PSOLA 방식을 이용한 화자인식 시스템의 처리시간 단축에 관한 연구

박 현 영, 서 지 호, 배 명 진
숭실대학교 정보통신공학과
전화 : 02-824-0906 / 핸드폰 : 016-237-9231

A Study on Reduction of the Processing time of Speaker Recognition using the PSOLA Method

Hyun-Young Park, Ji-Ho Seo, Myung-Jin Bae
Dept. of Information and Telecommunication Engr, Soongsil University
E-mail : phy717@hotmail.com

Abstract

화자인식은 음성의 특성을 이용해서 화자의 신원을 확인하는 기술이다. 이러한 기술은 등록된 화자집단중 화자를 식별하는 화자식별(speaker identification)과 지금 발성한 화자만을 비교하여 확인하는 화자확인(speaker verification)이 있다. 이러한 화자인식은 음성에 내재되어 있는 화자정보를 추출하여 개인을 확인하는 기술로 전화망을 통한 서비스가 확산되어 가고 있는 현대사회에 가장 효과적인 기술중 하나이다. 또한 PDA를 이용한 증건거래 시스템등 현대사회에서는 실시간으로 화자인식이 이루어져야한다.

본 논문에서는 이와 같이 실시간 화자인식을 위한 처리시간 단축에 관하여 연구하였다. 처리시간 단축을 위하여 우선 피치주기 단위로 음성 파형을 분해한 다음 분해된 피치 단위에 윈도우 함수를 곱해서 단구간 신호의 열로 만들고 분해된 단위를 조절하는 PSOLA 합성방식을 이용하여 인식 시스템의 전처리단을 재구성 하였다. 이와 같은 방식으로 제안한 인식시스템의 처리시간, 인식률을 기존의 화자인식 시스템과 비교하였다.

1. Introduction

화자의 특징을 이용하여 화자의 신원을 파악하는 기술인 화자 인식은 전화망이 설치된 이후부터 그 연구가 활발히 진행되었으며, 컴퓨터를 이용한 자동 화자인식(ASR: Automatic Speaker Recognition) 기술은 1976년 Atal의 연구를 시작으로 활발해지기 시작했으며, 부분적으로 상용화되기 시작했다. 이는 화자의 음성에서 나타나는 특성이 화자별로 두드러진 특성과 포괄적인 특성을 동시에 지니기 때문이다. 이러한 화자인식은 분실이나 도난 등과 같은 기존의 신원을 확인하는 방법의 문제점을 해결할 뿐만 아니라 처리시간, 원격자 확인등의 경우 가장 효과적인 기술이다. 본 논문에서는 이러한 화자인식 시스템의 처리시간 단축을 위하여 기존의 화자인식 시스템에 PSOLA 합성방식을 적용하여 인식 시스템의 전처리단을 재구성하였다. 논 논문에서는 제안한 인식 시스템 전처리단에서 PSOLA 합성방식을 사용하여 15%, 30%, 45% 로 각각 음성을 압축 합성한 후 화자인식 처리 과정을 수행하여 기존의 화자인식 시스템과 처리시간, 인식률을 각각의 경우에 대하여 비교 하였다. 2장에서는 화자인식 시스템에 일반적인 개요에 대해서 설명하고, 3장에서는 본 논문에서 제안한 PSOLA 합성방식을 적용하여 전처리단을 재구성한 화자인식 시스템에 대하여 실명하였고,. 4장에서는 실험결과를 보였으며 마지막으로 5장에서는 결론을 맺었다.

2. Speaker Recognition System

2.1 Classification of Speaker Recognition

화자인식은 방법상의 분류에 따라 크게 다음과 같이 분류된다. 발성화자가 누구인지를 구분해 내는 화자식별과 본인 여부를 판별하는 화자 확인으로 나누며, 인식에 사용하는 문장의 고정 여부에 따라 문장 고정형과 문장 독립형으로 나눈다. 화자확인(speaker verification)은 입력된 음성이 본인의 것인 지의 여부를 판정하는데 비해, 화자식별(speaker identification)의 경우는 입력된 미지의 음성이 이미 등록된 여러 명의 화자 중 어떤 화자에 위해 발성된 음성인지를 판정하는 것을 말한다.

따라서 화자 식별율은 화자 수에 비례하여 정확도가 감소하므로 화자확인에 비해 어려우며, 실제 응용에 있어서도 비협조적인 화자를 대상으로 하는 경우가 많으므로 화자의 정확한 판단에 어려움이 많다. 이에 비해 화자 확인은 대상화자의 허용 여부만을 판정하므로 오류율이 화자 수와 무관하지만 화자간 변이와 화자내 변이의 경계를 결정하는 한계값에 따라 본인 거부율과 사칭자 허용율이 달라지므로 한계값의 결정이 식별율을 좌우한다. 화자확인 기술은 음성을 이용한 신분확인 및 음성인식 기술과 조합하여 은행거래 등의 전화를 이용한 원격처리에, 화자식별은 범죄수사 등에서 용의자를 가려내는 일에 이용될 수 있다.

화자가 발성할 문장이 고정되어 있는 화자인식 시스템을 문장 의존(text dependent)형이라 하고, 문장이 정해져 있지 않고 자유롭게 발성하는 경우를 문장 독립(text independent)형이라 한다. 전자의 경우 음운성에 기반을 둔 공통적 특징의 개인 차이를 평가하게 되므로 미리 저장된 표준패턴과 비교를 수행하는 음성인식과 거의 동일한 방법을 사용한다. 또한, 선정 어휘에 따라 인식율이 영향을 받으므로 화자의 특성이 잘 나타나는 모음, 비음 등의 음운이 균형을 이룬 어휘 집합의 선택이 필요하다. 이에 비해 문장 독립형의 경우, 임의로 자유롭게 발성된 음성신호로부터 음운 정보를 제거한 화자 정보만을 사용해야 하므로 전자보다는 어려운 방법이다. 이 방법은 많은 음성 자료가 필요하며, 음성으로부터 음운정보를 제거하기 위해 통계 분석을 하거나, 음소 단위로 분리하여 각 음소와 관계되는 화자정보를 얻어내는 방법을 사용한다.

화자인식의 기본 동작의 요소는 그림 2-1에 보였다. 테스트할 화자의 발성음성은 먼저 특징 파라미터의 추출을 위해 분석된다. 추출된 파라미터는 기존에 저장되어 있는 학습용 파라미터 모델과 비교하는데 화자식별의 경우 모든 파라미터 모델간의 비교가 이루어지게 된다. 화자확인의 경우, 테스트 파라미터와 학습 파라미터의 비교는 단지 요구되는 신원에 대응하는 모델에만 이루어지게 된다. 비교에 의해 얻어지는 점수를 가지고서 화자식별과 화자확인에 대한 결정이 내려진다.

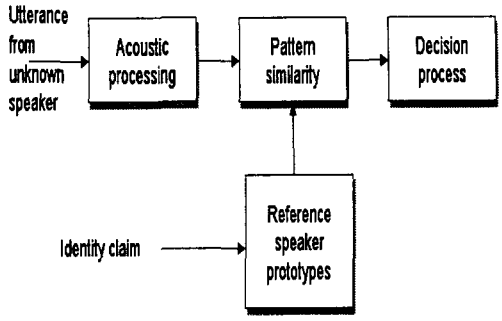

그림 2.1. 화자인식 기법의 기본 구성도

2.2 Type of Recognition Techniques

화자인식에 사용되는 DTW, VQ, HMM 그리고 GMM 등 여러 가지 많은 모델들이 있다. 텍스트 종속 시스템에서 좋은 효과가 있는 DTW는 텍스트 독립 시스템에는 적용하기 힘들다. VQ는 구조가 쉽고 좋은 성능을 갖으며 짧은 음성 신호로도 다른 인식모델과 대등한 성능을 낼 수 있다는 장점을 가진다.

이것은 VQ 코드북이 여러 개의 파라미터로 구성되기 때문에 테스트 음성신호가 짧아서 생기는 변이를 수용할 수 있기 때문이다. 또한, 사용어휘에 제한을 받지 않고 화자모델의 크기가 작아 실용 시스템에 유용한 방법이다. 그리고 VQ는 Speaker Identification과 Speaker Verification에 대해 공통구조를 가진다. 장기간에 걸친 음성신호 파라미터의 일시적인 변이는 상태와 상태 사이를 통계적 마코프 천이(stochastic markovian transition)에 의해서 표현할 수 있다.

HMM은 좋은 인식율을 보이나 작은 데이터를 이용할 경우에는 VQ를 기반으로 하는 방법이 연속적인 HMM방법보다 환경에 강인한 인식률을 보인다. 화자는 GMM에서 몇 개의 가우시안 혼합식으로 표현된다. GMM에 의존하는 한 화자의 개별 가우시안 요소는 일반적인 음성을 효과적으로 인식하는데 특징이 있는 스펙트럼 구조로 표현될 수 있다. 그러나 이 시스템은 큰 학습 데이터, 학습 데이터를 위해 적어도 30초 이상의 발성음을 필요로 한다. 이렇듯 HMM은 좋은 성능을 보이나 작은 학습 데이터를 이용할 경우 문제가

발생한다. 따라서 화자인식 시스템에 VQ모델을 사용하였다. 또한 화자인식의 근본적인 난점으로는 음성에서의 음운정보와 화자정보 분리의 어려움, 사칭자의 거부, 시간변화에 따른 인식율 저하등이 있다.

3. Proposed Algorithm

3.1.The PSOLA Synthesis Method

PSOLA 합성방식은 먼저 원래의 음성 파형을 피치 주기 단위로 분해한 다음 분해된 피치 단위에 윈도우 함수를 곱해서 단구간ST(Short-Term)신호의 열로 만든다. 분해된 단위의 운율조절을 하고 이렇게 조절된 단위로부터 음성을 합성한다

원래 음성 파형이 유성음인 경우에는 피치단위로 분해한 다음 윈도우 함수를 곱하여 ST신호의 열로 만든다. 무성음인 경우에는 10ms의 주기로 일정하게 분석한다. 분석 윈도우 함수에는 다음과 같은 Hanning, Hamming, Blackman 등의 형이 쓰인다. 이런 윈도우 함수를 원래의 음성 샘플에 곱함으로써 다음 식(3.1)과 같은 피치 단위로 분해된 샘플열들을 얻는다[4].

$$S_{analysis}(n) = W_{analysis}(m-n)S(n) \qquad (3.1)$$

$S_{analysis}(n)$: 피치주기 단위의 ST 신호

$W_{analysis}(n)$: 분석 윈도우 함수

m : m번째 피치

$S(n)$: 원 음성 파형

분석과정에서의 ST신호의 열은 원래의 음성 샘플의 피치단위로 배열되어있다. 따라서 피치를 변경하기 위해서는 이 간격들을 변경할 피치 간격들로 재배열하면 된다. 다음 식(3.2)은 피치가 변경된 신호를 나타낸 것이다.

$$S_{synthesis}(n) = S_{analysis}(n - m_a) \qquad (3.2)$$

$S_{synthesis}(n)$: 피치가 변경된 ST신호

m_a : 변경할 피치 간격

따라서 피치를 높일 때는 ST 신호의 간격을 작게 배열하고, 피치를 낮출 때는 ST신호의 간격을 크게 배열하면 된다. 하지만 이런 순차적인 배열사이에서 정확한 피치 동기화를 유지하는 것이 중요하다. 이렇게 재배열된 ST신호에서 겹쳐지는 부분을 더해주면 된다.

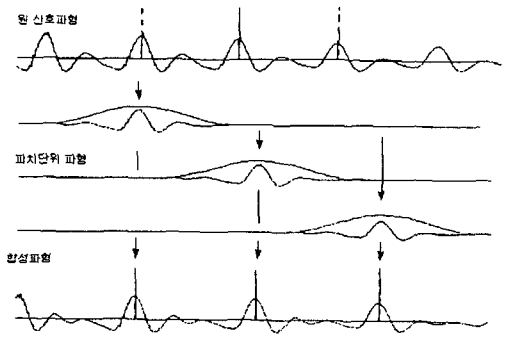

그림 3.1 PSOLA 합성법

3.2. Speaker Recognition using the PSOLA

본 논문에서는 기존의 화자인식 시스템에 앞서 설명한 PSOLA 합성방식을 이용하여 전처리단을 새로이 구성하였다. PSOLA 방식에 의하여 압축된 음성이 입력음성으로 들어가고 합성된 음성에서 특징 파라미터를 추출하여 기존에 저장되어 있는 학습용 파라미터 모델과 비교하여 인식 결과를 내리는 방식이다.

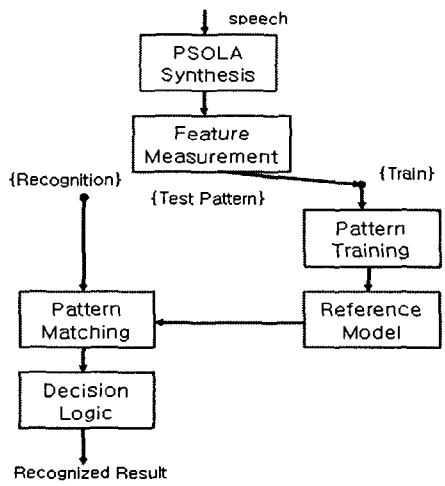

그림 3.2. 제안한 화자인식 구성도

4. Experiment and Results

본 논문의 성능을 평가하기 위하여 일반 IBM-PC 1.7GHz를 사용하여 11.025KHz 샘플링과 16bit 양자화를 수행하였으며 14차의 MFCC를 사용하였고 사용한 인식 알고리즘은 DTW이다. 본 논문에서 사용한 데이터 베이스를 위하여 일반 사무실 환경에서 65명의 등록자가 각각 1번씩 2회 발성(130개)하였다. 녹음에 사

용된 음성 시료는 /인수네 꼬마는 천재 소년을 좋아한다./ 이다.

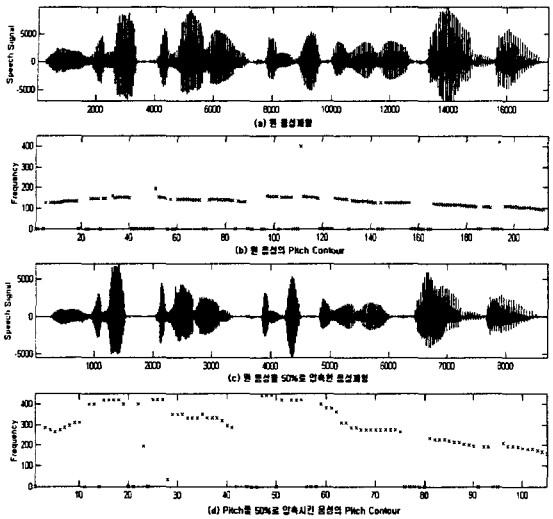

그림 4.1. PSOLA 합성 방식을 이용하여 음성을 50% 압축시킨 경우의 예

그림 4.1은 원래의 음성파형과 Pitch countour 그리고 PSOLA 합성법을 사용하여 50% 압축시킨 음성파형과 Pitch countour를 나타낸 것이다.

본 논문에서 제안한 방법의 효과를 알아보기 위해서 기존의 화자인식 시스템과 본 논문에서 제안한 알고리즘을 적용한 화자 인식 시스템의 전체 인식률과 65명 data의 평균 계산량을 비교해 보았다. 제안한 알고리즘의 계산량은 PSOLA 합성시 소요되는 계산시간을 포함한다. 제안한 알고리즘의 경우 25%, 50%, 75%로 압축된 음성에 대한 인식률과 평균 계산량은 표 4.1과 같다.

표 4.1 기존 DTW 화자인식 시스템과 제안한 알고리즘의 인식률과 계산량 비교

	인식률	계산량 (처리시간)	계산량 감소율
기존 DTW	87.8%	1.24초	-
25% 압축	87.4%	0.76초	38.7%
50% 압축	86.9%	0.53초	57.3%
75% 압축	68.3%	0.27초	78.2%

PSOLA 전처리단을 사용하여 25%와 50% 압축한 경우, 기존의 DTW 화자인식 시스템과 비교하여 인식률 측면에서는 유사한 결과를 나타내지만 계산량 측면에서보면 상당한 감소 효과가 있음을 알 수 있다. 75%로 압축한 경우에는 계산량 측면에서 78.2%로 많은 감소 효과를 가져왔지만 인식률에서 거의 20% 정도 저하가 발생하게 되므로 실제 응용에서는 적용하기 힘들다.

5. Conclusion

화자인식은 음성의 특성을 이용해서 화자의 신원을 확인하는 기술이다. 이러한 화자인식은 비교적 적은 수의 단어에 대한 실시간적인 인식에 적합한 DTW 알고리즘을 주로 사용한다. 그러나 DTW 화자인식 시스템은 시간축 상의 변동패턴을 모두 계산해야 하며, 화자의 수나 기준음성의 수에 비례하는 연산량이 필요하게 되고, 이는 실시간적인 처리를 요하는 일반적인 화자인식 시스템의 치명적 단점이라 할 수 있다. 따라서 본 논문에서는 이를 개선하기 위하여 기존의 화자인식 시스템에 PSOLA 합성방식을 적용하여 인식 시스템의 전처리단을 재구성하였다.

실험결과 본 논문에서 제안한 전처리단을 이용하여 음성을 25%, 50%로 압축한 결과 기존의 DTW 화자인식 시스템과 비교하여 유사한 인식률에 계산량 측면에서는 상당한 향상을 가져옴을 알 수 있었다. 이와같이 본 논문은 실시간 처리를 요하는 화자인식 시스템에 적용하여 상당한 성능의 향상을 가져올 것이다.

감사의 글

본 연구는 한국과학재단 특정기초연구 (과제번호 R01-2002-000-00278-0)의 지원에 의하여 이루어 졌습니다.

Reference

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

[2] A.M. Kondoz, *Digital Speech -Coding for Low Bit Rate Communications Systems*, John Wiley and Sons, 1994.

[3] B. E. Caspers, B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation," J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, 1983.

[4] 박 원 "음성신호의 실시간 운율 조절에 관한 연구." 숭실대학교 석사학위 논문 2001년 6월

[5] 배재옥 "인지 가중 필터를 이용한 화자 인식의 성능향상에 관한 연구." 숭실대학교 석사학위 논문 1998년 12월