국부 봉우리와 골에 의한 피치 검출과 퍼지를 이용한 화자 인식에 관한 연구

김연숙* · 김희주** · 김경재***

A Study on Speaker Recognition using the Peak and valley pitch detection and the Fuzzy

Yeoun-sook Kim* · Hee-joo Kim** · Kyoung-jae Kim ***

요 약

본 논문에서는 국부 봉우리와 골에 의한 피치 파라미터와 퍼지를 포함한 화자 인식 알고리즘을 제안한다. 음성의 패턴 인식에서 인식 성능을 저하시키는 시간 변동과 주파수 변동에 대한 문제를 해결하여 피치를 검출한다. 비선형적인 발성 시간에 따른 시간 변동의 폭을 모두 포함하기 위하여 음성 신호의 애매성을 보완할 수 있는 퍼지의 소속 함수를 이용하여 표준 패턴을 작성하고 퍼지 패턴 매칭을 이용하여 인식을 수행한다.

ABSTRACT

This paper proposes speaker recognition algorithm which includes the pitch parameter for the peak and valley. The time-frequency hybrid method for pitch extraction is valuable in that it can improve resolution in the time domain and accuracy in the frequency domain at the same time. It makes reference pattern using membership function and performs vocal track recognition of common character using fuzzy pattern matching in order to include time variation width for non-linear utterance for proposed method, speaker recognition experiments are carried out using vowels and number sounds.

키워드

Peak and valley, Formant, Pitch, Spectrum, Fuzzy, LPC

Ⅰ. 서 론

오늘날 세계 각국은 정보화 시대로 수많은 정보 교환을 위하여 인간과 기계 상호간의 정보통신에 대한 여러 가지 연구와 실험이 행하여져 오고 있다.

인간의 가장 간편한 의사 전달 방법은 음성을 이용하는 것으로, 세계 각국은 다른 나라의 언어를 알아듣기 위해 컴퓨터와 인간의 상호 대화를 위한 실용화 연구가 계속되어 오고 있다. 여러 외국어 가운데서도 일어는 우리 글의 문장 구조와 비슷한 문법 체계로 그 어떤 외국어보다도 익히고 사용하기가 쉽다.

한국어와 일어의 공통점은 모음과 자음이 있고, 어순이 비슷하며, 수사가 없고, 명사에 성의 구별이 없다. 차이점은 한국어는 자음을 받침으로 하여 끝날 수 있지만, 일어는 자음으로 끝나는 말이 없고 모두 모음으로 끝난다.

음성 인식과 화자 인식은 인간과 기계 상호 간

* 건국대학교
** 강원관광대학
*** 홍익대학교
접수일자 : 2003. 11. 5

의 의사 전달 방법이다.[1] 화자 인식에 대한 연구는 1963년 Bell 연구소의 Pruzansky가 화자 식별 실험을 하였고, 1981년 Furui가 텍스트 의존 화자 인식 실험을 하였다. 국내에서는 1991년 권석규가 DSP칩을 사용하여 하드웨어를 설계하였다.

음성은 화자의 성별, 나이, 발음 시 상태에 따라 변화하기 때문에 배경잡음이 없는 이상적인 음성 신호와의 경우와 한정된 대상의 발음의 경우에만 만족할만한 음성 인식 결과를 얻고 있는 실정이다.[2][3] 음성 인식의 실용성과 일반성에 대한 문제는 화자 인식에 사용되는 파라미터들을 통계적으로 적용할 수 있는 새로운 파라미터를 제안함으로써 해결될 것이다.[4][5]

본 논문에서는 비선형적인 발성 시간에 따른 시간 변동의 폭을 모두 포함할 수 있도록 퍼지의 소속 함수를 이용하여 표준 패턴을 작성하고 인식을 수행한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 음성 생성 시스템과 디지털 음성 신호 처리 과정을 설명하고, 3장에서는 제안한 피치 검출법과 퍼지 이론을 이용한 화자 인식을 살펴본다. 4장과 5장에서 본 논문에서 제안한 내용에 대한 실험을 평가하고 결론을 맺는다.

II. 음성 생성 시스템과 디지털 음성 신호 처리

1. 음성 생성 시스템

그림 1은 음성 생성 시스템의 모형도로 허파, 기관지, 호흡기관 등으로 구성된 하부-성문 시스템을 포함한다.

인두강과 구강을 합친 성도(vocal tract)는 후두의 출구에서 시작해서 입술에서 끝난다. 비강은 연구개로부터 콧구멍까지이다. 음성은 간단히 공기가 허파로부터 방출되고 결과적으로 성도에 있는 협착점에 의해 공기가 동요될 때 이 생성 시스템으로부터 방사되는 음향학적 파형이다.

피치는 강세, 억양 등의 요인에 따라 변한다. 강세는 음절이나 단어, 구에서 기본 주파수와 음의 강도를 변화시킨다.[6][7] 억양은 문법적 구조를 알리는 것으로 문장이나 절, 다른 경계에 대한 구별은 억양 패턴을 통해 이루어진다. 남자의 경우 피치 범위는 보통 50~250Hz 사이이고, 여성의 경우 120~500Hz 사이에 있다.

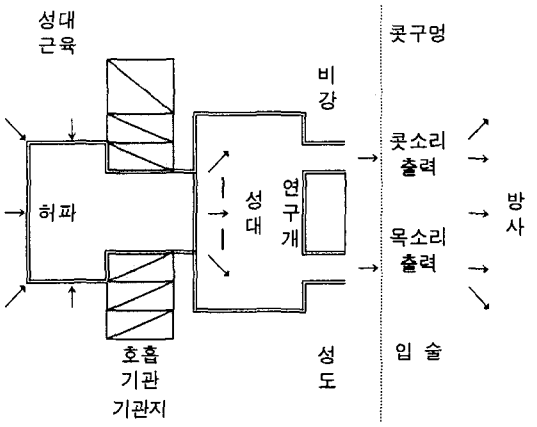


그림 1. 음성 생성 시스템의 모형도
Fig. 1 Schematized diagram of the vocal system

2. 디지털 음성 신호 처리

음성 신호를 표현하는 방법에는 에너지, 자기 상관 관계 함수, 평균 영 교차율 등이 있다.[5]

음성 신호의 단시간 처리 기법은 수식적으로 다음과 같이 나타낸다.

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \quad (1)$$

음성 신호는 선형 또는 비선형 전달 함수 T[ ]를 필요로 하고 조정 가능한 변수에 의존하게 된다. Q_n값은 T[x(m)]의 가중치가 적용된 평균 시퀀스이다.

음성 신호의 진폭은 시간에 따라 변하며 무성음의 진폭은 일반적으로 유성음 구간의 진폭보다 아주 작다. 진폭 변화를 반영하는 음성 신호에 대한 단시간 에너지는 다음과 같이 나타낸다.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad (2)$$

무성음 프레임에 대한 E_n값은 유성음 프레임에 대한 값보다 아주 작다. 에너지 함수는 유성음이 무성음으로 되는 시간과 반대로 무성음이 유성음으로 되는 시간을 근사적으로 나타낸다. 또한 고음질 음성에 대한 에너지는 묵음으로부터 음성을 구별해 내는데 사용될 수 있다.

이산 시간 결정론 신호의 자기 상관 관계 함수는 다음과 같다.

$$\phi_n = \sum_{m=-\infty}^{\infty} x(m)w(m+k) \tag{3}$$

만일 신호가 불규칙하거나 주기적이면 다음과 같이 나타낸다.

$$\phi(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{m=-N}^{N} x(m)x(m+k) \tag{4}$$

신호가 주기 P샘플로 주기적이라면 Φ(k)=Φ(k+P)로서 자기 상관 관계 함수는 같은 주기로 주기적이다. 자기 상관 관계 함수의 특성은 다음과 같다.

1) 우함수이다. 즉, Φ(k) = Φ(-k)
2) k=0에서 최대가 된다. 즉, | Φ(k) | ≤ Φ(0)
3) Φ(0)값은 결정론 신호의 또는 불규칙하거나 주기적인 신호에 대한 평균 에너지와 같다.

따라서 자기 상관 관계 함수는 에너지를 포함하고 있고 주기성을 나타낸다. 만약 특성을 고려하면 주기적인 신호에 대하여 자기 상관 관계 함수는 샘플 0, ±P, ±2P, ...에서 최대값이 된다. 즉, 신호에 대한 시간 기원에 무관하게 주기는 자기 상관 관계 함수에서 첫 번째 최대값을 찾음으로써 측정할 수 있다.

이산 시간 신호에서는 연속적인 샘플의 부호가 서로 다를 때 영 교차가 발생한다.

$$Z_n = \sum_{m=-\infty}^{\infty} | \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)] | w(n-m) \tag{5}$$

$$\mathrm{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

$$w(n) = \begin{cases} \frac{1}{2N}, & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

단시간 영 교차율은 단시간 에너지와 단시간 평균 진폭과 같은 일반적인 특성을 가지므로 영 교차율이 발생하는 곳을 결정하기 위해 샘플을 체크하고, 평균은 N개의 연속적인 샘플에 대해 계산한다. 이것은 가중치가 취해진 평균이고 만일 대칭적으로 유한한 길이의 창함수가 사용되면 지연은 정확하게 보상된다.

평균 영 교차율은 정현파의 주파수를 측정하는 방법으로 영 교차율 측정에서의 고려 사항은 샘플링 주기 T와 평균을 취할 구간이다. 샘플링 주기는 영 교차율 표현법에 대한 시간 분해력을 결정한다.

예를 들어, F_s비율로 샘플링된 F_o주파수에 대한 평균 영 교차율은 다음과 같이 정현파의 주파수를 측정하기에 적합하다.

$$Z = 2F_0 / F_s \tag{6}$$

III. 피치 검출법과 퍼지

1. 피치 검출법

음성 분석은 특징 파라미터를 구해서 음성의 전달 특성을 파악하는 것으로 LPC 분석, 피치 분석, 포만트 분석, 스펙트럼 분석 등이 있다.

피치 검출법에는 시간 영역법, 주파수 영역법, 혼성 영역법 등이 있으며, 메모리 절약을 위해서는 주파수 영역법을 사용하고 파라미터의 정확성을 위해서는 시간 영역법을 사용하고 있다. 신호의 상관 관계에 따른 봉우리와 골을 강조하여 피치를 검출하는 자기 상관 관계법은 분해력이 높으나 잡음에는 약하다. FFT를 수행하여 피치를 검출하는 주파수 영역법은 계산 시간이 방대하며 잡

음에 강하다. 따라서 본 논문에서는 계산량이 적고 잡음에 강인한 피치 검출법을 제안한다.

음성 신호를 발생원에 따라 분석을 해 보면 화자의 개성을 담고 있는 기본 주파수와 성도의 필터링 과정에서 발생되는 포만트들로 이루어져 있다. 봉우리와 골을 검출하는 식은 다음과 같다.

$$PV(n) = [s(n+1)-s(n)]*[s(n+2)-s(n+1)],\ n=1,2,..,k \qquad (7)$$

여기서 PV(n)은 검출된 봉우리와 골이고, s(n)은 음성 신호이다. 결과 값이 음의 값이면 봉우리와 골로 간주하고, 양의 값이나 영일 경우에는 상승이나 하강 중인 샘플로 간주한다.

여성이나 어린이의 음성에 적용할 때는 음의 형태 자체가 1차 데시메이션만으로도 검출이 가능하므로 한 번의 인터폴레이션을 적용하여 데시메이션을 적용하게 된다. 인터폴레이션의 지연 값은 검출된 봉우리와 골의 2.5ms이고, 크기는 두 이웃 검출 값의 중간 값을 적용한다. 따라서 모든 신호에 적용할 수 있는 강인한 알고리즘을 음성 신호에 적용하게 된다.

그림 2는 제안한 피치 검출법의 구성도이다.

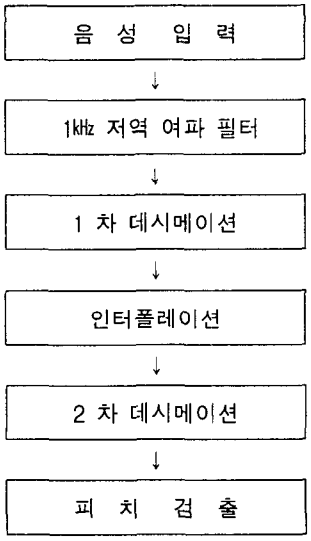


그림 2. 제안한 피치 검출법의 구성도
Fig. 2 Block diagram of proposed pitch detection method

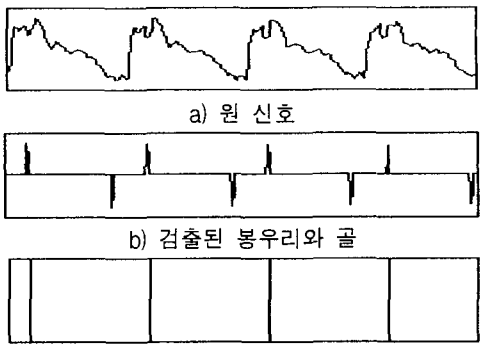


그림 3. 국부 봉우리와 골에 의한 피치 검출 알고리즘의 처리 과정
Fig. 3 Disposal processing of pitch detection algorithm

그림 3은 한국어 모음의 한 프레임에 대한 피치 검출을 나타낸 것이다.

피치 검출법은 성도의 공명 현상에 나타나는 포만트들의 영향을 제거시키고 여기원의 피치만을 강조하기 위한 목적으로 음성 파형의 주기성을 강조하며, 이런 성도의 영향을 제거시키면 결정 논리가 간단해진다. 피치 검출의 결정 논리법은 강조된 파형의 기본 주기를 실험적인 문턱 값이나 무게치를 적용하여 결정하는 것이다.

2. 퍼지

음성의 애매성으로 화자를 퍼지 추론에 의해 인식할 때, 인식을 위한 생성 규칙이 필요하다. 음성 신호의 피치 검출을 위해서는 퍼지 추론을 이용하여 특징 파라미터에 대한 소속도 함수를 구하고 퍼지 집합을 생성한다.

IF 조건 THEN 결론 END

조건 : 음성 신호내 모음을 선형 필터에 통과시킨 후 국부 봉우리와 골을 이용한 피치 검출법에서 얻은 피치 주파수 성분의 주파수 에너지가 퍼지 값을 갖고, 피치 인덱스가 퍼지 값을 갖는다면

결론 : 음성 신호 "아, 에, 이, 오, 우"

표 1. 피치 주파수 특징량의 퍼지화
TABLE. 1 A fuzzified features of pitch frequencies

퍼지 값 / 주파수[Hz]	표 준 패 턴		시 험 패 턴	
	퍼지값	에너지[dB]	퍼지값	에너지[dB]
대역 1 : 33	32	0.825	31	0.775
대역 2 : 66	30	0.750	29	0.750
:	:	:	:	:
대역 15 : 495	6	0.150	5	0.125
:	:	:	:	:
대역 30 : 1000	12	0.300	11	0.275

퍼지 이론에 의한 화자 인식은 화자가 발성한 음성에 대해 FFT를 수행한 후 행렬 양자화 인덱스와 각 주파수의 스펙트럼 양자화를 특징으로 사용하여 음성의 변동을 해결할 수 있도록 퍼지화 패턴으로 표현한다. 따라서 스펙트럼 양자화는 주파수를 채널로 나누어 각각의 중심 주파수에 해당하는 에너지에 대해 0.1dB 마다 퍼지 값을 주어 대응시킨다.

피치 주파수 특징량의 퍼지화를 표 1에 나타내었다.

3. 화자 인식

입력된 음성에 대해서 한 프레임을 512단위로 피치를 구하고 음성 구간에 대해서 평균 피치를 얻고 전체 구간에 대해서는 피치 패턴을 얻는다. 피치 패턴을 구하는 구성도는 그림 4와 같다.

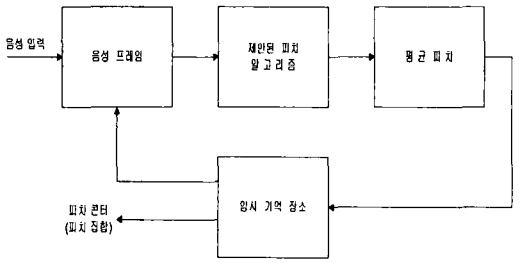


그림 4. 피치 콘터 검출의 구성도
Fig. 4 Block diagram of pitch contour detection

인간의 발성은 임의의 형태의 포락선으로 발음되어지는 경우가 많으므로 프레임별로 에너지를 구하고 평균 에너지를 구한 후 전체 에너지 패턴을 얻는다. 에너지 패턴을 구하는 구성도는 그림 5와 같다.

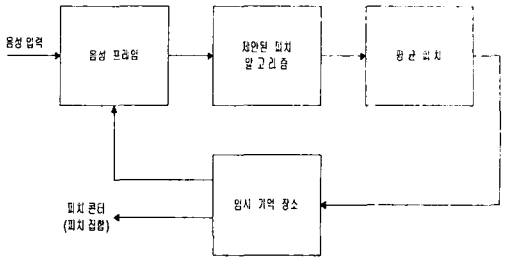


그림 5. 에너지 콘터 검출의 구성도
Fig. 5 Block diagram of energy contour detection

퍼지 이론을 사용할 확신도를 구하기 위해 표준 패턴과 시험 패턴의 주파수에 대한 퍼지 값의 소속도 함수 $S^c(i)$는 퍼지 추론의 합성 규칙을 적용한다.

$$S^c(i) = \vee(\mu^{ci}\mathrm{ref} \wedge \mu^{ci}\mathrm{test}) \qquad (8)$$

단, i = 1, 2, ... , n (i : 프레임 번호)

μ^{ci} ref : i번째 프레임 인덱스에 대한 표준 패턴의 소속도 함수

μ^{ci} test : i번째 프레임 인덱스에 대한 시험 패턴의 소속도 함수

$S^c(i)$: i번째 프레임 코드 북 인덱스에 대한 확신도 값

IV. 화자 인식 실험 및 고찰

본 논문에서의 실험은 한국어, 일어의 숫자 음을 사용하였다.

표 2. 한국어 및 일어 숫자 음 시료
TABLE. 2 Korean and Japanese number sounds data

숫 자	0	1	2	3	4	5	6	7	8	9
한국어 발 음	영	일	이	삼	사	오	육	칠	팔	구
일 어 발 음	れい	いち	に	さん	し	こ	ろく	しち	はち	く

표 3. 피치 검출(숫자 음)을 사용한 화자 인식
TABLE. 3 Result of speaker recognition using pitch detection(number sounds)

숫 자	0	1	2	3	4	5	6	7	8	9
한국어	83%	81%	84%	81%	84%	83%	81%	81%	84%	84%
일 어	82%	83%	85%	81%	84%	85%	82%	84%	84%	82%
평 균	82.5%	82%	84.5%	81%	84%	84%	81.5%	82.5%	84%	83%

표 4. 퍼지 추론과 피치 검출(숫자 음)을 사용한 화자 인식
TABLE. 4 Result of speaker recognition using fuzzy and pitch detection(number sounds)

숫 자	0	1	2	3	4	5	6	7	8	9
한국어	85%	86%	85%	84%	86%	84%	85%	88%	86%	83%
일 어	87%	84%	86%	82%	87%	85%	84%	86%	86%	84%
평 균	86%	85%	85.5%	83%	86.5%	84.5%	84.5%	87%	86%	83.5%

표 3은 숫자 음에 대한 피치 검출을 사용한 화자 인식을 나타낸 것이고, 표 4는 숫자 음에 대한 퍼지 추론과 피치 검출을 사용한 화자 인식을 나타낸 것이다.

표 5. LPC켑스트럼 방법과 제안된 방법의 한국인 화자가 발음한 한국어 숫자 음 인식율 비교와 개선율
TABLE. 5 Recognition rate comparison and improvement rate of Korean number sounds for Korean speakers using LPC cepstrum method and proposed method

숫자음		LPC켑스트럼	제안된 방법	인식 개선율
한국어	영	92%	94%	2%
	일	93%	94%	1%
	이	93%	96%	3%
	삼	91%	93%	2%
	사	94%	95%	1%
	오	93%	95%	2%
	육	96%	96%	0%
	칠	92%	95%	3%
	팔	91%	94%	3%
	구	94%	95%	1%
평 균		92.9%	94.7%	1.8%

표 6. LPC켑스트럼 방법과 제안된 방법의 한국인 화자가 발음한 일어 숫자 음 인식율 비교와 개선율
TABLE. 6 Recognition rate comparison and improvement rate of Japanese number sounds for Korean speakers using LPC cepstrum method and proposed method

숫자음		LPC켑스트럼	제안된 방법	인식 개선율
일어	れい	92%	94%	2%
	いち	90%	96%	6%
	に	93%	94%	1%
	さん	92%	95%	3%
	し	90%	94%	4%
	こ	93%	94%	1%
	ろく	94%	95%	1%
	しち	92%	93%	1%
	はち	90%	90%	0%
	く	90%	94%	4%
평 균		91.6%	93.9%	2.3%

표 5와 표 6은 LPC켑스트럼 방법과 본 논문에서 제안한 방법의 인식율을 나타낸 것이다.

V. 결 론

본 논문에서는 피치 패턴과 퍼지 추론을 이용하여 화자 인식 실험을 수행하여 기존의 인식율보다 우수한 결과를 얻을 수 있었다.

음성 인식 시스템은 입력 음성의 특성, 화자 등에 따라 구분할 수 있다. 음성으로부터 주파수 영역의 특징을 추출하기 위해 FFT방식을 사용한다. 또한 구강의 형태를 필터로 가정하여 필터 계수를 음성의 특징으로 삼아 주파수 필터에 적용하여 얻은 스펙트럼과 원래 신호의 차를 구하여 퍼지 집합을 구한다. 계산량이 적고 잡음에 강인한 국부 봉우리와 골에 의한 피치 검출과 퍼지 이론을 이용한 화자 인식을 제안함으로써 한국어 숫자 음 및 일어 숫자 음에 대한 화자 인식율이 개선되어 좋은 알고리즘임을 확인하였다.

참고 문헌

[1] Satoshi Takahashi, Sho-ichi Matsunaga and Shigeki Sagayama, "Isolated Word Recognition Using Pitch Pattern Information", Ieice Trans. Fundamentals, Vol. E76-A, No.2, pp.231-236, February 1993.

[2] Jean-Claude Junqua, Jean-Paul Haton, "Robustness in Automatic Pitch Recognition", Kluwer Academic Publishers, 1996.

[3] L R. Rabiner and B H.Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[4] Gil-Jin Jang, Te-Won Lee, and Yung-Hwan Oh, "Learning Statistically Efficient Features for Speaker Recognition", Neurocomputing, Special Issue on BSS/ICA, Vol 49 : Issue1-4, pp.329-348, December 2002.

[5] J. M. Baker, "A New Time-Domain Analysis of Human Speech and Other Complex Waveform", Ph.D Dissertation, Carnegie-Mellon Univ., Pittsburgh, PA., 1975.

[6] 김연숙, "피치 정보를 이용한 격리 단어 인식에 관한 연구", 한국학술진흥재단, September1995.

[7] Sungwook Chang, Y. Kwon, and Sung-il Yang, "Speech Feature Extracted from Adapted Wavelet for Speech Recognition", Electronics Letters, pp.2211-2213, November 1998.

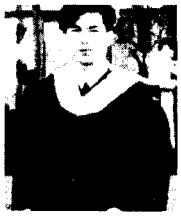
저자 소개

김연숙(Yeoun-sook Kim)

1981년 아주대학교 전자공학과(공학사)
1983년 아주대학교 전자공학과(공학석사)
1998년 건국대학교 전자공학과(공학박사)
1992년~1997년교육부-제1종교과용도서심의위원
1999년~현재 교육부 제1종 도서편찬 집필진
2003년~현재 서울시교육청 교수학습지원센터 교실수업지원단
2003년~현재 교육인적자원부- 사이버현장교원자문팀 자문위원
현재 상봉중학교
※ 관심분야 : 화자인식, 음성인식, VLSI Testing

김희주(Hee-joo Kim)

1987년 연세대학교 교육대학원(교육학석사)
1995년 성신여자대학교 식품영양학과(이학박사)
현재 강원관광대학 교수
※ 관심분야 : 화자인식, 음성인식



김경재(Kyoung-jae Kim)

홍익대학교 건축학과(공학석사)
1999년 홍익대학교 건축학과(공학박사)
※ 관심분야 : 화자인식, 음성인식