성량제한을 적용한 어구독립 화자증명 성능향상 방안

이태승[0] 최호진
한국항공대학교 항공전자공학과
thestaff@hitel.net, hjchoi@mail.hangkong.ac.kr

On a Method Which Improves Text Independent Speaker Verification Performance through Limiting Speech Production Loudness

Tae-Seung Lee[0] Ho-Jin Choi
Dept. of Avionics Eng., Hankuk Aviation Univ.

요 약

지속음(continuants) 단위로 화자간 차이를 식별하는 어구독립 화자증명(text-independent speaker verification) 방식에서 입력음성의 성량을 제한하여 보다 높은 인식률을 달성할 수 있는 화자인식 방법을 제안한다.

1. 서 론

음성에서 화자정보를 내포하는 파라미터로는 억양, 발성속도, 단구간 스펙트럼 등이 있지만, 이들 중 단구간 스펙트럼이 가장 안정적이면서도 강력한 화자특성을 보여준다. 그러나 스펙트럼은 똑같은 언어단위를 발성하는 경우에도 문장 중의 위치나 발성상황 및 환경에 따라 다른 성량과 스트레스로 발성함으로써 각 발성시기마다 상당한 차이를 보이게 되며, 특히 성량의 변화에 따른 영향이 크다. 음성인식에 비해 구별해야 하는 화자간 스펙트럼의 차이가 작은 화자인식에서 이러한 왜곡은 인식률에 상당한 영향을 줄 수 있다.

이 문제를 해결하기 위해 본 논문에서는 음성인식 기술을 사용하여 지속음(continuants) 단위로 화자간 차이를 식별하는 어구독립 화자증명(text-independent speaker verification) 방식에서 입력음성의 성량을 제한하여 보다 높은 인식률을 달성할 수 있는 방법을 제안한다.

논문의 구성은 다음과 같다. 먼저 인식 모델로 지속음을 사용하는 이유를 2장에서 설명하고, 성량이 화자인식 성능에 미치는 영향과 이에 대한 해결방안을 3장에서 제시한다. 그리고 4장에서 이 방법을 시험할 화자증명 시스템의 구성을 나열한 다음, 이 시스템을 이용하여 음성 데이터베이스를 시험한 실험내용과 결과를 5장에서 다루고, 최종적으로 6장에서 이들을 정리한다.

2. 지속음 모델 화자인식

화자인식시 음성 신호가 가지는 정보는 발성하는 문장에 대한 언어정보와 화자정보를 동시에 포함하고 있다. 다시 말해, 우리가 어떤 사람의 목소리를 구분할 때는 주어진 특정한 언어정보(즉, 문장이나 단어, 음절, 음소 등)에서 그 사람만의 독특한 특징을 인식하는 것으로 볼 수 있는 것이다. 보편적으로 사용되는 단구간 스펙트럼과 같은 음향 파라미터의 경우, 개인차가 단어나 음소에 따른 영향을 넘어설 정도로 크지 않아 음성으로부터 언어정보를 제거한 화자정보만을 추출하기는 어렵다. 따라서, 화자인식을 위해서는 언어정보를 완전히 제거하기보다는 화자특성을 언어에 무관한 것과 언어 종속적인 것으로 구분하여 이 둘을 동시에 사용하는 것이 분석이나 인식률 측면에서 바람직하다.

음성을 이루는 음소의 종류에 따라서는 다른 부분에 비해 화자인식에 더 유용하다는 사실이 밝혀졌다[1][2]. 따라서 화자인식에 어떤 음소범주를 사용하는 것이 가장 좋은지 명확히 알 수만 있다면 인식성능을 향상시킬 수 있음이 분명하다. 예를 들어 양호한 음소의 비중이 높은 비밀번호를 선택한다면 이러한 특성의 효과를 기대할 수 있다. 또는, 화자인식에 유리한 음소를 자동으로 식별하는 전단(front-end) 인식기를 사용할 수 있다면 최종 인식 단계에서 적절한 가중치를 부여함으로써 인식률 향상을 꾀할 수 있을 것이다.

Eatock 등[1]은 영어의 각 음소범주와 이들에 따른 화자간 인식능력의 차이를 연구하였다. 이들에 따르면 비음과 모음이 가장 뛰어난 성능을 보이고 그 뒤를 마찰음과 폐쇄음이 잇는다. 이 사실은 Delacrtaz 등[2]의 유사한 연구에 의해 뒷받침되고 있다.

이러한 결과와 음성발성의 무손실 튜브 모델(lossless tube model)[3]을 고려할 때 화자간 차이는 각 언어단위 발성시 화자의 구강구조 차이에 따른 공진 주파수의 차이로 볼 수 있으며, 모든 언어단위 가운데서도 지속적이며 공진 에너지가 큰 공명음(sororants)이 화자간 차이 정보를 양호하게 제공한다고 추론할 수 있다. 이에 따라 지속적인 부분이 비교적 많은 비음(nasals), 모음(vowels), 유사음(approximants)에서 지속부분(이후 지속음)을 채취하여 /a(아)/, /e(애)/, /v(어)/, /o(오)/, /u(우)/, /U(으)/, /i(이)/, /liq(종성ㄹ)/, /nas(종성ㅁ,ㄴ,ㅇ)/ 총 9개의 지속음을 화자간 구별의 기본 언어단위로 사용한다.

3. 성량에 따른 음성 파라미터 변화문제와 해결방법

통계적 유사도 비교방법[4]이나 신경망을 이용한 방법[5] 등 화자인식을 위한 대다수의 학습 및 인식방법은 의뢰화자의 인식점수 계산에 사용되는 배경화자가 모든 화자를 대표한다고 가정한다. 즉, 모든 화자가 발생시킬 수 있는 음성신호 범위를 배경화자의 음성이 충분히 표현할 수 있을 때 학습 및 인식방법의 유효성이 보장되는 것이다.

그러나 실제 화자인식 시스템의 설계에서는 배경화자의 음성을 사전에 제작한 데이터베이스에서 채취하는데, 이런 데이터베이스의 음성은 이상적인 조건(저잡음, 고성능 마이크, 일정한 성량 등) 하에서 녹음되는 것이 일반적이다. 따라서 배경화자의 음성은 발생 가능한 모든 범위에서 일부분만을 표현한다.

이 문제는 동종의 데이터베이스에서 배경화자의 음성과 성능시험용 음성을 채취하는 경우에는 심각하게 나타나지 않지만, 데이터베이스와 다른 조건에서 발성된 음성을 시험에 사용하는 경우에는 심각한 성능저하로 나타날 수 있다. 이런 문제 가운데 잡음과 음성입력용 마이크에 관련한 문제는 지금까지 상당부분 연구가 진행되었지만[6], 가장 혼하게 발생할 수 있는 일정하지 못한 성량에 의한 신호차이에 대한 연구는 충분히 다뤄지고 있지 못하다.

성량차는 발성중에 지역적으로 또는 광역적으로 나타난다. 한 문장을 발성하는 도중에도 구나 어휘의 강세에 따라 성량이 크게 변하며, 전달경로, 배경잡음, 정보의 중요정도에 따라서도 성량이 커지거나 작아질 수 있다. 한편, 원래부터 다른 사람에 비해 목소리가 큰 화자도 있을 수 있다. 이와 같은 성량의 변화는 스펙트럼의 형태와 포락선, 피치, SNR의 변화와 직접적인 관련이 있다[7].

본 연구에서는 정적인 스펙트럼의 형태와 포락선을 화자간 차이구별용 특징으로 사용하고 있으므로, 위의 현상들 가운데서도 성량차에 의한 스펙트럼의 변화가 중요하다. 성량에 따른 스펙트럼의 변화는 (1)포락선 높이의 변화와 (2)고주파 대역의 스펙트럼 형태 변화로 구분할 수 있다. 성량은 발성 에너지를 의미하므로 성량이 높을수록 스펙트럼의 포락선 위치가 높아진다. 또한 배경잡음이나 성도의 조음체적 차이로 인해 성량이 변할 수 있는데 이 때 스펙트럼의 고주파성분이 크게 달라지게 된다.

화자인식에서 이용하는 화자간 특징차이는 그 크기가 상대적으로 미미하므로 이와 같은 스펙트럼의 변화에 민감하게 영향받을 우려가 높다. 이에 따라 본 연구에서는 화자등록과 증명단계에서 받아들이는 음성을 배경화자 음성의 성량을 기준으로 선별하는 방식을 제안한다. 즉, 배경화자 음성의 성량분포를 분석하여 최대성량과 최소성량을 결정하고 이 범위 안에 들어가는 입력음성만 받아들이는 것이다.

이를 위해 먼저, 배경화자의 각 지속음에 대해 최대성량과 최소성량을 알아낸다. 성량측정치는 입력되는 음성의 일정간격 프레임마다 에너지를 계산하여 얻는다.

$$Loud(p,n)=\frac{1}{M}\sum_{i=0}^{M-1}\left|S(M\cdot n+i)\right| \tag{1}$$

여기서, S는 음성샘플을, p는 지속음을, M은 프레임의 음성샘플링 개수를, n은 단위 음성 내 프레임 번호를 나타낸다.

그런 다음, 화자등록과 증명단계에서 입력된 단어의 지속음 프레임들 가운데 배경화자 데이터를 대상으로 미리 측정한 성량범위를 벗어나는 프레임의 비율이 일정수준 이내일 때 그 프레임들을 이용하여 등록과 증명을 처리한다.

4. 화자증명 시스템 구성

본 연구에서는 어구독립 화자증명 시스템을 이용하여 성량범위를 제한하는 방법을 시험했다. 그림 1에서 보이는 각 처리부분의 기능은 아래와 같다.

(1) 16bit 16kHz로 샘플링된 입력음성을 20ms 오버랩시킨 30ms 프레임으로 나눈다.

(2) 각 프레임에 대해 16차 Mel 간격 필터뱅크[8]를 추출하여 고립단어 및 지속음 검출에 사용한다. 필터뱅크 계수는 전체 스펙트럼 포락에 미치는 성량의 영향을 제거하기 위해 1kHz까지의 계수를 평균하여 모든 계수에서 차감한다.

(3) 각 프레임에 대해 50차의 0-3kHz 대역 균등간격 Mel 필터뱅크를 추출하여 화자증명에 사용한다. 이 음성특징은 2차 포만트에 더 많은 화자정보가 집중된다는 연구결과[9]에 의한 것이다. 필터뱅크 계수는 전체 스펙트럼 포락에 미치는 성량의 영향을 제거하기 위해 1kHz까지의 계수를 평균하여 모든 계수에서 차감한다.

(4) 각 프레임에 대해 식(1)로 성량을 계산하여 성량크기 검사에 사용한다.

(5) 각 지속음과 묵음을 화자독립으로 검출하도록 학습된 MLP(multilayer perceptron)[10]를 사용하여 지속음이 포함된 고립단어를 검출한다.

(6) 검출된 고립단어 내의 지속음들의 프레임 성량을 검사하여 전체 지속음에 대해 성량검사를 통과한 지속음 프레임 비율을 계산한다. 비율이 일정 수치 이하면 화자등록 또는 증명처리를 취소한다.

(7) 화자등록 처리인 경우 성량검사를 통과한 각 지속음에 대해 MLP와 배경화자 데이터를 이용하여 등록화자를 학습한다.

(8) 화자증명 처리인 경우 성량검사를 통과한 각 지속음을 해당 MLP에 입력하여 프레임별 점수를 모아 평균을 낸다. 이 평균값과 사전 설정한 문턱값을 비교하여 거부 및 수락 여부를 결정한다.

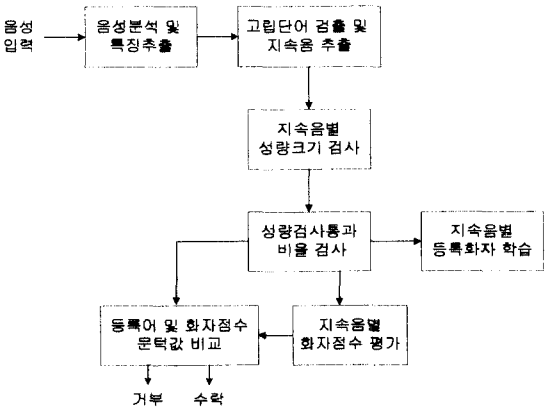

그림 1. 화자증명 시스템 구성도

5. 실험 및 결과

화자증명 시스템의 성능평가를 위해 한국과학기술대와 광운대에서 공동 제작한 음성 데이터베이스를 사용하였다. 이 데이터베이스는 단독숫자 및 지시어, 4연숫자, 단문, PBW(phone-balanced word)의 발성목록으로 구성되어 있는데, 본 연구에서는 이들 중 PBW와 4연숫자 목록을 지속음 인식 MLP와 화자증명 MLP에 각각 사용하였다. 이 목록들은 70명분의 음성이 방음 처리된 방에서 녹음되었으며 이를 16kHz로 샘플링하고 16bit로 양자화하여 DARPA-TIMIT 데이터 형식의 파일로 저장되었다.

4연숫자 목록은 4개의 0(공)~9(구)를 임의로 조합한 고립단어 35개를 단어마다 4회씩 발성한 것이다. 각 단어의 지속시간은 대략 1~1.5초 가량 된다.

화자증명 인식률을 측정하기 위해 4연숫자 목록에서 화자 29명을 배경화자로 사용하고 40명을 실험화자로 사용한다. 인식률은 각 실험화자를 한 번씩 등록화자로 했을 때 나머지 화자를 사칭화자로 간주하여 FR(false reject)과 FA(false accept)로 측정한다. 등록은 등록화자의 4회 발성 중 3회분을 이용하고 증명시험은 모든 등록화자의 마지막 1회를 이용한다. 따라서 FR은 35단어 X 40명(등록화자) X 1회 = 1,400회의 시도로 측정되며, FA는 35단어 X 40명(등록화자) X 39명(사칭화자) X 1회 = 54,600회의 시도로 측정된다.

성량범위 제한방법의 효과를 측정하기 위해 시험음성을 180%, 140%, 100%로 각각 증폭하여 성량을 임의조정하였다. 그리고 이들을 각각 조합하여 등록/검증한 뒤 각 조합에 대한 인식률을 표 1과 표2에 기록하였다. 표 내의 왼쪽과 오른쪽 기록치는 각각 성량검사를 하지 않았을 때와 성량통과 프레임비율을 70%로 했을 때의 인식률을 나타낸다.

표 1. FA 결과치

학습 \ 증명	180%	140%	100%
180%	9.6%/8.9%	7.4%/7.2%	3.2%/3.1%
140%	7.2%/6.7%	6.8%/6.4%	3.2%/3.0%
100%	2.7%/2.5%	2.7%/2.2%	2.7%/2.5%

표 2. FR 결과치

학습 \ 증명	180%	140%	100%
180%	1.4%/2.7%	3.7%/5.7%	32.3%/35.2%
140%	3.6%/5.5%	2.8%/4.0%	19.7%/20.7%
100%	29.5%/34.1%	19.6%/22.8%	2.7%/2.5%

인식률 결과에서 나타나듯이 FA는 전반적으로 향상되었으나 100% 조합을 제외한 FR은 전반적으로 나빠졌다. 이것은 성량범위 제한에 따라 좀더 엄격한 화자식별 기준이 적용됨을 의미하며, 이로 인해 보안성 면에서 중요한 FA가 향상된 것이다. 이 두 인식률과 더불어 증명시도에 대한 수락율을 측정하였다. 이 수락율은 미리 설정된 성량범위를 넘어 발성한 경우 이를 통보하고 증명처리를 취소하는 비율을 의미한다.

표 3. 증명시도 수락율

학습 \ 증명	180%	140%	100%
180%	99.7%/46.1%	99.5%/56.5%	99.5%/60.2%
140%	99.6%/56.6%	99.6%/75.6%	99.6%/82.8%
100%	99.2%/60.2%	99.3%/82.7%	99.5%/94.6%

이 결과에서 보듯이 적절한 성량의 발성음성만 받아들여 증명처리를 함으로써 보다 정확한 인식을 위해 의뢰화자에게 협조를 구할 수 있으므로 체감 인식률을 높일 수 있다.

6. 결론

음성에서 가장 안정적이면서도 강력한 화자특성을 보여주는 단구간 스펙트럼은 똑같은 언어단위를 발성하는 경우에도 문장 중의 위치나 발성상황 및 환경에 따라 다른 성량으로 발성함으로써 각 발성시기마다 상당한 차이를 보이게 된다.

이 문제를 해결하기 위해 본 논문에서는 입력음성의 성량을 제한하여 보다 높은 인식률을 달성할 수 있는 방법을 제안하였다. 지속음을 화자인식 단위로 사용하고, 이러한 지속음을 추출하고 등록화자의 특성을 학습하는 데 MLP를 사용하는 화자증명 시스템을 사용하여 제안한 방법을 실험한 결과 인식률의 향상을 확인하였으며, 아울러 부적절한 성량의 음성을 거부함으로써 사용자가 느끼는 체감 인식률을 높이는 결과를 얻었다.

참고문헌

[1] J. P. Eatock and J. S. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes," ICASSP, Vol. 1, pp. 133~136, 1994.

[2] D. P. Delacretaz and J. Hennebert, "Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs," ICASSP, Vol. 2, pp. 777~780, 1998.

[3] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.

[4] A. L. Higgins and R. E. Wohlford, "Speaker Verification Using Randomized Phrase Prompting," Digital Signal Processing, Vol. 1, pp. 89~106, 1991.

[5] N. Fakotakis and J. Sirigos, "A High Performance Text Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets," ICASSP, Vol. 2, pp. 661~664, 1996.

[6] R. M. Stern, et. al., "Signal Processing for Robust Speech Recognition," Automatic Speech and Speaker Recognition Advanced Topics, Kluwer Academic Publishers, pp. 357~384, 1996.

[7] D. Tapias, et. al., "On the Characteristics & Effects of Loudness during Utterance Production in Continuous Speech Recognition," ICASSP, Vol. 1, pp. 89~92, 1999.

[8] C. Becchetti and L. P. Ricotti, Speech Recognition, John Wiley & Sons, 1999.

[9] P. Cristea and Z. Valsan, "New Cepstrum Frequency Scale for Neural Network Speaker Verification," ICECS, Vol. 3, pp. 1573~1576, 1999.

[10] L. Fausett, Fundamentals of Neural Networks, Prentice Hall, 1994.