화자마다 그 모양이 다르며, 이러한 확률분포를 GMM을 이용하여 모델링하여 인식하고자 하는 화자의 모델로 사용함으로써 화자인식에 이용할 수 있다.

GMM에서 모델 λ가 주어져 있을 때 파라미터 벡터 x의 확률은 식 (1)과 같다.

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i N(x; \mu_i, \Sigma_i) \qquad (1)$$

여기서 x는 d차원 벡터, $N(x; \mu_i, \Sigma_i)$은 i번째 가우시안 분포, w_i는 혼합계수 그리고, λ는 GMM의 모델 파라미터를 나타낸다. 관측된 파라미터 벡터열 $X = x_1, x_2, \ldots, x_T$의 log-likelihood는 다음과 같다.

$$L(X|\lambda) = \log \prod_{t=1}^{T} p(x_t \mid \lambda) = \sum_{t=1}^{T} \log p(x_t \mid \lambda) \quad (2)$$

화자식별에서는 미지의 음성데이타의 신원을 주어진 N명의 화자의 모임에서 결정하는데, 일반적으로 likelihood값에 의한 결정은 화자 모델이 i^*일 때 다음 식(3)과 같다.

$$i^* = \arg\max_i L(X|\lambda_i) \qquad (3)$$

3. 화자인식 파라미터

(1) LPC 켑스트럼(LPCC)

파라미터로는 잡음에 강인하고, 일반적으로 많이 사용되는 LPC 켑스트럼 계수(LPCC)를 사용했다. LPC 켑스트럼 계수 $c(n)$은 선형예측계수 a_i로부터 다음의 관계식을 이용하여 계산한다[10].

$$c(n) = a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) c(i) a_{n-i} \qquad (4)$$

여기서 $c(n)$은 무한대의 구간에서 존재하며, n의 값이 커질수록 계수는 $1/|n|$에 비례해서 작아지고, 대신에 n의 값이 작을수록 계수의 중요성은 커진다. 따라서 켑스트럼의 차수를 p로 정하면, $c(1)$에서부터 $c(p)$까지의 계수만을 사용하게 되며, 이렇게 전체 켑스트럼의 일부만을 사용하여도 물리적으로는 파라미터를 추출한 음성구간에 들어 있는 스펙트럼 포락선에 관한 정보를 갖고 있다.

(2) 피치

음성의 기본 주파수인 피치(pitch)는 일반적으로 성문을 통한 공기흐름의 주기를 나타낸다. 이 피치정보는 LPC-잔차신호, 평균절대차함수(AMDF:average magnitude difference function) 또는 자기상관함수 등으로부터 추출할 수 있다. 본 논문에서는 피치의 효과적 추출을 위해 프레임의 길이에 추가 음성구간을 보강한 변형된 자기상관함수를 이용한다.

단구간 자기상관함수을 구하는 식은 다음과 같다.

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(n+m) w_1(m) x(n+m+k) w_2(m+k) \quad (5)$$

식 (5)에서 $w_i(m)$은 창 함수를 나타내고, k는 시간지연 차수를 나타낸다. 이때 창함수는 직사각창으로 하며 $w_2(m)$의 길이는 최대시간지연차수 만큼 음성구간을 추가로 연장시켜 시간지연에 따른, 상관계수값의 급격한 감소현상을 없앤다[11].

(3) LPCC와 피치를 통합한 벡터

피치 정보를 LPCC와 함께 나타낼 때, 로그를 취한 피치값을 첨가한다. 무성음 구간에서는 피치 값은 0가 되어 켑스트럼 벡터만 나타나게 되어 파라미터 벡터는 다음과 같이 두가지 형식으로 표현될 수 있다.

$$x_t^{voiced} = (c_{1t}, c_{2t}, \ldots, c_{dt}, \log F0_t)$$

$$x_t^{unvoiced} = (c_{1t}, c_{2t}, \ldots, c_{dt}) \qquad (6)$$

여기서 C_{it}와 $F0_t$는 각각 시간 t일 때 i번째 켑스트럼 계수와 피치의 로그값이다. 이때 $F0$대신 $\log F0$를 사용

하는 이유는 $F0$보다 $\log F0$의 분포가 가우시안 분포에 더욱 가깝기 때문이다[12].

4. 피치와 LPCC를 통합한 GMM

피치와 LPCC를 통합한 파라미터를 사용하기 위해서 피치의 유무에 따라 각 음성 프레임을 무성음 또는 유성음 프레임으로 나눈 후 각각을 EM(Expectation Maximization)알고리즘을 사용하여 유성음과 무성음의 GMM모델 λ_v, λ_{uv}를 각각 훈련시킨다.

일반적으로 피치검출 시에 유성음 구간의 처음이나 끝부분에서 피치가 검출이 안되거나, 반대로 무성음 구간에서 피치가 검출되는 경우가 있지만 이것은 전체 데이터에 비해 매우 적은 부분이기 때문에 인식에 큰 영향이 없다[9]. 또한 피치와 켑스트럼 계수의 상관 관계를 모델화하여 화자인식에 활용하기 위해 완전공분산행렬(full covaria-nce matrix)을 모델에 사용하는데 다음은 공분산행렬의 구조이다.

$$\Sigma=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} & \rho_1 \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} & \rho_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} & \rho_d \\ \rho_1 & \rho_2 & \cdots & \rho_d & \rho \end{bmatrix} \quad (7)$$

여기서 σ와 ρ는 각각 켑스트럼 계수와 피치의 분산 값이며, ρ_i는 피치와 켑스트럼의 공분산이다.

통합 파라미터에 의한 전체 음성데이타에 대한 인식실험을 위해서, 주어진 음성실험 데이터를 우선 유, 무성음으로 구분한 후, 각각의 GMM모델에 대한 log-likelihood을 계산한 다음 이들 두 개의 유사도값의 선형합을 취해 전체 유사도를 계산한다. 이때 가중치 α를 사용한 식은 다음과 같다.

$$L(X)=\alpha L(X_{uv} \mid \lambda_{uv})+(1-\alpha)L(X_v \mid \lambda_v) \quad (8)$$

여기서 X_{uv}와 X_v는 각각 유성음, 무성음의 벡터열을 나타낸다[9].

5. 실험 및 결과

본 논문에서는 유성음 구간에서 검출한 피치정보를 LPC켑스트럼과 함께 화자인식에 적용하면 인식 성능에 어떤 영향을 주는지 검토하였다. 실험에 사용한 화자식별시스템은 GMM을 이용하여 구성하였고, 이때 GMM의 혼합 가우시안 분포의 개수는 M=8으로 하였다. 또한 GMM에 사용한 공분산 행렬은 완전 행렬(full matrix)을 적용하였다. 또한 전체음성에 대한 인식실험에서 유, 무성음의 GMM모델을 통합한 유사도 계산에서 α값을 0.5로 정하였다[5]. 본 실험에서는 20대의 40명의 화자(남자 30명, 여자 10명)를 대상으로 음성데이타를 수집하였으며, 화자등록용으로 30초 정도, 그리고 인식 실험용으로 15초씩 각각 4번씩 실험실 환경에서 녹음하였다. 대상 문장은 일반 교양과목의 교재로 모든 음성 데이터는 서로 다른 문장을 사용하였으며, 화자 개인별로 총 1분30초 분량으로 데이터베이스를 구성하였다. 녹음에는 AKG D190마이크와 PC 사운드 카드를 사용했다. 음성데이터는 8kHz샘플링하여 16비트로 저장하였으며 음성의 한 프레임의 길이는 25ms로 200샘플/프레임이며, 15ms씩 중첩하여 끝점 검출과 전처리과정을 거쳐서 특징 벡터를 추출하였다. 또한 실험에서는 기준 파라미터로 12차의 LPC 켑스트럼사용하였으며 피치 검출시는 피치값을 반영한 13차의 켑스트럼을 사용하였다. 먼저 LPCC 켑스트럼과 피치의 상관관계를 이용했을 때 성능이 얼마나 개선 되는 지를 알아보는 실험을 하였고 그 결과는 표 1과 같다.

표1. 파라미터의 성능 비교

	전체음성	유성음
LPCC	93.9%	92.5%
LPCC+피치	86%	87.2%

표 1에서 유성음구간에서 추정한 LPCC 파라미터를 사용했을 때는 전체음성을 모델링한 경우에 비해 인식율이 약간 저하됨을 알 수 있다. 또한 인식파라미터에 피치를 추가할 경우에는 전체음성이나 유성음의 경우 모두 인식률이 저하되는데 이는 피치정보 추출의 정확도가 떨어져서 생기는 결과로 생각이 된다. 그러나 피치를 LPCC와 통합한 파라미터를 이용할 경우는 상대적으로 유성음 구간을 이용한 방법의 인식률이 약간 개선됨을 알 수 있다. 이는 유성음구간에서 피치와 LPCC 계수의 상관도를 이용한 것이므로 타당성이 있다고 보겠다. 또한 서로 다른 특성을 갖는 마이크에 대한 피치의 성능을 보기 위하여 Shure 마이크를 이용하여 같은 문장을 한번 더 녹음한 후 실험에 이용하였으며 결과는 표2에 있다. 결과에서 유성음에서 추출한 피치를 함께 이용한 경우에 성능이 향상됨을 알 수 있겠다.

표2. 서로 다른 마이크로 녹음한 음성에 대한 인식결과.

	전체음성	유성음
LPCC	91.3%	92.5%
LPCC+피치	92.0%	93.7%

그리고 시간 간격을 두고 발성한 음성데이터에 대한 피치의 성능을 보기 위하여 화자 14명이 2회에 걸쳐 녹

음한 데이터를 이용하여 실험을 하였으며 결과는 표2에 보인다. 이때 처음 음성시료를 얻은 후 40일 후에 두번째 음성 시료를 녹음하였다. 결과에서 보면 유성음구간의 정보를 이용하였을 때가 모든 경우에서 성능이 약간 개선됨을 보여준다.

표3. 녹음 시기가 다른 음성에 대한 성능비교

	전체음성	유성음
LPCC	82.3%	88.5%
LPCC+피치	83.0%	85.7%

6. 결론

본 논문에서는 GMM을 이용한 화자인식에서, GMM 모델을 훈련하고 인식실험을 할 때 사용하는 음성데이터의 크기에 따라 인식성능이 영향을 받는 것을 개선하기 위해 음성신호의 유성음구간만을 이용하여 인식성능을 개선할 수 있는지에 관한 실험을 하였다. 깨끗한 음성데이타에 대한 실험 결과에서 보면 피치값의 추출의 정확도가 성능에 많은 영향을 줌을 알 수 있었다. 그러나 피치와 켑스트럼계수와의 상관관계가 존재하여 이를 이용할 경우, 시간에 따른 변화 및 마이크의 변화에 대하여 피치정보가 스펙트럼 정보에 비해 덜 영향을 받음을 알 수 있었다. 추후 유, 무성음구간의 정확한 분리와 피치값의 추정을 통해서 환경의 변화에 강인한 화자인식시스템의 구성에 피치를 이용할 가능성을 확인할 수 있었다.

참고문헌

[1] B.S.Atal, "Automatic speaker recognition based on pitch contours," JASA vol.52, No.6. pp.1687-1697.

[2] J.D.Markel, B.T.Oshika and A.H.Gray, "Long-term feature averaging for speaker recognition," IEEE Trans. ASSP vol.ASSP-25, pp.330-337.

[3] T.Matsui and S.Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Proc. ICASSP, vol.II, pp.157-160, 1992.

[4] H.Gish and M.Schmidt, "Text-independent speaker identification," IEEE Signal processing magazine, pp.18-32, Oct. 1994.

[5] D.A.Reynolds and R.C.Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on SAP, vol.3, No.1, pp.72-83, 1995.

[6] P.Thevenaz and H.Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," Speech communication, vol.17, pp.145-157, Aug. 1995.

[7] J.He, L.Liu and F.Palm, "On the use of features from prediction residual signals in speaker identification," Proc. Eurospeech, pp.313-316, 1995.

[8] M.J.Carey, E.S.Parris, H.L.Thomas and S.Bennett, "Robust prosodic features for speaker identification," Proceedings ICSLP 96. Fourth International Conference on Spoken Language Processing IEEE. Part vol.3, pp.1800-1803, New York, NY, USA. 1996.

[9] Markov, Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," Journal of the Acoustical Society of Japan (E), vol.20, No.4, July 1999, pp.281-291.

[10] L. R. Rabiner and B. H. Juang, 1993. Fundamental of speech recognition, Prentice Hall, Englewood Cliffs, NJ

[11] L. R. Rabiner and R. W. Schafer, 1978. Digital Processing of Speech Signal, Prentice Hall, Englewood Cliffs, NJ

[12] M. K. Sonmez, L. Heck, M. Weintraub. and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition", in Proc. EUROSPEECH, 1997, pp.1391-1394.