한국지능시스템학회 논문지 2009, Vol. 19, No. 6, pp. 773-778

성도 정규화를 이용한 감정 변화에 강인한 음성 인식

Robust Speech Recognition using Vocal Tract Normalization for Emotional Variation

김원구* · 방현진**

Weon-Goo Kim* and Hyunjin Bang**

* 군산대학교 전기공학과

** 군산대학교 컴퓨터정보공학과

요 약

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 시스템을 구현하기 위하여 감정 변화의 영향을 최소화 하는 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화에 따른 음성 신호의 변화를 관찰하였다. 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 감소시키는 방법으로 성도 길이 정규화 방법을 사용한 감정 변화에 강인한 음성 인식 시스템을 개발하였다. HMM을 사용한 단독음 인식 실험에서 제안된 학습 방법을 사용하면 감정 데이터의 오차가 기존 방법보다 41.9% 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 variations on the speech signal 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, vocal tract normalization method is used to develop the robust speech recognition system for emotional variations. Experimental results from the isolated word recognition using HMM showed that the vocal tract normalization method reduced the error rate of the conventional recognition system by 41.9% when emotional test data was used.

Key words : 음성 신호, 강인한 음성 인식, 감정 변화, 성도 정규화, MFCC

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 환경 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 알고리듬을 주된 연구과제의 하나로 삼은 바 있다. 또한 국내에서도 음성 인식 기술이 많은 발전을 하여 실용화를 목표로 하면서 자동차 환경, 모발(mobile) 환경 등의 잡음 처리에 관한 연구가 오래 전부터 진행되어 왔다[1-7].

접수일자 : 2009년 6월 29일
완료일자 : 2009년 12월 1일
이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-521-D00376)

이와 함께 음성 인식 시스템의 성능에 영향을 미치는 요인으로 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 만들어 졌기 때문에 인간의 감정이 들어간 음성을 인식하는 경우에는 그 성능이 저하된다. 이와 관련된 외국의 연구로는 강세가 있는 음성(stressed speech)이나 롬바드 효과(Lombard effect)를 갖는 음성에 대한 인식 성능 향상에 관한 연구가 오래 전부터 진행되어 왔다. "인간의 감정이 음성에 어떠한 변화를 만들어 내는가"라는 음성과 감정과의 상관관계에 대한 연구는 서구의 음향학자들과 심리학자들에 의해 먼저 이루어졌다. 이러한 연구 결과를 바탕으로 공학자들이 다양한 응용 분야를 개발하고 있다[8-21]. 지금까지의 연구는 음성 합성시 인간의 감정을 포함시키는 감정 합성 분야와 음성에 포함된 감정을 추출하는 감정 인식에 관한 연구가 주로 진행되고 있다. 하지만 인간은 음성에 언어적인 정보뿐만 아니라 감정에 대한 정보도 함께 전달하기 감정 변화에 강인한 음성 인식 기술에 대한 필요성은 음성 인식 시스템의 실용화가 늘어남에 따라 더욱 증가될 것이다.

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 인간 감정 변화의 영향을 최소화 하는 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 신호에 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 최소화 하는 방법으로 성도 정규화 방법을 사용하여 감정 변화에 강인한 음성인식 시스템을 개발하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 성도 정규화 방법에 관하여 설명하고 3장에서는 주파수 와핑 방법에 대하여 설명한다. 4장에서는 다양한 실험을 통하여 감정이 음성에 미치는 영향을 분석하고 제안된 시스템의 성능을 비교 분석한다. 마지막으로 5장에서는 결론으로 끝을 맺는다.

2. 성도 정규화

성도 정규화 방법은 일반적으로 화자독립 음성 인식 시스템에서 화자의 성도길이 차이에 따른 음성 신호의 변화를 제거하기 위하여 각 화자의 성도 길이를 정규화 하는 방법이다. 성도의 길이를 변화시키는 방법은 음성 분석과정에서 스펙트럼의 주파수 축을 와핑하는 것이다[22]. 음성 인식 시스템의 학습 과정에서 사용되는 화자의 성도 길이를 정규화 하기 위해서는 각 화자의 성도 길이를 변화시킬 와핑 파라메터가 필요하다. HMM을 사용한 음성 모델을 λ라고 가정하고 X_i^{α}를 화자 i의 모든 음성에 와핑 파라메터 α를 적용하여 구한 특징 벡터의 집합이라고 한다면 최적의 와핑 파라메터 $\hat{\alpha}_i$는 문장 종속 확률 Pr을 최대화하도록 구하여진다.

$$\hat{\alpha}_i = \arg\max_{\alpha} Pr(X_i^{\alpha}|\lambda, W_i) \quad (1)$$

여기서 모델 λ는 보통 1개의 밀도함수를 갖는 낮은 해상도의 음성 모델이 사용된다. 일단 모든 화자의 와핑 파라메터가 결정되면 학습 데이터는 그 값에 따라 정규화되고 이렇게 정규화된 학습 데이터를 사용하여 정상적인 학습 알고리즘을 사용하여 모델 $\overline{\Lambda}$을 학습한다.

인식 단계에서는 학습 과정과 비슷하게 와핑 파라메터를 결정한다. 일반적으로 입력 화자의 신원을 알 수 없으므로 최적의 와핑 파라메터는 입력 문장 단위로 계산되어 진다. 또한 입력 음성 j의 문자열 W_j는 알 수 없으므로 초기 문자열 $\widehat{W}_j$는 정규화되지 않은 입력 특징 벡터 X_j와 정규화되지 않은 모델 Λ를 사용하여 첫 단계로 인식을 수행하여 구한다. 그다음 최적의 와핑 파라메터 $\hat{\alpha}_j$는 정규화된 음성 모델 $\overline{\Lambda}$을 사용하여 결정된다.

$$\hat{\alpha}_j = \arg\max_{\alpha} Pr(X_j^{\alpha}|\overline{\Lambda}, \widehat{W}_j) \quad (2)$$

마지막 단계에서는 입력 특징 벡터는 $\hat{\alpha}_j$에 의하여 정규화되고 정규화된 음성 모델 $\overline{\Lambda}$ 를 사용하여 두 번째 단계의 인식을 수행한다.

3. 주파수 와핑(frequency warping)

그림 1은 전통적인 멜 켑스트럼 분석 방법(a)과 주파수축 와핑을 추가한 분석 방법(b)을 나타낸다. 전통적인 방법인 그림 1(a)에서 음성 신호는 프리엠파시스(pre-emphasis)와 창 함수와 같은 일련의 전처리 과정을 거친 후 각 프레임마다 푸리에 파워 스펙트럼이 계산된다. 그 후 멜 주파수로 와핑된 후에 필터 뱅크와 로그 함수가 취해지고 마지막 단계에서 이산 코사인 변환(DCT) 적용되어 멜 켑스트럼 계수가 만들어 진다. 그림 1(b)에서는 전통적인 멜 켑스트럼을 구하는 과정 중에 근사화된 선형(piece-wise linear) 또는 이중선형(bilinear)와 같은 와핑 함수에 의하여 주파수 와핑이 된다. 이후에 필터뱅크와 로그함수가 취해지고 마지막 단계에서 이산 코사인 변환(DCT) 적용되어 멜 켑스트럼 계수가 만들어 진다.

지금까지 여러 가지 형태의 와핑 함수가 제안되었다. Wegmann[23]과 Welling [24] 등은 식 (3)과 같이 근사화된 선형 함수 $w_l(f)$를 사용하였다. 여기서 제한 주파수 f_0까지는 스펙트럼은 와핑 파라메터 α로 선형적으로 와핑되고 f_0부터 나이키스트 주파수까지는 다른 와핑 파라메터 α'가 적용되어 $w_l(f_N) = f_N$ 으로 생략되는 주파수 영역이 없도록 하였다. Acero, Stern 과 McDonough [25]는 이중선형 함수를 사용하였고 Eide와 Gish [26]는 파워 함수(power function)을 와핑 함수로 사용하였다. 이러한 와핑 함수는 스펙트럼을 위쪽으로 또는 아래쪽으로 이동시키는 역할을 수행한다. Molau[27] 등은 이러한 와핑 함수들의 성능을 비교한 연구를 수행하여 근사화된 선형 함수가 와핑 함수로 가장 우수한 성능을 나타냄을 보였다. 근사화된 선형 와핑 함수는 그림 2와 같다.

$$w_l(f) = \begin{cases} \alpha f & f \le f_0 \\ \alpha f_0 + \dfrac{f_N - \alpha f_0}{f_N - f_0}(f - f_0) & f > f_0 \end{cases} \quad (3)$$

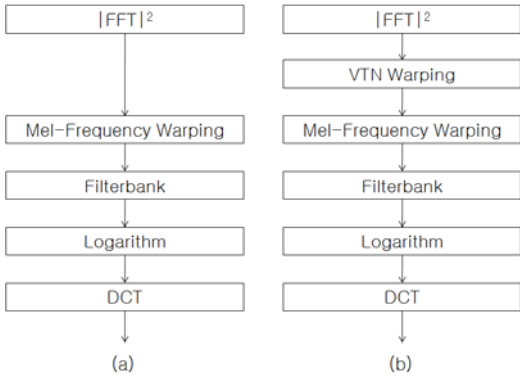

그림 1. 전통적인 멜 켑스트럼 연산(a)과 와핑 함수를 추가한 방법(b)

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

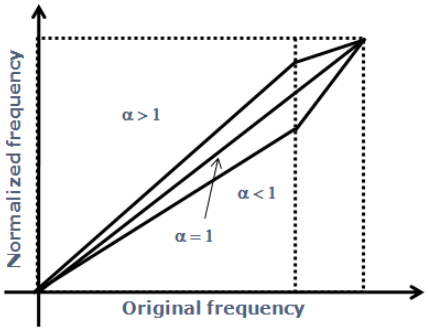

그림 2. 근사화된 선형 와핑 함수

Fig 2. piecewise linear warping function

4. 실험 및 결과

4.1 데이터 베이스(Data Base)

감정 변화에 강인한 음성 인식 시스템의 성능을 평가하기 위해서는 다양한 감정이 포함된 음성 데이터 베이스가 필요하다. 이러한 데이터 베이스는 다음과 같은 과정으로 구성되었다[28]. 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 표현능력을 심사하여 선별되었다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다. 각 화자는 45개의 문장을 네가지 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 파라메터 추출

음성 신호의 특징 파라메터 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 해밍창(Hamming window)을 사용하여 10ms씩 이동하면서 특징 파라메터를 구한다. 본 연구에서는 음성의 특징 파라메터로 멜 켑스트럼 계수를 사용하였다. 실험에 사용된 켑스트럼 계수는 12차를 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 블럭도는 그림 3과 같다. 음성 신호는 샘플링되어 고주파 성분이 보강된 후 음성구간 검출을 수행된다. 검출된 음성 신호를 사용하여 음성 파라메터를 구하고 음성에 포함된 편의(bias)를 제거하기 위한 CMS 방법이 사용되었다.

반연속 HMM 모델은 256개의 코드을 갖는 코드북을 사용하였고 반연속 HMM은 상태 당 4개의 가우시간 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 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성5명)이 각 문장을 3회 발음한 음성을 사용하였다.

학습 과정에서는 학습 데이터에 대하여 1개의 밀도함수를 갖는 낮은 해상도의 음성 모델을 사용하여 학습데이터를 정규화하고 정규화된 학습 데이터를 사용하여 정상적인 학습 알고리즘을 사용하여 모델을 학습한다. 인식 과정에서는 입력 음성을 성도 정규화하지 않은 음성모델로 인식하여 문자열을 파악한 후에 성도 정규화된 음성 도델을 사용하여 입력 음성을 정규화하고 마지막 단계에서 성도 정규화된 입력 음성과 모델을 사용하여 인식을 수행한다. 결정 법칙(decision rule)은 비교된 결과를 각 단어 당 기준 모델 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

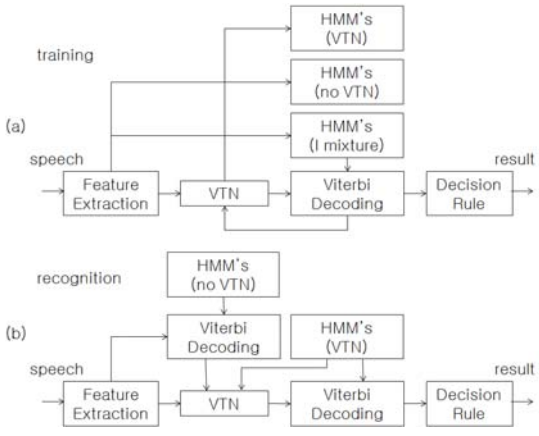

그림 3. 음성 인식 시스템 구조 (a) 학습 (b) 인식

Fig. 3 The Structure of speech recognition system (a) training (b) recognition

4.4 감정에 따른 음성의 변화

음성에 포함된 감정은 음성의 특성을 변화시킨다. 감정의

변화를 받는 파라메터로는 일반적으로 피치, 발음속도, 에너지, 스펙트럼 등 다양하다. 그림 4는 여성 화자의 감정별 스펙트럼의 변화를 나타내는 스펙트로그램을 보여준다. 그림 4(b)에서도 알 수 있듯이 감정이 평상인 그림 4(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 4(c)는 슬픈 감정이 포함된 경우로 스펙트럼의 변화가 평상시의 발음에 비하여 완만함을 알 수 있다. 그림 4(d)는 화남의 경우로 스펙트럼의 기복이 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.

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

1) 감정이 없는 평상 음성에 와핑 파라메터가 -0.88부터 1.12 까지 0.02 간격으로 13개의 음성 특징 파라메터를 생성한다.
2) 각 화자마다 동일한 문장에 대하여 와핑된 평상 감정의 음성들과 감정(기쁨, 슬픔, 화남)이 포함된 음성을 비교하여 최소의 값을 갖는 와핑 파라메터를 찾는다.
3) 최소값을 갖는 와핑 파라메터를 각 감정에 대하여 히스토그램을 그린다.

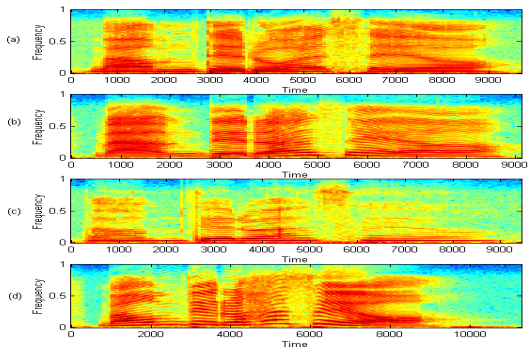

그림 4. 음성 “마음대로 하세요”의 감정별 스펙트로그램(여성화자)
(a) 평상 (b) 기쁨 (c) 슬픔 (d) 화남
Fig. 4 spectrogram of speech signal "ma-eum-dae-ro-ha-se-yo"(female) according to the emotion (a) neutral (b) happy (c) sad (d) angry

이와 같은 과정을 통하여 구하여진 히스토그램은 그림 5와 같다. 이 히스토그램은 각 화자마다 동일 문장에 대하여 감정의 변화에 따른 스펙트럼의 차이를 구한 것이다. 따라서 감정의 변화가 성도 길이에 변화를 주지 못하였다면 히스토그램은 1.0 부분에 집중될 것이고 성도 길이가 길어지거나 짧아지면 1,0보다 작거나 큰 값에 히스토그램이 집중된다. 그림 5(a)는 평상 감정의 음성을 서로 비교한 경우로 와핑 파라메터의 값이 1.0 부근에 집중되었다. 이는 동일 화자 동일 문장의 평상 음성은 성도 길이의 변화가 거의 없음을 나타낸다. 그림 5(b)는 평상 감정의 음성과 기쁨 감정의 음성을 비교한 경우로 와핑 파라메터의 값이 넓게 퍼지면서 1.0 이상의 값 부분에 넓게 분포되고 있다. 이는 기쁨 감정 음성의 경우 성도의 길이가 약간 짧아지는 특성을 나타내는 것이다. 그림 5(c)는 평상 감정 음성과 슬픔 감정 음성을 비교한 경우로 와핑 파라메터의 값이 0.96을 중심으로 넓게 분포하고 있다. 이는 슬픔 감정 음성의 경우 성도의 길이가 길어지는 특성을 나타내는 것이다. 그림 5(d)는 평상 감정 음성과 화남 감정 음성을 비교한 경우로 와핑 파라메터의 값이 1.0을 중심으로 넓게 분포하고 있다. 이는 화남 음성의 경우에는 성도 길이의 길이에 약간의 변화가 있다는 것을 의미한다.

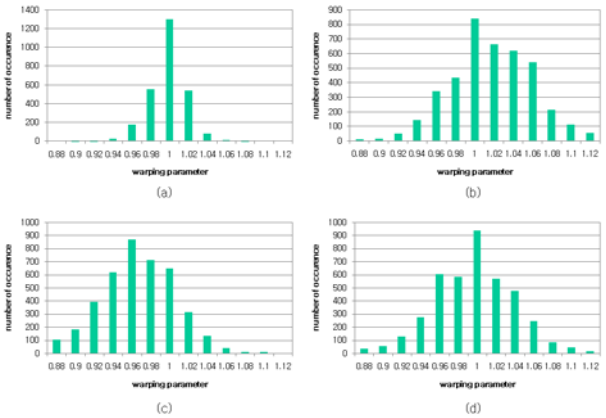

그림 5. 평상 음성과 감정 음성 사이의 차이를 나타내는 히스토그램
(a) 평상-평상 (b) 평상-기쁨 (c) 평상-슬픔 (d) 평상-화남
Fig. 5 Histogram that representing the difference between neutral and emotional speech (a) neutral-neutral (b) neutral-happy (c) neutral-sad (d) neutral-angry

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 시스템에 성도 정규화 방법을 사용하여 4가지 감정이 포함된 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다. 표 1은 기존 시스템과 성도 정규화 방법에 따른 감정별 인식 성능을 나타낸다. 여기서 기존 음성 인식 시스템(no VTN)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 오차는 3.41% 이다. 한편 성도 정규화를 통하여 학습과 인식 과정에서 정규화된 인식 시스템(VTN)의 경우에는 모든 감정의 데이터에서 인식 성능이 향상되는 것을 알 수 있다. 평상 감정음성의 경우에는 오차가 0.37%에서 0.07%로 81.1% 감소하였고 기쁨 감정 음성의 경우에는 5.63%에서 2.96%로 오차가 47.4% 감소하였고, 슬픔 감정 음성의 경우에는 4.15%에서 2.3%로 오차가 44.6% 감소하였다. 또한 화남의 경우에

는 3.48%에서 2.59%로 오차가 25.6% 감소하였다. 따라서 전체적인 평균 인식 오차는 3.41%에서 1.98%로 인식 오차가 41.9% 감소하였다.

표 1. 성도 정규화 방법에 따른 성능 평가(%)
Table 1. Recognition performance according to the VTN method(%)

감정 / 시스템	평상	기쁨	슬픔	화남	평균
Baseline (no VTN)	0.37	5.63	4.15	3.48	3.41
VTN	0.07	2.96	2.30	2.59	1.98

5. 결 론

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 감정 변화의 영향을 최소화 하는 방법에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화에 따른 음성 신호의 변화를 관찰하였다. 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 감소시키는 방법으로 성도 길이 정규화 방법을 사용한 감정 변화에 강인한 음성 인식 시스템을 개발하였다.

HMM 을 사용한 단독음 인식 실험에 성도 정규화를 사용한 경우에 평상 감정의 인식성능 뿐만 아니라 기쁨, 슬픔과 호남의 인식 성능도 크게 향상되는 것을 볼 수 있었다. 평상 감정음성의 경우에는 오차가 81.1% 감소하였고 기쁨 감정 음성의 경우에는 오차가 47.4% 감소하였고, 슬픔 감정 음성의 경우에는 오차가 44.6% 감소하였다. 또한 화남의 경우에는 오차가 25.6% 감소하였다. 따라서 전체적인 평균 인식 오차는 3.41%에서 1.98%로 인식 오차가 41.9% 감소하였다.

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 Proc. ICASSP*, pp. 421-424, 1994.

[5] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", in *Proceedings of EUROSPEECH,* vol. 3, pp. 1367-1370, 1991.

[6] P. Alexandre, P. Lockwood, "Root cepstral analysis: a unified view. application to speech processing in car noise environments", *Speech Communication,* vol. 12, no. 3, pp. 277-288, 1993.

[7] 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.

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

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

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

[11] J. E. Cahn, "The generation of affect in synthesized speech", *Journal of the American Voice I/O Society*, Vol. 8, pp. 1-19, July 1990.

[12] K. R. Scherer, D. R. Ladd, and K. E. A. Silverman, "Vocal cues to speaker affect: testing two models", *Journal Acoustical Society of America,* Vol. 76, No. 5, pp. 1346-1355, Nov. 1984.

[13] 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 Accoustal Society of America.*, pp. 1097-1108, Feb. 1993.

[14] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates", *Journal Acoustical Society of America*, Vol. 52, No. 4, pp. 1238-1250, 1972.

[15] M. Lewis and J. M. Haviland, *Handbook of Emotions*, The Guilford Press 1993.

[16] F. Dellaert, T. Polzin, A. Waibel, "Recognizing emotion in speech", in *Proceedings of the ICSLP 96*, Philadelphia, USA, Oct. 1996

[17] J. Sato, and S. Morishima, "Emotion modeling in speech production using emotion space", in *Proceedings of the IEEE International Workshop 1996*, pp. 472-477, Piscataway, NJ, USA., 1996.

[18] T. S. Huang, L. S. Chen and H. Tao, "Bimodal emotion recognition by man and machine", in *ATR Workshop on Virtual Communication Environments-Bridges over Art/Kansei and VR Technologies*, Kyoto, Japan, 1998.

[19] T. S. Polzin and A. H. Waibel, "Detecting emotions in speech", *Proceedings of the CMC* (Cooperative Multimodal Communication), 1998.

[20] J. Vroomen, R. Collier and S. Mozziconacci, "Duration and intonation in emotional Speech", in *Proceedings of Eurospeech '93,* Vol.1, pp.577-580, Berlin, Germany, 1993.

[21] B. Heuft, T. Portele, M. Rauth, "Emotions in time domain synthesis", in *Proceedings of ICSLP '96*,

Vol. 3, pp.1974-1977, Philadelphia, PA, USA, 1996.
[22] 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.
[23] 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.
[24] 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
[25] 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.
[26] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", in *Proceedings of ICASSP,* Atlanta, GA, pp.346-349, May 1996.
[27] Sirko Molau, Stephan Kanthak , Hermann Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition", in *Proceedings of* of the ESSV'00, Cottbus, Germany, pp. 209-216, 2000
[28] 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 : 연세대 전자공학과 박사
1994년 9월 ~ 현재 : 군산대 전기공학과 교수
1998년 9월 ~ 1999년 9월 : Bell Lab, Lucent Technologies(USA)객원연구원

관심분야 : 음성 신호처리, 음성 인식, 감성 인식, 음성 변환, 화자 인식
Phone : 063) 469-4745
Fax : 063) 469-4699
E-mail : wgkim@kunsan.ac.kr

방현진(Hyunjin Bang)

2001년 2월 : 전주교육대 수학과 학사
2006년 8월 : 군산대 교육대학원 컴퓨터과학과 석사
2007년 3월 ~ 현재 : 군산대 컴퓨터정보공학과 박사과정
2001년 3월 ~ 현재 : 군산해성초교~ 전주오송초교 교사

관심분야 : 음성마이닝, RFID/USN 응용
Phone : 063) 469-4557
Fax : 063) 469-4560
E-mail : bbangfamily@hanmail.net