한국음향학회지 제22권 제7호 pp. 528~538 (2003)

음성 문자 공용인식기를 위한 SSMS 기반 가변 파라미터 모델

A Variable Parameter Model based on SSMS for an On-line Speech and Character Combined Recognition System

석 수 영*, 정 호 열*, 정 현 열*
(Soo-Young Suk*, Ho-Youl Jung*, Hyun-Yeol Chung*)

*영남대학교 정보통신공학과

(접수일자: 2003년 6월 9일; 수정일자: 2003년 8월 7일; 채택일자: 2003년 9월 16일)

음성 문자 공용 인식 시스템은 PDA (Personal Digital Assistants)와 같은 휴대용 모빌 환경에서 음성인식과 문자인식을 적용하기에 적합하도록 개발되었다. 공용 인식 시스템은 특징 파라미터 추출에 있어서는 음성과 문자부분이 독립적으로 수행되나, 인식 과정은 단일 엔진으로 수행된다. CHMM (Continuous Hidden Markov Model)을 이용하는 인식엔진은 고정 파라미터 모델 구조 대신에 동일한 인식률을 유지하면서 모델의 파라미터의 수를 효과적으로 줄일 수 있는 가변 파라미터 모델 구조를 사용하는 것이 유리하다. 본 논문에서는 문맥 독립 가변 파라미터 모델을 생성하기 위해 SSMS (Successive State and Mixture Splitting) 방법을 제안한다. SSMS 알고리즘은 시간 방향 분할과 혼합수 방향 분할을 통해 적절한 상태수와 각 상태당 적절한 혼합수를 가지는 모델을 생성한다. 음성 인식 실험 결과 동일한 인식성능을 나타내는 경우 SSMS 기반 가변 파라미터 모델이 고정 파라미터 모델에 비해 GOPDD (Gaussian Output Probability Density Distribution)의 수가 40% 감소함을 확인할 수 있었다.

핵심용어: 음성인식, 문자인식, SCCRS, SSMS, 가변 파라미터
투고분야: 음성처리 분야 (2.5)

A SCCRS (Speech and Character Combined Recognition System) is developed for working on mobile devices such as PDA (Personal Digital Assistants). In SCCRS, the feature extraction is separately carried out for speech and for hand-written character, but the recognition is performed in a common engine. The recognition engine employs essentially CHMM (Continuous Hidden Markov Model), which consists of variable parameter topology in order to minimize the number of model parameters and to reduce recognition time. For generating context independent variable parameter model, we propose the SSMS(Successive State and Mixture Splitting), which gives appropriate numbers of mixture and of states through splitting in mixture domain and in time domain. The recognition results show that the proposed SSMS method can reduce the total number of GOPDD (Gaussian Output Probability Density Distribution) up to 40.0% compared to the conventional method with fixed parameter model, at the same recognition performance in speech recognition system.

Keywords: *Speech recognition, Character recognition, SCCRS, SSMS, Variable parameter*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

최근 이동형 정보단말기의 필요성이 증대됨에 따라 지능형 멀티모달 인터페이스에 관한 연구가 점증하고 있으며 이 결과를 적용한 상용단말기가 출현하고 있다. 특히 PDA (Personal Digital Assistants)와 같은 소형 모바일 기기에서는 음성인식 및 문자인식을 이용하여 편리하게 정보를 입력하는 방법을 채용하여 사용상의 편의성을 극대화하고 있다[1]. 그러나 현재까지 개발된 음성인식 및 문자인식을 이용하여 정보입력이 가능한 소형단말기는 각기 다른 인식모듈에 의해 동작하는 방법을 채용

책임저자: 정현열 (hychung@yu.ac.kr)
712-749 경북 경산시 대동 214-1 영남대학교 전자정보공학부 소재관 213호
영남대학교 정보통신공학과
(전화: 053-814-2496; 팩스: 053-814-5713)

하고 있기 때문에 소형단말기의 한정된 메모리용량으로 인하여 인식시스템 동작에 무리를 초래하고 있다. 메모리의 효율성을 제고하기 위한 한 방법으로서는 문자인식 및 음성인식을 하나의 엔진으로 수행하는 방법을 생각할 수 있다.

일반적으로 음성인식은 음성패턴의 다양성을 모델링하기 위해 HMM (Hidden Markov Model) 구조를 널리 채용하고 있으며, 이 구조는 온라인 필기체 문자인식의 경우에도 높은 인식성능을 나타내고 있다[2]. 따라서 HMM을 기반으로 하는 음성인식과 문자인식 공용 인식기를 구현할 경우 메모리가 제한된 소형단말기에서 높은 인식성능을 기대할 수 있을 것으로 생각된다. 또한, 소형단말기의 메모리 한계를 고려할 때 인식을 위한 기본 단위로서는 유사음소 또는 자소 단위의 CHMM (Continuous Hidden Markov Model) 모델로 구성하는 것이 유리할 것으로 생각된다. 이 경우 공용 인식기는 개별적인 인식기로 구성된 경우에 비해 인식률의 저하가 없도록 할 필요가 있으며 휴대용 기기에서 요구되는 인식 대기상태에서의 CHMM의 메모리 사용량의 효율성을 극대화시킬 필요가 있다. 또한 실시간으로 처리가 가능하기 위한 인식속도의 개선이 필수적이다.

일반적인 CHMM모델은 인식단위마다 동일한 상태수와 혼합수를 가지는 고정 상태수 모델 구조를 가지고 있다. 그러나 이 구조는 인식의 기본단위 사이의 다양성을 고려하지 않는 문제점을 가지고 있다. 예를 들어 온라인 문자인식의 경우, 한국어 자소 "ㄱ"의 경우 "ㄹ"에 비해 상대적으로 적은 수의 상태로 모델링하는 것이 효과적이라는 것이 저자들의 사전실험으로부터 확인한 바 있다[2]. 개별적인 인식 단위간의 적합한 모델의 구조를 결정하는 방법으로는 특징벡터가 출연하는 빈도로부터 작성되는 히스토그램을 통해 결정하는 방법과 확률과 정보이론으로 접근하는 AIC (AKAIKE Information Criterion)[3]과 BIC (Bayesian Information Criterion)[4] 방법 등이 있으며 이를 이용하여 가변 상태수 모델을 구성한 경우 고정 상태수 모델에 비해 동일한 성능에서 모델의 파라미터의 수를 감소시킬 수 있다. 그러나 가변 상태수 모델의 경우, 상태수를 결정하는데 있어 다른 인식모델의 상태수를 고려하지 않으므로 인하여 발생하는 인식률 저하를 고려하지 않고 있다. 또한 혼합수를 결정함에 있어서도 각 상태마다 동일한 혼합수를 가지도록 할 경우, 혼합수가 필요이상으로 증대되어 계산량 증가로 이어지는 문제점을 가지고 있다. 따라서 각각의 개별 인식단위가 각 인식단위에 적합한 상태수와 각 상태에 적합한 혼합수를 자동으로 결정하도록 하여 인식률을 제고하는 방법이 요구되어진다.

본 논문에서는 이를 위한 방법으로 GOPDD (Gaussian Output Probability Density Distribution)를 분할함으로 자동으로 모델 구조를 결정할 수 있는 방법을 이용한다. 이 방법은 일반적으로 상태단위로 공유된 문맥종속 모델링을 위한 SSS (Successive State Splitting)[5] 방법과 유사하나, 문맥방향 대신에 혼합수 방향 분할을 반복 수행한다. 본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 음성인식과 문자인식을 통합하기 위한 음성 문자 공용인식시스템의 구조에 대해 살펴보고, 3장에서는 가변 파라미터 모델 구조를 결정하는 기존의 방법에 대해서 기술하고 4장에서는 본 논문에서 제안한 GOPDD 분할방법에 대해 기술한 후, 5장에서는 성능시험 및 결과고찰에 대해 논의한 다음 6장에서는 결론과 향후 연구방향에 대해 기술한다.

II. 음성 문자 공용 인식 시스템

2.1. 시스템 구성

음성인식과 문자인식을 이용하는 PDA와 같은 소형-모바일 머신을 위하여 작은 메모리만으로 음성인식과 문자인식을 수행할 수 있는 음성 문자 공용 인식 시스템을 개발하였으며, 시스템의 구성은 그림 1과 같다[1].

마이크를 통해 입력된 음성과 터치 스크린으로부터 입력된 문자 데이터는 각각 다른 전처리 과정과 파라미터 추출 과정을 거쳐 동일한 파라미터 목적열로 만들어진다. 추출된 음성 및 문자 목적열은 레이블링 과정을 통해 음성은 48개의 문맥 독립 유사 음소 모델로, 문자는 67개의 자소 모델로 초기 CHMM모델이 구성된 후 총 115개의 M-mixture CHMM모델로 학습된다. 인식알고리즘은 OPDP (One Pass Dynamic Programming) 알고리즘을 이용한다[7].

2.2. 전처리

음성인식을 위한 전처리 과정으로 39차의 MFCC (Mel Frequency Cepstrum Coefficients) 파라미터를 추출한 후 잡음을 억제하기 위해 CMN (Cepstrum Mean Normalize) 처리과정을 거치며 문자인식을 위한 전처리과정은 평활화, 정규화, 재샘플링 과정을 거쳐 특징 파라미터로 6차의 위치 변화량 파라미터 및 9차의 비트맵 파라미터를 추출한다. 본 논문에서는 문자인식에 관한 전처리

과정만을 간략히 기술한다.

타블렛 혹은 터치 스크린으로부터 입력된 온라인 문자 데이터는 펜의 위치 정보를 나타내는 X, Y 좌표값이 100 샘플/초 이상으로 표본화 된다. 온라인 필기체 데이터는 필자에 따라 펜의 속도 및 모양에 있어서 다양한 형태로 표현되므로 이와 같은 변화를 전처리 과정에서 정규화할 필요가 있다. 그림 2에 문자인식을 위한 전 과정을 나타내었으며, 이하 각 과정을 간략한다.

A) 평활화 (Smoothing)

평활화는 필기시 손의 떨림 현상에 의해 울퉁불퉁하게 쓰인 문자를 미려하게 보정하여 위치좌표를 재추정한 후

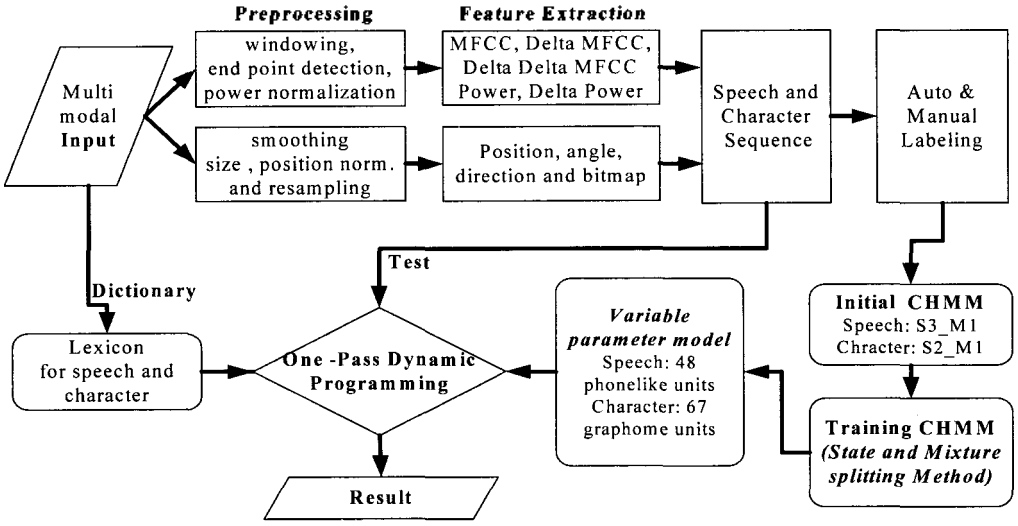

그림 1. 시스템 구성
Fig. 1. System architecture.

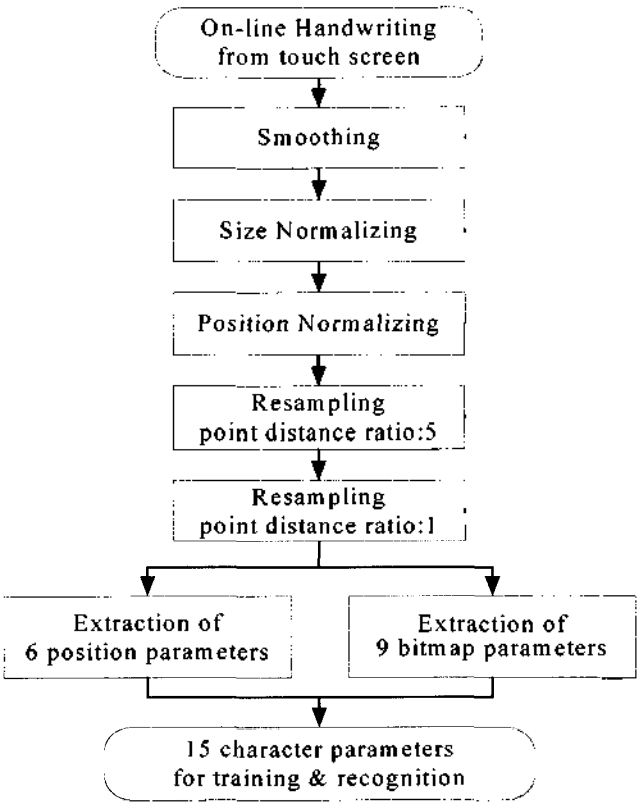

그림 2. 온라인 문자인식의 전처리
Fig. 2. Preprocessing of on-line character recognition.

표본화될 수 있도록 보정하는 방법으로 전후 이웃하는 점과의 평활화를 수행하는 과정이다. 재추정된 $\hat{x}_{ij}$는 식 (1)에 의해서 구해진다.

$$\hat{x}_{ij} = C_{j-n} \cdot x_{ij-n} + \ldots + C_j \cdot x_{ij} + \ldots + C_{j+m} \cdot x_{ij+m} \quad (1)$$

여기서, n과 m은 전향 및 후향 점들의 수를 나타내며, C_{j-n}, C_j, C_{j+n}는 각 점에 대한 가중치로써, 1/4, 1/2, 1/4를 각각 이용한다.

B) 정규화 (Normalization)

온라인 필기체의 경우 필자에 따라 다양한 필기 형태, 글자 크기 및 필기 위치가 나타난다. 이를 정규화하는 과정으로 크기 정규화와 위치 정규화를 수행한다. 크기 정규화는 글자단위의 외부분리가 이루어진 경우 x, y축의 최대/최소 값을 기준으로 정규화를 수행할 수 있으나, 단어단위로 외부 분리를 수행하는 경우에는 y축의 크기와 기준값과의 비로부터 x축과 y축의 정규화를 수행한다. 위치 정규화는 임의의 위치에 쓰인 필기의 좌표값을 최소값이 "0"이 되는 절대 좌표값으로 이동하는 정규화를 의미한다. 그림 3에 크기 및 위치 정규화 과정을 거친 단어를 나타내고 있다.

C) 재샘플링 (Resampling)

재샘플링은 필기속도의 불균일과 입력 장치의 샘플링 간격의 차이에 의해 발생되는 입력 점 사이의 불균등한 간격을 정규화하기 위한 과정으로 식 (2,3)을 이용하여 일정한 간격의 점들의 열로 새롭게 생성하게 한다.

$$px_i = \alpha \frac{(x_j - px_{i-1})}{Dis} + px_{i-1} \quad (2)$$

$$py_i = \alpha \frac{(y_i - py_{i-1})}{Dis} + py_{i-1} \quad (3)$$

$$Dis = \sqrt{(px_{i-1} - x_j)^2 + (py_{i-1} - y_j)^2} \quad (4)$$

여기에서 α는 샘플링 간격, j는 재샘플링 과정 전의 열, i는 재샘플링 과정 후 새롭게 만들어지는 열을 나타낸다. 전처리 과정이 완료되어 재샘플링된 각 점으로부터 국부적인 점의 위치를 나타내는 x, y좌표, 2차원의 국부적 각도, 2차원의 국부적 만곡을 나타내는 위치파라미터와 전역적인 정보를 나타내는 9차원의 비트맵 파라미터를 포함하여 총 15차의 파라미터를 추출하여 특징파라미터로 한다[8].

그림 3. 크기, 위치 정규화
Fig. 3. Size, position normalization.

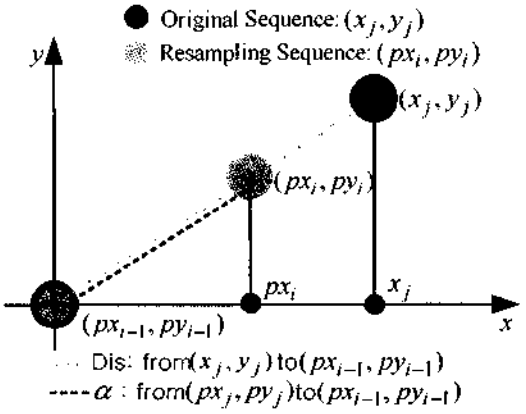

그림 4. 재샘플링
Fig. 4. Resampling.

III. 가변 파라미터 모델

문자인식 또는 음성인식을 위해 널리 이용되고 있는 CHMM 모델은 일반적으로 모델의 구조 (상태수, 혼합수)를 사전에 일정한 수로 고정하는 고정 파라미터 모델 구조를 가지고 있다. 그러나 고정 파라미터 모델구조로 이용하는 경우 각 인식 단위마다의 고유한 특성을 적절히 표현하기 어렵다. 따라서 그림 5와 같은 가변 파라미터

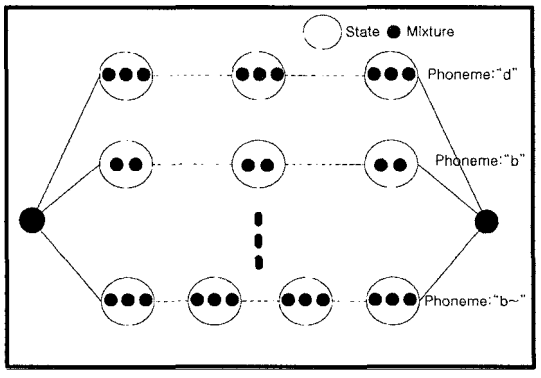

그림 5. 가변 파라미터 모델
Fig. 5. Variable parameter model.

모델을 이용하는 경우 인식률을 유지하면서 모델 파라미터의 수를 감소시킬 수 있다[4,9].

아하 본 절에서는 확률값에 근거한 일반적인 가변 파라미터 모델 구성방법인 MLC & BIC 알고리즘에 대해 기술하여, GOPDD 분할을 통해 가변 파라미터 모델 구성방법의 이해를 돕고자 한다[11-13].

3.1. 가변 파라미터 구조 결정방법

일반적인 CHMM의 모델 구조 T는 이를 구성하는 파라미터 $\theta=\{A,\ \mu,\ \Sigma,\ \omega\}$로 구성된다. 여기에서 A는 천이 확률을 나타내고, μ는 평균, Σ는 분산, ω는 확률분포의 가중치를 나타낸다. 이와 같은 모델 구조 T를 결정하는 문제는 모델의 상태수 S와 상태당 가우시안의 혼합수 M, 그리고 상태당 천이 구조를 결정하는 것을 의미하며, 선택된 모델 구조 $\hat{T}$는 데이터 집합 X와 구조 T로부터 식(5)와 같이 나타낼 수 있다. AIC, BIC와 같은 IC (Information Criterion) 방법에 있어서 모델구조 $\hat{T}$의 선택을 위한 베이시안 모델 선택 방법은 모델에 대한 사전 확률 $P(T)$이 동일하다고 가정할 경우 식 (6)의 $P(X|T)$를 최대로 하는 모델을 선택하는 것으로 나타낼 수 있다. 즉, 각 인식단위의 모델의 구조는 데이터의 집합으로부터 생성이 가능한 상태수와 가우시안 혼합수를 의미하는 것으로 현재 데이터를 효과적으로 표현이 가능한 하나의 구조를 식 (6)에 의해 선택한다.

$$\hat{T} = \arg\max_{T} P(T \mid X) = \arg\max_{T} p(T)P(X \mid T) \tag{5}$$

$$P(X \mid T) = \int p(X \mid T,\theta)p(\theta \mid T)d\theta \approx \log p(X \mid \theta_{ML}) - C(k,N) \tag{6}$$

여기에서 θ_{ML}은 MLE (maximum likelihood Estimator)를 이용하여 추정된 모델을 나타낸다. 모델 구조의 선택은 로그 우도항과 학습에 사용된 데이터의 수 N과 모델의 파라미터의 수 k에 의존하는 감쇄항 $C(k,N)$으로 나타낼 수 있다[4].

3.2. MLC에 의한 모델구조 선택방법

MLC (Maximum Likelihood Criterion)에 의해 모델의 구조를 선택하는 방법은 식 (7)과 같이 인식 단위마다 상태수와 혼합수에 따른 로그 우도값이 최대가 되는 모델 θ^*을 선택한다.

$$\theta^* = \max_{\hat{\theta}_i}\{\sum_{n=1}^{N}\log P(X_n \mid \hat{\theta}_i)\} \tag{7}$$

여기에서 $\hat{\theta}_i$는 MLE 방법에 의해 추정된 i번째 모델을 나타내고, X_n은 n번째 데이터, N은 데이터의 크기를 나타낸다. 그림 6에 모델 구조에 따른 로그 우도값의 예를 나타낸다. 한국어 "aa"음소의 경우, 3~6상태수와 1~4 혼합수를 가지는 모델 구조의 로그 우도항만을 고려하였을 때 5상태 4혼합 (S5_M4)모델이 최대가 됨으로 이를 이용하는 것이 유리함을 나타내고 있다.

3.3. BIC에 의한 모델구조 선택방법

IC 방법 중 널리 이용되고 있는 BIC 방법은 학습에 사용된 데이터의 집합 X가 모델 구성시 사용된 데이터의 수에 독립적이고 벡터 공간상에서 완전 분산되어 있다고 가정한다면 모델의 구조는 테일러 근사법을 이용하여 식(8)과 같이 ML항과 감쇄항의 합이 최대가 되는 모델 θ^{**}을 선택한다.

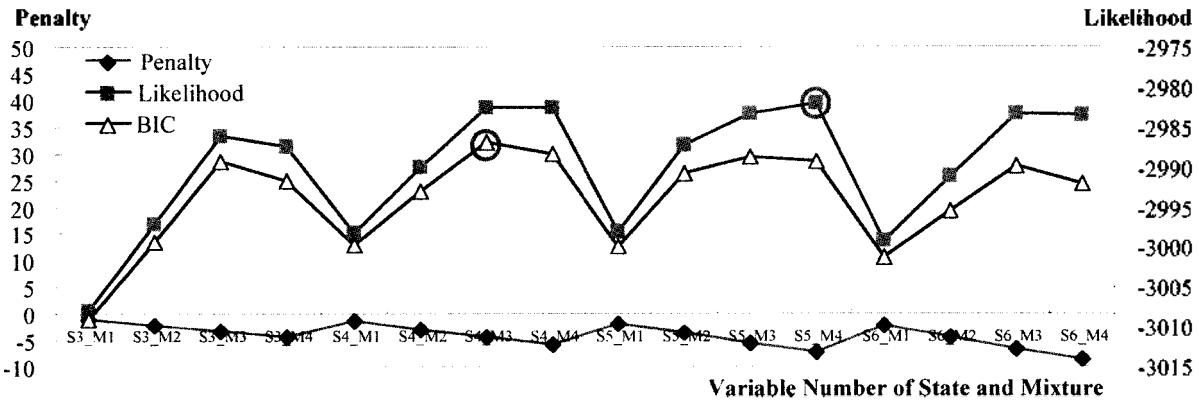

그림 6. 한국어 음소 "aa"의 로그 우도항, 감쇄항, BIC
Fig. 6. Log likelihood, penalty and BIC for Korean phoneme "aa".

$$\vartheta^{**} = \max_{\hat{\theta}_i}\{\sum_{n=1}^{N}\log P(X_n \mid \hat{\theta}_i) - \frac{k_i}{2}\log N\} \quad (8)$$

여기서 k_i는 i번째 모델의 파라미터의 수를 나타내며, N은 학습에 사용된 데이터의 수를 나타낸다. 이 방법을 이용한 결과를 그림 6에 나타내었다. 감쇄항을 고려한 BIC 모델 선택에서는 4상태 3혼합 (S4_M3) 모델이 우도값이 최대가 됨을 나타낸다.

IV. 상태 및 혼합 분할 모델

음소의 음향학적 특성은 각 음소의 특징, 화자의 특성, 발성빈도, 문맥적 정보, 시간적 정보와 같은 다양한 특성에 영향을 받는다. 이와 같은 특성을 상태가 공유된 문맥종속 모델로 효과적으로 모델링하기 위해 SSS-Free, ML (Maximum Likelihood)-SSS[10], DT (Decision Tree)-SSS[6] 등이 제안되었다. 일반적으로 문맥 종속 모델이 문맥 독립 모델에 비해 향상된 성능을 나타내나 상대적으로 많은 양의 메모리를 요구하여 소형 모바일 머신에 적용하기에는 어려운 면을 가지고 있으며, 문맥 독립 모델의 경우에도 모델의 파라미터 수가 증가하는 경우에는 필요 이상의 계산량이 증가하는 문제점을 가지고 있다.

기존의 가변파라미터 모델 선택방법인 MLC와 BIC는 우도항이 최대가 되는 모델 구조를 선택하는 방법으로 구조 선택 이전에 필요한 모델을 구성해야 함으로 필요이상의 많은 연산량을 요구하게 된다. 즉 필요한 모델구조를 생성하기 위한 소요 시간은 PC 2GHz 머신에서 국어정보학연구소 452 단어 데이터베이스 35명으로 학습을 수행하는 경우 대략 하나의 모델 구조 학습시 평균 2시간이 필요한 경우 전체 128 (= 2시간 * 4 (3상태~6상태) * 16 (1 GOPDD ~ 16 GOPDD)) 시간을 요구하게 된다. 따라서 대어휘 데이터를 이용하기 위해서는 사전에 모델 구조를 생성하지 않는 고속의 구조 결정방법이 요구된다.

이에 본 연구에서는 가변 파라미터 문맥 독립모델을 자동으로 생성하기 위해 GOPDD의 반복 분할을 통해 생성 하는 SSMS (Successive State and Mixture Splitting) 알고리즘을 제안한다. SSMS 알고리즘은 문맥종속 음향모델을 생성하는 SSS 알고리즘과는 달리 문맥독립 음향모델을 GOPDD 분할을 이용하여 음소별로 적합한 상태수와 각 상태에 적합한 혼합수를 가지는 모델을 구성한다. 또한 SSS 알고리즘의 경우 시간방향 분할과 상태방향 분할을 반복 수행하나, SSMS 알고리즘의 경우 문맥방향 분할대신 혼합수 방향 분할을 수행한다. 전체적인 구조는 그림 7과 같으며, 이하 SSMS 알고리즘을 각 단계별로 간략한다.

Step 1: 초기 학습 모델

SSMS의 초기모델은 음성의 경우 유사음소단위의 지속시간을 고려하여 3상태 1혼합 모델 구조를 이용하고, 문자의 경우는 "ㄱ", "ㄴ"과 같은 가로획과 세로획으로 구성된 기본자소의 형태를 고려하여 2상태 1혼합 모델 구조를 이용한다.

Step 2: 분할을 위한 GOPDD 선택

모든 상태 중에서 GOPDD의 분산가 가장 큰 상태를 분할할 상태로 선택한다. GOPDD 분산은 단일 가우시안의 분산값과 파라미터 추정에 이용한 학습 샘플수를 곱한 것을 기준으로 모든 상태별로 계산 후 크기가 최대인 하나의 상태를 선택한다. 각 상태 $S(i)$는 M개의 GOPDD로 구성되어 있으므로 정규화된 분산의 크기 d_i는 식 (9)로 나타낼 수 있다.

$$d_i = \sum_{k}^{K}\frac{\sigma_{ik}^{\,2}}{\sigma_{Tk}^{\,2}}\sqrt{n_i}$$

$$\sigma_{ik}^{\,2} = \sum_{m}^{M}\lambda_{im}\sigma^{2}_{imk} + \sum_{m}^{M}\sum_{m'=m+1}^{M}\lambda_{im}\lambda_{im'}(\mu_{imk} - \mu_{im'k})^2 \quad (9)$$

여기에서 K는 특징 파라미터의 차원수를 나타내며, $\lambda_{im}\lambda_{im'}$는 GOPDD의 혼합계수로써 MLE 학습시 생성되며, n_i는 i 상태 학습에 적용된 샘플의 수, $\sigma_{ik}^{\,2}$은 모든 학습에 사용된 샘플의 k번째 분산을 나타낸다.

Step 3: 상태 및 혼합 분할

분산의 크기가 최대인 상태는 시간 방향 분할과 혼합수 방향 분할을 수행한 후 각 분할 방향으로 Baum-Welch알고리즘을 수행한 후 두 개의 분할 방향 중 우도가 큰 모델 구조를 선택한다. 그림 8은 시간 방향 분할과 혼합수 방향 분할의 예를 나타내고 있다. 큰 원은 하나의 상태를 나타내며, 작은 원은 하나의 GOPDD를 나타낸다. 그림에서 2 GOPDD를 가지는 2 상태의 분산의 크기가 최대인 경우 시간 방향 분할은 동일한 상태를 시간열상의 2개의 상태로 분할하는 것을 의미하며, 혼합수 방향 분할은 2 GOPDD 가운데 분산의 크기가 큰 GOPDD를 하나를 증가시키는 것을 의미한다.

이에 비해 SSS 알고리즘의 경우 그림 9에서와 같이 각 상태는 문맥방향과 시간방향 분할을 수행한다. 이 경우

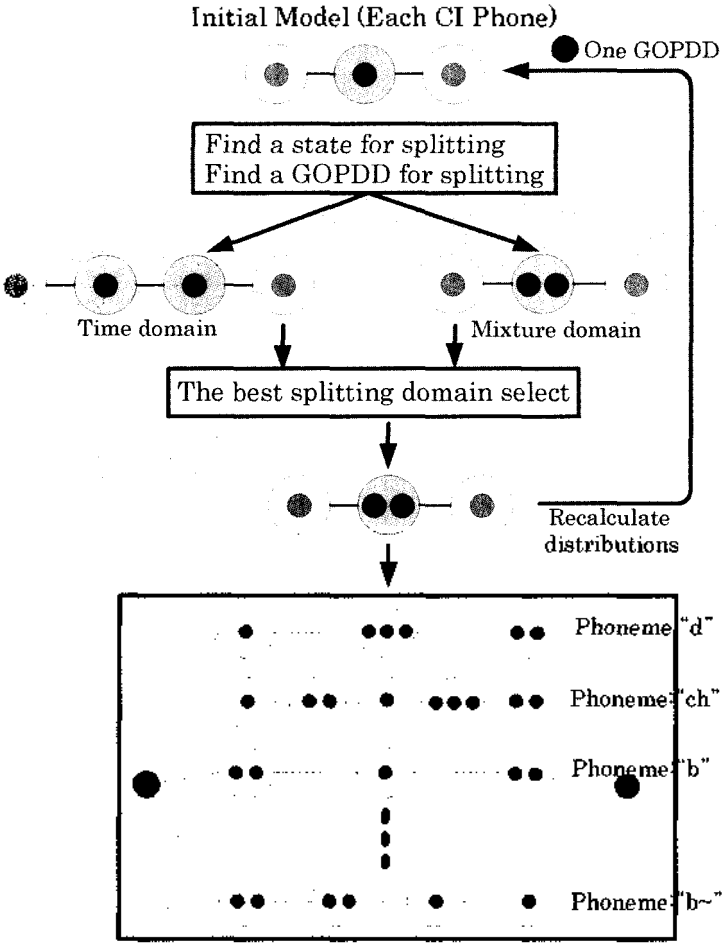

그림 7. SSMS 모델 생성
Fig. 7. Construction of a SSMS model.

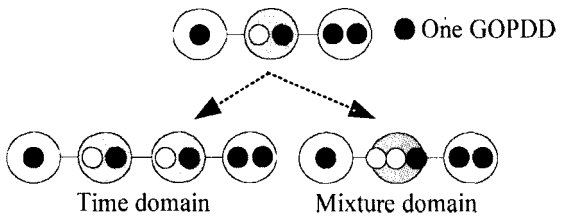

그림 8. 시간, 혼합수 방향 분할의 예
Fig. 8. Examples of splitting in time and mixture domain.

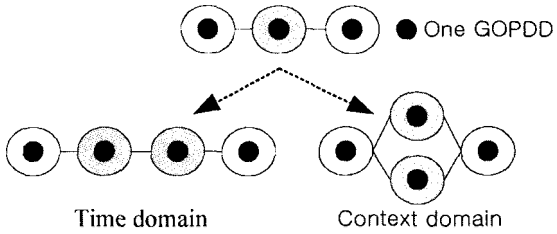

그림 9. 시간, 문맥 방향 분할의 예
Fig. 9. Examples of splitting in time and context domain.

모든 상태는 하나의 GOPDD만을 가지고 있다.

이후 적절한 최종 상태수에 이를 때까지 단계 2~3 과정을 반복한다. 위와 같은 단계를 거쳐 생성된 SSMS 모델은 각 인식 단위별로 적합한 상태수를 가지고 있으며 각 상태는 적합한 혼합수로 분할이 이루어진다. 따라서 SSMS 알고리즘을 일반화된 가변 파라미터 모델 생성 방법이라 할 수 있다.

V. 실 험

화자 독립 모델을 구성하기 위한 음성 데이터는 국어공학연구소에서 작성된 38명의 452단어 데이터베이스를 소형 모바일 기기에 적합하도록 8 KHz로 다운샘플링한 후 사용하였으며, 문자데이터는 KAIST에서 작성된 필기체 한글 데이터 중 10인의 1회 필기분을 이용하였다. 음성 및 문자 데이터의 분석 조건은 표 1과 같다.

SSMS기반 가변 파라미터 모델의 유효성을 확인하기

표 1. 음성/문자 데이터 분석 조건
Table 1. Analysis conditions for speech/character data.

	Speech	On-line Character
Preprocessing	8 KHz Sampling, 16 bits 16 ms Hamming Window 5 ms frame shift	100 samples/sec smoothing size/position normalization distance resampling
Feature	12 MFCCs, 1 Power, 12 Delta MFCCs, 1 Delta Power, 12 Delta Delta MFCCs, 1 Delta Delta Power	2 Absolute X,Y positions, 2 Angles 2 Curvatures 9 Modified bitmaps
DB	KLE Korean Words	KAIST Korean Written Characters
Model	M Mixture Variable Parameter CHMM	

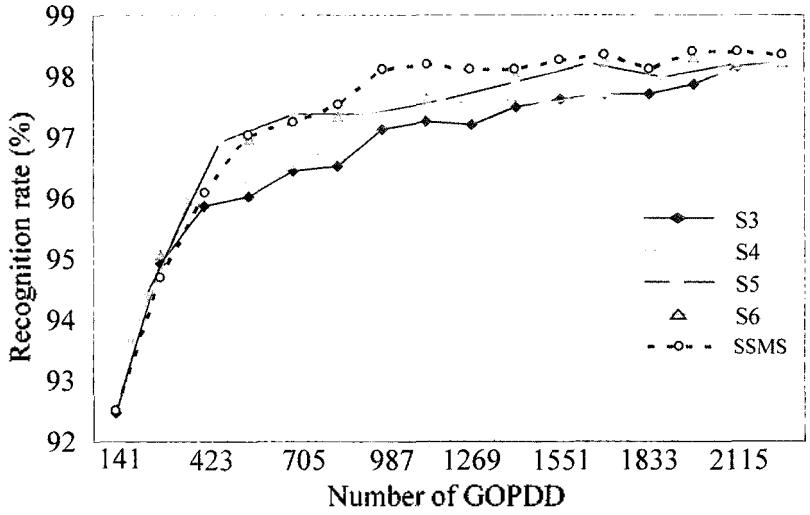

그림 10. 3~6상태의 고정 파라미터 모델과 SSMS 기반 가변 파라미터 모델의 인식률의 비교
Fig. 10. Comparison of recognition rates between fixed parameter models from 3 states (S3) to 6 states (S6) and variable parameter model by SSMS.

위해 고정 파라미터 모델, DT-SSS 기반 HM-net 모델과 비교하였다. 38명의 데이터베이스 가운데 35명을 모델훈련에, 3명을 화자독립 인식실험에 이용하였다. 그림 10은 고정 파라미터 모델과 SSMS 기반 가변 파라미터 모델과의 화자독립 인식률을 나타내고 있다. 고정 파라미터 모델의 경우 3~6상태 1혼합 모델로부터 1혼합수씩 증가시키면서 실험을 수행하였으며, 소요 메모리와 계산량을 고려하여 2256 GOPDD의 수를 최대값으로 설정하였다.

여기에서 총 GOPDD의 수는 음소의 수, 각 음소의 상태의 수, 각 상태의 혼합수의 곱으로 나타난다. 각 모델은 GOPDD가 증가함에 따라 인식률이 향상되고 있다. 그러나 점선으로 나타난 SSMS 기반 가변 파라미터 모델이 고정 파라미터 모델에 비해 최대 인식률인 98.2%에 빠르게 근접함을 보이고 있다.

기존의 가변 파라미터 생성방법인 MLC 모델은 2252개의 GOPDD 수로 구성되어 98.01%의 인식성능을 나타내었으며, BIC 모델은 2131개의 GOPDD 수로 97.86%를 나타내어 고정 파라미터 모델에 비해 성능 향상이 없음을 나타내었다. 이는 각 상태당 4 GOPDD 수 이하로 제한한 사전 실험에 비해 각 상태당 16 GOPDD를 가질 수 있는 경우에는 우도항만으로 모델을 선택하는 방법은 유효하지 않음을 나타내고 있다.

표 2. 98.2%의 최대 인식률에 도달하는 각 모델의 GOPDD의 수 (#GOPDD: GOPDD의 수)
Table 2. Number of GOPDD of each model reaching to the maximum recognition accuracy of 98.2% (#GOPDD: number of GOPDD).

Model	S3	S4	S5	S6	SSMS
#GOPDD	2115	2256	1645	1692	987

표 2는 고정 파라미터 모델의 최대인식률인 98.2%에 도달하는 GOPDD의 수를 나타내고 있다. 테이블에서 5상태의 고정 파라미터 모델의 경우 1645개의 GOPDD로

표 3. SSMS, MLC, BIC 기반 모델 구조의 예 (#M: 혼합수, #S: 상태수)
Tabel 3. Examples of model topology by SSMS, MLC, BIC Model (#M: number of Mixture, #S: number of State).

Phoneme		SSMS Model							MLC	BIC
		#S \ M	1	2	3	4	5	6	#S_#M	#S_#M
ㄱ	g	5	4	4	7	5	6	-	S6_M8	S5_M9
	g~	4	7	5	6	4	-	-	S3_M16	S3_M16
	g+	3	5	7	6	-	-	-	S3_M16	S3_M15
ㄴ	n	3	6	8	11	-	-	-	S3_M16	S3_M16
ㄷ	d	4	4	4	7	5	-	-	S6_M8	S6_M8
	d~	4	5	6	5	4	-	-	S6_M8	S5_M9
	d+	4	6	6	6	5	-	-	S3_M16	S3_M16
ㄹ	r	5	5	5	4	3	4	-	S3_M16	S3_M16
	l	3	7	9	7	-	-	-	S3_M16	S3_M15
ㅁ	M	3	6	11	9	-	-	-	S3_M16	S3_M12
ㅂ	b	4	7	5	5	5	-	-	S4_M12	S3_M14
	b~	4	5	5	4	5	-	-	S4_M12	S4_M11
	b+	3	6	7	6	-	-	-	S4_M12	S4_M12
ㅅ	s	3	3	10	7	-	-	-	S6_M8	S6_M7
ㅇ	ng	3	5	9	9	-	-	-	S3_M16	S3_M15
ㅈ	z	6	4	2	4	2	3	4	S6_M8	S6_M8
	z~	3	7	5	6	-	-	-	S6_M8	S5_M9
ㅊ	ch	6	5	3	4	2	2	2	S6_M8	S6_M8
ㅋ	k	4	7	5	4	5	-	-	S6_M8	S5_M9
ㅌ	t	3	9	4	6		-	-	S6_M8	S6_M8
ㅍ	p	4	5	4	6	5	-	-	S6_M8	S5_M10
ㅎ	hh	4	4	5	5	4	-	-	S4_M12	S4_M12
	hh~	3	9	8	5	-	-	-	S3_M16	S3_M15
ㄲ	gg	6	7	4	4	3	3	2	S4_M12	S4_M12
ㄸ	dd	5	7	4	5	4	3	-	S6_M8	S6_M8
ㅃ	bb	5	7	4	3	3	3	-	S6_M8	S5_M9
ㅆ	ss	6	6	9	2	2	2	2	S6_M8	S6_M8
ㅉ	zz	6	6	3	2	1	1	2	S6_M8	S6_M7
ㅏ	aa	3	3	7	8	-	-	-	S4_M12	S4_M11
ㅓ	axr	4	3	6	7	7	-	-	S4_M12	S3_M15
ㅗ	ao	3	3	10	10	-	-	-	S3_M16	S3_M16
ㅜ	uh	3	5	8	12	-	-	-	S3_M16	S3_M15
ㅡ	U	3	6	9	10	-	-	-	S3_M16	S3_M15
ㅣ	ih	3	4	9	11	-	-	-	S3_M16	S3_M12
ㅐ	ae	3	4	7	8	-	-	-	S3_M16	S3_M14
ㅔ	eh	3	3	7	14	-	-	-	S4_M12	S4_M11
ㅑ	ja	4	3	2	3	7	-	-	S6_M8	S5_M9
ㅕ	jv	3	7	4	7	-	-	-	S4_M12	S4_M11
ㅛ	jo	3	4	3	6	-	-	-	S4_M12	S3_M14
ㅠ	ju	3	7	5	7	-	-	-	S4_M12	S4_M11
ㅘ	wa	3	5	6	8	-	-	-	S4_M12	S4_M12
ㅝ	wv	3	7	5	4	-	-	-	S6_M8	S6_M8
ㅚ	we+	4	5	5	7	8	-	-	S4_M11	S4_M10
ㅞ,ㅙ	we	3	5	7	9	-	-	-	S3_M16	S4_M12
ㅟ	wi	3	5	7	10	-	-	-	S3_M16	S3_M15
ㅖ, ㅒ	je	4	5	5	5	8	-	-	S4_M12	S4_M12
ㅢ	wi+	4	3	3	5	7	-	-	S6_M8	S5_M9
Total #GOPDD		987							2252	2131

표 4. DT-SSS 기반 HM-net 모델의 인식률 (#GOPDD)
Table 4. Recognition rate of HM-net based DT-SSS (#GOPDD).

#S \ #M	1	2	4
300	95.28 (300)	97.42 (600)	98.08 (1200)
600	97.49 (600)	98.20 (1200)	98.71 (2400)
1000	98.01 (1000)	98.67 (2000)	98.97 (4000)
2000	98.75 (2000)	98.75 (4000)	99.19 (8000)

구성되나, SSMS 의 경우 987개로 구성되어 고정 파라미터 모델에 비해 40%의 모델 파라미터가 감소됨을 나타내고 있다.

표 3은 SSMS, MLC, BIC 가변 파라미터 모델의 예로써, "g" 음소의 경우 SSMS 모델은 총 상태수가 5이며, 첫 번째 상태는 4 혼합, 두 번째 상태 4 혼합, 세 번째 상태 7 혼합으로 구성되어 있음을 나타내며, MLC 모델은 총 상태수가 6이며, 각 상태는 동일한 8 혼합으로 구성되어 있으며, BIC 모델은 총 상태수가 5이며, 각 상태는 동일한 9 혼합으로 구성되어 있음을 나타내고 있다.

표 4는 DT-SSS 기반 문맥 의존 HM-net 모델의 화자독립인식률을 나타내고 있다. 실험 결과 문맥 의존모델이 문맥 독립모델에 비해 높은 인식률을 나타내고 있음을 확인할 수 있다. 그러나 이 경우에도 98.2%이상의 인식률을 위해서는 1000개의 GOPDD 이상으로 상태 공유 모델을 작성해야 함을 확인할 수 있다.

VI. 결 론

현재까지 구현된 소형-모바일 머신에서의 음성인식과 문자인식은 개별적인 개체로 구현되어 있어 많은 메모리와 계산량을 필요로 하고 있다. 이에 본 논문에서는 개별적인 인식과정을 수행하는 음성인식과 문자인식을 하나의 CHMM으로 모델을 구성한 후 동일한 처리과정으로 인식을 수행할 수 있는 공용 인식 시스템을 제안하였다. 본 시스템은 전처리단과 특징 파라미터 추출과정은 음성과 문자 처리부가 분리되어 독립적으로 수행되나, 인식 과정은 단일한 과정으로 구성됨으로 소형-모바일 머신에서 작은 메모리로 음성인식과 문자인식을 동시에 제공할 수 있다.

일반적인 CHMM은 인식단위마다 동일한 상태수와 동일한 혼합수를 가지는 고정 파라미터 모델을 사용함으로 인식단위의 다양한 복잡도를 고려하지 못하였다. 따라서 인식률을 유지하면서 파라미터의 수를 감소시킬 수 있는 가변 파라미터 모델을 사용하는 것이 유리하며, 이를 위해 기존의 MLC, BIC 방법의 경우에는 사전에 선택을 위한 모델 구조를 생성이 필요하다. 그러나 이는 많은 연산량을 요구하여 대어휘 데이터를 이용하기 위해서는 사전에 모델 구조를 생성하지 않는 고속의 구조 결정방법이 요구된다.

본 논문에서는 가변 파라미터 모델을 고속으로 생성하기 위해 SSMS 방법을 제안하였다. 이 방법은 문맥독립모델 생성을 위해 문맥 방향 분할 대신에 혼합수 방향 분할을 수행함으로 각 상태당 혼합수를 감소시킬 수 있다. SSMS기반 가변파라미터 모델의 유효성을 확인하기 위해 음성 인식 실험결과 SSMS기반 가변 파라미터 모델이 동일한 인식성능을 가지면서도 고정 파라미터 모델에 비해 모델 파라미터 수가 40% 감소함을 확인할 수 있었다. 이는 제안한 SSMS 방법이 PDA와 같은 소형의 머신에서 효과적으로 적용가능함을 의미한다.

참 고 문 헌

1. S.-Y. Suk, M.-J. Kim, and H.-Y. Chung, "An on-line speech and character combined recognition system for multimodal interfaces," *EALPIIT Proc.*, 89-92, 2002.
2. B.-K. Sin, and J. Kim, "A statistical approach with HMMs for on-line cursive hangul (Korean Script) recognition," *Second International Conference on Document Analysis and Recognition Proc.*, 147-150, Zuchuba, Japan, 1993.
3. H. Tong, "Determination of the order of a markov chain by Akaike`s information criterion," *Journal of Applied Probability*, 12, 488-497, 1975.
4. D. Li, A. Biem, and J. Subrahmonia, "Hmm topology optimization for handwriting recognition," *ICASSP Proc.*, 2001.
5. J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *ICASSP-92 Proc.*, 1, 573-576, 1992.
6. H. Takaki, K. Mashahru, I. Akinori, and K. Masaki, "A Study on HM-Nets using Decision Tree-based Successive Splitting," *ICSP-97 Proc.*, 383-387, 1997.
7. S. Nakagawa, "A connected spoken word recognition method by O(n) dynamic programming pattern matching algorithm," *ICASSP Proc.*, 296-299, 1983.
8. G. Ralph, M. Stefan, and W. Alex, "Run-on recognition in an on-line handwriting recognition system," Carnegie Mellon Univ Press, 1997.
9. J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *ICASSP-92 Proc.*, 1, 573-576, 1992.
10. H. Takaki, K. Mashahru, I. Akinori, and K. Masaki, "A study on HM-Nets using decision tree-based successive splitting," *ICSP-97 Proc.*, 383-387, 1997.

11. 석수영, 김민정, 김광수, 정호열, 정현열, "한국어 음성/문자 공용인식기의 성능향상을 위한 가변 상태수 CHMM 모델의 구성," 한국음향학회 학술발표대회 논문집, 21 (1), 95-98, 2002.
12. 석수영, 김민정, 정호열, 정현열, "Local Maximum 방법을 이용한 가변 파라미터 CHMM 모델의 구성," 음성통신 및 신호처리 학술대회 논문집, 19 (1), 211-214, 2002.
13. S. Y. Suk, M. J. Kim, H. Y. Jung, and H. Y. Chung, "An On-Line speech and character combined recognition system using CHMM with different model parameter," *HWESPAC 8 Proc.*, WB32, 2003.

저자 약력

● 석 수 영 (Soo-Young Suk)

1998년 2월: 계명대학교 물리학과 (이학사)
2000년 2월: 영남대학교 일반대학원 멀티미디어 통신공학과 (공학석사)
2000년 3월~현재: 영남대학교 일반 대학원 정보통신공학과 (박사수료)
※ 주관심분야: 디지털신호처리, 음성인식, 문자인식

● 정 호 열 (Ho-Youl Jung)

1988년 2월: 아주대학교 전자공학과 (공학사)
1990년 2월: 아주대학교 전자공학과 (공학석사)
1993년 2월: 아주대학교 전자공학과 (박사수료)
1998년: (프)리옹국립응용과학원 (INSA de Lyon) 전자공학전공(공학박사)
1998년 4월~1998년: 12월 (프)CREATIS 박사후 과정
1999년 3월~현재: 영남대학교 전자정보공학부 전임강사
※ 주관심분야: 음성, 영상 신호처리, 인공지능, 디지털 워터마킹

● 정 현 열 (Hyun-Yeol Chung)

1975년: 영남대학교 전자공학과 (공학사)
1989년: 일본 동북대학교 정보공학과 (공학박사)
1989년 3월~현재: 영남대학교 전자정보공학부 교수
1992년 7월~1993년: 7월 미국 CMU Robotics 연구소 객원연구원
1994년 12월~1995년 2월: 일본 토요하시기술과학대학 외국인 연구자
2000년 6월~2000년 8월: 미국 Qualcomm Inc. 수석 엔지니어
※ 주관심분야: 음성인식, 화자인식, 음성합성 및 DSP 응용분야