1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

성문파형을 이용한 문장독립 화자 인식기

양기혁, 전범기, 백성준, 강상기, 성굉모
서울대학교 전기공학부

Text-Independent Speaker Recognition Using Glottal Flow Waveform

Ki-Hyuk Yang, Bumki Jeon, SeongJoon Baek, Sang-Ki Kang, and Koeng-Mo Sung
School of Electrical Engineering, Seoul National University, Seoul 151-742, KOREA
TEL : 02-880-7263, FAX : 02-882-4657

요 약

본 논문에서는 성문파에서 화자특성 계수를 추출하여 화자 인식기에 적용하고자 한다. 공분산 방법으로 음성의 잔류신호를 추정하고 이를 적분하여 성문파를 얻어낸다. 하나의 성문파 구간을 성문닫힘순간 사이가 아닌 잔류신호의 오차가 최대가 되는 순간 사이로 잡았다. 구해진 성문파를 M개의 데이터로 다시 샘플링하여 특성 벡터로 삼고 VQ기반 인식기를 사용하여 인식률을 측정하였다. 4초의 test data와 30차의 특성벡터를 사용한 경우 남성의 경우 평균 96.08%, 여성에 대하여 93.61%의 평균 인식률을 얻었다.

1. 서론

음성신호에는 언어학적 정보 (linguistic information)와 개별화자 정보(specific speaker information)가 들어있다. 성도(vocal tract)의 특성에 근간한 특성계수인 LP계수와 캡스트럼계수 등은 언어학적 정보와 개별화자 정보를 모두 가지고 있으므로 음성인식(speech recognition)과 화자인식(speaker recogni-tion)에 널리 사용되어 왔다. 그런데 성문파(glottal flow waveform)도 개별화자별로 그 모양이 서로 다르고 또한 음성신호를 선형 예측한 후에 남은 잔류신호에서부터 추정되므로, 성문파가 선형예측에서 얻어지는 특성계수들과 독립적인 성분을 가질 것이라고 기대할 수가 있다.

개별화자의 성문파 특성을 화자인식 시스템에 응용하기 위해 우선 음성으로부터 성문파를 추정해 내야한다. 성문파는 폐에서 발생한 공기압력에 의해 성문이 열린 후 닫힐 때까지 발생하는 공기의 흐름이므로, 성문 닫힘 순간을 시작으로 하고 다음 성문 닫힘 순간을 끝으로하는 구간(그림 1-b)을 하나의 독립된 과정으로 보는 것이 타당하며, 따라서 이 구간에서 성문파를 묘사하는 것이 일반적이다. 이를 위해 성문 닫힘 순간을 추정해 내는 방법들이 제안되어왔는데 [2],[3] 이 방법들은 시간이 오래 걸리기 때문에 실시간으로 작동되는 화자인식기에 적용하기에는 부적절하다. 또한 추정한 구간의 신뢰성이 떨어질 때는 그 구간의 정보가 잡음의 역할을 하므로 인식기의 인식률 향상을 기대할 수 없다.

본 논문에서는 성문파 추정 시간을 줄이기 위해서 유성음에서는 인접한 두 성문파는 거의 같은 모습을 가지므로 성문파 구간을 잔류신호의 최소값이 발생하는 순간에서 다음 최소값이 발생하는 순간(그림 1-a)까지로 설정하였다. 따라서 잔류신호의 최소값이 발생하는 순간은 그 근처의 피치정보를 이용하면 비교적 빠르고 정확하게 찾아낼 수 있다.

다음절에서는 성문파 추정과 파라미터 추출법 및 이 파라미터를 인식기에 적용하는 방법에 대해 기술하고, 3절에서는 이를 이용한 실험결과를 제시한다. 그리고 마지막으로 결론과 앞으로의 연구방향을 제

시한다.

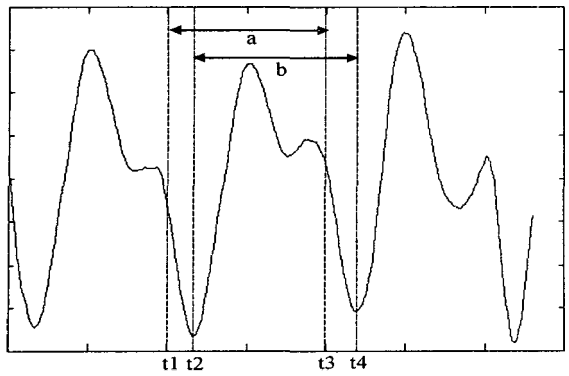

그림 1. 추정된 성문파형
a : 추정된 성문파 구간, b : 실제 성문파 구간

2. 성문파라미터 추출과 화자 인식기의 구현

성문파는 성문이 진동하면서 발생하는 공기의 흐름이므로 유성음에서만 추정할 수가 있다. 무성음을 제거하기 위해 영교차율(zero-crossing rate)과 음성의 에너지를 순차적으로 이용하였다.

성문파를 미분한 유효구동함수(effective driving function)는 음성에서 성도의 영향을 제거한 잔류신호로 간주된다. 따라서 성문파형을 추정하기 위해서는 선형 예측 방법을 이용한 잔류신호의 추정이 필요하다. 음성신호는 일반적으로 비정적(non-stationary)신호이므로 자기상관(autocorrelation) 방법에 비해 공분산(covariance) 방법을 사용하는 것이 효과적이다 [7].

서론에서 언급한 바와 같이 성문파 구간을 결정하는 경계는 잔류신호의 한 피치구간에서 최소값을 갖는 순간이다. 이 순간을 찾기위해 근처의 피치값을 활용할 수 있다. 피치는 정규자기상관(normalized autocorrelation) 방법에 의해 찾아내었다 [8]. 공분산(covariance)에 의한 선형예측방법으로 얻어낸 잔류신호를 적분함으로서 성문파를 얻어낼 수 있다.

본 논문에서는 개별화자 정보가 성문파의 크기(magnitude)뿐만 아니라 위상(phase)에도 포함되어 있으므로 성문파에 대한 LP 계수와 켑스트럼 계수보다는 화자인식을 위한 파라미터로 추정된 성문파의 파형 자체를 사용하였다. 추정된 성문파의 샘플수(P)는 피치간격에 따라 다른 값을 가지기 때문에 그림 2와 같이 이를 일정한 갯수(M)로 다시 샘플링하였다. 성문파는 적분기를 통과한 신호이므로 음성의 샘플링 주파수에 비해서 낮은 주파수 영역에 에너지가 집중되어 있다. 따라서 M이 P보다 작은 값을 갖더라도 손실없이 성문파를 표현할 수 있다. 하지만 성문파가 높은 주파수 대역의 성분을 가질 수 있기 때문에 샘플수를 줄이는 과정에서 aliasing이 생길 수 있으므로 성문파에 간단한 median 필터를 통과시킨 후에 다운 샘플링 하였다.

본 논문에서 제안한 성문파 파라미터가 개별화자 정보를 가지고 있다는 것을 설명하기 위해서 간단한 VQ 기반 화자인식 실험을 하였다. 제안된 성문파 파라미터는 시간 영역 특성이므로 화자인식 파라미터로 사용하기 위해서는 추정된 성문파의 DC 성분을 제거하고 에너지를 정규화(normalization)시켜야 한다. 그림 3은 성문파 파라미터 추출 및 인식 시스템에 대한 전체적인 블록도이다.

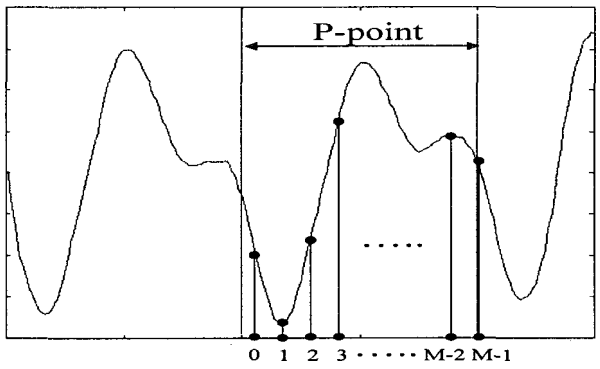

그림 2. M-샘플로의 다운 샘플링 과정

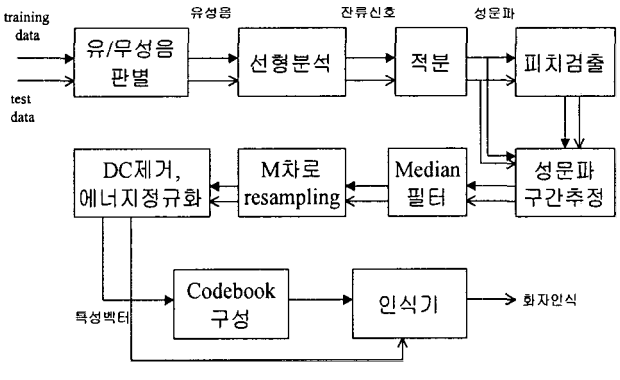

그림 3. 화자 인식 블록도

3. 실험 및 토론

14명의 여성화자와 14명의 남성화자에 대하여 추출된 성문파의 파라미터만을 이용하여 화자식별 실험을 하였다. 샘플링 주파수는 16kHz 이고 화자별로 3분의 길이의 음성을 이용하여 학습시켰다. 한 개의 성문파에서 M개의 특성계수를 추출하고 codeword

의 크기는 256개로 하였다. 남성과 여성의 Test 음성의 길이에 대한 평균 인식률을 그림 4 ,5에 나타내었다 (M=20, 30). 모두 M이 30인 경우가 20인 경우에 비해 더 좋은 인식률을 나타내었다. 표 1은 남성 화자에 대해 M의 개수를 변화 시켰을 때의 인식율이다.

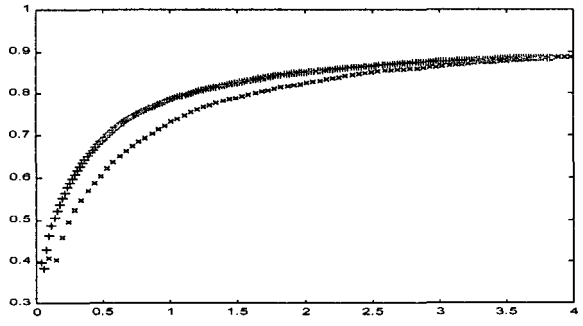

그림 4. 테스트 음성길이에 따른 인식률(x:남자 +:여자 M=20)

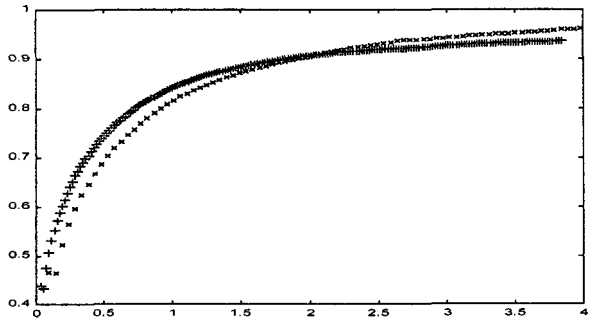

그림 5. 테스트 음성길이에 따른 인식률(x:남자 +:여자 M=30)

표 1. M의 개수에 대한 남성 화자의 인식률

	M=20	M=30	M=50	M=70
1초	73.23	81.60	82.54	84.611
2초	82.49	90.59	91.85	93.57
3초	86.47	94.34	95.23	96.82
4초	88.78	96.08	96.67	97.65

이번 실험에서는 성문파 신호 자체를 특성계수로 사용하였는데 계수의 숫자를 줄이기 위해 저역 성분만을 사용하였다. 인식률을 향상시키기 위해서는 다양한 종류의 feature를 이용하여 실험해 볼 필요가 있다. 성문파의 cepstrum 계수와 Fourier series 계수 등을 사용하여 인식기를 만드는 방법도 가능하지만 이 때에는 위상성분을 반드시 고려해 주어야 할 것이다.

본 논문에서는 성문파의 개별 화자 특성만을 화자 인식에 적용하였지만, 성문파라미터와 성도(vocal tract) 파라미터와 결합하여 인식기에 사용하는 방법도 연구되어야 할 것이다. GMM(Gaussian Mixture Model)을 이용한 인식기는 독립적으로 얻은 두 파라미터 set을 결합하는데 용이하고, 그 성능이 VQ에 기반을 둔 인식기보다 우수한 것으로 알려져 있다 [1].

성도 파라미터와 독립적인 성문 파라미터를 찾아내고 빠르게 계산할 수 있는 알고리듬을 개발하는 것이 앞으로의 연구과제이다.

4. 결론

본 논문을 통해 성문파를 이용하여 화자 특성 계수를 뽑아내고 이를 화자 인식기에 적용하는 방법에 대하여 살펴 보았다. 잔류신호를 적분하여 성문파를 추정해 내고, 빠른 인식 시스템 구현을 위해 성문파 한구간의 시작과 끝이 잔류신호의 값이 최소가 되는 시각이라고 가정하였다. 추정한 성문파형을 decimation하여 VQ에 기반을 둔 인식기로 실험을 하였다. 실험 결과에 의하면 성문파만으로도 화자인식이 가능하다는 사실과 특성계수의 숫자를 높일수록 인식률이 증가한다는 사실을 확인할 수 있었다. 그러나 성문 파라미터의 차수를 줄이기 위해 다른 종류의 feature set을 찾을 필요가 있다.

참고 문헌

[1] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. SAP*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[2] Y. M. Cheng and D. O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period." *IEEE Trans. ASSP*, vol. 37, no. 12, pp. 1805-1814, Dec. 1989.

[3] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of Vocal-Tract System Characteristics from Speech Signals," *IEEE Trans. SAP*, vol. 6, no. 4, pp. 313-326, July 1998.

[4] D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Mag.*, pp.4-17, Oct. 1986.

[5] K. R. Farrell and R. J. Mammone, "Speaker Recognition Using Neural Networks and Conventional Classifiers," *IEEE Trans. SAP,* vol. 2, no. 1, pp. 194-204, Jan. 1994.

[6] Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Communications*, vol. COM-28, no. 1, Jan. 1980.

[7] Deller, Proakis and Hansen, "Discrete-Time Processing of Speech Signals," *Prentice-Hall,* 1987.

[8] ITU-T, Draft Recommendation G.723 "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 & 6.3 kbit/s,"