음소별 GMM을 이용한 화자식별

권 석 봉, 김 회 린
한국정보통신대학원대학교 공학부

Speaker Identification using Phonetic GMM

Sukbong Kwon, Hoi-Rin Kim
School of Engineering, ICU
E-mail: sbkwon@icu.ac.kr, hrkim@icu.ac.kr

Abstract

In this paper, we construct phonetic GMM for text-independent speaker identification system. The basic idea is to combine of the advantages of baseline GMM and HMM. GMM is more proper for text-independent speaker identification system. In text-dependent system, HMM do work better. Phonetic GMM represents more sophistgate text-dependent speaker model based on text-independent speaker model. In speaker identification system, phonetic GMM using HMM-based speaker-independent phoneme recognition results in better performance than baseline GMM. In addition to the method, N-best recognition algorithm used to decrease the computation complexity and to be applicable to new speakers.

I. 머리말

지금까지 화자인식 시스템에서 HMM(Hidden Markov Model)과 GMM(Gaussian Mixture Model)이 효과적인 모델로 많이 사용되고 있다. 문맥 독립형 화자식별 시스템에서는 GMM이 좋은 성능을 보이고 있고[1], 문맥 종속형 화자식별 시스템에서는 HMM이 효과적으로 사용되고 있다[4]. 최근에는 보다 좋은 성능 얻기 위해 GMM과 HMM외에 다른 추가적인 정보를 이용하여 인식률 향상 시키고 있다. 본 논문에서는 문맥 독립형 화자인식 시스템에 GMM의 장점과 HMM의 장점을 결합한 음소별 GMM을 제안한다. 음소별 GMM은 각 화자에 보다 적합하도록 음소별로 세분화 하여 GMM을 만드는 방법으로써 HMM기반의 음소인식 시스템을 활용하여 훈련과 인식을 하는 방법을 제시한다. 음소별 GMM의 사용은 기본적인 GMM을 이용한 것보다 화자간의 확률값에 대한 구분을 더욱 분명히 할 수 있어 화자 인증에도 활용 될 수 있는 가능성을 제시하고 있다. 2장에서는 음소별 GMM에 대한 이론적 배경을 설명하고, 3장에서는 음소별 GMM을 만드는 방법과 화자식별 과정에 대한 설명을 하고 , ETRI의 화자인식용 DB를 통한 훈련 과정 및 실험 과정에 대해 기술한다. 4장에서는 음소별 GMM에 대한 실험방법과 결과에 대해 기술하고 5장에서는 결론과 함께 차후 연구 방향을 제시한다.

II. 음소별 GMM

2.1 HMM 기반 음소 단위 인식 시스템

음소별 GMM을 형성하기 위해서는 훈련에 사용되는 음성파일를 각 음소별로 분할하는 것이 중요한 작업이다. 하지만 여기서 아주 정확한 음소별 음성 인식을 요구하지 않는다. 음소별 GMM의 mixture 수를 증가시킴으로서 음소별 음성 인식 시스템의 낮은 인식률을 극복하여 효과적인 음소별 GMM을 형성할 수 있다. 즉, 음소별 GMM에서는 한 음소에 대한 정보만을 가지고 있지 않고 유사한 음소에 대한 정보를 일부 포함하여 지니고 있다. 이런 의미에서 보면 baseline GMM은 각 화자에 대한 모든 음소에 대한 정보를 가지고 있다고 볼 수 있다. 하지만 음소별 GMM이 효과적인 이유는 HMM 기반 음소단위 인식 시스템의 분할 정보를 이용하기 때문에 음소에 대한 보다 정확한 정보를 음소별 GMM을 통해 이용할 수 있는 장점이 있다. 따라서 HMM 기반 음소단위 음성인식 시스템을 구축할 때 모른 음소에 대

한 정보를 지니고 있는 HMM을 형성하는 것이 중요하다. 그리고 화자식별을 위해 화자 모델을 형성하기 위해서는 음성 구간만을 사용해야 한다. 특히, 화자를 식별하기 위해서 사용되는 특징 벡터는 유성음 구간에서 지배적이다. 따라서 HMM 기반 음성 인식에 앞서 음성 구간 검출기를 통하여 유성음, 무성음, 묵음 구간을 확실히 구분하고 묵음 구간을 제거하여 작업하는 것이 효과적이다.

2.2 Baseline GMM

GMM은 각 화자에 대해서 화자를 특징짓는 모델이 각 화자가 발성한 모든 음성구간에 대한 특징 벡터의 Gaussian mixture 확률 분포를 의미한다.[1] 즉, GMM은 M개의 연속확률밀도 분포의 합으로 이루어지는데 식은 다음과 같다.

$$p(\mathbf{x}\mid\lambda)=\sum_{k=1}^{M}c_k b_k(\mathbf{x}) \tag{1}$$

여기서, $\mathbf{X}$는 D-dimensional 특징 벡터를 나타내고, $b_k(\mathbf{x})$, $k=1,2,\ldots M$ 는 k 번째 연속확률밀도 분포를 나타낸다. c_k, $k=1,2,\ldots M$ 는 k 번째 mixture의 가중치를 나타내며 k 번째 연속확률밀도 함수는 다음과 같다.

$$b_k(\mathbf{x})=\frac{1}{(2\pi)^{D/2}\,|\Sigma_k|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\} \tag{2}$$

여기서 $\boldsymbol{\mu}_k$는 평균 특징 벡터의 평균을 나타내고, Σ_k는 covariance 행렬을 나타낸다. 각 mixture의 가중치는 다음과 같은 성질을 만족한다.

$$\sum_{k=1}^{M}c_k=1 \tag{3}$$

따라서, 화자 s마다 GMM의 확률밀도 함수의 파라미터를 구하면 다음과 같은 모델을 만들 수 있다.

$$\lambda_s=\{c_k^s,\mu_k^s,\Sigma_k^s\}\quad,\quad k=1,2,\ldots,M \tag{4}$$

각 covariance 행렬은 diagonal에만 값을 가지는 경우와 모든 element가 값을 갖는 형태가 있을 수 있다. 본 연구에는 각 mixture마다 covariance를 갖는 nodal covariance 행렬과 diagonal 행렬을 사용한다. 선택의 기준은 화자식별의 실험 결과에 바탕을 두고 있다.

2.3 음소별 GMM

각 화자의 훈련 DB에 대해 HMM 기반 음소별 음성인식 시스템에서의 음소별 분할 정보를 활용하여 음소별 GMM을 형성한다. 즉, 각 화자 모델은 음소별 GMM의 집합으로 표현된다.

$$\lambda_s=(\lambda_{s1},\lambda_{s2},\ldots,\lambda_{sp}) \tag{5}$$

여기서, m1, m2, ..., mn 은 음소 모델의 index를 나타내고, 각 음소 모델은 다음과 같이

$$\lambda_{sp}=(c_k^{sp},\mu_k^{sp},\ldots,\Sigma_k^{sp})\quad k=1,2,\ldots,M \tag{6}$$

GMM으로 형성이 된다. λ_{sp}에 사용되는 mixture의 수는 baseline GMM에 사용되는 것보다 적게 사용된다. 본 연구에서는 HMM에 사용되는 46개의 모델을 사용하였다. 물론 묵음 모델은 제외되었다. 7개의 mixture구조를 가진 음소별 GMM을 사용하면 총 322개의 mixture를 사용하는 효과를 나타낸다. 하지만 훈련과정에서 보면 EM algorithm을 사용하여 훈련을 하게 되는데 baseline GMM에서는

$$c_k=\frac{1}{T}\sum_{t=1}^{T}p(k\mid\bar{x}_t,\lambda) \tag{7}$$

이다. 즉, 각 mixture의 가중치는 훈련 DB의 음소 빈도수에 대한 확률 분포를 나타낸다. 하지만 음소별 GMM에서는 일단 음소별 인식을 HMM에서 정보를 얻기 때문에 음소에 대한 빈도수를 활용하지 않는다. 하지만 음소 단위 인식이 정확한 것이 아니기 때문에 음소별 GMM에도 훈련 DB에 나타나는 음소의 빈도수를 포함하고 있게 된다. 따라서 HMM을 이용한 음소단위의 음성인식 시스템의 인식 성능이 음소별 GMM의 성능에 많은 영향을 준다.

2.4 그룹화된 음소별 GMM

음소별 GMM을 형성할 때 각 음소별로 만들 특징벡터가 부족하거나, 훈련 DB에 특정 음소가 없지만 인식과정에서 훈련되지 않은 음소 모델이 나타날 수 있기 때문에 음소별 GMM을 그룹화가 필요가 있다. 실제 화자식별에서 유성음 구간에서의 인식이 지배적이고 훈련 DB에서도 모음은 항상 존재하기 때문에 모음외의 다른 음소에 대해서 그룹화를 해야 한다. 그룹화 하는 방법은 HTK의 방식을 사용한다. 그룹화 하는 만큼 mixture의 수를 증가시켜 주어야 한다. 만약 충분한 길이의 음성 데이터가 인식과정에 사용될 때에는 무성음에 대한 GMM에는 인식과정에서 무시해도 인식률에 큰 영향을 주지 않는다.

Ⅲ. 훈련 및 화자식별 시스템 구현

3.1 Database

화자식별 시스템에는 ETRI에서 비영리용으로 수집된

화자 인식용 DB 중 본 연구에서는 중가용 마이크용 DB를 사용하였다. HMM 기반 음소단위 음성인식 시스템에는 모든 음소에 대한 HMM을 만들 수 있는 PBW452를 사용한다. 중가마이크 화자인식용 음성 DB는 PC 환경에서의 음성 정보 처리 시스템의 개발을 위해 사무실 환경에서 중가의 마이크를 사용하여 50명(20명-주차, 20명-월차, 10명-3개월차)의 화자가 발성한 2연 숫자, 4연 숫자, 문장으로 구성된 DB이다. 문장 음성의 발성목록은 개인정보와 관련된 10개의 질문과 3어절 이내로 구성된 단문 10개로 구성되어 있다. 한 화자당 동일한 목록을 5회 발성하고, 주차/월차/3개월차로 구분하여 4회 반복하였다. 두 음성 DB는 16kHz/16bit, linear PCM으로 저장되어 있다.

3.2 훈련

먼저 각 DB 파일에 대해 음소단위의 분할 정보를 얻기 위해 PBW452 DB를 이용하여 12차 MFCC의 특징 벡터에 derivative와 acceleration를 추가하여 총 36차의 특징 벡터를 사용하여 HMM을 만들고, 화자 인식용 DB를 음성 구간 검출기를 사용하여 음성 구간에서의 특징 벡터를 추출하여 앞에서 만든 HMM을 이용하여 음소별 분할 정보를 얻는다. 이렇게 형성된 분할 정보를 이용하여 각 화자에 대한 음소별 GMM을 만든다. 그림 1은 음소별 GMM을 만드는 전체 흐름도를 보여 주고 있다.

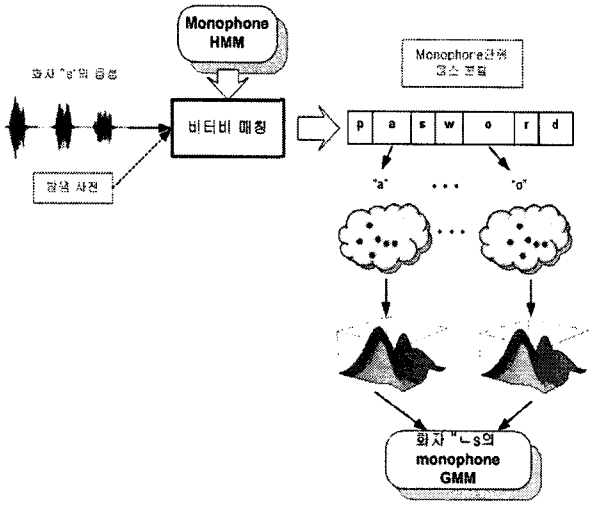

그림 1. 각 화자에 대한 음소별 GMM 훈련과정

3.3 화자식별 시스템

화자식별 시스템의 인식 과정은 음소단위 인식 시스템에 쓰이는 HMM, N-best 화자를 얻기 위한 baseline GMM과 N-best 화자로부터 최종 인식 화자를 선택하기 위한 음소별 GMM을 사용한다. 그림 2는 본 논문에서 구현된 화자식별 시스템의 인식 과정을 나타낸다. Baseline GMM의 mixture수를 80이상으로 GMM을 쓰게 되면 화자의 수가 증가하였을때 각 화자별 확률값을 구하는데 많은 시간이 소요된다. Baseline GMM를 사용한 화자식별에서 5-best 화자인식을 사용하면 거의 100%에 가까운 인식률을 보이기 때문에 음소별 GMM의 사용에 있어 5명의 화자에 대한 인식만을 하면 좋은 결과를 얻을 수 있다.[3]

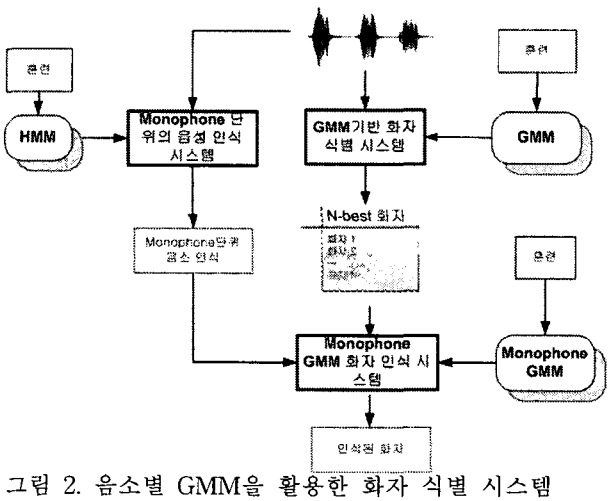

그림 2. 음소별 GMM을 활용한 화자 식별 시스템

Ⅳ. 실험 및 결과

중가마이크 화자인식용 음성 DB는 총 80,000개의 파일로 이루어져 있고, 이 중 64,000을 훈련용으로 사용하고 16,000개를 테스트용으로 사용하였다. 훈련에 사용된 DB에서 음성 구간의 길이는 화자당 대략 30분 정도이고 테스트에서는 0.5 ~ 2초 이내의 음성에 대한 인식 결과를 얻었다. 표 1은 baseline GMM에서의 mixture 수에 따른 화자식별 인식결과를 나타낸 것으로써 그림 3에서 도식화하였다. 그림 3을 보면 mixture의 수가 40 이상이 되면 인식률의 증가가 더디게 상승한다. Baseline GMM을 사용하면 mixture 수가 40이고 5-best에서의 인식률이 99.90%로서 음소별 GMM을 재검증에 사용하기에 충분한 인식률을 얻을 수 있다. 표 2은 baseline GMM을 인식을 통해 얻은 5명의 후보화자에 대해서 음소별 GMM을 사용하여 재인식한 결과를 나타낸 것으로서 각 음소에 대해 mixture수 가 3, 7, 14에 대한 인식결과를 나타낸다. 현재 모든 GMM에 대해 36차의 특징 벡터를 쓰고 있는데 MFCC와 derivative만을 이용한 24차의 특징 벡터와의 인식률의 차이는 거의 없다. 하지만 HMM 기반 음소단위 인식 시스템에서의 높은 인식률을 얻기 위해 36차를 사용했기 때문에 GMM에서도 36차를 사용하였다.

표 1. Baseline GMM의 인식률(%)

Mixture 수	1-best 인식률	5-best 인식률
3	44.92	76.48
5	65.66	89.72
7	80.14	96.66
14	87.39	98.17
28	92.96	99.17
40	95.49	99.90
80	97.41	99.95

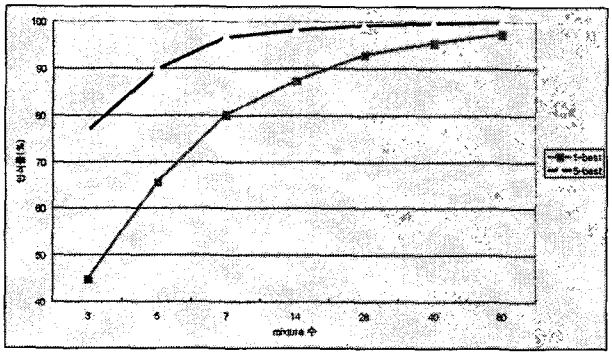

그림 3. Baseline GMM의 mixture 수 변화에 따른 성능 비교

표 2. 음소별 GMM의 인식률(%)

mixture 수	인식률
3	95.25
7	99.00
14	99.46

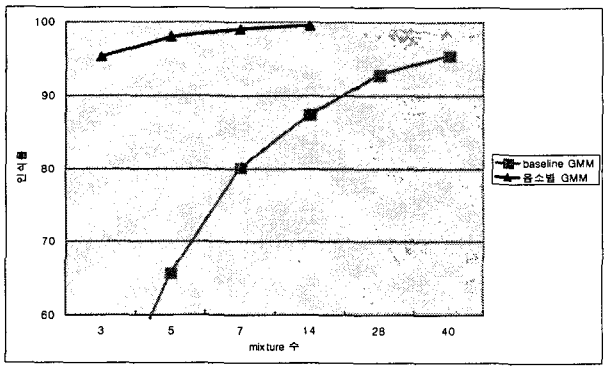

그림 4. 음소별 GMM의 mixture 수 변화에 따른 성능 비교

인식에 사용된 입력 음성의 길이에 따라 인식률의 차이를 많이 보이고 있는데 0.5초 가량의 입력 음성에 대한 에러률이 1초 가량의 입력 음성에 대한 에러률보다 4배 가량의 높은 에러률을 보이고 있다. 따라서 실제 화자식별 시스템에서 인식을 위한 입력 음성의 길이는 최소 1초 이상의 되도록 하는 것이 효과적이다.

V. 맺음말

본 논문에서는 GMM과 HMM의 장점을 결합한 문맥독립형 화자인식 시스템을 제안하였다. 즉, baseline GMM처럼 화자의 전체적인 모델을 사용하여 대략적인 화자 인식의 정보를 얻고, HMM 기반의 음소단위 음성인식 시스템을 이용하여 얻은 음소단위의 분할 정보와 음소단위의 분할 정보를 이용하여 훈련된 음소별 GMM을 사용함으로써 화자에 대한 모델을 좀 더 정밀하게 구현할 수 있었으며, baseline GMM의 mixture의 수를 무작정 늘이는 것 보다 인식 속도나 인식률 면에 효과적인 결과를 내고 있다. 본 논문에서 사용된 훈련용 DB의 양이 화자당 30분 정도로 많은 편인데, 앞으로 훈련에 사용되는 DB의 량을 줄이면서도 효율적인 음소별 GMM 모델을 구현할 수 있는 방안에 대해 좀 더 연구를 할 것이며, 실제 잡음 환경에서의 효과적인 모델의 형성 방법에 대한 연구를 추가할 것이다.

감사의 글

본 연구는 한국정보통신대학교 디지털미디어연구소의 정보통신연구개발사업의 연구비 지원에 의하여 수행되었음.

참고문헌

[1] D.A. Reynolds and R.C. Rose, "Roubust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Processing, pp. 72-83, Jan. 1995.

[2] C. Taji, P. Dumouchel and Y. Fang, "N-best GMM's for Speaker Identification," Proceedings of Eurospeech vol. 5, pp. 2295-2298, 1997.

[3] Mary A. Kohler, "Phonetic speaker recognition," Signals, Systems and Computers, 1557-1561 vol 2, 2001.

[4] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communications pp. 91-108, 1995.