ICA 기반의 특징변환을 이용한 화자적응

박만수(ICU), 김회린(ICU)

<차 례>

<Abstract>

Speaker Adaptation using ICA-based Feature Transformation

Mansoo Park, Hoi-Rin Kim

The speaker adaptation technique is generally used to reduce the speaker difference in speech recognition. In this work, we focus on the features fitted to a linear regression-based speaker adaptation. These are obtained by feature transformation based on independent component analysis (ICA), and the transformation matrix is learned from a speaker independent training data. When the amount of data is small, however, it is necessary to adjust the ICA-based transformation matrix estimated from a new speaker utterance. To cope with this problem, we propose a smoothing method through a linear interpolation between the speaker-independent (SI) feature transformation matrix and the speaker-dependent (SD) feature transformation matrix. We observed that the proposed technique is effective to adaptation performance.

* 주제어: ICA, 특징변환, 화자적응

1. 서 론

음성인식 시스템에서 화자 또는 환경의 변이성 문제를 해결하기 위하여 적응기술이 유용하게 사용된다. 일반적으로 화자의 변이성 문제를 해결하기 위해서 MAP(Maximum A Posteriori) 방식과 MLLR(Maximum Likelihood Linear Regression)[1,2,3] 방식을 화자 적응에 널리 이용하고 있다. 이러한 방법들은 모델 파라미터를 테스트 화자의 특성에 유사하게 변이시키는 작용을 한다. 특히, HMM(Hidden Markov Model) state에서의 Gaussian 평균값들에 대해 적용을 하게 된다. 일반적으로 MAP 방식은 적응 데이터가 충분히 많을 경우에 사용되며 MLLR 방식은 적응 데이터가 적은 경우에 사용된다. 본 논문에서는 적응시간을 고려하여 MLLR 방식을 기반으로 하였다. Leggetter과 Woodland에 의해 제안된 MLLR[4] 방식은 HMM 파라미터들의 선형변환을 기반으로 한다.

본 논문은 ICA(Independent Component Analysis)[5] 기반의 특징변환 기술이 linear regression을 이용하는 화자적응 기술인 MLLR에 좀 더 적합한 특징을 추출하는데 유용함을 기술한다. 즉, ICA 기반의 특징변환 기술은 MLLR에서 화자 적응을 위한 모델 변환 regression matrix를 추정하는데 강인함을 보인다.

본 논문은 다음과 같은 개요로 기술된다. 2장에서는 ICA를 위한 infomax learning rule의 간략한 소개와 ICA를 이용한 특징변환에 대해 기술한다. 3장에서는 화자적응에 적합한 특징 변환 방법에 대해 기술하고 4장에서는 실험 결과에 대한 소개를 한다. 마지막으로 5장에서는 실험 결과를 바탕으로 제안된 방법의 실효성 및 추후의 연구 방향을 제시한다.

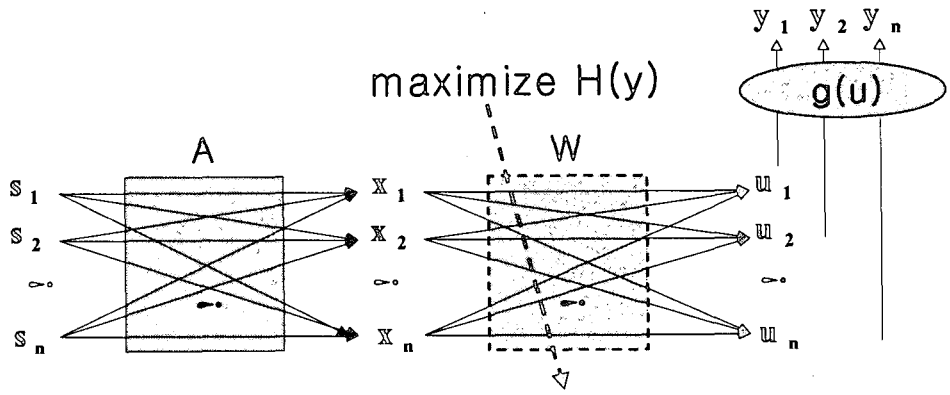
2. ICA 기반의 특징변환

실제적으로 ICA는 BSS(Blind Souce Separation)[7] 문제를 해결하는 방법 중 하나로 독립적인 소스 신호들이 선형적으로 혼합되어 있는 경우에 각 독립적인 신호들을 복원해내는 기술이다. 여기에서 blind의 의미는 소스 신호들과 소스 신호들이 서로 혼합되는 channel을 알 수 없다는 것이다. 하지만 ICA를 수행함으로써 관측된 신호들을 통계적으로 서로 독립적이 되도록 선형변환을 수행하여 소스 신호들을 복원할 수 있다.

2.1. Infomax 알고리즘

Infomax 알고리즘[8]은 독립적인 소스들을 분리하기 위하여 정보량의 최대화를 이용하는 learning 알고리즘이다. 이 알고리즘은 출력신호들의 joint entropy $H(\mathbf{y})$의

최대화는 출력 신호들의 mutual information의 최소화하는 것과 동일함을 보인다[8]. 그림 1은 infomax learning rule의 전반적인 과정을 나타낸다. 여기에서 소스 신호를 나타내는 s_i는 서로 독립적이며 소스 신호와 서로 혼합되는 채널을 나타내는 matrix A는 blind이다. 궁극적으로 관측된 신호 x_i로부터 독립적인 신호 s_i를 복원하기 위한 deconvolution 채널 matrix W가 matrix A의 inverse가 될 수 있도록 infomax 알고리즘을 적용하는 것이다.



<그림 1> ICA를 위한 infomax learning rule

출력신호들의 joint entropy는 식 (1)과 같이 나타낼 수 있다.

$$H(y_1,\ldots,y_N) = H(y_1) + \cdots + H(y_N) - I(y_1,\ldots,y_N) \tag{1}$$

여기에서 $H(y_i)$는 출력신호의 marginal entropy를 나타내고 $I(y_1,\ldots,y_N)$는 출력신호들의 mutual information을 나타낸다. Joint entropy $H(y_1,\ldots,y_N)$를 최대화함으로써 marginal entropy들을 최대화할 수 있고 mutual information 값을 최소화할 수 있다[8]. 이 알고리즘의 learning rule[8]은 식 (2)와 같다.

$$\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = (\mathbf{W}^{\mathrm{T}})^{-1} + \left(\frac{\partial p(\mathbf{u})/\partial \mathbf{u}}{p(\mathbf{u})}\right)\mathbf{x}^{\mathrm{T}} \tag{2}$$

joint entropy를 최대화 하기위한 좀 더 효율적인 방법으로써 'natural' gradient방식이 사용된다. 식 (3)과 같이 $\mathbf{W}^{\mathrm{T}}\mathbf{W}$를 식 (2)의 앞단에 곱해줌으로써 natural gradient 형식으로 변환할 수 있다[8].

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^{\mathrm{T}} \mathbf{W} = \left[\mathbf{I} + \left(\frac{\partial p(\mathbf{u}) / \partial \mathbf{u}}{p(\mathbf{u})} \right) \mathbf{u}^{\mathrm{T}} \right] \mathbf{W} \tag{3}$$

$\mathbf{I}$ 는 identity matrix를 나타내고 비선형성 맵핑 함수로써 식 (4)와 같이 정의한다.

$$\varphi(\mathbf{u}) = -\frac{\partial p(\mathbf{u}) / \partial \mathbf{u}}{p(\mathbf{u})} \tag{4}$$

따라서 식 (3)을 식 (5)와 같이 나타낼 수 있다.

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - \varphi(\mathbf{u}) \mathbf{u}^{\mathrm{T}} \right] \mathbf{W} \tag{5}$$

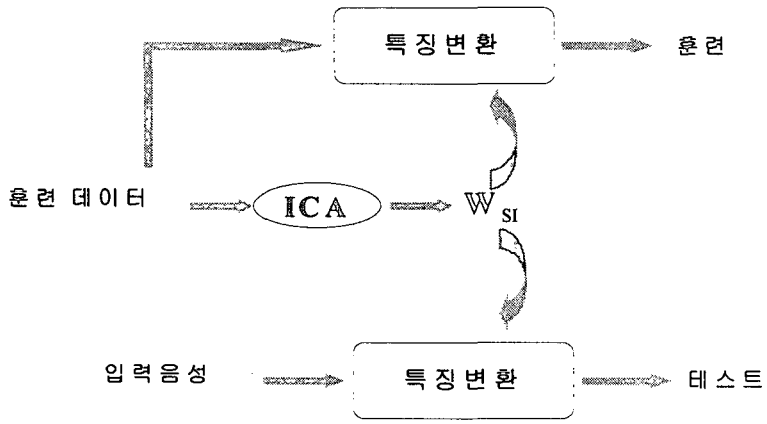
여기에서 $p_i(u_i)$ 이 소스들의 로그 분포의 편미분 값으로 나타낼 수 있다면 이 알고리즘의 수렴은 보장하게 된다[8]. 이러한 이유로 비선형 함수 $\varphi(\mathbf{u})$ 를 $2\tanh(\mathbf{u})$ 로 정의한다면 learning rule은 식 (6)과 같이 나타낼 수 있다.

$$\Delta \mathbf{W} \propto \left[\mathbf{I} - 2\tanh(\mathbf{u}) \mathbf{u}^{\mathrm{T}} \right] \mathbf{W} \tag{6}$$

2.2. ICA 기반의 특징변환

ICA는 가정에서와 같이 homomorphic 시스템에서만 적용이 가능하다[5,6]. 음성 특징으로 사용되는 켑스트럼 벡터는 로그 스펙트럼 영역에서 glottal pulse, vocal tract, mouth radiation, 그리고 전송 채널 왜곡 함수들의 선형적인 합으로 나타 낼 수 있다[5]. 각 화자의 켑스트럼 성분들은 이 필터들의 특성들이 서로 혼합되는 결과를 초래하기 때문에 화자 적응 문제를 더욱 더 어렵게 만든다. 켑스트럼 벡터에 ICA를 적용함으로써 각 필터들의 특성을 서로 독립적으로 풀 수 있기 때문에 위의 문제를 간단히 해결할 수 있는 것과 동시에 노이즈 특성과 같은 불필요한 변이성을 제거할 수 있게 된다. 켑스트럼은 orthogonal 하기 때문에 독립적인 projection에서 상대적으로 효율적이다. $\mathbf{C}' = \mathbf{W} \cdot \mathbf{C}$는 위의 내용을 수식으로 나타낸 것이다. 여기에서 $\mathbf{C}'$는 projection한 특징벡터를 나타내고 $\mathbf{C}$는 본래의 특징벡터를 나타낸다. 그리고 $\mathbf{W}$는 ICA 기반의 특징변환 matrix이다. 그림 2는 ICA를 이용한 기본적인 특징변환 과정을 나타낸다. 이것은 화자 독립 인식시스템에서 훈련데이

터로부터 추정한 화자 독립 특징변환 matrix W_{SI}를 이용한 것이다.



<그림 2> 기본적인 ICA 기반의 특징변환 과정

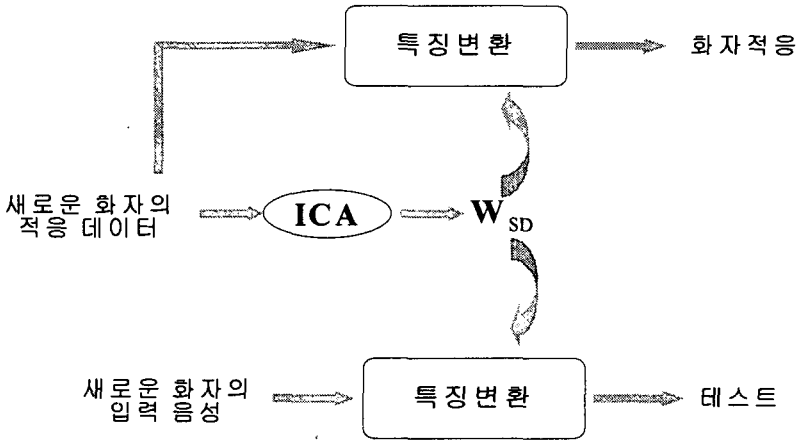
3. 화자적응에 적합한 특징변환

켑스트럼 영역에 ICA를 적용함으로써 독립적인 특징 벡터를 구할 수 있다. 일반적으로 독립적인 특징벡터는 패턴 분류에 있어서 효과적이다. 또한, 특징벡터 성분들 간의 종속적인 영향 없이 환경적 요인들을 독립적으로 표현함으로써 화자나 잡음의 적응과 같은 특정 목적에 좀 더 효과적으로 이용될 수 있다. 즉, full regression matrix를 사용하는 MLLR 방법에서 불충분한 화자적응 데이터 때문에 성분들 간의 상호 관계를 나타내는 off-diagonal 성분들이 불안정하게 추정되는 문제를 해결할 수 있다. 특히, 특징벡터의 성분들 간의 독립적인 특성 때문에 diagonal regression matrix를 사용하는 MLLR 방법에서는 매우 효과적이다. full regression matrix를 사용하는 MLLR과 비교해서 diagonal regression matrix를 사용하는 MLLR 방법은 일반적인 특징벡터를 사용할 경우에 약간의 낮은 성능을 보이지만 계산 량을 줄일 수 있는 장점이 있다. 제안된 ICA 기반의 독립적인 특징 벡터는 diagonal regression의 가정을 만족할 수 있고 적응 데이터의 적은 양을 요구하는 온라인 적응 시스템에 적합할 것이다.

그림 3에서처럼 화자 적응을 위해서 새로운 화자의 적응데이터로부터 추정한 특징 변환 matrix W_{SD}가 필요하다. 그러나 적은 적응 데이터로부터 얻어진 W_{SD}는 특정 환경에 편중될 수 있어서 정보의 손실이 발생하기 때문에 항상 신뢰할 수는 없다. 이 문제를 해결하기 위해서 적응과 인식 과정에서 필요한 특징변환 matrix를 조정할 필요가 있다. 본 논문에서는 적응과 인식 과정에서 필요한 특징변환 matrix 조정 방법으로써 식 (7)과 같이 W_{SI}과 W_{SD}의 선형보간 과정을 통해 smoothing 할

것을 제안한다. 선형보간 과정을 거침으로써 주요 정보의 손실을 막을 수 있다.

(7) $$\mathbf{W}_{smooth} = (1-\alpha)\mathbf{W}_{SI} + \alpha\mathbf{W}_{SD}, \qquad 0 \le \alpha \le 1$$



<그림 3> 화자 종속 특징변환 과정

4. 실험 및 결과

4.1. 실험 데이터

실험을 위해 70명(남자: 38명, 여자: 32명)의 화자가 어휘 내용이 다른 452 균일 음소 분포 단어(Phonetically Balanced Words, PBW)를 2회씩 발성한 데이터베이스를 사용하였다. 음성신호는 16kHz로 샘플링 되어있고 16bit로 양자화 되어있다. 본 실험을 위해 통신 채널 특성에 맞게 8kHz로 다운샘플링하여 사용하였다. 훈련을 위해 화자 63명 분량의 DB를 사용하였고 나머지 7명(남자: 4명, 여자: 3명) 분량을 적응데이터와 테스트를 위하여 사용하였다. 7명의 화자의 1회 발성 분량을 적응 데이터로 나머지를 테스트를 위하여 사용하였다. 화자 독립 특징변환 matrix $\mathbf{W}_{SI}$는 훈련 데이터로부터 각 화자마다 100 단어씩 총 6,300 단어를 발췌하여 추정하였다. 화자 종속 특징변환 matrix $\mathbf{W}_{SD}$를 추정하기 위해 각 테스트 화자의 적응 데이터를 사용하였다.

음성신호는 8ms 단위의 프레임 마다 총 39차 특징벡터로 표현하였다. 기반 시스템에서의 39차 특징벡터는 12차 MFCC(Mel Frequency Cepstral Coefficients)와 로그 에너지, delta 및 delta-delta로 구성되어있다. 녹음환경의 영향을 줄이기 위하여

CMN(Cepstral Mean Normalization)을 수행하였다. 화자 적응을 위한 특징벡터는 12차 MFCC-CMN로부터 ICA를 수행하여 얻은 12차 ICA 기반의 특징에 로그에너지와 delta 및 delta-delta의 특징을 더한 총 39차 특징을 사용하였다. 본 실험에서의 모델은 triphone 단위의 1 Gaussian mixture를 사용하여 3 state의 left-to-right 방식의 연속 밀도 HMM(Hidden Markove Model) 기반으로 하였다.

4.2. Matched case의 화자적응: Clean Speech

여기에서는 노이즈가 없는 음성에 대해 3가지 다른 특징변환 matrix에 따른 화자적응의 성능을 비교하기 위한 실험이다. 비교를 위해 39차 MFCC 특징을 사용한 화자적응 성능을 기반으로 하였다. 훈련과정에서는 $\mathbb{W}_{SD}$를 사용하여 특징변환을 수행하였고 적응과 테스트를 위하여 3가지($\mathbb{W}_{SD}$, $\mathbb{W}_{SI}$, $\mathbb{W}_{smooth}$) 다른 특징변환 matrix를 적용하여 화자 적응 성능을 비교하였다.

<표 1> Matched case에서 서로 다른 특징변환 matrix를 적용한 경우의 ERR

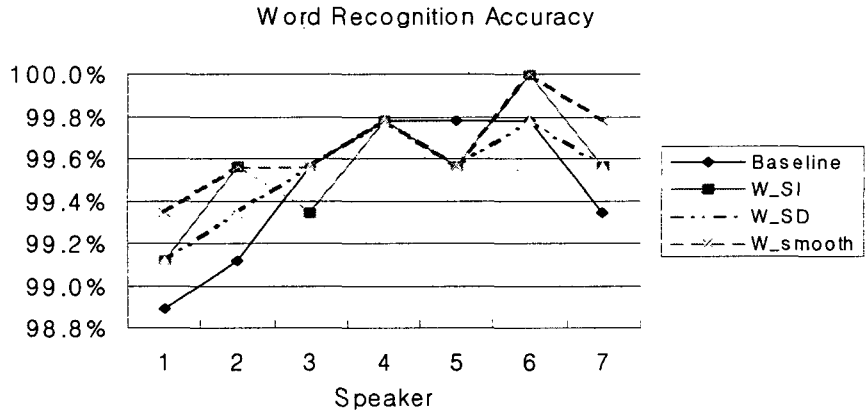
특징 변환 matrix	평균 ERR
Baseline	-
$\mathbf{W}_{SI}$	17.9 %
$\mathbf{W}_{SD}$	12.0 %
$\mathbf{W}_{smooth}$	35.4 %

표 1에서는 smoothing 특징 변환 matrix가 가장 좋은 성능을 나타냄을 보인다. 성능 비교를 위하여 ERR(Error Reduction Rate)로 표현하였다. 여기에서 $\mathbb{W}_{smooth}$는 각 화자 마다 최대 성능을 나타내는 α를 적용하여 얻어진 matrix이다. 그림 4는 각 화자마다 화자적응 후의 인식 성능을 나타내고 있다. 그림 4에서 나타나듯이 제안된 방법이 대부분의 화자에 대해 비교적 나은 성능을 보이고 있다. 하지만 엄밀히 말해서 노이즈가 없는 음성을 가지고 matched case에 적용한 경우 인식 성능이 충분히 높기 때문에 절대적으로 나은 성능을 보인다고 말하기는 어렵다.

4.3. Mismatched case의 화자적응: Clean vs. Noisy Speech (SNR: 15dB)

이번 실험은 위의 경우와 유사하지만 적응 데이터와 테스트 데이터에 SNR 15dB의 AWGN(Additive White Gaussian Noise)를 첨가한 것만 다르다. 이 경우 훈련은 노이즈가 없는 음성으로 하였고 적응과 테스트 환경에는 SNR 15dB의 노이즈가

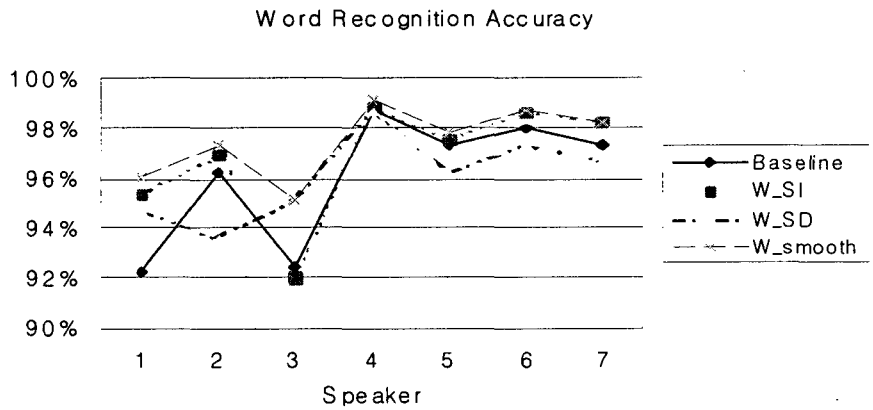
삽입된 mismatched case이다.



<그림 4> Matched case에서 각 화자에 따른 화자적응 결과

<표 2> Mismatched case에서 서로 다른 특징변환 matrix를 적용한 경우의 ERR

특징 변환 matrix	평균 ERR
Baseline	-
W_{SI}	19.1 %
W_{SD}	0.0 %
W_{smooth}	35.9 %



<그림 5> Mismatched case에서 각 화자에 따른 화자적응 결과

결과는 표 2에 나타나듯이 표 1과 유사하다. 여기에서도 α값은 각 화자마다 최고의 성능을 나타내는 값을 사용하였다. 그림 5는 모든 화자에 대해서 본 논문에서 제안한 방법이 더 효과적임을 나타내고 있다. 본 논문에서 제안한 화자적응을 위한 특징변환 방법은 matched case에서 대체적으로 나은 결과를 보였고 mismatched case에서는 더 효율적인 결과를 보였다.

5. 결 론

본 논문에서 화자적응에 좀 더 효과적인 특징을 얻기 위하여 ICA 기반의 특징변환 기술을 제안하였다. 화자독립 모델을 화자종속 모델로 변이시키는 화자적응을 위해서는 새로운 화자의 특징을 반영하는 것이 효과적일 것이다. 새로운 화자의 특징변환 matrix를 제한된 적응 데이터로 구하기 때문에 특정 환경에 편중되는 경우가 발생하여 정보 손실이 발생할 수 있다. 이러한 문제를 해결하기 위해 훈련과정에서 추정한 화자독립 특징변환 matrix와 적응데이터로부터 추정한 화자종속 특징변환 matrix 사이의 선형 보간을 수행하여 적응과 인식과정에서 사용할 특징변환 matrix를 추정하였다. 이로써 적은 양의 적응 데이터 때문에 발생 가능한 정보의 손실을 줄일 수 있다. 특히 mismatched case에서는 모든 화자에 대해 제안된 방법이 화자적응에서는 효과적이었다.

향후에는 다양한 노이즈 환경에서 제안된 방법의 효율성에 대해서 검증하고 화자마다 최고성능을 나타내는 α값의 결정을 위한 알고리즘의 연구를 유사도 값을 이용하여 보안할 것이다.

참 고 문 헌

[1] Reddy, Raj (2001), *Spoken Language Processing.*

[2] Young Steve (2000), *The HTK BOOK (for HTK Version 3.0).*

[3] Doh, Sam-Joo (2000), *Enhancements to Transformation-Based Speaker Adaptation: Principal Component and Inter-Class Maximum Likelihood Linear Regression,* Ph.D. Thesis, CMU.

[4] Leggetter, C. J. and P. C. Woodland (1995), Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models, *Computer Speech and Language* Vol.9, pp.171~185.

[5] Jang, Gil-Jin, Seong-Jin Yun, and Yung-Hwan Oh, (1999), Feature Vector Transformation using Independent Component Analysis and Its application to speaker identification, *Proc. of EUROSPEECH*, pp.767~770.

[6] Potamitis, L., G. Kokkinakis Fakotakis (2000), Independent component analysis applied to feature extraction for robust automatic speech recognition, *ELECTRONICS LETTERS* Vol.36. No.23, pp.1977~1978.

[7] Cardoso, J.-F. (1998), Blind signal separation: Statistical principles, *Proc. IEEE* Vol. 86. Oct., pp.2009~2025.

[8] Lee, Te-Won (1998), *Independent Component Analysis: Theory and Applications,* Boston, MA: Kluwer.

접수일자: 2002년 5월 7일
게재결정: 2002년 5월 24일

▶ 박만수(Mansoo Park)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학교 공학부
소속: 한국정보통신대학교 공학부 음성인식기술연구실
전화: 042) 866-6207
Fax: 042) 866-6245
E-mail: mansoo@icu.ac.kr

▶ 김회린(Hoi-Rin Kim)
주소: 305-732 대전광역시 유성구 화암동 58-4번지 한국정보통신대학교 공학부
소속: 한국정보통신대학교 공학부 음성인식기술연구실
전화: 042) 866-6139
Fax: 042) 866-6245
E-mail: hrkim@icu.ac.kr