필터 뱅크 최적화에 의한 멜켑스트럼의 성능 향상

Performace Improvement of Mel-Cepstrum Through Optimzing Filter Banks

현 동 훈*, 이 철 희*
(Dong Hoon Hyun*, Chul Hee Lee*)

※본 연구는 정통부 대학기초연구지원사업의 지원을 받아 이루어졌습니다.

요 약

본 논문에서는 현재 음성 인식에서 널리 사용되고 있는 멜켑스트럼의 성능 향상 방안을 제안한다. 일반적으로 멜켑스트럼은 인접한 필터간의 중심 간격과 필터의 대역폭이 일정한 critical band 필터들을 사용하여 구한다. 그러나 필터의 특성에 따라 멜켑스트럼의 값들이 달라지게 되고, 이에 따라 인식 성능도 변하게 된다. 본 논문에서는 삼각형과 사각형 모양의 critical band 필터를 사용하여 인접한 필터간의 중심 간격과 필터의 대역폭을 각각 변화시키면서 멜켑스트럼을 구하고 이에 따른 인식 성능을 분석한다. 또한 최적화 알고리즘인 simplex 방법을 사용하여 필터의 중심 주파수와 대역폭을 각각 변화시키면서 최적의 성능을 나타내는 필터를 구하는 방법을 제안한다. 인식 알고리즘으로 DTW (dynamic time warping)를 사용하고, 남자 10명과 여자 10명이 발음한 한국어 숫자음을 인식 대상으로 하여 실험을 수행하였다. 사각형 모양의 필터가 삼각형 모양의 필터 보다 우수한 성능을 보여 주었고 제안된 방법으로 최적화된 필터를 사용하여 구한 멜켑스트럼은 기존의 critical band 필터를 사용하는 것보다 향상된 인식 성능을 나타내었다.

ABSTRACT

In this paper we propose a method to improve the performance of the mel-cepstrum that is widely used in speech recognition. Typically, the mel-cepstrum is obtained by critical band filters that have fixed center spacing and bandwidth. However different filter characteristics produce a different mel-cepstrum, resulting in a different performance. In this paper we analyze triangular-shaped and rectangular-shaped filters. By changing the characteristics of filters such as center frequency and bandwidth, we analyze the performance of the mel-cepstrum. Then utilizing the simplex method, we propose a method to optimize the critical band filters. Using the dynamic time warping, we performed speaker independent recognition experiments with Korean digit words pronounced by 10 males and 10 females. Experiments show that the rectangular-shaped filters show good performance and the mel-cepstrum obtained by the optimized filters shows better performance than filters that have fixed center spacing and bandwidth.

I. 서 론

음성 신호는 화자의 변화, 음의 강약, 주변 잡음 등으로 인해 같은 단어라도 많은 차이를 나타내고, 음성 신호 자체가 수천 내지 수만 샘플들로 구성되어 있으므로 원래 음성 신호를 직접 사용하여 인식을 수행하기에는 많은 문제점이 있다. 따라서 음성 인식을 위해서는 고차원의 데이터인 음성 신호를 그 특징을 잘 나타낼 수 있는 저차원의 피쳐로 전환하는 과정을 거치게 된다. 일반적으로 음성의 피쳐는 음성 신호를 일정한 프레임(수십ms)으로 나누어 각 프레임에서 추출하게 되는데, 현재 널리 사용되고 있는 피쳐로 LPC 계수와 켑스트럼 등이 있다.

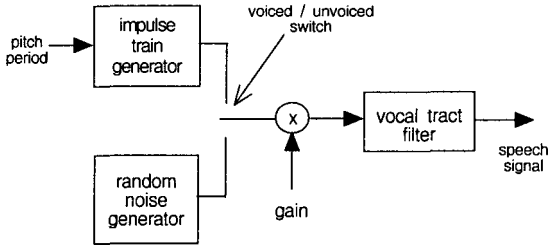


그림 1. 음성 발생 모델
Fig. 1. A model for the generation of speech signal.

음성 신호의 발생은 그림 1과 같이 모델링 될 수 있다[1]. 일반적으로 그림 1에서 성도에 대한 정보(vocal tract

* 연세대학교 전자공학과
접수일자 : 1998년 10월 21일

필터)가 음성 인식에 적절한 피춰이나 음성 신호에는 성도에 대한 정보 외에 여기 신호(vocal tract 필터의 입력)가 포함되어 있다. 음성 신호부터 성도에 대한 정보를 얻기 위하여 LPC 분석을 통해서 성도 필터의 계수들을 구할 수 있는데 이를 LPC 계수라 한다. 이것은 성도에 대한 정보를 나타내므로 음성 인식을 위한 피춰로 사용될 수 있다. 여기 신호를 분리하는 또 다른 피춰로 켑스트럼이 있는데, LPC 계수보다 우수한 인식 성능을 나타낸다 [2]. 최근에는 인간의 청각 특성을 이용한 켑스트럼인 멜켑스트럼을 피춰로 많이 사용하고 있다. 멜켑스트럼은 critical band 필터를 사용하여 구하는데, Davis와 Mermelstein는 그림 2와 같이 critical band 필터 20개를 사용하여 구한 멜켑스트럼이 다른 피춰들보다 우수한 성능을 나타내는 것을 보여 주었다 [2].

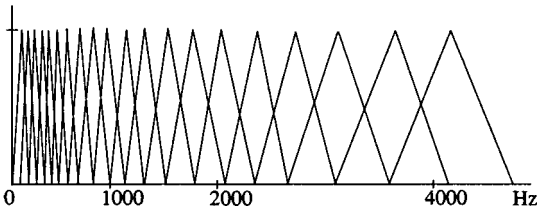


그림 2. 멜켑스트럼을 구하기 위한 필터 뱅크
Fig. 2. Filter banks for the mel-cepstrum.

이러한 피춰를 사용하여 인식을 수행하는 알고리즘으로는 DTW(dynamic time warping)와 HMM(hidden Markov model) 등이 있다. DTW는 고립어(isolated word) 인식에 적당하지만, 입력 단어와 비교하는 기준 단어의 수가 많아지면 계산량이 늘어나고 그만큼 인식 시간이 증가하는 단점이 있다. HMM은 확률적인 모델을 이용한 것으로, 현재 가장 널리 사용되고 있는 인식 알고리즘의 하나이다. 본 논문에서는 숫자음을 인식 대상으로 하였으므로 DTW를 인식 알고리즘으로 사용하여 그림 2와 같은 critical band 필터의 중심 주파수, 대역폭, 필터 형태 등을 변화시켜 구한 멜켑스트럼의 성능을 분석하여 최적화 기법을 제안한다. 본 논문의 구성은 다음과 같다.

Ⅱ절에서 최적화 기법에 필요한 멜켑스트럼을 구하는 방법에 대하여 간략하게 기술하고, Ⅲ절에서 최적화 방법으로 사용한 simplex 방법에 대해서 설명한다. Ⅳ절에서 simplex 방법을 이용하여 멜켑스트럼의 최적화 알고리즘을 제시한 뒤, Ⅴ절에서는 숫자음에 대한 인식 실험 결과를 보여주고 Ⅵ절에서 결론을 제시한다.

II. 피춰 추출

음성 신호의 발생은 그림 1과 같이 모델링 할 수 있으며, 여기서 성도 필터(vocal tract filter)의 임펄스 응답(impulse response)은 성도의 모양에 대한 정보를 나타낸다. 음성의 발생은 성도의 모양에 직접적인 영향을 받으므로 성도 필터의 임펄스 응답이 음성 인식에 적절한 피춰로 간주된다. 성도 필터의 입력, 즉 여기 신호(excitation signal)를 $e(n)$, 성도 필터의 임펄스 응답을 $\theta(n)$이라 하면 음성 신호 $s(n)$은 다음과 같이 나타낼 수 있다.

$$s(n) = e(n) * \theta(n) \tag{1}$$

이때 켑스트럴 분석을 통하여 각 단어마다 고유한 $\theta(n)$을 음성 신호로부터 분리할 수 있다. 켑스트럴 분석의 첫 번째 단계로서, 식 (1)을 주파수 영역에서 나타내면 다음과 같다.

$$S(\omega) = E(\omega)\Theta(\omega) \tag{2}$$

주파수 영역에서의 크기만 고려하여 식 (2)의 양변에 log를 취하면,

$$\log|S(\omega)| = \log|E(\omega)| + \log|\Theta(\omega)| \tag{3}$$

이 된다. 이때 식 (3)을 역푸리에 변환하면 다음과 같다.

$$c_s(n) = c_e(n) + c_\theta(n) \tag{4}$$

식 (4)에서 구한 $c_s(n)$을 켑스트럼이라 한다. 켑스트럼에서 낮은 차수 대역은 성도 필터의 임펄스 응답에 해당하고 높은 차수 대역은 여기 신호에 해당하므로 적절한 smoothing 과정을 통하여 여기 신호 부분을 제거할 수 있으며 이를 liftering이라 한다 [3]. 켑스트럼은 주파수 영역에서 단순히 log 연산을 수행하여 구하는데, 여기에 인간의 청각 특성까지 고려한 것이 멜켑스트럼이다. 멜켑스트럼을 구하는 과정이 그림 3에 나타나 있다 [4]. 그림 3에서 멜켑스트럼을 구하는 과정은 켑스트럼을 구하는 과정과 유사하지만 critical band 필터를 사용하는 점이 다르다.

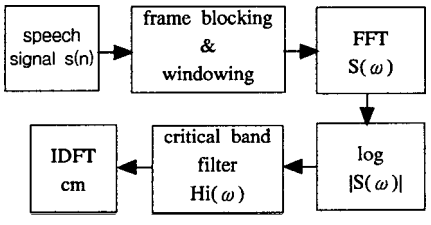


그림 3. 음성 신호로부터 멜켑스트럼을 구하는 과정
Fig. 3. The procedure for computing the mel-cepstrum.

멜켑스트럼을 구하기 위한 첫 번째 과정으로 음성 신호의 각 프레임에 대해서 N-point FFT를 수행하여 N개의 성분을 구한다. 이때 각 성분에 해당하는 주파수는 다음과 같다.

$$f_k = k\frac{F_s}{N}, \quad \text{for} \quad 0 \le k \le N-1$$

여기서 F_s는 샘플링 주파수이고 k는 주파수 인덱스를 나타낸다. 이때 i번째 critical band 필터의 출력 $Y(i)$는

$$Y(i)=\sum_{k=0}^{N/2}\log|S(k)|H_i(k)\,,\qquad i=1,\ldots,N_f$$

이고, 여기서 N_f는 필터의 개수를 나타낸다. N_f개 필터의 출력들을 토대로 다음과 같은 새로운 주파수 성분을 형성한다.

$$\tilde{Y}(k)=\begin{cases} Y(i), & k=k_i \\ 0, & 0\le k\le N-1 \quad (k\neq k_i)\end{cases}$$

여기서 k_i는 i번째 필터의 중심에 해당하는 주파수 인덱스이다. 이때 $\tilde{Y}(k)$를 다음과 같이 역푸리에 변환하여 멜 켑스트럼 c_m을 구한다.

$$c_m=\frac{1}{N}\sum_{k=0}^{N-1}\tilde{Y}(k)e^{j(2\pi/N)km}\,,\qquad m=1,\ldots,n$$

여기서 n은 멜켑스트럼의 차수를 나타낸다. 실신호(real signal)를 푸리에 변환하여 스펙트럼의 크기만 고려하면 우함수가 된다. 이를 샘플링하여 N-point DFT를 하면 주파수 인덱스 $N/2$에 대해서 대칭이 된다 [5]. $\tilde{Y}(k)$가 이러한 성질을 가진다고 가정하여 아래의 식으로 역푸리에 변환을 수행한다.

$$\begin{aligned} c_m &= \frac{1}{N}\sum_{k=0}^{N-1}\tilde{Y}(k)\cos\left(\frac{2\pi}{N}km\right) \\ &= \frac{2}{N}\sum_{k=1}^{N/2-1}\tilde{Y}(k)\cos\left(\frac{2\pi}{N}km\right) \\ &= \frac{2}{N}\sum_{i=1}^{N_f}\tilde{Y}(k_i)\cos\left(\frac{2\pi}{N}km\right)\end{aligned}$$

여기서 $\tilde{Y}(0)=\tilde{Y}(N/2)=0$으로 가정하였다. 위의 과정에서 critical band 필터의 모양, 개수, 중심 주파수, 대역폭 등을 변화시켜 같은 음성이라도 각각 다른 멜켑스트럼을 구할 수 있고, 이에 따른 인식률의 변화를 관찰할 수 있다. 본 논문에서는 필터의 변화에 따른 인식 성능의 변화를 분석한 뒤, 향상된 인식 성능을 나타내도록 필터를 최적화 하였다.

III. 최적화 방법

입력이 N차원 벡터 v이고 출력이 k인 함수 f를 생각할 때, 이를 그림 4와 같이 나타낼 수 있다.

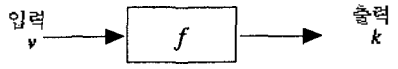


그림 4. 함수 f의 입력과 출력
Fig. 4. The input and output of function f.

이때 최적화(optimization) 방법을 통해 함수 f의 최대 값 또는 최소 값을 구할 수 있다. 최적화 방법에는 여러 가지가 있으나, 함수 f의 gradient가 주어져 있지 않을 때는 direct search 방법을 사용하며, 이러한 방법에는 Nelder와 Mead가 제안한 simplex 방법이 있다 [6]. Simplex는 N차원 공간에서 $N+1$개의 점으로 구성되는 다면체를 의미한다. 예를 들면, 2차원 공간에서 simplex는 삼각형이다. 그림 4에서 입력 v가 다음과 같은 N차원 벡터로 주어질 때,

$$v=[x_1\ \cdots\ x_N]^T$$

최소를 찾기 위해서 먼저 N차원 공간에 초기 simplex를 설정한다. 초기 simplex를 일련의 과정에 따라 변화시키고, 이러한 과정을 반복함으로써 함수 f의 최소를 찾게 된다. 이때 N차원 공간에서 $N+1$개의 점으로 구성되는 초기 simplex는 아래 식에 의해서 설정한다.

$$v_i = v_0 + a_i e_i\,,\qquad i=1,\ldots,N \tag{5}$$

$$\textit{where}\quad \begin{cases} e_1=[1\ 0\ 0\ \cdots\ 0]^T \\ e_2=[0\ 1\ 0\ \cdots\ 0]^T \\ \qquad\vdots \\ e_N=[0\ 0\ 0\ \cdots\ 1]^T \end{cases}$$

여기서 v_0는 N차원에서 임의로 잡은 벡터, e_i는 N차원의 단위 벡터, 그리고 a_i는 simplex의 각 모서리의 길이를 결정하는 상수이다. 이와 같이 초기 simplex를 설정한 뒤, $N+1$개의 점들에서 함수 f값을 계산하여 최대 값과 최소 값을 각각 f_G, f_S로 표시하고, 그에 해당되는 점을 각각 G, S라 할 때, G를 제외한 나머지 점들이 구성하는 다면체의 중심 X를 다음 식에 의해 구할 수 있다.

$$X=\frac{1}{N}\left(\sum_{i=0}^{N}v_i-G\right)$$

여기서 X는 G를 이동하기 위한 기준점으로 볼 수 있다. 초기 simplex를 설정한 뒤 함수 f의 최소를 찾기 위해서 simplex를 변형시키는데, 이 방법에는 크게 두 가지가 있다. 첫 번째 방법은 simplex를 구성하는 $N+1$개의 점 중에서 1개의 점을 이동하는 것인데, 여기에는 reflection, expansion, contraction 등의 단계가 있다. 각각의 단계에 해당되는 점을 각각 R, E, C라 하면 아래와 같은 식으로 구할 수 있다.

$$R = X + (X - G), \qquad f_R = f(R)$$
$$E = R + (X - G), \qquad f_E = f(E)$$
$$C = X + \frac{1}{2}(G - X), \qquad f_C = f(C)$$

여기서 f_R, f_E f_C,는 각 점에 해당하는 함수 값이다. 이때 이동된 점에서의 함수 값 f_R, f_E, f_C를 f_G, f_S와 비교하여 그림 6에 제시된 몇몇 조건들에 의하여 이동된 점으로 새로운 simplex를 형성한다. 다른 이동 방법은 S를 제외한 모든 점을 이동하는 것인데, 이를 multiple contraction이라 한다. 이것은 아래 식에서 구한 $N+1$개의 점으로 simplex를 변형시키는 것이다.

$$v_i = \frac{v_i + S}{2}, \qquad i = 0, \ldots, N$$

그림 5는 2차원 공간에서 simplex를 고려할 때, 즉 입력 v가 2차원 벡터일 때 위에 열거한 모든 단계를 도식적으로 나타낸다.

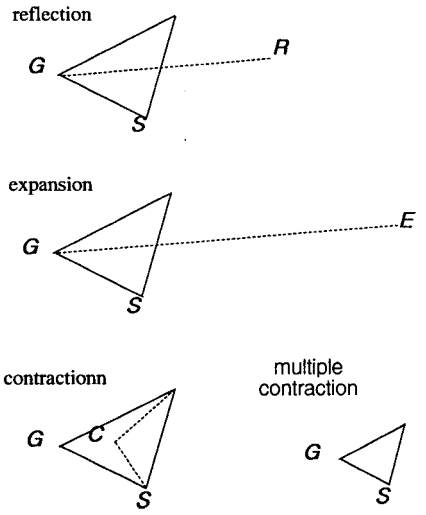


그림 5. Simplex를 변화시키는 단계들
Fig. 5. Various operations to change a simplex.

이러한 단계들을 반복함으로써 함수 f의 최소를 가지는 N차원 영역으로 simplex가 이동하게 된다. 종료 조건으로 사용할 수 있는 조건으로 반복 횟수를 제한하여 일정 횟수를 초과할 때까지 최적화를 수행하는 것과 아래 식과 같이 simplex의 각 모서리 길이가 미리 정해 놓은 값 ε보다 작을 때까지 최적화를 수행하는 것 등이 있다.

$$\max | v_i - v_j | < \varepsilon \qquad i, j = 1, \ldots, N$$

본 논문에서는 반복 횟수를 제한하여 최적화를 수행하였다. 지금까지 설명한 simplex 방법을 요약하여 도식적으로 나타내면 그림 6과 같다 [7].

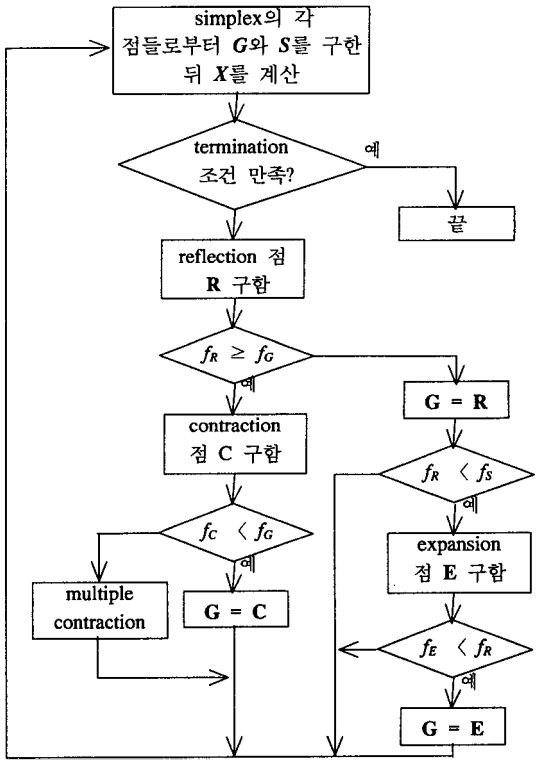


그림 6. 최소값을 찾기 위한 simplex 방법의 흐름도
Fig. 6. A flowchart for the simplex method to find a minimum.

IV. 멜켑스트럼의 최적화

같은 단어를 발음한 음성이라도 critical band 필터의 중심 주파수, 대역폭 등에 따라 다른 멜켑스트럼 값을 갖게 되며 인식 성능도 변화하게 된다. 본 연구에서는 앞서 기술한 simplex 방법으로 필터를 최적화한다.

Critical band 필터의 중심 주파수와 대역폭을 변화시킬 수 있다고 가정할 때, N_f개의 critical band 필터가 있는 경우 $2N_f$개의 파라미터가 존재하게 된다. 따라서 음성 인식률을 함수 f의 출력으로, $2N_f$개의 파라미터는 다음과 같이 입력 벡터 v의 원소들로 생각할 수 있다.

$$v = [\, c_1 \cdots c_{Nf} \; b_1 \cdots b_{Nf} \,]^T$$

여기서 c_i, b_i는 각각 i번째 필터의 중심 주파수와 대역폭을 나타낸다.

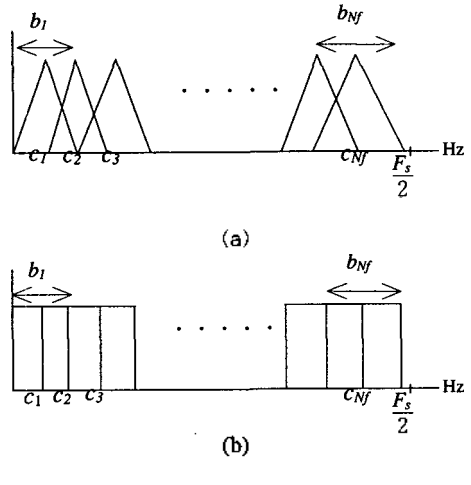
Ⅲ절에서 최소를 찾는 것을 가정하여 simplex 방법을 설명하였으나 함수 f의 최소는 함수 $-f$의 최대와 같다. 그러므로 최소를 찾는 것과 최대를 찾는 것은 동일한 과정으로 생각할 수 있다. 따라서 Ⅲ절에서 설명한 과정을 통하여 최대의 인식률을 구할 수 있고 이때의 입력 v를 통하여 최적의 critical band 필터를 얻을 수 있다. 이때 중요한 문제점의 하나는 $2N_f$차원 공간에서 초기 simplex의 위치이다. 초기 simplex가 최대를 나타내는 영역에 근

접해 있으면 빠르고 정확하게 최대 값을 찾을 수 있으나, 그렇지 않으면 최대 값에 도달하는 시간이 오래 걸리거나 찾지 못하는 경우도 발생하게 된다.

본 논문에서는 최대를 나타내는 영역에 근접하는 초기 simplex를 찾기 위하여 그림 7과 같이 필터의 대역폭(B)과 인접한 필터의 중심 간격(C)을 변화시킨 여러 개의 critical band 필터를 사용하여 멜켑스트럼을 구하고 이에 따른 인식 결과를 관찰하여 우수한 인식 성능을 나타내는 critical band 필터를 선택하였다. 이때 선택된 필터의 중심 주파수와 대역폭을 초기입력 v_0로 설정하여 식 (5)에 의해서 초기 simplex를 구성하였다. 그림 7에서 각 필터의 대역폭(B)과 인접한 필터의 중심 간격(C)은 각각 mel 단위의 값을 가진다. Mel 단위와 Hz 단위의 대응 관계는 아래 식과 같다 [8].

$$F_{mel} = 2595 \log_{10}\left(1 + \frac{F_{Hz}}{700}\right)$$

또한 critical band 필터의 개수 N_f는 각 필터의 통과 대역이 $F_s/2$보다 작은 부분에 위치하도록 결정하였다. 즉, C 와 B 의 값에 따라서 필터의 개수가 결정된다.



$c_i = c_{i-1} + C$, $b_i = B$ $(C, B$: *mel unit*$)$

그림 7. 여러 가지 critical band 필터들
Fig. 7. Various critical band filters.

초기 simplex를 결정한 후 III절에서 설명한 reflection, expansion, contraction, multiple contraction 과정을 통하여 simplex를 변형시키게 된다. 입력 v를 필터의 중심 주파수와 대역폭으로 설정하였기 때문에 simplex를 구성하는 $2N_f+1$ 개의 점들 중에서 부적절한 점이 발생할 수 있으며 다음과 같이 두 가지 경우를 생각할 수 있다.

첫째로, 필터의 중심 주파수 c_i와 대역폭 b_i가 음수로 되는 경우이다. 즉, simplex를 구성하는 한 점 v의 원소 중에서 음수가 존재하는 경우이다. Simplex를 변형시키게 되면 이러한 경우가 발생할 수 있는데, 이때 멜켑스트럼을 구하는데 문제점이 생긴다. 대역폭 b_i가 음수로 되는 경우에는 이에 해당되는 필터가 존재하지 않는 것으로 생각하여 멜켑스트럼을 구할 수 있다. 그러나 이러한 경우에 필터의 개수가 변하게 되는데, 이것은 필터의 개수를 고정시키고 중심 주파수와 대역폭을 파라미터로 설정하여 최적화를 수행하는 가정에 위배된다. 따라서 본 논문에서는 대역폭이 음수가 되는 경우에 처음과 동일한 필터의 개수를 유지하기 위해서 $|b_i|$를 대역폭으로 사용하였다. 또한 필터의 중심 주파수 c_i가 음수로 되는 경우를 고려할 수 있는데, 실험 결과 상에서는 드물게 발생하였다. 이러한 경우는 다음과 같이 처리한다.

둘째로, 필터의 통과 대역이 $F_s/2$보다 큰 부분이나 0보다 작은 부분을 포함하는 경우가 발생한다. 다시 말하면, $c_i - \frac{b_i}{2} < 0$이 되거나 $c_i + \frac{b_i}{2} > F_s/2$이 되는 v가 simplex를 구성하는 한 점이 될 수 있다. 음성 신호의 스펙트럼은 $0 \sim F_s/2$에 분포하는 것으로 가정하므로 필터의 통과 대역이 $F_s/2$보다 크거나 0보다 작은 대역은 필터의 이득을 0으로 하여 멜켑스트럼을 구하였다. 따라서 필터의 중심 주파수 c_i가 음수로 되는 경우에도 대역폭을 고려하여 필터의 이득을 조정함으로써 필터의 개수를 처음과 같이 유지하여 최적화를 수행할 수 있다.

V. 실험 결과 및 고찰

본 논문에서는 숫자음을 실험 대상으로 하여 멜켑스트럼을 최적화하였다. 화자 독립 인식 실험을 수행하기 위해 남자 10명과 여자 10명의 화자가 숫자음 0에서 9까지 각각 10번씩 발음하여 총 2000개의 음성 데이터를 구축하였다. 단, 샘플링 주파수를 11.025kHz로 하여 PC에서 음성을 녹음하였고 음성의 시작점과 끝점은 수동으로 검출하였다. 300개의 샘플들을 하나의 프레임으로 하여 피춰를 추출하였고, 프레임 간의 간격은 100 샘플로 하였다. 또한 각 프레임에서 스펙트럼을 구하기 위해서 1024-point FFT를 수행하였고 인식 알고리즘으로 DTW를 사용하였다.

첫 번째 실험은 화자 2명(남자 1명, 여자 1명), 화자 4명(남자 2명, 여자 2명), 화자 6명(남자 3명, 여자 3명)이 한번씩 발음한 숫자음을 기준 음성으로 사용하여 각각 최적화를 수행하는 것이다. IV절에서 설명한 바와 같이 초기의 simplex를 설정하기 위해서 mel 단위의 (C, B)를 다음과 같은 8가지 경우로 설정하여 초기 실험을 수행하였다.

(100, 100), (100, 200), (125, 125), (125, 250),
(150, 150), (150, 300), (175, 175), (175, 350)

인접한 필터간의 중심 간격과 대역폭을 위의 (C, B)로 사용하고, 필터의 모양을 삼각형과 사각형으로 하였을 때 인식 결과를 그림 8(a), 8(b)에 각각 나타내었다. 기준 음성으로 남자 1명, 여자 1명이 한번씩 발음한 음성을 임의로 한 개 선택하였고 10, 12, 14, 16차 멜켑스트럼에 대한 인식률을 구하였다.

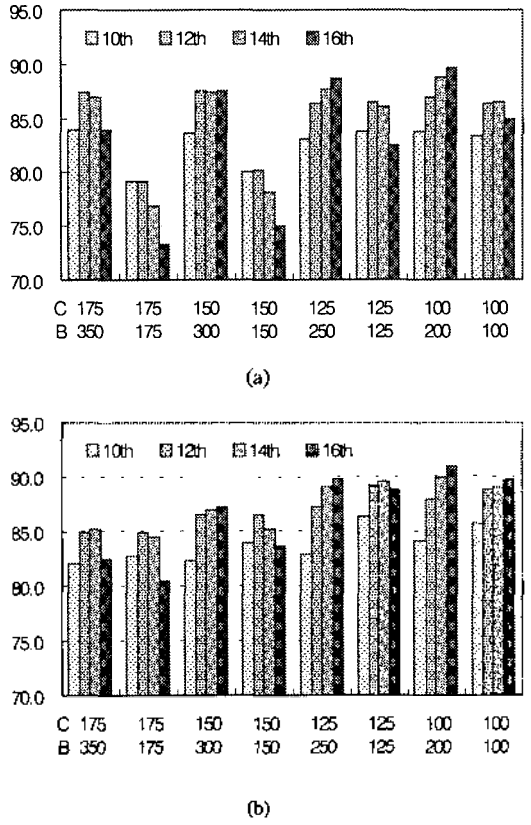


그림 8. 여러 가지 (C, B)에 의한 멜캡스트럼의 인식률
(a) 삼각형 모양의 필터 (b) 사각형 모양의 필터
Fig. 8. Recognition rates of the mel-cepstrum by various (C,B)s;
(a) Triangular-shaped filter, (b) Rectangular-shaped filter.

위와 인식 결과에 의하면 필터의 모양은 사각형이 삼각형보다 대체적으로 인식 성능이 우수하였고 사각형 모양의 필터에서 (C, B)가 (125, 125)와 (100, 100)일 때 인식률이 양호하였다. 따라서 본 실험에서는 초기의 simplex를 구성하기 위하여 (C, B)를 (100, 100)으로 하여 초기의 입력 $\boldsymbol{v}_0$를 설정하였다. 샘플링 주파수를 고려하면 24개의 필터가 $F_s/2$ 범위 안에 존재하므로 초기의 입력 $\boldsymbol{v}_0$는 다음과 같이 나타난다.

$$\boldsymbol{v}_0 = [100\ \ 200\ \cdots\ 2400\ \ 100\ \ 100\ \cdots\ 100]^T$$

이것은 초기 simplex를 구성하는 한 점이 되고 나머지 점들은 아래의 식으로 구한다.

$$\begin{cases} \boldsymbol{v}_i = \boldsymbol{v}_0 + 60\boldsymbol{e}_i, & i = 1, \cdots, N_f \\ \boldsymbol{v}_i = \boldsymbol{v}_0 + 150\boldsymbol{e}_i, & i = N_f + 1, \cdots, 2N_f \end{cases}$$

이와 같이 초기 simplex를 설정하였고, 반복 횟수를 100회로 정하여 최적화를 수행하였다. 10, 12, 14, 16차의 멜캡스트럼에 대해서 각각 최적화를 수행하였으며 (C, B)가 (100, 100)일 때의 인식률과 최적화된 인식률을 비교하면 그림 9와 같다. 그림에서 볼 수 있듯이 4~5%의 성능 향상이 관찰되었다. 또한 화자 4명(남자 2명, 여자 2명)과 화자 6명(남자 3명, 여자 3명)이 한번씩 발음한 숫자음을 기준 음성으로 각각 사용하여 마찬가지 방법으로 최적화를 수행한 결과도 그림 9에 나타내었다. 최적화된 필터에서 저주파 대역의 필터는 대역폭이 대부분 100 mel 보다 증가하고 고주파 대역의 필터는 대역폭이 100 mel 보다 감소하는 경향이 있는데, 그림 10은 기준화자 6명에 대해서 최적화된 필터의 대역폭의 변화를 나타낸다.

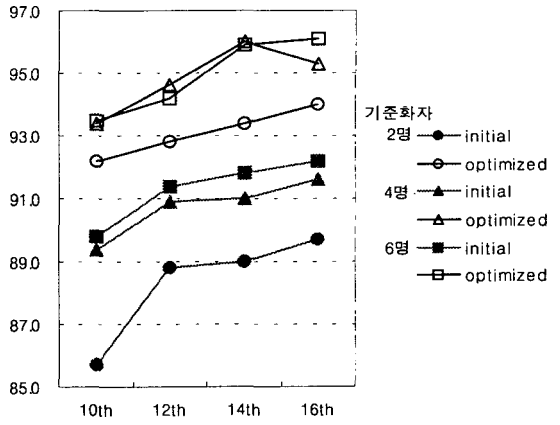


그림 9. (C,B)가 (100, 100)일 때와 최적화된 인식률의 비교 (기준화자 2, 4, 6명)
Fig. 9. Performance comparison between (100,100) and optimized filters (2, 4, and 6 reference speakers).

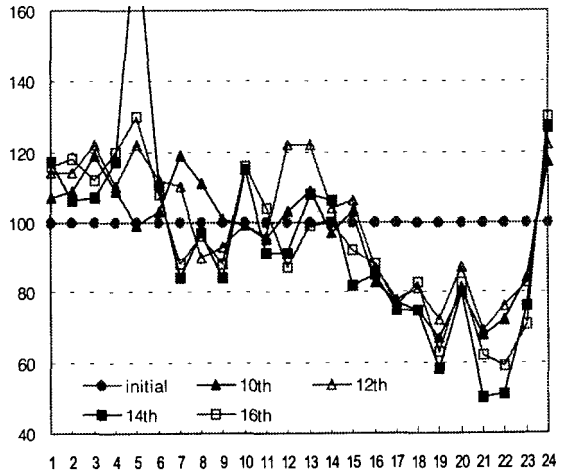


그림 10. 기준화자 6명에 대해서 최적화된 필터의 대역폭의 변화 (mel 단위)
Fig. 10. Bandwidth characteristics of the optimized filters for 6 reference speakers (mel unit).

두 번째 실험에서는 앞에서 최적화된 필터가 첫 번째 실험과 다르게 선택된 기준 음성에 대해서 향상된 인식률을 나타내는지 검증하였다. 6명의 기준화자(남자 3명,

여자 3명)를 임의로 35회 선택하여 앞에서 최적화된 필터를 사용하여 인식 실험을 수행하였다. 이때 최적화를 수행하기 전의 필터를 사용한 경우, 즉 (C, B)가 (100, 100)일 때와 최적화된 필터를 사용한 경우의 인식률의 평균값을 그림 11에 나타내었다. 그림에서 optimal(2), optimal(4), optimal(6)은 각각 기준화자 2명, 기준화자 4명, 기준화자 6명에 대해서 최적화된 필터에 의한 인식률을 의미한다. 또한 최적화된 필터에 의하여 증가된 인식률의 최대, 최소, 평균, 표준편차를 표 1에 나타내었다. 최적화 필터에 의한 인식률의 증가는 선택된 기준 음성에 따라서 차이를 보였으며 몇 개의 기준 음성에 대해서는 인식률이 감소하는 경우도 발생하였다. 그러나 대부분의 경우 인식률이 증가하였고 평균적으로 볼 때 표 1에 나타난 바와 같이 2~4%의 인식률 향상을 나타내었다.

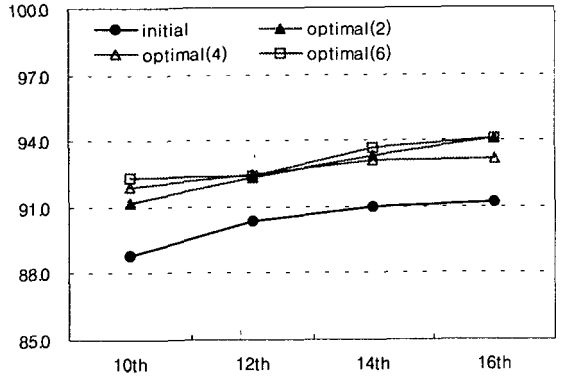


그림 11. 기준화자 2명, 4명, 6명에 대해서 최적화된 필터를 각각 사용할 때와 (C,B)가 (100, 100)일 때의 인식률 비교 (기준화자 6명을 임의로 35회 선택)

Fig. 11. Performance comparison between (100,100) and optimized filters for 2, 4, and 6 reference speakers (35 random selections of 6 reference speakers).

표 1. 기준화자 2명, 4명, 6명에 대해서 최적화된 필터를 각각 사용할 때 증가된 인식률의 최대값, 최소값, 평균, 표준편차 (기준화자 6명을 임의로 35회 선택, 단위 %)

Table 1. Maximum, minimum, average, and standerd deviation of improved rates by optimized filter for 2, 4, and 6 reference speakers (35 random selections of 6 reference speakers).

	기준화자 2명에 대한 최적화 필터				기준화자 4명에 대한 최적화 필터				기준화자 6명에 대한 최적화 필터			
	최대	최소	평균	표준편차	최대	최소	평균	표준편차	최대	최소	평균	표준편차
10th	5.5	0.4	2.4	1.4	5.7	0.7	3.1	1.2	6.1	1.8	3.5	1.1
12th	4.6	0.7	2.0	1.0	4.8	-0.3	2.1	1.3	4.7	0.0	2.1	1.1
14th	5.4	0.2	2.4	1.2	5.7	-0.1	2.1	1.4	5.3	0.7	2.7	1.0
16th	6.0	0.8	2.8	1.1	3.6	0.8	1.9	0.8	4.5	1.1	2.8	0.8

VI. 결 론

본 논문에서는 simplex 알고리즘에 의하여 멜켑스트럼을 최적화하여 음성 인식률의 향상을 모색하였다. 삼각형 모양의 필터 뱅크보다 사각형 모양의 필터 뱅크를 사용하여 구한 멜켑스트럼이 보다 좋은 인식 성능을 나타내었으며, 사각형 모양의 필터 뱅크의 중심 주파수와 대역폭을 변화시키면서 멜켑스트럼의 최적화 가능성을 보여주었다. 최적화된 멜켑스트럼은 인접한 필터간의 중심 간격이 일정한 기존의 필터 뱅크에 의해서 구한 멜켑스트럼보다 향상된 인식 성능을 나타내었다. 최적화된 필터는 저주파 성분의 경우 대역폭이 증가하였고, 고주파 성분의 경우 대역폭이 감소하는 것을 관찰할 수 있었다.

참 고 문 헌

1. L. R. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, PTR Prentice Hall, 1993, pp. 97-122.
2. S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-28, pp. 357-366, Aug. 1980.
3. B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-35, pp. 947-954, July 1987.
4. J. R. Deller J. R. J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals, Prentice* Hall, 1987, pp. 352-386.
5. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989, pp. 535-547.
6. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, 2nd.*, Cambridge University Press, 1992, pp. 408-412.
7. J. L. Buchanan, P. R. Turner, *Numerical Methods and Analysis*, McGraw-Hill, INC., 1992, pp. 362-371.
8. D. O'Shaughnessy, *Speech Communication*, Addison-Wesley Publishing Company, 1987, pp. 148-151.

▲현 동 훈(Donghoon Hyun)

1997년 2월 : 연세대학교 전자공학과 졸업(공학사)

1997년 3월~현재 : 연세대학교 전자공학과 석사과정

※ 주관심분야 : 음성신호처리, 음성인식

▲이 철 희(Chulhee Lee)

1980년 3월～1984년 2월 : 서울대학교 전자공학과 졸업(공학사)

1984년 3월～1986년 2월 : 서울대학교 대학원 전자공학과 (공학석사)

1986년 9월～1987년 3월 : Technical University of Denmark (Researcher)

1987년 8월～1992년 12월 : Purdue University Electrical Engineering(Ph. D.)

1993년 7월～1996년 8월 : National Institutes of Health, Maryland, USA(Visiting fellow)

1996년 9월～현재 : 연세대학교 기계전자공학부 조교수

※주관심분야 : 신호처리, 영상처리, 패턴인식, 음성합성