Journal of Institute of Control, Robotics and Systems (2010) 16(12):1137-1142
DOI:10.5302/J.ICROS.2010.16.12.1137 ISSN:1976-5622

감정에 강인한 음성 인식을 위한 음성 파라메터

Speech Parameters for the Robust Emotional Speech Recognition

김 원 구*
(Weon-Goo Kim[1])
[1]Kunsan National University

Abstract: This paper studied the speech parameters less affected by the human emotion for the development of the robust 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 and frequency warped mel-cepstral coefficient were used as feature parameters. And CMS (Cepstral Mean Subtraction) method were used as a signal bias removal technique. Experimental results showed that the HMM based speaker independent word recognizer using vocal tract length normalized mel-cepstral coefficient, its derivatives and CMS as a signal bias removal showed the best performance of 0.78% word error rate. This corresponds to about a 50% word error reduction as compare to the performance of baseline system using mel-cepstral coefficient, its derivatives and CMS.

Keywords: robust speech recognition, speech parameter, vocal tract length normalization

I. 서론

인간과 기계사이의 보다 편리한 인터페이스로 음성 인식 기술의 사용이 급격히 증가하고 있다. 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 음성 인식 기술이 아직도 가지고 있는 문제점은 이러한 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[1-4]. 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 한 가지로 연구하고 있으며, 유럽 국가들의 ESPRIT 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, 모바일 환경 등의 잡음 처리에 관한 연구가 오래 전부터 진행되어 왔다[1-7].

이와 함께 음성 인식 시스템의 성능에 영향을 미치는 요인으로 인간의 심리적 변화가 있다. 음성 신호의 형태는 인간의 감정 상태에 따라서 변화하여 평상시 발음과 기쁨, 슬픔, 화남, 우울 등의 상태에서 발음한 것이 크게 다르다. 현재의 음성 인식 시스템들이 평상시 감정 상태에서 발음한 음성 데이터를 사용하여 만들어졌기 때문에 인간의 감정이 포함된 음성을 인식하는 경우에는 그 성능이 저하된다. 이와 관련되어 외국에서도 감정이 포함된 음성에 대한 음성 인식 시스템의 성능을 향상시키기 위한 연구가 오래전부터 진행되어 왔다. 강세가 있는 음성이나 롬바드 효과를 갖는 음성에 대한 인식 성능 향상에 관한 연구나 감정이 포함된 음성 모델을 사용한 연구가 진행되어 왔다[8-10]. 그러나 기쁨, 화남, 슬픔, 두려움, 혐오감 등의 감정들을 표현할 때 독특한 형태로 변화하는 음성의 특성 때문에 발생하는 음성 인식 시스템의 성능 저하에 관하여는 연구가 아직 체계적으로 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 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술을 개발하기 위하여 감정 변화의 영향을 적게 받는 음성 파라메터 개발에 관한 연구를 수행하였다. 이를 위하여 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정이 음성 신호와 음성 인식 시스템의 성능에 미치는 영향을 관찰하였다. 또한 이러한 변화가 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 관찰하였다. 우선 감정 변화에 강인한 특징 파라메터에 대한 연구를 수행하여 기존 음성 파라메터를 비교하고 감정 변화에 강인한 파라메터를 찾는 연구를 수행하였다. 본 연구에서 사용된 음성 파

* 책임저자(Corresponding Author)
논문접수: 2010. 9. 10., 수정: 2010. 10. 7., 채택확정: 2010. 12. 1.
김원구: 군산대학교 전기공학과(wgkim@kunsan.ac.kr)
※ 본 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구결과임(중견연구자지원사업[핵심] No. 2007-0056927).

라메터는 멜 켑스트럼 계수, 델타 켑스트럼 계수, RASTA 멜 켑스트럼 계수,와 주파수 와핑(frequency warping)된 멜 켑스트럼 계수 등이 사용되었고 음성 신호에 포함된 편의(bias) 제거를 위하여 켑스트럼 평균 차감법이 사용되었다.

본 논문의 구성은 다음과 같다. II 장에서는 여러 가지 음성 파라메터에 관하여 설명하고 III 장에서는 감정이 음성에 미치는 영향에 관하여 분석한다. IV 장에서는 다양한 실험을 통하여 여러 가지 음성 파라메터의 성능을 비교 분석한다. 마지막으로 V 장에서는 결론을 맺는다.

II. 음성 파라메터

음성 인식에 널리 사용되고 있는 특징 벡터로는 멜 켑스트럼 계수가 주로 사용되고 있으며 시간 정보를 사용한 델타 켑스트럼 계수와 RASTA 멜 켑스트럼 계수가 널리 사용되고 있다. 또한 화자의 성도 길이 차이를 보상하는 주파수 와핑을 사용한 음성 파라메터도 음성 인식에서 우수한 성능을 나타내고 있으며 음성 신호에 포함된 편의를 제거를 위하여 켑스트럼 평균 차감법이 사용되고 있다.

1. RASTA (RelAtive SpecTrAl) 멜 켑스트럼 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 1과 같다[3,5].

그림 1의 흐름도에서 필터링 블록은 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링하는 것과 같다. 이 대역 통과 필터의 전달 함수는 다음과 같다.

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

2. 주파수 와핑된 멜 켑스트럼 계수

그림 2는 전통적인 멜 켑스트럼 분석 방법(a)과 주파수 축 와핑을 추가한 분석 방법(b)을 나타낸다. 전통적인 방법인 그림 2(a)에서 음성 신호는 프리엠파시스(pre-emphasis)와

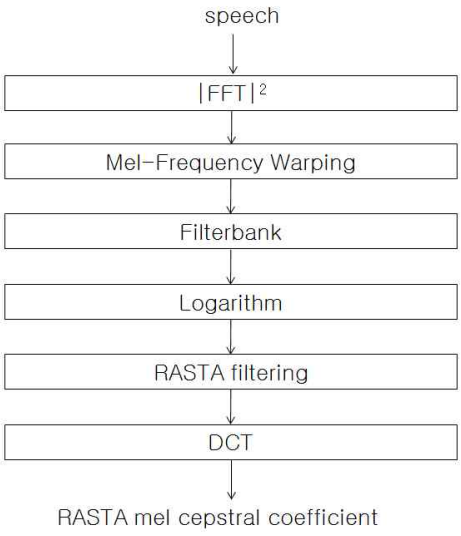


그림 1. RASTA 멜 켑스트럼 계수 추출 방법.

Fig. 1. Extraction method of RASTA mel-cepstral coefficient.

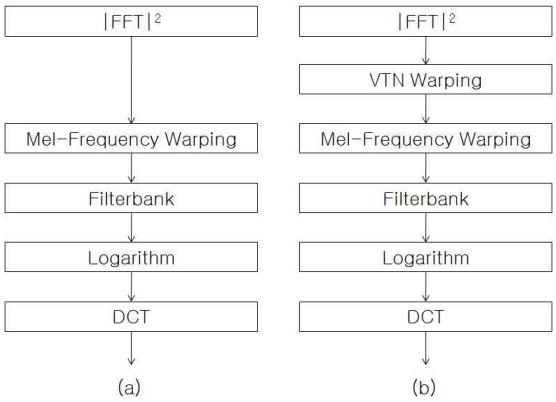


그림 2. 전통적인 멜 켑스트럼 (a)과 와핑 함수를 추가한 방법 (b).

Fig. 2. Scheme of traditional MFCC (a) and integrated method with warping function (b).

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 2(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 [12]과 Welling [13] 등은 식 (2)와 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. 이러한 와핑 함수는 스펙트럼을 위쪽으로 또는 아래쪽으로 이동시키는 역할을 수행한다. Molau[14] 등은 이러한 와핑 함수들의 성능을 비교한 연구를 수행하여 근사화된 선형 함수가 와핑 함수로 가장 우수한 성능을 나타냄을 보였다. 근사화된 선형 와핑 함수는 그림 3과 같다.

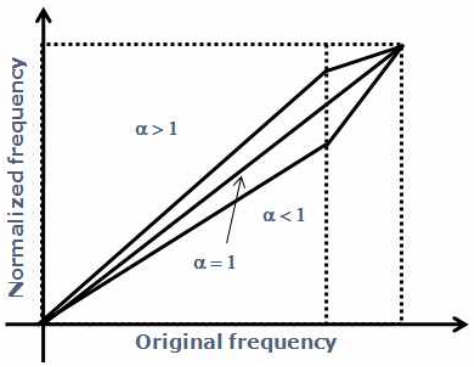


그림 3. 근사화된 선형 와핑 함수.

Fig. 3. piecewise linear warping function.

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

3. 성도길이 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[11]. 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^α를 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.

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

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.

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

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

4. 켑스트럼 평균 차감법

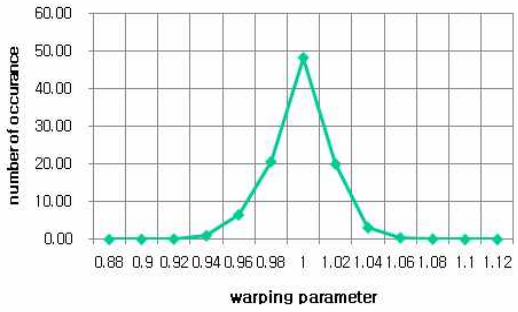
채널왜곡 특성이 음성신호의 관찰구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 켑스트럼의 추정치는 관찰된 신호의 켑스트럼의 평균으로 구해질 수 있다. 이와 같이 긴 시간 동안의 켑스트럼의 평균을 빼줌으로써 채널왜곡의 영향을 제거하는 방식을 CMS (Cepstral Mean Subtraction)이라고 부르며, 다음 수식으로 표현될 수 있다.

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

여기서 my 는 음성의 모든 프레임에서 켑스트럼의 평균이고, N(s)는 입력음성의 전체 프레임 수이며, CCMS는 t번째 프레임에서 CMS를 통해 보상된 켑스트럼을 의미한다.

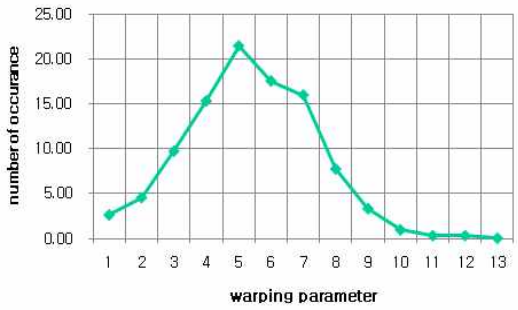
III. 감정에 따른 음성 변화

본 논문에서는 감정에 따른 음성 스펙트럼의 변화를 연구하였다. 감정이 포함되지 않은 음성과 포함된 음성 스펙트럼간의 차이를 알아보기 위하여 주파수 와핑 파라메터의

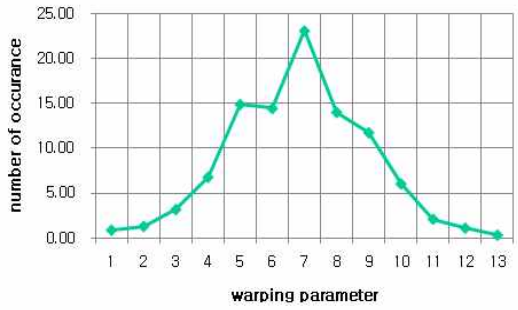


(a) Neutral-neutral.

(b) Neutral-happy.



(c) Neutral-sad.



(d) Neutral-angry.

그림 4. 평상 음성과 감정 음성 사이의 차이를 나타내는 히스토그램.

Fig. 4. Histogram that representing the difference between neutral and emotional speech.

변화를 관찰하였다. 이를 위하여 다음과 같은 과정을 통하여 음성 스펙트럼의 차이를 알아보았다.

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 그래프를 그린다.

이와 같은 과정을 통하여 구하여진 그래프는 그림 4와 같다. 그림 4(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 4(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 4(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 4(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.

따라서 입력 음성에 감정이 포함된 경우에 평상 음성으로 학습한 음성인식 시스템과 불일치가 발생하여 음성인식 시스템의 성능이 저하되는 문제를 발생시킨다. 이러한 감정에 따른 주파수 스펙트럼의 변화를 음성 인식 시스템에 적용함으로서 감정이 포함된 음성을 인식하는데 강인한 음성 인식 시스템을 개발할 수 있다.

IV. 실험 및 결과

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[15]. 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.

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 창을 사용하여 10ms씩 이동하면서 특징 파라메터를 구한다. 본 연구에서는 음성의 특징 파라메터로 멜 켑스트럼 계수, RASTA 처리를 한 멜 켑스트럼 계수를 사용하였다. 또한 특징 파라메터의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼을 사용하였다. 실험에 사용된 켑스트럼 계수는 12차를 사용하였고 음성에 포함된 편의를 제거하는 방법으로 CMS 방법을 사용하여 그 성능을 비교하였다.

3. 감정에 따른 음성의 변화

감정은 음성의 특성을 변화시킨다. 감정에 따라 음성 신호의 피치, 발음속도, 에너지, 스펙트럼 등 다양하게 영향을 받는다. 그림 5는 감정에 따른 스펙트럼의 변화를 나타내는 스펙트로그램이다. 그림에서 감정이 평상인 그림 5(a) 경우에 비하여 그림 5(b)의 기쁨인 경우에 스펙트럼의 변화가 큰 것을 알 수 있다. 그림 5(c)는 슬픈 감정이 포함된 경우로 스펙트럼의 변화가 평상시의 발음에 비하여 완만함을 알 수 있다. 그림 5(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.

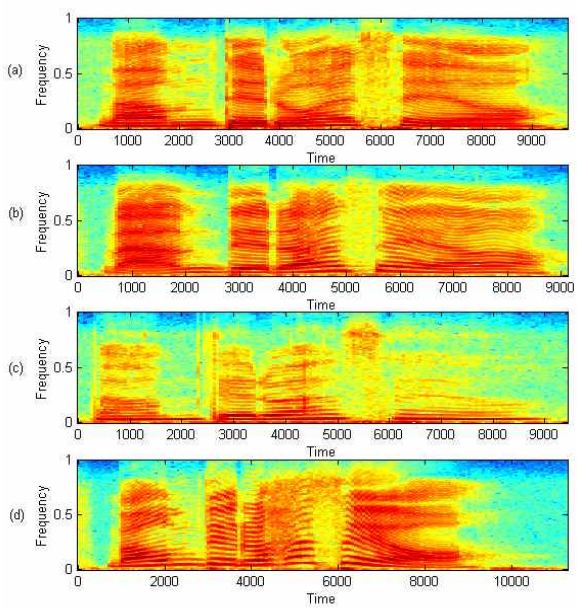


그림 5. 여성 화자의 음성 “마음대로 하세요”의 감정별 스펙트로그램 (a) 평상 (b) 기쁨 (c) 슬픔 (d)화남.

Fig. 5. Spectrogram of female speech signal “ma-eum-dae-ro-ha-se-yo” according to the emotion (a) Neutral (b) Happy (c) Sad (d) Angry.

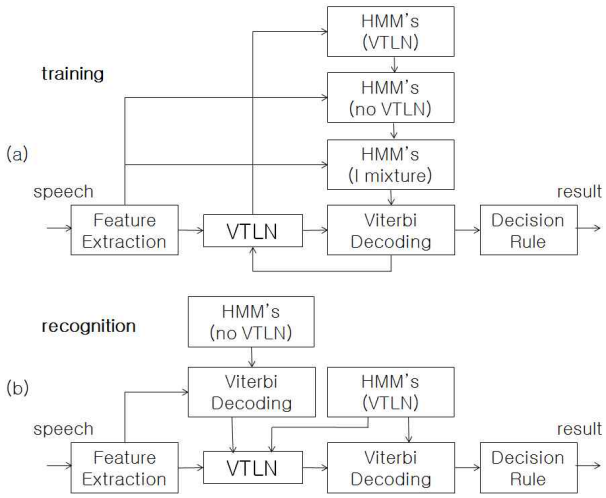


그림 6. 성도길이 정규화 방법 기반의 음성 인식 시스템 구조 (a) 학습 (b) 인식.

Fig. 6. The Structure of speech recognition system based on VTLN (a) Training (b) Recognition.

4. 음성 인식 시스템의 구성

본 연구에서는 감정에 따른 음성 파라메터의 성능을 비교하기 위하여 반연속 HMM을 기본으로 하는 화자 독립 단독음 인식 시스템을 구현하였다. 실험에 사용된 음성 인식 시스템은 일반적인 형태의 시스템과 성도 길이 정규화 방법을 사용한 인식 시스템(그림 6)으로 구현하였다. 음성 신호는 샘플링되어 고주파 성분이 보강된 후 음성구간 검출을 수행된다. 검출된 음성 신호를 사용하여 음성 파라메터를 구하고 음성에 포함된 편의 제거 방법을 사용하였다. 음성 인식 시스템의 구성에 사용된 반연속 HMM 모델은 256개의 코드을 갖는 코드북을 사용하였고 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 여성 5명)이 각 문장을 3회 발음한 음성을 사용하였다.

그림 6의 성도 길이 정규화 방법을 사용한 인식 시스템의 학습 과정에서는 학습 데이터에 대하여 1개의 밀도함수를 갖는 저해상도의 음성 모델을 사용하여 학습데이터를 정규화하여 정상적인 학습 알고리즘을 사용하여 모델을 학습한다. 인식 과정에서는 입력 음성을 성도 정규화하지 않은 음성모델로 인식하여 문자열을 파악한 후에 성도 정규화된 음성 모델을 사용하여 입력 음성을 정규화하고 마지막 단계에서 성도 길이 정규화된 입력 음성과 모델을 사용하여 인식을 수행한다. 결정 법칙은 비교된 결과를 각 단어당 기준 모델 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

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 1은 멜 켑스트럼 계수(MEL), RASTA 멜 켑스트럼 계수(RASTA_MEL)와 켑스트럼 평균 차감법을 적용한 멜 켑스트럼 계수(CMS_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 인식 성능이 급격히 저하된다. 표에서 평균값은 4가지 감정에 대한 평균 오차율을 나타낸다. 실험에 사용된 3가지의 음성 파라메터 중에서는 CMS_MEL이 3.41%로 가장 우수한 성능을 나타내었다. 이러한 것은 음성에 포함된 편의를 제거하는 켑스트럼 평균 차감법이 감정이 포함된 음성을 인식하는데도 효과가 있다고 볼 수 있다.

표 1. 감정에 따른 특징 파라메터의 인식 오차(%).

Table 1. Recognition error rate of feature parameters according to emotions(%).

감정 / 특징파라메터	평상	기쁨	슬픔	화남	평균
MEL	3.33	14.37	12.30	16.59	11.65
RASTA_MEL	1.63	7.85	8.28	6.00	5.92
CMS_MEL	0.37	5.63	4.15	3.48	3.41

표 2. 감정에 따른 델타 파라메터의 인식 오차(%).

Table 2. Recognition error rate of delta parameters according to emotions.

감정 / 특징 파라메터	평상	기쁨	슬픔	화남	평균
MEL	3.33	14.37	12.30	16.59	11.65
MEL+DMEL	0.52	5.56	4.96	6.67	4.43
RASTA_MEL	1.63	7.85	8.28	6.00	5.92
RASTA_MEL+DMEL	0.15	2.81	4.44	3.33	2.68
CMS_MEL	0.37	5.63	4.15	3.48	3.41
CMS_MEL+DMEL	0.0	2.22	2.52	1.48	1.56

다음은 델타 켑스트럼을 사용했을 때의 성능 평가 실험을 수행하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었다. 델타 켑스트럼은 켑스트럼과 결합하여 사용되었다. 표 2에서 알 수 있듯이 멜 켑스트럼의 경우에는 델타 켑스트럼과 결합하여 사용한 경우(MEL+DMEL)에 평균 오차율이 4.43%로 감소하고 RASTA 멜 켑스트럼과 결합한 경우(RASTA_MEL+DMEL)는 2.68%로 감소하였고 켑스트럼 평균 차감법과 결합한 경우(CMS_MEL+DMEL)에는 1.56%로 가장 우수한 성능을 보여주었다. 이러한 것은 스펙트럼의 시간적 변화 정보를 가지는 델타 켑스트럼이 감정이 포함된 음성의 인식에 도움이 되는 것을 알 수 있다.

표 3은 성도 정규화 방법에 따른 감정별 인식 성능을 나타낸다. 여기서 성능을 비교할 기준 시스템으로 켑스트럼 평균 차감법을 사용한 멜 켑스트럼 계수와 델타 켑스트럼을 결합한 경우(CMS_MEL+DMEL)을 사용하였다. 여기에 성도 길이 정규화 방법을 결합한 경우(CMS_MEL+DMEL+VTLN)의 평균 인식 오차는 0.78%로 기준 시스템에 비하여

표 3. 성도 정규화 방법에 따른 인식 오차(%).

Table 3. Recognition error rate according to the VTLN method(%).

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
CMS_MEL+DMEL	0.0	2.22	2.52	1.48	1.56
CMS_MEL+DMEL+VTLN	0.0	1.15	1.07	0.89	0.78

약 50%정도 인식 성능이 향상 되었다.

V. 결론

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. 본 연구에서는 멜 켑스트럼 계수와 RASTA 처리를 한 멜 켑스트럼 계수를 사용하였으며 특징 파라메터의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼을 사용하였다. 또한 음성에 포함된 편의를 제거하는 방법으로 켑스트럼 평균 차감법을 사용하여 그 성능을 비교하였다. 감정의 변화에 따른 주파수 분석에서 주파수 와핑이 감정 변화에 따른 변화를 보상해 줄 것으로 관찰하여 주파수 와핑 과정을 포함하는 성도 길이 정규화 방법을 사용하여 그 성능을 평가하였다.

실험 결과에서 켑스트럼 평균 차감법을 사용한 멜 켑스트럼 계수와 델타 켑스트럼을 결합한 음성 파라메터에 성도 길이 정규화 방법을 함께 사용한 경우에 오차율 0.78%의 우수한 성능을 나타내었다. 이러한 것은 성도 길이 정규화를 사용하지 않은 기준 시스템의 오차율 1.56%보다 약 50%정도 오차율이 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 계산양이 기존 방법에 비하여 많으므로 감정에 따른 변화만을 음성 인식 시스템에 적용하는 방법에 관한 연구가 차후에 이루어져야 할 것이다.

참고문헌

[1] M. Benzeghiba and et al., “Impact of variabilities on speech recognition,” *in SPECOM'2006, 11th International Conference Speech and Computer,* pp. 3-16, June 2006.

[2] D. O'Shaughnessy “Invited paper: Automatic speech recognition: History, methods and challenges,” Pattern Recognition, vol. 41, no 10, pp. 2965-2979, Oct. 2008.

[3] H. Hermansky, N. Morgan, and H. G. Hirsch, “Recognition of speech in additive and convolutional noise based RASTA spectral processing,” *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2, pp. 83-86, 1993.

[4] 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 2009*, pp. 1-4, Dec. 2009.

[5] 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.

[6] B. Schuller, J. Stadermann1, and G. Rigoll, “Affect-robust speech recognition by dynamic emotional adaptation,” *Speech Prosody 2006*, May 2006.

[7] J. H. Hansen and S. Patil, “Speech under stress: Analysis, modeling and recognition,” Berlin, Heidelberg: Springer-Verlag, pp. 108-137, 2007.

[8] N. Amir, “Classifying emotions in speech: a comparison of methods,” *Proc. of Eurospeech 2001*, Aalborg, Denmark, vol. 1, pp. 127-130, 2001.

[9] A. Nogueiras, etc, “Speech emotion recognition using Hidden Markov Models,” *Proc. of Eurospeech 2001*, Aalborg, Denmark, vol. 4, pp. 2679-2682, 2001.

[10] R. W. Picard, “Affective computing,” The MIT Press 1997.

[11] 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.

[12] S. Wegmann, D. McAllaster, J. Orlofl, and B. Peskin, “Speaker normalization on conversational telephone speech,” *Proc. 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,” *Proc. of ICASSP*, Seattle, WA, vol. 2, pp. 797-800, May 1998

[14] S. Molau, S. Kanthak, and H. Ney, “Efficient vocal tract normalization in automatic speech recognition,” in *Proc. of of the ESSV'00*, Cottbus, Germany, pp. 209-216, 2000

[15] 강봉석, “음성 신호를 이용한 문장독립 감정 인식 시스템,” 연세대학교 석사학위 논문, 2000.

김 원 구

1987년 연세대 전자공학과 졸업. 1989년 동 대학원 석사. 1994년 동 대학 박사. 1994년~현재 군산대학교 전기공학과 교수. 1998년~1999년 Bell lab, Lucent Technologies (USA) 객원 연구원. 2008년~2009년 호주 Griffith 대학 교환교수. 관심분야는 음성신호처리, 음성인식, 음성변환, 감정인식, 화자인식.