Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

감정 변화에 강인한 음성 인식

Robust Speech Recognition for Emotional Variation

김 원 구

전북 군산시 군산대학교 전자정보공학부
E-mail: wgkim@kunsan.ac.kr

요 약

본 논문에서는 인간의 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 영향과 감정 변화의 영향을 적게 받는 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.

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

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. 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.

음성 인식 시스템의 성능에 영향을 미치는 요인가운데 하나로 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 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]. 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 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. 잡음에 강인한 거리 측정 방법으로는 가중 켑스트랄 거리 측정 방법(weighted cepstral distance measure) 방법이 주로 사용되고 있다.

2.1 루트 켑스트럼 계수

Lockwood 등은 Mel-based 루트 켑스트럼 계수가 잡음에 의한 변형에 강인한 것을 관찰하였고 root 함수로 일반적인 로가리듬 역컨볼루션을 근사화하였다.. 그림 1은 일반적인 LFCC와 LPCC를 루트 호모모픽 접근 방법으로 통합된 음성 신호 분석 블록도이다.

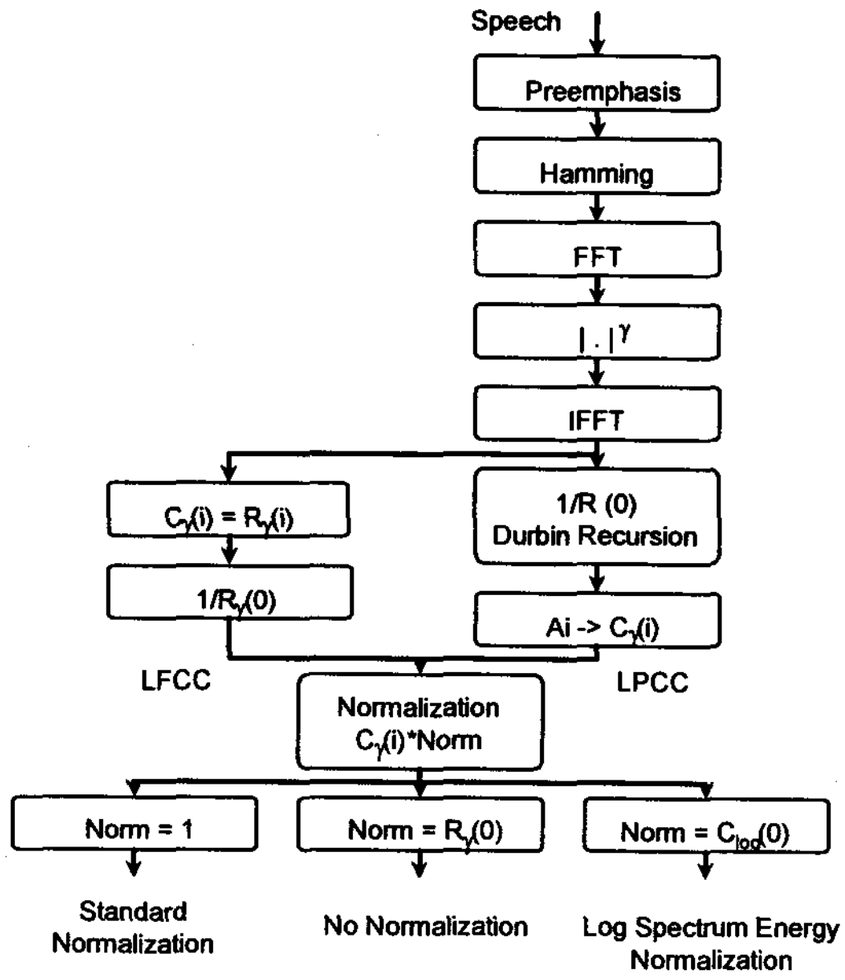

그림 1. 루트 켑스트럼 분석 과정

2.2 PLP 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. 위에서 설명한 PLP 분석 방법의 흐름도를 그림 2에서 보여주고 있다.

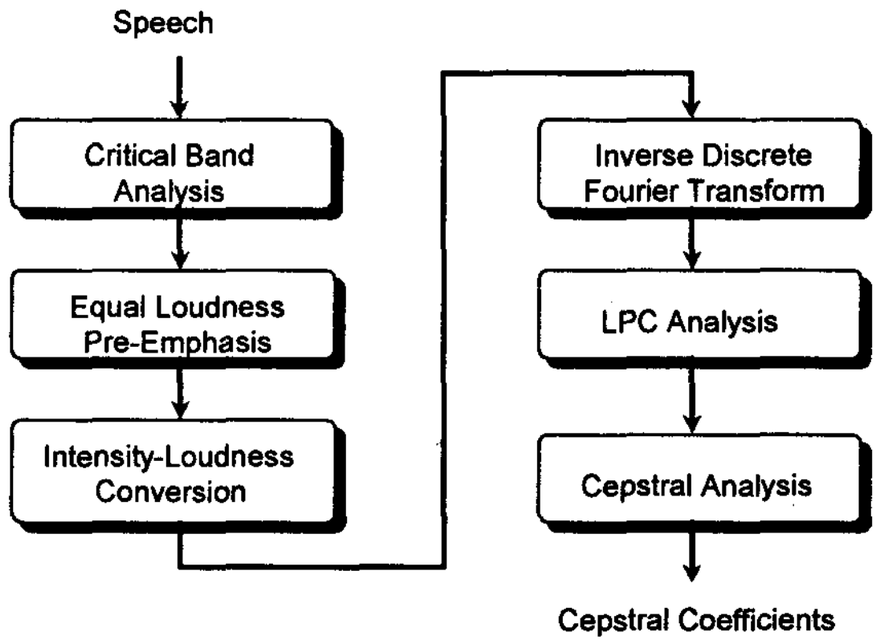

그림 2. PLP 분석 방법

2.3 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 3과 같다.

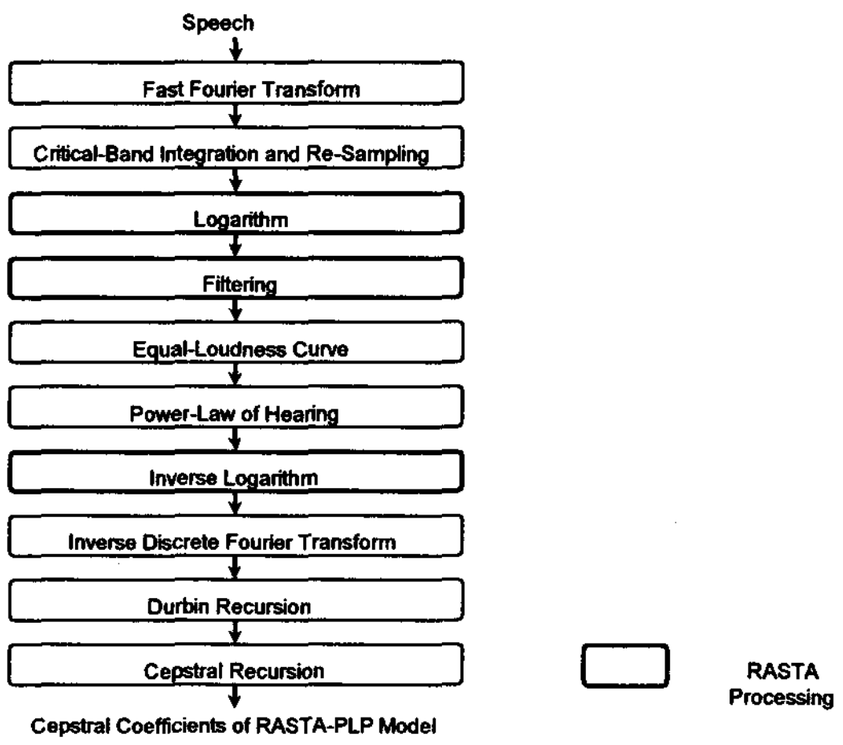

그림 3. RASTA-PLP 분석 방법
(RASTA 분석 구간은 굵은 선)

그림 3의 흐름도에서 필터링 블록은 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링(bandpass filtering)하는 것과 같다.

2.4 ML 방법에 의한 SBR

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

$$\tilde{x}_t = y_t - b \quad (1)$$

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

$$z_t = \mu_i = \arg\max_j p(y_t|b,\lambda_j) = \arg\max_j p(\tilde{x}_t|\lambda_j) \quad (2)$$

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - z_t) \quad (3)$$

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

3. 시뮬레이션 및 결과 고찰

3.1 데이터베이스

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. 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. 실험에 사용된 켑스트럼 계수는 12차를 사용하였고 PLP 계수는 5차를 사용하였다.

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.

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.

3.4 실험 결과

표 1은 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 86.8%로 가장 우수한 성능을 나타내었다. 표에 사용된 기호는 다음과 같다.

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

표 1. 감정에 따른 특징 파라메터의 성능 평가

감정 / 특징파라메터	평상	기쁨	슬픔	화남	평균
CEP	89.3	72.7	73.8	73.1	77.2
MEL	92.9	74.9	78.9	75.6	80.6
ROOT_MEL	91.1	64.7	66.9	68.7	72.8
RASTA_MEL	97.1	82.2	82.9	85.1	86.8
PLP	92.9	6667	75.6	62.2	74.3

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 2에서 BPL이 가장 89.6%로 가중을 사용하지 않은 경우보다 약 3%정도 인식 성능이 향상되었다.

표 2. 거리측정 방법에 따른 특징 파라메터의 성능 평가

감정 / 특징 파라메터		평상	기쁨	슬픔	화남	평균
MEL	EUC	92.9	74.9	78.9	75.6	80.6
	BPL	94.4	80.2	81.3	85.1	85.3
	RPS	92.7	76.0	76.4	81.6	81.7
RASTA_MEL	EUC	97.1	82.2	82.9	85.1	86.8
	BPL	97.3	86.2	82.9	92.0	89.6
	RPS	96.4	83.1	81.6	88.2	87.3

다음은 신호 편의 제거 방법에 따른 인식 성능 평가를 수행하였다. 표 3에서 알 수 있듯이 편의 제거를 수행하면 인식 성능이 향상되는 것을 알

수 있다. 특히 CMS가 SBR에 비하여 우수한 성능을 나타내어서 93.3%의 인식률을 보였다.

표 3. 감정에 따른 신호편의 제거방법의 성능 평가

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.4	80.2	81.3	85.1	85.3
MEL+SBR	96.4	86.4	84.4	90.0	89.3
MEL+CMS	97.8	89.1	86.2	93.8	91.7
RASTA_MEL	97.3	86.2	82.9	92.0	89.6
RASTA_MEL +SBR	95.8	86.4	81.8	90.0	88.5
RASTA_MEL +CMS	97.3	89.6	90.9	95.3	93.3

다음은 위에서 수행한 여러 가지 처리를 모두 결합하여 최적화 한 경우의 성능을 평가하였다. 표 4에서 알 수 있듯이 RASTA_MEL 멜 켑스트럼과 델타 켑스트럼을 사용하고 거리측정 방법으로는 BPL을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 94.0%로 가장 우수한 성능을 나타내었다.

표 4. 감정 변화에 강인한 특징 파라메터의 성능 평가

특징 파라메터 \ 감정	평상	기쁨	슬픔	화남	평균
MEL	94.4	80.2	81.3	85.1	85.3
MEL+DMEL+ CMS	97.6	83.8	79.6	86.4	86.8
RASTA_MEL +CMS	97.3	89.6	90.9	95.3	93.3
RASTA_MEL +DMEL+CMS	99.1	91.8	89.3	95.6	94.0

4. 결 론

본 연구에서는 감정 변화에 영향을 적게 받는 음성 특징 파라메터에 관한 연구를 수행하였다. 실험 결과에서 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 거리측정 방법으로는 BPL을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 94.0%로 가장 우수한 성능을 나타내었다. 이러한 것은 멜 켑스트럼을 사용한 경우 인식 성능 85.3%를 기준 시스템으로 할 때 8.7%의 인식률 향상을 나타내고 오차의 감소율로 계산하면 약 59%정도 감소되었다.

감사의 글

이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(R11-2002-105-04004-0(2007))

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.