CASA 기반 음성분리 성능 향상을 위한 형태 분석 기술의 응용*

이윤경(충북대), 권오욱(충북대)

<차 례>

<Abstract>

Application of Shape Analysis Techniques for Improved CASA-Based Speech Separation

Yun-Kyung Lee, Oh-Wook Kwon

We propose a new method to apply shape analysis techniques to a computational auditory scene analysis (CASA)-based speech separation system. The conventional CASA-based speech separation system extracts speech signals from a mixture of speech and noise signals. In the proposed method, we complement the missing speech signals by applying the shape analysis techniques such as labelling and distance function. In the speech separation experiment, the proposed method improves signal-to-noise ratio by 6.6 dB. When the proposed method is used as a front-end of speech recognizers, it improves recognition accuracy by 22% for the speech-shaped stationary noise condition and 7.2% for the two-talker noise condition at the target-to-masker ratio lower than or equal to −3 dB.

* Keywords: Computational auditory scene analysis, Speech separation.

* 본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2007-S-036-01, 신성장동력산업용 대용량/대화형 분산/내장처리 음성인터페이스기술 개발]

1. 서 론

사람은 아무리 많은 목소리가 섞여서 들려도 관심 있는 소리를 들을 수 있다. 이를 칵테일 파티 문제(cocktail party problem)라고 하며 2개 이상의 음원이 동시에 제시될 때 주의를 기울인 음원만을 선택적으로 청취할 수 있는 청각적인 성질을 말한다. 이를 해결하기 위하여 2개 이상의 음성 또는 잡음이 혼합되어 있는 신호에서 원하는 음성신호만을 추출하는 분리방법이 넓은 범위에서 연구되어 왔다. 특히 최근에는 동적인 잡음이 존재하는 환경에서의 강인한 음성인식을 위하여 이러한 음성분리 기술을 응용하고자 하는 연구가 진행되고 있다[1]-[3].

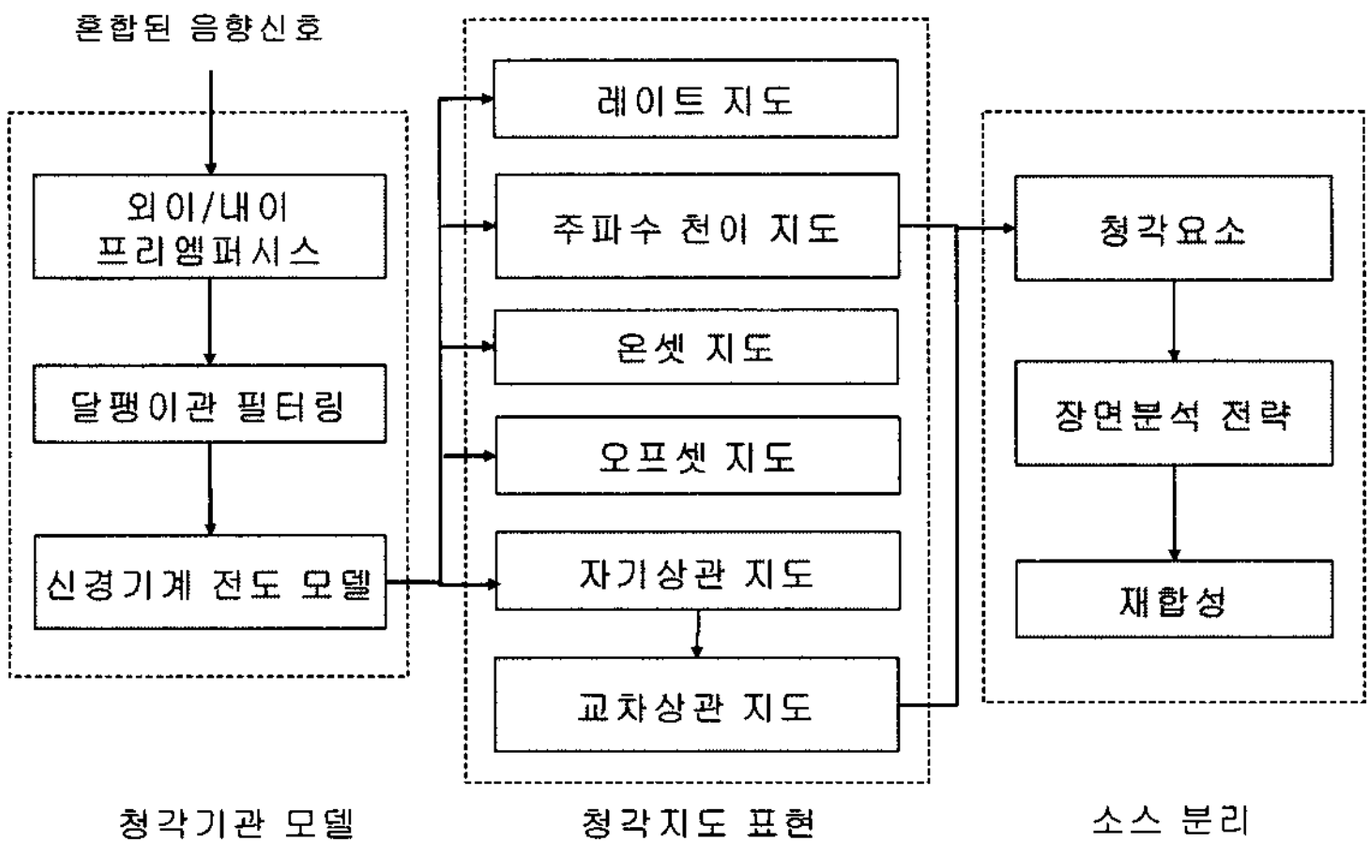
전산 청각 장면 분석(computational auditory scene analysis: CASA)는 귀에 도달한 혼합신호로부터 인간의 청각특성을 이용하여 음성신호를 분리하는 기술이다. 음성신호를 시간-주파수 영역으로 변환하고, 피치, 자기상관(autocorrelation), 온셋/오프셋(onset/offset) 특성을 이용하여 동일 음원으로부터 발생한 음향요소들을 찾아내는 방법이다[1]. 이 방법은 음성학적 지식과 휴리스틱이 요구되는 단점이 있지만, 단일 채널의 입력신호에 대하여 적용할 수 있다는 장점이 있다. 최근 CASA 알고리듬을 음성인식 전처리에 적용하여 음성인식률 향상에 응용하는 연구가 진행되고 있다. Runqiang의 알고리듬은 청각 특징과 피치 궤적을 이용하여 음성과 잡음 신호를 분리한 후 스펙트럼에서의 음성의 손실을 특징에 기초한 클러스터의 방법을 사용하여 보완한 후 음성인식을 수행한다[4]. Srinivasan의 알고리듬은 동시 그루핑(simultaneous grouping)과 순차 그루핑(sequential grouping)의 2 단계로 구성되어 있으며 각 과정에서 주기 정보를 이용하여 유성음을 분리하고 온셋/오프셋 정보를 이용하여 무성음을 분리한다[5]. 하지만 기존의 방법들은 음성분리를 수행하는 과정에서 신호의 끊어짐으로 인한 음성신호의 손실이 있다. CASA 알고리듬을 사용한 음성 분리에서는 피치 궤적에 따라 입력 신호를 잡음, 음성신호의 그룹으로 분리한다. 따라서 손실된 음성신호가 많을수록 피치 궤적을 구하기가 어려워 음성분리의 성능이 떨어지는 경우가 종종 있다.

본 논문에서는 이러한 손실된 음성 신호를 보완하기 위해 입력신호로부터 영상처리의 형태 분석(shape analysis) 기술[6][7]의 하나인 레이블링(labelling)과 거리함수(distance function)를 사용하여 계산된 각 청각 요소 그룹의 수와 두께 정보를 이용하여 손실된 음성을 보완한다.

본 논문은 2장에서 기존의 CASA 알고리듬을 간략히 소개하고 3장에서 개선된 알고리듬에 대하여 설명하였다. 4장에서 음성분리 실험결과 및 Interspeech 음성데이터를 이용한 음성인식 결과를 제시하고 5장에서 결론을 맺는다.

2. CASA 알고리듬

CASA 기반 음성분리 알고리듬의 구성도는 <그림 1>과 같으며 세 단계로 이루어져 있다[1]. 첫 번째 단계는 청각의 바깥둘레인 사람의 귀(외이/중이)와 달팽이관, 신경의 특징을 필터뱅크와 내유모세포(inner hair cell)의 출력으로 모델링하는 것이다. 두 번째 단계에서는 청각 신경의 주기 정보, 주파수 전달 특징, 온셋/오프셋 정보를 추출한다. 마지막 단계는 추출된 청각 특징들을 이용하여 청각 신호를 상징적으로 묘사하고 피치 궤적을 이용하여 잡음과 음성 그룹으로 분리하는 단계이다. 분리된 음성 그룹은 청취 또는 음성인식기 입력 용도를 위하여 시간영역 파형으로 재합성된다[1].



<그림 1> CASA 기반 음성분리 알고리듬

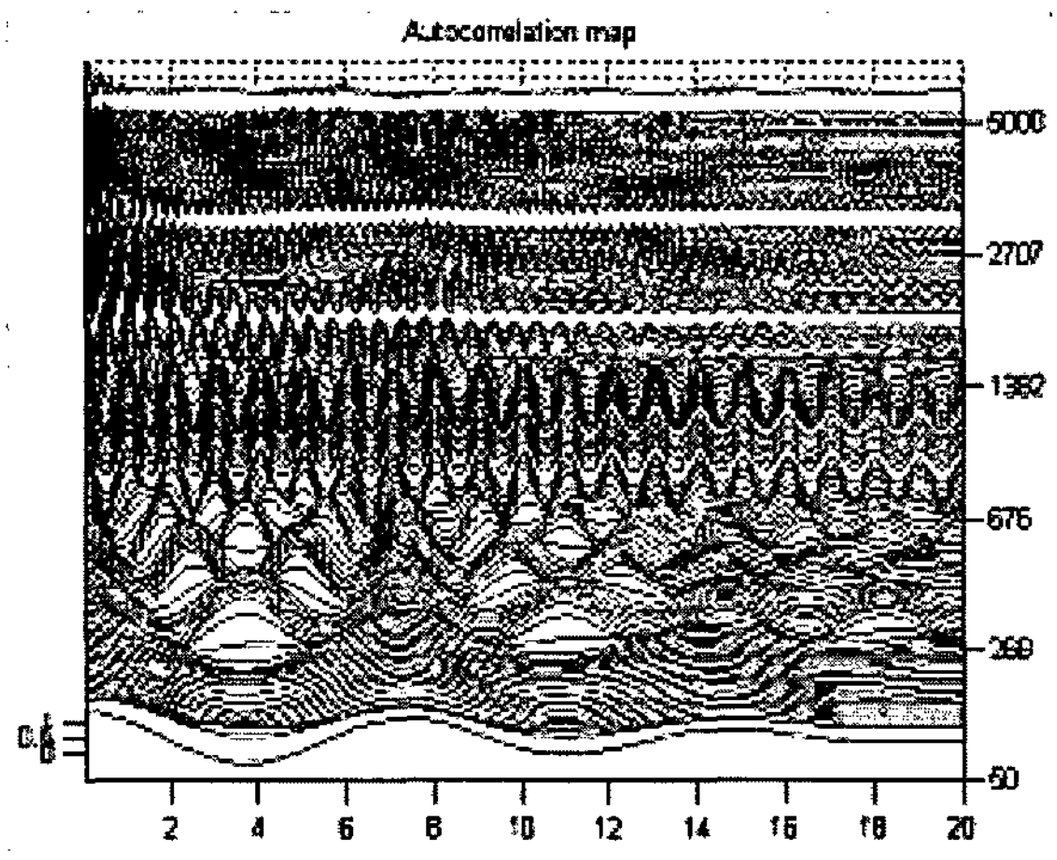
2.1. 청각 기관 모델 (Auditory periphery model)

청각의 외부기관에 해당하는 외이와 중이, 달팽이관과 신경의 특성을 계산한다. 외이와 중이에서 작은 소리에서부터 적당한 소리까지의 강도가 거의 선형으로 되어 있기 때문에 간단한 선형 필터로 모델링할 수 있다. 주파수 전달 특징은 필터뱅크에 의해 설계되며 각 필터응답은 달팽이관을 따라서 특정한 구간의 주파수 응답을 모사한다. 필터뱅크는 128 채널을 가진 감마 톤 필터로 구성되며 채널은 50~5000 Hz를 등가 사각 대역폭(equivalent rectangular bandwidth: ERB) 비율에 따라 나누어 정한다[1]. 각 필터의 출력을 입력신호로 하여, 청각 신경 섬유를 흉내낸 메디스 모델(Meddis model)[1]을 사용하여 신경의 특성을 묘사한다.

2.2. 고차 청각 표현 (Higher auditory representations)

2.2.1. 자기상관 지도 (Autocorrelation map)

음성신호의 주기, 주파수 전달 특성, 온셋과 오프셋 등의 청각 신호의 특성 정보를 구한다. 자기상관 지도는 각 필터에서의 응답의 자기상관을 계산하여 128개의 각 채널에서 계산된 신경패턴의 주기에 관한 정보를 한꺼번에 출력한 것으로 피크들이 겹쳐있는 부분이 피치 주기에 해당한다[1]. <그림 2>는 20 ms의 해밍 윈도우를 씌워 10 ms 간격으로 계산한 자기상관 지도의 출력 예로서, 약 7.5 ms의 피치 주기를 가짐을 볼 수 있다.



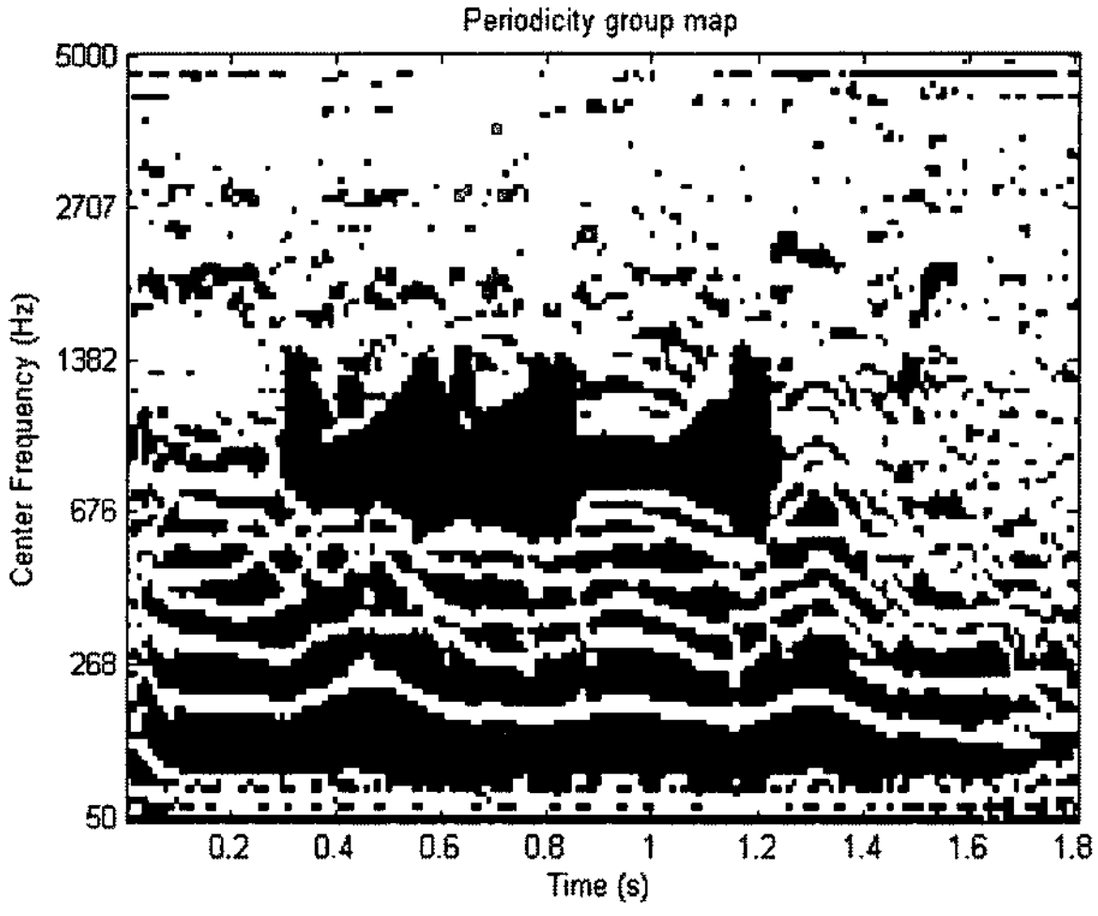
<그림 2> 자기상관 지도 출력

2.2.2. 교차상관 지도 (Cross-correlation map)

자기상관의 각 출력을 입력으로 하여 두 채널의 유사한 정도를 계산하며 주기의 패턴의 유사성의 정보를 확인한다. 각 채널의 중심주파수 f_1, f_2에서의 유사도는 다음과 같이 정의된다[1].

$$sim(f_1,f_2,t)=\frac{2\sum_{\Delta t}a_n(t,f_1,\Delta t)a_n(t,f_2,\Delta t)}{\sum_{\Delta t}a_n(t,f_1,\Delta t)^2+\sum_{\Delta t}a_n(t,f_2,\Delta t)^2} \quad (1)$$

여기서 $a_n(t,f,\Delta t)$는 자기상관의 응답이며, 계산된 유사도 결과는 정규화된 응답으로 0~1 사이의 값을 갖는다. 이웃한 두 채널의 주기의 패턴이 비슷할수록 유사도 결과는 1에 가까운 값을 가지며 상관성이 없을수록 0에 가까운 값을 가지기



<그림 3> 교차상관 지도 출력

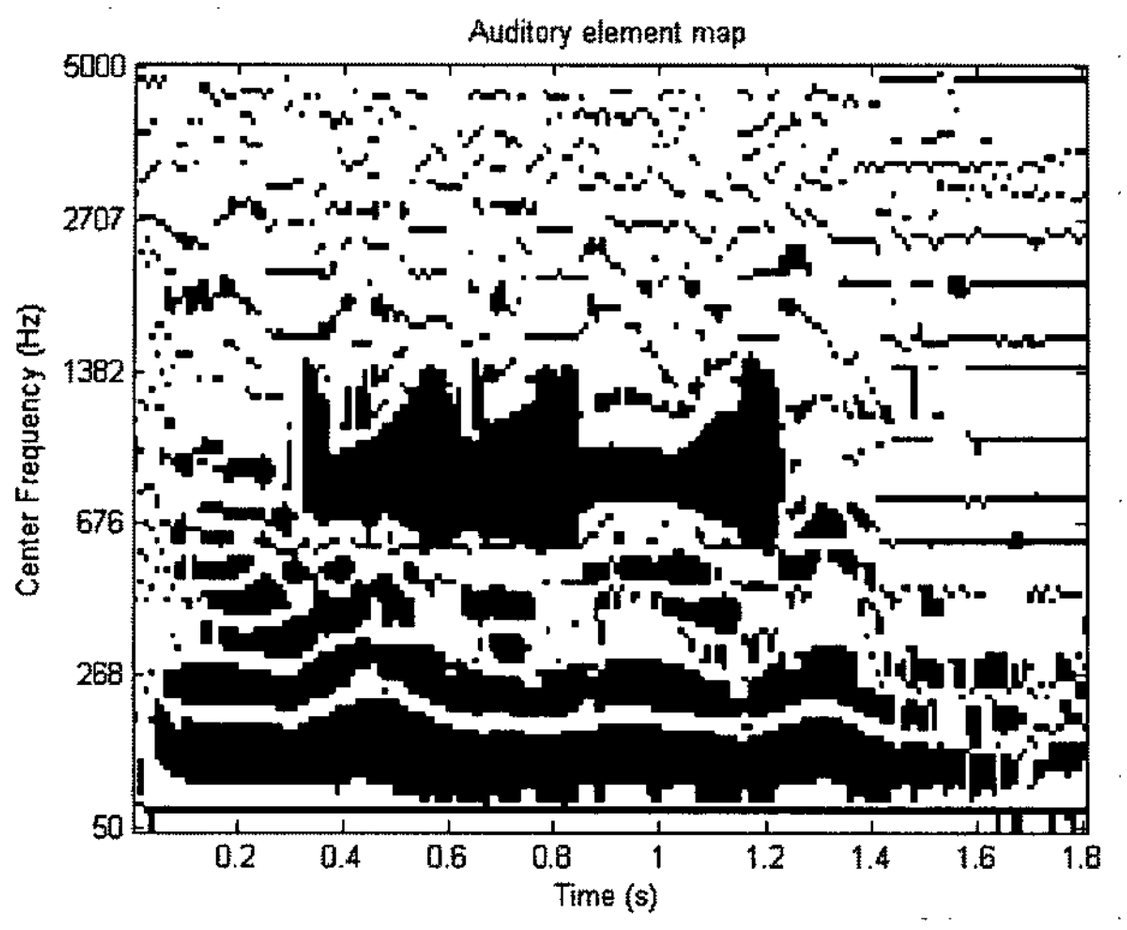
때문에 교차상관 지도에서 비슷한 특징을 가지는 채널의 응답이 검은색으로 표현되는 청각 요소의 그룹으로 나타난다. <그림 3>은 1 kHz 톤 잡음이 더해진 음성 데이터의 교차상관 지도 출력을 나타낸다.

2.3. 청각 요소 (Auditory elements)

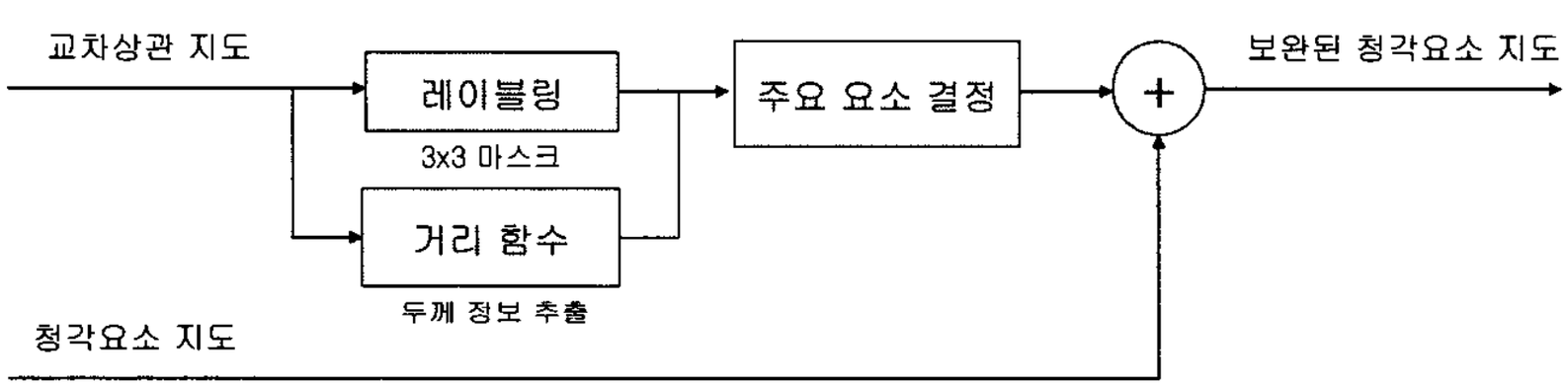
청각 특징 표현들을 조합하여 음성신호와 잡음신호를 분리하고 각 그룹을 재합성, 복원한다. 청각 요소 지도(auditory element map)는 주파수 천이 지도(frequency transition map)와 교차상관 지도의 조합으로 구해진다[1]. 여기서, 주파수 천이 지도는 피크의 유무와 방향성의 정보를 표현한 것으로 신경의 패턴 응답으로부터 계산되며 스펙트럼의 피크가 일직선이 됐을 때 최대 응답을 갖는다. <그림 4>는 1 kHz 톤 잡음이 더해진 음성 데이터의 청각 요소 지도를 보여준다.

피치궤적을 사용하여 음성신호와 잡음신호를 분리한다. 음성신호를 판별하는 기준으로 사용하는 피치 궤적은 자기상관의 가중 합으로 계산된 피크 값으로 구한다. 좀더 정확한 피치 궤적을 구하기 위해 각 시간프레임에서의 부분적인 가중합 자기상관을 계산하여 자기상관의 가중 합으로 구한 피크 값과 조합한다. 최적의 피치 궤적은 피치 궤적을 계산하는 과정에서 끊어진 부분을 이전의 피치 궤적을 참조하여 보완하는 동적 프로그래밍을 사용하여 구한다.

피크를 계산하는 과정에서 발생하는 피크오류는 지연시간을 길게 줌으로써 감소시킬 수 있다. 또, 피크 값에 임계값을 주어 임계값 이상의 값으로 계산되는 피크들만을 구하여 사용함으로써 작은 값임에도 불구하고 피크로 계산되는 부분을 제거하였다.



<그림 4> 청각 요소 지도 출력



<그림 5> 제안된 음성신호 보완 알고리듬

최적의 피치 궤적과 각 프레임 간의 피치 궤적의 유사도를 계산하여 유사도가 0.9 이상인 경우 같은 그룹으로 판정하는 방법으로 입력 신호를 음성신호와 잡음신호로 분리한다.

3. 개선된 알고리듬

CASA의 방법으로 청각 요소 지도를 구하는 과정에서 피치정보에 따라 해당하는 프레임의 모든 정보가 손실되기 때문에 필요한 음성 부분임에도 불구하고 함께 끊어져 정보를 얻고 신호를 분리하는 것이 어려워진다. 이러한 점을 보완하기 위하여 본 논문에서는 <그림 5>와 같이 음성의 주요 부분이라고 생각되는 청각 요소(element)를 입력 신호에서부터 구하여 청각 요소 지도와 조합하여 손실된 음성신호를 보완하였다. 음성의 주요 청각 요소는 영상처리 기법의 하나인 레이블링(labelling)과 거리 함수(distance function)를 사용하여 추출된 각 청각 요소 그룹의

수와 두께 정보를 이용하여 계산하였다. 잡음을 음성의 주요 부분으로 잘못 계산할 수 있는 것을 방지하기 위해 계산된 결과가 음성의 주파수에 비해 높은 주파수 대역에 있다면 결과를 배제시켜 음성 부분을 위주로 하여 청각 요소를 보완하였다. 보완된 출력 청각요소 지도는 장면분석전략 모듈로 입력된다.

3.1. 음성신호 보완 알고리듬

3.1.1. 레이블링 (Labelling)

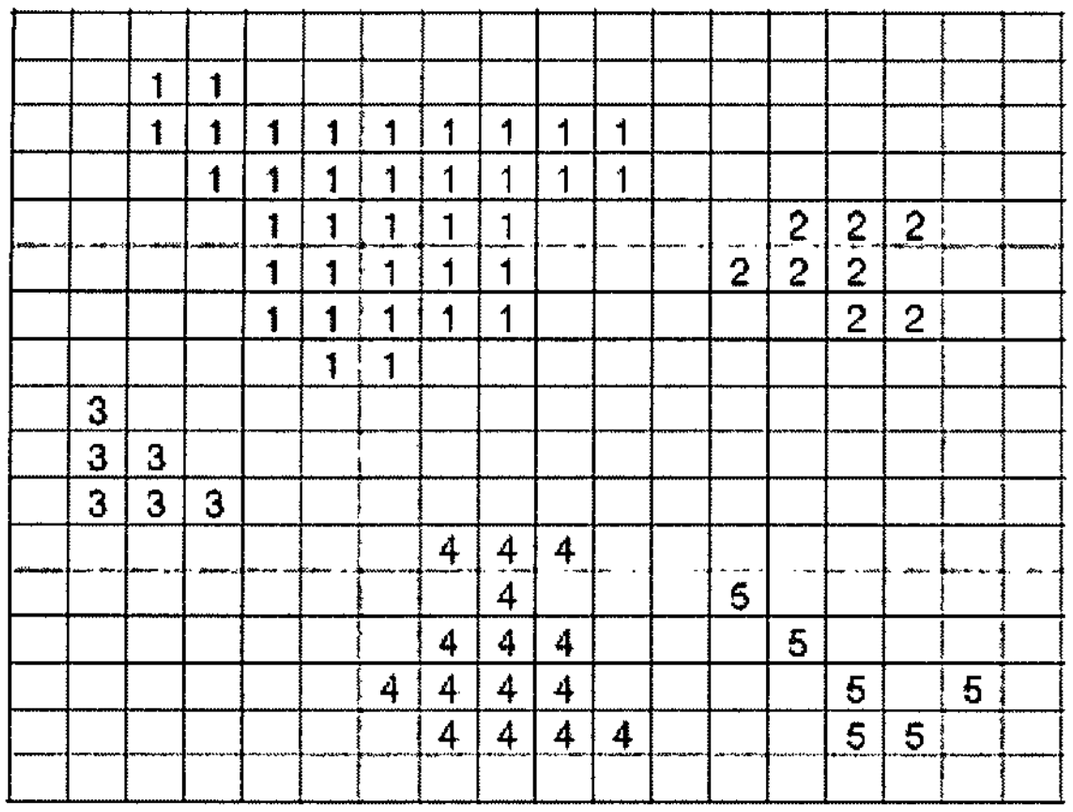
교차상관 지도에서의 출력을 입력으로 하여 각 청각 요소를 따로 떼어 사용할 수 있고 구분할 수 있도록 순서대로 번호를 붙인다. 청각 요소별로 총 수와 위치하고 있는 주파수 대역을 구하여 비중치 정보와 주파수 정보를 확인할 수 있도록 하였다. 레이블링을 구하는 방법은 다음을 따른다.

S1. 교차상관 지도의 모든 청각 요소를 순서대로 읽으면서 레이블링되지 않은 청각 요소를 찾는다. 이 때, 'visited elements'를 체크하여 만약 레이블링이 되었거나 검사를 한 청각 요소라면 건너 뛰어 다음 청각 요소를 검사한다.
S2. 찾은 청각 요소와 연결된 청각 요소들을 찾아 같은 번호로 레이블링한다. 현재 위치의 청각 요소를 기준으로 하여 3x3 마스크를 씌웠을 때 주위에 청각 요소가 있다면 같은 그룹으로 판단하는 방법으로 연결된 청각 요소를 구한다.
S3. 모든 그룹이 찾아질 때까지 S2를 반복한다.
S4. 만약 연결된 모든 청각 요소를 찾아 레이블링하였다면 찾는 과정을 끝낸다.
S5. 레이블링의 숫자 정보를 1 증가시킨다.
S6. S1으로 돌아가 다음 청각 요소 그룹을 찾는다.

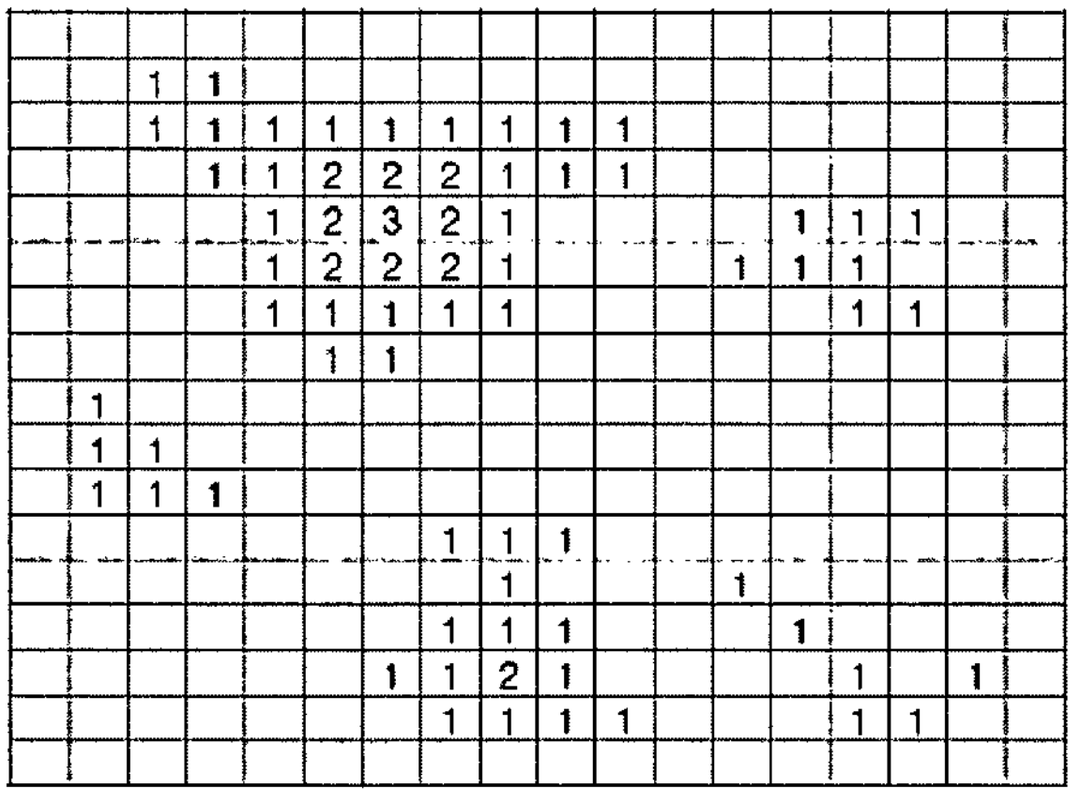
<그림 6>에 5개의 레이블링 그룹을 가지는 예를 나타내었다.

3.1.2. 거리 함수 (Distance function)

거리 함수는 각 청각 요소 그룹의 두께정보를 알 수 있는 것으로, 바깥쪽으로부터 어느 정도의 거리가 있는지를 계산한다. 청각 요소를 찾은 후 검출 되어진 청각 요소에 3x3 마스크를 씌웠을 때 마스크 안에 청각 요소가 모두 존재하고 모든 청각 요소가 같은 두께 값을 가지고 있다면 두께 정보를 1 증가시키고, 그렇지 않다면 현재의 두께 정보를 유지하는 방법으로 구한다. 이 과정을 최대 거리의 정



<그림 6> 레이블링 출력 예



<그림 7> 거리 함수 출력 예

보가 더 이상 변하지 않을 때까지 반복한다. 바깥쪽으로부터의 거리는 각 그룹에서 다음과 같이 표현된다.

0: 배경(비활성 요소)
1: 모서리 청각 요소
2: 거리 1인 청각 요소(모서리 요소)에 인접한 청각 요소
3: 거리 2인 청각 요소에 인접한 청각 요소
...

거리 함수의 예를 다음 <그림 7>에 나타내었다. 여기서, 첫 번째 그룹의 최대 거리는 3이다.

3.1.3. 음성신호의 주요 요소

레이블링과 거리 함수를 이용하여 계산된 각 청각 요소의 수와 두께 정보를 이용하여 음성신호의 주요 부분을 계산한다. 충분한 수와 충분한 두께를 가지고 있는 청각 요소를 분리하여 음성신호의 주요 부분으로 계산하며 만약 계산된 그룹이 고주파수 대역에 위치하고 있다면, 결과를 배제하였다. 본 논문에서는 청각 요소의 수의 임계값을 100, 두께 정보의 임계값을 3으로 정하여 사용하였으며 주파수 채널은 약 700 Hz에 해당하는 50 채널 이상의 주파수 대역을 고주파수 대역으로 정하여 적용하였다. 최종적으로 음성신호의 주요 부분을 계산한 후, 청각 요소 지도를 계산된 결과와 논리합(OR) 연산을 사용하여 보완하였다.

3.2. 보완된 음성신호의 예

3.2.1. 음성신호의 주요 부분 출력

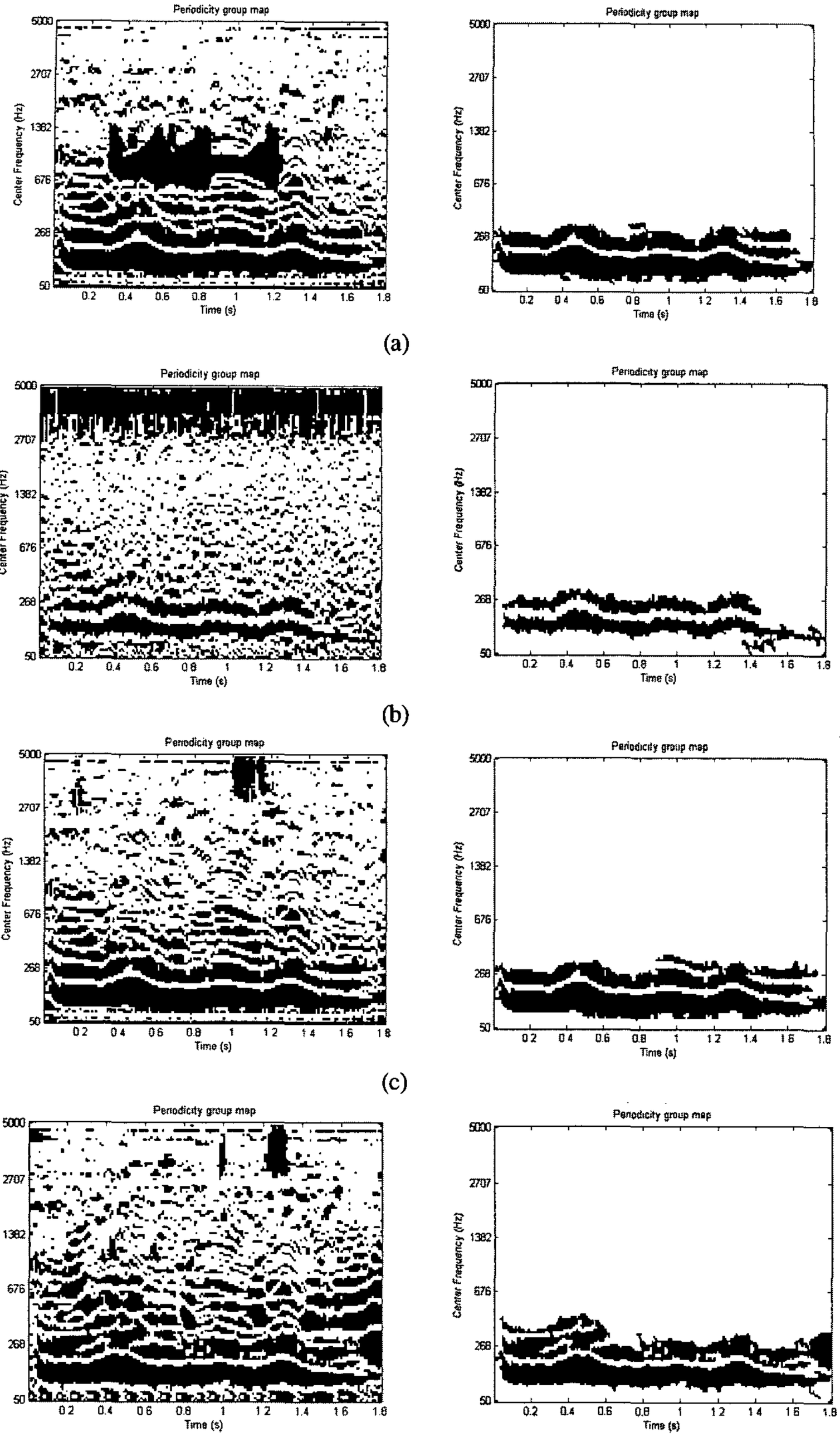
<그림 8>에 교차 상관 지도에서의 출력과 계산된 음성신호의 주요 부분의 결과를 나타내었다. 음성 데이터는 1 kHz 톤 신호와 랜덤 잡음, 그리고 남성화자의 음성신호를 잡음신호로 하여 음성신호에 첨가된 데이터로 출력된 청각 요소를 분석한 결과 음성신호에 해당하는 주요 부분이 거의 출력되었으며 잡음신호는 음성신호와 청각 요소에 의해 연결된 경우를 제외하고는 출력되지 않았다.

3.2. 보완된 청각 요소 지도 출력

<그림 9>에 음성신호가 손실되어 출력된 원래의 청각 요소 지도의 출력과 음성신호를 보완한 청각 요소 지도의 출력을 나타내었다. 원래의 청각 요소 지도의 출력은 100~300 Hz의 주파수 대역에 위치하고 있는 음성 신호 청각 요소가 끊어진 부분에 의해 신호의 손실이 있는 반면 보완된 청각 요소 지도의 출력은 음성신호의 손실 없이 출력되었다.

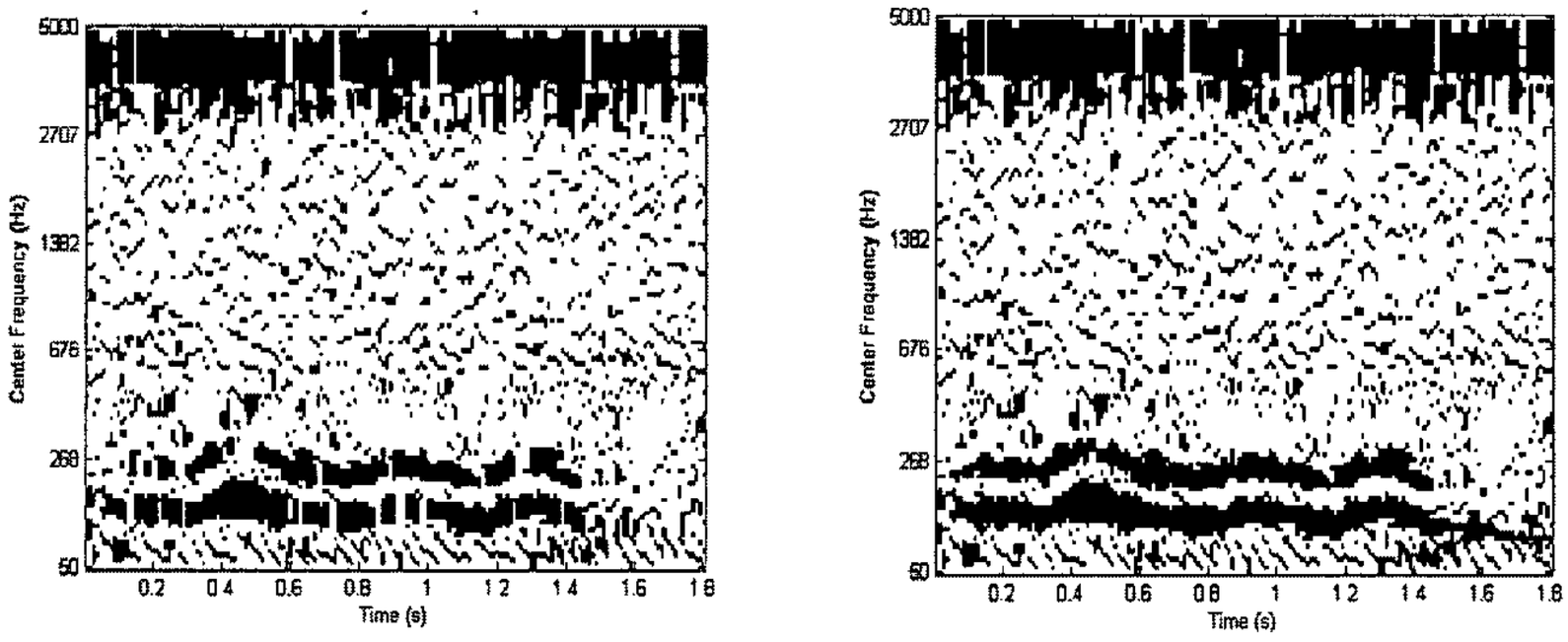
4. 실험결과

음성분리의 성능의 정도를 측정하기 위하여 잡음 혼합 음성데이터[8]를 사용하여 음성분리를 수행하는 것과 공개된 음성 데이터베이스[9]를 분리하여 음성 인식률을 측정해 보는 두 가지의 실험을 수행하였다.



(a) (b) (c) (d)

그림 8> 음성신호의 주요 부분 출력: (a) 1 kHz 톤 신호, (b) 랜덤 잡음(random noise
(c) 남성화자의 음성신호, (d) 여성화자의 음성신호



(a) (b)

<그림 9> 보완된 청각 요소 지도 출력: (a) 청각 요소 지도 출력, (b) 보완된 청각 요소 지도 출력

4.1. 음성 분리 실험 결과

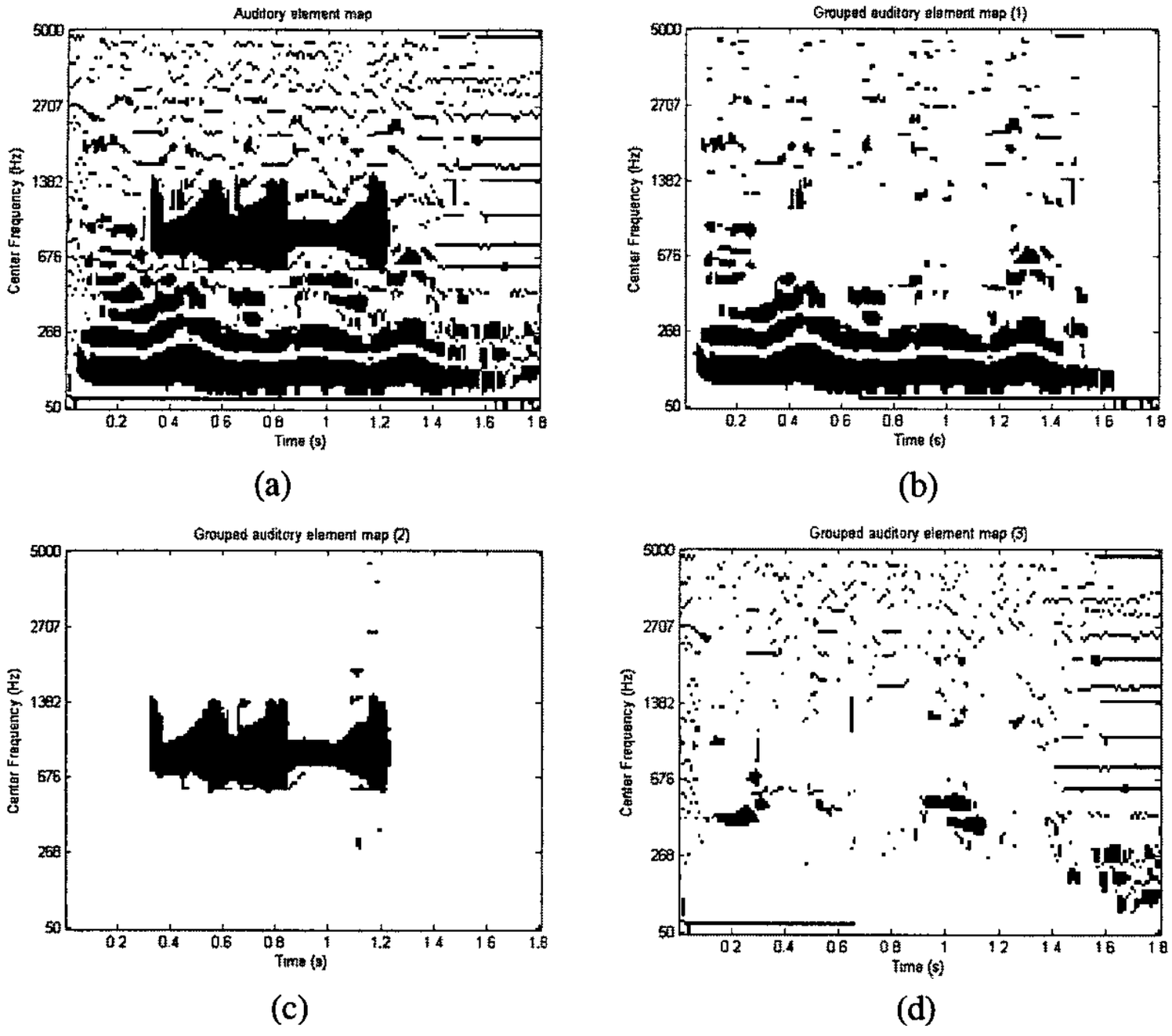
음성을 분리하기 위해 사용된 음성데이터는 Cooke가 제공하는 10개의 문장(2명의 화자에 의해 발음되는 5문장)의 각각에 10 가지의 잡음이 섞인 100개의 음성데이터이다. 잡음신호는 1 kHz의 톤 신호, 랜덤 잡음(random noise), 버스트 잡음(bursts noise), 록 뮤직(rock music), 사이렌 소리, 전화벨 소리, 여성 화자와 남성 화자의 음성신호로 구성되며, <표 1>에서 n0 ~ n9로 표기하였다. 음성과 잡음의 분류된 정도를 확인하기 위하여 음성신호, 잡음신호, 배경잡음의 세 그룹으로 분리하여 각 청각 요소 지도를 출력하였다. <그림 10>은 1 kHz 톤이 더해진 신호를 분리한 결과로서, 음성 분리를 수행한 후 음성신호를 청취하였을 때에도 잡음은 거의 들리지 않았다.

음성 분리의 결과를 수치적으로 보기 위하여 음성신호의 분리 전과 분리한 후의 신호대잡음비(SNR)를 계산하여 비교하였다.

$$SNR = 10\log_{10}\left[\frac{\sum_{t=0}^{T-1} s^2(t)}{\sum_{t=0}^{T-1}(r(t)-s(t))^2}\right] \quad (2)$$

여기서 $s(t)$는 원래의 음성신호, $r(t)$는 복원된 음성신호를 나타낸다.

각 잡음에 대하여 분리 전후의 SNR을 비교한 결과, <표 1>과 같이 기존의 CASA 알고리듬으로 분리하였을 때 약 2.8 dB의 증가가 있었으며 제안된 알고리듬은 약 6.6 dB의 증가가 있었다. 제안된 알고리듬을 사용하여 음성 분리를 수행한 후의 SNR을 분리하기 전의 SNR과 비교하였을 때 전체적으로 음성신호가 증가



<그림 10> 청각 요소 그룹 출력: (a) 입력 신호, (b) 추출된 음성신호, (c) 잡음 신호, (d) 배경 잡음

하였음을 볼 수 있었으며, 제안된 알고리듬은 특히 1 kHz 톤 신호와 버스트 잡음이 더해진 음성데이터에서 효과적이었다.

<표 1> 잡음 종류에 따른 신호대잡음비 결과 (dB)

잡음신호 종류	원 음성	CASA 결과	제안 방법 결과
n0	0.5	2.7	8.9
n1	-2.9	1.8	2.5
n2	2.1	3.3	16.6
n3	0.8	2.6	6.8
n4	0.5	4.4	6.8
n5	-4.8	2.8	3.8
n6	1.6	3.6	4.2
n7	2.9	4.4	4.9
n8	3.2	5.1	9.0
n9	0.6	2.6	6.9
평균	0.5	3.3	7.0

4.2. 음성인식 실험 결과

음성 인식률을 측정하기 위하여 음성분리대회(speech separation challenge)[9]에서 제공하는 음성형태 잡음(speech-shaped noise: SSN)와 경쟁 화자의 신호가 원하는 음성 신호와 혼합되어 있는 음성 데이터베이스[9]를 사용하여 음성을 분리한 후, 분리된 음성 데이터를 음성입력기에 입력하여 인식률을 계산하였다. 학습과 테스트를 위한 음성 데이터는 단일 채널 입력의 데이터로 25 kHz에서 샘플링 되었으며 학습 데이터는 34명의 남성, 여성 화자로부터의 17,000 문장으로 이루어져 있다. 음성은 GRID 형식으로 6 가지 단어로 표현되어 있으며 다음과 같이 구성되어 있다.

<command:4> <color:4> <preposition:4> <letter:25> <number:10> <adverb:4>

각 단어의 구성은 다음과 같다.

command : bin, lay, place, set
color : blue, green, red, white
preposition : at, by, in, with
letter : W를 제외한 A부터 Z까지의 알파벳
number : 0, 1부터 9까지의 숫자
adverb : again, now, please, soon

4.2.1. 음성 형태 잡음 데이터 실험 결과

SSN 데이터는 정적 잡음(stationary noise)이 첨가된 음성신호로 SNR은 Clean과 6, 0, -6, -12 dB로 구성되어 있다. <표 2>에 음성 인식률을 나타내었으며 인식률을 측정한 결과 6 dB에서 약 20%의 증가를 보였고, 0 dB와 -6 dB에서 약 25%, 그리고 -12 dB에서 약 15%의 증가를 보였다. 이러한 결과는 SNR이 높을수록, 그리고 음성신호와 유사성이 떨어지는 정적 잡음과 혼합될수록 분리가 잘 된다는 것을 나타낸다.

<표 2> 정적 잡음의 음성인식 결과

SNR(dB)	원 음성	제안 방법
Clean	93.9	87.4
6	29.5	50.3
0	16.2	41.1
-6	12.5	35.5
-12	13.0	28.4

<표 3> 분리 전 음성 인식률 (%) 결과

TMR (dB)	ST	SG	DG	Avg.
6	67.0	65.0	61.5	64.5
3	54.0	53.0	42.0	49.7
0	37.0	38.5	32.0	35.8
-3	21.0	25.0	23.0	23.0
-6	14.0	16.0	18.0	16.0
-9	10.5	12.0	12.0	11.5

<표 4> 제안된 알고리듬으로 분리 후 음성 인식률 (%) 결과

TMR (dB)	ST	SG	DG	Avg.
6	33.3	34.6	41.2	35.3
3	26.7	31.5	35.2	31.1
0	21.0	29.2	31.4	27.2
-3	15.4	28.3	34.5	26.3
-6	19.6	25.5	33.7	26.3
-9	14.2	20.2	23.7	19.4

4.2.2. 2화자 문장 데이터 실험 결과

2화자(Two-talker) 테스트 데이터는 6, 3, 0, -3, -6, 그리고 -9 dB에서의 6개의 목표대마스커비(target-to-masker ratio: TMR)를 가진다. 두 화자 문장 데이터는 경쟁 음성의 종류에 따라서 Same Talker (ST), Same Gender (SG), Different Gender (DG)로 구분된다. 여기서 ST는 같은 화자의 두 음성이 음성과 잡음신호로 사용된 경우, SG는 같은 성별의 두 화자의 음성이 음성과 잡음신호로 사용된 경우, 그리고 DG는 다른 성별의 두 화자의 음성이 음성과 잡음신호로 사용된 경우이다. <표 3>과 <표 4>에 원래의 음성에 대한 음성 인식률과 음성분리 후의 음성 인식률을 나타내었다. CASA 기반의 음성 분리 과정에서 청각 요소 그룹은 하나의 청각 요소를 통하여 연결되어 있다고 하더라도 같은 그룹으로 계산이 된다. 또한 음성을 분리하는 과정에서 다른 그룹으로 떨어지거나 따로 계산되지 않기 때문에 두 화자 문장 데이터의 인식률은 음성과 잡음의 유사도가 작은 단일톤이나 정적 잡음에 비하여 취약하였다. 음성분리 결과를 살펴보면 0 dB 이상의 TMR에서 제안한 방법을 사용하여 측정한 인식률이 분리 전의 인식률보다 낮게 측정되었다. 이는 음성을 분리하는 과정에서 음성신호, 특히 무성음의 스펙트럼이 손실되기 때문이다. 이러한 환경에서는 낮은 TMR의 음성분리 성능이 더 좋으며 본 논문에서는 -3 dB

이하의 TMR에서 약 7.2%의 인식률 증가가 있었다. CASA를 기반으로 하는 다른 논문의 연구 결과[4][5]에서도 TMR이 낮을수록 음성분리의 성능 개선 정도가 높게 측정되었다. 2화자 음성데이터베이스에서의 인식률 증가를 위해 향후 통계적 모델을 사용한 손실된 무성음 보완과 멀티피치를 이용한 피치궤적 계산의 방법이 필요하다. 전체적으로 음성과 잡음 신호로 사용된 화자의 음성이 비슷한 신호일수록 분리가 어려웠으며 이성(DG) 화자의 혼합 음성이 동일화자(ST) 또는 동성(SG) 화자의 혼합 음성보다 비교적 잘 분리되었다.

5. 결 론

본 논문에서는 음성인식기의 전처리기로 사용하기 위한 CASA 기반 음성분리기의 성능향상을 위하여 형태 분석 기술을 응용하였다. 제시한 방법에서는 레이블링과 거리 함수를 사용하여 음성 분리 과정에서 손실되는 음성 부분을 보완하였다. 음성신호 분리 성능을 평가하기 위하여 인공 혼합신호의 음성 분리 실험과 음성 인식 응용 실험을 수행하였다. Cooke의 음성데이터를 사용한 음성 분리 실험에서 음성의 분리 전과 분리 후의 SNR을 비교한 결과 6.6 dB의 SNR 증가가 있었다. 음성인식 응용 실험에서는 Interspeech의 정적 잡음 및 2화자 음성데이터베이스를 사용하여 인식률을 측정한 결과 정적 잡음의 경우 인식률이 약 22% 증가하였으며 2화자 음성데이터베이스의 인식률은 −3 dB 이하의 TMR에서 약 7.2% 증가하였다. 향후 음성 인식률을 높이기 위하여 교차 상관 지도와 청각 요소 지도를 절충하여 필요 없는 잡음신호를 정리하고 음성신호를 끊어지지 않도록 보완하고 피치 궤적을 보완하여 손실되는 음성을 최소화하는 것이 필요하다.

참 고 문 헌

[1] G. J. Brown, M. Cooke, "Computational auditory scene analysis", *Computer Speech and Language*, Vol. 8, No. 4, pp. 297-326, 1994.

[2] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system", *Proc. Interspeech*, pp. 1775-1778, 2006.

[3] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space", *Proc. Interspeech*, pp. 1850-1853, 2006.

[4] H. Runqiang, Z. Pei, G. Qin, Q. Zhiping, W. Hao, W. Xihong, "CASA based speech separation for robust speech recognition", *Proc. Interspeech*, pp. 2068-2071, 2006.

[5] S. Srinivasan, Y. Shao, Z. Jin, D. Wang, "A computational auditory scene analysis system for robust speech recognition", *Proc. Interspeech*, pp. 1547-1550, 2006.

[6] E. R. Davies, *Machine Vision*, Elsevier, pp. 158-193, 2005.
[7] A. McAndrew, *Introduction to Digital Image Processing*, Course Technology, 2004.
[8] http://www.dcs.shef.ac.uk/~martin/.
[9] M. Cooke, T.-W. Lee. *Speech Separation and Recognition Competition,* Avai http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

접수일자: 2008년 2월 25일
게재결정: 2008년 3월 24일

▶ 이윤경(Yun-Kyung Lee)
주소: 361-763 충북 청주시 흥덕구 성봉로 410 (개신동)
소속: 충북대학교 제어계측공학과
전화: 043)261-3374
E-mail: yklee@cbnu.ac.kr

▶ 권오욱(Oh-Wook Kwon) : 교신저자
주소: 361-763 충북 청주시 흥덕구 성봉로 410 (개신동)
소속: 충북대학교 전기전자컴퓨터공학부
전화: 043)261-3374
E-mail: owkwon@cbnu.ac.kr