한국음향학회지 제21권 제8호 pp. 735~743 (2002)

강건한 문맥독립 화자식별을 위한 프레임 선택방법, 복합방법, 수정된 가중모델순위 방법

Frame Selection, Hybrid, Modified Weighting Model Rank Method for Robust Text-independent Speaker Identification

김 민 정*, 오 세 진**, 정 호 열*, 정 현 열*
(Min-Jung Kim*, Se-Jin Oh**, Ho-Youl Jung*, Hyun-Yeol Chung*)

*영남대학교 정보통신공학과, **대구과학대학 디지털 정보통신계열
(접수일자: 2002년 10월 17일; 채택일자: 2002년 10월 30일)

본 논문에서는 세 가지 문맥독립 화자식별방법을 제안한다. 먼저, 화자 식별시 성도의 특성을 충분히 표현하지 못한 프레임이 포함되지 않도록 하는 프레임선택 (Frame Selection; FS)방법을 제안한다. 이 방법은 각 프레임에서 가장 큰 유사도와 두 번째로 큰 유사도의 차이를 평가하여 중요 프레임을 선택한 후, 선택된 프레임만을 이용하여 유사도를 계산하는 방법이다. 두 번째로 제안하는 복합 (Hyrid)방법은 FS와 가중모델순위 (Weighting Model Rank; WMR)방법을 결합시킨 것으로, FS방법을 이용하여 중요 프레임을 선택한 후, 지수함수 가중치를 이용하여 식별화자를 결정하는 것이다. 마지막으로 제안하는 수정된 가중모델순위 (Modified WMR; MWMR)방법은 식별화자를 결정할 때 유사도의 상대적 위치만을 고려하였던 기존의 WMR방법과는 달리 유사도와 유사도의 상대적 위치를 함께 고려하는 방법이다. 화자식별실험결과 제안한 방법들이 기존의 ML방법보다 향상된 식별률을 보였으며, 복합 방법 및 MWMR방법의 경우에는 WMR방법보다 각각 약 2%와 3%의 향상된 식별률을 나타내어 제안한 방법들의 유효성을 확인할 수 있었다.
핵심용어: 화자식별·검증, 최대유사도, 프레임선택, 가중모델순위, 수정된 가중모델순위
투고분야: 음성처리 분야 (2.5)

In this paper, we propose three new text-independent speaker identification methods. At first, to exclude the frames not having enough features of speaker's vocal from calculation of the maximum likelihood, we propose the FS (Frame Selection) method. This approach selects the important frames by evaluating the difference between the biggest likelihood and the second in each frame, and uses only the frames in calculating the score of likelihood. Our secondly proposed, called the Hybrid, is a combined version of the FS and WMR (Weighting Model Rank). This method determines the claimed speaker using exponential function weights, instead of likelihood itself, only on the selected frames obtained from the FS method. The last proposed, called MWMR (Modified WMR), considers both original likelihood itself and its relative position, when the claimed speaker is determined. It is different from the WMR that take into account only the relative position of likelihood. Through the experiments of the speaker identification, we show that the all the proposed have higher identification rates than the ML. In addition, the Hybrid and MWMR have higher identification rate about 2% and about 3% than WMR, respectively.
Keywords: *Speaker identification · verification, Maximum likelihood, Frame selection, Weighting model Rank, Modified weighting models rank*
ASK subject classification: *Speech signal processing (2.5)*

책임저자: 김민정 (manjuk@orgio.net)
주 2-749 경북 경산시 대동 214-1
영남대학교 정보통신공학과 소재관 213호
(전화: 053-813-2496; 팩스: 053-814-5713)

I. 서 론

최근 정보통신 기술의 급격한 발전과 더불어 각 개인의 정보보호를 위한 인증 수단에 관한 연구가 점증하고 있다. 음성을 이용한 개인 확인은 카드, 키 등과 같은 인공적인 수단보다는 매우 편리할 뿐만 아니라 분실위험이나 도난위험이 전혀 없어 매우 안전하다. 또 음성을 이용한 화자식별은 손이나 다른 도구를 전혀 필요로 하지 않으므로 급속히 발전하는 정보화 시대의 각종 시스템 구현에 중요한 기술로서 각광받고 있다.

이를 위한 화자식별 시스템은 음성을 이용하여 고객의 요구를 만족시키는 여러 분야에서 중요한 역할을 할 것으로 기대된다. 예를 들어 전화선 또는 인터넷을 이용한 은행간 송금, 잔고조회, 쇼핑, 음성메일, 개인정보조회, 예약, 컴퓨터의 원격접근, 극비장소에의 접근통제 등에 편리하게 이용될 수 있기 때문이다. 특히 이와 같은 요구에 따라 화자식별 기술이 인터넷을 이용한 각종 개인의 인증방법에 이용되기 시작되었으며 법정 증빙도구로서도 이용되기 시작되고 있다[1].

그러나 이와 같은 음성을 이용한 화자식별의 단점은 음성의 물리적 특성은 항상 불변이 아니며 마이크 특성, 전송선로 또는 배경잡음 등에 의해 쉽게 변할 수 있다는 것이다. 만약 시스템이 어떤 고객 음성의 다양한 변화를 수용한다고 하면 이는 또한 목소리가 비슷한 다른 사람의 음성을 수용한다는 의미가 되므로 화자식별은 실패하게 된다. 따라서 화자식별에 있어서 무엇보다 중요한 것은 쉽게 모방할 수 없으며 전송특성에 영향을 받지 않는 안정된 물리적 특징을 특징 파라미터로 사용하는 것이라고 할 수 있다.

화자식별 시스템은 발성의 종류에 따라 문맥종속 및 문맥독립 화자식별로 나눌 수 있는데, 문맥독립 화자식별의 경우 보안성이 높아 이에 관해 많은 연구가 진행 중이다[1]. 문맥독립 화자식별 방법으로서는 장시간 (Long-term) 통계에 기반한 방법[2], VQ (Vector quantization)에 기반한 방법[3], HMM (Hidden Markov Model)이나 GMM (Gaussian mixture model)에 기반한 방법[4] 등이 있으며, 이러한 방법들 중 화자특성 변화의 표현과 화자식별 성능면에서 좋은 결과를 나타내고 있는 GMM에 기반한 방법이 가장 유리한 것으로 알려져 있다[5].

한편, 화자식별 및 검증과 같은 화자인식은 고정도의 인식률을 요구하기 때문에 실제 인식 시스템을 구현하는 데는 많은 어려움이 있다. 따라서 강건한 화자인식성능을 위하여 최대 유사도 (Maximum Likelihood; ML)를 이용한 식별방법을 많이 이용하고 있다. 또한 최근 K.Markov 등[6]에 의해 제안된 가중모델순위방법 (Weighting Model Rank; WMR)의 경우, ML방법보다 평균 3%의 향상된 식별성능을 나타내고 있다.

ML방법의 경우에는 화자들 사이의 변별력이 유사도의 변별력에만 의존하므로 비교화자의 수가 많거나 비슷한 특성을 가지는 화자가 존재할 경우, 변별력이 떨어지는 단점이 있었다. WMR방법의 경우, ML방법보다는 화자들 사이의 변별력을 높일 수 있지만, 유사도 본래의 변별력을 고려하지 않고 유사도의 상대적 위치에 따라 결정된 가중치만을 화자식별에 이용하므로 화자의 특성을 잘 표현하지 못하는 프레임에서도 상대적인 유사도 값만 크다면 큰 가중치가 부여될 수 있는 문제점이 있다.

이와 같은 문제점을 해결하는 방법으로는 특정화자에 의해서 발성된 음성 중 특정 프레임이 그 화자의 특징정보를 많이 보유하고 있는 경우 그 프레임은 다른 화자 모델들과의 유사도 차이가 클 것이므로 이와 같은 유사도 차이가 큰 프레임만을 이용하여 식별화자를 결정하는 방법을 생각할 수 있다. 또한 이들 프레임에 가중치를 부여하여 식별에 이용함으로써 화자들 사이의 변별력을 더욱 높이는 방법도 고려할 수 있으며 식별화자를 결정하기 위한 전체 점수를 계산할 때, 프레임 단위로 계산된 유사도와 유사도의 상대적 크기에 따라 결정할 수 있는 지수함수 가중치를 함께 이용하는 방법도 고려할 수 있다.

따라서 본 논문에서는 기존의 화자식별 시스템 가운데에서도 성능이 우수한 GMM방식을 기반으로 하여 위에서 언급된 변별력이 높은 프레임 선택방법, 여기에 가중치를 부여하여 식별에 이용하는 방법 등을 이용하여 화자식별 성능을 제고하는 방법, 그리고 프레임 단위로 계산된 유사도와 유사도의 상대적 위치도 함께 고려하는 방법을 제안하고 실험을 통하여 제안한 방법의 유효성을 확인하고자 한다.

본 논문의 구성은 다음과 같다. II장에서는 기존의 화자식별방법에 대해 살펴보고, III장에서는 본 논문에서 제안하는 화자식별 방법들에 대해서 자세히 설명한다. IV장에서 식별실험을 실시한 후 그 결과를 고찰한 후, 마지막으로 V장에서 결론을 맺는다.

II. 기존의 화자식별방법

2.1. 프레임 단위 최대 유사도 (Frame Level Maximum Likelihood)방법

일반적인 화자식별 방법은 Bayes의 정리[7]에 따라 식 (1)과 같이 N명의 화자 중 사후확률 $P(\lambda_i|X)$, $1 \le i \le N$를 최대로 하는 모델 λ_i의 화자 i^*를 찾는 것이다.

$$P(\lambda_i|X) = \frac{p(X|\lambda_i)P(\lambda_i)}{p(X)} \tag{1}$$

여기서 사전정보가 없기 때문에 사전확률 $P(\lambda_i)$는 식 (2)와 같이 주어진다.

$$P(\lambda_i) = \frac{1}{N}, \quad 1 \le i \le N \tag{2}$$

식 (1)의 분모인 $p(X)$는 발성 X의 빈도에 대한 무조건적인 유사도를 나타내며 모든 화자에 대해 동일한 값을 가진다. 따라서 식 (1)에서 식별화자는 $p(X|\lambda_i)$의 사후확률이 최대가 될 때의 화자가 되며 다음과 같이 결정된다.

$$i^* = \arg\max_i p(X|\lambda_i) \tag{3}$$

일반적인 화자식별 시스템에서는 식별화자를 결정하는데 있어서 발성된 음성전부로부터 유사도를 계산한다. 그러나 이 경우 발성문장의 내용이 달라질 경우 문제가 있다. 이를 개선하기 위해 제안된 프레임단위를 이용한 화자식별의 경우에는 백그라운드 화자모델에 의한 유사도 정규화를 통해 발성문장의 내용변화에 따른 특징변화를 최소화 할 수 있기 때문에 시스템의 성능을 향상시킬 수 있었다[8]. 프레임단위 유사도를 식 (1)에 적용하면 식 (4)와 같이 된다.

$$p_{norm}(x_t|\lambda_i) = \frac{p(x_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(x_t|\lambda_b)} \tag{4}$$

여기에서 $p(x_t|\lambda_b)$는 t 프레임에서의 b 백그라운드 화자모델의 유사도이며, $p_{norm}(x_t|\lambda_i)$는 t 프레임에서 백그라운드 화자모델에 의해 정규화 된 i번째 화자의 유사도를 나타낸다.

식 (4)를 이용하여 각 화자의 입력발성 x_t, ($t=1,2,\cdots,T$)의 각 프레임별 유사도를 계산하여 합한 각 화자모델 i에 대한 점수로부터 식별화자는 식 (5)를 이용하여 결정한다.

$$Sc_i(X|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T} \log p_{norm}(x_t|\lambda_i) \tag{5}$$

이러한 최대 유사도 방법은 비교화자의 수가 작을 때는 높은 화자 식별성능을 나타내나 비교화자의 수가 많아지거나 비슷한 특성을 가지는 화자가 존재하면 화자들 사이의 변별력이 떨어지는 단점이 있다.

2.2. 가중모델순위 (Weighting Model Rank) 방법

WMR (Weighting Model Rank) 방법[6]은 식별화자를 결정하는 점수를 계산할 때 테스트 음성과 화자모델들과의 프레임 유사도를 그대로 사용하지 않고, 계산된 유사도들의 상대적 위치에 따라 정해진 가중치를 점수 계산에 사용한다. 즉 프레임단위에서 높은 유사도 값을 가지는 화자모델은 더 높은 값을, 낮은 유사도 값을 가지는 화자모델은 더 낮은 값을 부여하여 화자들간의 변별력을 더 높이는 방법이다.

WMR 방법을 단계별로 간략하면 다음과 같다. 첫 번째 단계에서는 각 테스트 벡터 x_t, ($t=1,2,\cdots,T$)에서 프레임 유사도 $p(x_t|\lambda_i)$, $i=1,\cdots,N$ 계산하고 이를 내림차순으로 정렬한다. 즉 가장 큰 유사도를 가지는 화자모델은 최상위에 위치시키고, 가장 낮은 유사도를 가지는 화자모델은 최하위에 위치시킨다. 표 1은 각 프레임에서의 화자모델의 순위와 가중치의 관계를 나타낸 것이다.

두 번째 단계에서는 화자모델의 각 순위에 따라 가중치 $w(r)$을 결정한다. 이때, 가중치는 식 (6)과 같은 지수함수를 이용하는 방법이 많이 이용되고 있다[6].

$$w(r_\lambda) = \exp(A - Br_\lambda), \quad r_\lambda = 1,\cdots,N \tag{6}$$

여기서 A와 B는 $w(1) \approx N$이 되도록 설정한다.

세 번째 단계에서는 각 모델 λ_i의 순위에 해당하는 가중치 $w_t(r_{\lambda_i})$를 유사도 $p(x_t|\lambda_i)$ 대신 이용하여 전체 점수 $Sc(X|\lambda_i)$를 계산한다. $Sc(X|\lambda_i)$는 식 (7)과 같이 $t=1,\cdots,T$에서 모든 가중치를 더하여 구한다.

$$\log Sc(X|\lambda_i) = \sum_{t=1}^{T} w_t(r_{\lambda_i}) \tag{7}$$

여기서 $w_t(r_{\lambda_i})$는 시간 t에서 순위가 r_{λ_i}인 모델 i의 가중치를 나타내다. WMR방법은 프레임단위 최대 유사도 방법보다 화자들 사이의 변별력을 높일 수 있는 장점이 있지만, 유사도 본래의 변별력은 고려되지 않고 유사도의 상대적 위치에 따라 결정된 가중치만을 화자식별에

표 1. 화자모델의 N-best 유사도 list
Table 1. N-best likelihood list for speaker model.

1	p_j^t	$w(1)$	Model λ_j (최대 유사도)
2	p_j^t	$w(2)$	Model λ_j
...	...	...	...
m	p_k^t	$w(m)$	model λ_k
...	...	...	...
N	p_b^t	$w(N)$	Model λ_b (최소유사도)

이용하므로 화자의 정보를 많이 포함하고 있지 않은 프레임에서도 유사도의 상대적 위치만 높다면 큰 가중치가 결정된다는 문제점이 있다.

III. 제안 방법

앞 절에서 살펴본 기존의 화자식별방법들의 단점을 보완하고 강건한 화자식별성능을 얻기 위하여 본 논문에서는 다음의 방법들을 제안한다.

3.1. 유사도 차를 이용한 프레임 선택 (Frame Selection; FS)방법

입력음성이 식별하고자 하는 화자의 정보를 많이 포함하고 있다고 가정한다면, 식별하고자 하는 화자모델과의 유사도가 다른 화자모델들과의 유사도보다는 훨씬 크게 나타난다. 이 경우 다른 화자들과의 유사도 차이가 큰 프레임만을 선택할 수 있다면, 식별하고자 하는 화자의 특성이 많이 포함되어 있는 프레임들을 선택 가능하다. 유사도 차를 이용한 프레임 선택방법은 각 프레임에서 가장 큰 유사도를 가지는 화자와 두 번째 큰 유사도를 가지는 화자의 프레임 유사도 값이 일정값 이상의 차이가 있는 프레임은 해당 프레임의 각 화자들의 유사도를 전체 점수의 계산에 이용하고, 일정값 이하인 프레임은 제외시키면 결과적으로 전체 점수 계산시 화자의 정보를 많이 포함하고 있는, 즉 변별력이 큰 프레임들만을 이용하게 되는 것이다.

이 경우 먼저 식 (4)를 이용하여 계산된 프레임단위 유사도로부터 상위 두 유사도는 식 (8)에 의해 찾을 수 있다.

$$r^t_{m1} = \max\{\, p_{norm}(x_t|\lambda_1),\ p_{norm}(x_t|\lambda_2),\ \cdots,\ p_{norm}(x_t|\lambda_N)\}$$

$$r^t_{m2} = \max\{\, p_{norm}(x_t|\lambda_1),\ p_{norm}(x_t|\lambda_2),\ \cdots,\ p_{norm}(x_t|\lambda_{m1-1}),\ p_{norm}(x_t|\lambda_{m1+1}),\ \cdots,\ p_{norm}(x_t|\lambda_N)\} \quad (8)$$

여기서 유사도 차이는 다음과 같이 나타난다.

$$R^t = r^t_{m1} - r^t_{m2} \quad (9)$$

이렇게 얻어진 유사도 차이를 사전실험을 통하여 얻어진 문턱치와 비교하여 식 (5)의 점수를 계산에 사용여부를 판단한다.

3.2. 복합방법 (Hybrid Method)

복합방법은 3.1절에서의 FS 방법과 기존 방법인 WMR 방법을 혼합한 방법으로 프레임들 중에서 화자정보를 많이 포함하고 있는 프레임을 선택할 수 있는 FS방법의 장점과 선택된 프레임을 이용하여 화자 식별시 화자들의 유사도 값 대신 지수함수 가중치를 이용함으로써 화자들 간의 변별력을 더 높일 수 있다는 장점을 혼합한 방법이다. 복합방법을 간략히 설명하면 다음과 같다.

먼저 식 (9)로부터 얻어진 유사도 차이 R^t를 문턱치와 비교한 후 이 문턱치를 넘어서는 프레임만을 선택하게 된다.

$$Threshold < r^t_{m1} - r^t_{m2} \quad (10)$$

선택된 프레임들로부터 유사도 $p(x_t|\lambda_N)$를 계산하여 표 1에서와 같이 각 유사도에 해당하는 가중치 $w_t(r_{\lambda_N})$를 식 (7)의 점수 계산에 이용한다.

이때 가중치를 결정하는 식 (6)의 A, B 값은 등록화자 35명을 기준으로 하였을 경우 가장 높은 식별률을 나타낼 때를 기준으로 등록 화자수에 따라 re-scaling 하여 이용하였다.

3.3. 수정된 가중모델순위 (Modified Weighting Model Rank; MWMR)방법

WMR 방법은 프레임 단위에서 화자모델들 사이의 변별력을 크게 할 수는 있지만, 각 프레임의 유사도에 의한 변별력은 무시되었다. 따라서 화자의 성도 특성을 잘 표현하지 못한 프레임의 경우에도 유사도의 순위만 높다면 큰 가중치를 부여함으로써 결과적으로 화자의 변별력을 감소시킬 수 있다. 또한 i_{th} 프레임의 유사도 순위가 $(i-1)_{th}$ 프레임의 유사도 순위와 상대적으로 동일한 위

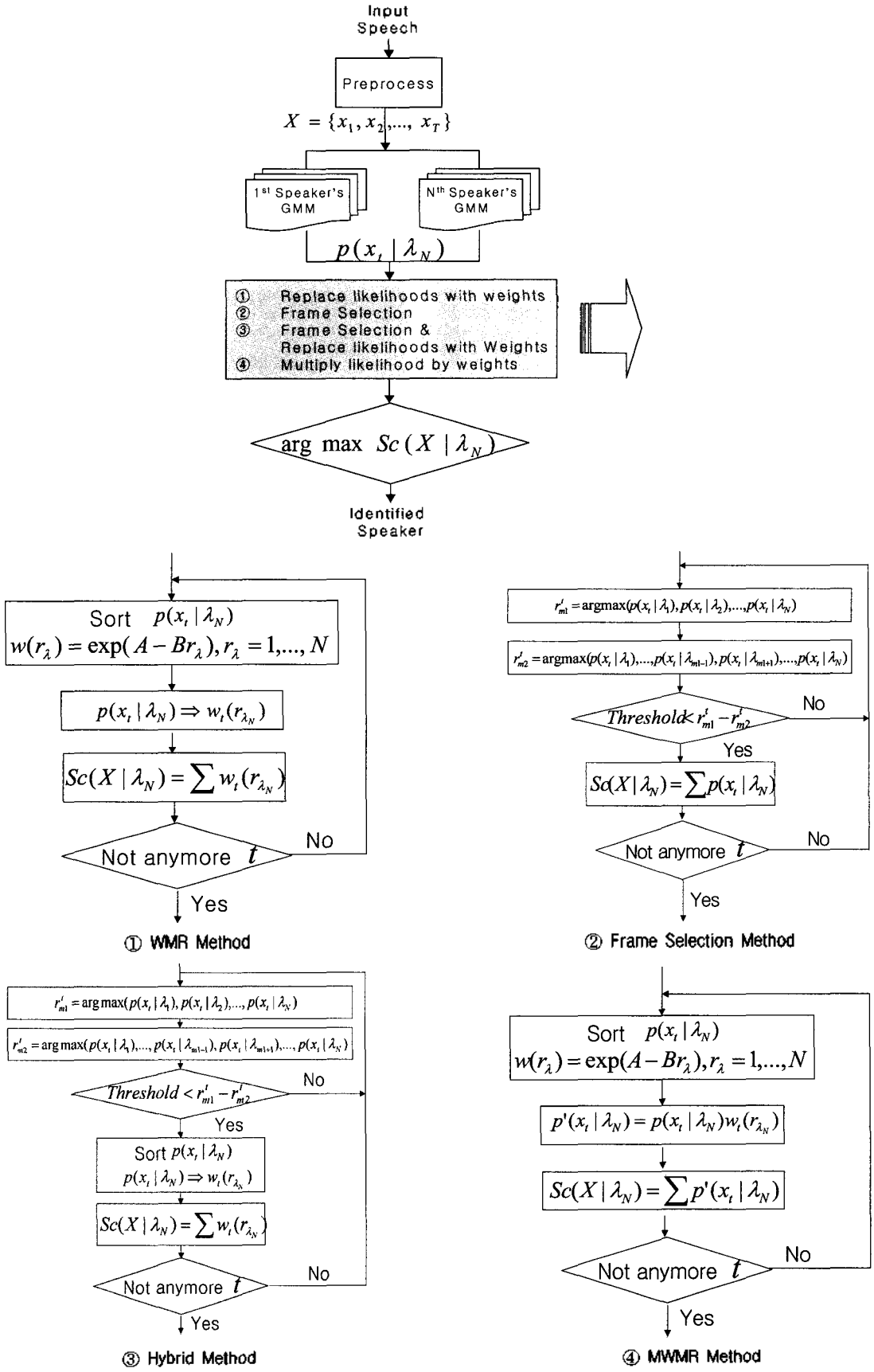

그림 1. 식별방법에 따른 식별 시스템 구성
Fig. 1. System configuration for each method.

치를 차지할 경우 $(i-1)_{th}$ 프레임과 동일한 가중치를 갖게 된다. 그러므로 전체 프레임에서 순위의 합계가 동일한 두 화자모델이 존재할 경우 동일한 가중치 합계가 나온다는 단점이 있다. 만일 가중치를 결정하는데 있어서 프레임 유사도의 크기까지 고려할 수 있다면 화자모델들 사이의 변별력을 좀 더 크게 할 수 있을 것으로 기대된다. 따라서 여기서는 프레임 단위 유사도의 상대적 위치에 따라 결정된 가중치와 프레임단위 유사도를 곱한 값을 식별화자를 결정하는 점수 계산에 이용하기로 한다. 이 방법을 수정된 가중모델순위 (MWMR) 방법이라 칭한다.

즉 식 (6)의 가중치 $w_i(r_{\lambda_N})$을 다음과 같이 수정하여 유사도 값 $p'(x_i|\lambda_N)$을 구한다.

$$p'(x_i|\lambda_N) = p(x_i|\lambda_N) w_i(r_{\lambda_N}) \tag{11}$$

IV. 식별실험

4.1. 식별시스템

그림 1에 화자식별 시스템을 종류별로 나타내었다. 식별방법에 따라 ①기존의 WMR방법, ②는 프레임 선택 방법, ③은 복합방법, ④는 MWMR방법을 나타낸 것이다.

본 논문에서 제안한 방법들의 유효성을 확인하기 위해 본 논문에서 언급한 ML, WMR방법과 제시한 복합, FS, MWMR 방법을 적용하여 식별실험을 수행하였다.

4.2. 음성 데이터 및 분석

식별실험에서 GMM의 혼합수는 식별률 및 계산량을 고려하여 16으로 고정하였으며, 특징 파라미터는 켑스트럼 계수 10차와 회귀계수 10차만을 사용하였다. 음성 특징 파라미터의 분석조건을 표 2에 나타내었다.

식별실험을 위한 데이터베이스는 단일 데이터베이스와 혼합 데이터베이스의 두 종류로 구성하여 이용하였다.

표 2. 전처리를 위한 분석조건
Table 2. Analysis condition for pre-processing.

Sampling Rate	16 kHz
Pre-emphasis coefficient	0.98
Hamming Windows	yes
Frame length	256 points
Frame Shift	120 points
Cepstrum vector dimension	10

단일 데이터베이스는 남성화자들로만 이루어진 데이터베이스로서 채집기간과 발성내용이 서로 다른 음성 데이터로 구성되어 있으며, 혼합 데이터베이스는 각각 다른 종류의 단일 데이터베이스들을 통합한 것으로서 남녀화자가 포함된 데이터베이스이다.

화자모델 학습과 평가를 위한 음성 데이터는 각 데이터베이스로부터 무작위로 추출한 단어를 사용하였다. 평가화자는 단일 데이터베이스에서는 남성 35명과 남성 131명의 화자를 추출하여 이용하였고, 복합 데이터베이스로부터는 남성 및 여성이 혼합된 102명과 316명의 화자를 추출하여 이용하였다. 실험에서는 화자 모델 학습을 위한 프레임 수 (1 프레임 =256 points, 16 ms)는 학습을 위한 단어집합의 최소 길이로 생각되는 4,000 프레임 및 최대 길이로 생각되는 10,000 프레임으로 하였고, 실제 식별실험을 위한 프레임 수는 실제의 화자식별환경을 고려하여 350 프레임으로 설정하였다.

4.3. 실험결과 및 고찰

각 화자식별방법에 대해 식별실험을 수행한 결과를 표 3에 나타내었다. 표 3으로부터 본 논문에서 채택한 학습용 프레임 수에 따라서는 10,000 프레임이 4,000 프레임보다는 모든 식별실험에서 높은 식별률을 나타내었다. 또한 기존의 화자식별방법 중 WMR 방법이 ML 방법보다 평균 2% 이상 향상된 식별률을 보여 보다 효과적인 방법임을 알 수 있었다.

기존의 방법들과 제안한 방법들과의 식별률 비교에서

표 3. 테스트 화자수와 및 학습 프레임 수에 따른 각 식별방법들의 식별률 (350 테스트 프레임)
Table 3. Identification rates of each method according to number of speakers for test and number of frames for training (350 Test Frame).

테스트데이터	데이터종류	학습 프레임수	ML	프레임선택	WMR	복합	MWMR
35 명	단일데이터 소수화자	f4k	91.43	94.29	94.29	94.29	94.29
		f10k	100	100	97.15	100	100
131 명	단일데이터 다수화자	f4k	87.79	87.79	93.13	93.13	100
		f10k	89.31	89.31	98.47	98.47	98.47
102 명	복합데이터 소수화자	f4k	85.29	85.29	94.12	95.1	95.1
		f10k	97.06	98.03	99.02	100	100
316명	복합데이터 다수화자	f4k	87.66	87.66	89.97	90.82	90.82
		f10k	93.35	93.35	97.78	97.78	98.1

는 FS방법의 경우, 기존의 WMR방법과 거의 동등한 결과를 나타내었으며 복합방법의 경우에는 약 2%의 향상된 식별률을 얻어 FS방법을 병용할 경우 더욱 효과적임을 확인할 수 있었다. 또 MWMR 방법은 기존의 WMR 방법보다 약 3%의 향상된 식별성능을 보여 이 방법의 유효성을 확인할 수 있었다. 특히 MWMR 방법의 경우, 316명의 화자를 대상으로 한 화자식별실험 결과, 학습 프레임길이를 10,000 프레임으로 할 경우, 평균 98.1%의 식별률을 얻어 다른 방법들에 비해 상대적으로 높은 식별성능을 보였다.

위의 결과로부터 MWMR방법이 화자 수에 관계없이 가장 우수한 화자식별성능을 나타내었으므로 MWMR방법에 대하여 학습 프레임의 변화에 따른 식별률의 변화를 실험하였다. 이 결과를 다른 식별방법들과 함께 표 4에 나타내었다. 표 4로부터 알 수 있는 바와 같이, 학습 프레임 수의 변화에 관계없이 본 논문에서 제안한 MWMR 방법이 전체적으로 가장 높은 식별 성능을 나타내어 MWMR방법이 학습프레임의 변화에 대해서도 강건함을 알 수 있었고 학습 프레임수가 많을수록 향상된 식별률을 나타냄을 알 수 있었다.

테스트 프레임 수의 변화에 따른 식별률의 변화와 식별시간을 알아보기 위하여 테스트 프레임 수를 350 프레임으로부터 1000 프레임까지 변경해 가면서 식별실험을 수행하였다. 이때 학습 프레임 수는 비교적 양호한 식별결과를 나타내는 4,000 프레임을 하한으로 설정하였으며, 사용된 데이터는 보다 실제적인 환경을 고려하여 채집기간과 발성단어, 성별이 모두 다른 복합 데이터만을 사용하였다. 실험을 위한 환경은 750 MHz CPU, 256 메모리 공간, Windows XP 이며 식별방법으로는 가장 우수한 식별성능을 보이는 MWMR 방법을 이용하였다.

표 5와 그림 2에 프레임 수 변화에 따른 식별률을, 그림 3에 식별시간을 측정한 결과를 나타내었다. 표 5와 그림 2로부터 화자 수에 관계없이 테스트 프레임수가 증가함에 따라 식별률이 증가함을 확인할 수 있으며, 314명 화자

표 4. 식별방법, 화자수, 학습 프레임수에 따른 식별률 (350 테스트 프레임)
Table 4. Identification rates according to the number of frames for training, the number of speakers, and identification methods (350 test frame).

화자수	식별방법	4k	5k	6k	7k	8k	9k	10k
35 (명)	ML	91.43	94.28	97.14	97.14	97.14	94.28	100
	WMR	94.29	94.29	94.29	94.29	97.14	97.14	97.14
	MWMR	94.29	94.29	94.29	94.29	97.14	97.14	100
131 (명)	ML	87.79	84.73	83.96	84.73	85.45	87.79	89.31
	WMR	93.13	94.66	93.13	96.18	96.18	96.95	98.47
	MWMR	93.89	95.41	93.89	96.95	96.65	96.18	98.47
102 (명)	ML	85.29	89.21	89.21	94.12	96.08	93.14	97.06
	WMR	94.12	96.08	96.08	97.06	98.04	98.04	99.02
	MWMR	95.1	96.08	96.08	97.06	98.04	98.04	100
316 (명)	ML	87.66	88.61	89.24	90.82	91.77	90.82	93.35
	WMR	89.87	92.72	93.35	93.35	95.57	95.87	97.78
	MWMR	90.18	93.99	93.99	94.3	96.52	96.2	98.1

표 5. 수정된 가중모델순위 방법의 테스트 프레임 수에 따른 학습프레임의 식별률
Table 5. Identification rates for models having different training frames vs. models having different test frames in modified weighting model rank method.

no. of speakers	no. of test frames	f4k	f5k	f6k	f7k	f8k	f9k	f10k
102	f350	95.09	96.07	96.07	97.05	98.03	98.03	100
	f600	98.03	97.05	99.01	98.03	100	100	100
314	f350	94.90	95.85	95.85	96.81	97.77	98.08	97.77
	f400	95.54	95.85	97.13	97.45	98.72	98.40	98.72
	f500	97.45	96.49	98.09	98.09	98.09	98.40	98.72
	f600	98.40	96.81	97.45	98.40	99.36	98.72	99.04
	f700	97.13	97.77	98.72	98.72	98.72	98.72	98.72
	f800	97.77	98.40	98.72	98.72	98.72	99.04	99.68
	f900	98.72	98.72	99.36	99.36	99.36	99.36	99.68
	f1000	98.40	99.36	99.36	99.36	99.36	99.36	99.68

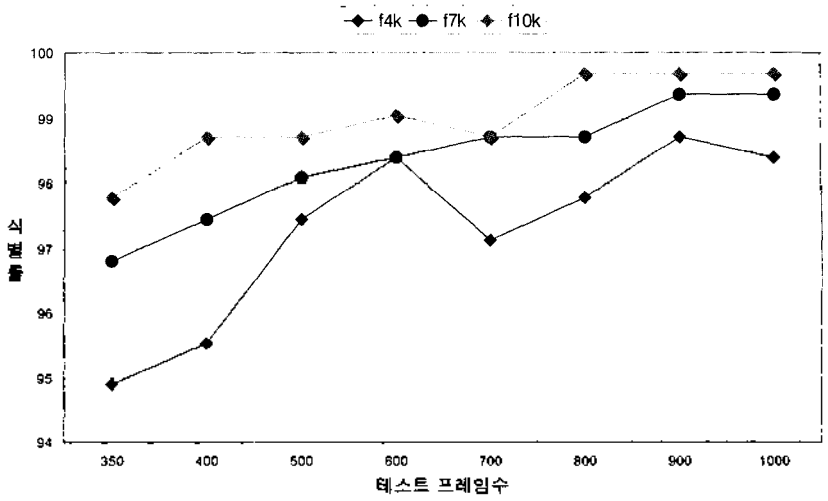

그림 2. 수정된 가중모델순위 방법의 테스트 프레임 수에 따른 학습프레임 (f4k, f7k, f10k)의 식별률
Fig. 2. Identification rates for models having different training frames (f4k, f7k, f10k) vs models having different test frames in modified weighting model rank method.

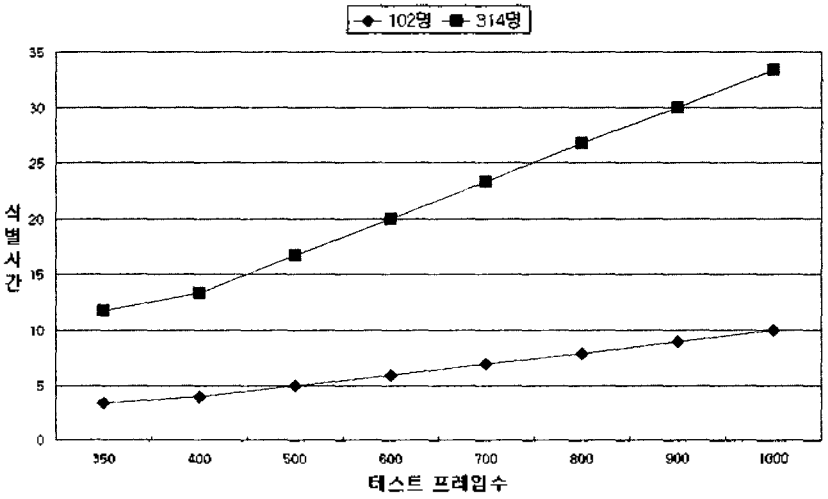

그림 3. 테스트 프레임 수에 따른 식별시간
Fig. 3. Identification time according to test frame length.

를 대상으로 한 경우 학습용으로 10000 프레임을, 시험용으로 800 프레임 이상을 이용한 경우가 가장 높은 99.68%의 화자 식별률을 나타내었다.

참고로 테스트 프레임수의 증가에 따른 식별시간의 변화를 조사하여 그림 3에 나타내었다. 그림 3에서 확인할 수 있는 바와 같이 식별 대상 화자 수 및 테스트 프레임 수의 증가에 따라 선형적으로 증가함을 볼 수 있어 실용화를 위해서는 이에 대한 연구가 필요할 것으로 생각된다.

V. 결 론

본 논문에서는 강건한 화자식별을 위하여 FS 방법, 복합방법, MWMR방법을 제안하고 제안한 방법의 유효성을 확인하기 위하여 기존의 화자식별 방법들과 식별성능을 비교하였다. 기존의 ML 방법은 모든 입력 프레임으로부터 유사도를 계산하므로 화자의 성도 특성이 잘 표현되지 못한 프레임의 유사도까지도 식별화자를 결정하는 점수 계산에 사용하는 문제점이 있었다. 이러한 문제점을 해결하기 위해서 FS 방법에서는 각 프레임에서 상위 두 유사도의 차를 이용하여 화자의 특성이 잘 표현되고 있는 프레임만을 선택한 후, 선택된 프레임을 식별화자를 결정하는 점수 계산에 이용한다. 복합방법은 FS 방법과 WMR 방법의 장점만을 이용하는 방법이다. 즉 FS 방법을 이용하여 입력 프레임에서 화자의 성도특성이 잘 표현된 프레임만을 선별한 후, 이들 프레임에서 계산된 유사도 대신에 프레임 유사도의 상대적 위치에 따른 가중치를 이용하여 전체 점수를 계산하는 방법이다. 이렇게 함으

로써 각 프레임에서 낮은 프레임 유사도 값을 가지는 화자는 더욱 낮은 값을 부여하고 높은 프레임 유사도를 가지는 화자는 더욱 높은 값을 부여하여 프레임 단위에서 화자의 변별력을 더욱 크게 하는 방법이다. MWMR 방법은 각 화자의 프레임 유사도와 지수함수를 이용한 가중치를 곱한 값을 이용하여 전체 점수를 계산하는 방법으로 기존의 WMR 방법이 각 화자에 대한 프레임 유사도의 순위에 따라 지수함수 가중치로 대치시키는 방법을 사용하고 있으므로, 유사도 본래의 변별력이 전체 계산에서 고려되지 않는 문제점이 있다. MWMR방법은 이를 해결하기 위해 유사도와 가중치를 함께 이용하는 방법이다.

제안방법들의 유효성을 확인하기 위해 남녀 화자수가 각각 다른 5종류의 음성 데이터베이스에 대해 화자식별 실험을 수행한 결과, 제안한 FS방법의 경우 기존의 ML방법보다 약 2%의 향상된 성능을 보였으며, 복합방법의 경우 WMR방법보다 약 2% 향상된 결과를 나타내어 FS방법을 다른 방법들과 함께 사용할 경우 더욱 효과적임을 확인하였다. 제안한 MWMR 방법의 경우 WMR 방법보다 약 3%의 향상된 식별결과를 보여 제안한 방법들 중 가장 우수한 식별 성능을 보여 전체적으로 제안한 방법들의 유효성을 확인하였다. 특히 MWMR방법을 이용한 경우 314명의 화자를 식별대상으로 하였을 경우, 학습 프레임 수가 10,000, 테스트 프레임수가 800일 때 99.68%의 화자 식별률을 얻어 제안방법의 유효성을 확인할 수 있었다.

그러나 실용화 가능한 시스템 구현을 위해서는 화자식별을 위한 학습 프레임 길이의 감소, 테스트 프레임 수의 증가에 따른 계산량 감소에 대한 연구가 필요하다. 이에 대해서는 향후 과제로 하기로 한다.

감사의 글

본 연구의 진행에 많은 도움을 주신 Toyohashi Univ. 의 Nakagawa 교수님께 감사드립니다.

참 고 문 헌

1. 정현열, "음성을 이용한 화자인식 기술의 현황과 전망," 정보과학회지, **19** (7), 32-44, 2001.
2. S. Furui, F. Itakura, and S. Saito, "Talker recognition by longtime averaged speech spectrum," *Trans. IECE*, **55-A** (1), 549-556, 1972.
3. A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent models," *Computer Speech and Language*, 2, 143-157, 1987.
4. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on SAP*, **3** (1), 72-83, 1995.
5. S. Furui, "An overview of speaker recognition technology," *in Acoustic Speech and Speaker Recognition* (C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), Ch. 2, 31-56, Kluwer Acad. Pub., 1996.
6. K. Markov and S. Nakagawa, "Text-Independent speaker identification on TIMIT database," *Proceedings of Acoustical Society of Japan*, 83-84, March 1995.
7. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Inc., 1990.
8. D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, **17** (1-2), 91-108, 1995.

저자 약력

● 김 민 정 (Min-Jung Kim)

1999년: 영남대학교 일반대학원 멀티미디어 통신공학과 (공학석사)
1999년~현재: 영남대학교 일반 대학원 정보통신공학과 (박사수료)
※ 주관심분야: 디지털신호처리, 음성처리, 음성인식, 화자 인식

● 오 세 진 (Se-Jin Oh)

1996년 2월: 영남대학교 전자공학과 (공학사)
1998년 2월: 영남대학교 대학원 전자공학과 (공학석사)
2002년 2월: 영남대학교 대학원 전자공학과 (공학박사)
2001년 9월~현재: 대구과학대학 디지털정보통신계열 전임강사
※ 주관심분야: 음성분석 및 인식, 언어처리

● 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
※ 주관심분야음성인식, 화자인식, 음성합성 및 DSP 응용분야