감정 음성 인식을 위한 강인한 음성 파라미터

Robust Speech Parameters for the Emotional Speech Recognition

이규현 · 김원구*

Guehyun Lee and Weon-Goo Kim*

군산대학교 전기공학과

Department of Electrical Engineering, Kunsan National University

요 약

본 논문에서는 강인한 감정 음성 인식 시스템을 개발하기 위하여 감정의 영향을 적게 받는 음성 파라미터에 대한 연구를 수행하였다. 이러한 목적을 위하여 다양한 감정이 포함된 데이터를 사용하여 감정이 음성 인식 시스템과 음성 파라미터에 미치는 영향을 분석하였다. 본 연구에서는 멜 캡스트럼, 델타 멜 캡스트럼, RASTA 멜 캡스트럼, 루트 캡스트럼, PLP 계수와 성도 길이 정규화 방법에서 주파수 와핑된 멜 캡스트럼 계수를 사용하였다. 또한 신호 편의 제거 방법으로 CMS 방법과 SBR 방법이 사용되었다. 실험결과에서 성도정규화 방법을 사용한 RASTA 멜 캡스트럼, 델타 멜 캡스트럼 및 CMS 방법을 사용한 경우가 HMM 기반의 화자독립 단독음 인식 실험 결과에서 가장 우수한 결과를 나타내었다.

키워드 : 음성인식, 음성 파라미터, 감정변화, 성도길이 정규화, 주파수 와핑

Abstract

This paper studied the speech parameters less affected by the human emotion for the development of the robust emotional speech recognition system. For this purpose, the effect of emotion on the speech recognition system and robust speech parameters of speech recognition system were studied using speech database containing various emotions. In this study, mel-cepstral coefficient, delta-cepstral coefficient, RASTA mel-cepstral coefficient, root-cepstral coefficient, PLP coefficient and frequency warped mel-cepstral coefficient in the vocal tract length normalization method were used as feature parameters. And CMS (Cepstral Mean Subtraction) and SBR(Signal Bias Removal) method were used as a signal bias removal technique. Experimental results showed that the HMM based speaker independent word recognizer using frequency warped RASTA mel-cepstral coefficient in the vocal tract length normalized method, its derivatives and CMS as a signal bias removal showed the best performance.

Key Words : Speech Recognition, Speech Parameter, Emotional Variation, Vocal Tract Length Normalization, Frequency Warping

1. 서 론

음성 인식 시스템의 실용화가 늘어나면서 인간과 기계사이의 보다 편리한 인터페이스로 음성 인식 기술이 사용되는

응용 제품들이 속속 개발되고 있다. 현재 음성 인식 기술은 상당히 발전하여 수십만 단어의 어휘를 인식하고 사용자의 질문에 대답하는 기능을 가지고 인식 성능도 실용화가 가능할 정도로 향상되고 있다. 그러나 이러한 기술이 아직도 가지고 있는 문제점은 주변 잡음이나 채널 특성 등의 환경 변화와 감정 상태와 같은 심리적 변화에 시스템의 성능이 크게 저하된다는 것이다. 음성 인식 시스템의 환경 변화에 대한 연구는 음성 인식의 실용화를 위하여 오래 전부터 연구되어왔다[1-4]. 외국에서 이러한 연구는 음성 인식 시스템을 실용화하기 위한 중요한 기술로 연구되어왔다. 유럽 국가들의 ESPRIT 공동 프로그램에서도 잡음을 고려한 음성 인식 알고리즘을 주된 연구과제로 연구하였고 일본도 음성 에 관하여 잡음에서의 음성처리를 8가지 핵심 기술 분야의 한가지로 삼은 바 있다. 국내에서도 음성 인식 기술이 많이 발전하여 자동차 환경, 모바일 환경 등의 잡음 처리에 관한 연구를 통한 음성 인식 시스템의 실용화에 관하여 오래 전부터 연구해오고 있다[1-3].

음성 인식 시스템의 성능을 저하시키는 또 다른 요인으

접수일자: 2012년 10월 12일

심사(수정)일자: 2012년 10월 19일

게재확정일자: 2012년 11월 30일

* Corresponding author

본 논문은 본 학회 2012년도 추계학술대회에서 선정된 우수논문입니다

이 논문은 2012년도 정부(교육과학기술부) 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(Nb. 2012R1A1A4A01014421)

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

로 인간의 감정 변화가 있다. 음성 신호는 인간의 감정 상태에 따라서 변화하여 평상시 음성과 기쁨, 슬픔, 화남, 우울 등의 감정 상태에서 발음한 것이 크게 다르다. 대부분의 음성 인식 시스템들은 감정이 없는 음성 데이터를 사용하여 만들어졌기 때문에 감정이 포함된 음성을 인식하는 경우에는 성능이 나빠진다. 외국에서도 감정이 포함된 음성에 대한 음성 인식 시스템의 성능을 향상시키기 위한 연구가 오래전부터 진행되어 왔다. 강세가 있는 음성이나 롬바드 효과를 갖는 음성에 대한 음성 인식 연구나 감정 음성 모델에 관한 연구가 진행되어 왔다[4,5]. 인간은 음성에 언어 정보뿐만 아니라 감정 정보도 함께 전달하기 때문에 감정에 강인한 음성 인식 기술에 대한 필요성은 음성 인식 시스템의 실용화가 늘어남에 따라 더욱 증가될 것이다.

본 논문에서는 강인한 감정 음성 인식 기술을 개발하기 위하여 감정 변화의 영향을 적게 받는 음성 파라미터 개발에 관한 연구를 수행하였다. 이러한 목적을 위하여 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정의 변화에 따른 음성 신호와 음성 인식 시스템의 성능 변화를 관찰하였다. 우선 감정 변화에 강인한 특징 파라미터에 대한 연구를 수행하여 기존 음성 파라미터들의 성능을 비교하고 감정 변화에 강인한 파라미터를 찾는 연구를 수행하였다. 본 연구에서 사용된 음성 파라미터는 멜 캡스트럼 계수, 델타 멜 캡스트럼 계수, RASTA 멜 캡스트럼 계수, 루트 캡스트럼 계수, PLP 계수와 성도 길이 정규화 방법에서 주파수와 평면 멜 캡스트럼 계수 등이 사용되었다. 또한 음성 신호에 포함된 편의 제거를 위한 CMS(Cepstral Mean Subtraction)와 SBR(Signal Bias Removal) 방법의 성능이 비교되었다[13].

본 논문의 구성은 다음과 같다. 2장에서는 여러 가지 음성 파라미터에 관하여 설명하고 3장에서는 감정이 음성에 미치는 영향에 관하여 분석하고 다양한 실험을 통하여 여러 가지 음성 파라미터의 성능을 비교 분석하였다. 마지막으로 4장에서는 결론을 맺는다.

2. 음성 파라미터

음성 파라미터인 멜 캡스트럼 계수는 음성 인식에 널리 사용되고 있으며 시간 정보를 나타내는 델타 캡스트럼 계수와 RASTA 멜 캡스트럼도 많이 사용되고 있다. 또한 잡음에 강인한 파라미터로는 루트 캡스트럼과 PLP 계수들이 있고 주파수와 평면을 사용하여 화자의 성도 길이 차이를 보상하는 방법도 음성 인식에서 우수한 성능을 나타내고 있다. 음성 신호에 포함된 편의를 제거를 위하여 CMS와 SBR 방법이 사용되고 있다[11].

2.1 루트 캡스트럼 계수

멜 기반의 루트 캡스트럼 계수는 잡음에 의한 변형에 강인한 것이 관찰되었고 루트(root) 함수로 일반적인 로그라디스 역전볼루먼을 근사화하였다[12]. 그림 1은 루트 캡스트럼 계수를 구하는 과정이다.

2.2 PLP 계수

PLP 분석 방법은 1982년 Hermansky에 의해 제안되었으며, 음성 신호의 파워 스펙트럼을 변화시켜 청각 특성이 고려된 스펙트럼을 이용한다[13]. 이러한 단계를 거쳐 얻어지

는 저차의 스펙트럼은 인간이 실제 감지하는 소리와 유사한 특성을 갖게 되며, 음성 인식에 적용되어 좋은 성능을 보여 주었다. 그림 2는 PLP 분석 방법의 흐름도이다.

2.3 RASTA 멜 캡스트럼 계수

RASTA 분석 방법에서는 일반적인 단구간 스펙트럼을 사용하는 대신 스펙트럼 성분 중 시간에 따라 천천히 변화하는 성분을 배제하는 대역 통과 스펙트럼을 사용한다. RASTA 분석 방법의 흐름도는 그림 3과 같다[3]. 그림의 흐름도에서 필터링 블록은 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링하는 것과 같다. 이 대역 통과 필터의 전달 함수는 다음과 같다.

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98z^{-1})} \quad (1)$$

그림 1. 루트 캡스트럼 분석과정
Fig. 1. Root cepstrum analysis procedure

그림 2. PLP 분석 과정
Fig. 2. PLP analysis procedure

그림 3. RASTA 멜 캡스트럼 계수
Fig. 3. RASTA mel cepstral coefficient

2.4 주파수 와핑된 멜 캡스트럼 계수

그림 4는 전통적인 멜 캡스트럼 분석 방법에 주파수축 와핑을 추가한 분석 방법을 나타낸다. 그림에서는 전통적인 멜 캡스트럼을 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 캡스트럼 계수가 만들어 진다.

그림 4. 주파수 와핑된 멜 캡스트럼 계수

Fig. 4. Frequency Warped mel cepstral coefficient

지금까지 여러 가지 형태의 와핑 함수가 제안되었다. Wegmann [7]과 Welling [8] 등은 식 (2)와 같이 근사화된 선형 함수 $w_l(f)$ 를 사용하였다.

$$w_l(f) = \begin{cases} \alpha f & f \leq f_0 \\ \alpha f_0 + \frac{f_N - \alpha f_0}{f_N - f_0} (f - f_0) & f > f_0 \end{cases} \quad (2)$$

여기서 제한 주파수 f_0 까지는 스펙트럼은 와핑 파라미터 α 로 선형적으로 와핑되고 f_0 부터 나이퀴스트 주파수까지는 다른 와핑 파라미터 α' 가 적용되어 $w_l(f_N) = f_N$ 으로 생략되는 주파수 영역이 없도록 하였다. 이러한 와핑 함수는 스펙트럼을 위쪽으로 또는 아래쪽으로 이동시키는 역할을 수행한다. Molau[9] 등은 이러한 와핑 함수들의 성능을 비교한 연구를 수행하여 근사화된 선형 함수가 와핑 함수로 가장 우수한 성능을 나타냄을 보였다. 근사화된 선형 와핑 함수는 그림 5와 같다.

2.5 성도길이 정규화

성도길이 정규화 방법은 화자독립 음성 인식 시스템에서 화자의 성도길이 차이에 따른 화자간 차이를 제거하기 위하여 각 화자의 성도 길이를 정규화 하는 방법이다. 성도의 길이를 변화시키는 방법은 음성 분석과정에서 스펙트럼의 주파수 축을 와핑하는 것이다[6]. 음성 인식 시스템의 학습 과정에서 사용되는 화자의 성도 길이를 정규화 하기 위해서는 각 화자의 성도 길이를 변화시킬 와핑 파라미터가 필요하다. HMM을 사용한 음성 모델을 λ 라고 가정하고 X_i^α 를 화자 i 의 모든 음성에 와핑 파라미터 α 를 적용하여 구한 특징 벡터의 집합이라고 한다면 최적의 와핑 파라미터 $\hat{\alpha}_i$

는 문장 종속 확률 P_r 을 최대화하도록 구하여진다.

$$\hat{\alpha}_i = \arg \max_{\alpha} Pr(X_i^\alpha | \lambda, W_i) \quad (3)$$

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

인식 단계에서는 학습 과정과 비슷하게 와핑 파라미터를 결정한다. 일반적으로 입력 화자의 신원을 알 수 없으므로 최적의 와핑 파라미터는 입력 문장 단위로 계산되어 진다. 또한 입력 음성 j 의 문자열 W_j 는 알 수 없으므로 초기 문자열 $\hat{W}_j$ 는 정규화되지 않은 입력 특징 벡터 X_j 와 정규화되지 않은 모델 $\bar{\lambda}$ 를 사용하여 첫 단계로 인식을 수행하여 구한다. 그다음 최적의 와핑 파라미터 $\hat{\alpha}_j$ 는 정규화된 음성 모델 $\bar{\lambda}$ 을 사용하여 결정된다.

$$\hat{\alpha}_j = \arg \max_{\alpha} Pr(X_j^\alpha | \bar{\lambda}, \hat{W}_j) \quad (4)$$

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

그림 5. 근사화된 선형 와핑 함수

Fig. 5. piecewise linear warping function

2.6 캡스트럼 평균 차감법

채널왜곡 특성이 음성신호의 관찰구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 캡스트럼의 추정치는 관찰된 신호의 캡스트럼의 평균으로 구해질 수 있다. 이와 같이 긴 시간 동안의 캡스트럼의 평균을 빼줌으로써 채널왜곡의 영향을 제거하는 방식을 CMS(Cepstral Mean Subtraction)이라고 부르며, 다음 수식으로 표현될 수 있다.

$$C_{CMS}^t = c_y^t - m_y, \text{ where } m_y = \frac{1}{N} \sum_{t=1}^N c_y^t \quad (5)$$

여기서 m_y 는 음성의 모든 프레임에서 캡스트럼의 평균

이고, N 는 입력음성의 전체 프레임 수이며, C_{CMS} 는 t 번째 프레임에서 CMS를 통해 보상된 캡스트럼을 의미한다.

2.7 신호 편의 제거

편의(bias)를 제거하기 위한 방법은 ML(Maximum Likelihood) 추정에 의해 유사도를 최대화하는 방법을 이용한다[13]. 현재 추정된 바이어스를 b 라고 하면, b 를 이용하여 보상된 신호 $\tilde{x}_t$ 는

$$\tilde{x}_t = y_t - b \quad (6)$$

이고 보상된 신호에 대한 가장 가까운 모델과 추정된 편의는 다음과 같다.

$$z_t = \mu_i = \arg\max_j p(y_t | b, \lambda_j) = \arg\max_j p(\tilde{x}_t | \lambda_j) \quad (7)$$

$$\bar{b} = \frac{1}{T} \sum_{t=1}^T (y_t - z_t) \quad (8)$$

위의 방법을 이용하여 반복적으로 편의를 구하면 편의는 어떠한 값에 수렴하게 된다.

3. 실험 및 결과

3.1 감정에 따른 파라미터 변화

감정에 따른 음성 파라미터의 변화, 즉 감정이 포함되지 않은 음성과 포함된 음성 스펙트럼간의 차이를 알아보기 위하여 주파수 와핑 파라미터의 변화를 관찰하였다. 이를 위하여 다음과 같은 과정을 통하여 음성 스펙트럼의 차이를 알아보았다.

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

이와 같은 과정을 통하여 구하여진 그래프는 그림 6과 같다. 그림에서 (a)는 평상 감정의 음성을 서로 비교한 경우로 와핑 파라미터의 값이 1.0 부근에 집중되었다. 이는 동일 화자 동일 문장의 평상 음성은 성도 길이의 변화가 거의 없음을 나타낸다. (b)는 평상 감정의 음성들과 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. (c)는 평상 감정 음성들과 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. (d)는 평상 감정 음성들과 화남 감정을 비교한 경우로 와핑 파라미터의 값이 1.0을 중심으

로 넓게 분포하고 있다. 이는 화남 음성의 경우에는 성도 길이의 길이에 약간의 변화가 있다는 것을 의미한다.

그림 6. 감정에 따른 와핑 파라미터 차이 히스토그램
(a) 평상-평상 (b) 평상-기쁨
(c) 평상-슬픔 (d) 평상-화남

Fig. 6. Histogram representing the difference of warping parameter according to the emotions
(a) neutral-neutral (b) neutral-happy
(c) neutral-sad (d) neutral-angry

3.2 데이터베이스

감정 변화에 강인한 음성 인식 시스템의 성능을 평가하기 위해서 데이터 베이스는 다음과 같은 과정으로 구성되었다[10]. 본 연구에서는 인간의 주요 감정인 기쁨, 슬픔, 화남의 3가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4가지 감정을 인식 대상 감정으로 결정하였다. 음성의 녹음은 평상 감정 표현을 훈련하는 아마추어 연극단원 남/녀 각 15명을 대상으로 하였다. 녹음작업은 조용한 사무실 환경에서 이루어졌고 각 화자는 45개의 문장을 4가지 감정으로 녹음하였다. 녹음 동안에 감정 표현이 미흡하다고 판단된 경우에는 다시 녹음을 하였다. 본 연구를 위하여 사용된 데이터의 규모는 16,200(30명×4감정×45문장×3회)문장이다.

3.3 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 멜 캡스트럼 계수, RASTA 처리를 한 멜 캡스트럼 계수, 루트 캡스트럼 계수, PLP 계수를 사용하였다. 또한 특징 파라미터의 시간적인 변화에 대한 정보를 포함하는 델타 캡스트럼을 사용하였다. 실험에 사용된 캡스트럼 계수는 12차를 사용하였고 PLP 계수는 6차를 사용하였다.

3.4 음성 인식 시스템의 구성

본 연구에서는 감정에 따른 음성 파라미터의 성능을 비교하기 위하여 반연속 HMM을 기본으로 하는 화자 독립 단독음 인식 시스템을 구현하였다. 실험에 사용된 음성 인식 시스템은 두 가지로 일반적인 형태의 인식 시스템과 성도

길이 정규화 방법을 사용한 인식 시스템으로 구현하였다. 일반적인 음성 인식 시스템에서 음성 신호는 샘플링되어 고주파 성분이 보강된 후 음성구간 검출이 수행된다. 검출된 음성 신호를 사용하여 음성 파라미터를 구하고 음성에 포함된 편의를 제거하였다. 음성 인식 시스템의 구성에 사용된 반연속 HMM 모델은 256개의 코드를 갖는 코드북을 사용하였고 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 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성 5명)이 각 문장을 3회 발음한 음성을 사용하였다.

성도 길이 정규화 방법을 사용한 인식 시스템의 학습 과정에서는 학습 데이터에 대하여 한개의 밀도함수를 갖는 저해상도의 음성 모델을 사용하여 학습데이터를 정규화하여 정상적인 학습 알고리즘을 사용하여 모델을 학습한다. 인식 과정에서는 입력 음성을 성도 길이 정규화하지 않은 음성모델로 인식하여 문자열을 파악한 후에 성도 길이 정규화된 음성 모델을 사용하여 입력 음성을 정규화하고 마지막 단계에서 성도 길이 정규화된 입력 음성과 모델을 사용하여 인식을 수행한다. 결정 법칙은 비교된 결과를 각 단어 당 기준 모델 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

3.5 실험 결과

본 실험에서는 우선 감정이 포함되지 않은 음성으로 학습한 인식 시스템을 대상으로 4가지 감정이 포함된 테스트 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다. 표 1은 멜 캡스트럼 계수(MEL), 루트 캡스트럼 계수(ROOT_MEL), PLP 계수와 RASTA 멜 캡스트럼 계수(RASTA_MEL)의 감정별 인식 오차를 나타낸다. 여기서 음성 인식 시스템은 평상의 감정만 포함된 데이터로 학습되었기 때문에 인식 데이터가 평상인 경우에 가장 성능이 우수하고 감정이 포함되면 성능이 급격히 저하된다. 표에서 평균값은 4가지 감정에 대한 평균 오차율을 나타낸다. 실험에 사용된 3가지의 음성 파라미터 중에서는 RASTA_MEL이 5.92%로 가장 우수한 성능을 나타내었다.

표 1. 감정에 따른 특징 파라미터의 인식 오차(%)
Table 1. Recognition error rate of feature parameters according to emotions(%)

특징 파라미터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	3.33	14.37	12.30	16.59	11.65
ROOT_MEL	3.89	21.56	19.79	18.47	15.93
PLP	3.33	19.87	14.34	21.22	14.69
RASTA_MEL	1.63	7.85	8.28	6.00	5.92

다음은 신호 편의 제거 방법에 따른 인식 성능 평가를 수행하였다. 편의 제거 방법으로는 SBR과 CMS 방법을 사용하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성으로 학습되었다. 표 2에서 알 수 있듯이 편의 제거를 수행하면 인식 성능이 향상되는 것을 알 수 있다. 특히 CMS가 SBR에 비하여 우수한 성능을 나타내어서 RASTA_MEL과 CMS를 결합한 경우에 3.05%의 오차를 나타내었다.

표 2. 감정에 따른 신호 편의 제거 방법의 인식 오차(%)
Table 2. Recognition error rate of bias removal techniques according to emotions

특징 파라미터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	3.33	14.37	12.30	16.59	11.65
MEL+SBR	0.97	7.75	8.11	6.89	5.93
MEL+CMS	0.37	5.63	4.15	3.48	3.41
RASTA_MEL	1.63	7.85	8.28	6.00	5.92
RASTA_MEL+SBR	1.86	7.72	8.54	6.87	6.25
RASTA_MEL+CMS	1.63	5.58	2.35	2.65	3.05

표 3은 델타 캡스트럼과 성도 길이 정규화 방법에 따른 감정별 인식 성능을 나타낸다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성으로 학습되었다. 델타 캡스트럼은 캡스트럼과 결합하여 사용되었다. 표 3에서 알 수 있듯이 델 캡스트럼의 경우에는 델타 캡스트럼과 결합하여 사용한 경우(MEL+DMEL+CMS)에 평균 오차율이 1.56%로 감소하고 RASTA 델 캡스트럼과 결합한 경우(RASTA_MEL+DMEL+CMS)는 1.27%로 감소하여 가장 우수한 성능을 나타내었다. 이러한 것은 스펙트럼의 시간적 변화 정보를 가지는 델타 캡스트럼이 감정이 포함된 음성의 인식에 도움이 되는 것을 알 수 있다.

표 3에서는 성도 길이 정규화(VTLN) 방법에 따른 성능 평가도 수행하였다. 여기서 성능을 비교할 기준 시스템으로 캡스트럼 평균 차감법을 사용한 멜 캡스트럼 계수와 델타 캡스트럼을 결합한 경우(MEL+DMEL+CMS)와 RASTA 델 캡스트럼(RASTA_MEL+DMEL+CMS)을 사용하였다. 여기에 성도 길이 정규화 방법을 결합한 경우, MEL+DMEL+CMS+VTLN의 평균 인식 오차는 0.78%로 기준 시스템에 비하여 약 50%정도 인식 오차가 감소되었고 RASTA_MEL+DMEL+CMS+VTLN의 평균 인식 오차는 0.67%로 약 47.2% 인식 오차가 감소되었다.

표 3. 감정에 따른 델타 파라미터와 성도 길이 정규화 방법의 인식 오차(%)

Table 3. Recognition error rate of delta parameters and VTLN method according to emotions

특징 파라미터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL+DMEL	0.52	5.56	4.96	6.67	4.43
MEL+DMEL+CMS	0.0	2.22	2.52	1.48	1.56
RASTA_MEL+DMEL	0.15	2.81	4.44	3.33	2.68
RASTA_MEL+DMEL+CMS	0.0	1.78	1.98	1.33	1.27
MEL+DMEL+CMS+VTLN	0.0	1.15	1.07	0.89	0.78
RASTA_MEL+DMEL+CMS+VTLN	0.0	1.05	0.98	0.65	0.67

4. 결 론

본 연구에서는 감정 변화가 음성 인식 시스템의 성능에 미치는 영향을 조사하고, 감정 변화에 영향을 적게 받는 음성 특징 파라미터에 관한 연구를 수행하였다. 본 연구에서

는 멜 캡스트럼 계수, 루트 캡스트럼 계수, PLP 계수와 RASTA 처리를 한 멜 캡스트럼 계수를 사용하였으며 특징 파라미터의 시간적인 변화에 대한 정보를 포함하는 델타 캡스트럼을 사용하였다. 음성에 포함된 편의를 제거하는 방법으로 캡스트럼 평균 차감법(CMS)과 신호 편의 제거법(SBR)을 사용하여 성능을 비교하였다. 감정의 변화에 따른 음성 스펙트럼의 변화를 관찰하여 주파수 왜곡을 포함하는 성도 길이 정규화 방법을 사용하여 성능을 평가하였다.

실험 결과에서 캡스트럼 평균 차감법을 사용한 RASTA 멜 캡스트럼 계수와 델타 캡스트럼을 결합한 음성 파라미터에 성도 길이 정규화 방법을 함께 사용한 경우에 오차를 0.67%의 우수한 성능을 나타내었다. 이러한 것은 성도 길이 정규화를 사용하지 않은 기존 시스템의 오차를 1.27%보다 약 47.2%정도 오차를 감소된 것이다. 이러한 성능 향상의 결과는 성도길이 정규화 방법이 화간의 차이뿐만 아니라 감정에 따른 음성의 차이를 보상해 준 것으로 생각된다. 그러나 성도 길이 정규화 방법은 계산량이 기존 방법에 비하여 많으므로 감정에 따른 변화만을 음성 인식 시스템에 적용하는 방법에 관한 연구가 차후에 이루어져야 할 것이다.

References

- [1] M. Benzeghiba *et al.*, "Impact of variabilities on speech recognition," in *SPECOM'2006, 11th International Conference Speech and Computer*, pp. 3-16, June 2006.
- [2] Y. Sun, Y. Zhou, Q. Zhao, and Y. Yan, "Acoustic Feature optimization for emotion affected speech recognition," *International Conference on Information Engineering and Computer Science*, pp. 1-4, Dec. 2009.
- [3] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, and G. Tong, "Integrating RASTA-PLP into Speech Recognition," *Proc. ICASSP*, pp. 421-424, Apr. 1994.
- [4] B. Schuller, J. Stadermann and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," *Speech Prosody 2006*.
- [5] J. H. Hansen and S. Patil, *Speech under stress: Analysis, modeling and recognition*, Berlin, Heidelberg: Springer-Verlag, pp. 108-137, 2007.
- [6] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech & Audio Processing*, vol. 13, no. 5, pp. 930-944, 2005.
- [7] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," *Proc. of ICASSP, Atlanta, GA*, pp. 339-342, May. 1996.
- [8] L. Welling, R. Haeb-Umbach, X. Aubert and N. Haberland, "A study on speaker Normalization using vocal tract normalization and speaker adaptive training," *Proc. of ICASSP, Seattle, WA*, vol. 2, pp. 797-800, May. 1998.
- [9] S. Molau, S. Kanthak, and H. Ney, "Efficient vocal

tract normalization in automatic speech recognition," *Proc. of the ESSV'00*, Cottbus, Germany, pp. 209-216, 2000.

- [10] B. S. Kang, *Context Independent Emotion Recognition System using Speech Signal*, Yonsei University, Master of Engineering Thesis, 2000.
- [11] M. G. Rahim and 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.
- [12] P. Alexandre, *etc.*, "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments," *Speech Communication*, vol. 12, no. 3, pp. 277-288, 1993.
- [13] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech(RASTA-PLP)," in *Proc. EUROSPEECH*, vol. 3, pp. 1367-1370, Sep. 1991.

저 자 소 개

이규현(Guehyun Lee)
2011년 : 군산대학교 전기공학과 공학사
2011년~현재 : 군산대학교 대학원
전자정보공학부 석사과정

관심분야 : 음성신호처리, 음성인식, 감정인식
Phone : +82-63-469-4745
E-mail : lovely_guehyun@hanmail.net

김원구(Weon-Goo Kim)
1987년 2월 : 연세대 전자공학과 공학사
1989년 8월 : 연세대 전자공학과 공학석사
1994년 2월 : 연세대 전자공학과 공학박사
1994년 9월~현재 : 군산대학교 전기공학과 교수
1998년 9월~1999년 9월 : Bell Lab, Lucent Technologies(USA) 객원연구원
2008년 9월~2009년 12월 : Griffith University(Australia) 방문교수

관심분야 : 음성신호처리, 음성인식, 감정인식, 음성변환, 화자 인식
Phone : +82-63-469-4745
E-mail : wgkim@kunsan.ac.kr