

## 주파수 왜핑을 이용한 감정에 강인한 음성 인식 학습 방법

# A Training Method for Emotionally Robust Speech Recognition using Frequency Warping

김원구

Weon-Goo Kim

군산대학교 전기공학과

### 요 약

본 논문에서는 인간의 감정 변화의 영향을 적게 받는 음성 인식 시스템의 학습 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 신호와 음성 인식 시스템의 성능에 미치는 영향에 관한 연구를 수행하였다. 감정이 포함되지 않은 평상의 음성으로 학습된 음성 인식 시스템에 감정이 포함된 인식 데이터가 입력되는 경우 감정에 따른 음성의 차이가 인식 시스템의 성능을 저하시킨다. 본 연구에서는 감정의 변화에 따라 화자의 성도 길이가 변화한다는 것과 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 관찰하였다. 본 연구에서는 이러한 음성의 변화를 포함하는 학습 방법을 제안하여 감정 변화에 강인한 음성 인식 시스템을 개발하였다. HMM을 사용한 단독음 인식 실험에서 제안된 학습 방법을 사용하면 감정 데이터의 오차가 기존 방법보다 28.4% 감소되었다.

### Abstract

This paper studied the training methods less affected by the emotional variation for the development of the robust speech recognition system. For this purpose, the effect of emotional variation on the speech signal and the speech recognition system were studied using speech database containing various emotions. The performance of the speech recognition system trained by using the speech signal containing no emotion is deteriorated if the test speech signal contains the emotions because of the emotional difference between the test and training data. In this study, it is observed that vocal tract length of the speaker is affected by the emotional variation and this effect is one of the reasons that makes the performance of the speech recognition system worse. In this paper, a training method that cover the speech variations is proposed to develop the emotionally robust speech recognition system. Experimental results from the isolated word recognition using HMM showed that propose method reduced the error rate of the conventional recognition system by 28.4% when emotional test data was used.

**Key Words** : 음성 신호, 강인한 음성 인식, 감정 변화, MFCC, 주파수 왜핑

## 1. 서 론

음성 인식 기술은 인간의 언어를 해석하여 적절한 행동을 수행할 수 있는 기계를 만드는 것을 목적으로 한다. 최근에는 이러한 기술들이 발달함에 따라 인간과 기계사이의 보다 편리한 인터페이스로의 사용이 급격히 증가하고 있다. 특히 최근에는 음성 인식 시스템의 실용화가 늘어나면서 실생활에 유용하게 사용될 수 있는 응용 제품들이 개발되고 있다. 현재 음성 인식 기술은 상당히 발전하여 수십만 단어의 어휘를 인식하고 실용화가 가능할 정도로 인식 성능도 향상되고 있다.

그러나 이러한 기술이 아직도 가지고 있는 문제점은 음성 인식 시스템의 성능이 주변 잡음 및 채널 특성 등의 환

경 변화와 감정 상태와 같은 심리적 변화에 크게 좌우된다는 것이다. 이중에 환경 변화에 대한 연구는 음성 인식 시스템의 실용화를 위하여 오래 전부터 연구되어왔다. 그러한 이유는 잡음이 없거나 비교적 조용한 실험실 환경에서 우수한 성능을 나타내는 음성 인식 시스템의 성능은 주위에 잡음이 존재하거나 인식 시스템의 학습 환경과 다른 환경에서 사용될 때 그 성능이 급격히 떨어지기 때문이다. 현재 외국의 이러한 연구는 음성 인식 시스템을 실용화하기 위한 중요한 기술로 연구되어 지고 있다. 일본은 음성에 관하여 잡음에서의 음성처리를 8가지 핵심 기술 분야의 한가지로 연구하고 있으며, 유럽 국가들의 ESPRIT(European strategic program for research and development in information technology) 공동 프로그램에서도 잡음을 고려한 음성 인식 알고리즘을 주된 연구과제의 하나로 삼은 바 있다. 또한 국내에서도 음성 인식 기술이 많은 발전을 하여 실용화를 목표로 하면서 자동차 환경, 모바일(mobile) 환경 등의 잡음 처리에 관한 연구가 오래 전부터 진행되어 왔다 [1-7].

이와 함께 음성 인식 시스템의 성능에 영향을 미치는 요

접수일자 : 2009년 12월 21일

완료일자 : 2010년 8월 3일

감사의 글 : 본 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구결과임(중견연구자지원사업[핵심] No.2007-0056927)

인으로 인간의 심리적 변화가 있다. 즉 음성 신호의 형태가 인간의 감정 상태에 따라서 변화하여 평상시 발음과 기쁨, 슬픔, 화남, 우울 등의 상태에서 발음한 것이 크게 다르다는 점이다. 현재의 음성 인식 시스템들이 평상시 감정 상태 (neutral state)에서 발음한 음성 데이터를 사용하여 만들어졌기 때문에 인간의 감정이 들어간 음성을 인식하는 경우에는 그 성능이 저하된다. 이와 관련된 외국의 연구로는 강제가 있는 음성(stressed speech)이나 롬바드 효과(Lombard effect)를 갖는 음성에 대한 인식 성능 향상에 관한 연구가 오래 전부터 진행되어 왔다[8-10].

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 인간의 감정 변화의 영향을 적게 받는 음성 인식 시스템의 학습 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 신호와 음성 인식 시스템의 성능에 미치는 영향에 관한 연구를 수행하였다. 본 연구에서는 감정의 변화에 따라 화자의 성도 길이가 변화한다는 것을 분석하였고 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 관찰하였다. 본 연구에서는 이러한 음성의 변화를 포함하는 학습 방법을 제안하여 감정 변화에 강인한 음성 인식 시스템을 개발하였다.

본 논문의 구성은 다음과 같다. 2장에서는 감정에 따른 음성 변화와 주파수 왜핑에 관하여 설명하고 3장에서는 제안된 감정 변화에 강인한 학습 방법에 대하여 설명한다. 4장에서는 다양한 실험을 통하여 감정이 음성에 미치는 영향을 분석하고 제안된 시스템의 성능을 비교 분석한다. 마지막으로 5장에서는 결론을 맺는다.

## 2. 감정에 따른 음성 변화와 주파수 왜핑(frequency warping)

화자독립 음성 인식 시스템은 음성 신호에 포함된 다양한 변화를 극복해야 한다. 이러한 변화는 전송 채널, 잡음, 화자와 발음 형태 등 다양한 형태로 존재한다. 이 중에서 음성 신호에서 중요한 변화중의 한 가지가 각 화자의 성도 길이 차이에 의한 것이다. 성도 길이 정규화(Vocal Tract Length Normalization : VTLN) 방법은 각 화자의 성도 길이 차이에 따른 영향을 줄이기 위하여 음성 신호 스펙트럼의 주파수 축을 왜핑하는 것이다[11]. 그림 1은 전통적인 멜 캡스트럼 분석 방법(a)과 주파수축 왜핑을 추가한 분석 방법(b)을 나타낸다. 전통적인 방법인 그림 1(a)에서 음성 신호는 프리엠퍼시스(pre-emphasis)와 창 함수와 같은 일련의 전처리 과정을 거친 후 각 프레임마다 푸리에 파워 스펙트럼이 계산된다. 그 후 멜 주파수로 왜핑된 후에 필터 बैं크와 로그 함수가 취해지고 마지막 단계에서 이산 코사인 변환(DCT) 적용되어 멜 캡스트럼 계수가 만들어진다. 그림 1(b)에서는 전통적인 멜 캡스트럼을 구하는 과정 중에 근사화된 선형(piece-wise linear) 또는 이중선형(bilinear)와 같은 왜핑 함수에 의하여 주파수 왜핑이 된다. 이후에 필터 बैं크와 로그 함수가 취해지고 마지막 단계에서 이산 코사인 변환(DCT) 적용되어 멜 캡스트럼 계수가 만들어진다.

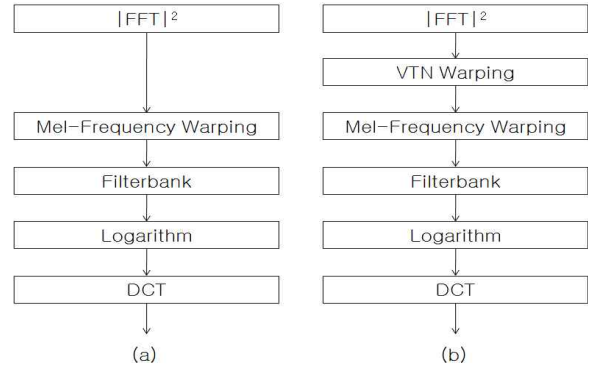


그림 1. 전통적인 멜 캡스트럼 연산(a)과 왜핑 함수를 추가한 방법(b)

Fig. 1. Scheme of traditional MFCC computation(a) and integrated method with warping function(b)

지금까지 여러 가지 형태의 왜핑 함수가 제안되었다. Wegmann[12]과 Welling[13] 등은 식 (1)과 같이 근사화된 선형 함수  $w_l(f)$ 를 사용하였다. 여기서 제한 주파수  $f_0$ 까지는 스펙트럼은 왜핑 파라미터  $\alpha$ 로 선형적으로 왜핑되고  $f_0$ 부터 나이퀴스트 주파수까지는 다른 왜핑 파라미터  $\alpha$ 가 적용되어  $w_l(f_N) = f_N$ 으로 생략되는 주파수 영역이 없도록 하였다. Acero, Stern 과 McDonough[14]는 이중선형 함수를 사용하였고 Eide와 Gish[15]는 파워 함수(power function)을 왜핑 함수로 사용하였다. 이러한 왜핑 함수는 스펙트럼을 위쪽으로 또는 아래쪽으로 이동시키는 역할을 수행한다. Molau[16] 등은 이러한 왜핑 함수들의 성능을 비교한 연구를 수행하여 근사화된 선형 함수가 왜핑 함수로 가장 우수한 성능을 나타냄을 보였다. 근사화된 선형 왜핑 함수는 그림 2와 같다.

$$w_l(f) = \begin{cases} \alpha f & f \leq f_0 \\ \alpha f_0 + \frac{f_N - \alpha f_0}{f_N - f_0} (f - f_0) & f > f_0 \end{cases} \quad (1)$$

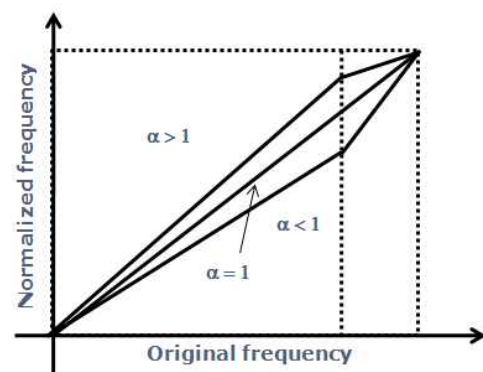


그림 2. 근사화된 선형 왜핑 함수  
Fig. 2. piecewise linear warping function

본 연구에서는 감정에 따른 음성의 변화로 화자의 성도 길이가 변화를 연구하였다. 우선 감정 변화에 따라 성도의 길이가 변화한다는 것을 실험적으로 증명하기 위하여 다음과 같은 과정의 실험을 수행하였다.

- 1) 감정이 없는 평상 음성에 왜핑 파라미터가 -0.88부터

- 1.12 까지 0.02 간격으로 13개의 음성 특징 파라미터를 생성한다.
- 2) 각 화자마다 동일한 문장에 대하여 와핑된 정상 감정의 음성들과 감정(기쁨, 슬픔, 화남)이 포함된 음성을 비교하여 최소의 값을 갖는 와핑 파라미터를 찾는다.
- 3) 최소값을 갖는 와핑 파라미터를 각 감정에 대하여 히스토그램을 그린다.

이와 같은 과정을 통하여 구하여진 히스토그램은 그림 3과 같다. 그림 3(a)는 정상 감정의 음성을 서로 비교한 경우로 와핑 파라미터의 값이 1.0 부근에 집중되었다. 이는 동일 화자 동일 문장의 정상 음성은 성도 길이의 변화가 거의 없음을 나타낸다. 그림 3(b)는 정상 감정의 음성과 기쁨 감정의 음성을 비교한 경우로 와핑 파라미터의 값이 넓게 퍼지면서 1.0 이상의 값 부분에 넓게 분포되고 있다. 이는 기쁨 감정 음성의 경우 성도의 길이가 약간 짧아지는 특성을 나타내는 것이다. 그림 3(c)는 정상 감정 음성과 슬픔 감정 음성을 비교한 경우로 와핑 파라미터의 값이 0.96을 중심으로 넓게 분포하고 있다. 이는 슬픔 감정 음성의 경우 성도의 길이가 길어지는 특성을 나타내는 것이다. 그림 3(d)는 정상 감정 음성과 화남 감정 음성을 비교한 경우로 와핑 파라미터의 값이 1.0을 중심으로 넓게 분포하고 있다. 이는 화남 음성의 경우에는 성도 길이의 길이에 약간의 변화가 있다는 것을 의미한다.

따라서 입력 음성에 감정이 포함된 경우에 정상 음성으로 학습한 음성인식 시스템과 불일치가 발생하여 음성인식 시스템의 성능이 저하되는 문제를 발생시킨다. 이러한 감정에 따른 성도 길이의 변화를 학습 과정에서 음성 인식 시스템에 적용함으로써 감정이 포함된 음성을 인식하는데 강한 음성인식 시스템을 개발하였다.

### 3. 감정 변화에 강한 학습 방법

일반적으로 음성 인식 시스템은 감정이 포함되지 않은 정상 음성으로 학습되기 때문에 감정이 포함된 음성이 입력되면 인식 성능이 저하되는 문제점이 있다. 이러한 것은 감정이 음성의 특성을 변화시켜 음성인식 시스템 성능을 저하시킨다는 것을 의미한다. 본 연구에서는 감정의 변화에 따라 화자의 성도 길이가 변화한다는 것을 실험을 통하여 관찰하였고 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 분석하였다.

이러한 음성의 변화를 감정 인식 시스템에 포함하여 감정 인식 시스템의 성능을 향상하기 위하여 본 연구에서는 다음과 같은 학습 방법을 제안하여 감정 변화에 강한 음성 인식 시스템을 개발하였다.

제안된 학습 방법은 그림 4와 같다. 고립단어 음성인식 시스템의 경우, 각 단어의 학습에 사용될  $N$ 개의 학습 데이터  $S_k, k=1, \dots, N$ 는 특징 추출 과정을 거쳐  $N$ 개의 특징벡터열  $X_k, k=1, \dots, N$ 로 변환된다. 제안된 방법에서는 이러한 특징벡터 추출과정에서 와핑 파라미터  $\alpha$ 를 사용하여 학습데이터를 음성 특징 벡터열  $X_k^\alpha, k=1, \dots, N$ 로 변환한다. 이때 사용되는 와핑 파라미터  $a$ 는 감정에 따른 성도 길이의 변화를 포함할 수 있는 정도로 변화시키기 위하여  $-0.88 \leq a \leq 1.12$  범위에서 0.02 간격으로 샘플링하여 사용하였다. 이렇게 하여  $M$ 개의 와핑 파라미터가 사용된다면  $N$ 개의 학습 데이터는 각각의 와핑 파라미터에 의하여 변형되어

$M \times N$ 개의 특징 벡터열로 변환된다. 이렇게 만들어진 학습 데이터는 각 음성을 학습하는데 사용된다. 이렇게 학습된 음성 모델은 감정에 따라 변화하는 스펙트럼의 변형을 포함하는 감정 변화에 강한 음성 인식 시스템의 모델로 사용된다.

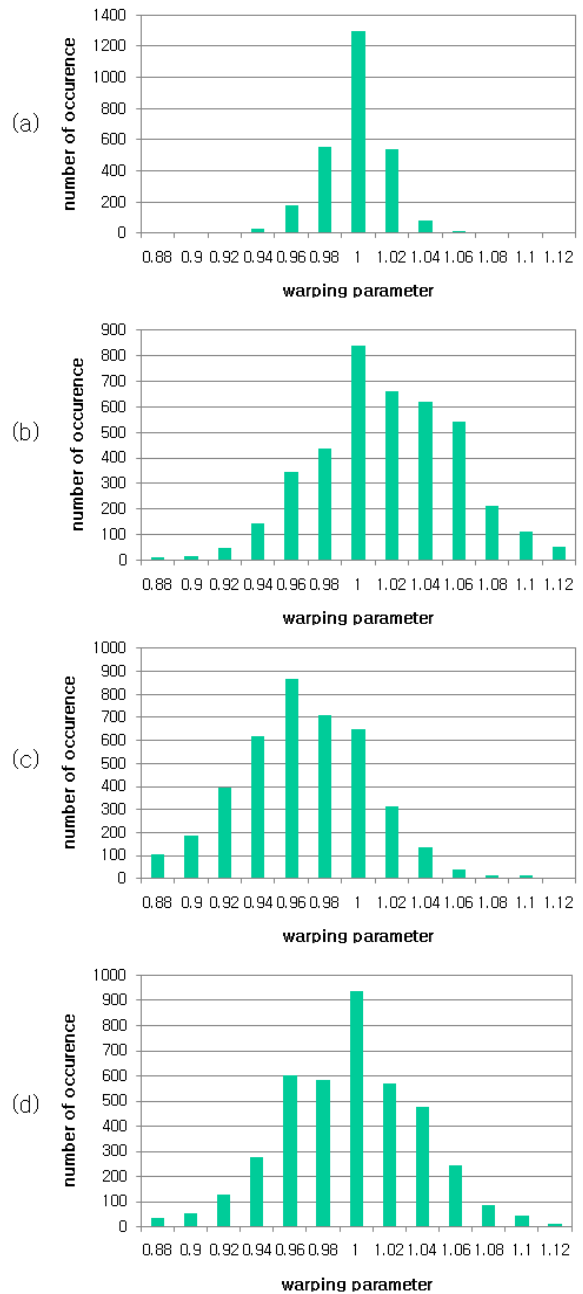


그림 3. 정상 음성과 감정 음성 사이의 차이를 나타내는 히스토그램 (a) 정상-정상 (b) 정상-기쁨 (c) 정상-슬픔 (d) 정상-화남

Fig. 3. Histogram that representing the difference between neutral and emotional speech (a) neutral-neutral (b) neutral-happy (c) neutral-sad (d) neutral-angry

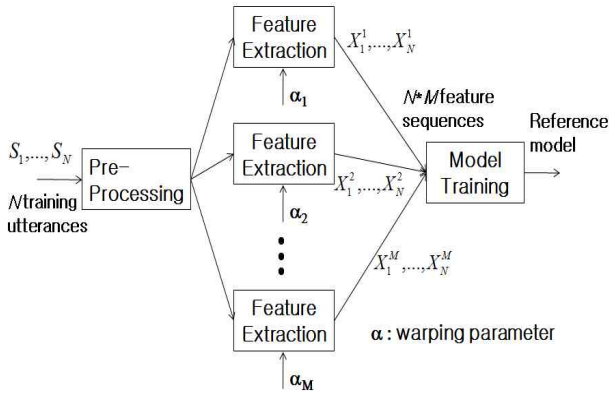


그림 4. 감정 변화에 강인한 모델 학습 방법  
Fig. 4. a robust training method for emotional variation

### 4. 실험 및 결과

#### 4.1 데이터 베이스

감정 변화에 강인한 음성 인식 시스템의 성능을 평가하기 위해서는 다양한 감정이 포함된 음성 데이터 베이스가 필요하다. 이러한 데이터 베이스는 다음과 같은 과정으로 구성되었다[17]. 데이터 베이스를 구성하기 위해서는 사용 용도를 고려한 감정 선정, 문장 선정, 녹음 대상 선정, 녹음 환경, DB 규모 등의 결정 작업이 필요하다. 본 연구에서는 인간의 주요 감정인 기쁨, 슬픔, 화남의 3가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4가지 감정을 인식 대상 감정으로 결정하였다. 음성의 녹음은 평소 감정 표현을 훈련하는 아마추어 연극단원 남/녀 각 15명을 대상으로 하였고, 모든 참여자에 대해서 표준어 사용여부 및 감정 표현 능력을 심사하여 선별되었다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다. 각 화자는 45개의 문장을 4가지 감정으로 녹음하였고 녹음 동안에 감정 표현이 미흡하다고 판단된 경우에는 다시 녹음을 하였다. 본 연구를 위하여 사용된 데이터의 규모는 16,200(30명×4감정×45문장×3회)문장이다.

#### 4.2 특징 파라미터 추출

음성 신호의 특징 파라미터 추출 과정은 다음과 같다. 전처리를 통하여 16KHz, 16비트로 샘플링하고, 고주파 성분을 보강한다. 이렇게 샘플링된 신호는 음성 구간과 묵음 구간을 구별하기 위하여 음성 구간 검출을 수행하고 특징 벡터를 구한다. 검출된 음성 신호는 20ms(320샘플)의 길이를 갖는 해밍 창을 사용하여 10ms씩 이동하면서 특징 파라미터를 구한다. 본 연구에서는 음성의 특징 파라미터로 12차의 멜 캡스트럼 계수와 12차의 델타 멜 캡스트럼 계수로 총 24개를 사용하였다. 또한 음성에 포함된 편의(bias)를 제거하는 방법으로 CMS(Cepstrum Mean Subtraction) 방법을 사용하였다.

#### 4.3 음성 인식 시스템의 구성

본 연구에서는 우선 감정 변화에 강인한 음성 인식 시스템 개발을 위하여 우선 반연속 HMM을 기본으로 하는 화자 독립 단독음 인식 시스템을 구현하였다. 실험에 사용된 음성 인식 시스템의 블록도는 그림 5와 같다. 음성 신호는

샘플링되어 고주파 성분이 보강된 후 음성구간 검출을 수행된다. 검출된 음성 신호를 사용하여 음성 파라미터를 구하고 음성에 포함된 편의(bias)를 제거하기 위한 CMS 방법이 사용되었다.

반연속 HMM 모델은 256개의 코드를 갖는 코드북을 사용하였고 반연속 HMM은 상태 당 4개의 가우시안 결합 분포를 사용하였다. 또한 각 모델의 상태 수는 학습에 사용된 문장의 평균길이에 비례하게 할당하였다. 모델의 학습에는 20명(남성 10명과 여성 10명)이 각 문장을 3회 발음한 음성이 사용되었고(20명×45문장×3회=2700문장) 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성 5명)이 각 문장을 3회 발음한 음성(10명×45문장×3회×4감정=5400문장)을 사용하였다.

학습 과정에서는 다수의 와핑 파라미터를 사용하여 입력 음성에 대하여 다수의 특징 벡터 열이 생성된다. 이렇게 생성된 특징 파라미터들은 반연속 HMM을 학습하는데 사용된다. 인식 과정에서는 입력된 음성에 대하여 전처리 과정을 거쳐 특징 파라미터를 구한 후 기준 모델과 유사도를 비교한다. 결정 법칙(decision rule)은 비교된 결과를 각 단어당 기준 모델 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

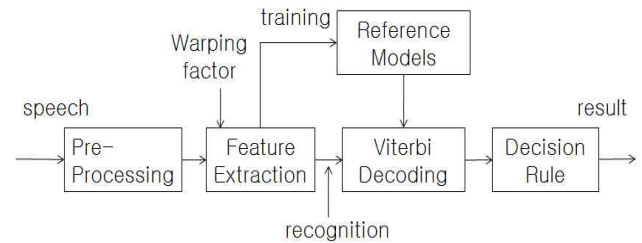


그림 5. 음성 인식 시스템 구조.  
Fig. 5. The structure of speech recognition system

#### 4.4 감정에 따른 음성의 변화

음성에 포함된 감정은 음성의 특성을 변화시킨다. 감정의 변화를 받는 파라미터로는 일반적으로 피치, 발음속도, 에너지, 스펙트럼 등 다양하다. 그림 6은 여성 화자의 감정별 스펙트럼의 변화를 나타내는 스펙트로그램을 보여준다. 그림 6(b)에서도 알 수 있듯이 감정이 평상인 그림 6(a) 경우에 비하여 기쁨인 경우에 스펙트럼의 변화가 큰 것을 알 수 있다. 또한 음성의 특정 부분에서는 비슷한 스펙트럼의 형태를 가지지만 다른 부분에서는 상당히 다른 스펙트럼의 모양과 변화를 나타내고 있다. 이러한 점은 감정이 한 문장의 전체에 영향을 미치기도 하지만 특정 부분이나 단어에만 영향을 미친다는 것을 의미한다. 그림 6(c)는 슬픈 감정이 포함된 경우로 스펙트럼의 변화가 평상시의 발음에 비하여 완만함을 알 수 있다. 그림 6(d)는 화남의 경우로 스펙트럼의 기폭이 매우 심하고 유성음의 길이도 짧음을 알 수 있다. 특히 다른 감정과 다르게 스펙트럼의 변화가 음성의 전 구간에서 나타나고 있으며 문장의 끝 부분에서 스펙트럼의 변화가 매우 급격한 특성을 나타낸다. 따라서 감정 변화에 강인한 음성 인식 시스템을 구현하기 위해서는 이러한 음성의 변화를 파악하여 음성 인식 시스템이 처리할 수 있어야 한다.

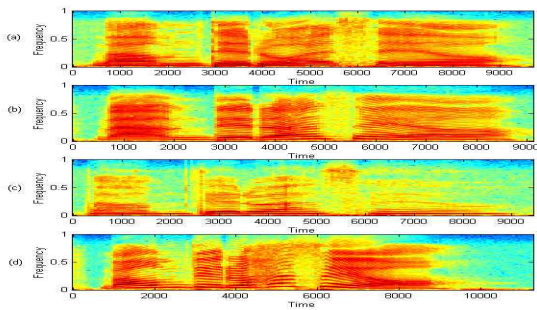


그림 6. 음성 “마음대로 하세요”의 감정별 스펙트로그램(여성화자) (a) 평상 (b) 기쁨 (c) 슬픔 (d) 화남

Fig. 6. spectrogram of speech signal “ma-eum-dae-ro-ha-se-yo”(female) according to the emotion (a) neutral (b) happy (c) sad (d) angry

#### 4.5 실험 결과

평상시 발음한 음성을 대상으로 구축한 음성 인식 시스템에 감정이 포함된 음성이 사용되면 감정에 의한 음성 신호의 변형으로 인하여 인식 시스템의 성능이 크게 저하된다. 본 실험에서는 제안된 학습 방법의 성능을 평가하기 위하여 감정이 포함되지 않은 음성으로 학습한 인식 시스템과 와핑을 사용하여 학습한 인식 시스템을 대상으로 4가지 감정이 포함된 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다. 표 1은 각 학습 방법에 따른 감정별 인식 성능을 나타낸다. 여기서 기존 음성 인식 시스템(no warping)은 평상의 감정만 포함된 데이터로 학습되었기 때문에 인식 데이터가 평상인 경우에 가장 성능이 우수하고 감정이 포함되면 인식 성능이 저하된다. 표에서 평균값은 4가지 감정에 대한 평균 인식률을 나타낸다. 기존 인식 시스템의 평균 인식률은 98.45% 이다. 한편 학습과정에서 학습데이터에 와핑 함수를 사용하여 학습된 인식 시스템(warping)인 경우에는 기쁨과 슬픔에서 인식 성능이 크게 향상되는 것을 볼 수 있다. 기쁨 음성의 경우에는 97.78%에서 99.19%로 인식 성능이 향상되어 오차가 63.5% 감소하였고, 슬픔의 경우에는 97.48%에서 98.22%로 인식 성능이 향상되어 오차가 29.4% 감소하였다. 한편 화남의 경우에는 25% 인식 오차가 증가하였다. 따라서 전체적인 인식 성능은 98.45%에서 98.89%로 향상되어 인식 오차가 28.4% 감소하였다.

표 1. 학습 방법에 따른 성능 평가

Table 1. Recognition performance according to the training methods

학습방법 \ 감정	평상	기쁨	슬픔	화남	평균
no warping	100.0	97.78	97.48	98.52	98.45
warping	100.0	99.19	98.22	98.15	98.89

그림 7은 제안된 방법을 사용할 때 학습에 사용된 와핑 파라미터의 개수에 따른 인식 성능을 변화를 나타낸다. 실

험에서 사용된 와핑 파라미터는 -0.88부터 1.12 까지 0.02 간격으로 총 13개를 사용하였다. 학습 과정에서는 사용된 와핑 파라미터의 개수는 표 2와 같이 사용하였다. 표에서 와핑된 학습 데이터가 1개인 경우는 값이 1.0인 와핑 파라미터 1개만을 사용하였고, 3개의 경우는 -0.98, 1, 1.02, 5개의 경우에는 -0.96, -0.98, 1.0, 1.02, 1.04 과 같이 와핑 파라미터 1.0을 기준으로 증가하면서 총 13개의 개수까지 늘려나갔다. 그림 7에서 인식 시스템의 성능은 와핑된 학습데이터의 개수를 증가시킬수록 향상되어 13개를 사용하였을 때 가장 우수한 성능을 나타내었다

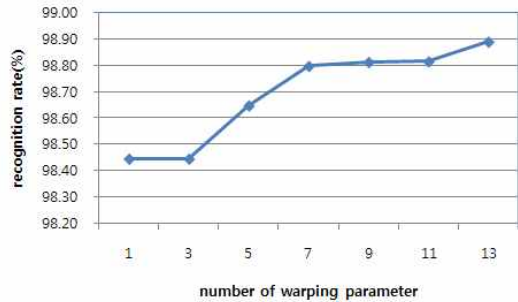


그림 7. 와핑 파라미터 개수에 따른 인식 성능  
Fig. 7. Recognition performance according to the number of warping factors

표 2. 제안된 학습 방법에 사용된 와핑 파라미터

Table 2. warping parameter used in the propose training method

학습에 사용된 와핑 파라미터 개수	와핑 파라미터 값 (x는 0.02 간격)
1	1.0
3	-0.98 ≤ x ≤ 1.02
5	-0.96 ≤ x ≤ 1.04
7	-0.94 ≤ x ≤ 1.06
9	-0.92 ≤ x ≤ 1.08
11	-0.90 ≤ x ≤ 1.10
13	-0.88 ≤ x ≤ 1.12

#### 5. 결 론

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 인간의 감정 변화의 영향을 적게 받는 음성 인식 시스템의 학습 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 신호와 음성 인식 시스템의 성능에 미치는 영향에 관한 연구를 수행하였다. 본 연구에서는 감정의 변화에 따라 화자의 성도 길이가 변화한다는 것을 분석하였고 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 관찰하였다. 본 연구에서는 이러한 음성의 변화를 포함하는 학습 방법으로 학습 데이터에 여러 가지의 와핑 파라미터를 사용하여 다수의 특징 파라미터를 생성한 후 학습하는 방법을 제안하였다.

HMM 을 사용한 단독음 인식 실험에 제안된 학습 알고

리즘을 사용한 경우 기쁨과 슬픔에서 인식 성능이 크게 향상되는 것을 볼 수 있었다. 기쁨 음성의 경우에는 97.78%에서 99.19%로 인식 성능이 향상되어 오차가 63.5% 감소하였고, 슬픔의 경우에는 97.48%에서 98.22%로 인식 성능이 향상되어 오차가 29.4% 감소하였다. 한편 화남의 경우에는 25% 인식 오차가 증가하였다. 평상, 기쁨, 슬픔과 화남의 4가지 감정을 포함하는 인식 데이터에 대하여 제안된 학습 방법을 사용하여 기존 인식 시스템의 인식 오차를 28.4% 감소시킬 수 있었다.

### 참 고 문 헌

[1] J. C. Junqua, and J. P. Haton, *Robustness in Automatic Speech Recognition - Fundamental and Applications*, Kluwer Academic Publishers, 1996.

[2] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proceedings of ICASSP*, pp. 849-852, April 1990.

[3] H. Hermansky, N. Morgan, H. G. Hirsch, "Recognition of speech in additive and convolutional noise based RASTA spectral processing", in *Proceedings of ICASSP*, pp. 83-86, 1993.

[4] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong, "Integrating RASTA-PLP into Speech Recognition", in *Proceedings of ICASSP*, pp. 421-424, 1994.

[5] M. G. Rahim, B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. Speech & Audio Processing*, vol. 4, No. 1, pp. 19-30, 1996.

[6] N. Amir, "Classifying emotions in speech: a comparison of methods", in *Proceedings of Eurospeech '2001*, Vol. 1, pp. 127-130, Aalborg, Denmark, 2001

[7] A. Nogueiras, etc, "Speech emotion recognition using Hidden Markov Models", in *Proceedings of Eurospeech '2001*, Vol. 4, pp. 2679-2682, Aalborg, Denmark, 2001

[8] R. W. Picard, *Affective Computing*, The MIT Press 1997.

[9] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of Acoustical Society of America*, pp. 1097-1108, Feb. 1993.

[10] 조영임, 장성순, "응급상황에서의 음성인식을 위한 필터기 구현", *한국지능시스템학회 논문지*, 20권 2호, pp. 208-213, 2010.4

[11] M. Pitz, H. Ney, "Vocal tract normalization equals linear transformation in cepstral space", *IEEE Trans. Speech & Audio Processing*, vol. 13, No. 5, pp. 930-944, 2005.

[12] S. Wegmann, D. McAllaster, J. Orlofl and B. Peskin, "Speaker Normalization on

Conversational Telephone Speech, in *Proceedings of ICASSP*, Atlanta, GA, pp. 339-342, May 1996.

[13] L. Welling, R. Haeb-Umbach, X. Aubert and N. Haberland, "A study on speaker Normalization using vocal tract normalization and speaker adaptive training", in *Proceedings of ICASSP*, Seattle, WA, pp. 797-800, May 1998

[14] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space", in *Proceedings of ICASSP*, Toronto, pp. 893-896, May 1991.

[15] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", in *Proceedings of ICASSP*, Atlanta, GA, pp.346-349, May 1996.

[16] Sirko Molau, Stephan Kanthak, Hermann Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition", in *Proceedings of the ESSV'00*, Cottbus, Germany, pp. 209-216, 2000

[17] 강봉석, "음성 신호를 이용한 문장독립 감정 인식 시스템", *연세대학교 석사학위 논문*, 2000.

### 저 자 소 개



#### 김원구(Weon-Goo Kim)

1987년 2월 : 연세대 전자공학과 학사  
 1989년 8월 : 연세대 전자공학과 석사  
 1994년 2월 : 연세 전자공학과 박사  
 1994년 9월 ~ 현재 : 군산대 전기공학과 교수

1998년 9월 ~ 1999년 9월 : Bell Lab,  
 Lucent Technologies(USA) 객원연구원

2008년 7월 ~ 2009년 12월 : 호주 Griffith 대학교 교환교수

관심분야 : 음성 신호처리, 음성 인식, 감정 인식, 음성 변환, 화자 인식

Phone : 063) 469-4745

Fax : 063) 469-4699

E-mail : wgkim@kunsan.ac.kr