RASTA-PLP의 음소 모델 단어 인식기 적용

허창원, 전성채, 하판봉
창원대학교 전자공학과

Phoneme-Model Word Recognizer on RASTA-PLP

Changwon Heo, Sungchae Jeon, Panbong Ha

Dept. of Electronics Engineering, Changwon National University
e-mail : bilite@sarim.changwon.ac.kr, jusc@sarim.changwon.ac.kr, pha@sarim.changwon.ac.kr

요 약

대부분의 음성 파라미터 추정 기법은 통신 채널의 주파수 응답에 의해 쉽게 영향을 받는다. 이 논문에서 우리는 음성에서 그러한 안정상태의 스펙트럼 계수에 있어서 좀더 강인한 기법인 RASTA-PLP 방법을 적용하여 파라미터를 추출하고 그 파라미터를 연속 HMM 인식기의 입력으로 사용하여 문맥독립 (context-independent) 음소 모델을 훈련하는 과정에서 최적의 모델을 찾게 된다. 여기서는 ETRI 445 DB에 RASTA-PLP를 적용하였을 때 가장 좋은 성능을 나타내는 재추정 횟수와 mixture 수를 찾는 데 목표를 둔다.

문맥독립 음소 모델은 한국어의 발성학적 근거를 토대로 하고, 여기에 묵음 (silence)을 추가하여 총 40개로 정의하였다. 문맥독립 음소 모델은 3개의 상태를 가지는 전형적인 left-to-right CHMM(Continuous Hidden Markov Model)을 이용하여 훈련한다. 그리고 훈련 시간을 줄이기 위해 Viterbi beam 탐색법을 적용한다.

1. 서 론

음성인식 시스템을 구현할 때 특징 파라미터로서 LPC (Linear Predictive Coefficients), LPC-CC (Cepstral Coefficients), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Prediction) 등 여러 가지가 저마다의 이점을 가지고 이용되어 왔다. 음성 발성 모델의 계수로 초기 음성인식 파라미터로 많이 쓰였던 것이 LPC이다. 음성의 여기신호가 비선형적으로 결합, 즉 콘볼루션 (convolution) 결합되어 있다고 할 때 음성신호에서 각 성분을 분리해 내기 위해 사용한 것이 켑스트럼 계수(CC)인데 이를 LPC 계수로부터 계산해 내었기 때문에 LPC-CC라 부른다. MFCC는 사람의 청각 계통이 인식하는 주파수가 물리적인 주파수와 선형적으로 대응하지 않는다는 것을 해결하기 위해 mel-scale 주파수를 도입한 방법이다. 이런 계수들은 프레임 간의 동적인 특성을 반영하기 위해 delta MFCC, delta-delta MFCC를 사용하기도 한다. PLP 분석법은 음성 신호의 고주파 해상도를 감소시키며 LPC 보다 적은 차수의 계수로 비슷한 인식율을 나타낸다. 이는 사람이 듣는 주파수 대역을 강조시켜서 최종적으로 전극점 모델링하는 방법을 구사하고 있다. 이 PLP는 Hermansky에 의해 제안되었다.

본 논문에서는 역시 Hermansky에 의해 제안된 RASTA-PLP 방법으로 구한 계수를

분백독립 음소 모델 기반의 연속 HMM 인식기의 특징 파라미터로 인가할 때 가장 좋은 성능을 나타내는 연속 HMM의 mixture 수와 재추정 횟수를 구하는 기초연구를 행하는데 초점을 둔다. 이것을 바탕으로 문맥종속 모델로 확장하여 연속음 인식 시스템으로 확장하려 한다.

2. RASTA-PLP 특징 파라미터 추출

PLP 음성해석 기법은 음성의 단구간 스펙트럼에 기초하고 있다. 음성의 단구간 스펙트럼은 나중에 몇가지 인체 물리학적인 기반의 스펙트럼 변환법에 의해 수정될지라도 단구간 스펙트럼 값이 통신채널의 주파수 특성에 의해 변화될 때에는 약점이 잡히게 된다. 사람의

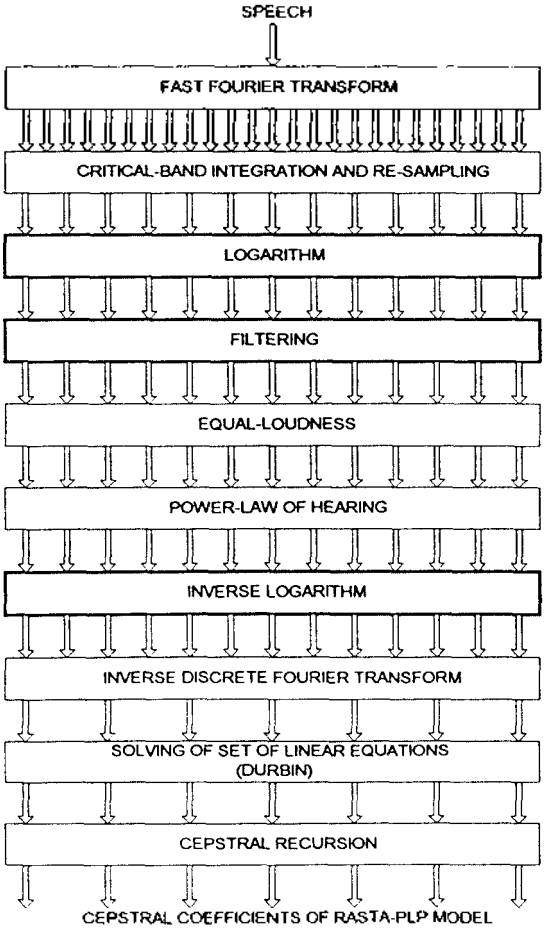

그림 1. RASTA-PLP 파라미터 추출방법

음성 지각은 그러한 정상적 (steady-state)인 스펙트럼 계수에 대해 덜 예민하다. 여기서 선형 스펙트럼 왜곡에 대해 더 강인한 PLP (그리고 어떤 다른 단구간 스펙트럼 기반의 방법에 대해서도 가능함)를 만드는 RelAtive SpecTrAl (RASTA) 방법을 도입했다.

평범한 단구간 절대 스펙트럼을 대역통과 필터링된 것으로 교체했다. 여기서 필터는 영 (zero) 주파수에서 예리하게 영 (zero)으로 떨어지는 대역통과 필터이다. 각 주파수 채널에서 어떤 상수 성분이나 천천히 변하는 성분이 이 동작에 의해 제거되기 때문에 새로운 스펙트럼 추정값은 단구간 스펙트럼에서 천천히 변하는 것에 대해 덜 민감하다. 로그 스펙트럼 영역에서 필터링이 이루어졌을 때, 제거된 상수 스펙트럼 성분은 입력 음성신호에 대해서는 읽혀있는 효과를 반영한다. 그림 1은 RASTA-PLP 파라미터 추출과정을 나타내고 있다.

음성 데이터는 그림 1과 같은 추출과정을 거친후 9차의 RASTA-PLP 모델의 계수를 추출하게 된다.

3. 음소 모델 구성 및 훈련

음소 모델은 한국어 발성시 나올 수 있는 음소 단위로 분류, 표기하고, 여기에 묵음 (silence) 모델을 포함하여 40 개로 정의하였다.

정의된 음소 모델은 15명이 발성한 ETRI 445 데이터를 이용하여 훈련을 수행하며, 모든 음소모델은 3개의 상태를 갖는 전형적인 left- to-right CHMM으로 정의되어 있다. 또한 훈련시에 각 단어의 전후에 묵음 부분을 첨가, 사용하였다. 음소 모델은 정의되어진 음소사전과 레이블링 되지 않은 훈련 데이터를 가지고 Baum-Welch 재추정 과정을 통하여 각 모델에 대한 초기화 작업 및 훈련이 이루어진다.

지금까지 언급한 훈련순서는 그림 2와 같다.

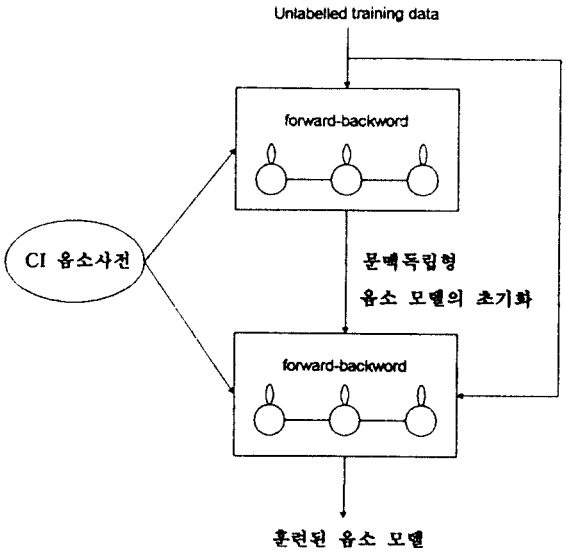

그림 2. 음소 모델 훈련과정

4. 실험

학습에 사용된 데이터는 7 kHz의 LPF를 거쳐 16 kHz 표본화 주파수로, 16 bit로 A/D 변환되어진 ETRI에서 배부된 445개의 한국어 고립단어 데이터베이스를 이용했다. ETRI 445 균일 음소분포 단어 중에서 15명 (2회 발성분)이 발성한 데이터를 사용하였다.
일반적으로 재추정 횟수를 고정시켜 훈련하는데, 이 실험에서는 likelyhood 값이 일정한 값으로 수렴하는 데까지의 횟수를 측정했다. 그리고 이와 같은 방법으로 mixture 수를 변화시켜 가면서 최적의 mixture 수를 찾아보았다.

5. 결론

본 논문에서는 RASTA-PLP 파라미터를 음성인식에 적용할 때, HMM의 구조를 결정하는 데 있어서 적절한 재추정 횟수와 mixture 수를 찾고자 했다. 이는 앞으로의 음성인식 실험에 있어서 기초 자료로 활용될 수 있을 것이다. 비록 여기서 정한 정보가 화자독립 인식실험에서 꼭 최상의 정보를 제공한다고는 말할 수 없지만 일반적인 기준은 될 수 있을 것으로 본다. 앞으로의 연구방향은 여기서 얻은 자료를 가지고 RASTA-PLP를 이용해 화자독립 인식실험을 해보고, MFCC를 이용한 인식실험의 결과[9]와 비교하고자 한다. 그리고 다른 특성의 마이크와 다양한 SN 비의 음성을 대상으로 실험함으로써 잡음환경에 많이 노출된 음성에는 어떤 방법이 더 효과적인지를 살펴보고자 한다.

* 본 논문에 사용한 DB는 한국통신이 출연하여 한국전자통신연구소가 구축한 445 단어 데이터베이스를 사용하였습니다.

6. 참고문헌

[1] Hynek Hermansky and Nelson Morgan, "RASTA Processing of Speech", IEEE Trans. on Speech and Audio Processing, vol. 2, pp. 578-589, 1994.

[2] 서영주, 성철재, 이정철, 한민수, 이영직, "음성학적 지식에 기반한 한국어 변이음 집단화 수형도의 구현", 제13회 음성통신 및 신호처리 워크샵, pp. 344-347, 1996.

[3] 허 웅, 국어 음운학, 샘문화사.

[4] S. J. Young, N. H. Russell and J. H. S. Thornton, "Token Passing: a Conceptual Model for Connected Speech Recognition Systems", Sued Technical Report F_INENG/TR38, Cambridge University, 1989.

[5] B. H. Juang and L. R. Rabiner, "Mixture Autoregressive Hidden Markov Models for Speech Signals", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 33, pp. 1404-1413, 1985.

[6] Yunxin Zhao, "A Speaker-Independent Continuous Speech Recognition System

Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units", IEEE Trans. on Speech and Audio Processing, vol. 1, pp. 345-361, 1993.

[7] L. Labiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[8] Mei-Yuh Hwang and Xuedong Huang, "Shared-Distribution Hidden Markov Models for Speech Recognition", IEEE Trans. on Speech and Audio Processing, vol. 1, pp. 414-420, 1993.

[9] 전성채, 하판봉, "어휘독립 한국어 단어 인식기", 한국음향학회 학술발표대회 논문집, vol. 15, pp. 55-58, 1996.