한국음향학회지 제22권 제8호 pp. 711~716 (2003)

화자인식에서 연속밀도 은닉마코프모델의 혼합밀도 결정방법

Gaussian Density Selection Method of CDHMM in Speaker Recognition

서 창 우*, 이 주 헌**, 임 재 열***, 이 기 용****
(ChangWoo Seo*, Joohun Lee**, JaeYeol Rheem***, KiYong Lee****)

*(주)인스모바일 기술연구소, **동아방송대학 인터넷방송과,
***한국기술교육대학교 정보기술공학부, ****숭실대학교 정보통신전자공학부
(접수일자: 2003년 7월 30일; 수정일자: 2003년 11월 6일; 채택일자: 2003년 11월 14일)

본 논문은 연속밀도 은닉마코프모델에서 각 상태별 혼합성분 개수를 결정하는 방법을 제안한다. 지금까지의 대부분의 연구가 연속밀도 은닉마코프모델에서 화자의 스펙트럼 특성에 상관없이 각 상태별 동일한 혼합성분 개수를 적용하였다. 이런 접근방법은 많은 계산량을 요구할 뿐만 아니라, 각 상태의 특성을 무시하고 있기 때문에 각 상태별 음성신호의 정확한 모델링을 할 수 없다. 따라서 본 논문에서 제안한 연속밀도 은닉마코프모델의 파라미터 추정은 각 상태별 혼합성분에 대한 발생 확률값에 따라서 결정하였다. 또한 혼합성분의 개수를 줄이는 과정에서 신호의 상관성을 줄이고 시스템의 전체적인 안정성을 얻기 위해서 주성분 분석을 이용하였다. 제안한 방법은 기존의 은닉마코프모델에 비해서 평균 10% 작은 혼합성분 개수를 이용했을 때를 기준으로 실험하였다. 실험결과에서 혼합성분 결정만을 적용했을 때 거의 비슷한 성능을 얻을 수 있었다. 그리고 주성분 분석을 이용했을 때, 특징벡터가 16차일 때 평균 0.35%의 성능감소가 일어났지만, 25차에서는 평균 0.65%의 성능개선을 얻을 수 있었다.
핵심용어: 화자인식, 연속밀도 은닉마코프모델, 혼합밀도, 주성분 분석
투고분야: 음성처리 분야 (2.5)

This paper proposes the method to select the number of optimal mixtures in each state in Continuous Density HMM (Hidden Markov Models). Previously, researchers used the same number of mixture components in each state of HMM regardless spectral characteristic of speaker. To model each speaker as accurately as possible, we propose to use a different number of mixture components for each state. Selection of mixture components considered the probability value of mixture by each state that affects much parameter estimation of continuous density HMM. Also, we use PCA (principal component analysis) to reduce the correlation and obtain the system' stability when it is reduced the number of mixture components. We experiment it when the proposed method used average 10% small mixture components than the conventional HMM. When experiment result is only applied selection of mixture components, the proposed method could get the similar performance. When we used principal component analysis, the feature vector of the 16 order could get the performance decrease of average 0.35% and the 25 order performance improvement of average 0.65%.
Keywords: *Speaker recognition, CDHMM, Gaussian density, Principal component analysis*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

HMM (Hidden Markov Models)은 음성인식 (speech recognition)뿐만 아니라 화자인식 (speaker recognition)에서도 성공적으로 사용되고 있다. 상태전이와 통계적 모델링을 사용하여 음성의 연속성과 정지 스펙트럼 특성을 모델링하는 HMM은 동적인 음성패턴을 다루기에 아주 유익한 접근방법이다[1,2]. 음성신호에서 각 화자의 스펙트럼 특성 변화를 모델링하는 것은 화자인식을 위한 HMM에서 중요한 문제이다. 각 화자별 스펙트럼 변화를 증가시키는 많은 요소들이 존재한다. 특히 발성 방법,

책임저자: 서창우 (cwseo@insmobile.com)
463-840 경기도 성남시 분당구 야탑동 537-2 덕인빌딩 304호
(주)인스모바일 기술연구소
(전화: 031-703-7301; 팩스: 031-703-7302)

액센트, 감정, 그리고 배경 잡음 등이 많은 영향을 미친다. 이러한 변화를 모델링하기 위한 통계적인 접근방법이 연구되었다[3,4].

지금까지의 많은 연구가 HMM의 학습과정에서 파라미터를 추정할 때 동일한 혼합성분 (mixture) 개수를 사용하였다. 이런 기법은 많은 계산량을 요구할 뿐만 아니라 각 상태의 특성을 무시하고 있기 때문에 각 화자별 음성신호의 정확한 모델링을 할 수 없다. 이러한 문제점을 해결하기 위해서 HMM 선택문제 (selection problem) 혹은 HMM 위상 최적화 (topology optimization)과 같은 연구가 진행되었다[5,6].

본 논문에서는 계산량을 줄일 뿐만 아니라 화자의 스펙트럼 특성을 정확히 모델링하기 위해서 각 상태별 혼합성분의 기대치를 이용해서 혼합성분 개수를 결정하는 방법을 제안한다. 그리고 각 상태별 혼합성분 성분을 줄이는 과정에서 특징벡터의 상관성으로 인한 시스템 안정화 문제가 발생할 수 있다. 따라서 이러한 상관성은 주성분 분석 (principal component analysis)을 이용함으로써 줄일 수 있었다. 주성분 분석은 신호의 상관성을 제거시킬 수 있을 뿐만 아니라 차원을 감소시킬 수 있다[7,8].

논문은 다음과 같이 구성되었다. II장은 논문에서 제안한 연속밀도 HMM에서 각 상태별 혼합성분 개수를 결정하는 방법을 설명하였다. 그리고 실험결과는 III장에서 설명되며, IV장에는 결론을 서술하였다.

II. 연속밀도 HMM의 혼합밀도 결정

HMM은 음성인식뿐만 아니라 화자인식에서도 성공적으로 사용되고 있다. 상태전이와 통계적 모델링을 사용하여 음성의 연속성과 정지 스펙트럼 특성을 모델링하는 HMM은 동적인 음성패턴을 다루기에 아주 유익한 접근방법이다[1,2].

지금까지의 많은 연구가 HMM의 학습과정에서 파라미터를 추정할 때 각 화자의 스펙트럼 특성에 상관없이 동일한 혼합성분 개수를 사용하였다. 그러나 이런 기법은 음성신호의 다양성과 각 상태의 특성을 무시하고 있기 때문에 화자별 음성신호의 정확한 모델링을 할 수 없다. 최근, 이러한 문제점을 해결하기 위해서 HMM 선택문제 혹은 HMM 위상 최적화와 같은 연구가 진행되고 있다.

본 논문에서는 화자의 스펙트럼 특성을 정확히 모델링하기 위해서 연속밀도 HMM의 파라미터에서 각 상태별 혼합성분의 발생 확률을 이용해서 혼합성분 개수를 결정하는 방법을 제안한다. Baum Welch 알고리즘에 의한 HMM의 파라미터, 즉 가중치 c_{jm}, 평균벡터 μ_{jm} 그리고 공분산 행렬 Σ_{jm}은 기대치 $\gamma_t(j,m)$로 구성된다. 정규화를 적용한 기대치는 순향-역향 (forward-backward) 알고리즘을 이용하여 다음과 같이 구할 수 있다.

$$\gamma_t(j,m) = P(s_t = j, m_t = m \mid Y, \lambda) = \frac{P(Y, s_t = j, m_t = m \mid \lambda)}{P(Y \mid \lambda)}$$

$$= \frac{\sum_i \alpha_{t-1}(i) a_{ij} c_{jm} b_{jm}(y_t) \beta_t(j)}{\sum_i \alpha_t(i) \beta_t(i)}, \quad for\ 1 < t \le T \tag{1}$$

여기서 $\gamma_t(j,m)$는 관측 열 Y와 HMM 모델 λ가 주어질 때, 시간 t에서 상태 j와 혼합성분 m이 일어날 기대치이다. 따라서 논문에서 고려된 방법은 식 (1)의 각 상태별 혼합성분이 발생할 확률값을 이용하였다.

각 상태별 혼합성분의 개수를 결정하기 위한 방법은 그림 1과 같이 진행되었다.

과정 1. 초기화: 각 상태별 혼합성분 성분의 개수는 8개로 시작한다.

과정 2. 학습과정 반복: HMM의 학습과정이 반복된다.

과정 3. 혼합성분 개수를 결정: 각 상태별 혼합성분 개수는 혼합성분의 파라미터를 결정하는데 가장 큰 영향을 미치는 식 (1)의 $\gamma_t(j,m)$의 값을 시간에 대한 합의 값으로 결정을 하였다. 시간 t에 대한 합 $\gamma(j,m) = \sum_{t=1}^{T} \gamma_t(j,m)$은 상태 j 와 혼합성분 m에 대한 전체 기대치이다. 기대치 $\gamma_t(j,m)$를 각 상태별 전체 혼합성분의 값으로 정규화했을 때, 임계값 δ_T 보다 작은 경우가 발생하면 가장 작은 값으로 나타난 혼합성분을 줄인다. 만약 모든 혼합성분의 발생 확률이 임계값보다 크면 혼합성분을 줄이지 않고 반복과정이 계속된다. 각 반복과정에서 각 상태별 혼합성분의 최소 할당 개수는 1개로 결정하였다.

과정 4. 반복과정의 종료: 반복과정의 종료는 새로운 추정값을 이전의 값으로 차감 $p(Y \mid \bar{\lambda}) - p(Y \mid \lambda) < \delta$ 했을 때, 기준값 δ 보다 작고 그리고 전체 상태별 혼합성분의 개수가 변함없이 5회 이상 반복과정을 수행할 때 종료한다. 만약 과정 4의 조건이 만족되지 않으면, 과정 2가 계속된다.

과정 1에서 혼합성분의 시작 개수를 8개로 한 것은 음성인식과 달리 화자인식에서는 화자의 음성만을 사용하

기 때문에 사용자의 편리성과 일반성을 고려했을 때, 충분한 학습 데이터를 얻는 것은 사실상 불가능하기 때문에 성분을 8개로 시작하였다. 과정 2의 반복이 시작되었을 때, 처음 반복의 5회까지는 알고리즘 수렴을 고려해서 혼합성분의 개수 변화는 고려하지 않았다. 5회 이상 반복학습을 수행한 후 혼합성분 수를 결정하기 위해서 과정 3이 진행된다. 과정 3에서는 HMM의 파라미터에 가장 많은 영향을 미치는 기대치 $\gamma_t(j,m)$을 시간에 대한 합으로 결정을 하였다. 식 (1)을 시간 t에 대한 합으로 나타내면 상태 j와 혼합성분 m에 대한 전체 기대치를 구할 수 있다.

$$\gamma(j,m)=\sum_{t=1}^{T}\gamma_t(j,m) \tag{2}$$

위에서 구해진 각 상태별 혼합성분의 기대치는 각 상태별 전체 혼합성분의 값으로 정규화를 하였다. 이렇게 구해진 값이 임계값 δ_T 보다 작으면, 가장 작은 값에 해당하는 혼합성분을 줄인다.

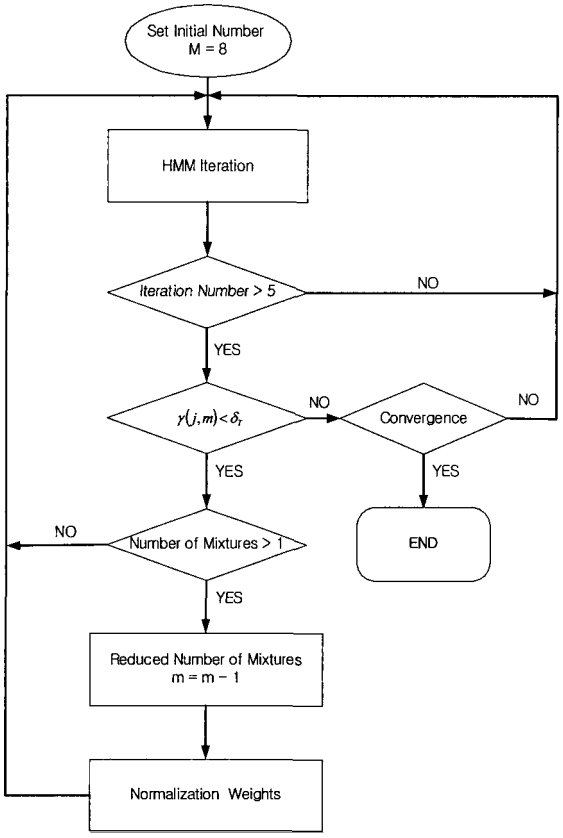


그림 1. 각 상태별 혼합성분을 결정하기 위한 순서도
Fig. 1. Flowchart for decision of mixture components for each state.

$$if \quad \gamma(j,m)<\delta_T, \quad m=m-1 \tag{3}$$

이런 반복과정에서 고려해야 하는 것은 기대치 $\gamma(j,m)$가 임계값보다 작은 경우가 여러 상태에서 동시에 발생할 수 있다. 그런 경우에는 전체 상태에서 가장 작은 기대치를 갖는 혼합성분만을 줄인다. 왜냐하면, 처음 충분한 수렴이 이루어지지 않은 상태에서는 임계값보다 작은 값이 동시에 여러 개 발생할 수 있기 때문이다. 따라서 임계값보다 작은 혼합성분이 동시에 여러 개 발생했을 때, 많은 혼합성분 수를 줄이는 것은 시스템의 수렴에 많은 영향을 미칠 수 있다.

만약 이런 반복과정에서 모든 혼합성분의 발생 확률이 임계값보다 크면 반복과정은 혼합성분의 개수에 별다른 영향없이 계속 반복된다. 각 반복과정에서 각 상태별 혼합성분의 수가 1개인 경우 기대값의 크기에 상관없이 해당 혼합성분 수는 줄이지 않는다. 각 상태별 최소 1개의 혼합성분 수를 할당해야만 한다. 반복과정의 종료는 새로운 추정값을 이전의 값으로 차감 $p(Y|\bar{\lambda})-p(Y|\lambda)<\delta$ 했을 때, 기준값 δ보다 작고 그리고 전체 상태별 혼합성분 수가 변함없이 5회 이상 반복과정을 수행했을 때 종료한다. 만약 과정 4의 조건이 만족되지 않으면, 과정 2로부터 계속된다.

위와 같은 과정으로 혼합성분 개수가 변하면, 이전의 반복으로부터 얻어진 파라미터의 값을 변화시켜야 한다. 특히 가중치 c_{jm}는 다음 조건을 만족시켜야 한다.

$$\sum_{m=1}^{\tilde{M}}c_{jm}=1 \tag{4}$$

여기서 $\tilde{M}$는 줄어든 혼합성분의 개수다. 먼저 가중치 c_{jm}를 계산하기 전에 식 (1)의 $\gamma_t(j,m)$의 확률값을 줄어든 혼합성분 수에 맞게 계산하고 그리고 그 값을 토대로 가중치 c_{jm}, 평균벡터 μ_{jm}, 그리고 공분산 행렬 Σ_{jm}을 계산해야 한다.

III. 실험결과

실험을 위해서 사용된 데이터는 대학원 실험실 환경에서 수집하였으며, 한국어 문장 종속 연속음 화자인식을 위해서 "열려라 참깨"를 사용하였다. 수집된 데이터는 1주 간격의 시간차를 가지고 있으며 3주에 걸쳐서 수집하였다. 첫째 주 5회 발성에서 수집한 데이터는 학습을 위해

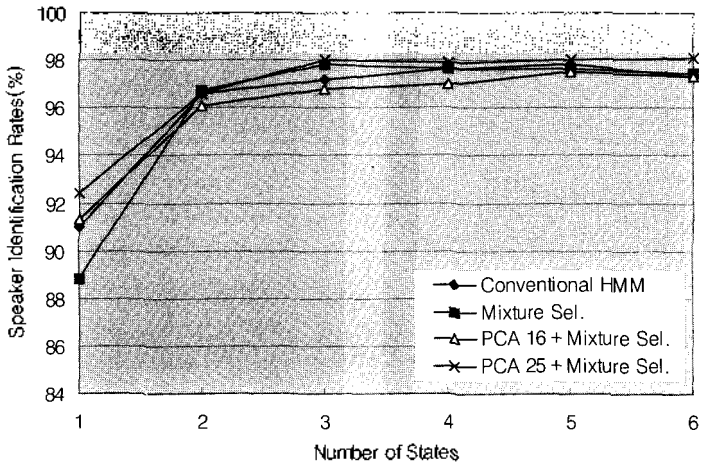


그림 2. 일반적인 HMM과 δ_T = 0.12에서의 제안한 방법의 성능 비교
Fig. 2. Performance comparison between the conventional HMM and proposed method for δ_T = 0.12.

서 사용하였으며, 나머지 2주간 수집한 데이터는 테스트에 사용하였다. 따라서 개인별 전체 발성된 데이터 수는 15개이고, 수집된 데이터의 화자 인원수는 200명으로 남/여 각각 100명이고 샘플링 주파수는 16 kHz이고 분해능은 16 bit이다.

먼저 실험에서 사용된 음성의 프레임 길이는 256 샘플을 한 프레임으로 간주하고 그리고 프레임간 50% 중첩을 적용했다. 특징벡터는 MFCC를 사용했으며, 이때 FFT는 256 point 그리고 필터뱅크는 24개를 사용하였다. 전체 특징벡터는 MFCC를 통해서 얻어진 12차 켑스트럼 계수, 델타 에너지 그리고 12차 델타 켑스트럼의 25차원으로 구성되었다. 또한 MFCC를 통해서 얻어진 켑스트럼은 채널의 영향을 없애기 위해서 CMS (cepstral mean subtraction)와 고차 성분을 강조하기 위해서 켑스트럼 가중치 (cepstral weighting)를 적용하였다. 실험에서 연속밀도 HMM의 학습 파라미터를 추정하기 위해서 사용한 방법은 Baum Welch 재추정 알고리즘을 사용하였다. 실험을 통한 제안한 방법의 우수성을 확인하기 위해서 화자식별에서 비교하였으며, 일반적인 방법에서의 혼합성분 개수는 학습 데이터와 성능차를 고려하여 5개를 사용하였다. 그리고 주성분 분석의 차수는 정보율이 97%에 가장 가까운 16차를 기준으로 실험하였다.

그림 2에서 일반적인 HMM은 각 상태별 혼합성분 5개를 사용하였고 제안한 혼합성분 결정방법은 주성분 분석을 적용한 경우와 그렇지 않은 경우를 나타낸 것이다. 제안한 방법에서 사용된 식 (3)의 임계값 δ_T는 0.12이다. δ_T 는 각 상태별 혼합성분의 확률값이 0.12보다 작은 경우에 혼합성분 성분을 줄이기 위해서 사용한 임계값이며, 또한 이것은 전체 확률값의 12%를 의미한다. 즉 기대치 값이 12% 이하인 경우에는 그 해당 혼합성분을 줄이는 방법을 사용하였다. 표 1은 제안한 방법에서 각 상태별 평균 혼합성분 성분의 개수를 나타낸 것이다.

그림과 표를 통해서 알 수 있듯이, 제안한 혼합성분 결정은 일반적인 HMM 방법보다 혼합성분 수를 평균 10% 작게 사용했지만 단지 0.2%의 성능감쇠가 있었다. 또한 제안한 방법에서 주성분 분석을 사용했을 때, 특징벡터 차수가 16차와 25차일 때는 혼합성분의 개수가 약 10%이상 줄었음에도 불구하고 단지 0.25%의 성능감쇠와 0.65%의 성능개선을 얻을 수 있었다.

표 2는 제안한 방법에서 각 상태별 사용된 평균 혼합성분 수이다. 표를 통해서 알 수 있듯이 각 상태별 선택된 혼합성분의 평균 개수는 각 상태별 동일한 개수의 혼합성분 성분이 사용되지 않고 있다는 것을 알 수 있다. 표에서 나타난 결과를 보면 각 상태에서 사용된 혼합성분 개수의

표 1. 제안한 방법에서 평균 혼합성분 성분의 개수 비교
Table 1. Comparison of the number of average mixture components between the proposed methods.

States	Number of average mixtures		
	Mixture sel.	PCA 16 + mixture sel.	PCA 25 + mixture sel.
1	4.5	4.49	4.5
2	4.53	4.51	4.48
3	4.43	4.47	4.38
4	4.46	4.44	4.42
5	4.47	4.45	4.38
6	4.45	4.48	4.44
Average	4.45	4.47	4.43

표 2. 제안된 방법에서의 각 상태별 평균 혼합성분 개수
Table 2. Average number of mixture components by each state in the proposed methods.

State		Number of average mixtures by each state					
		1	2	3	4	5	6
1	a	3.895					
	b	3.79					
	c	3.76					
2	a	4.01	3.68				
	b	3.93	3.705				
	c	3.945	3.6				
3	a	3.83	3.805	3.57			
	b	3.79	3.83	3.625			
	c	3.725	3.86	3.675			
4	a	3.76	3.92	3.82	3.37		
	b	3.81	3.875	3.735	3.52		
	c	3.59	3.865	3.905	3.23		
5	a	3.595	4.015	3.93	3.76	3.2	
	b	3.665	3.755	3.825	3.735	3.575	
	c	3.375	3.91	3.935	3.925	3.91	
6	a	3.54	3.85	3.97	3.98	3.775	3.115
	b	3.58	3.78	3.815	3.845	3.77	3.62
	c	3.49	3.795	3.96	3.98	3.795	2.975

a : Mixture sel., b : PCA 16 + mixture sel., c : PCA 25 + mixture sel.

차이는 상태 4일 때, 최대 20% 정도의 차이를 보였다. 또한 개인별 상태에서 나타난 평균 혼합성분의 개수 차이는 상태 수 5일 때, 최대 4.015개에서 최소 3.2개로 각 상태별 최대 1개 정도의 혼합성분의 사용개수 차이가 나타났다. 화자의 혼합성분의 개수 차이가 이렇게 두드러지게 나타난 것은 각 상태별 스펙트럼 분포를 모델링하는데 있어서 스펙트럼 분포가 일정하지가 않다는 것을 나타내고 있다.

실험을 통해서 알 수 있듯이 가장 좋은 결과는 주성분 분석 25차에서 제안한 방법을 사용했을 때, 기존의 방법보다 평균적으로 0.65%이상 성능개선이 일어났음을 알 수 있다. 또한 주성분 분석을 이용했을 때 특징벡터의 차수가 36% 줄었지만 성능감소는 단지 0.35%만 일어났다.

지금까지의 실험결과를 토대로 각 상태별 혼합성분 개수 결정은 기존의 동일한 혼합성분 개수를 할당하는 것보다 화자의 스펙트럼 특성을 모델링하는데 효과적인 방법임을 확인할 수 있다. 이것은 화자의 스펙트럼 분포를 모델링하는데 있어서 각 상태별 분포의 모델링이 일정하지가 않다는 것을 나타내고 있다.

IV. 결 론

본 논문에서 제안한 방법은 연속밀도 HMM에서 각 상태별 혼합성분 개수를 결정하는 방법을 제안한다. 지금까지의 대부분의 연구가 HMM에서 각 화자의 스펙트럼 특성에 상관없이 각 상태별 동일한 혼합성분 개수를 적용하였다. 이런 접근 방법은 많은 계산량을 요구할 뿐만 아니라, 개인별 특성을 무시하고 있기 때문에 각 상태별 음성신호의 정확한 모델링을 할 수 없다. 연속밀도 HMM에서 각 화자의 스펙트럼 특성을 효과적으로 모델링하기 위해서 각 상태별 혼합성분 개수 결정방법을 제안한다.

논문에서 제안한 방법은 연속밀도 HMM의 파라미터를 추정할 때, 각 상태별 혼합성분의 확률값을 적용해서 혼합성분 개수를 결정하였다. 또한 혼합성분의 개수를 줄이는 과정에서 신호의 상관성을 줄이고 시스템의 전체적인 안정성을 얻기 위해서 주성분 분석을 이용하였다. 제안한 방법은 기존의 HMM 방법과 비교해서 평균 10% 작은 혼합성분 개수를 이용했을 때를 기준으로 실험하였지만, 혼합성분 결정만을 적용했을 때 거의 비슷한 성능 결과를 얻을 수 있었다. 그리고 주성분 분석을 이용했을 때, 특징벡터가 16차일 때 평균 0.35%의 성능감소가 일어났지만,

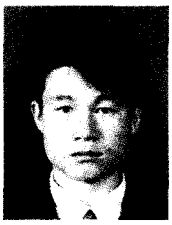
25차에서는 평균 0.65%의 성능개선을 얻을 수 있었다.

참 고 문 헌

1. A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Sub-word talker verification using hidden Markov models," *IEEE ICASSP*, 269-272, 1990.
2. S. Furui, "An overview of speaker recognition technology," *ESCA workshop on Automatic Speaker Recognition Identification Verification*, 1-9, 1994.
3. K. F. Lee, *Automatic Speech Recognition: the development of the SPHINX system*, Kluwer Academic, 1989.
4. X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University, 1990.
5. R. C . Vasko, A. El-Jarudi, J. R. Boston, "An algorithm to determine hidden Markov model topology," *ICASS 96*, 6, 3577-3580, May, 1996.
6. A. Biem, H. Jin-Young, and J. Subrahmonia, "A bayesian model selection criterion for HMM topology optimization," *ICASSP 02*, 1, 989-992, May 2002.
7. Y. Ariki, S. Tagashira, and M. Nishijima, "Speaker recognition and speaker normalization by projection to speaker subspace," *ICASSP 96*, Atlanta, USA, 319-322, 1996.
8. L. Liu and J. He, "On the use of orthogonal GMM in speaker recognition," *Pro. ICASSP*, 845-849, 1999.

저자 약력

● 서 창 우 (ChangWoo Seo)



1996년 2월: 창원대학교 전자공학과 (공학사)
1998년 2월: 창원대학교 전기전자제어공학부 (석사)
2003년 2월: 숭실대학교 정보통신전자공학부 (박사)
2000년 3월~2003년 5월: ㈜웹프로텍 음성개발팀 팀장
2003년 5월~현재: ㈜인스모바일 기술연구소 S/W 개발팀 선임연구원
※ 주관심분야: 음성신호처리, 멀티미디어, 모바일

● 이 주 헌 (Joohun Lee)

한국음향학회지 제21권 제1호 참조
현재: 동아방송대학 인터넷방송과 부교수

● 임 재 열 (JaeYeol Rheem)

1986년 2월: 서울대학교 전자공학과 (공학사)
1988년 2월: 서울대학교 전자공학과 (공학석사)
1995년 2월: 서울대학교 전자공학과 (공학박사)
1995년 9월~현재: 한국기술교육대학교 정보기술공학부 부교수
※ 주관심분야: DSP, 음성신호처리, 생체인식, 잡음처리

● 이 기 용 (KiYong Lee)

한국음향학회지 제21권 제1호 참조
현재: 숭실대학교 정보통신전자공학부 부교수