韓國 컴퓨터情報學會 論文誌
第8卷 第3號, 2003. 9.

2003-8-3-2-2

피치 검출과 퍼지화 패턴을 이용한 숫자음 화자 인식에 관한 연구

김 연 숙*, 김 희 주**, 김 경 재***

A Study on Number sounds Speaker recognition using the Pitch detection and the Fuzzified pattern

Yeoun-Sook Kim*, Hee-Joo Kim**, Kyoung-Jae Kim***

요 약

본 논문에서는 피치 검출과 퍼지화 패턴 매칭을 포함하는 화자 인식 알고리즘을 제안한다. 음의 개성을 표현하는 피치를 이용한 피치 패턴을 사용하고 음성의 파라미터는 2진화 스펙트럼을 사용한다. 비선형적인 발성 시간에 따른 시간 변동의 폭을 모두 포함할 수 있도록 음성 신호의 애매성을 보완할 수 있는 퍼지의 소속 함수를 이용하여 표준 패턴을 작성하고 퍼지화 패턴 매칭을 이용하여 인식을 수행한다.

Abstract

This paper proposes speaker recognition algorithm which includes both the pitch detection and the fuzzified pattern matching. This study utilizes pitch pattern using a pitch and speech parameter uses binary spectrum. In this paper, makes reference pattern using fuzzy membership function in order to include time variation width for non-utterance time and performs vocal track recognition of common character using fuzzified pattern matching.

▸ Keywords : Pitch detection, Fuzzifed pattern, LPC

* 건국대학교 전자공학과
** 강원관광대학
*** 홍익대학교

Ⅰ. 서 론

인간의 기계 사용으로 상호간의 정보 교환이 절실한 정보화 시대에서 여러 가지 연구가 행하여져 왔다.

인간은 의사 전달로 음성, 몸짓, 부호 등을 사용하지만, 그 중에서 음성 수단이 가장 활용도가 높고 효율적인 방법이다. 음성은 두 가지 범주로 생각해 볼 수 있는데, 공기가 성도를 통과하면서 그 흐름이 저항을 받지 않는 모음과 상당한 저항을 받아서 모음보다 크기가 더 작고 종종 잡음과 같은 자음이 있다. 모음과 자음의 차이는 시간 파형에서 쉽게 관찰할 수 있다. 시간 파형에서 얻을 수 있는 정보는 신호의 주기성, 강세 등이 있다. 모음은 모두 유성음으로 단독으로 발음되는 경우와 초성 또는 종성에 자음을 동반하는 경우가 있다. 이때 단독으로 모음을 사용하는 경우는 쉽게 구분할 수 있으나 초성 또는 종성에 자음이 오는 경우에는 모음 검출이 쉽지 않다. 더욱이 유성 자음을 동반할 경우에는 검출에 많은 어려움이 있다. 따라서 정확한 모음을 검출하여 각 화자별로 모음의 특징을 찾아 실제 음성에 적용하여 모음 검출 알고리즘을 구성한다[1][2].

화자 인식 연구는 1963년 Bell 연구소의 Pruzansky가 시간 평균 스펙트럼을 사용해서 화자 식별 실험을 하였고, 1981년 Bell 연구소의 Furui는 텍스트 의존 화자 인식 실험을 하였다. 국내에서는 1991년 권석규가 DSP 칩을 사용하여 H/W 설계를 하였다.

음성 인식에 사용하는 발음은 단어의 수에 제한을 받지 않아야 실용성과 일반성을 가질 수 있다[3][4][5].

이러한 문제는 화자 인식에 사용되는 파라미터들을 통계적으로 추출, 적용함으로써 해결할 수 있다.

본 논문에서는 피치 검출과 퍼지의 소속 함수를 이용하여 표준 패턴을 작성하고 퍼지 패턴 매칭을 이용하여 인식을 수행한다.

본 논문은 다음과 같이 구성되어 있다. 제 2장에서는 디지털 음성 신호 처리의 포만트 주파수 측정과 선형 예측 켑스트럼을 설명하고, 제 3장에서는 피치 검출과 퍼지 이론을 이용한 화자 인식을 살펴보기로 한다. 제 4장에서는 본 논문에서 제안한 내용과 실험을 평가하고, 제 5장에서는 결론을 맺는다.

Ⅱ. 디지털 음성 신호 처리

1. 포만트 주파수 측정

음성 분석은 파형의 중요한 특성을 표현하는 파라미터를 정확하게 추정하는 것으로 기본적인 파라미터인 포만트를 측정하는 것이다. 포만트 주파수 추출법에는 루트-솔빙법과 피크-피킹법 등이 있다.

음성의 각 프레임은 길이의 수열로 나타내지는데, 생성되는 데이터를 루트-솔빙법으로 계산하면 임의의 복소수근 z에 대한 대역폭 $\hat{A}$와 주파수 $\hat{E}$는 s평면에서부터 z평면으로 변환 $z = \exp(sT)$에 의해 구해진다. 여기서 $s = -\pi\hat{A} \pm j2\pi\hat{E}$이다.

복소수근의 실수부와 허수부를 표현하면 다음과 같다.

$$\hat{A} = -(fs/\pi)\log|z| \text{ [Hz]} \quad \cdots\cdots (1)$$

$$\hat{E} = -(f_s/2\pi)\tan^{-1}[\mathrm{Im}(z)/\mathrm{Re}(z)] \text{ [Hz]} \quad \cdots\cdots (2)$$

(단, 샘플링 주파수 : $fs = 1/T$)

피크-피킹법과 포물선 보간법에 의해 구한 값을 첨두값으로 사용할 때 포물선의 형태는 다음과 같다.

$$y(\lambda) = a\lambda^2 + b\lambda + c \quad \cdots\cdots (3)$$

y(0)이 이산적인 첨두값을 가지며 y(0)의 왼쪽, 오른쪽을 각각 y(-1)과 y(1)로 정의하면 이 점을 통과한 포물선은 다음 식의 점들을 갖는다.

$$c = y(0) \quad \cdots\cdots (4)$$

$$b = [y(-1) + y(1)]/2 \quad \cdots\cdots (5)$$

$$a = [y(-1) + y(1)]/2 - y(0) \quad \cdots\cdots (6)$$

zero인덱스에 대한 첨두 위치는 $\lambda_p = -b/2a$ 로, 만약 이산적 스펙트럼 위치가 n_p에 위치한다면 보간해서 생성된 데이터는 다음과 같다.

$$\hat{E} = (n_p + \lambda_p) f_s/N \quad \cdots\cdots (7)$$

모음에 대한 포만트 주파수와 성도간의 관계 규칙에 대하여 정리하면 다음과 같다.

① 길이규칙 : 포만트의 평균 주파수는 인두-구강 성도의 길이에 반비례

② F1규칙-구강수축 : 구강의 앞쪽이 수축할수록 F1주파수는 낮아진다.

③ F1규칙-인두수축 : 인두가 수축할 때 F1주파수는 높아진다.

④ F1규칙-후설수축 : 후설이 수축할수록 F2주파수는 낮아진다.

⑤ F1규칙-전설수축 : 전설이 수축할수록 F1주파수는 높아진다.

⑥ 원순규칙 : 입술을 둥글게 함으로써 모든 포만트 주파수는 낮아진다.

2. 선형 예측 켑스트럼(Linear Prediction Cepstrum)

입력된 음성 신호의 전력 밀도 스펙트럼을 P(w)라고 할 때 다음과 같이 나타낼 수 있다.

$$P(w) = |G(w)|^2 |H(w)|^2 = |V(w)|^2 |E(w)|^2 \quad \cdots\cdots (8)$$

단, S(w) : 신호의 스펙트럼
G(w) : 음성의 발생 신호
V(w) : 이상적인 음성의 발생 신호(impulse train)
H(w) : 스펙트럼의 envelope
E(w) : 스펙트럼의 envelope

스펙트럼 envelope E(w)는 대수적 파워 스펙트럼을 스므드하게 근사시킬 수 있다. 스므딩된 스펙트럼 F(w)는 다음과 같다.

$$F(w) = \int_0^{\infty} L(w-\xi)\log P(\xi)d\xi \quad (0 \le w \le \infty) \quad \cdots\cdots (9)$$

단, L(w) : I(t)의 퓨리에 변환

따라서, C(t)는 다음과 같다.

$$C(t) = I(t) \cdot C*(t) \overset{\mathcal{F}}{\leftrightarrow} F(w) \quad \cdots\cdots (10)$$

이때 대수적 연산자는 G와 H를 곱셈의 끝에서 덧셈의 형태로 바꾸어 줌으로써 F(w)는 P(w)보다 피치와 같은 여기원의 영향을 쉽게 제거할 수 있다.

선형 예측 코딩에 의해 구해진 스펙트럼을 $\sqrt{a/A(z)}$ 라고 하면 신호 스펙트럼 S(w)는 다음과 같다.

$$S(w) = \sqrt{a/A(z)} \quad \cdots\cdots (11)$$

단, $A(z) = \sum_{i=0}^{M} a_i z^{-1}$

여기서 M은 선형 예측 코딩의 차수이다. 따라서 P(w) 대신에 스펙트럼 근사를 이용하여 켑스트럼을 구할 수 있다.

$$C_n = \frac{1}{2\pi}\int_0^{2\pi} \log|S(w)| e^{jnw} dw \quad \cdots\cdots (12)$$

III. 피치 검출과 퍼지 이론

1. 피치 검출

음성 합성이나 인식 파라미터로서 피치 정보를 구하는데는 시간 영역법, 주파수 영역법, 혼성 영역법 등이 있으며, 메모리 절약을 위해서는 주파수 영역법을, 파라미터의 정확성을 위해서는 시간 영역법을 주로 사용하고 있다.

시간 주파수 혼성 영역법은 시간 영역법의 계산 시간 절감과 피치의 정밀성, 주파수 영역법의 배경 잡음이나 음소 변화에 대해 피치를 정확히 구할 수 있는 장점이 있다.

본 논문에서는 인식에 필요한 파라미터를 얻기 위해 정확한 피치를 검출하는 새로운 알고리즘을 제안한다.

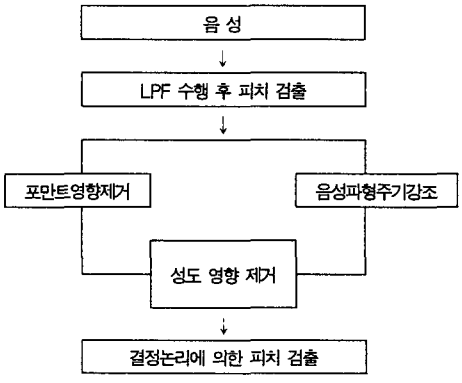

그림 1. 제안한 피치 검출의 구성도
Fig. 1 Block diagram of proposed pitch detection

음성 신호에서 피치를 검출하는 식은 다음과 같다.

$$PV(n) = [s(n+1) - s(n)] * [s(n+2) - s(n+1)] \quad \cdots\cdots (13)$$

단, $n = 1, 2, 3, \ldots, k$

음성 신호 S(n)에 의한 피치 검출 값을 확대시켜 적용하면 다음과 같다.

$$\overline{PV2} = [PV2 - PV1] * [PV3 - PV2] \quad \cdots\cdots (14)$$

$$= \begin{cases} 0, & [PV2 - PV1] * [PV3 - PV2] > 0 \\ 1, & [PV2 - PV1] * [PV3 - PV2] < 0 \end{cases}$$

PV1은 0이 아닌 첫 번째 PV(n) 값이고, PV2는 0이 아닌 두 번째 PV(n) 값이고, PV3은 0이 아닌 세 번째 PV(n) 값이다. 여성이나 어린이의 음성에서는 한 번의 인터폴레이션으로 데시메이션을 적용하고 인터폴레이션의 지연값은 두 이웃 검출 값의 중간 값을 적용한다. 따라서 모든 신호에 적용할 수 있는 알고리즘으로, 입력 버퍼에 512 샘플이 채워지면 한 프레임으로 인정하여 1㎑에 해당하는 LPF를 수행한 후 피치 검출을 수행한다. 성도의 공명 현상에 나타나는 포만트들의 영향을 제거시키고 여기원의 피치만을 강조하기 위한 목적으로 음성 파형의 주기성을 강조하며, 이러한 성도의 영향을 제거시키면 결정 논리가 간단해진다.

2. 퍼지 이론

화자 인식을 위하여 대부분 확률을 기본으로 하는 통계적인 기법들을 사용하여 왔다. 그러나 확률 밀도는 함수의 파라미터에 대한 전제가 정확하지 않다.

측정하고자 하는 특징들은 실제로 애매한 것들이기 때문에 임의의 측정으로 분포를 알아보기 위해 확률 밀도에 의한 통계적인 수법만을 사용하는 것은 정확하지 않다. 따라서 본 논문에서는 음성 신호의 애매성을 극복하기 위해 퍼지 이론을 인식에 수행한다.

화자가 발성한 모음에 대한 피치 검출에서 얻은 피치 주파수를 사용하여 퍼지 집합을 형성한다. 스펙트럼 정보를 사용하여 퍼지 값을 할당하고 각 음성 신호 별로 주파수 값에 대한 퍼지화 값을 할당한다. 음성의 변동을 해결하기 위해 다음과 같이 퍼지화 패턴으로 표현한다.

$$A' = \text{fuzzifier}(x) \quad \cdots\cdots (15)$$

입력 변수들의 대집합 U에 있는 값 x를 하나의 퍼지 집합 A′로 매핑한다.

n개의 입력 데이터 Xci(단, i=1,2,...n)의 평균을 $\bar{X}$, 표준편차를 σx라 할 때, 입력 데이터에 대응하는 A′를 퍼지수로 정의하는 방법이 그림 2에 나타나 있다. 이때 Xt = 2σx이다.

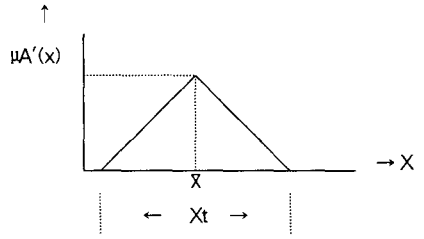

그림 2. 입력 데이터에 대응하는 퍼지수 A′
Fig. 2 Fuzzy number A′ corresponding with input data

각 음성 신호의 특징량은 각 모음에서 추출한 차분 주파수와 스펙트럼 양자화로 한다. 이때 스펙트럼 양자화를 퍼지 값으로 나타내기 위해 주파수를 채널로 나누어 각각의 중심 주파수에 해당하는 에너지만을 나타낼 수 있도록 한다.

퍼지 이론에 의한 화자 인식은 화자가 발성한 음성에 대해 FFT를 수행한 후 행렬 양자화 인덱스와 각 주파수의 스펙트럼 양자화를 특징량으로 사용하여 음성의 변동을 해결할 수 있는 퍼지화 패턴으로 표현한다. 〈표 1〉은 주파수 특징량의 퍼지화를 나타내고 있다.

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

퍼지 값 / 주파수 (Hz)	표준 패턴		시험 패턴	
	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 29 : 957	16	0.400	19	0.475
대역 30 : 1000	12	0.300	11	0.275

표준 패턴과 시험 패턴의 스펙트럼 양자화 값에 대한 확신도는 다음과 같다.

$$S_e(i) = \vee(\mu_{ei}\,\mathrm{ref} \wedge \mu_{ei}\,\mathrm{test}) \quad \cdots\cdots (16)$$

단, i = 1, 2, ... , N (i : i번째 채널)

μei ref : i 번째 표준 패턴의 소속도 함수

μei test : i 번째 시험 패턴의 소속도 함수

표준 패턴과 시험 패턴의 퍼지 값이 모두 같으면 확신도 값은 1이 된다. 이것은 두 패턴들의 의미가 일치한다는 것을 나타낸다.

표 2. 대역1의 피치 에너지에 대한 확신도 결과
TABLE. 2 Result of certainty factors of 1st band pitch energy

확신도 / 퍼지값	표준 패턴의 소속도 함수	시험 패턴의 소속도 함수	확신도(1)
1	0.0	0.0	0.0
2	0.0	0.0	0.0
.	.	.	.
20	0.9	0.9	0.9
21	1.0	1.0	1.0
.	.	.	.
30	0.1	0.1	0.0
.	.	.	.
40	0.0	0.0	0.0

표 3. 대역10의 피치 에너지에 대한 확신도 결과
TABLE 3. Result of certainty factors of 10th band pitch energy

확신도 / 퍼지값	표준 패턴의 소속도 함수	시험 패턴의 소속도 함수	확신도(10)
1	0.0	0.0	0.0
2	0.0	0.0	0.0
.	.	.	.
16	0.6	0.9	0.9
17	0.7	1.0	1.0
.	.	.	.
37	0.1	0.1	0.0
40	0.0	0.0	0.0

〈표 2〉와 〈표 3〉은 각각 대역 1과 대역 10의 피치 에너지에 대한 확신도 결과이다.

Ⅳ. 실험 및 고찰

〈표 4〉는 본 논문에서 사용된 시료를 나타낸 것이다.

표 4. 본 논문에서 사용한 숫자 음 시료
TABLE. 4 Used Korean and English number sounds data

숫 자	0	1	2	3	4	5	6	7	8	9
한국어	영	일	이	삼	사	오	육	칠	팔	구
영 어	zero	one	two	three	four	five	six	seven	eight	nine

〈표 5〉는 피치 검출을 이용한 인식율이고, 〈표 6〉은 피치 검출과 퍼지 이론을 이용한 인식율을 나타낸 것이다. 즉, 퍼지 이론을 이용한 것이 피치 검출만을 이용한 것보다 더 높은 인식율을 나타내고 있다.

표 5. 피치 검출을 이용한 화자 인식(숫자 음)
TABLE. 5 Results of speaker recognition using pitch detection (number sounds)

숫 자	0	1	2	3	4	5	6	7	8	9
한국어	85%	81%	84%	81%	84%	83%	81%	81%	84%	84%
영 어	84%	85%	83%	82%	83%	82%	83%	82%	82%	82%
평 균	84.5 %	83%	83.5 %	81.5 %	83.5 %	82.5 %	82%	81.5 %	83%	83%

표 6. 피치 검출과 퍼지 이론을 이용한 화자 인식(숫자 음)
TABLE. 6 Results 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%
영 어	86%	88%	87%	83%	85%	84%	86%	85%	85%	84%
평 균	85.5 %	87%	86%	83.5 %	85.5 %	84%	85.5 %	86%	85.5 %	83.5 %

〈표 7〉과 〈표 8〉은 LPC 켑스트럼을 사용했을 때의 인식율과 제안한 방법을 사용했을 때의 인식율을 나타낸 것이다.

표 7. LPC 켑스트럼 방법과 제안된 방법의 화자가 발음한 한국어 숫자 음 인식율 비교와 개선율
TABLE. 7 Recognition rate comparison and improvement rate of Korean number sounds for speakers using LPC cepstrum method and proposed method

시료 \ 방법		LPC 켑스트럼 방법	제안된 방법 (피치+퍼지 추론)	인식 개선율
데이터	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 8. LPC 켑스트럼 방법과 제안된 방법의 화자가 발음한 영어 숫자 음 인식율 비교와 개선율
TABLE. 8 Recognition rate comparison and improvement rate of English number sounds for speakers using LPC cepstrum method and proposed method

시료 \ 방법		LPC 켑스트럼 방법	제안된 방법 (피치+퍼지 추론)	인식 개선율
데이터	zero	90 %	94 %	4 %
	one	91 %	93 %	2 %
	two	92 %	96 %	4 %
	three	90 %	94 %	4 %
	four	93 %	93 %	0 %
	five	89 %	92 %	3 %
	six	88 %	90 %	2 %
	seven	92 %	94 %	2 %
	eight	91 %	95 %	4 %
	nine	93 %	93 %	0 %
평 균		90.9 %	93.4 %	2.5 %

한국어 숫자 음은 평균 1.8% 인식율이 개선되었고, 영어 숫자 음은 평균 2.5% 인식율이 개선되었다. 〈그림 3〉은 LPC 켑스트럼을 사용했을 때의 인식율과 제안한 방법을 사용했을 때의 인식율을 그래프로 나타낸 것이다.

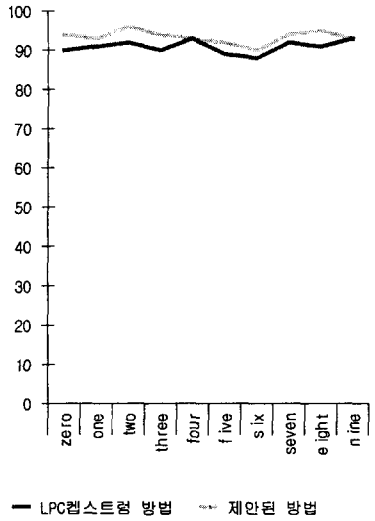

그림 3. LPC 켑스트럼 방법과 제안된 방법별 영어 숫자 음 인식율 비교
Fig. 3 Comparison of recognition rate of English number sounds according to LPC cepstrum method and proposed method

V. 결 론

본 논문에서는 피치 패턴과 퍼지 이론을 이용하여 화자 인식 실험을 수행하여 기존의 인식율 보다 우수한 결과를 얻을 수 있었다.

피치 검출과 퍼지 이론을 이용한 화자 인식을 제안함으로써 한국어 및 영어를 인식하는데 있어서 얼마만큼의 오인식율을 향상시킬 수 있는지에 대한 것으로서 실험과 시뮬레이션을 수행하였다. 인식 알고리즘으로는 본 논문에서 제안한 피치 검출법과 퍼지 이론을 사용하여 인식에 사용하였다.

입력 버퍼에 512샘플이 채워지면 한 프레임으로 인정하여 1㎑에 해당하는 LPF를 수행한 후 피치 검출을 수행하여 다시 결정 논리에 의해 정확한 피치를 검출한다. 화자가 발성한 음성에 대한 피치 검출에서 얻은 피치 주파수를 행렬 양자화 인덱스와 각 주파수의 스펙트럼 양자화를 특징량으로 사용하여 음성의 변동을 해결할 수 있는 퍼지화 패턴으로 표현한다.

본 논문에서는 인식율을 개선하기 위한 방법으로 본 논문에서 제안한 피치 검출법과 주파수 필터에 의한 퍼지 이론을 겸용하여 한국어 및 영어 화자 인식을 위한 좋은 알고리즘임을 확인하였다.

참고문헌

[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] L. Y. Kim, H. Y. Cho and Y. H. Oh, "Missing data techniques using voicing probability for robust automatic speech recognition", Electronics Letters, Vol. 37, No. 11, pp. 723-724, 2001.

[3] Per Hedelin and Dieter Huber, "PITCH PERIOD DETERMINATION OF A PERIODIC SPEECH SIGNALS", IEEE, CH2847-2/90/0000-0361, pp. 361-364, 1990.

[4] 김연숙, "피치 정보를 이용한 격리 단어 인식에 관한 연구", 한국학술진흥재단, SEPTEMBER 1995.

[5] Kai-Fu Lee, "AUTOMATIC SPEECH RECOGNITION," KLUWER ACADEMIC PUBLISHERS, Boston/Dordreecht/London, 1989.

저 자 소 개

김 연 숙

1983 아주대학교 전자공학과 (공학석사)

1998 건국대학교 전자공학과 (공학박사)

1992 ~ 1997 교육부 교과용 도서 심의위원

1999 ~ 현재 교육부1종 도서편찬 집필진

2003 ~ 현재 교육인적자원부 사이버현장 교원자문팀 자문위원

2003 ~ 현재 서울특별시교육청 교수학습지원센터 교실수업지원단원

현 재 상봉중학교

김 희 주

1987 연세대학교 교육대학원 (교육학석사)

1995 성신여자대학교 식품영양학과 (이학박사)

현 재 강원관광대학 교수

김 경 재

홍익대학교 건축학과(공학석사)

1999 홍익대학교 건축학과 (공학박사)