퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 6, pp. 655-660

감정 변화에 강인한 음성 인식 파라메터

Robust Speech Recognition Parameters for Emotional Variation

김 원 구

Weon-Goo Kim

군산대학교 전자정보공학부

요 약

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 하여 감정 변화의 영향을 적게 받는 음성 인식 시스템의 특징 파라메터에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 인식 시스템의 성능에 미치는 영향에 관한 연구와 감정 변화의 영향을 적게 받는 음성 인식 시스템의 특징 파라메터에 관한 연구를 수행하였다. 본 연구에서는 LPC 켑스트럼 계수, 멜 켑스트럼 계수, 루트 켑스트럼 계수, PLP 계수와 RASTA 처리를 한 멜 켑스트럼 계수와 음성의 에너지를 사용하였다. 또한 음성에 포함된 편의(bias)를 제거하는 방법으로 CMS와 SBR 방법을 사용하여 그 성능을 비교하였다.

실험 결과에서 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 HMM 기반의 화자독립 단어 인식기의 오차가 7.05%로 가장 우수한 성능을 나타내었다. 이러한 것은 멜 켑스트럼을 사용한 기준 시스템과 비교하여 59%정도 오차가 감소된 것이다.

Abstract

This paper studied the feature parameters less affected by the emotional variation for the development of the robust speech recognition technologies. For this purpose, the effect of emotional variation on the speech recognition system and robust feature parameters of speech recognition system were studied using speech database containing various emotions. In this study, LPC cepstral coefficient, mel-cepstral coefficient, root-cepstral coefficient, PLP coefficient, RASTA mel-cepstral coefficient were used as a feature parameters. And CMS and SBR method were used as a signal bias removal techniques.

Experimental results showed that the HMM based speaker independent word recognizer using RASTA mel-cepstral coefficient and its derivatives and CMS as a signal bias removal showed the best performance of 7.05% word error rate. This corresponds to about a 52% word error reduction as compare to the performance of baseline system using mel-cepstral coefficient.

Key words : 음성 신호, 음성 인식, 감정 변화, HMM, 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, 유럽국가들의 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[12-18].

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

접수일자 : 2005년 10월 10일
완료일자 : 2005년 12월 10일
감사의 글 : 본 논문은 한국과학재단 지역대학우수과학자지원연구(R05-2003-000-12043-0)연구비 지원으로 수행되었습니다.

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 왔으나 여러 가지 감정이 포함된 음성에 대한 연구는 아직 초보 단계이다. "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[1-11]. 지금까지의 연구는 음성 합성시 인간의 감정을 포함시키는 감정 합성 분야와 음성에 포함된 감정을 추출하는 감정 인식에 관한 연구가 주로 진행되고 있다. 그러나 기쁨, 화남, 슬픔, 두려움, 혐오감 등의 감정들을 표현할 때 독특한 형태로 변화하는 음성의 특성 때문에 발생하는 음성 인식 시스템의 성능 저하에 관하여는 연구가 거의 이루어지지 않고 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 연구에서는 인간의 감정 변화에 강인한 음성 인식 기술 개발을 목표로 하여 감정 변화의 영향을 적게 받는 음성 인식 시스템의 특징 파라메터에 관한 연구를 수행하였다. 우선 감정 변화에 강인한 특징 파라메터에 대한 연구를 수행하여 기존 특징 파라메터를 비교하고 감정 변화에 강인한 파라메터를 찾는 연구를 수행하였다.

2. 음성 특징 파라메터

음성 인식에 널리 사용되고 있는 특징 벡터로는 오래 전부터 사용되어온 LPC 켑스트럼 계수(LPC cepstral coefficient)와 멜(mel) 켑스트럼 계수가 주로 사용되고 있으며 잡음에 강인한 특징 벡터로 루트(root) 켑스트럼 계수, PLP(Perceptually Linear Prediction) 계수와 RASTA (RelAtive SpecTrAl) 처리를 한 특징 파라메터 특징 벡터들이 있다[12-18]. 잡음에 강인한 거리 측정 방법으로는 가중 켑스트랄 거리 측정 방법(weighted cepstral distance measure) 방법이 주로 사용되고 있다.

2.1 켑스트럼 계수

켑스트럼 계수를 구하는 과정은 그림 1과 같다. 여기서 켑스트럼 계수 $\{c_n\}$은 선형 예측 계수와 최소 자승 평균 오차를 사용하여 순환적으로 구할 수 있다. 이렇게 구한 계수를 LPC 켑스트럼 계수라 한다.

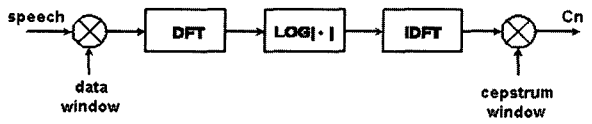

그림 1. 켑스트럼 분석 과정
Fig. 1. cepstrum analysis procedure

멜(mel)을 기반으로 한 켑스트럼은 DFT 또는 FFT 크기를 멜과 주파수 사이의 대응 관계에 따라 주파수 축에서 와핑(warping)하여 이의 대수 값을 역 DCT하여 8에서 14차 정도의 계수를 구한다. 예를들어, 로그 에너지 출력을 X_k라 하면 M개의 멜 켑스트럼 계수는 다음과 같이 나타내어진다.

$$c_n = \frac{1}{20}\sum_{k=1}^{20} X_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{20}\right],\quad n=1,\cdots,M. \qquad (1)$$

2.2 델타(delta) 켑스트럼 계수

델타 켑스트럼 계수 $d_k(t)$는 t번째 구간의 k번째 켑스트럼 계수를 $c_k(t)$라 할 때, 다음과 같이 나타낼 수 있다. 여기서, δ는 시간 간격을 나타낸다.

$$d_k(t) = c_k(t+\delta) - c_k(t-\delta) \qquad (2)$$

2.3 루트 켑스트럼 계수

Lockwood 등은 Mel-based 루트 켑스트럼 계수가 잡음에 의한 변형에 강인한 것을 관찰하였고 root 함수로 일반적인 로가리듬 역컨볼루션을 근사화하였다.[17]. 그림 2는 일반적인 LFCC와 LPCC를 루트 호모모픽 접근 방법으로 통합된 음성 신호 분석 블록도이다.

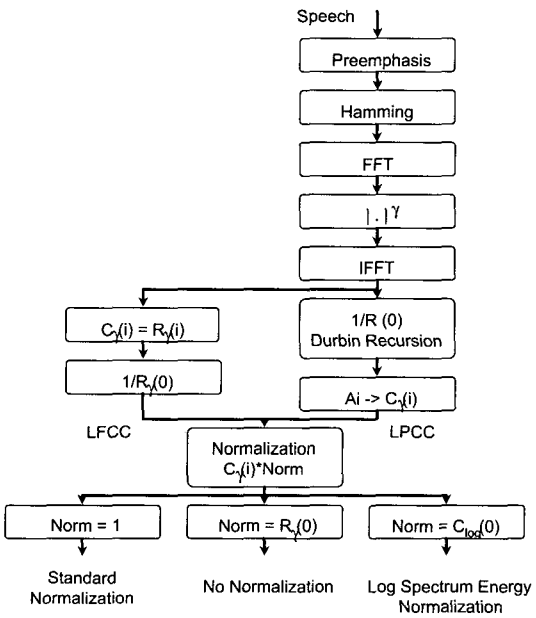

그림 2. 루트 켑스트럼 분석 과정
Fig. 2. root cepstrum analysis procedure

2.4 PLP(Perceptually Linear Prediction) 계수

PLP 분석 방법은 1982년 Hermansky에 의해 제안되었으며, 음성 신호의 파워 스펙트럼을 변화시켜 청각 특성이 고려된 스펙트럼을 이용한다. 이러한 단계를 거쳐 얻어지는 저차의 스펙트럼은 인간이 실제 감지하는 소리와 유사한 특성을 갖게 되며, 음성 인식에 적용되어 좋은 성능을 보여주었다[16]. 위에서 설명한 PLP 분석 방법의 흐름도를 그림 3에서 보여주고 있다.

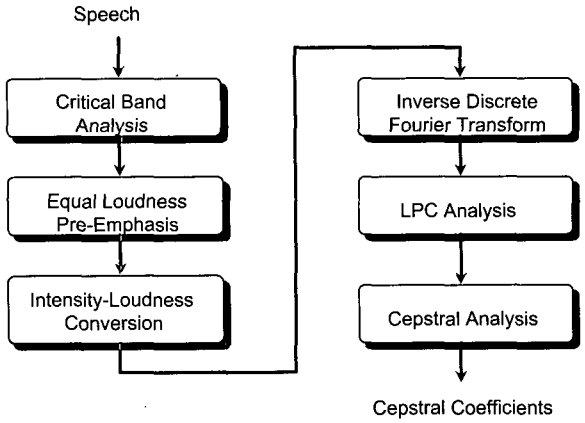

그림 3. PLP 분석 방법
Fig. 3. PLP analysis method

2.5 RASTA(RelAtive SpecTrAl) 처리

RASTA-PLP 분석 방법에서는 일반적인 단구간 스펙트럼(short-term absolute spectrum)을 사용하는 대신 스펙트럼 성분 중 시간에 따라 천천히 변화하는 성분을 배제하는 대역 통과 스펙트럼(band-pass filtered spectrum)을 사용한다. RASTA-PLP 분석 방법의 흐름도는 그림 4와 같다[15,16].

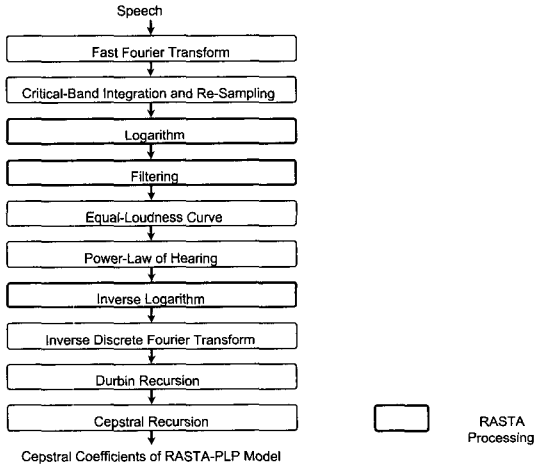

그림 4. RASTA-PLP 분석 방법
(RASTA 분석 구간은 굵은 선)
Fig. 4. RASTA-PLP analysis method

그림 4의 흐름도에서 필터링 블록은 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링(bandpass filtering)하는 것과 같다. 이 대역 통과 필터의 전달 함수는 다음과 같다.

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

2.6 신호 편의 제거 방법

2.6.1 캡스트랄 평균 차감법

채널왜곡 특성이 음성신호의 관찰구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 켑스트럼의 추정치는 관찰된 신호의 켑스트럼의 평균으로 구해질 수 있다. 이와 같이 긴 시간 동안의 켑스트럼의 평균을 빼줌으로써 채널왜곡의 영향을 제거하는 방식을 CMS (Cepstral Mean Subtraction)이라고 부르며, 다음 수식으로 표현될 수 있다. 여기서 m_y 는 음성의 모든 프레임에서 켑스트럼의 평균이고, $N(s)$는 입력 음성의 전체 프레임 수이며, C^t_{comp}는 t 번째 프레임에서 CMS를 통해 보상된 켑스트럼을 의미한다.

$$C^t_{comp} = c^t_y - m_y, \text{ where } m_y = \frac{1}{N(s)}\sum_{t=1}^{N(s)} c^t_y \quad (4)$$

2.6.2. ML(Maximum likelihood) 방법에 의한 SBR

편의(bias)를 제거하기 위한 방법은 ML(Maximum Likelihood) 추정에 의해 유사도를 최대화하는 방법을 이용한다[18]. 현재 추정된 바이어스를 b 라고 하면, b를 이용하여 보상된 신호 $\bar{x}_t$는

$$\bar{x}_t = y_t - b \quad (5)$$

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

$$z_t = \mu_i = \arg\max_j p(y_t|b,\lambda_j) = \arg\max_j p(\bar{x}_t|\lambda_j) \quad (6)$$

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - z_t) \quad (7)$$

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

3. 실험 및 결과 고찰

3.1 데이터베이스

감정 변화에 강인한 음성 인식 시스템의 성능을 평가하기 위해서는 다양한 감정이 포함된 음성 데이터베이스가 필요하다. 이러한 데이터베이스는 다음과 같은 과정으로 구성되었다.

데이터베이스를 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 4가지 감정으로 녹음하였고 녹음 동안에 감정 표현이 미흡하다고 판단된 경우에는 다시 녹음을 하였다. 본 연구을 위하여 사용된 데이터의 규모는 5400(30명×4감정×45문장×1회)문장이다.

3.2 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 해밍 창(Hamming window)을 사용하여 10ms씩 이동하면서

특징 파라메터를 구한다. 본 연구에서는 음성의 특징 파라메터로 LPC 켑스트럼 계수, 멜 켑스트럼 계수, 루트 켑스트럼 계수, PLP 계수와 RASTA 처리를 한 멜 켑스트럼 계수와 음성의 에너지를 사용하였다. 또한 특징 파라메터의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼과 델타 에너지를 사용하였다. 실험에 사용된 켑스트럼 계수는 12차를 사용하였고 PLP 계수는 5차를 사용하였다. 또한 음성에 포함된 편의를 제거하는 방법으로 CMS와 SBR 방법을 사용하여 그 성능을 비교하였다.

3.3 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. 음성 신호는 샘플링되어 고주파 성분이 보강된 후 음성구간 검출을 수행된다. 검출된 음성 신호를 사용하여 음성 파라메터를 구하고 음성에 포함된 편의(bias)를 제거하기 위한 편의 제거 방법을 사용하였다.

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)의 음성이 사용되었고 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성5명)을 사용하였다.

입력 특징 파라메터는 다양한 거리 측정 방법과 반연속 HMM을 사용하여 기준 패턴과 유사도를 측정한다. 이때 기준 패턴은 각 문장마다 4가지 감정이 모두 포함된 하나의 HMM 모델을 사용하는 경우와 각 문장마다 각각의 감정으로 학습된 4개의 모델을 사용하는 경우로 구분하였다. 스펙트럼간의 비교 또는 매칭 방법으로 가중 켑스트럼에 의한 거리 측정 방법이 사용되었고 결정 법칙은 비교된 결과를 각 단어당 기준 패턴 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 패턴을 입력 음성의 단어로 결정한다.

3.4 감정 음성의 특징 분석

일반적으로 평상, 기쁨, 슬픔, 화남의 각 감정이 포함된 음성 신호와 몇 가지 파라메터(피치 변화, 피치범위, 피치 최대값, 발음속도)를 비교해 보면 대체로 표 1과 같은 특징을 발견할 수 있다. 가장 두드러진 특징은 기쁨과 화남의 경우에는 전체적으로 피치(기본 주파수)와 에너지가 높고 발음 속도가 빠른 반면, 슬픔의 경우에는 전체적으로 피치(기본 주파수)와 에너지가 낮고 발음 속도가 느리다는 것을 알 수 있다.

표 1. 감정별 특징 비교

Table 1. feature comparison according to emotions

특성 / 감정	피치 변화	피치 범위	피치 최대값	발음 속도
평상	완만	좁음	낮음	보통
기쁨	급격	넓음	높음	빠름
슬픔	완만	좁음	낮음	느림
화남	급격	매우 넓음	매우 높음	매우 빠름

그림 5에서는 본 연구에 사용된 데이터베이스로부터 평상 감정의 음성 길이를 1로 정규화였을 때 기쁨, 슬픔, 화남의 평균 음성 길이 변화를 나타낸 것이다. 그림에서 기쁨 및 화남이 감정이 포함된 음성은 평상보다 약간 10%정도 길이가 증가하였고 슬픔의 경우에는 20%정도 길이가 증가되었다. 이것은 같은 문장에 대하여 감정에 따른 음성 길이 변화를 구한 것이므로 음성의 길이가 길다는 것은 발음 속도가 느리다는 것을 의미한다. 따라서 음성 슬픔의 감정이 포함된 음성의 발음 속도가 약 20%정도 느리다는 것을 알 수 있다. 그러나 기쁨과 화남의 경우에는 일반적으로 평상시보다 발음속도가 빨라진다는 것과 다르게 약 10%정도 문장 길이가 긴 것을 알 수 있다. 이러한 이유는 본 연구에서 사용된 데이터베이스는 화자로 하여금 동일한 문장에 대하여 4가지 감정을 포함하여 발음하도록 하였기 때문에 음성에 기쁨과 화남의 감정을 포함시키기 위하여 음성의 특정 부분을 강조하여 길게 발음하는 특성 때문이라고 판단된다.

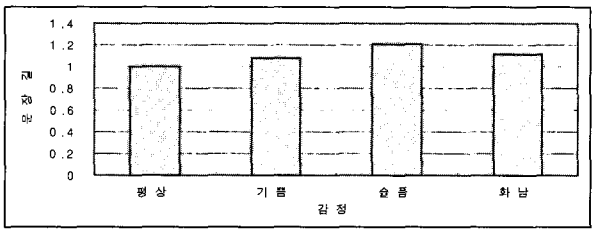

그림 5. 감정별 문장 길이 비율

Fig. 5. sentence length ratio according to emotions

3.5 실험 결과

본 실험에서는 우선 감정이 포함되지 않은 음성으로 학습한 인식 시스템을 대상으로 테스트 음성에 4가지 감정이 포함된 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다. 표 2는 각 음성 파라메터와 감정별 인식 성능을 나타낸다. 여기서 음성 인식 시스템은 평상의 감정만 포함된 데이터로 학습되었기 때문에 인식 데이터가 평상인 경우에 가장 성능이 우수하고 감정이 포함되면 인식 성능이 급격히 저하된다. 표에서 평균값은 4가지 감정에 대한 평균 인식률을 나타낸다. 실험에 사용된 5가지의 음성 파라메터 중에서는 RASTA 멜 켑스트럼이 86.83%로 가장 우수한 성능을 나타내었다. 이러한 것은 멜 켑스트럼의 인식률 80.56% 보다 약 6%이상 높은 인식 성능인데, 이것은 RASTA 처리 과정이 음성의 감정 변화에 따른 스펙트럼의 변화를 보상해 주는 효과가 있다고 볼 수 있다. 표에 사용된 기호는 다음과 같다.

CEP : LPC 켑스트럼 계수,
MEL : 멜 켑스트럼 계수,
ROOT_MEL : 루트 켑스트럼 계수
RASTA_MEL : RASTA 처리를 한 멜 켑스트럼 계수
PLP : PLP 계수
ENG : 에너지

표 2. 감정에 따른 특징 파라메터의 성능 평가
Table 2. Performance of feature parameters according to emotions

특징파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
CEP	89.33	72.67	73.78	73.11	77.22
MEL	92.89	74.89	78.89	75.56	80.56
ROOT_MEL	91.11	64.67	66.89	68.67	72.84
RASTA_MEL	97.11	82.22	82.89	85.11	86.83
PLP	92.89	66.67	75.56	62.22	74.34

다음은 거리 측정 방법에 따른 성능 평가 실험을 수행하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었다. 여기서 EUC는 가중이 모두 1이고 RPS 는 선형 리프터이고 BPL은 밴드 패스 리프터이다. 표 3에서 BPL이 가장 89.61%로 가중을 사용하지 않은 경우보다 약 3%정도 인식 성능이 향상되었다. 이러한 것은 특징 파라메터에 곱하여주는 가중 함수가 스펙트럼을 스무딩 하는 효과가 있기 때문에 음성의 감정 변화에 따른 스펙트럼의 변화를 어느 정도 보상해 주는 효과가 있다고 볼 수 있다.

표 3. 거리측정 방법에 따른 특징 파라메터의 성능 평가
Table 3. Performance of distance measures and feature parameters according to emotions

특징 파라메터 \ 감정		평상	기쁨	슬픔	화남	평균
MEL	EUC	92.89	74.89	78.89	75.56	80.56
	BPL	94.44	80.22	81.33	85.11	85.28
	RPS	92.67	76.00	76.44	81.56	81.67
RASTA_MEL	EUC	97.11	82.22	82.89	85.11	86.83
	BPL	97.33	86.22	82.89	92.00	89.61
	RPS	96.44	83.11	81.56	88.22	87.33

다음은 에너지 파라메터를 사용하였을 때의 성능 평가 실험을 수행하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었고 거리 측정 방법의 가중은 BPL이 사용되었다. 에너지 파라메터는 켑스트럼과 결합하여 사용되었다. 표 4에서 알 수 있듯이 감정이 포함된 음성을 사용한 경우에 에너지 파라메터의 사용은 음성 인식 시스템의 성능을 크게 저하시킨다.

표 4. 감정에 따른 에너지 파라메터의 성능 평가
Table 4. Performance of energy parameters according to emotions

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.44	80.22	81.33	85.11	85.28
MEL+ENG	88.89	67.78	62.67	70.67	72.50

다음은 델타 켑스트럼을 사용했을 때의 성능 평가 실험을 수행하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었고 거리 측정 방법의 가중은 BPL이 사용되었다. 델타 켑스트럼은 켑스트럼과 결합하여 사용되었다. 표 5에서 알 수 있듯이 멜 켑스트럼의 경우에는 델타 켑스트럼과 결합하여 사용한 경우에 평균 인식률이 1.5%정도 감소하였으나 RASTA 멜 켑스트럼의 경우에는 델타 켑스트럼을 사용한 경우에 인식률이 1.8% 정도 증가하여 91.39%의 성능을 나타내었다. 이러한 것은 RASTA 처리 과정이 감정의 변화 때문에 발생하는 스펙트럼의 시간적인 변화를 어느 정도 보상해 주는 효과가 있다고 볼 수 있다.

표 5. 감정에 따른 델타 파라메터의 성능 평가
Table 5. Performance of delta parameters according to emotions

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.44	80.22	81.33	85.11	85.28
MEL+DMEL	95.33	80.44	75.56	84.22	83.89
RASTA_MEL	97.33	86.22	82.89	92.00	89.61
RASTA_MEL+DMEL	98.67	90.22	83.78	92.89	91.39

다음은 신호 편의 제거 방법에 따른 인식 성능 평가를 수행하였다. 편의 제거 방법으로는 ML 방법을 사용한 SBR과 CMS 방법을 사용하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었고 거리 측정 방법의 가중은 BPL이 사용되었다. 표 6에서 알 수 있듯이 편의 제거를 수행하면 인식 성능이 향상되는 것을 알 수 있다. 특히 CMS가 SBR에 비하여 우수한 성능을 나타내어서 93.28%의 인식률을 보였다. 이러한 것은 감정의 변화에 따라 음성에 편의가 발생한다는 것을 의미하고 편의 제거 과정을 통하여 이러한 변화를 정도 보상해 주는 효과가 있다고 볼 수 있다.

표 6. 감정에 따른 신호편의 제거 방법의 성능 평가
Table 6. Performance of bial removal techniques according to emotions

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.44	80.22	81.33	85.11	85.28
MEL+SBR	96.44	86.44	84.44	90.00	89.33
MEL+CMS	97.78	89.11	86.22	93.78	91.72
RASTA_MEL	97.33	86.22	82.89	92.00	89.61
RASTA_MEL+SBR	95.78	86.44	81.78	90.00	88.50
RASTA_MEL+CMS	97.33	89.56	90.89	95.33	93.28

다음은 위에서 수행한 여러 가지 처리를 모두 결합하여 최적화 한 경우의 성능을 평가하였다. 표 7에서 알 수 있듯이 RASTA_MEL 멜 켑스트럼과 델타 켑스트럼을 사용하고 거리측정 방법으로는 BPL을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 93.95%로 가장 우수한 성능을 나타내었다. 이러한 것은 멜 켑스트럼과 거리측정 방법으로

표 7. 감정 변화에 강인한 특징 파라메터의 성능 평가
Table 7. Performance of robust feature parameters for emotional variation

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.44	80.22	81.33	85.11	85.28
MEL+DMEL+CMS	97.56	83.78	79.56	86.44	86.84
RASTA_MEL+CMS	97.33	89.56	90.89	95.33	93.28
RASTA_MEL+DMEL+CMS	99.11	91.78	89.33	95.56	93.95

BPL을 사용한 경우의 인식 성능 85.28%를 기준 시스템으로 할 때 8.67%의 인식률 향상을 나타내고 오차의 감소율로 계산하면 약 59%정도 오차가 감소된다고 볼 수 있다.

4. 결 론

본 연구에서는 다양한 감정이 포함된 음성 데이터를 사용하여 감정 변화가 음성 인식 시스템의 성능에 미치는 영향을 조사하고, 감정 변화에 영향을 적게 받는 음성 특징 파라메터에 관한 연구를 수행하였다. 먼저 감정 변화가 음성 인식 시스템의 성능에 미치는 영향에 관하여 연구를 수행하였다. 또한 감정 변화의 영향을 적게 받는 음성 인식 시스템의 특징 파라메터에 관한 연구를 수행하였다. 본 연구에서는 LPC 켑스트럼 계수, 멜 켑스트럼 계수, 루트 켑스트럼 계수, PLP 계수와 RASTA 처리를 한 멜 켑스트럼 계수와 음성의 에너지를 사용하였다. 또한 특징 파라메터의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼과 델타 에너지를 사용하였다. 또한 음성에 포함된 편의를 제거하는 방법으로 CMS와 SBR 방법을 사용하여 그 성능을 비교하였다. 실험 결과에서 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 거리측정 방법으로는 BPL을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 93.95%로 가장 우수한 성능을 나타내었다. 이러한 것은 멜 켑스트럼과 거리측정 방법으로 BPL을 사용한 경우 인식 성능 85.28%를 기준 시스템으로 할 때 8.67%의 인식률 향상을 나타내고 오차의 감소율로 계산하면 약 59%정도 감소되었다.

참 고 문 헌

[1] Noam Amir,"Classifying Emotions in Speech: a Comparison of Methods", *Proceedings of Eurospeech '2001,* Vol. 1, pp. 127-130, Aalborg, Denmark, 2001

[2] A. Nogueiras, etc,"Speech Emotion Recognition using Hidden Markov Models", *Proceedings of Eurospeech '2001,* Vol. 4, pp. 2679-2682, Aalborg, Denmark, 2001

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

[4] Janet E. Cahn, "The Generation of Affect in Synthesized Speech", *Journal of the American Voice I/O Society,* Vol. 8, pp. 1-19, July 1990.

[5] 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.

[6] Iain R. Murray and John 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.

[7] 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.

[8] Michael Lewis and Jeannette M. Haviland, *Handbook of Emotions,* The Guilford Press 1993.

[9] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition,* Prentice-Hall Inc., 1993.

[10] S. Young, "A Review of Large-Vocabulary Continuous-Speech Recognition", *IEEE Signal Processing Magazine,* Vol. 13, No. 5, pp. 45-47, 1996.

[11] L. R. Rabiner, "A Tutorial on HMMs and Selected Applications in Speech Recognition", *Proc. IEEE,* Vol. 77, No. 2, pp. 257-285, 1989.

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

[13] A. Acero, ect, "Environmental Robustness in Automatic Speech Recognition," *in Proc. ICASSP,* pp. 849-852, April 1990.

[14] H. Hermansky, N. Morgan, H. G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise based RASTA Spectral Processing", *in Proc. ICASSP,* pp. 83-86, 1993.

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

[16] 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 Proc. EUROSPEECH,* vol. 3, pp. 1367-1370, Sep. 1991.

[17] P.Alexandre, ect. "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments", *Speech Communication,* vol. 12, no. 3, pp. 277-288, 1993.

[18] 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.

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 전자정보공학부 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