한국음향학회지 제27권 제8호 pp. 418~426 (2008)

음성 분리를 위한 스펙트로그램의 마루와 골을 이용한 시간-주파수 공간에서 소리 분할 기법

A Method of Sound Segmentation in Time-Frequency Domain Using Peaks and Valleys in Spectrogram for Speech Separation

임 성 길*, 이 현 수*
(Sung-Kil Lim*, Hyon-Soo Lee*)

*경희대학교 컴퓨터공학과
(접수일자: 2008년 8월 25일; 수정일자: 2008년 10월 14일; 채택일자: 2008년 11월 18일)

본 논문에서는 스펙트로그램에서 마루와 골을 이용한 주파수 채널 분할 알고리즘을 제안한다. 주파수 채널 분할 문제는 동일한 음원으로부터 발생한 음성이 포함된 주파수 채널들을 하나의 그룹으로 묶는 것을 의미한다. 제안된 알고리즘은 입력 신호의 평탄화된 스펙트럼에 기반한 알고리즘이다. 평탄화된 스펙트럼에서 마루와 골은 각각 세그먼트의 중심과 경계를 판단하기 위해 사용된다. 각 세그먼트를 하나의 소리로 묶는 그룹핑 단계 이전에 제안된 알고리즘에 의한 세그멘테이션 결과가 유용함을 평가하기 위하여 이상적인 마스크에 의한 세그멘테이션 결과와 제안된 방법을 비교한다. 제안된 방법을 협대역 잡음, 광대역 잡음, 다른 음성신호와 혼합된 음성신호에 대하여 실험하였다.
핵심용어: 주파수 채널 분할, 음성 분리, 소리 분할, 마루, 골, 스펙트로그램
투고분야: 음성처리 분야 (2.3)

In this paper, we propose an algorithm for the frequency channel segmentation using peaks and valleys in spectrogram. The frequency channel segments means that local groups of channels in frequency domain that could be arisen from the same sound source. The proposed algorithm is based on the smoothed spectrum of the input sound. Peaks and valleys in the smoothed spectrum are used to determine centers and boundaries of segments, respectively. To evaluate a suitableness of the proposed segmentation algorithm before that the grouping stage is applied, we compare the synthesized results using ideal mask with that of proposed algorithm. Simulations are performed with mixed speech signals with narrow band noises, wide band noises and other speech signals.
Keywords: Frequency Channel Segmentation, Speech Separation, Sound Segmentation, Peak, Valley, Spectrogram
ASK subject classification: Speech Signal Processing (2.3)

I. 서 론

잡음에 강인하여 실생활에 적용 가능한 음성 인식 시스템을 구현하기 위해서는 잡음과 음성이 혼합된 음향 신호로부터 음성을 분리하는 작업이 선행되어야 한다. 실제의 음향 환경에서는 수많은 잡음이 존재하기 때문에, 음성 이외의 잡음을 제거하거나 인식의 대상이 되는 음성에 선택적인 주의 집중이 이루어 져야 하며, 이러한 문제는 각테일 파티 효과로 널리 알려져 있다. 지난 수년간 각테일 파티 효과를 해결하기 위하여 많은 시스템이 제안되었으며, 통계적인 방법 [1][2]이나 잡음 제거 알고리즘 [3]에 기반한 시스템들이 제안되었다. 다른 접근 방법으로는 음향 심리학적 발견과 생리학적 발견에 근간을 둔 계산적 음향 장면 분석 방법 (CASA : Computational Auditory Scene Analysis)이 사용되고 있다 [4].

CASA의 접근 방법에서 음성의 분리는 음향 신호를 시간-주파수 영역에서 비슷한 특징을 가지는 영역으로 분할하는 세그멘테이션 (segmentation) 단계와 분할된 영역을 음원에 따라 그룹화하는 그룹핑 (grouping) 단계의 두 부분으로 구성된다 [5]. 본 논문에서는 시간-주파수 공간에서 소리의 분할 문제만을 다루고 있는데, CASA적인 접근 방법에서 음성의 분리 성능은 그룹핑 단계 보다 세그멘테이션 단계의 결과에 많은 영향을 받기 때문이

책임저자: 이 현 수 (leehs@khu.ac.kr)
446-701 경기도 용인시 기흥구 서천동 1 경희대학교 컴퓨터공학과
(전화: 031-201-2947; 팩스: 031-202-1723)

다. 또한 주파수 채널의 분할 문제는 beam-forming에서 채널별 가중치를 계산하는 문제, 잡음 환경에서 음원의 위치 추적 문제 [6], MBE (Multi Band Exitation) 모델을 이용한 소리 부호화 (Sound coding) [7] 문제 등에도 적용 가능하다.

주파수 채널 분할 문제를 해결하고자 하는 기존의 연구들은 CASA에 기반한 음성 분리 방법의 일부분으로 포함되어 있다. Wang과 Brown의 모델 [5]에서는 입력 신호의 주파수 분석을 위하여 ERB (Equivalent Rectangular Band-width) Filter Bank을 사용하였고, 청각 신경신호의 발화를 통계적으로 모델링 하기위하여 내이세포 모델 (inner hair cell model)을 적용하였다. 신경세포의 발화에 대한 running autocorrelation으로 구성되는 correlogram을 계산한 후에, correlogram으로부터 각 주파수 채널의 주기성을 계산 하였다. 계산된 주기성은 음성을 분리하기 위한 신경망 모델인 oscillatory neural network의 연결 강도 설정을 위하여 사용하였다. 주파수 채널의 분할은 oscillatory neural network에서 동기화 되어 발화하는 뉴런의 집합으로 결정된다. 이후 혼합된 음향 신호로부터 피치를 계산하여 계산된 피치를 이용하여 분할된 세그먼트 중 음성에 해당하는 부분을 그룹핑 과정을 수행하게 된다.

또 다른 주파수 분석 방법인 푸리에 분석을 이용한 음성 분리 및 주파수 채널 분할 방법도 제안되었다. Kan-kanhalli와 Srinivasan의 모델 [8]은 주파수 분석을 위하여 단구간 푸리에 변환 (Short Time Fourier Transform : STFT)을 사용하였으며, 채널간의 유사성을 평가하기 위하여 채널의 진폭 (amplitude)의 변화를 파라메터화 한 채널 변화량 (channel dynamics)을 사용하였다. 주파수 분할을 위한 알고리즘으로는 normalized cut 방법에 기반한 클러스터링 알고리즘을 사용하였다.

ERB Filter Bank는 계산적 청각 시스템 (Computational Auditory System)에서 인간의 청각 기관을 모델링 하기위해 널리 사용되고 있지만, 고주파 영역에서 너무 넓은 대역폭을 가지고 있으며, 저주파 영역에서는 채널들이 중첩되어 표현되므로 대역별 분석을 하기 어렵다. 또한 correlogram은 음성의 중간 단계 표현 (mid-level representation)으로 매우 유용하여 음성의 여러 특성을 잘 나타낼 수 있지만 계산량이 많다는 단점을 가지고 있다. 한편 채널 변화량은 시간-주파수 공간에서 주파수 슬라이딩과 주파수 진폭의 변화가 동시에 일어나는 현상을 표현하기 적합하지 않다.

본 논문에서는 CASA 접근 방법의 음성 분리에서 분리 성능을 향상시키기 위하여, 시간-주파수 영역에서 주파수 채널 분할을 위한 새로운 알고리즘을 제안한다. 제안하는 알고리즘은 계산의 효율성을 위하여 주파수 분석에 STFT를 사용하였으며, 평탄화된 스펙트로그램 (smoothed spectrogram)에서 마루 (peak)와 골 (valley)를 이용하여 분할 영역의 중심과 경계를 판단하는 방법을 제안한다.

2장에서는 음성 분리를 위한 기존의 연구에 대하여 기술 하고 본 연구에서 제안하는 주파수 채널 분할 방법을 이용하는 이진 마스크 (Binary Mask) 음원 분리 방법에 대하여 기술한다. 3장에서는 본 연구에서 제안하는 주파수 채널 분할 방법을 설명하고, 4장에서는 제안하는 방법에 의한 주파수 채널 분할 결과를 평가하는 방법에 대하여 논의한다. 5장에서는 실험 환경과 실험 결과, 결과에 대한 분석을 기술하며, 본 논문의 결론은 6장에서 기술하였다.

II. CASA 접근에서 마스크를 이용한 음성 분리 시스템 구조

배경 잡음과 혼합된 음성을 분리하기 위한 시스템에 대한 연구는 대표적으로 신호의 통계적인 특성에 기반한 접근 방법인 ICA (Independent Component Analysis) [9]와 생물학적 발견 및 심리학적 연구에 기반한 접근 방법인 CASA [1]가 주로 연구되고 있다. 두 방법에서 공통적으로 사용하고 있는 기본 시스템은 그림 1과 같다 [10].

그림 1에서 입력되는 음향 신호의 표현으로는 시간 공간에서 음압을 나타낸 음향신호, 시간-주파수 공간에서 주파수 성분의 크기를 나타내기 위한 STFT 스펙트로그램, 음향 신호의 주기성, 연속성 등을 표현한 correlogram 등이 사용되고 있다. 분리 방법은 각 시스템에서 사용하고 있는 알고리즘을 나타내는 것으로 CASA 접근 방법에서는 시간-주파수 공간에서 마스크 처리를 사용하고 있다. 특히 마스크의 성분이 1과 0으로 구성되어 있는 이진 마스크는 짧은 시간으로 나뉘어진 프레임에 대하여 가변적인 대역통과 필터를 적용하는 방법이다. 음성 분리를 위한 시스템에서 평가/제어 신호 부분은 음성을 분리하

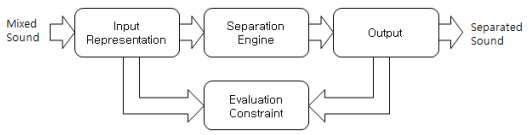


그림 1. 음향 분리를 위한 프레임워크
Fig. 1. Framework for Sound Separation.

기 위한 제약사항 및 규칙을 나타낸다. 음원의 위치, 기본 주파수와 배음구조, 진폭/주파수 변조 등을 음원 분리를 위한 제약으로 사용하고 있다.

본 연구에서 제안하는 시간-주파수 공간에서 주파수 채널 분할 방법은 CASA 접근 방법에서 이진 마스크를 생성하기 위한 단계의 일부분으로써 사용된다. 이진 마스크를 생성하는 방법은 우선 음향 신호를 주파수 분석을 이용하여 시간-주파수 공간의 신호로 변환한다. 이산 신호의 주파수 변환은 주파수 공간을 이산 채널로, 시간 공간을 프레임으로 구분하여 각 프레임별 각 채널을 기본 단위로 하여 분석하게 되는데, 이후 본 논문에서는 이 기본 단위를 음향 화소 (sound pixel)라고 한다.

입력된 음성/잡음 혼합 음향 신호는 주파수 분석 단계를 거쳐 음향 화소로 표현되고, 같은 음원에서 발생한 음향 화소는 시간-주파수 공간에서 지역적으로 밀집한 영역을 이루게 된다. 주파수 채널에서 이 영역을 찾는 문제가 주파수 채널 분할 문제이다.

시간-주파수 공간에서 지역적으로 나눠진 영역 혹은 세그멘트는 기본 주파수와 배음구조, 영역의 시작점 (onset), 밀집된 에너지의 공통된 변화, 밀집된 영역의 공통된 주파수 변화 등의 성질을 이용하여 전역적인 영역으로 그룹핑되며, 하나의 그룹에 포함되는 음향 화소에는 1의 값을, 나머지 음향 화소에는 0의 값을 지정한 것이 음향 신호 분리를 위한 이진 마스크이다. 이진 마스크는 시간-주파수 공간의 신호에 적용되어, 가변적인 대역 통과 필터 역할을 하며, 마스크가 적용된 신호를 역변환 하여 분리된 음향 신호를 획득한다.

III. 스펙트로그램에서 마루와 골 기반의 주파수 채널 분할 방법

본 논문에서는 평탄화된 스펙트로그램에서 골을 이용하여 분할 영역의 경계를 결정하여 주파수 채널을 분할하는 방법과 마루를 이용하여 분할 역역의 중심을 결정하여 채널을 분할하는 방법을 제안한다.

3.1. 제안하는 세그멘테이션 시스템 구조 및 전처리

본 논문에서 제안하는 마루와 골을 이용한 두 가지 분할 방법의 시스템 구조는 그림 2와 같다. 골과 마루를 이용한 두 가지 분할 방법 모두 STFT를 이용하여 주파수 분석을 수행하고, 보다 연속된 영역으로 분할하기 위하여 평탄화 방법을 사용하였다.

STFT를 이용한 주파수 분석은 셈플링 주파수와 분석 윈도우의 크기에 따라 시간 공간의 해상도와 주파수 공간의 해상도가 결정되고, 각각의 해상도에 따라 시간 공간에서의 연속성, 주파수 공간에서의 음성 배음구조 등의 정보가 명확해 지거나 불명확해 진다. 따라서 음향 신호를 주파수 분석하는 경우에는 분석 목적에 맞는 셈플링 주파수, 분석 윈도우의 크기, 분석 윈도우의 중첩 정도의 선택이 매우 중요한 문제이다. 본 연구에서는 음성의 분리가 최종적인 목적이므로 음성 정보를 충분히 표현 할 수 있도록 16 KHz 셈플링을 하였고, 음성의 기본 주파수가 80-300 Hz 사이에 분포한다는 특성을 이용하여 한 채널의 주파수 대역폭이 16 Hz 정도가 되도록 분석 윈도우의 크기를 1024셈플 포인트 (64 msec)로 하였다. 또한 채널을 지역적인 영역으로 나누는데 주파수 채널 변화의 시간 공간에서 연속성 파악이 중요한 역할을 하기 때문에 4 msec마다 윈도우를 설정하여 분석하였다.

주파수 분석된 신호의 주파수 성분 진폭을 시간-주파수 공간에 나타낸 스펙트로그램에서 같은 음원에서 발생한 지역적인 영역을 보다 부드럽게 표현하기 위하여 시간 방향과 주파수 방향으로 평탄화 과정을 수행하였다. 음향 신호의 스펙트로그램 특성은, 진폭이 큰 부분은 비교적 부드러운 곡선을 그리며 진폭이 작은 부분은 잡음 특성을 가지고 있다. 따라서 평탄화를 수행하면, 진폭이 큰 마루 부분의 채널 위치는 유지하고, 진폭이 작은 골 부분은 부드럽게 표현하여 신호의 미세한 요철에 영향을 받지 않고 골의 위치를 결정할 수 있다. 평탄화 방법으로는 식 (1)과 같이 이동 평균 (moving average) 방법을 사용하였다.

$$\bar{x}(n) = \frac{1}{2m+1} \sum_{i=-m}^{m} x(n+i) \tag{1}$$

식 (1)에서 x는 스펙트로그램에서의 진폭을, $\bar{x}$는 평탄

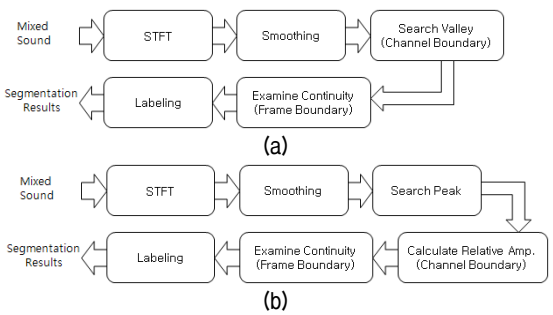


그림 2. 제안하는 주파수 분할 시스템 (a) 골 기반 (b) 마루 기반
Fig. 2. Proposed Frequency Segmentation System (a) Valley based (b) Peak based.

화된 진폭을 나타내며, $(2m+1)$은 이동 평균 계산에 사용된 윈도우의 크기를 나타낸다. 본 논문에서는 $m=2$로 설정하여 전체 윈도우의 크기를 5로 설정하였다.

3.2. 골을 이용한 분할 방법

스펙트로그램에서 특정 프레임에 해당하는 채널방향의 데이터는 분석 윈도우에 포함되어 있는 주파수 성분을 나타낸다. 특정 주파수는 푸리에 분석의 근접한 채널에서 중첩되어 표현되고, 이산 푸리에 변환은 연속적인 주파수를 이산적인 채널들의 합으로 근사 (approximation)하여 표현하므로, 복잡한 주파수 성분을 가진 음향뿐만 아니라, 단순한 주파수 성분을 가진 음향도 스펙트로그램에서는 포물선의 형태를 가지게 된다. 서로 다른 주파수 성분을 가진 음향 신호가 섞여 있는 경우, 너무 인접하지 않은 두 주파수 성분은 각각의 마루를 형성하며 두 주파수 성분의 영향이 거의 같아지는 채널에서 골을 형성하게 된다. 따라서 스펙트로그램의 채널 방향의 골은 서로 다른 음원으로부터 발생한 주파수 영역들의 경계를 판단하는데 중요한 단서를 제공하고 있다.

그림 3은 혼합된 음성 신호에서 채널 방향의 골의 위치를 표시한 것이다. 그림 (d)의 아래 부분은 남성의 음성만 존재하는 부분으로 골의 위치가 음성의 에너지가 집중되어 있는 부분을 연속적로 감싸고 있음을 볼 수 있다. 좌측 중앙 부분은 남성의 음성과 여성의 음성이 중첩되어 있는데 1로 표기한 타원 이 여성의 음성, 2로 표기한 타원이 남성 음성의 부분을 나타내고 있다. 서로 다른 음원으로부터 발생한 영역의 경우는 중첩되어 있는 부분을 제외하고 불연속적인 골이 나타나고 있음을 확인 할 수 있다. 이러한 골의 위치를 이용한 주파수 채널 분할 방법은 아래와 같다.

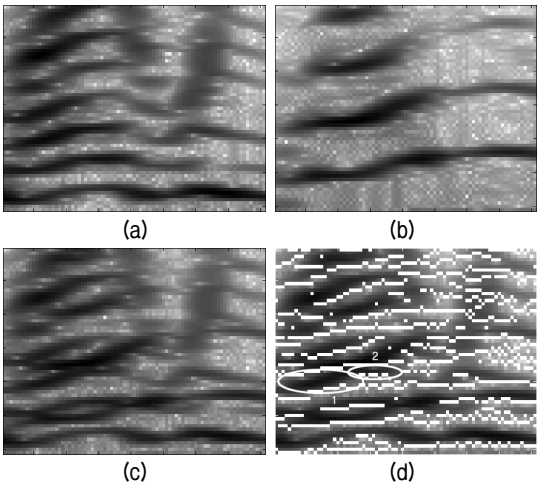


그림 3. 스펙트로그램에서 골의 위치 (a) 남성 음성 신호 (b) 여성 음성 신호 (c) 혼합된 음성 신호 (d) 혼합된 음성 신호에서 골의 위치

Fig. 3. Valleys in Spectrogram. (a) male speech (b) female speech (c) mixture (d) valleys in mixture

Step1. (채널방향 경계결정) 평탄화된 스펙트로그램에서 채널 방향으로 모든 골을 탐색하여 저장 함

$$V_{valley}(i,j)=\begin{cases}0 & \text{if } j \text{ is local minima}\\ 1 & \text{otherwise}\end{cases} \quad (2)$$

where j : channel number

i : frame number

Step2. (프레임방향 경계결정) 골을 저장한 V_{valley}에서 경계의 연속성을 이용하여 프레임방향 경계결정. i번째 프레임에서 인접한 골 j_k, j_{k+1}을 선택, 다음의 조건을 만족하는 경우 (j_k, j_{k+1}) 사이의 모든 값을 1로 설정 함

cond1. 이전 $(i-1)$프레임에서 $j_k \pm 1$범위에 골 존재

cond2. 이전 프레임에서 $j_{k+1} \pm 1$범위에 골 존재

cond3. 이전 프레임에서 (j_k, j_{k+1}) 범위에 골 없음

$$H_{valley}(i,j)=1 \text{ for } \forall j \in (j_k, j_{k+1}), \text{ if above conditions are satisfied} \quad (3)$$

$$H_{valley}(i,j)=0 \text{ for } \forall j \in (j_k, j_{k+1}), \text{ otherwise}$$

Step3. V_{valley}와 H_{valley}를 이용하여 채널/프레임 방향의 최종 결정 경계 결정

$$Seg_{valley} = V_{valley} \vee H_{valley} \quad (4)$$

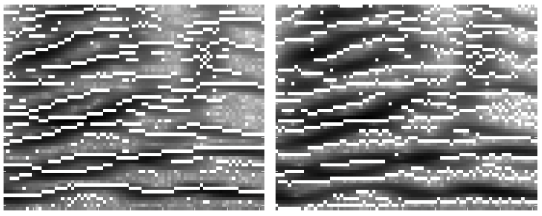
where $\vee$: Logical OR

Step4. Seg_{valley}에서 1로 표시된 영역을 레이블링 (Labeling)

3.3. 마루를 이용한 분할 방법

골이 시간-주파수 공간에서 동일한 음원으로 발생한 음향 화소 영역의 경계를 결정하는데 중요한 정보를 제공하고 있기는 하지만, 골을 이루는 음향 화소의 경우 주파수 공간에서 진폭이 작게 나타나므로 잡음의 영향을 많이 받는다. 반면에 마루는 음향 화소의 밀집된 영역의 중심에 위치하고 있으며, 진폭도 주변의 영역보다 크기 때문에 잡음의 영향을 비교적 적게 받는다.

그림 4는 여성 음성과 남성 음성이 혼합된 신호에서 골의 위치와 마루의 위치를 비교한 것이다. 골의 위치가



(a) (b)

그림 4. 스펙트로그램에서 골과 마루의 비교 (a) 혼합된 음성 신호에서 골의 위치 (b) 혼합된 음성에서 마루의 위치

Fig. 4. Comparison Peaks with Valleys (a) valleys in mixture (b) peaks in mixture

1-2 음향화소 차이로 부드럽지 못한 경계를 나타내고 있는 반면 마루의 위치는 비교적 부드러운 변화를 보이고 있고, 그림 (b)에서 좌측 중앙부분의 남녀 음성이 혼합되어 있는 모습도 더 명확하게 표현되고 있다. 마루를 이용한 주파수 채널 분할 방법은 먼저 밀집된 영역의 중심이 되는 마루들을 탐색하고, 마루의 연속성을 이용하여 시간 프레임 영역의 경계를 결정, 상대적인 진폭을 이용하여 주파수 채널 영역의 경계를 결정하는 순서로 수행한다. 마루를 이용한 주파수 채널 분할 방법은 아래와 같다.

Step1. (상대적인 진폭계산) 식(5)와 같이 상대적인 진폭을 계산

$$amp_{relative}(i,j) = (amp(i,j) - amp(i,j_v))/(amp(i,j_p) - amp(i,j_v)) \quad (5)$$

where $amp(i,j)$: amplitude of sound pixel

j_v : index of adjacent valley

j_p : index of adjacent peak

Step2. (마루 탐색) 평탄화된 스펙트로그램에서 채널 방향으로 모든 마루를 탐색하여 저장 함

$$P_{peak}(i,j) = \begin{cases} 1 & \text{if } j \text{ is local maxima} \\ 0 & otherwise \end{cases} \quad (6)$$

where j : channel number

i : frame number

Step3. (채널방향 경계결정) 마루의 위쪽 채널과 아래쪽 채널의 상대적인 진폭이 특정 임계치보다 작아지는 영역을 하나의 영역으로 결정

$$V_{peak}(i,j) = \begin{cases} 1 \ from \ j_k \ until \ amp_{realtive}(i,j) < \theta \\ 0 \ otherwise \end{cases} \quad (7)$$

Step4. (프레임방향 경계결정) i번째 프레임에서 하나의 마루 j_k를 선택, 이전 $(i-1)$ 프레임 에서 $j_k \pm 1$ 범위에 마루가 존재하지 않으면 프레임 방향의 경계로 설정, Step2에서 채널 경계로 결정된 영역에 대하여 경계 값으로 설정

$$H_{peak}(i,j) = \begin{cases} 0 \ from \ j_k \ until \ V_{peak}(i,j) = 0 \\ \quad \text{if } i \text{ is frame boundary} \\ 1 \ otherwise \end{cases} \quad (8)$$

Step5. V_{peak}와 H_{peak}를 이용하여 채널/프레임 방향의 최종 결정 경계 결정

$$Seg_{peak} = V_{peak} \vee H_{peak} \quad (9)$$

where $\vee$: *Logical OR*

Step6. Seg_{peak}에서 1로 표시된 영역을 레이블링 (Labeling)

각각의 분할 방법에서 레이블링된 영역은 동일한 음원으로부터 발생한 영역을 나타내며, 이후 영역의 시작 프레임 (음향의 시작점), 영역에 포함되는 에너지의 변화 등의 특성을 이용하여 하나의 음원으로부터 발생한 영역을 다른 배경 잡음으로부터 분리 하는데 사용된다.

IV. 분할 결과 평가 방법

기존의 연구에서는 세그멘테이션과 그룹핑을 모두 수행하여 배경 잡음으로부터 목적하는 음향신호를 얼마나 정확하게 분리하였는가를 분리된 신호와 혼합되기 이전의 신호에 대하여 신호 대 잡음비를 계산하여 평가하였다. 그러나 각 단계별로 성능을 평가하는 방법은 사용되지 않았기 때문에 두 부분으로 구성된 음향 분리 시스템의 각 부분을 독립적으로 분석하는데 어려움이 있었다. 본 논문에서는 혼합되기 이전의 음향신호를 이용하여 이상적인 분할 결과를 계산하고 분할 결과에 이상적인 그룹핑 방법을 적용함으로써 이진 마스크를 이용한 음성 분리에서 이론적으로 최상의 분리 결과를 계산하는 방법을 제안한다.

분리의 목적이 되는 음향 신호와 잡음이 섞여있는 혼합된 음향 신호의 경우, 특정 시간-주파수 공간의 영역은 잡음과 목적 음향이 섞여있게 된다. 그러나 이 경우에도 두 음향신호 중 보다 우세한 신호가 존재하게 된다. 목적 음향신호가 우세한 음향 화소에 1 값을 잡음이 우세한 음향 화소에 0값을 지정하여 이상적인 이진 마스크 $Mask_{ideal}$를

만들 수 있다.

$$Mask_{ideal}(i,j)=\begin{cases}1 & \text{if } a_{target}(i,j)>a_{noise}(i,j)\\ 0 & \text{if } a_{target}(i,j)<a_{noise}(i,j)\end{cases} \quad (10)$$

(식10)에서 $a_{target}(i,j)$는 음향 화소 (i,j)의 목적 음향의 진폭을 나타내며 $a_{noise}(i,j)$는 잡음의 진폭을 나타낸다. 이상적인 마스크를 혼합된 신호에 적용하여 재합성한 신호를 이진 마스크를 이용한 이상적인 분리 음성으로 정의 한다. 이상적인 분리 음성 Rec_{ideal}은 식 (11)과 같이 정의된다.

$$Rec_{ideal}=IFFT(MIX\times Mask_{ideal}) \quad (11)$$

식 (11)에서 $IFFT$는 역푸리에 변환 (Inverse Fourier Transform)을, MIX는 혼합된 음향 신호의 푸리에 변환 결과를 나타낸다.

본 논문에서 제안한 주파수 채널 분할 방법의 성능을 평가하기 위해서는 나누어진 영역과 이상적인 이진 마스크를 비교하여 일치하는 부분과 일치하지 않는 부분의 음향 화소 수를 측정하는 방법을 사용 할 수 있다. 그러나 이 방법은 음향 화소의 에너지가 고려되지 않은 방법이기 때문에 정확한 분리 성능을 예측하기 어렵다. 따라서 제안하는 방법에 의해 나누어진 영역에 목적 신호의 에너지가 크면 그 영역 전체를 1로, 잡음 신호의 에너지가 크면 그 영역 전체를 0으로 설정한 마스크를 식 (12)과 같이 생성하여, 혼합된 음향 신호에 적용, 재합성된 신호를 식 (13)에 의해 생성하고, 혼합되기 이전의 음성과 신호 대 잡음비를 계산하여 분할 영역의 정확성을 평가하였다.

$$Mask_{proposed}=\begin{cases}0 & \text{if } E_t(Seg(l))<E_n(Seg(l))\\ 1 & \text{if } E_t(Seg(l))>E_n(Seg(l))\end{cases} \quad (12)$$

$$Rec_{proposed}=IFFT(MIX\times Mask_{proposed}) \quad (13)$$

V. 실험 및 결과

본 논문에서는 제안된 채널 분할 방법의 동작을 검증하고, 기존의 채널 분할 방법과 비교하기 위한 실험을 수행하였다. 실험에 사용한 음향 신호는 목적 신호로 남성의 음성을 사용하였으며, 10개의 다양한 특성을 가지는 잡음 신호를 사용하였다. 10개의 잡음 신호는 각각 N0 : 1 kHz 순음 (pure tone), N1 : 백색잡음 (white noise), N2 : 폭발적인 잡음 (noise bust), N3 : 칵테일 파티 잡음 (cocktail party noise), N4 : 록 음악 (rock music), N5 : 사이렌 (siren), N6 : 전화벨 소리 (trill telephone), N7 : 여성음성 (female speech), N8 : 남성 음성 (male speech), N9 : 또 다른 여성 음성 (another female speech) 로 구성되어 있다. 각각의 데이터는 16 KHz로 셈플링되었으며, 이 데이터들은 Cooke [11]에 의해 수집된 데이터로써, 기존의 CASA 기반의 음성 분리 성능을 평가하기 위해 널리 사용되고 있는 데이터이다. 목적 신호와 잡음 신호의 합성은 두 신호의 에너지가 같도록 하였다.

5.1. 분할 결과

동일한 음원으로부터 발생한 소리가 시간–주파수 공간에서 밀집된 영역을 이루고 있다는 것은 주파수 분석 결과를 관찰함으로써 밝혀진 사실이지만 그 영역의 경계를 찾는 것은 매우 어려운 일이다. 또한 어떻게 영역을 나누는 것이 원래의 소리를 잘 표현 하는 것인가에 대한 기준이나 평가 방법도 정의하기 어렵다. 본 논문에서는 제안하는 두 가지 방법에 의한 세그멘테이션 결과를 비교함으로써 보다 좋은 세그멘테이션 방법의 특성들을 살펴본다.

그림 5는 10개의 잡음과 혼합된 음향 신호 중에서 사이렌 잡음 (N5)에 대한 분할 결과를 나타낸 것이다. 그림 5 (c)와 그림 5 (d)는 각각 골 기반의 분할 결과와 마루 기반의 분할 결과 중 4 KHz 이하의 저주파 부분을 나타낸 것으로, 음성에 해당하는 세그먼트는 진한 회색으로, 잡음에 해당하는 세그먼트는 연한 회색으로 나타낸 것이다. 골 기반의 분할 결과는 1263개의 세그먼트로, 마루 기반의 분할 결과는 676개의 세그먼트로 분할되었다.

그림 5 (c)에서 100–120 프레임, 100–150 채널 사이와 같이 주파수 슬라이딩이 일어나는 경우, 골을 이용한 방법에서는 분할 영역의 경계에 대한 연속성이 끊어져 많은 세그먼트가 생성된 반면, 마루를 이용한 방법에서는 시간적으로 오래 지속되는 적은 수의 세그먼트가 생성되고 있음을 볼 수 있다. 또한 200–220 프레임, 120–160 채널 사이의 영역을 보면 잡음의 주파수 슬라이딩이 매우 급격하게 일어나는 경우, 골을 이용한 방법에서는 세그먼트가 생성되지 않고 있음을 알 수 있다. 이러한 현상은 골의 위치가 음향 화소 진폭의 국부 최소점이기 때문에 잡음과 같은 외부 요인에 영향을 많이 받기 때문에 발생하는 것으로 보인다. 고주파 영역의 분할 결과를 살펴보면 골 기반보다 마루 기반의 결과에서 채널 방향으로 더 넓은 세그먼트들을 형성하고 있는데, 이 현상은 마루 기반의 방법에서 세그먼트의 경계를 판단하는데 상대적인 진폭을

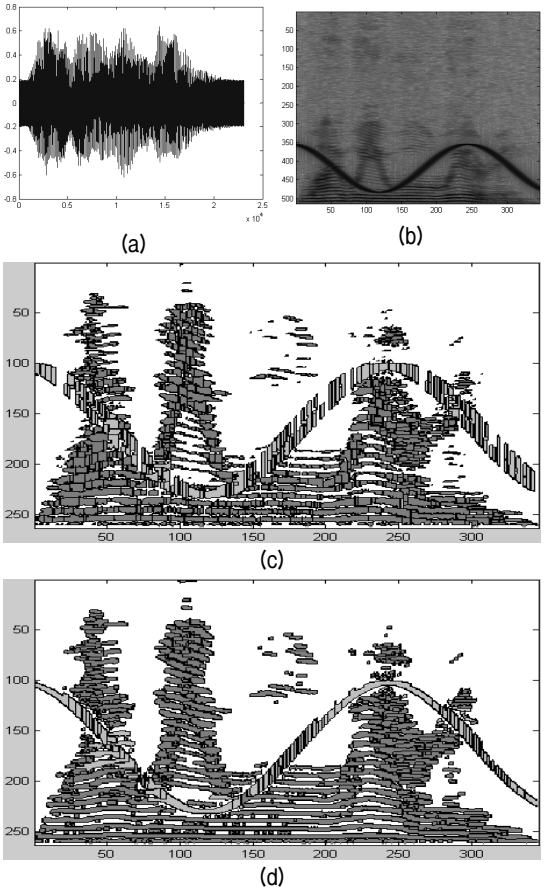


그림 5. 사이렌 잡음 (n5)에 대한 분할 결과 (a) 혼합된 음향 (b) 로그 스케일 스펙트로그램 (c) 골 기반의 분할 결과 (d) 마루 기반의 분할 결과

Fig. 5. Results of Segmentation for Siren Noise. (n5) (a) mixed sound. (b) log spectrogram (c) segments by valley based method (d) segments by peak based method

사용했기 때문이다.

표 1은 여러 가지 잡음에 따른 세그먼트의 수를 나타내고 있다. N0과 N5 잡음의 경우 비교적 적은 수의 세그먼트로 분할되었는데, 이는 두 잡음이 각각 순음과 사이렌 소리로 좁은 대역폭을 가지는 잡음이기 때문이다. N1의 경우 다른 잡음과는 달리 매우 많은 세그먼트로 분할되었는데, N1은 백색잡음이기 때문에 연속적이지 않은 매우 많은 단편으로 이루어져 있기 때문이다. 더 적은 수의 세그먼트가 항상 더 좋은 결과라고 판단 할 수는 없지만, 동일한 특성을 가지는 영역을 여러 개의 세그먼트로 분할하게 되면, 이후 그룹핑 과정의 계산을 복잡하게 만드는 원인이 된다.

표 1. 잡음에 따른 세그먼트의 수

Table 1. The Number of Segments for each Noise.

Noise	N0	N1	N2	N3	N4
Valley	1106	4590	2303	2390	2440
Peak	353	3095	1631	1521	1595
Noise	N5	N6	N7	N8	N9
Valley	1263	2285	2121	1920	1540
Peak	676	1360	1210	1140	910

5.2. 기존 모델과의 비교

기존의 연구에서 시간-주파수 공간에서 주파수 분할 문제는 음성 분리 알고리즘의 일부분으로 연구되었다. 따라서 제안하는 알고리즘은 [5][12]와 같은 음성 분리 알고리즘의 일부분과 비교되어야 하는데, [12]의 경우에는 세그멘테이션 단계와 그룹핑 단계가 명확하게 분리되어 있지 않으므로 직접적인 비교가 어렵다. 따라서, 본 논문에서는 [5]에서 제안된 Wang-Brown모델의 세그멘테이션 부분과 제안된 알고리즘을 비교한다.

표 2는 분리 이전의 혼합신호, Wang -Brown 모델에서 생성한 마스크, 제안된 골과 마루의 방법에 의한 마스크를 이용하여 혼합된 음향 신호로부터 분리된 음성, 이진 마스크를 사용한 경우의 이상적인 분리 신호의 신호 대 잡음비를 dB로 나타낸 것이다. 신호 대 잡음비는 (식14)과 같이 계산하였다.

$$SNR = 10\ \log_{10}\frac{S_{original}^2}{(S_{original}-S_{resyn})^2} \tag{14}$$

식 (14)에서 S_{resyn}은 혼합된 신호에 각 마스크를 적용하여 재합성한 분리 음성 신호이고, $S_{original}$은 혼합되기 이전의 음성 신호를 나타낸다.

N1, N3, N4의 경우는 광대역 잡음으로 음성의 영역과 중첩된 부분이 많아서 모든 알고리즘에서 신호 대 잡음비

표 2. 모델과 잡음에 따른 SNR 비교

Table 2. Improved SNR for each Noise and Model.

Noise	N0	N1	N2	N3	N4
Mixture	2.36	2.35	7.36	3.80	4.05
Wang	17.18	10.32	16.15	6.56	10.97
Valley	24.41	12.24	18.54	8.54	13.12
Peak	27.32	12.61	19.33	8.91	13.85
Ideal	30.67	15.07	23.21	11.91	17.19
Noise	N5	N6	N7	N8	N9
Mixture	2.36	2.05	5.12	9.41	5.44
Wang	15.74	16.34	10.66	14.02	7.31
Valley	20.57	22.45	15.12	17.04	11.16
Peak	16.82	23.16	15.83	17.05	10.24
Ideal	23.76	25.35	20.67	20.94	17.93

가 낮게 나타나고 있다. N0, N2, N5, N6의 경우는 협대역 잡음이거나 짧은 시간에 강하게 발생하는 잡음으로, 상대적으로 높은 분리 성능을 보이고 있으며, N7, N8, N9의 경우에는 잡음이 다른 음성 신호로 분리 목적이 되는 신호와 거의 같은 대역에 분포하고 있고, SNR도 비교적 낮게 나타나고 있다. N5를 제외한 모든 잡음에 대하여 마루를 이용한 방법이 골을 이용한 방법보다 좋거나 비슷한 성능을 보이고 있으며, 기존의 모델보다 좋은 분리 성능을 보이고 있다.

N5의 경우에는 골기반의 방법이 마루 기반의 방법보다 좋은 성능을 보이고 있는데, 이는 그림 5 (d)의 150-200 채널, 20-60 프레임에서 마루의 중첩 현상이 발생하여 잡음에서 발생한 음향화소와 음성에서 발생한 음향화소가 하나의 세그먼트로 분할되어 발생하는 현상이다. 반면 골 기반의 방법에서는 세그먼트가 작게 나누어지기 때문에 이러한 중첩부분의 영향이 작게 나타나고 있다.

그림 6은 잡음이 음성인 경우 중에서 분리 성능이 비교적 좋은 N7의 분리 음성 파형과 잡음 파형을 나타낸 것이다. 이상적인 이진 마스크를 사용한 경우에는 음성과 잡음 모두 약간의 왜곡만이 발생하고 있으며, 제안한 마루

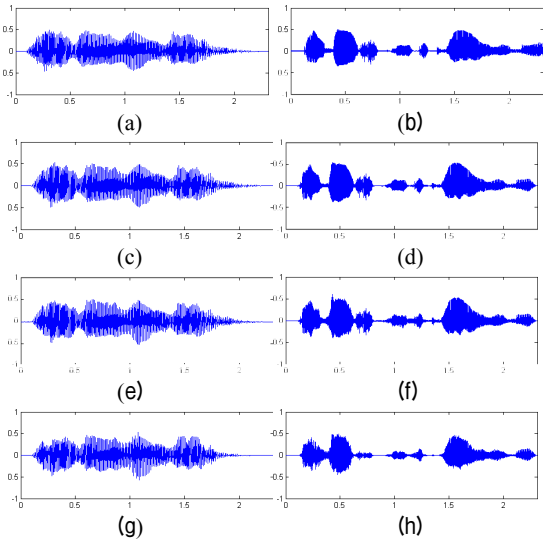


그림 6. 분리된 음성 신호 (a) 혼합하기 전 목적 음성 (b) 혼합하기 전 잡음 음성 (c) 이상적인 분리 음성 (d) 이상적인 분리 잡음 음성 (e) 마루 기반의 분리 목적 음성 (f) 마루 기반의 분리 잡음 음성 (g) 기존 모델의 분리 목적 음성 (h) 기존 모델의 분리 잡음 음성

Fig. 6. The separated speech signals. (a) original target speech (b) original noise speech (c) separated target speech by ideal mask (d) separated noise speech by ideal mask (e) separated target speech by peak based method (f) separated noise speech by peak based method (g) separated target speech by Wang-Brown model (h) separated target speech by Wang-Brown model

기반의 방법, 기존의 모델 순서로 음성의 왜곡이 심해지고 있음을 알 수 있다.

그림 7은 N7 잡음에 대한 분리 결과 중 4 KHz 이하 부분을 시간 주파수 공간에서 나타낸 것이다. 진한 회색이 분리 목적 음성, 연한 회색이 잡음 음성을 나타내고

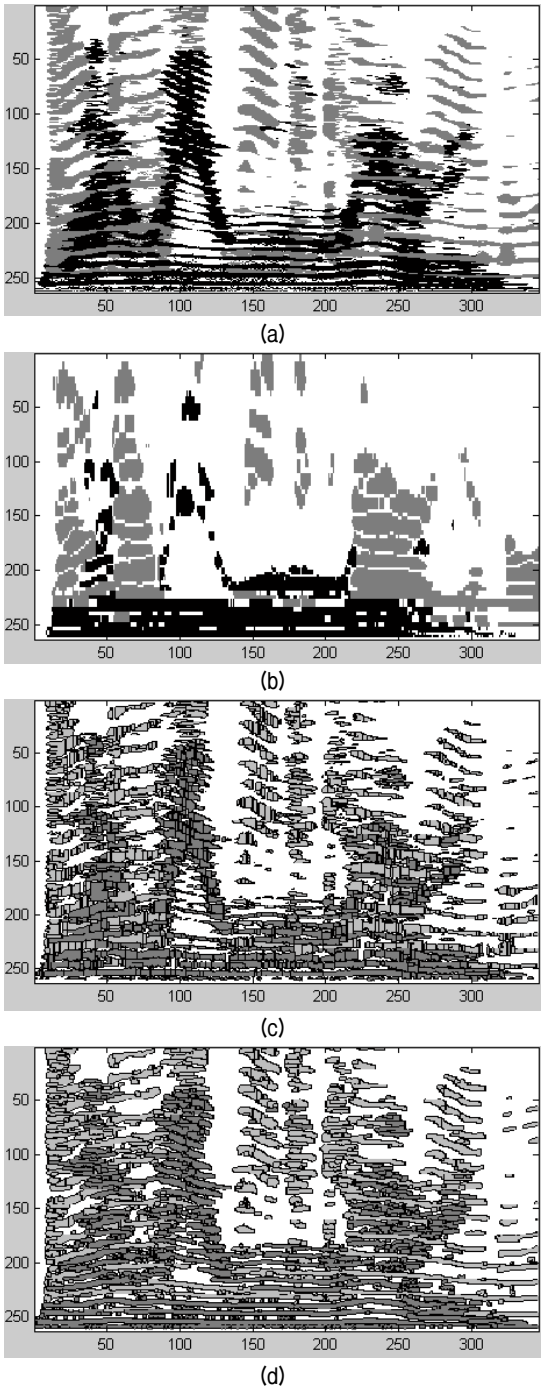


그림 7. 주파수 공간에서 분리된 음성 신호 (a) 이상적인 이진 마스크 (b) Wang-Brown 모델 (c) 골 기반 방법 (d) 마루 기반 방법

Fig. 7. Separated Results in Time-Frequency Domain. (a) ideal binary mask (b) Wang-Brown model (c) valley based model (d) peak based model

있다. Wang-Brown 모델의 경우, 주파수 해상도가 비등간격인 ERB Scale의 마스크를 등간격인 FFT Scale로 변환하여 표현한 것이다. ERB의 경우에는 저주파 영역에서 각 채널이 높은 빈도로 중첩되어 있고 해상도도 높기 때문에 각 세그먼트가 주파수 채널 방향으로 넓게 분포하고 있다. 골 기반의 결과는 골이 세그먼트의 경계를 결정하기 때문에 세그먼트의 끝부분이 비교적 날카로운 모양을 하고 있고, 마루 기반의 방법은 상대적인 진폭에서 임계값을 이용하여 경계를 결정하기 때문에 세그먼트 끝부분이 부드럽게 형성되고 있음을 알 수 있다. 또한 골의 연속성 보다는 마루의 연속성이 잡음이 섞인 경우에도 잘 보존되어, 전체적인 세그먼트가 시간 방향으로 오래 유지하고 있다는 특성을 가지고 있음을 확인 할 수 있다.

Ⅵ. 결 론

본 논문에서는 시간-주파수 영역에서 동일한 음원에서 발생한 인접한 주파수 채널을 분할하는 새로운 알고리즘을 제안하였다. 제안된 알고리즘은 평탄화된 스펙트로그램에서 마루와 골을 이용하여 분할 영역의중심과 경계를 결정하는 방법을 사용하고 있다. 제안된 알고리즘의 분할 결과는 골을 이용한 방법보다 마루를 이용한 방법이 세그먼트의 수가 적게 나오면서도 분리 성능이 높아 마루를 이용한 방법이 주파수 채널 분할에 더 적당함을 알 수 있었다. 또한, 이상적인 마스크를 이용한 그루핑 방법을 적용했을 때, 제안된 방법이 기존 모델보다 모든 잡음에 대하여 성능이 향상됨을 실험을 통하여 알 수 있었다.

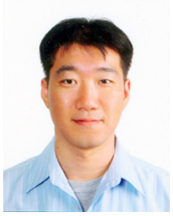
실험 결과에서 일부 혼합 신호에 대해서는 잡음과 음성의 마루가 겹침으로써 분할 성능을 감소시키고 있는데, 주파수 채널 분리에 마루와 골을 모두 사용함으로써 보다 정확하게 주파수 채널을 분할하는 문제는 향후 해결해야 할 과제이다. 본 논문에서 제안한 음성 분할 알고리즘은 음성 분리의 일부분으로 사용된다. 혼합된 음향 신호로부터 음성을 분리하기 위한 시스템을 구현하기 위해서는 분할된 세그먼트 단위로 특징을 추출하여 하나의 음원으로부터 발생한 세그먼트들을 하나로 그루핑 하는 방법에 대한 후속 연구가 필요하다.

참 고 문 헌

1. Walsh, J.M., Kim, Y.M., Doll, T.M., "Joint Iterative Multi-Speaker Identification and Source Separation using Expectation Propagation", ASPAA 2007, 283-286, 2007.
2. Mohammed, U.S., Mahmmoud, M.F., "A Blind Signal Separation Technique using Combination of Second-Order and Higher-Order Approaches", ICICT 06, 1-2, 2007.
3. Abdollahpouri, M., Khaki-Sedigh, A., Khaloozadeh, H., "A New Method for Active Noise Cancellation in the Presence of Three Unknown Moving Sources", AICMS 08, 1006-1011, 2008.
4. Bregman, A. S., "Auditory Scene Analysis : The Perceptual Organization of Sound, MIT Press", (1991).
5. Brown, G. J. and Wang, D. L. "The separation of speech from interfering sounds based on oscillatory correlation", Trans. on Neural Networks, **10**, I. 3, pp.684-697, 1999.
6. Jin, C., van Schaik, A., Carlile, S.,"The integration of acoustical cues during human sound localisation of band-pass filtered noise", ICONIP 1999, **2**, 483-488, 1999.
7. Chan, C.F., Yu, E.W.M., "Improving pitch estimation for efficient multiband excitation coding of speech", Electronics Letters, **32**, I. 10, pp.870-872, 1996.
8. Srinivasan, S.H. and Kankanhalli, M, "Harmonicity and dynamics based audio separation", ICASSP 03, **5**, V-640-3, 2003.
9. Jen-Tzung Chien, Bo-Cheng Chen, "A new independent component analysis for speech recognition and separation", Trans. on Audio, Speech and Language Processing, V. 14, I. 4, pp.1245-1254, 2006
10. Dan Ellis, "Computational Auditory Scene Analysis", Talk Slids, http://www.ee.columbia.edu /~dpwe/ talks/ oldenburg-casa-2005-06.pdf, 2005.
11. Cook, M.P., "Modeling Auditory Processing and Organization", Cambridge Univ. Press, 1993.
12. Hu, G. and Wang, D.L., "Monaural speech segregation based on pitch tracking and amplitude modulation", Trans. on Neural Networks, V. 15, I 5, pp.1135-1150, 2004.

저자 약력

•임 성 길 (Sung-Kil Lim)



1997년 2월 : 경희대학교 수학과 (이학사)
1999년 2월 : 경희대학교 전자계산공학과 (공학석사)
1999년 3월~현재: 경희대학교 컴퓨터공학과 (박사과정)
※주관심분야 : ASA (Auditory Scene Analysis), 패턴인식, 신경망

•이 현 수 (Hyon-Soo Lee)

1979년 2월 : 경희대학교 전자공학과 (공학사)
1982년 4월 : 일본 게이오대학원 전기공학과 (공학석사)
1985년 4월 : 일본 게이오대학원 전기공학과 (공학박사)
1999년 9월~2000년 8월 : 미국 오레곤 주립대학교 전기 및 컴퓨터공학과 방문연구원, 미국 캘리포니아대학교(U.C.I) 전기 및 컴퓨터공학과 방문연구원
1985년~현재 : 경희대학교 컴퓨터공학과 교수
2005년~현재 : 경희대학교 전자정보대학 학장 겸 정보통신대학원 원장
※주관심분야 : 컴퓨터구조 및 VLSI, 병렬처리, 패턴인식, 신경망, 음성처리