음성인식에서 화자 내 정규화를 위한 진폭 변경 방법

An Amplitude Warping Approach to Intra-Speaker Normalization for Speech Recognition

김 동 현* 홍 광 석**
Dong-Hyun Kim Kwang-Seok Hong

요 약

기존의 성도 정규화 방법은 화자 간 정규화의 정확성을 개선하기 위한 매우 좋은 방법이다. 본 논문에서는 피치 변경 발성에 기반을 둔 새로운 화자 내 warping 인수 추정 방법을 제안한다. 화자 내 피치 변경 발성은 성문과 성도에 의해 발생되는 음성의 음향학적 차이 때문에 음성의 특징 공간 분포는 다르게 나타날 것이다. 발성의 변동은 frequency 성분과 amplitude 성분의 두가지 유형이 있다. 성도 정규화는 화자 간 정규화 방법들 중에서 주파수 정규화 방법이다. 여기에서는 화자 내 정규화를 위하여 진폭 변동을 정규화하는 방법을 제안한다. 참조 피치와 입력 피치의 역비례 계산에 의해서 진폭 warping 인수를 결정하는 것이 가능하다. 성능 평가를 위한 인식 실험 결과 숫자와 단어 인식에서 0.4%~2.3% 정도의 인식 오류가 감소되었다.

Abstract

The method of vocal tract normalization is a successful method for improving the accuracy of inter-speaker normalization. In this paper, we present an intra-speaker warping factor estimation based on pitch alteration utterance. The feature space distributions of untransformed speech from the pitch alteration utterance of intra-speaker would vary due to the acoustic differences of speech produced by glottis and vocal tract. The variation of utterance is two types: frequency and amplitude variation. The vocal tract normalization is frequency normalization among inter-speaker normalization methods. Therefore, we have to consider amplitude variation, and it may be possible to determine the amplitude warping factor by calculating the inverse ratio of input to reference pitch. As the recognition results, the error rate is reduced from 0.4% to 2.3% for digit and word decoding.

Key words : speaker normalization, speaker adaptation, vocal tract normalization, speech recognition

1. 서 론

성대의 성문에서는 목소리의 피치를 제어한다. 반면에 성도는 성도의 포먼트를 통하여 모음들을 결정하고 또한 자음들을 조음시킨다. 피치와 포먼트 성분들은 음성 신호에서 거의 서로 독립적이다.

화자들간의 성도 모양 변이로 인한 음성인식 성능 저하를 줄이기 위한 노력으로 화자 정규화 방법인 주파수 warping 기법이 연구되었다. 또한 화자 상호간의 효과를 줄이기 위한 음성신호의 파리미터 성분 표현들을 정규화하기 위한 기술들이 연구되었다. 여기에서는 주로 화자들 간의 포먼트 위치 변동을 보상하기 위하여 선형과 비선형 주파수 warping 함수들을 사용하여 정규화가 수행되었다. 이러한 방법들은 각 화자의 실제 성도 모양에 해당하는 포먼트 위치를 추정하고 이러한 차이들을 위한 보상에 의해서 해결하려는 시도가 되었다[1]. Hidden Markov Model에서 출력 확률로써 가우시안 mixture들을 사용할 때 가장 중요한 문제들 중의 하나는 다양한 화자 의존적인 스케일 인수들은 mixture 분포들의 다형질에 의해 모델이 구성되는 경향이 있다[2].

* 준회원 : 성균관대학교 정보통신공학부 석사과정
super1621@daum.net(제 1저자)
** 정회원 : 성균관대학교 정보통신공학부 부교수
kshong@skku.ac.kr(공동저자)

또한, 화자 내 인수도 음성인식을 위하여 중요하다. 화자 간 정규화는 화자 간 정규화를 필요로 하는 화자 적응을 위해 성도 정규화(Vocal Tract Normalization)에 기반을 두었다. 화자 내 warping 인수를 선택하기 위해 본 논문에서 제안된 방법은 피치 변경 발성에 기반을 둔다. 이는 감정의 상태에 따른 피치 변경 발성에서 변동 보상에 의해 화자 내 음성의 변동을 줄이기위해 시도하였다. 본 논문에서 화자 간 정규화와 화자 내 정규화의 실험은 SKKU-1(SKKU: sungkyunkwan University) 음성 DB를 사용하여 수행하였다.

2. 화자간 정규화를 위한 주파수 축 정규화

VTN의 제일 중요한 개념은 인식 과정에서 음향학적 벡터들에서 화자 의존적 가변성을 제거하기 위해 각각의 화자를 위한 음향학적 벡터들의 주파수 축을 정규화한다[2]. 주어진 소리의 발성에 대하여 스펙트럼의 formant정점들의 위치는 성도의 길이에 반비례한다. 성도의 길이는 대략 13cm에서 18cm까지 다양하다. 포먼트 중간 주파수는 화자들 사이에 25%만큼 다양하게 변화한다. 이런 변이 요소들은 화자 종속부터 화자 독립 음성 인식 성능의 주요한 저하 요소이다. 최적의 warping factor는 $0.88 \le \alpha \le 1.12$사이에서 균등한 간격으로 13개 인수들의 검색으로 얻어진다. α의 범위는 어른들에서 발견되는 성도 길이들에서 대략 25%범위의 변화를 반영하기 위해 선택된다[1].

인식에서 최적의 주파수 warping 크기의 결정을 위해서 많은 방법들이 제안되었다. 음성인식에서 음향학적인 벡터들의 sequence는 시간 $t=1,...,T$에 걸쳐서 관찰된다. 즉,

$$X = x_1 \ldots x_t \ldots x_T \tag{1}$$

각각의 가정한 단어 sequence W를 위해 적절한 참조모델 parameters θ를 가지고 모델 $p(X|W;\theta)$ 분포를 가정한다. 화자 적응 음향학적 모델링에서 어떤 화자의 특성 파라미터 α분포에 다음과 같은 의존성이 있다는 것을 가정한다.

$$p(X|W;\theta,\alpha) \tag{2}$$

이는 전형적으로 두가지 변수의 변환으로 구분하여 설명된다.

□ 모델 파라미터 θ의 변환:

상호 화자의 특정한 파라미터 α의 각각의 값을 위해 정규화 되지않은 모델 파라미터 θ를 정규화한 파라미터 θ^{α}로 변환한다.

$$\theta \rightarrow \theta^{\alpha} \tag{3}$$

따라서 분포는

$$p(X|W;\theta,\alpha) = p(X|W;\theta^{\alpha}) \tag{4}$$

이다.

□ 관측 벡터 X의 변환:

이것은 음향학적 벡터들의 매핑으로 공식화 할 수 있다.

$$X \rightarrow X^{\alpha} \tag{5}$$

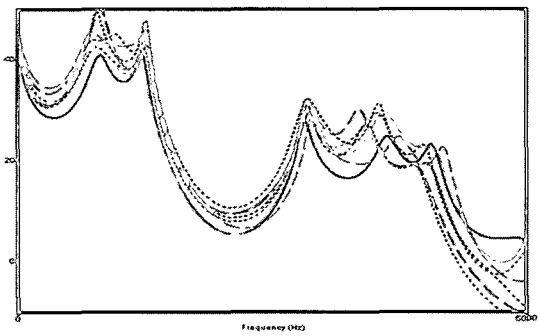
분포는

$$p(X|W;\theta,\alpha) = p(X^{\alpha}|W;\theta) \tag{6}$$

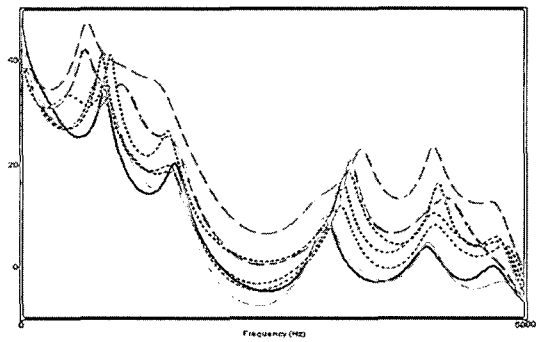
이다.

3. 화자 내 정규화를 위한 진폭 정규화

화자 내 정규화에서 진폭 warping 방법을 수행하는데 사용하는 처리과정에 대하여 설명을 한다. 이러한 처리과정들은 감정에 따른 피치 변경 발성에서 변동의 보상에 의해 화자 내 음성의 변동



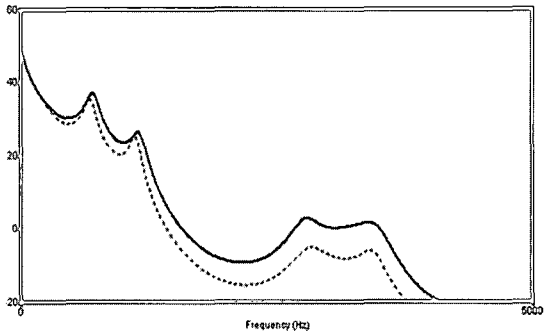
(그림 3) 남자가 발성한 유성음의 LPC스펙트럼 포락 (113~251Hz의 pitch를 갖는 모음 /a/)



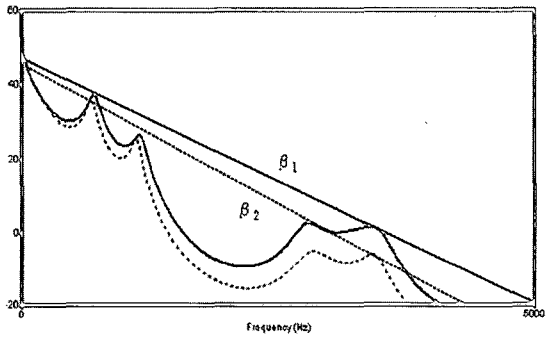
(그림 4) 여성이 발성한 유성음의 LPC 스펙트럼 포락(194~342Hz의 피치를 갖는 모음 /a/)

을 줄이기 위한 시도이다. 피치 변경 발성들에 의한 왜곡들은 음성 신호의 주파수 도메인에서 간단한 선형 warping에 의해 설계 될 수 있기 때문에 정규화 절차는 적절하게 추정된 warping factor에 의해 진폭 축을 조절한다. 운율은 감정의 음향학적인 특성들의 표현으로 알려져 있다. 따라서 음성 파형 데이터의 유성음 구역으로부터 특징 파라미터를 분석한다. 이는 화자 내 인수를 위한 중요한 점이다. 그림 3과 그림 4는 피치 변경 발성에따른 남자와 여자가 발성한 유성음의 선형 예측 계수(LPC) 스펙트럼 포락들을 나타내었다.

높은 harmonics에서 에너지 이득은 성문의 기류 파형의 비교에 의해서 나타낼 수 있다. 발음의 세기가 증가함에 따라 성문의 폐쇄비율이 증가한다. 일반적으로 남성의 목소리는 여성의 목소리보다 낮은 기본 주파수와 강한 harmonics를 가지는



(그림 5) pitch 변경 발성의 화자 내 특성



(그림 6) pitch 변경 발성에 따른 화자 내 특성 파라미터 β

경향이 있다.

그림 5는 남성 음성 /a/의 정상(굵은선) 발성과 피치를 낮춰서(점선) 발성한 음성 스펙트럼들을 보여준다. 화자 내 피치 변경 발성으로부터 변형되지 않은 음성의 특징 공간 분포들은 성문과 성도에 의해 발생하는 음성의 음향학적 차이 때문에 다양하다. 그러므로, 참조 피치 입력의 반비례율 계산에 의해 warping factor를 고려하는 것이 가능하다.

개념적으로 warping 인수는 입력 화자의 피치와 참조 피치 사이의 역비율을 나타낸다. 음성은 추정된 진폭 warping 인수를 사용하여 warping한다. 그리고, 결과 특징 벡터들은 HMM 디코딩을 위해 사용한다. 목표는 정규화된 HMM 모델에 진폭 크기에 match하기위해 각 시험 발성의 진폭 크기를 warping시키는 것이다. 그림 6은 pitch 변경 발성에 따라 화자 내 특성 파라미터 β를 보여

준다. 화자 내 특정한 scale 인수는 스펙트럼의 에너지와 밀접하게 관련이 있다. 그림 6은 pitch와 에너지를 사용하여 β1 =ββ2를 만족하는 β를 추정한다.

화자 내 음향학적 모델링의 정규화에서, 어떤 화자 내 특정 파라미터 β분포에 의존 한다는 것을 다음과 같이 가정한다.

$$p(X|W;\theta,\alpha,\beta) \tag{7}$$

이는 전형적으로 두가지 변환으로 구분할 수 있다.

□ 모델 파라미터 θ의 변환:

화자 내 특징 파라미터 β값에 대해 비 정규화된 모델 파라미터 θ를 정규화한 파라미터 모델 파라미터 θ^{β}로 변환한다.

$$\theta \rightarrow \theta^{\alpha,\beta}(\text{or } \theta \rightarrow \theta^{\beta}) \tag{8}$$

분포는 다음과 같다.

$$p(X|W;\theta,\alpha,\beta) = p(X|W;\theta^{\alpha,\beta}) \tag{9}$$

or

$$p(X|W;\theta,\alpha,\beta) = p(X^{\alpha}|W;\theta^{\beta}) \tag{10}$$

□ 관측벡터 X의 변환:

음향학적 벡터들의 mapping에 따라 다음과 같이 공식화 할 수 있다.

$$X \rightarrow X^{\alpha,\beta}(\text{or } X \rightarrow X^{\beta}) \tag{11}$$

분포는 다음과 같다.

$$p(X|W;\theta,\alpha,\beta) = p(X^{\alpha,\beta}|W;\theta) \tag{12}$$

or

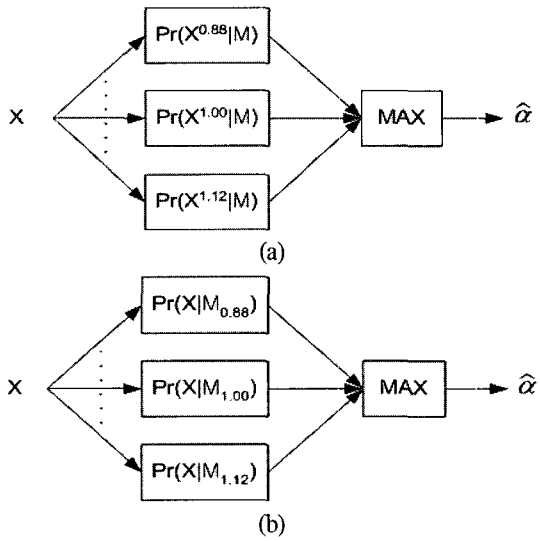
$$p(X|W;\theta,\alpha,\beta) = p(X^{\beta}|W;\theta^{\alpha}) \tag{13}$$

진폭 축의 화자 내 scale 인수 β는 음성 인식을 위한 음향벡터를 계산하기전에 진폭 축을 rescale 하는데 사용된다.

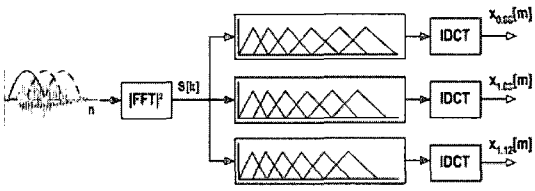
4. 실험 및 결과

실험은 SKKU-1음성 DB를 사용하여 실행하였다. SKKU-1음성 DB의 어휘는 한국어 숫자음, 성명, PBW(phonetically balanced word), PRW(phonetically rich word)로 이루어져 있다. 음성 신호는 1-0.95z-1 고역강조 하였다. 그리고 20ms의 해밍 윈도우를 취하여 10ms단위로 분석하였다. 각각의 프레임은 39차원의 특징벡터를 추출하였다. 특징들은 12차 MFCC(mel-frequency cepstrum coefficient)벡터, 12차 delta-MFCC벡터, 12차 delta-delta-MFCC벡터, log 에너지, delta log 에너지, delta-delta log 에너지이다.

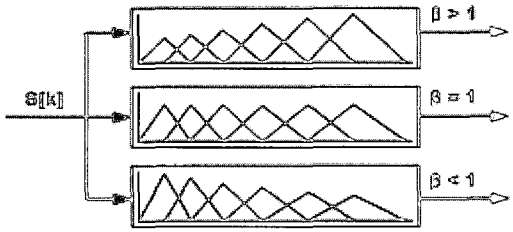
그림 7은 최적의 주파수 warping 인수 추정에 기반을 둔 mixture를 보여준다[1]. 음성은 추정된 warping 인수를 사용하여 warping되고, 특징 벡터들의 결과는 HMM 인식을 위하여 사용된다. 그림 8은 주파수 warping을 하는 mel-filter bank 분석을 보여준다.



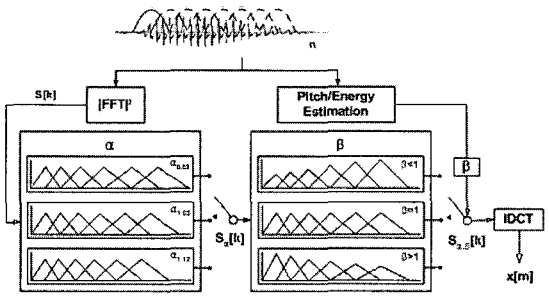
(그림 7) 최적의 주파수 warping 인수 추정



(그림 8) 주파수 warping을 하는 mel filterbank 분석



(그림 9) 진폭 warping 하는 mel filterbank 분석



(그림 10) α와 β의 순차 적용예

진폭 정규화를 위하여 첫번째로 발성으로부터 pitch와 에너지를 추출한다. 그리고 두번째로 화자 내 파라미터를 결정한다. 그림9는 화자 내 정규화를 위해 진폭 warping을 하는mel filterbank를 보여준다. 그림 10은 주파수 정규화와 진폭 정규화를 순차적으로 적용하는 예를 나타내고 있다.

표 1은 기본 인식기를 사용한 경우, 화자간 정규화를 적용한 인식기를 사용한 경우, 그리고 화자 간과 화자 내 정규화를 적용한 인식기를 사용하여 SKKU-1 DB에서 숫자와 단어의 인식 단어 에러율을 보여준다. 인식 결과에 따르면 단어 인식율은 96.4%와 98.2%이다. 에러율은 숫자와 단어 인식에 대하여 최소0.4%에서 최대 2.3%까지 감소되었다.

(표 1) 단어 에러율

	digits	words
Baseline	5.7%	4.1%
with α	4.1%	2.2%
with αand β	3.6%	1.8%

5. 결 론

화자 정규화를 위한 새로운 화자 내 warping 인수 추정 방법을 제안하였다. 화자 내 정규화 음향학적 모델에 따르면, 어떤 화자 내 특징 파라미터 분포에 의존하고 있다는 것을 가정한다. 화자 내 warping 인수의 선택을 위해 본 논문에서 제안된 방법은 pitch 변경 발성에 따른 진폭 정규화에 기반을 둔다. 실험 결과 제안된 알고리듬은 화자 간과 화자 내 정규화를 사용하여 한국어 숫자음과 단어에 대해 효과적인 인식을 할 수 있다는 것을 알수 있었다.

〈감사의 글〉

본 연구는 한국과학재단 목적기초연구(R05-2002-001007-0) 지원으로 수행되었음.

참 고 문 헌

[1] L. Lee and R.C. Rose, "A frequency warping approach to speaker normalization," IEEE Trans. Speech Audio Processing, Vol. 6, No. 1, pp. 49~60, 1998.

[2] L. Welling, H. Ney and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," IEEE Trans. Speech Audio Processing, Vol. 10, No. 6, pp. 415~427, 2002.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, Vol. 9, pp. 171~185, 1995.

[4] J. McDonough and W. Byrne, "Speaker adaptation

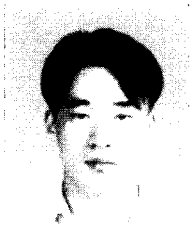
with all-pass transforms," Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 2, Phoenix, AZ, pp. 757~760, 1999.

[5] J. S. Youn, K. W. Chung and K. S. Hong, "A Continuous Digit Speech Recognition Applied Vowel Sequence and VCCV Unit HMM," Proceeding of the Acoustical Society of Korea, Vol. 20, No. 2, pp. 25~28, 2001.

[6] T.D. Rossing, P. Wheeler and F.R. Moore, The Science of Sound, Addition Wesley, 2002.

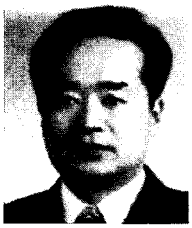
[7] L. Rabiner and R. Schafer, "Digital Processing of speech Signals", Prentice-Hall, 1978.

◐ 저 자 소 개 ◑



김 동 현

2003년 강원대학교 컴퓨터 정보통신공학부 졸업(학사)
2003년~현재 : 성균관대학교 정보통신공학부 재학(석사)
관심분야 : 음성합성, 음성인식
E-mail : super1621@daum.net



홍 광 석

1985년 성균관대학교 전자공학과 졸업(학사)
1988년 성균관대학교 대학원 전자공학과 졸업(공학석사)
1992년 성균관대학교 대학원 전자공학과 졸업(공학박사)
1990년~1993년 서울보건전문대학 전임강사
1993년~1995년 제주대학교 전임강사
1995년~현재 : 성균관대학교 정보통신공학부 부교수
관심분야 : 음성인식 및 음성합성
E-mail : kshong@skku.ac.kr