GMM을 이용한 프레임 단위 분류에 의한 우리말 음성의 분할과 인식

권호민*, 한학용*, 고시영**, 허강인*
동아대학교 전자공학과*, 경일대학교 전자정보공학과**

Korean Speech Segmentation and Recognition by Frame Classification via GMM

Ho-Min Kwon*, Hag-Yong Han*, Si-young Koh**, Kang-In Hur*
Dept. of Electronics, Dong-A University*
Dept. of Electronic & Information Engineering, Kyung-il University**
hmkwon@donga.ac.kr

Abstract

In general it has been considered to be the difficult problem that we divide continuous speech into short interval with having identical phoneme quality. In this paper we used Gaussian Mixture Model (GMM) related to probability density to divide speech into phonemes, an initial, medial, and final sound. From them we performed continuous speech recognition. Decision boundary of phonemes is determined by algorithm with maximum frequency in a short interval. Recognition process is performed by Continuous Hidden Markov Model(CHMM), and we compared it with another phoneme divided by eye-measurement. For the experiments result we confirmed that the method we presented is relatively superior in auto-segmentation in korean speech.

Ⅰ. 서론

연속된 음성신호를 동일한 음운 특성을 갖는 소구간으로 나누는 것을 세그멘테이션(이하분할)이라 하며 음성신호처리의 주요한 과제 중의 하나이다. 그러나, 음성 분할은 분할 단위에 대한 정확한 정보와 지식이 필요하며, 발화자의 발음 습관 혹은 심리상태 등과 같은 발화자간에 존재하는 개인성 때문에 각 분할 단위에 존재하는 공통적인 음성 정보와 조음 결합 등이 고려된 정확한 분할 단위의 경계점을 찾는다는 것은 어려운 작업이다.

최근의 대부분의 음성인식 시스템의 경우에는 음성의 시변성이 포함된 모델을 이용함으로써 이러한 분할에 관한 연구가 상대적으로 미흡한 실정이다. 대표적인 음성인식도구인 HTK(HMM Tool Kit)의 경우에도 인식 단위에 대한 모델을 생성하는 단계에서 발음 사전을 바탕으로 음성인식의 단위를 얼라인먼트하여 초기 음소모델을 구성하고 다이폰이나 트라이폰 모델 등으로 확장하여 사용하기도 하지만 안정된 발음 사전의 구성과 데이터의 확보에 어려움이 있다.

그러나, 프레임 단위의 인식을 통한 인식시스템의 경우, 조음결합 부분의 처리와 자음의 인식률 저하가 치명적인 단점으로 지적되고 있지만 참조 모델의 수를 최소로 하는 음소모델을 이용할 수 있으며 연속음성인식으로 확장이 용이하다는 등의 이점이 있다. 또한, 전처리 단계로 음소분할을 통하여 HMM과 같은 인식기로 인식을 하는 방법도 고려할 수 있는데 이 경우에도 최소의 모델을 이용할 수 있으며 신뢰할 만한 음소단위의 분할이 이루어진다면 기존의 인식시스템의 성능향상 뿐만 아니라 적절한 후처리를 통하여 연속음성인식 시스템으로의 확장도 가능할 것이다.

음성분할로의 접근은 확률모델, Fuzzy, ANN, HMM 등의 패턴 매칭 방법으로 음소인식을 통하여 훈련된 데이터에 의해 이루어지는 간접적인 처리방법과 시간영역의 음성 파라메터인 ZCR(Zero Crossing Rate), LCE(Level Crossing Rate), PVR(Peak Valley Rate), 피치 그리고 음성의 지속시간과 같은 정보와 주파수 영역의 스펙트럼 동적 변화정보와 같은 음향학적인 특징규칙에 의한 방법 등이 널리 사용되어져 왔다. 간접적인 처리방법은 근본적

으로 음성인식 절차와 동일한 것으로 특징벡터로 표시된 음향패턴을 입력음성과 비교하여 얻은 정보를 이용하여 분할을 행하기 때문에 표준패턴의 작성시 발생하는 문제점 및 학습용 데이터의 양에 의존한다. 반면에 규칙에 의한 방법은 음소의 음향학적인 특징으로 이루어진 분할 파라메터를 설정하고 이들의 임계값에 의해 사전훈련 없이 자동으로 분할할 수 있는데 반하여 임계값의 설정이 정량적이지 않는 단점이 있다.

본 논문은 전통적인 패턴분류 알고리즘인 GMM을 이용하여 프레임 단위 분류를 통한 인식과 음소구간을 결정하는 방법에 관한 연구이다. 2장에서는 GMM에 관하여 간단히 소개하고 3장에서 실험결과 및 고찰, 그리고 4장에서 결론을 맺는다.

II. 본론

2.1 GMM(Gaussian Mixture Model)

통계적인 GMM 모델은 가우시안 확률분포가 가중치를 가지면서 합해진 전체 확률분포를 보인다. 모델의 수가 k 개라고 했을 때 전체 확률분포 $p(x|\lambda_k)$는 다음 식 (1)과 같다.

$$p(x|\lambda_k) = \sum_{m=1}^{M} p_m b_m(x) \tag{1}$$

여기서, λ_k는 모델 k에 대한 GMM 음소모델 파라메터를 기호로 다음 수식 (2)로 표현된다.

$$\lambda_k = \{p_m, \mu_m, \Sigma_m\},\ m = 1,...,M \tag{2}$$

여기서, M은 가우시안 확률분포의 개수, 즉 혼합수의 크기를 의미하고, p_m은 m번째 가우시안 혼합수의 가중치가 된다. 또, μ_m과 Σ_m은 각각 가우시안 평균과 분산이다. 표준편차 행렬이 다이고날(diagonal)이고 각 성분이 $\sigma_1, \sigma_2, ..., \sigma_D$이라고 하면 m번째 가우시안은 다음 식 (3)과 같다.

$$b_m(x) = N(\mu_m, \Sigma_m) \tag{3}$$

$$N(\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)'\Sigma_m(x-\mu_m)}$$

$$= \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2|\Sigma_m|}(x-\mu_m)'(x-\mu_m)}$$

여기서, $|\Sigma_m| = \prod_{i=1}^{D} \sigma_i$이고 D는 벡터의 차원이 된다.

이러한 통계적 모델에 대한 학습은 주어진 학습데이터로부터 GMM 모델의 우도(likelihood)를 최대로 만드는 EM(Expectation and Maximization) 알고리즘을 사용한다. EM 알고리듬은 먼저 통계적 모델의 파라메터를 추정(Expectation)하고 다음에 이를 수정(Maximization)하여 우도가 최대값을 가지도록 학습한 후 최종 파라메터를 구하게 된다.

T개의 학습벡터 X가 존재한다고 할 때 GMM 모델에 대한 우도는 다음과 같다.

$$X = \{\vec{x_1}, \vec{x_2}, ..., \vec{x_T}\}$$
$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x_t}|\lambda) \tag{4}$$

현재의 GMM 모델 파라메터를 λ라 하고, 업데이트 하려는 새로운 모델 파라메터를 λ_n이라고 할 때 EM 알고리즘에 의해 다음 식 (5)의 조건을 만족하도록 λ_n을 구하는 것이다.

$$p(X|\lambda_n) \geq p(X|\lambda) \tag{5}$$

학습 데이터 X의 확률 값은 아래 식 (6)과 같다.

$$p(i|\vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{k=1}^{M} p_k b_k(\vec{x_t})} \tag{6}$$

식 (6)의 결과를 이용, 새로운 모델 파라메터의 값을 다음 수식들을 통해서 구할 수 있다.

$$p_n(i|\vec{x_t}, \lambda) = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda) \tag{7}$$

$$\mu_{n,i} = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)} \tag{8}$$

$$\sigma_{n,i} = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)\vec{x_t}^2}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)} - \mu_{n,i}^2 \tag{9}$$

GMM 모델 $(\lambda_1, \lambda_2, ..., \lambda_S)$들로부터 음성 특징벡터 X가 발생할 확률이 $1 < k < S$에서 가장 높은 모델 λ_p를 찾기 위한 수식은 아래와 같다.

$$P(\lambda_p|X) = argmax P(\lambda_k|X) = argmax \frac{P(X|\lambda)P(\lambda)}{P(X)} \quad (10)$$

2.2 후처리 알고리듬을 통한 분류-인식과 과 음소 분할 알고리즘

GMM을 이용한 프레임별 인식결과를 바탕으로 다음에서 제안하는 후처리 알고리즘을 통하여 프레임 분할-인식과 음소 경계를 결정하게 된다.

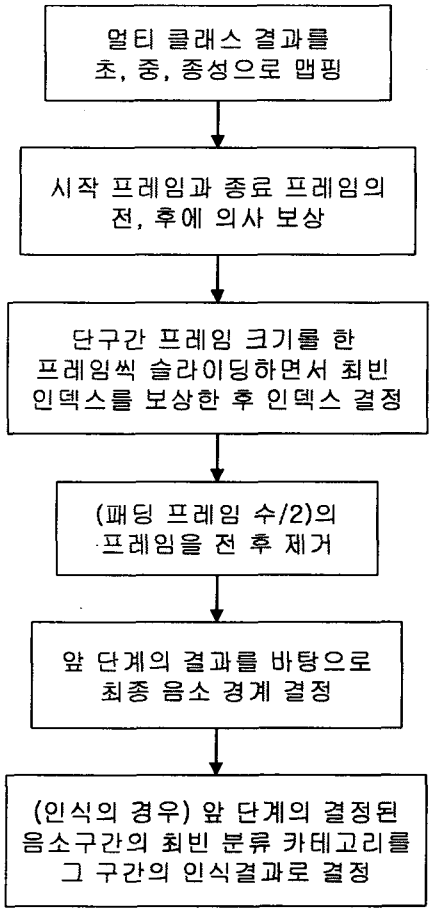

그림 3 후처리 알고리즘 및 음소 단위 인식

III. 실험결과 및 고찰

3.1 실험조건

프레임별 분할-인식과 음소 경계 결정을 위하여 CVC형 108음절 DB를 구성하였다. 본 DB는 우리말 음성의 CVC형 음절로 초성(ㅂ ㄷ ㄱ ㅍ ㅌ ㅋ), 중성(ㅏ ㅓ ㅗ ㅔ ㅜ ㅣ), 종성(ㄴ ㄹ ㅁ)으로 이루어진 108개의 유사 음절로 구성되어져 있으며 5명의 화자가 5회 발성하여 3회분은 학습용으로 나머지 2회분은 평가용으로 이용하였다. 목측에 의하여 초성, 중성, 종성을 분리한 음소 DB를 별도 구성하여 이를 기준으로 하여 인식실험을 행하였다. 음성분석조건은 표1과 같다.

표 1. 음성 데이타의 분석조건

A/D	16kHz, 16bit
Filtering	LPF , 7 kHz
Step Size	60 point
Window Length	256 point
Feature Parameter	MFCC 15th

3.2 실험방법 및 결과

프레임 분류-인식과 음소 분할 모두 인식 방법은 GMM으로 하고 HMM 음소 인식과 그 결과를 비교하였다. 표 2은 목측 분할을 기준으로 GMM을 이용하여 분할된 음소의 프레임 길이와 음소 경계와의 차이에 대한 평균과 표준편차이다.

표 2. 자동분할 성능

Frame	초성	중성	종성
평균	6.33	4.97	4.69
표준편차	6.88	6.29	6.36

초성과 중성의 구분은 상당히 어려운 작업이며 표 2에서 보듯이 분할된 각 음소의 길이에 대한 평균과 표준편차가 중성과 종성보다 초성에서 값이 큰 것을 확인할 수 있다.

표3은 실험결과이다. GMM분류-인식은 GMM 프레임별 분류만을 통하여 후처리를 통하여 분할 경계를 기준으로 인식한 결과이고, GMM-HMM은 분할 경계를 기준으로 HMM으로 인식한 결과이다. HMM 음소인식은 목측에 의하여 분할하여 음소 인식한 결과이다.

표 3. 분류 인식 및 목측 인식률 비교(음소)

단위 %		GMM 분류-인식	HMM 음소인식	GMM-HMM
초성	ㅂ	62.78	86.11	70.00
	ㄷ	65.00	84.44	71.67
	ㄱ	83.33	87.78	81.11
	ㅍ	66.11	81.11	72.78
	ㅌ	73.89	89.44	81.11
	ㅋ	68.33	87.78	80.56
	평균	69.91	86.11	76.20
중성	ㅏ	94.44	96.67	95.56
	ㅖ	98.33	96.11	95.00
	ㅣ	99.44	98.33	100.00
	ㅗ	91.67	93.89	89.44
	ㅜ	92.78	92.78	91.67
	ㅝ	98.89	98.33	95.56
	평균	95.93	96.02	94.54
종성	ㄹ	97.22	99.17	98.61
	ㅁ	87.22	88.33	82.22
	ㄴ	97.50	96.39	94.72
	평균	93.98	94.63	91.85

표 4. 음소인식을 통한 음절인식률

GMM	69.91(%)
HMM	86.11(%)

Ⅳ. 결론 및 향후과제

우리말은 외국어와 달리 초성, 중성, 종성이 합해져서 음절을 이루고 이 음절이 단어와 문장을 이루기 때문에 인식단위, 특히 음소와 같은 최소 단위로의 안정된 분할은 연속음성인식을 위한 주목할만한 연구과제이다. 본 논문에선 GMM을 이용해서 자동 음소분할을 시도하고 이를 통하여 연속음성인식을 이루고자 하는 시도이다. 실험결과 음성의 시변성을 고려한 모델인 HMM에 비하여 다소 떨어지는 인식률을 나타내었지만 HMM 모델의 많은 연산과 메모리 요구사항 그리고 구현상의 효율성을 감안하여 평가할 경우, 비교적 좋은 인식률을 나타내었다. 향후, 최근 많은 연구가 진행되고 있는 SVM을 본 논문의 GMM 분류기를 대체하여 연구할 계획이다.

참고문헌

[1] Joachim M. Buhmann, "Learning and Data Clustering",

[2] Frederick Jelinek, "Statistical Methods for Speech Recognition", The MIT Press, 1999

[3] L. R. Rabiner , R. W. Schafer, "Digital Precossing of Speech Signals", Prentice Hall, 1978

[4] Xuedonga Huang, Alex Acero, Hsiao-Wuen Hon, "Spoken Language Processing/A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001

[5] http://dynamo.ecn.purdue.edu/~bouman/software/cluster/cluster-3.5.1.tar.gz

[6] 김상경, "GMM에 기반한 화자인식에서 모음을 이용한 인증 발성 감축에 관한 연구", 석사학위 논문, 1996

[7] 한학용, 권호민, 이광석, 고시영, 허강인, "유/무성음 척도를 포함한 재구성 특징 파라메터의 음성인식 성능평가", 한국음향학회 추계학술발표대회 제21권 제2호, pp. 27-30

[8] 한학용, 고시영, 허강인, "우리말 연속음성의 음절 분할법", 한국음향학회지 제20권 제3호 pp. 70-75