음성 변환을 사용한 감정 변화에 강인한 음성 인식

Emotion Robust Speech Recognition using Speech Transformation

김 원 구 Weon-Goo Kim

군산대학교 전기공학과

요 약

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 시스템을 구현하기 위하여 음성 변환 방법 중의 한가지인 주파수 와 핑 방법을 사용한 연구를 수행하였다. 이러한 목표를 위하여 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정의 변화에 따라 음성의 스펙트럼이 변화한다는 것과 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 관찰하였다. 본 논문에서는 이러한 음성의 변화를 감소시키는 방법으로 주파수 와핑을 학습 과정에 사용하는 방법을 제 안하여 감정 변화에 강인한 음성 인식 시스템을 구현하였고 성도 길이 정규화 방법을 사용한 방법과 성능을 비교하였다. HMM을 사용한 단독음 인식 실험에서 제안된 학습 방법은 사용하면 감정이 포함된 데이터에 대한 인식 오차가 기존 방법보다 감소되었다.

키워드: 강인한 음성 인식, 주파수 와핑, 성도 길이 정규화

Abstract

This paper studied some methods which use frequency warping method that is the one of the speech transformation method to develope the robust speech recognition system for the emotional variation. For this purpose, the effect of emotional variations on the speech signal were studied using speech database containing various emotions and it is observed that speech spectrum is affected by the emotional variation and this effect is one of the reasons that makes the performance of the speech recognition system worse. In this paper, new training method that uses frequency warping in training process is presented to reduce the effect of emotional variation and the speech recognition system based on vocal tract length normalization method is developed to be compared with proposed system. Experimental results from the isolated word recognition using HMM showed that new training method reduced the error rate of the conventional recognition system using speech signal containing various emotions.

Key Words: robust speech recognition, frequency warping, vocal tract length normalization

1. 서 론

음성 인식 기술이 발달함에 따라 인간과 기계사이의 보다 편리한 인터페이스로의 사용이 급격히 증가하고 있다. 최근에는 음성 인식 시스템의 실용화가 늘어나면서 실생활에 유용하게 사용될 수 있는 응용 제품들이 개발되고 있다. 현재 음성 인식 기술은 상당히 발전하여 수십만 단어의 어휘를 인식하고 실용화가 가능할 정도로 인식 성능도 향상되고 있다.

그러나 음성 인식 기술이 아직도 가지고 있는 문제점은 이러한 시스템의 성능이 주변 잡음 및 채널 특성 등의 환경 변화와 감정 상태와 같은 심리적 변화에 크게 좌우된다는 것이다. 이중에 환경 변화에 대한 연구는 음성 인식 시스템의 실용화를 위하여 오래 전부터 연구되어왔다[1-7]. 이와함께 음성 인식 시스템의 성능에 영향을 미치는 요인으로

접수일자: 2010년 8월 9일 완료일자: 2010년 9월 21일

감사의 글 : 본 논문은 2008년도 군산대학교 교수장기

해외연수경비의 지원에 의하여 연구되었음

인간의 심리적 변화가 있다. 음성 신호의 형태는 인간의 감정 상태에 따라서 변화하여 평상시 발음과 기쁨, 슬픔, 화남, 우울 등의 상태에서 발음한 것이 크게 다르다. 현재의음성 인식 시스템들이 평상시 감정 상태에서 발음한 음성데이터를 사용하여 만들어졌기 때문에 인간의 감정이 포함된 음성을 인식하는 경우에는 그 성능이 저하된다. 이와 관련된 외국의 연구로는 강세가 있는 음성이나 롬바드 효과를 갖는 음성에 대한 인식 성능 향상에 관한 연구가 오래 전부터 진행되어 왔다[8.9].

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 기술을 개발하기 위하여 감정 변화의 영향을 적게 받는 음성 인식 시스템에 관한 연구를 수행하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 신호와 음성 인식 시스템의 성능에 미치는 영향에 관한 연구를 수행하여 감정의 변화에 따라 화자의 음성 스펙트럼이 변화한다는 것을 분석하였다. 또한 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의하나임을 관찰하였다[10]. 본 연구에서는 이러한 음성의 변화를 포함하는 음성 인식 시스템에 포함시키는 학습 방법을 제안하였고 성도 길이 정규화 방법을 사용한 음성 인식 시스템과 성능을 비교하였다.

본 논문의 구성은 다음과 같다. 2장에서는 감정 변화에 강인한 인식 시스템을 만들기 위한 음성 변환을 이용한 학습 방법에 관하여 설명하고 3장에서는 성도 길이 정규화 방법에 대하여 설명한다. 4장에서는 다양한 실험을 통하여 감정이 음성에 미치는 영향을 분석하고 제안된 시스템의 성능을 비교 분석한다. 마지막으로 5장에서는 결론을 맺는다.

2. 음성 변환을 사용한 학습 방법

일반적으로 음성 인식 시스템은 감정이 포함되지 않은 평상 음성으로 학습되기 때문에 감정이 포함된 음성이 입력 되면 인식 성능이 저하되는 문제점이 있다. 이러한 것은 감 정이 음성의 특성을 변화시켜 음성 인식 시스템의 성능을 저하시킨다는 것을 의미한다. 본 연구에서는 감정의 변화에 따라 화자의 음성 스펙트럼이 변화한다는 것을 실험을 통하여 관찰하였고 이러한 변화는 음성 인식 시스템의 성능을 저하시키는 원인 중의 하나임을 분석하였다[10].

이러한 음성의 변화를 감정 인식 시스템에 포함하여 감정 인식 시스템의 성능을 향상하기 위하여 본 연구에서는 다음과 같은 학습 방법을 제안하여 감정 변화에 강인한 음성 인식 시스템을 개발하였다.

제안된 학습 방법은 그림 1과 같다. 고립단어 음성인식 시스템의 경우, 각 단어의 학습에 사용될 N개의 학습 데이 터 S_k , $k=1,\cdots,N$ 는 특징 추출 과정을 거쳐 N개의 특징 벡 터열 X_k , $k=1,\dots,N$ 로 변환된다. 제안된 방법에서는 특징 벡터 추출 과정에서 와핑 파라메터 α 를 사용하여 학습 데 이터를 음성 특징 벡터열 $X_k^{\alpha}, k=1,\cdots,N$ 로 변환한다. 그 릮 2는 주파수축으로 스펙트럼을 변형시키는 전형적인 와 핑 함수를 나타낸다. 여기서 와핑 파라메터 $\alpha=1$ 은 스펙 트럼의 변화가 없는 경우이고 $\alpha>1$ 는 스펙트럼을 고주파 영역으로 확장시키고 $\alpha < 1$ 인 경우에는 스펙트럼을 저주파 영역으로 압축하도록 변환시킨다. 이때 사용되는 와핑 파라 메터 α는 감정에 따른 성도 길이의 변화를 포함할 수 있는 정도로 변화시키기 위하여 -0.88 ≤ a ≤ 1.12 범위에서 0.02 간격으로 샘플링하여 사용하였다. 이 과정에서 M개의 와핑 파라메터가 사용된다면 N개의 학습 데이터는 각각의 와핑 파라메터에 의하여 변형되어 MXN개의 특징 벡터열 로 변환된다. 특징 벡터 추출 과정에서 만들어진 학습 데이 터는 각 음성을 학습하는데 사용된다. 학습된 음성 모델은 감정에 따라 변화하는 스펙트럼의 변형을 포함하는 감정 변 화에 강인한 음성 인식 시스템의 모델로 사용된다. 본 논문 에서는 이러한 방법을 왑 학습(warp training) 방법이라고 정의하였다.

그림 3의 왑 학습 방법을 사용하여 구성한 음성 인식 시스템의 학습과 인식 시스템 구조를 나타낸다. 감정 변화에 강인한 학습 방법을 사용한 시스템의 학습 과정에서는 다수의 와핑 파라메터를 사용하여 하나의 입력 음성에 대하여 다수의 특징 벡터 열이 생성된다. 이때 시용되는 주파수 외핑 파라메터는 식(1)과 같이 근사화된 선형 함수 $w_l(f)$ 를 사용하였다(12-16).

$$w_l(f) = \begin{cases} \alpha f & f \le f_0 \\ \alpha f_0 + \frac{f_N - \alpha f_0}{f_N - f_0} (f - f_0) & f > f_0 \end{cases} \tag{1}$$

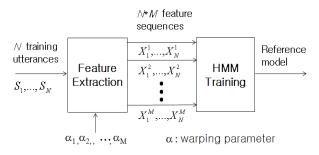


그림 1. 감정 변화에 강인한 모델 학습 방법 Fig. 1. a robust training method for emotional variation

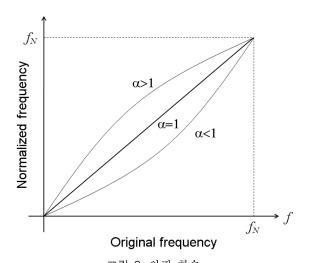


그림 2. 와핑 함수 Fig. 2. warping function

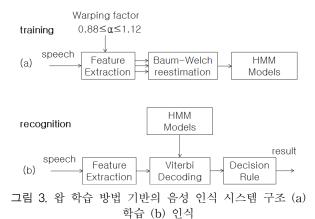


Fig. 3. The structure of speech recognition system based on warp training (a) training (b) recognition

여기서 제한 주파수 f_0 까지는 스펙트럼은 와핑 파라메터 α 로 선형적으로 와핑되고 f_0 부터 나이키스트 주파수까지는 다른 와핑 파라메터 α' 가 적용되어 $w_l(f_N)=f_N$ 으로 생략되는 주파수 영역이 없도록 하였다. 근사화된 선형 와핑 함수는 그림 4와 같다.

이렇게 생성된 특징 파라메터들은 HMM 모델을 학습하는데 사용된다. 인식 과정에서는 기존 방법과 동일하게 수

행된다. 입력된 음성에 대하여 전처리 과정을 거쳐 특징 파라메터를 구한 후 기준 모델과 유사도를 비교한다. 결정 법칙은 비교된 결과를 각 단어 당 기준 모델 수를 고려하여최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

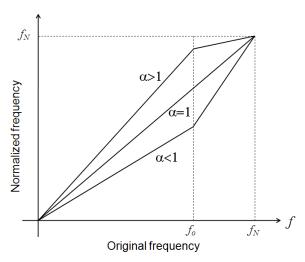


그림 4. 근사화된 선형 와핑 함수 Fig. 4. piecewise linear warping function

3. 성도 길이 정규화 기반의 음성 인식

성도 정규화 방법은 일반적으로 화자독립 음성 인식 시스템에서 화자의 성도길이 차이에 따른 음성 신호의 변화를 제거하기 위하여 각 화자의 성도 길이를 정규화 하는 방법이다. 성도의 길이를 변화시키는 방법은 음성 분석과정에서스펙트럼의 주파수 축을 와핑하는 것이다[11].

성도 길이 정규화 기반의 음성 인식 시스템 구조는 그림 5와 같다. 화자의 성도 길이를 정규화하여 음성 인식 시스템의 학습 과정에 사용하기 위해서는 각 화자의 성도 길이를 변화시킬 와핑 파라메터가 필요하다. HMM을 사용한 음성 모델을 λ 라고 가정하고 X_i^{α} 를 화자 i의 모든 음성에 와핑 파라메터 α 를 적용하여 구한 특징 벡터의 집합이라고한다면 최적의 와핑 파라메터 $\hat{\alpha_i}$ 는 문장 종속 확률 Pr을 최대화하도록 구하여진다.

$$\hat{\alpha_i} = \underset{\alpha}{\arg\max} Pr(X_i^{\alpha} | \lambda, W_i)$$
(2)

여기서 모델 λ 는 보통 1개의 밀도함수를 갖는 낮은 해상도의 음성 모델이 사용된다. 일단 모든 화자의 와핑 파라메터가 결정되면 학습 데이터는 그 값에 따라 정규화되고 이렇게 정규화된 학습 데이터를 사용하여 정상적인 학습 알고리즘을 사용하여 모델 Λ 을 학습한다.

인식 단계에서는 학습 과정과 비슷하게 와핑 파라메터를 결정한다. 일반적으로 입력 화자의 신원을 알 수 없으므로 최적의 와핑 파라메터는 입력 문장 단위로 계산되어 진다. 또한 입력 음성 j의 문자열 W_j 는 알 수 없으므로 초기 문자열 \widehat{W}_j 는 정규화되지 않은 입력 특징 벡터 X_j 와 정규화되지 않은 모델 Λ 를 사용하여 첫 단계로 인식을 수행하여

구한다. 그다음 최적의 와핑 파라메터 $\hat{\alpha_j}$ 는 정규화된 음성 모델 Λ 을 사용하여 결정된다.

$$\widehat{\alpha_j} = \arg\max_{\alpha} Pr(X_j^{\alpha} | \overline{\Lambda}, \widehat{W_j})$$
(3)

마지막 단계에서는 입력 특징 벡터는 $\hat{\alpha_j}$ 에 의하여 정규화되고, 정규화된 음성 모델 $\overline{\Lambda}$ 를 사용하여 두 번째 단계의 인식을 수행한다. 결정 법칙은 비교된 결과를 각 단어 당 기준모델 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 모델을 입력 음성의 단어로 결정한다.

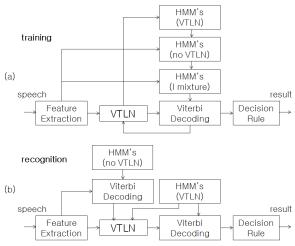


그림 5. 성도길이 정규화 방법 기반의 음성 인식 시스템 구조 (a) 학습 (b) 인식

Fig. 5. The Structure of speech recognition system based on VTLN (a) training (b) recognition

4. 실험 및 결과

4.1 데이터 베이스

감정 변화에 강인한 음성 인식 시스템의 성능을 평가하기 위해서는 다양한 감정이 포함된 음성 데이터 베이스가 필요하다. 이러한 데이터 베이스는 다음과 같은 과정으로 구성되었다[17]. 본 연구에서는 인간의 주요 감정인 기쁨, 슬픔, 화남의 3가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4가지 감정을 인식 대상 감정으로 결정하였다. 녹음된 음성은 아마추어 연극단원 남/녀 각 15명을 대상으로 하였고, 모든 참여자에 대해서 표준어 사용여부 및 감정표현능력을 심사하여 선별되었다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다. 본연구를 위하여 사용된 데이터의 규모는 16,200(30명×4감정×45문장×3회)문장이다.

4.2 특징 파라메터 추출

음성 인식을 위한 특징 파라메터 추출 과정은 다음과 같다. 전처리를 통하여 16KHz, 16비트로 샘플링하고, 고주파성분을 보강한다. 이렇게 샘플링된 신호는 음성 구간과 묵음 구간을 구별하기 위하여 음성 구간 검출을 수행하고 특징 벡터를 구한다. 검출된 음성 신호는 20ms(320샘플)의 길이로 10ms씩 이동하면서 특징 파라메터를 구한다. 본 연구

에서는 음성의 특징 파라메터로 12차 멜 켑스트럼 계수를 사용하였다. 또한 음성에 포함된 편의를 제거하는 방법으로 CMS (Cepstrum Mean Subtraction) 방법이 사용되었다.

4.3 음성 인식 시스템의 구성

본 연구에서는 우선 감정 변화에 강인한 음성 인식 시스템 개발을 위하여 우선 반연속 HMM을 기본으로 하는 화자 독립 단독음 인식 시스템을 구현하였다. 실험에 사용된음성 인식 시스템은 감정 변화에 강인한 학습 방법을 사용한 시스템(그림 3)과 성도 길이 정규화 방법을 사용한 인식시스템(그림 5)으로 구현하였다..

음성 인식 시스템의 구성에 사용된 반연속 HMM 모델은 256개의 코드를 갖는 코드북을 사용하였고 상태 당 4개의 가우시간 결합 분포를 사용하였다. 또한 각 모델의 상태 수는 학습에 사용된 문장의 평균길이에 비례하게 할당하였다. 모델의 학습에는 20명(남성 10명과 여성 10명)이 각 문장을 3회 발음한 음성이 사용되었고 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성 5명)이 각 문장을 3회 발음한 음성을 사용하였다.

4.4 실험 결과

평상시 발음한 음성을 대상으로 구축한 음성 인식 시스템에 감정이 포함된 음성이 사용되면 감정에 의한 음성 신호의 변형으로 인하여 인식 시스템의 성능이 크게 저하된다. 본 실험에서는 제안된 학습 방법의 성능을 평가하기 위하여 감정이 포함되지 않은 음성으로 학습한 인식 시스템, 주파수 와핑을 사용하여 학습한 인식 시스템과 성도 길이정규화 방법을 사용한 인식 시스템을 대상으로 4가지 감정이 포함된 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다.

그림 6은 각 학습과 인식 방법에 따른 감정별 인식 성능을 나타낸다. 여기서 기존 음성 인식 시스템(no warping)은 평상의 감정만 포함된 데이터로 학습되었기 때문에 인식 데이터가 평상인 경우에 가장 성능이 우수하고 감정이 포함되면 인식 성능이 저하된다. 한편 학습과정에서 학습데이터에 와핑 함수를 사용하여 학습된 인식 시스템(warp training)인 경우에는 기존 음성 인식 시스템보다 인식 성능이 향상되는 것을 볼 수 있다. 기쁨 음성의 경우에는 94.4%에서 95.9%로 인식 성능이 향상되었고, 슬픔의 경우에는 96.0%에서 97.1%로 인식 성능이 향상되었다. 화남의 경우에는 96.5%에서 97.0%로 인식 성능이 향상되었다.

또한 성도 길이 정규화를 통하여 학습과 인식 과정에서 정규화된 인식 시스템(VTLN)의 경우에는 모든 감정의 데이터에서 인식 성능이 향상되는 것을 알 수 있다. 평상 감정음성의 경우에는 인식률이 99.6%에서 99.9%로, 기쁨 감정 음성의 경우에는 94.4%에서 97.0%로, 슬픔 감정 음성의 경우에는 96.5%에서 97.7%로, 화남 감정 음성의 경우에는 96.5%에서 97.4%로 인식 성능이 향상되었다.

그림 7은 세 가지 학습 및 인식 방법을 사용한 인식 시스템에 대한 4가지 감정을 모두 포함한 음성에 대한 평균 인식률을 나타낸다. 기존 인식 시스템(no warping)의 평균 인식률은 96.6% 이고 학습과정에서 학습데이터에 와핑 함수를 사용하여 학습된 인식 시스템(warp training)인 경우에전체적인 인식 성능은 96.6%에서 97.3%로 향상되었다. 그리고 성도 길이 정규화를 통하여 학습과 인식 과정에서 정규화된 인식 시스템(VTLN)의 경우에는 전체적인 평균 인식 성능은 98.0%로 향상되었다.

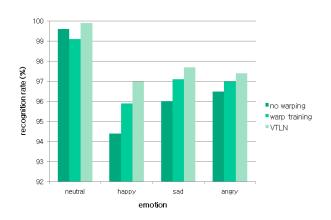


그림 6. 학습 방법에 따른 감정별 인식 성능 Fig. 6. Recognition performance according to emotions and the training methods



그림 7. 학습 및 인식 방법에 따른 인식 시스템 성능 Fig. 7 Recognition performance according to the training and recognition methods

5. 결 론

본 논문에서는 인간의 감정 변화에 강인한 음성 인식 시스템을 구현하기 위하여 주파수 와핑 방법을 사용한 방법에 관하여 연구하였다. 이러한 목표를 위하여 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정에 따른 음성 신호의 변화를 관찰하였고 감정의 변화에 따라 음성의 스펙트럼이 변화한다는 것과 이러한 변화는 음성 인식 시스템의성능을 저하시키는 원인 중의 하나임을 관찰하였다.

본 논문에서는 이러한 음성의 변화를 감소시키는 방법으로 음성 변환 기법중의 하나인 주파수 와핑을 사용한 학습방법을 개발하고 성도 길이 정규화 방법을 사용한 감정 변화에 강인한 음성 인식 시스템과 성능을 비교하였다.

HMM을 사용한 단독음 인식 실험에 제안된 학습 알고리 금을 사용한 경우 감정이 포함된 음성에 대한 인식 성능이 향상되는 것을 볼 수 있었다. 그러나 제안된 학습 알고리즘은 성도 길이 정규화 방법보다는 인식 성능의 향상이 적었다. 하지만 제안된 주파수 와핑을 사용한 방법은 학습 과정에만 적용되는 것이므로 인식 시스템의 형태가 변형되지 않고 인식 시간도 변경되지 않는 장점이 있다. 또한 성도 길

이 정규화를 사용한 경우에 평상 감정의 인식성능 뿐만 아니라 기쁨, 슬픔과 화남의 인식 성능도 향상되는 것을 볼수 있었다.

참 고 문 헌

- [1] J. C. Junqua, and J. P. Haton, Robustness in Automatic Speech Recognition Fundamental and Applications, Kluwer Academic Publishers, 1996.
- [2] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," *Proc. of ICASSP*, pp. 849–852, April 1990.
- [3] H. Hermansky, N. Morgan, H. G. Hirsch, "Recognition of speech in additive and convolutional noise based RASTA spectral processing", *Proc. of ICASSP*, pp. 83–86, 1993.
- [4] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong, "Integrating RASTA-PLP into Speech Recognition", Proc. of ICASSP, pp. 421-424, 1994.
- [5] M. G. Rahim, B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. Speech & Audio Processing*, vol. 4, No. 1, pp. 19–30, 1996.
- [6] N. Amir, "Classifying emotions in speech: a comparison of methods", *Proc. of Eurospeech* '2001, Vol. 1, pp. 127–130, Aalborg, Denmark, 2001
- [7] A. Nogueiras, etc, "Speech emotion recognition using Hidden Markov Models", Proc. of Eurospeech '2001, Vol. 4, pp. 2679–2682, Aalborg, Denmark, 2001
- [8] R. W. Picard, *Affective Computing*, The MIT Press 1997.
- [9] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of Acoustical Society of America*, pp. 1097–1108, Feb. 1993.
- [10] 김원구, 방현진, "성도 정규화를 이용한 감정 변화에 강인한 음성 인식", 한국 지능시스템학회 논문지, 19권 6호, pp. 773-338, 2009
- [11] M. Pitz, H. Ney, "Vocal tract normalization equals linear transformation in cepstral space", *IEEE Trans. Speech & Audio Processing*, vol. 13, No. 5, pp. 930–944, 2005.
- [12] S. Wegmann, D. McAllaster, J. Orlofl and B. Peskin, "Speaker Normalization on Conversational Telephone Speech, Proc. of ICASSP, Atlanta, GA, pp. 339–342, May 1996.
- [13] L. Welling, R. Haeb-Umbach, X. Aubert and N. Haberland, "A study on speaker Normalization using vocal tract normalization and speaker adaptive training", *Proc. of ICASSP*, Seattle,

- WA, pp. 797-800, May 1998
- [14] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space", Proc. of ICASSP, Toronto, pp. 893–896, May 1991.
- [15] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", *Proc. of ICASSP*, Atlanta, GA, pp.346–349, May 1996.
- [16] Sirko Molau, Stephan Kanthak, Hermann Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition", *Proc. of* the ESSV'00, Cottbus, Germany, pp. 209–216, 2000
- [17] 강봉석, "음성 신호를 이용한 문장독립 감정 인식 시스템", *연세대학교 석사학위 논문*, 2000.

저 자 소 개



김원구(Weon-Goo Kim)

1987년 2월: 연세대 전자공학과 학사 1989년 8월: 연세대 전자공학과 석사 1994년 2월: 연세대 전자공학과 박사 1994년 9월~현재: 군산대 전기공학과 교수 1998년 9월~1999년 9월: Bell Lab, Lucent Technologies(USA) 객원연구원 2008년 7월~2009년 12월 호주 Griffith

대학교 교환교수

관심분야 : 음성 신호처리, 음성 인식, 감성 인식, 음성 변환,

화자 인식

Phone : 063) 469-4745 Fax : 063) 469-4699 E-mail : wgkim@kunsan.ac.kr