화자인식에 효과적인 특징벡터에 관한 비교연구

박 태 선, 김 상 진, 문 광, 한 민 수
한국정보통신대학원대학교

A study on Effective Feature Parameters Comparison for Speaker Recognition

TaeSun Park, Sang-Jin Kim, Moon Kwang, Minsoo Hahn
Information and Communication University
E-mail : {tspark, sangjin, mlight, mshahn}@icu.ac.kr

Abstract

In this paper, we carried out comparative study about various feature parameters for the effective speaker recognition such as LPC, LPCC, MFCC, Log Area Ratio, Reflection Coefficients, Inverse Sine, and Delta Parameter. We also adopted cepstral liftering and cepstral mean subtraction methods to check their usefulness. Our recognition system is HMM based one with 4 connected-Korean-digit speech database. Various experimental results will help to select the most effective parameter for speaker recognition.

Ⅰ. 서 론

화자 인식은 생체 인식 시스템 중의 하나로, 음성 파형으로부터 수집된 정보를 이용하여 발성 화자가 누구인지 자동적으로 인식하는 것이다. 개인이나 특정 단체의 정보 보안을 위해서 사용자의 확인 과정이 필요한 음성 다이얼링(voice dialing), 전화 네트워크를 이용한 은행거래(banking transactions over a telephone network), 전화 쇼핑(telephone shopping), 데이터베이스 접근 서비스(database access services) 등 다양한 서비스에 화자인식이 응용될 수 있다.

본 연구의 목적은 기존의 다양한 특징벡터 추출기법을 적용하고 이들에 대한 비교실험을 수행하여 화자인식에 보다 효과적인 특징벡터를 찾는데 있다. 다루어진 특징벡터 및 전처리 기법은 다음과 같다.

1) LPC / LPCC / MFCC
2) Log Area Ratio / Inverse Sine / PARCOR
3) Dynamic Feature (Delta Parameter)
4) Cepstral Liftering
5) Cepstral Mean Subtraction

즉, 위의 각 특징벡터 및 전처리 방법을 음성데이터베이스에 대해 적용하고 화자인식 결과를 비교하여 가장 좋은 결과를 보인 특징벡터를 제시하는 것이 목적이다.

2장에서 각각의 특징벡터 추출에 대해 간략한 설명을 하였으며, 3장에서 사용된 음성 데이터베이스 및 기본 인식 시스템을 설명하고, 4장에서 실험 및 결과를 제시하고 분석한 뒤, 5장에서 결론을 맺겠다.

Ⅱ. 화자인식에 효과적인 특징벡터

2.1 LPC (Linear Prediction Coefficient)

선형 예측 분석을 사용하면 음성 신호, 혹은 음성 스펙트럼이 가진 특성을 상대적으로 적은 수의 파라미터 만으로 정확하게 표현할 수 있다는 장점이 있다. 또한 선형 예측 분석 자체가 그리 많은 연산량을 요구하지 않는 장점도 갖는다. 시간 t 에서의 음성 샘플을

x_t라고 하자. 이 때, 선형 예측 분석법에서는 현재의 음성 샘플을 이전 예측값과 실제값 사이의 차이를 p 개의 샘플로부터 예측한다. 이 때, 예측값과 실제값과의 차이를 e_t 라고 하면 아래의 식이 성립한다.

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + e_t \quad (1)$$

여기서 오차 e_t의 값을 최소화하는 p개의 a_i 값들을 선형예측계수(LPC)라고 한다.

2.2 LPC Cepstrum

LPC 켑스트럼 계수 $c(n)$은 선형예측계수 a_i로부터 다음의 관계식을 이용하여 계산한다[1].

$$c(n) = a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) c(i) a_{n-i} \quad (2)$$

여기서 $c(n)$은 무한대의 구간에서 존재하며, n의 값이 커질수록 계수는 $\frac{1}{|n|}$ 에 비례해서 작아지고, 대신에 n의 값이 작을수록 계수의 중요성은 커진다.

2.3 PARCOR(Partial Autocorrelation Coefficient)

선형 예측 계수는 예측 차수에 따라 그 값이 변하는 단점을 가지고 있다. 또한 선형 예측 계수를 표현하는 비트(bit)가 적을 경우에 선형 예측 계수 합성 시스템을 구성할 때 불안정한 발진 현상을 일으킬 수 있다. 이런 현상을 최소화하기 위해 등장한 것이 부분 자기상관 계수(PARCOR)이다. 선형 예측 계수는 각 샘플들 사이의 상관관계로부터 출발하지만 부분 자기상관 계수는 각 샘플의 나머지 값, 즉 예측된 값과 실제값 사이의 오차들끼리의 상관관계로부터 출발한다[4]. 반사계수(Reflection Coefficients)와 동일한 의미로 사용되며 LPC를 구하는 과정에서 얻을 수 있다. 반사계수는 다음과 같은 식으로 정의된다.

$$k_n = \frac{A_{n-1} - A_n}{A_{n-1} + A_n} \quad (3)$$

2.4 LAR(Log Area Ratio)

LPC나 PARCOR계수는 주파수 민감도(spectral sensitivity)가 일정하지(flat) 않다. 즉, unity 부근에서 양자화 오차를 줄이기 위해서는 더 많은 bit을 요구한다. LAR계수 g_i 는 아래와 같은 공식을 이용하여 PARCOR 계수로부터 구할 수 있으며, 특히 양자화에 적합한 파라미터로 알려져 있다[4].

$$g_t = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1-k_i}{1+k_i}\right], \quad 1 \le i \le p \quad (4)$$

2.5 IS(Inverse Sine)

IS계수 s_i 도 LAR 계수와 마찬가지로 주파수 민감도가 비교적 일정한 파라미터이며 양자화에 적합하다고 알려져 있다. 아래 식을 이용하여 PARCOR 계수로부터 구할 수 있다[4].

$$s_t = \sin^{-1}(k_i), \quad 1 \le i \le p \quad (5)$$

2.6 Cepstral Liftering

노이즈와 같은 원하지 않는 성분은 줄이거나 제거하지만, 포먼트 구조와 같은 필수성분은 그대로 유지하여, 거리 측정시 오류를 야기할 수 있는 민감도를 줄여주는 효과를 얻기 위하여 liftering을 사용한다[5].

$$c'_t = \left(1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right) c_t \quad (6)$$

여기서 L값은 4kHz 대역폭의 음성의 경우 전형적으로 10~16의 값을 가진다[5].

2.7 MFCC (Mel-Frequency Cepstral Coefficient)

인간의 외이/중이의 주파수 특성을 모델링하기 위하여 고대역 필터링을 한다. 이는 입술에서의 방사에 의하여 20 dB/decade로 감쇄되는 것을 보상하게 되어 음성으로부터 성도 특성 만을 얻게 된다. 또한 청각시스템이 1 kHz 이상의 스펙트럼 영역에 대하여 민감하다는 사실을 어느 정도 보상하게 된다[5].

2.8 Dynamic Feature (Delta Parameter)

음성 인식 시스템의 성능은 기본적인 확률 파라미터에 시간 축에서의 미분계수를 부가함으로서 인식률을 향상시킬 수 있다[6].

$$d_t = \frac{\sum_{\Delta=1}^{2} \Delta \times (c_{t+\Delta} - c_{t-\Delta})}{2\sum_{\Delta=1}^{2} \Delta^2} \quad (7)$$

여기서, d_t 는 시간 t 에서의 delta coefficient이다.

2.9 CMS (Cepstral Mean Subtraction)

CMS는 전체 음성구간에 대하여 켑스트럼의 평균을 구하고 이를 차감하여 채널의 효과를 제거하는 방법이다. 음성인식 시스템에 입력된 음성은 마이크를 통해 A/D변환 및 양자화를 통해 시스템에 도달할 때까지 채널왜곡이 일어난다. CMS는 채널왜곡을 보상하는데 효과적이라고 알려져 있으며, 적용할 음성구간이 비교적 길어야 효용성이 높아진다.

$$c_{cms,t} = c_{Y,t} - \frac{1}{T}\sum_{n=1}^{T} c_{Y,n} \ , \ 1 \le t \le T \qquad (8)$$

여기서, T는 전체 프레임의 수이며 c_t 는 t번째 프레임의 켑스트럼이다.

Ⅲ. 음성 데이터베이스 및 인식 시스템

3.1 음성 데이터베이스

본 논문에서 사용한 음성 데이터베이스는 국어공학센터의 주관으로 원광대에서 제작된 것으로 방음 부스에서 녹음되었고, 16 kHz로 추출되어 16 bit로 양자화되었다. 한국어 숫자음 "공", "일", "이", "삼", "사", "오", "육", "칠", "팔", 그리고 "구"의 조합으로 사연숫자를 발성하였다. 실험에 사용된 음성샘플은 총 2,800개로 남성화자 및 여성화자 각각 10명이 35개의 사연숫자를 4회씩 발성하였다. 훈련용으로 남성화자 및 여성화자가 각각 3회 발성한 2,100개의 음성샘플을 사용하였고, 훈련에 사용되지 않은, 남성화자 및 여성화자가 각각 1회 발성한 700개의 음성샘플을 테스트용으로 사용하였다.

3.2 인식 시스템

HMM 기반의 인식기를 사용하였다. 음성 특징 파라미터는 LPC, LPCC, PARCOR, LAR, IS, MFCC 각각에 대하여 12차 및 log energy, delta, delta-delta 의 조합으로 구성되었다. 먼저 0.97로 pre-emphasis를 한 뒤, 20 msec 의 해밍 윈도우를 이용하여 10 msec씩 이동하며 추출하였다. 기본 모델링은 whole-word을 사용했으며, word별 상태 수는 7개를 사용하였다[4].

표 1. log Energy 사용에 따른 인식결과 비교

특징벡터	w/o logE	with logE
LPC	59.68	64.82
Ref.Coeff.	78.32	81.00
LAR	79.25	81.68
IS	78.86	81.61
LPCC	**86.79**	**89.46**
MFCC	87.36	88.39

표 2. Dynamic Feature의 사용에 따른 인식결과 비교

특징벡터	w/o Delta	with Delta	with D-Delta
LPC	**64.82**	64.00	58.57
Ref.Coeff.	81.00	**82.07**	80.75
LPCC	89.46	**91.11**	89.32
MFCC	88.39	**90.21**	89.18

Ⅳ. 실험 및 결과

각각의 특징벡터에 대하여, log Energy, Cepstral Liftering의 사용 유무, Delta 및 Delta-Delta와 같은 Dynamic Feature의 사용 유무, CMS (Cepstral Mean Subtraction)의 사용 유무에 따른 화자인식 성능을 비교하였다.

표 1은 각 12차의 특징파라미터에 대하여 log Energy를 부가한 경우에 대한 결과를 나타낸다. 이 결과를 보면 12차의 특징파라미터만을 사용한 것보다 log Energy를 더하여 사용한 것이 전체적으로 보다 좋은 인식률을 보임을 알 수 있다. 또한 LPC보다 Reflection Coefficients나 Log Area Ratio 등 LPC에 변화를 준 특징벡터들이 더 좋은 인식결과를 보였다. 그리고, 음성인식에 일반적으로 많이 사용되는 MFCC보다 LPCC가 화자인식에 더 좋은 결과를 가져옴을 볼 수 있다. 이하 실험에서는 log Energy를 포함한 13차 특징벡터를 사용하여 실험하였다.

표 2는 Dynamic Feature인 Delta 및 Delta-Delta의 사용 유무에 따른 결과를 나타내었다. 이 결과를 보면 12차의 각 특징파라미터와 log Energy에 대해 Delta 특징을 사용한 경우가 사용하지 않은 경우나 Delta-Delta 특징을 사용한 경우보다 더 좋은 인식률

표 3. Cepstral Liftering 사용에 대한 인식결과 비교

특징벡터(26차)	w/o Liftering	with Liftering
LPC	64.00	64.00
Ref.Coeff.	82.07	82.07
LPCC	91.11	91.11
MFCC	90.21	90.21

표 4. CMS 사용에 대한 인식결과 비교

특징벡터(26차)	w/o CMS	with CMS
LPC	**64.00**	60.07
Ref.Coeff.	**82.07**	79.64
LPCC	**91.11**	88.54
MFCC	**90.21**	86.61

을 보임을 알 수 있다. 이하 실험에서는 Delta 특징을 포함한 26차 특징벡터를 사용하여 실험하였다.

표 3의 결과를 보면 Cepstral Liftering의 사용한 경우와 그렇지 않은 경우를 비교할 수 있다. L값은 14차의 켑스트럼에 대해 가장 좋은 결과를 보인다는 22값을 사용하였다[6]. 표에 제시된 4종류의 26차 특징벡터에 대해 Cepstral Liftering을 사용한 결과, 사용하지 않았을 때와 비교하여 인식율 향상이 없었다. 즉, Liftering은 화자간 변별력의 증가에 아무런 영향을 미치지 않는다고 판단된다. 이하 실험에서는 Liftering을 적용하지 않았다.

표 4의 결과를 보면 CMS를 적용한 경우와 그렇지 않은 경우를 비교할 수 있다. 표에 제시된 4종류의 26차 특징벡터에 대해 CMS를 적용하였다. 적용한 결과 인식 성능이 떨어지는 것을 볼 수 있다. 대표적인 채널 정규화 방법인 CMS가 화자인식 향상에 기여하지 못함을 나타내며 이는 화자간 변별력을 감소시키는 효과를 보이기 때문이다[2].

V. 결 론

다양한 특징파라미터 및 전처리에 대하여 화자 인식률 변화를 확인할 수 있었다. 가장 좋은 인식률을 보인 조합은 LPCC + log Energy + Delta를 사용한 경우였다. Cepstral Liftering은 화자 인식율 향상에 도움을 주지 않았으며, CMS 전처리는 화자간 변별력을 감소시키는 결과를 가져와 인식율이 감소하였다.

PLP나 LSF와 같은 특징벡터에 대해서 실험중이며, 사무실 잡음이나 전화망 잡음환경에 대해 확장 실험을 준비 중이다. 또한, F-ratio와 같은 켑스트럼의 가중 적용 등, 화자간 변이를 극대화할 수 있는 방법을 연구 중이다.

참고문헌

[1] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol.55, no.6, pp.1304-1312, 1974.

[2] R.J. Mammone, X. Zhang, R.P. Ramachandran, "Robust Speaker Recognition, A Feature-based Approach," *IEEE signal processing mag.*, pp58-71, Sep. 1996.

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

[4] A.M. Kondoz, *Digital Speech*, John Wiley&Sons, 1994.

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

[6] S. Young, et al., *The HTK Book (for HTK version 3.1)*, Entropic Ltd., 1995-2001.