일반논문 (Regular Paper)
방송공학회논문지 제17권 제6호, 2012년 11월 (JBE Vol. 17, No. 6, November 2012)
http://dx.doi.org/10.5909/JBE.2012.17.6.1061
ISSN 1226-7953(Print)

# 음성 명료도 향상을 위한 학습 기반의 신호 대 잡음 비 추정을 이용한 이산 마스크 추정 방법

김 기 백[a)‡]

# Binary Mask Estimation using Training-based SNR Estimation for Improving Speech Intelligibility

Gibak Kim[a)‡]

요 약

본 논문에서는 시간-주파수 영역에서의 이산 마스킹을 이용하여 잡음환경 음성의 음성 명료도를 높이는 방법에 대해 다루고자 한다. 잡음이 섞여 있는 음성신호를 시간-주파수 영역으로 분해하여, 상대적으로 잡음이 많이 섞여 있는 시간-주파수 영역의 신호를 마스크 “0”을 할당하여 제거함으로써 음성명료도를 향상시킬 수 있다. 이러한 이산 마스크를 추정하기 위해서는 각 시간-주파수 영역에서 신호 대 잡음 비를 추정하여 문턱값과 비교해야 하는데, 본 논문에서는 학습 기반의 신호 대 잡음 비 추정방법을 사용하여 문턱값과 비교하여 이산 마스크를 추정한다. 신호 대 잡음 비와 비교하기 위한 문턱값은 모든 주파수 대역에 대해 동일한 값을 이용하는 고정 문턱값 외에도 주파수 대역에 따라 학습 데이터의 분포로부터 최적의 값을 사용하는 최적 문턱값을 제안한다. 제안된 이산 마스크 추정 방법은 잡음 환경 데이터에 적용한 후, 피험자에게 들려주어 음성 명료도를 측정한다.

Abstract

This paper deals with a noise reduction algorithm which uses the binary masking approach in the time-frequency domain to improve speech intelligibility. In the binary masking approach, the noise-corrupted speech is decomposed into time-frequency units. Noise-dominant time-frequency units are removed by setting the corresponding binary masks as “0”s and target-dominant units are retained untouched by assigning mask “1”s. We propose a binary mask estimation by comparing the local signal-to-noise ratio (SNR) to a threshold. The local SNR is estimated by a training-based approach. An optimal threshold is proposed, which is obtained from observing the distribution of the training database. The proposed method is evaluated by normal-hearing subjects and the intelligibility scores are computed by counting the number of words correctly recognized.

Keyword : Binary mask, Noise reduction, Speech intelligibility

a) 숭실대학교 전기공학부(School of Electrical Engineering, Soongsil University)
‡ Corresponding Author : 김기백 (Gibak Kim)
E-mail: imkgb27@ssu.ac.kr
Tel: +82-2-828-7266 Fax: +82-2-828-7266
※이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행되었습니다 (2012-0003455).
· Manuscript received August 30, 2012 Revised November 14, 2012 Accepted November 14, 2012

# Ⅰ. 서 론

잡음 환경에서 녹음된 음성 신호에 대해 잡음을 제거하는 연구들이 많이 진행되어 왔다. 지금까지 시도되었던 잡음 제거 기법들은 주로 잡음 환경에서의 음성 통신을 위해 사용자가 좀 더 편안하게 들을 수 있도록 음질을 개선하는데 초점이 맞춰져 있다[1-4]. 또한, 음성 인식에 대한 연구가 활발해지고 음성 인식의 실용화를 위해 생활 잡음 환경에서 취득된 음성 신호의 잡음을 억제하고자하는 연구들도 많이 진행되어 왔다[5,6]. 이때까지의 연구들을 통해 잡음을 제거하여 음질 (quality)을 개선하거나 음성 인식의 성능을 높여왔지만 주로 시간에 따라 특성이 빨리 변하지 않는 잡음, 즉 정상 잡음 (stationary noise)에 국한되어 왔다. 음성 명료도 (intelligibility) 향상은 잡음 제거의 또 다른 목적으로서 음질 향상과는 달리 사람이 들었을 때 그 의미를 얼마나 잘 이해할 수 있는가를 판단하는 문제이다. Hu 와 Loizou의 연구결과에 의하면 기존의 대표적인 방법들을 적용하여 car, street, babble, train 등의 잡음 환경에서 취득된 음성 신호의 잡음을 제거한 결과, 음질은 어느 정도 개선할 수 있었으나 명료도를 의미있게 개선하지는 못함을 알 수 있다[7-9].

본 논문에서는 음성의 명료도를 개선하는 알고리즘을 개발하고자 하며, 이산 마스크를 이용하는 방법을 사용하고자 한다. 이산 마스크는 CASA[10-12] (Computational Auditory Scene Analysis)로부터 연구되어온 것으로서 각각의 시간-주파수 영역에서 목적 신호 (음성)의 에너지가 간섭 잡음 신호의 에너지보다 큰 경우는 그대로 두고, 그렇지 않은 경우는 신호를 제거한다. 이전의 연구결과들로부터 이상적인 (ideal) 이산 마스크를 극심한 잡음 환경 데이터에 적용하였을 때 음성 명료도를 획기적으로 높일 수 있음을 확인하였다[13-15]. 이러한 연구들에서는 이상적인 이산 마스크를 얻기 위해 잡음이 없는 음성 신호에 잡음을 섞어 잡음환경 데이터를 만들고, 이미 알고 있는 음성과 잡음으로부터 신호 대 잡음 비의 참값을 구해서 정해진 문턱값과 비교하였다. 그러나 실제 환경에서는 잡음이 섞인 신호의 신호 대 잡음 비의 참값을 구할 수 없으므로 추정값에 의존할 수밖에 없다. Hu와 Loizou는 각각의 시간-주파수 영역에서 Decision-Directed 방법[3]으로 추정된 신호 대 잡음 비를 이용하여 이산 마스크를 추정하였다[16]. 이들의 연구에서 추정된 이산 마스크와 이상적인 이산 마스크를 비교하여 성능을 비교한 결과, 신호 대 잡음 비가 낮은 경우 (0 dB)는 잡음 제거 알고리즘으로 사용할 수 없을 정도로 낮은 성능을 보여주었다. Decision-Directed 방법은 적은 계산량으로 신호 대 잡음 비를 추정하는 비교적 간단한 방법이나 비정상 잡음이나 신호 대 잡음 비가 낮은 경우는 좋은 성능을 기대하기 어렵다. 정상적인 청력을 가진 사람들에게는, 신호 대 잡음 비가 그리 낮지 않은 경우 (> 0 dB)는 잡음이 어느 정도 있는 경우, 청취하는데 불편함을 느끼기는 하지만 음성이 전달하고자 하는 언어적 의미를 이해하는 데는 큰 문제가 없다는 것이 잘 알려진 사실이다. 그러나 신호 대 잡음 비가 0 dB보다 작은 경우는 청취하는데 불편을 느낄 뿐만 아니라 그 의미를 이해하기도 힘들어 진다. 따라서 음성 명료도 향상이라는 주제는 대부분의 경우 신호 대 잡음 비가 0 dB보다 낮은 잡음 환경을 대상으로 하고 있어 기존의 신호 대 잡음 비 추정 방법을 적용해서는 좋은 성능을 기대하기 어렵다.

본 논문에서는 신호 대 잡음 비 추정의 대표적 방법인 Decision-Directed 방법과는 달리 특정한 잡음 환경에서 취득한 학습 데이터를 이용하는 방법을 사용하고자 한다. 특정 환경에서 수집된 데이터를 학습하여 결과를 얻으므로 환경에 따른 제약이 있지만 좋은 성능을 기대할 수 있다. 학습에 기반한 방법은 학습 데이터로부터 추출된 특징벡터를 입력으로 하고 신호 대 잡음 비를 출력으로 하는 분류기를 사용한다. 본 논문에서는 진폭 변조 스펙트로그램 (Amplitude Modulation Spectrogram)을 특징벡터로 하고 분류기로는 인공 신경망 (artificial neural network)을 사용한다[17-19]. 특징벡터로 사용하는 진폭 변조 스펙트로그램은 피치와 같은 시간 영역의 정보와 주파수 영역의 정보를 모두 포함하고 있다[20,21]. 이와 같은 신호 대 잡음 비 추정 방법은 학습 데이터를 이용하여 학습하여야 하므로 학습 데이터와 같은 잡음 환경에서만 동작하는 단점이 있지만 Decision-Directed와 같은 비훈련 알고리즘에 비해 낮은 신호 대 잡음 비를 보이는 잡음 환경 데이터의 신호 대 잡음 비 추정에서 높은 성능을 갖는 것으로 알려져 있다. 본 논문에서는 이러한 신호 대 잡음 비 추정 방법을 통해 추정된 각 시간-주파수 영역의 신호 대 잡음 비와 문턱값을 비교하

여 이산 마스크를 추정하고 이를 잡음이 섞인 원래의 신호에 적용한다. 문턱값은 고정 문턱값 (fixed threshold)과 학습 데이터의 분포로부터 구한 최적 문턱값 (optimal threshold)을 적용하였다. 이렇게 이산 마스크가 적용된 신호와 처리하지 않은 신호를 정상 청력을 가진 피험자들에게 들려주어 음성 명료도를 계산하고 향상 정도를 확인한다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 특징벡터의 추출과정 및 인공 신경망을 이용한 신호 대 잡음 비 추정에 대해 설명하고, 3장에서는 추정된 신호 대 잡음 비로부터 이산 마스크를 추정하기 위한 고정 문턱값과 최적 문턱값에 대해 설명한다. 정상 청력을 가진 피험자들로부터 얻어진 실험결과는 4장에서 제시한다.

## II. 학습 기반의 신호 대 잡음 비 추정

본 장에서는 잡음 환경에서 녹음된 음성 신호의 신호 대 잡음 비를 추정하는 것에 대해 설명한다. 신호 대 잡음 비 추정은 학습 데이터로부터 특징벡터를 추출하여 분류기를 학습하는 과정을 거치게 된다. 본 논문에서 사용하는 Tchorz와 Kollmeier가 사용했던 방법을 기초로 하고 있으며 특징벡터로는 진폭 변조 스펙트로그램, 분류기로는 인공 신경망을 사용한다[17-19].

### 1. 진폭 변조 스펙트로그램

진폭 변조 스펙트로그램을 추출하는 과정을 정리하면 다음과 같다.

입력 신호에 대해 4ms 크기의 Hann 윈도우를 적용한다. 본 논문의 실험에서 사용하는 데이터의 표본 주파수는 12kHz이므로 4ms는 48샘플에 해당한다. 80개의 제로를 추가하여 128 포인트 FFT (Fast Fourier Transform)을 적용한 후, FFT 계수의 절대값에 제곱을 취하여 전력 스펙트럼 (power spectrum)을 얻는다. 입력 신호에 4ms의 Hann 윈도우를 적용하는 것은 0.25ms가 진행될 때마다 수행하게 되며, 이렇게 되면 스펙트럼의 크기에 대한 표본 주파수는 4kHz(1/0.25ms)가 된다.

전력 스펙트럼은 인간의 청각 특성을 반영한 mel 스케일 주파수[22]를 기반으로 한 25개의 필터로 구성된 필터뱅크를 통과시키게 된다. 이제 25개의 주파수 채널에 대해 4kHz로 표본화된 전력 스펙트럼을 얻게 된다.

각 주파수 채널에 대해 32ms (128샘플) 동안의 전력 스펙트럼들을 모아 다시 Hann 윈도우를 적용하게 되고, 이러한 과정을 16ms씩 진행하여 수행하게 된다. 이제 Hann 윈도우를 적용한 128개의 데이터에 128개의 제로를 추가하여 256 포인트 FFT를 적용한 후, 절대값을 취하여 진폭 변조 스펙트로그램을 얻는다.

특징벡터의 크기를 줄이기 위해 15개의 삼각필터로 이루어진 필터뱅크를 적용하여 크기가 15인 진폭 변조 스펙트로그램을 얻게 되고, 주파수 채널이 모두 25개이므로 전체 주파수 대역에 대한 진폭 변조 스펙트로그램 특징벡터의 크기는 25×15가 된다.

특징벡터를 추출하는 이상의 과정을 그림 1에 정리하였다.

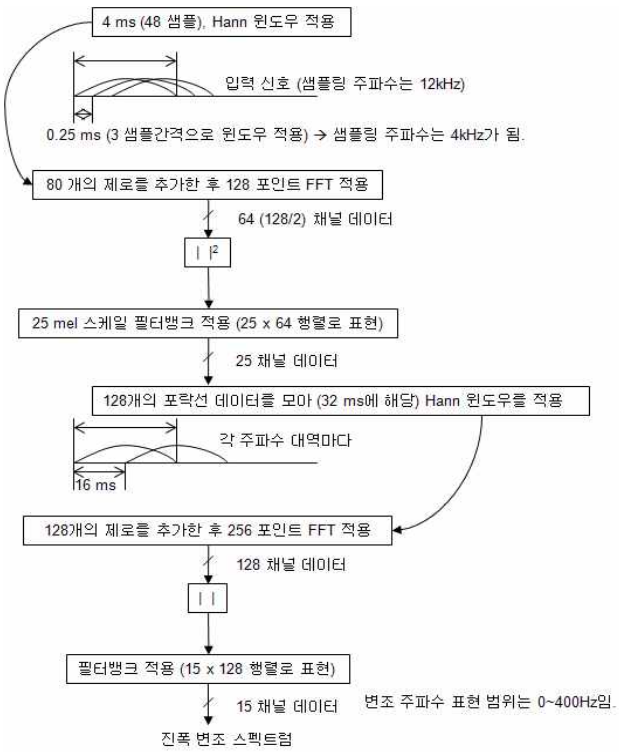


그림 1. 특징벡터로 사용되는 진폭 변조 스펙트로그램의 추출 과정
Fig. 1. Extraction of amplitude modulation spectrogram

### 2. 밝기 정보를 이용한 특징값 추출

신호 대 잡음 비를 추정하는 문제를 패턴 인식을 이용하여 풀 수 있는데, 본 논문에서는 앞에서 구한 특징벡터를 입력으로 하고 신호 대 잡음 비를 출력으로 하는 인공 신경망을 구성하고자 한다. 인공 신경망은 패턴 분류 문제에서 다양하게 사용되는 방법으로서 본 논문에서는 표준 feed-forward 신경망이다[23]. feed-forward 신경망은 그림 2와 같이 은닉층 (hidden layer)이 입력층 (input layer)과 출력층 (output layer) 사이에 존재한다.

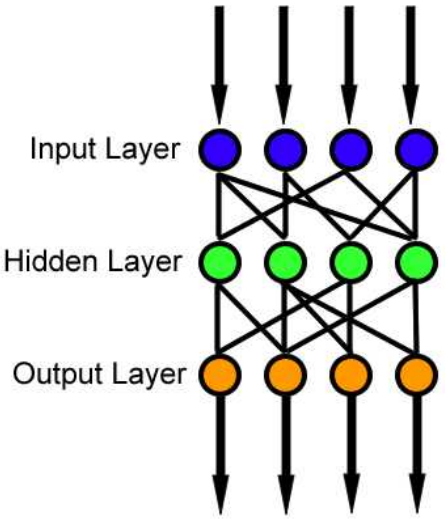


그림 2. Feed-forward 신경망의 구조 (출처: Wikipedia)
Fig. 2. Structure of feed-forward neural networks

입력층은 375 (25×15)개의 뉴런 (neuron)을 갖고, 은닉층에는 225개의 뉴런을 배치하였다. 225개의 은닉층 개수는 실험에 의해 결정되었으며, 이는 일반적인 경우에 비해 많은 편에 속한다고 볼 수 있다. 이와 같이 많은 은닉층 개수가 필요한 이유로 추정되는 것 중 하나는 잡음 환경에 종속적으로 신경망을 학습한다는 것이다. 출력층은 하나의 뉴런을 갖는데, 학습 시 이 뉴런의 값은 신호 대 잡음 비 -60dB에서 25dB를 0에서 1의 값으로 선형매핑하였다 (그림 3). 테스트 데이터의 특징벡터를 입력층에 넣어 신경망의 출력으로 나온 값으로부터 신호 대 잡음 비를 얻기 위해서는 그림 3의 매핑함수를 역으로 적용하면 된다.

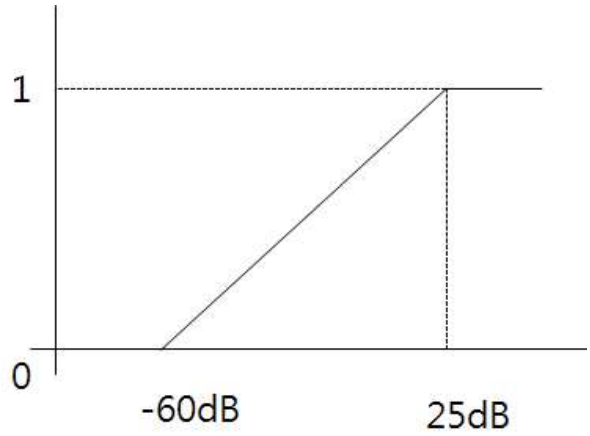


그림 3. 신호 대 잡음 비와 출력층 뉴런값 매핑함수
Fig. 3. Mapping fuction of output neuron according to SNR

## III. 이산 마스크 추정

각 시간-주파수 영역에서 신호 대 잡음 비를 추정한 후에는 이산 마스크를 추정하기 위해 문턱값과 비교하여야 한다. 즉, 추정된 신호 대 잡음 비가 문턱값보다 큰 경우는 마스크 값을 1로 설정하고 문턱값보다 작은 경우는 마스크 값은 0으로 설정한다. 이제 문턱값을 어떻게 설정해야 하는지에 대한 문제가 발생한다. 이상적인 이산 마스크를 이용한 Li와 Loizou의 연구에 따르면, 신호 대 잡음 비의 문턱값을 -20에서 0 dB 사이에서 설정하여 얻은 이산 마스크를 잡음이 섞인 음성에 적용하여 정상청력을 가진 피험자들에게 들려주었을 때 100%에 가까운 인식률을 나타냄을 알 수 있었다[14]. 본 연구에서는 모든 주파수 대역에 대해 동일한 문턱값을 적용하는 고정 문턱값으로는 -8 dB를 사용하였다[1]. 이러한 고정 문턱값과는 달리 주파수 대역에 따라 다른 문턱값을 갖는 경우도 생각해 볼 수 있다. 본 논문에서는 각 주파수 대역마다 학습 데이터의

1) -20~0 dB의 몇 가지 문턱값을 선택하여 실험하여 가장 좋은 결과를 나타내는 -8 dB를 고정 문턱값으로 선택하였음.

분포를 고려하여 문턱값을 정하는 최적 문턱값을 제안한다. 각 주파수 대역에 대해 최적 문턱값은 다음과 같은 방법을 이용하여 구한다.

각 학습 데이터를 시간-주파수 영역으로 분해하고 인공 신경망을 이용하여 신호 대 잡음 비를 추정한다.

각 주파수 대역에 대해, 실제 신호 대 잡음 비 ($\xi_{true}$)가 고정 문턱값 -8 dB 보다 큰 경우에 대해 추정 신호 대 잡음 비 ($\xi_{est}$)의 히스토그램을 구한다.

각 주파수 대역에 대해, 실제 신호 대 잡음 비가 고정 문턱값 -8 dB 보다 작은 경우에 대해 추정 신호 대 잡음 비의 히스토그램을 구한다.

각 주파수 대역에 대해 이렇게 얻어진 두 히스토그램이 교차하는 신호 대 잡음 비 값을 최적 문턱값으로 설정한다.

$H_1$을 실제 신호 대 잡음 비 ($\xi_{true}$)가 문턱값보다 큰 경우라고 하고 $H_0$를 실제 신호 대 잡음 비가 문턱값보다 작은 경우라고 하면, ②번 과정에서 구한 히스토그램은 이산 마스크 "1"에 대한 추정된 신호 대 잡음 비의 조건부 확률분포 ($p(\xi_{est}|H_1)$)를 나타내고 ③번 과정에서 구한 히스토그램은 이산 마스크 "0"에 대한 추정된 신호 대 잡음 비의 조건부 확률분포 ($p(\xi_{est}|H_0)$)를 나타낸다. 따라서 두 분포가 교차하는 지점의 신호 대 잡음 비를 최적 문턱값으로 설정할 수 있다. 본 연구에서 사용한 학습데이터로부터 구해진 최적 문턱값은 그림 4와 같다.

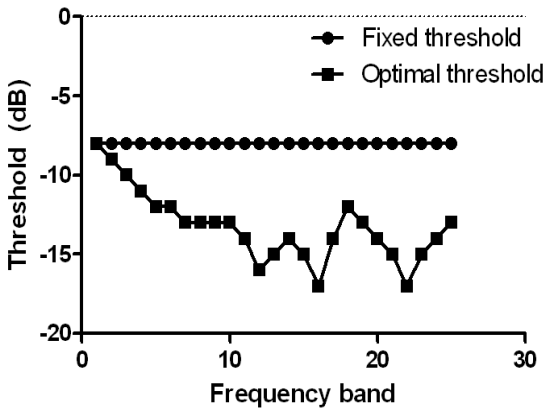


그림 4. 주파수 대역에 따른 고정 문턱값과 최적 문턱값
Fig. 4. Fixed and optimal thresholds according to frequency band

## IV. 실험 결과

### 1. 실험 환경

성능 검증을 위한 실험에 사용한 문장은 영어로 구성되어 있고, 영어 원어민에 의해 발성된 것을 12kHz 샘플링 주파수로 녹음하였다. 잡음 데이터는 Babble 잡음[2]을 사용하였다. 한 문장에는 10 개 내외의 단어들이 포함되어 있으며 학습 데이터로는 80개의 문장을 -5dB, 0dB, 5dB 등으로 잡음을 섞어 사용하였다. 테스트를 위해서는 -5dB, 0dB 각 10개의 문장을 사용하였다. 테스트를 위해 사용한 데이터는 학습 데이터와는 겹치지 않는다.

### 2. 실험 결과 I: 정검출율과 오검출율

먼저, 제안한 방법에 따른 이산 마스크 추정 성능 평가를 위해 정검출율 (Hit)과 오검출율 (FA)을 계산하였다. 정검출율은 실제 신호 대 잡음 비가 문턱값보다 높아 실제 이산 마스크가 "1"인 경우 중에, 추정한 신호 대 잡음 비도 역시 문턱값보다 높아 이산 마스크가 "1"로 추정된 경우의 비를 말한다. 오검출율은 실제 신호 대 잡음 비가 문턱값보다 낮아 실제 이산 마스크가 "0"인 경우 중에, 추정한 신호 대 잡음 비가 문턱값보다 높아 이산 마스크가 "1"로 잘못 추정된 경우의 비를 말한다. 그림 5에 정검출율 (Hit, 가장 위에 있는 그래프)와 오검출율 (FA, 가운데 있는 그래프)을 나타내었다. 정검출율과 오검출율 모두 최적 문턱값의 경우가 고정 문턱값의 경우보다 높게 나왔다. 정검출율은 높을수록 좋은 성능을 나타내고, 오검출율은 낮을수록 좋은 성능을 나타내므로 이 두 그래프 결과만으로는 최적 문턱값과 고정 문턱값 중 어떤 것이 좋은 결과를 나타내는지 가늠하기 어렵다. 앞에서 언급한 것처럼 정검출율은 높을수록 좋고, 오검출율은 낮을수록 좋은 것이므로 두 측정값의 차이 (Hit-FA)가 클수록 성능이 좋은 것이라고 할 수 있다. 그림

2) 남녀 각각 10명이 동시에 서로 다른 문장을 읽는 것을 녹음하여 생성한 잡음을 사용하였다.

5의 가장 아래에 있는 그래프에는 정검출율과 오검출율의 차이 (Hit-FA)를 나타내었다. 그림에서 보는 것처럼 최적 문턱값을 적용했을 때가 고정 문턱값을 적용했을 때에 비해 높은 Hit-FA를 나타냄을 알 수 있다. 저주파 대역에서는 큰 차이를 보이지 않으나 고주파 대역에서는 성능개선이 크게 나타남을 알 수 있다. 마찰음이나 파열음 계열의 자음이 고주파 대역에 분포하고 있으므로 고주파 대역에서의 성능개선은 이러한 자음의 명료도를 높여 준다고 볼 수 있다.

## 3. 실험 결과 II: 음성 명료도 측정

제안한 알고리즘의 성능을 검증하기 위해 잡음 환경의 음성을 정상 청력을 가진 영어 원어민 10명의 피험자를 통해 음성 명료도를 측정하였다. 잡음 환경의 음성을 그대로 들려주는 실험과 제안한 알고리즘을 통해 이산 마스크를 추정하고, 추정된 이산 마스크를 잡음 환경 음성에 적용하여 잡음 제거 처리를 한 신호를 피험자에게 들려주는 실험을 수행하였다. 이산 마스크를 이용한 잡음 제거 방법의 최고 성능을 가늠하기, 실제 신호 대 잡음 비를 이용하여 이상적인 이산 마스크를 구하고 이를 잡음 환경 음성에 적용하여 피험자에게 들려주는 실험도 같이 수행하였다. 음성 명료도 측정을 위해서는 피험자들이 헤드폰을 통해 청취하고 이해한 문장을 그대로 받아 적게 하여 주어진 문장에 있는 단어 중 몇 %를 정확히 받아 적었는지를 계산하였다.

명료도 측정 결과는 표 1에 나타내었다. 잡음 제거 처리 전 0dB인 경우는 66.1%로서 어느 정도 알아듣는 것으로 나왔으나 -5dB에서는 11.3%로서 거의 알아듣지 못한다고 할 수 있다. 이상적인 이산 마스크를 적용한 경우는 고정 문턱값이든 최적 문턱값이든 0dB에서는 100%에 가깝고, -5dB에서는 95%에 가까운 높은 결과를 나타냄을 알 수 있다. 본 논문에서 제안한 방법을 이용하여 시간-주파수 영역의 이산 마스크를 추정하고, 추정된 이산 마스크를 잡음 환경의 음성에 적용한 후 측정된 음성 명료도는 잡음 제거

표 1. 피험자를 통한 음성 명료도 측정
Table 1. Speech intelligibility measured by listening test

| 신호 대 잡음 비 | 잡음 제거 처리 전 | 추정 이산 마스크 적용 | | 이상적인 이산 마스크 적용 | |
|---|---|---|---|---|---|
| | | 고정 문턱값 | 최적 문턱값 | 고정 문턱값 | 최적 문턱값 |
| -5dB | 11.3% | 35.7% | 51.6% | 95.0% | 94.0% |
| 0dB | 66.1% | 81.2% | 82.3% | 98.2% | 99.1% |

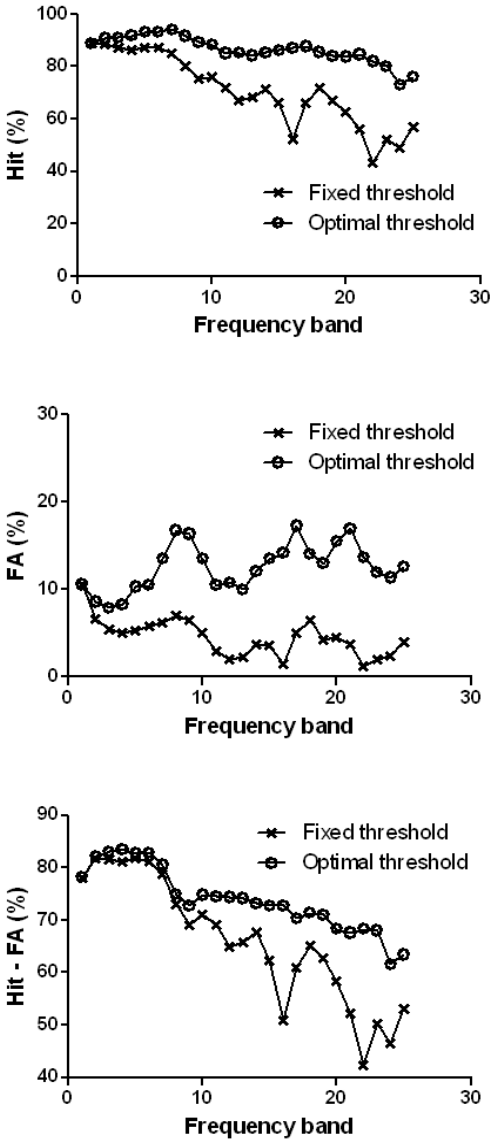


그림 5. 고정 문턱값 (Fixed threshold)과 최적 문턱값 (Optimal threshold)을 적용했을 때의 정검출율 (Hit), 오검출율 (FA) 성능 비교
Fig. 5. Performance comparison

처리 전의 결과에 비해 많이 향상되었음을 알 수 있다. 0dB 환경에서는 고정 문턱값에 비해 제안한 최적 문턱값의 결과가 별 차이가 없는 것으로 나타났지만 -5dB 환경에서는 고정 문턱값에 비해 최적 문턱값을 적용했을 때 16%정도 더 높은 결과를 보여 의미있는 향상을 보인다고 할 수 있다.

## V. 결 론

본 논문에서는 잡음이 심한 상황 (0dB, -5dB)에서의 잡음 제거를 통한 음성 명료도 향상을 도모하였고, 그 방법으로서 시간-주파수 영역의 이산 마스크를 추정하여 적용하는 연구에 대해 논의하였다. 이산 마스크 추정을 위해서 필요한 신호 대 잡음 비 추정은 진폭 변조 스펙트로그램을 기반으로 하는 특징벡터와 인공 신경망을 이용하여 특정 잡음 환경에서 학습하는 방법을 선택하였다. 추정된 신호 대 잡음 비를 이용하여 이산 마스크를 추정하기 위해서는 문턱값과 비교하여야 하는데, 본 논문에서는 모든 주파수 대역에서 일정한 문턱값을 사용하는 고정 문턱값외에 학습 데이터의 분포를 이용하여 주파수 대역별로 최적의 문턱값을 사용하는 방법을 제안하였다. 본 논문에서 제안하는 이산 마스크 적용을 통한 잡음 제거 방법의 성능을 평가하기 위해 잡음 제거된 음향 신호를 정상 청력을 가진 피험자들에게 들려주는 실험을 수행하였다. 실험 결과 본 논문에서 제안하는 잡음 제거 방법이 의미있는 음성 명료도 향상을 보여줌을 알 수 있었고, 최적 문턱값을 적용한 경우 보다 향상된 성능을 보임을 확인할 수 있었다.

## 참 고 문 헌

[1] J. S. Lim and a. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586 – 1604, 1979.

[2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-27, no. 2, pp. 113 – 120, 1979.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-32, no. 6, pp. 1109 – 1121, 1984.

[4] Y. Ephraim and H. Van Trees, “A signal subspace approach for speech enhancement,” IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 251 – 266, 1995.

[5] J. Huang and Y. Zhao, “An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises,” Speech Communication, vol. 26, no. 3, pp. 165 – 181, Nov. 1998.

[6] K. Hermus, P. Wambacq, and H. Hamme, “A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition,” EURASIP Journal on Advances in Signal Processing, vol. 2007, no. 1, p. 045821, 2007.

[7] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” Speech communication, vol. 49, no. 7, pp. 588 – 601, Jul. 2007.

[8] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Transactions on Speech and Audio Processing, vol. 16, no. 1, pp. 229 – 238, 2008.

[9] Y. Hu and P. C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms.” The Journal of the Acoustical Society of America, vol. 122, no. 3, p. 1777, Sep. 2007.

[10] G. Brown and M. Cooke, “Computational auditory scene analysis,” Computer speech and language, vol. 8, pp. 297 – 336, 1994.

[11] D. Wang and G. Brown, Computational Auditory Scene Analysis : Principles, Algorithms, and Applications, Wiley, Hoboken, NJ, 2006.

[12] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” In Divenyi P. (ed.), Speech Separation by Humans and Machines, pp. 181-197, Kluwer Academic, Norwell MA, 2005.

[13] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” The Journal of the Acoustical Society of America, vol. 120, no. 6, p. 4007, 2006.

[14] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction.,” The Journal of the Acoustical Society of America, vol. 123, no. 3, pp. 1673 – 82, Mar. 2008.

[15] N. Li and P. C. Loizou, “Effect of spectral resolution on the intelligibility of ideal binary masked speech.,” The Journal of the Acoustical Society of America, vol. 123, no. 4, pp. EL59 – 64, Apr. 2008.

[16] Y. Hu and P. Loizou, “Techniques for estimating the ideal binary mask,” in Proc. 11th Int. Workshop Acoust. Echo Noise Control, 2008.

[17] J. Tchorz and B. Kollmeier, “Estimation of the signal-to-noise ratio with amplitude modulation spectrograms,” Speech Communication, vol. 38, no. 1 – 2, pp. 1 – 17, Sep. 2002.

[18] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 3, pp. 184 – 192, May 2003.

[19] M. Kleinschmidt and V. Hohmann, “Sub-band SNR estimation using auditory feature processing,” Speech Communication, vol. 39, no. 1 – 2, pp. 47 – 63, Jan. 2003.

[20] G. Langner and C. E. Schreiner, “Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms.,” Journal of neuro-

physiology, vol. 60, no. 6, pp. 1799 - 822, Dec. 1988.
[21] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.,” The Journal of the Acoustical Society of America, vol. 95, no. 3, pp. 1593 - 602, Mar. 1994.
[22] S. Stevens, J. Volkmann, and E. Newman, “A scale for the measurement of the psychological magnitude pitch,” The Journal of the Acoustical Society of America, vol. 8, no. 3, pp. 185-190, 1937.
[23] C. Bishop, Neural Networks for Pattern Recognition, New York: Oxford Univ. Press, 1995.

## 저 자 소 개

### 김 기 백

- 1994년 : 서울대학교 전자공학과 학사
- 1996년 : 서울대학교 전자공학과 석사
- 2007년 : 서울대학교 전기컴퓨터공학부 박사
- 1996년 ~ 2000년 : LG전자기술원 연구원
- 2000년 ~ 2003년 : (주)보이스웨어 선임연구원
- 2008년 ~ 2010년 : Univ. of Texas at Dallas, Research Associate
- 2010년 ~ 2011년 : 대구대학교 전자공학부 전임강사
- 2011년 ~ 현재 : 숭실대학교 조교수
- 주관심분야 : 음성신호처리, 영상신호처리, 멀티모달신호처리, 어레이신호처리