말소리와 음성과학 제2권 제4호 (2010)
pp. 145~151

가중 ARMA 필터를 이용한 강인한 음성인식

Robust Speech Recognition Using Weighted Auto-Regressive Moving Average Filter

반 성 민[1] · 김 형 순[2]
Ban, Sung Min · Kim, Hyung Soon

ABSTRACT

In this paper, a robust feature compensation method is proposed for improving the performance of speech recognition. The proposed method is incorporated into the auto-regressive moving average (ARMA) based feature compensation. We employ variable weights for the ARMA filter according to the degree of speech activity, and pass the normalized cepstral sequence through the weighted ARMA filter. Additionally when normalizing the cepstral sequences in training, the cepstral means and variances are estimated from total training utterances. Experimental results show the proposed method significantly improves the speech recognition performance in the noisy and reverberant environments.

Keywords: robust feature compensation, speech recognition, temporal modulation filter, Auto-Regressive Moving Average

1. 서론

음성인식에서 훈련환경과 인식환경의 차이는 인식 성능저하의 주요한 요인이다. 특히 음원과 마이크 사이의 거리가 멀어질수록 부가 잡음과 반향 때문에 이러한 불일치는 더욱 커진다. 대부분의 음성인식 시스템은 음원과 마이크 사이의 거리가 가까운 환경에서 안정적인 성능을 보인다. 하지만 다양한 응용분야에서 음성인식 시스템을 적용하기 위해서는 환경 불일치 문제를 해결해야 한다. 본 논문에서는 이러한 문제를 해결하기 위해 부가잡음과 반향에 강인한 특징 보상 알고리즘을 제안한다.

환경 불일치 문제를 해결하기 위해서 많은 연구가 진행되었는데, 모델 보상(model compensation), 음질 개선(speech enhancement), 반향 제거(dereverberation), 강인한 특징 보상(robust feature compensation)등이 있다. 모델 보상은 parallel model combination(PMC)과 vector Taylor series(VTS) 방식처럼 사전에 인식환경의 정보를 이용하여 깨끗한 음성으로부터의 음향모델을 보상할 수 있다[1], [2]. 이러한 방식은 인식환경의 정보를 충분히 가지고 있다면 효과적으로 사용될 수 있지만, 보통 인식환경을 예측하기 어렵기 때문에 실제로 이러한 모델 보상 방식을 사용하는 데에는 제한이 있다. 이에 비해 음질 개선은 인식환경에 대한 사전정보 없이 잡음을 추정하여 음질을 개선시킨다. 음질 개선은 부가잡음으로 인한 왜곡 감소에 초점을 맞춘 것으로 지금까지 많은 연구가 진행 되었다[3], [4]. 음질 개선 방식으로 부가 잡음을 효율적으로 제거할 수 있지만 반향까지 함께 제거하기는 어렵다. 그래서 반향을 줄이기 위한 연구가 별도로 진행되기도 했는데, 최근에는 음질 개선과 반향 제거를 결합하여 음성인식에서 우수한 성능을 보이고 있다. 반향 제거 알고리즘은 역필터(inverse filtering), 피치, 선형 예측(linear prediction)등을 사용하는 방식들이 있는데, 이 중 다단 선형예측(multi-step linear prediction, MSLP)이 우수한 성능을 보인다[5]. 하지만 다단 선형 예측은 음성 파형의 상관도를 이용하기 때문에 계산량이 많은 단점을 가지고 있다. 실제 음성인식 시스템은 마이크의 개수와 계산량이 제한되는데, 본 논문에서는 단일 채널 마이크를 사용하면서 적은 계산량으로 우수한 성능을 내는 특징 보상 알고리즘을 제안한다.

부가 잡음과 반향에 강인한 특징을 추출하기 위해 어떤 정보가 음성의 명료성에 큰 영향을 미치는지는 매우 중요하며, 이를 알아내기 위한 연구들이 진행되었다. 최근에 temporal modulation structure(TMS)가 음성의 명료성에 있어 중요하며, 여러 잡음환

1) 부산대학교 bansungmin@pusan.ac.kr
2) 부산대학교 kimhs@pusan.ac.kr, 교신저자

접수일자: 2010년 10월 30일
수정일자: 2010년 12월 21일
게재결정: 2010년 12월 22일

경에 강인하다는 연구 결과가 나왔다[6]. 음성은 FFT를 통해 진폭 변조(amplitude modulation)된 협대역(narrowband)들로 표현할 수 있는데, 시간에 따라 켑스트럼의 크기가 변화하는 TMS는 음성의 중요한 정보를 나타낸다. 음성에서 대부분의 유용한 정보는 1~16Hz의 주파수 영역에 존재하며, 특히 3~4Hz구간은 음성의 음절률을 반영한다. 반면 16Hz 이상의 주파수 대역은 음성의 특성을 왜곡시키므로, 이러한 성분을 억제하면서 음성이 존재하는 대역을 잘 보존하면 음성인식에 있어서 강인한 특징을 추출할 수 있을 것이다. 이미 이러한 접근방법으로 강인한 특징 보상에 대한 연구가 진행되었고, ARMA필터, 에지 보존(edge-preserved) 필터를 사용한 방식 등이 제안되었다[7], [8]. 이러한 방식들을 temporal modulation filter(TMF)라 부른다.

본 논문은 기존의 ARMA 필터를 켑스트럼 시계열에 적용할 때 각 프레임별 특징벡터에 음성의 존재에 대한 가중치를 주어 특징을 보상하는 알고리즘을 제안한다. 또한 특징 정규화 과정에서 신뢰도가 높은 켑스트럼의 평균과 분산을 구하기 위해 전체 훈련 데이터에 대한 전역 평균과 분산을 구하여 특징을 정규화하는 방식을 도입한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 다양한 TMF를 설명하고, 3장에서 제안한 알고리즘에 대해 설명한다. 4장에서 부가잡음과 반향이 존재하는 환경에서의 음성인식률로 성능을 평가하고, 5장에서 논문의 결론을 맺는다.

2. 기존의 TMF

TMF 알고리즘들은 켑스트럼 시계열을 cepstral mean and variance normalization(CMVN)으로 정규화 시킨 후 다양한 필터를 통과시킬 수 있다. 먼저 기존의 TMF중 ARMA필터와 에지 보존 필터 방식에 대해 간단히 살펴도록 하겠다.

2.1 ARMA 필터

저대역 통과 특성을 가지는 ARMA필터를 켑스트럼 시계열에 적용하면 잡음으로 인한 고주파수 성분들을 제거할 수 있다. Chen은 ARMA 필터를 식 (1)과 같이 정의하였다[7].

$$C_{ARMA}^{(t,k)} = \frac{\sum_{i=1}^{m} C_{ARMA}^{(t-i,k)} + \sum_{i=0}^{m} \hat{C}^{(t+i,k)}}{2m+1} \tag{1}$$

여기서 $C_{ARMA}^{(t,k)}$는 t번째 프레임의 k번째 켑스트럼의 ARMA 필터링된 결과이며, m은 ARMA필터의 차수이다. $\hat{C}^{(t,k)}$는 t번째 프레임의 k번째 켑스트럼의 CMVN 결과이며, CMVN 과정은 식 (2)와 같다.

$$\hat{C}^{(t,k)} = \frac{C^{(t,k)} - \overline{C}^{k}}{\sigma^{k}}, \quad k = 0,1,\cdots 12 \tag{2a}$$

여기서 켑스트럼의 평균과 표준편차 $\overline{C}^{k}$와 σ^{k}는 각각 식 (2b), (2c)와 같다.

$$\overline{C}^{k} = \frac{1}{T}\sum_{t=1}^{T} C^{(t,k)} \tag{2b}$$

$$\sigma^{k} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(C^{(t,k)} - \overline{C}^{k}\right)^{2}} \tag{2c}$$

여기서 $C^{(t,k)}$는 t번째 프레임의 k번째 켑스트럼의 값이고, T는 문장의 프레임수이다.

2.2 에지 보존 필터

에지 보존 필터는 영상처리 분야에서 먼저 사용된 것으로 바이레터럴 필터(bilateral filter)로 불리기도 한다. 에지 보존 필터는 특징이 변화(transient)하는 부분에서의 에지정보를 보존하면서 평활화(smoothing)하는 방식이다. t번째 프레임의 k번째 켑스트럼의 에지 보존 필터링된 결과 $C_{Edge}^{(t,k)}$는 식 (3)과 같다[8].

$$C_{Edge}^{(t,k)} = \frac{\sum_{i=1}^{m} w_{k}^{(t,-i)} C_{Edge}^{(t-i,k)} + \sum_{i=0}^{m} w_{k}^{(t,i)} \hat{C}^{(t+i,k)}}{\sum_{i=-m}^{m} w_{k}^{(t,i)}} \tag{3}$$

여기서 $w_{k}^{(t,i)} = w_{k_T}^{(t,i)} \cdot w_{k_I}^{(t,i)}$ 이며, $w_{k_T}^{(t,i)}$와 $w_{k_I}^{(t,i)}$는 각각 temporal, intensity 평활화 계수로서 식 (4)와 같다.

$$w_{k_T}^{(t,i)} = \exp\left\{-\frac{d^{2}(t,t+i)}{2\sigma_{T}^{2}}\right\} \tag{4a}$$

$$w_{k_I}^{(t,i)} = \exp\left\{-\frac{d^{2}(C^{(t,k)}, C^{(t+i,k)})}{2\sigma_{I}^{2}}\right\} \tag{4b}$$

이 때 $d(\cdot,\cdot)$은 유클리디안(Euclidean) 거리이다. Temporal 평활화 계수 $w_{k_T}^{(t,i)}$는 현재 프레임으로부터 시간 상으로 거리가 먼 켑스트럼이 평활화에 영향을 덜 미치게 한다. Intensity 평활화 계수 $w_{k_I}^{(t,i)}$는 현재 프레임의 켑스트럼과 $t+i$번째 프레임의 켑스트럼의 차이가 클수록 평활화에 영향을 미치지 못하게 하여 특징의 경계 정보를 보존한다. 식 (4)의 σ_T와 σ_I는 실험적으로 구할 수 있다.

3. 제안한 특징 보상 알고리즘

에지 보존 필터는 필터뱅크와 프레임별로 가중치를 구해야

하므로 계산량이 많아 실제 사용하는 데에 제한이 있다. 이에 비해 ARMA 필터는 적은 계산량으로 높은 성능을 내기 때문에 실시간으로 충분히 동작할 수 있는 장점이 있다[7]. 본 논문에서는 ARMA필터를 기반으로 하여 음성의 존재에 대한 가중치를 적용하는 방식과 켑스트럼의 전역 평균과 분산으로 특징을 정규화하는 방식을 제안한다.

3.1 가중 ARMA 필터

잡음에서 음성으로 또는 음성에서 잡음으로 변화하는 구간에서 기존 ARMA 필터는 음성구간의 특징을 보상할 때 인접한 잡음구간의 특징을 사용하게 되므로 잡음구간의 특징이 음성구간의 특징에 영향을 미치며, 이는 ARMA 필터의 성능을 떨어뜨리는 요인이 될 수 있다. 그래서 본 논문에서는 특징 보상 시 인접한 프레임의 잡음으로 인한 왜곡을 줄이기 위해 프레임별로 음성의 존재여부에 대한 가중치를 계산하여 ARMA필터의 계수로 두었다. 제안한 가중 ARMA필터는 아래의 식 (5)와 같다.

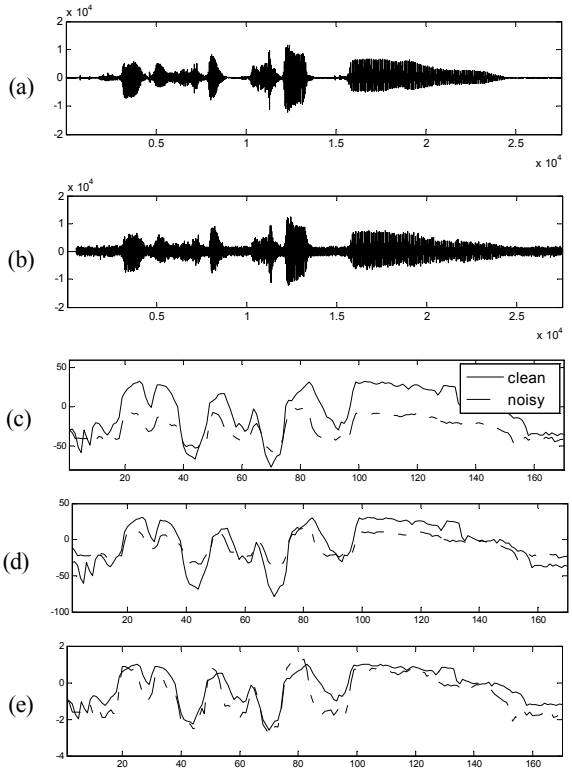
$$C_w^{(t,k)} = \frac{\sum_{i=1}^{m} w(t-i)C_w^{(t-i,k)} + \sum_{i=0}^{m} w(t+i)\hat{C}^{(t+i,k)}}{2m+1} \quad (5a)$$

여기서

$$w(t) = \frac{1}{1+e^{-\alpha \cdot x(t)}} \quad (5b)$$

이고, 이 때

$$x(t) = C^{(t,0)} - \overline{C}^{0} \quad (5c)$$



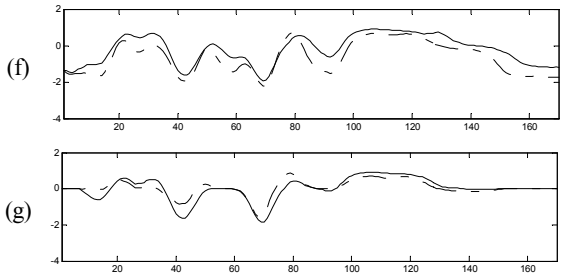


그림 1. 깨끗한 음성과 잡음 섞인 음성의 켑스트럼 궤적 비교
(a) 깨끗한 음성의 파형 (b) 잡음 섞인 음성의 파형
(c) No processing (d) CMN (e) CMVN (f) ARMA (g) 제안방식
Figure 1. Cepstrum contour comparison between clean and noisy speech (a) Clean speech wave (b) Noisy speech wave
(c) No processing (d) CMN (e) CMVN (f) ARMA
(g) Proposed method

이다. 식 (5a)에서 $k=1,\cdots,12$로 켑스트럼의 차수를 나타내고, $C_w^{(t,k)}$는 t번째 프레임의 k번째 정규화된 켑스트럼에 가중 ARMA 필터를 적용한 결과이다. 식 (5b)의 $w(t)$는 t번째 프레임에서의 음성의 존재에 대한 가중치이며 [0,1]의 범위를 갖는 시그모이드(sigmoid) 함수이다. 식 (5c)의 $C^{(t,0)}$는 t번째 프레임의 0번째 켑스트럼으로 에너지의 크기를 나타내며, $\overline{C}^{0}$는 0번째 켑스트럼의 평균이다. 따라서 $x(t)$를 사용한 $w(t)$는 음성의 존재여부의 정도를 나타낸다. 여기서 α는 양의 상수이고 m은 필터의 차수이다. 제안한 방식의 효과를 <그림 1>에서 살펴보았다. <그림 1>에서 깨끗한 음성과 10dB SNR로 백색잡음을 더한 음성을 사용하여 첫번째 켑스트럼(C_1)의 궤적을 비교하였다. 아무런 처리를 하지 않은 (c)에서 잡음으로 인해 궤적의 차이가 큰 것을 관찰할 수 있다. CMN과 CMVN으로 이러한 차이를 많이 줄일 수 있지만 여전히 차이가 있으며, 이러한 차이는 켑스트럼의 1,2차 미분값에 대해 더 큰 차이를 가져올 수 있다. 이러한 차이는 인식률을 저하시키는데, 켑스트럼의 궤적에 ARMA필터를 적용함에 따라 고주파수 성분이 사라지고 궤적의 차이가 많이 줄어든다. ARMA필터에 제안한 가중치를 적용하면 인접한 잡음구간으로부터의 영향을 줄일 수 있고, 이는 앞서 언급한 기존 ARMA 필터의 약점을 보완하여 (g)에서처럼 궤적의 차이를 더 줄일 수 있다. 이러한 궤적 차이의 감소는 잡음 환경으로부터의 인식 음성 켑스트럼과 음향모델 사이의 불일치가 줄어든 것을 의미하며 따라서 음성인식의 성능을 향상시킬 수 있다. 대부분의 환경보상 알고리즘은 변화하는 잡음의 특성을 추정하지만 TMF는 잡음을 추정하지 않고, 잡음으로 인한 환경불일치에 민감하지 않은 특징을 추출하여 적은 계산량으로 효율적인 특징 보상을 할 수 있다.

3.2 전역 평균과 분산을 사용한 정규화

특징을 정규화하기 위해서 켑스트럼의 평균을 사용한 정규화 과정(CMN)을 많이 사용한다[9]. CMN은 채널의 영향을 제

거하기 위한 목적으로 사용되는데, 훈련환경에서 켑스트럼 영역의 채널 성분과 음성 사이의 관계는 식 (6)과 같다.

$$\mathbf{y}_{tr}^{(n,t)} = \mathbf{x}_{tr}^{(n,t)} + \mathbf{h}_{tr} \tag{6a}$$

$$\overline{\mathbf{y}}_{tr}^{n} = \frac{1}{T_n}\sum_{t=1}^{T_n}\mathbf{y}_{tr}^{(n,t)} = \frac{1}{T_n}\sum_{t=1}^{T_n}(\mathbf{x}_{tr}^{(n,t)} + \mathbf{h}_{tr}) = \overline{\mathbf{x}}_{tr}^{n} + \mathbf{h}_{tr} \tag{6b}$$

$$\hat{\mathbf{y}}_{tr}^{(n,t)} = \mathbf{y}_{tr}^{(n,t)} - \overline{\mathbf{y}}_{tr}^{n} = \mathbf{x}_{tr}^{(n,t)} - \overline{\mathbf{x}}_{tr}^{n} \tag{6c}$$

인식환경에서 켑스트럼 영역의 채널 성분과 음성 사이의 관계는 식 (7)과 같다.

$$\mathbf{y}_{ts}^{(n,t)} = \mathbf{x}_{ts}^{(n,t)} + \mathbf{h}_{ts} \tag{7a}$$

$$\overline{\mathbf{y}}_{ts}^{n} = \frac{1}{T_n}\sum_{t=1}^{T_n}\mathbf{y}_{ts}^{(n,t)} = \frac{1}{T_n}\sum_{t=1}^{T_n}(\mathbf{x}_{ts}^{(n,t)} + \mathbf{h}_{ts}) = \overline{\mathbf{x}}_{ts}^{n} + \mathbf{h}_{ts} \tag{7b}$$

$$\hat{\mathbf{y}}_{ts}^{(n,t)} = \mathbf{y}_{ts}^{(n,t)} - \overline{\mathbf{y}}_{ts}^{n} = \mathbf{x}_{ts}^{(n,t)} - \overline{\mathbf{x}}_{ts}^{n} \tag{7c}$$

위의 식 (6), (7)에서 $\mathbf{x}^{(n,t)}$, $\mathbf{y}^{(n,t)}$는 각각 깨끗한 음성과 채널을 통과한 음성으로 n번째 문장의 t번째 프레임의 켑스트럼 벡터를 나타내며, $\mathbf{h}$는 채널 성분의 켑스트럼 벡터를 나타낸다. 또한 $\bar{\mathbf{y}}^{n}$는 n번째 문장의 켑스트럼 평균을 나타내며, T_n는 n번째 문장의 프레임수이다. $\hat{\mathbf{y}}^{(n,t)}$는 n번째 문장의 t번째 프레임에서 보상된 켑스트럼 벡터를 나타낸다. 또한 위 식에서 사용한 아래 첨자 tr과 ts는 각각 훈련환경과 인식환경을 뜻한다. 채널성분이 들어있는 $\mathbf{y}^{(n,t)}$에서 켑스트럼의 평균 $\bar{\mathbf{y}}^{n}$을 빼면 채널성분을 제거할 수 있다. 하지만 훈련환경과 인식환경에 사용한 서로 다른 음성의 평균 $\bar{\mathbf{x}}^{n}$이 존재해서 훈련환경과 인식환경의 불일치를 가져올 수 있다. 즉, $\bar{\mathbf{x}}_{tr}^{n} \neq \bar{\mathbf{x}}_{ts}^{n}$이다. 특히 T_n의 값이 작을 때 $\bar{\mathbf{x}}^{n}$은 특정 음성의 켑스트럼 성분에 많이 편중된다. 본 논문에서는 이러한 문제를 해결하기 위해 전체 훈련 데이터의 켑스트럼에 대한 전역 평균을 구하여 CMN을 수행한다. 이에 대한 과정은 아래와 같다. 훈련과정에서의 전역 평균 $\bar{\mathbf{y}}_{g_tr}$와 인식과정에서의 전역 평균 $\bar{\mathbf{y}}_{g_ts}$는 각각 식 (8), (9)과 같다.

$$\begin{aligned}\overline{\mathbf{y}}_{g\_tr} &= \frac{1}{\sum_{n=1}^{N}T_n}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\mathbf{y}_{tr}^{(n,t)} \\ &= \frac{1}{\sum_{n=1}^{N}T_n}\sum_{n=1}^{N}\sum_{t=1}^{T_n}(\mathbf{x}_{tr}^{(n,t)} + \mathbf{h}_{tr}) = \overline{\mathbf{x}}_{g\_tr} + \mathbf{h}_{tr}\end{aligned} \tag{8}$$

$$\begin{aligned}\overline{\mathbf{y}}_{g\_ts} &= \frac{1}{\sum_{n=1}^{N}T_n}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\mathbf{y}_{ts}^{(n,t)} \\ &= \frac{1}{\sum_{n=1}^{N}T_n}\sum_{n=1}^{N}\sum_{t=1}^{T_n}(\mathbf{x}_{ts}^{(n,t)} + \mathbf{h}_{ts}) = \overline{\mathbf{x}}_{g\_ts} + \mathbf{h}_{ts}\end{aligned} \tag{9}$$

여기서 $\sum_{n=1}^{N}T_n$은 N개 문장의 전체 프레임수이다. 위와 같이 구한 $\bar{\mathbf{y}}_{g_tr}$과 $\bar{\mathbf{y}}_{g_ts}$를 사용하여 CMN을 수행하면 $\bar{\mathbf{x}}_{g_tr} \approx \bar{\mathbf{x}}_{g_ts}$이기 때문에 환경 불일치를 더 줄일 수 있다. 본 논문에서는 훈련 데이터의 전역 평균과 분산을 사용하여 식 (10)과 같이 훈련 데이터를 정규화시킨다.

$$\hat{y}_{tr}^{(n,t,k)} = \frac{y_{tr}^{(n,t,k)} - \bar{y}_{g\_tr}^{k}}{\sigma_{g\_tr}^{k}} \tag{10}$$

여기서 $\bar{y}_{g_tr}^{k}$는 식 (8)의 켑스트럼 평균 벡터 $\bar{\mathbf{y}}_{g_tr}$의 k번째 스칼라 값이다. 이 때 전역 분산 $(\sigma_{g_tr}^{k})^2$는 아래와 같이 구한다.

$$(\sigma_{g\_tr}^{k})^2 = \sqrt{\frac{1}{\sum_{n=1}^{N}T_n}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(y_{tr}^{(n,t,k)} - \bar{y}_{g\_tr}^{k}\right)^2} \tag{11}$$

실제로 인식환경에서는 테스트 데이터를 미리 알 수 없기 때문에 이전 인식실험까지의 누적된 입력 문장의 전역 평균과 분산을 사용하여 식 (12)와 같이 특징을 정규화시킨다.

$$\hat{y}_{ts}^{(n,t,k)} = \frac{y_{ts}^{(n,t,k)} - \bar{y}_{g\_ts}^{k}(1:n-1)}{\sigma_{g\_ts}^{k}(1:n-1)} \tag{12}$$

여기서 사용한 평균과 분산은 각 각 식 (13), (14)와 같다.

$$\bar{y}_{g\_ts}^{k}(1:n) \cong \lambda_n \cdot \bar{y}_{g\_ts}^{k}(1:n-1) + (1-\lambda_n)\cdot \bar{y}_{g\_ts}^{k}(1:n) \tag{13}$$

$$\sigma_{g\_ts}^{k}(1:n)^2 \cong \lambda_n \cdot \sigma_{g\_ts}^{k}(1:n)^2 + (1-\lambda_n)\cdot \sigma_{g\_ts}^{k}(1:n)^2 \tag{14}$$

여기서 $\bar{y}_{g_ts}^{k}(1:n)$와 $\sigma_{g_ts}^{k}(1:n)^2$는 1번째 문장부터 n번째 문장을 사용한 전역 평균과 분산을 나타낸다. 또한 λ_n은 n번째 문장을 추가함에 따라 식 (15)와 같이 구한다.

$$\lambda_n = \begin{cases} \sum_{p=1}^{n-1}T_p / \sum_{p=1}^{n}T_p & n < n' \\ \alpha & otherwise \end{cases} \tag{15}$$

위 식에서 n'과 α는 실험을 통해서 구한다. 제안한 전역 평균과 분산을 이용한 정규화 방식은 잡음의 통계적 특성이 변화가 없다면 효과적으로 특징을 보상할 수 있지만, 잡음의 통계적 특성이 변화한다면 추정된 전역 평균과 분산의 신뢰도가 떨어질 수 있다. 본 논문에서는 인식 환경의 변화가 크지 않은 환경을 가정하여 제안한 알고리즘의 성능을 평가하였다.

4. 성능평가

제안한 특징 보상 알고리즘 성능평가에 앞서 식 (5)의 $w(t)$에서 상수 α의 영향을 알아본다. 그 후 두 종류의 시뮬레이션 데이터와 real 데이터를 사용하여 인식실험을 하였다. 시뮬레이션 데이터는 깨끗한 음성에 부가잡음을 더하거나, room impulse response(RIR)를 convolution하여 구성하였다. 시뮬레이션 데이터를 사용한 실험에서는 제안한 특징 보상 알고리즘의 각 모듈별 성능향상 기여도를 확인하였다. Real 데이터는 마이크와 스피커 사이의 거리를 1m로 떨어뜨린 후 스피커로 깨끗한 음성을 재생한 후 녹음하였다[11]. 마이크 주변에는 백색 잡음 또는 에어컨 잡음이 존재하도록 했다. Real 데이터를 사용한 실험에서는 제안한 특징 보상 알고리즘을 에지 보존 필터에 적용하고 기존의 다른 알고리즘과 성능을 비교하였다. 성능평가에서 훈련 DB는 ETRI의 POW DB를 사용하였고, 테스트 DB는 국어공학 연구 센터의 PBW DB를 사용하였다. 특징은 에너지를 제외한 38차 MFCC를 사용하였고, 인식기는 HTK를 사용하였다. 음향 모델은 tied-state triphone모델 2291개를 사용하였다.

4.1 가중치에서 상수 α의 영향

제안한 가중치 $w(t)$를 계산하는 데 사용한 시그모이드 함수는 상수 α에 따라 <그림 2>와 같이 변한다. 또한 식 (5b)에서 $\alpha \cdot x(t)$의 α값의 크기를 조절하면 음성의 에너지 gain의 영향을 살펴볼 수 있다. α를 0.2에서 0.9까지 변화시키면서 함수의 곡선을 그렸다. 이러한 α값의 변화에 따른 인식률은 <그림 3>에서 나타내었다. 테스트에는 10dB SNR로 백색잡음을 더한 시뮬레이션 데이터를 사용하였다. 실험에서 α가 0.2 이상의 값일 때 성능이 비슷한 수준을 유지하는데, 이를 통해 제안한 방식이 에너지의 gain에 큰 영향을 받지 않음을 알 수 있다. 본 논문에서 α는 0.4로 하였다.

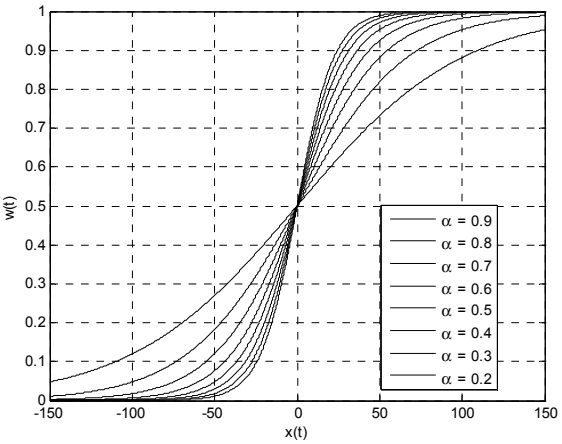


그림 2. α값에 따른 시그모이드 함수
Figure 2. Sigmoid function according to α

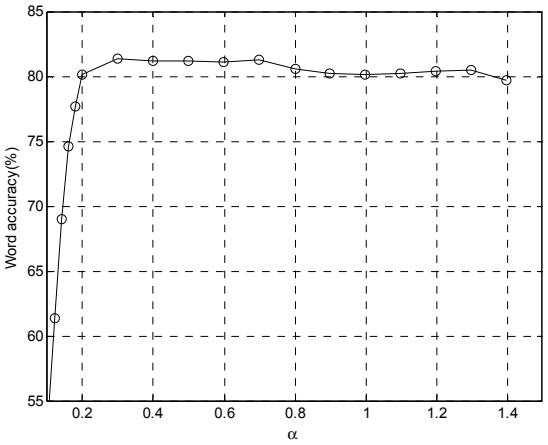


그림 3. α 값에 따른 단어 인식률
Figure 3. Word accuracy according to α

4.2 시뮬레이션 데이터에 의한 실험결과

깨끗한 음성에 백색 잡음을 SNR별로 더하여 테스트 데이터를 구성한 후, <그림 4>와 같이 단어 인식률로 제안한 특징 보상 알고리즘을 평가하였다. 이 때 SNR별로 다른 인식 환경으로 가정하여 다른 전역 평균과 분산을 구하였다. <그림 4>에서 문장별로 특징의 평균과 분산을 이용하여 정규화한 것을 uCMN과 uCMVN으로 나타내었다. 이렇게 정규화한 특징을 ARMA필터에 통과시켰을 때의 결과를 uARMA로 표시하였다. 제안한 가중치를 사용한 방식을 uwARMA, 전역 평균과 분산을 사용하여 정규화한 것을 gwARMA로 표시하였다. 또한 에지 보존 필터를 사용한 특징 보상 방식은 Edge로 표시하였다. <그림 4>에서 uCMN과 uCMVN을 적용함에 따라 성능이 향상됨을 확인할 수 있다. 그런데 uARMA는 uCMVN에 비해 성능이 약간 떨어진다. 이는 ARMA필터링하는 과정에서 비음성 구간의 특징이 음성구간의 특징을 왜곡시킬 수 있기 때문이다. 이러한 왜곡의 영향은 음성의 존재에 대한 가중치를 적용함으로 줄일 수 있다. 또한 전역 평균과 분산을 함께 사용했을 때 가장 높은 성능을 보인다.

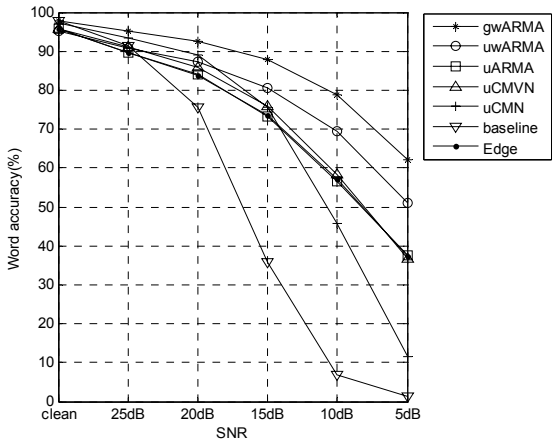


그림 4. 백색잡음을 더한 데이터에서의 단어 인식률
Figure 4. Word accuracy for the data added by white noise

원거리(distance talking) 환경에서 반향의 영향을 고려하기 위해서 깨끗한 음성에 RIR을 convolution하여 두번째 시뮬레이션 데이터를 구성하였다. 사용한 RIR은 mirror image method를 사용하여 생성하였다[10]. 이 때 거리별로 다른 인식 환경으로 가정하여 다른 전역 평균과 분산을 구하였다. <그림 5>에서 baseline 성능은 0.5m의 거리에서도 성능이 심각하게 떨어짐을 알 수 있다. 그리고 부가잡음을 더한 데이터의 결과와는 달리 CMN을 적용했을 때 성능의 감소가 있었다. 하지만 제안한 특징 보상 알고리즘을 사용했을 때 큰 성능 향상을 보인다.

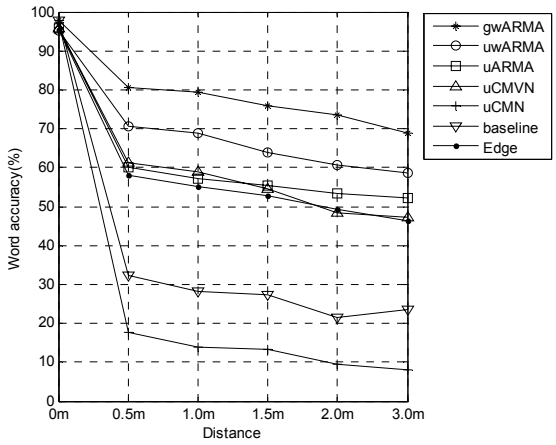


그림 5. 반향이 있는 음성에서 거리에 따른 단어 인식률
Figure 5. Word accuracy for the reverberant data from different distance

4.3 Real 데이터에 의한 실험결과

본 논문에서는 real 데이터를 크기가 540cm x 1050cm x 330cm 인 실험실에서 <그림 6>과 같이 마이크와 스피커를 1m 거리로 떨어뜨리고 스피커로 깨끗한 음성을 play시켜 녹음하였다. 마이크 주변에는 백색잡음 또는 에어컨 잡음이 존재하도록 하였다. 이 때 마이크로부터 취득한 백색잡음과 에어컨 잡음이 섞인 음성의 SNR은 약 20dB, 10dB 수준이다.

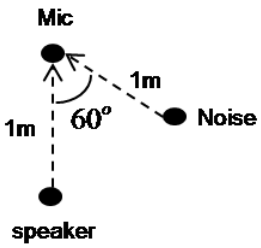


그림 6. Real 데이터 취득환경
Figure 6. Environment for acquiring the real data

<그림 7>에서는 제안한 알고리즘을 기존의 다른 알고리즘과 성능을 비교하였다. 성능 비교에는 에지 보존 기반의 TMF와 ETSI의 advanced front-end(AFE) 특징 보상 알고리즘을 사용하여 단어인식률로 비교하였다[12]. <그림 7>의 결과로부터 제안한 알고리즘이 다른 알고리즘에 비해 우수한 성능을 보임을 알 수 있다.

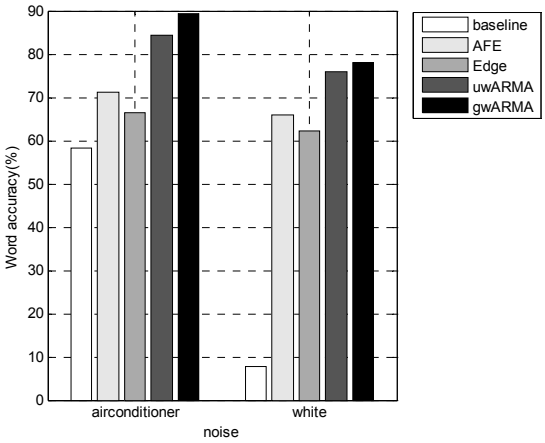


그림 7. 다른 특징 보상 알고리즘과의 성능 비교
Figure 7. Performance comparison of the other feature compensation algorithms

5. 결론

본 논문에서는 원거리 환경에서 부가잡음과 반향에 강인한 특징 보상 알고리즘을 제안하였다. 제안한 알고리즘에서는 기존의 ARMA필터 기반의 TMF 방식에서 음성의 존재여부 정도에 따른 가중치를 적용하였다. 또한 켑스트럼의 평균과 분산을 전역으로 구한 후 특징을 정규화하여 환경 불일치를 더 감소시켰다. 제안한 알고리즘의 성능은 부가잡음과 반향에 대한 시뮬레이션 데이터와 실제 환경에서 녹음한 데이터로 평가하였고, 기존의 AFE에 비해 우수한 성능을 보임을 확인하였다. AFE의 복잡한 처리과정을 고려할 때 제안한 특징 보상 알고리즘은 실제 사용하는 데에 유용할 것이다. 향후 다른 음질개선 모듈과 효과적으로 결합하여 더 열악한 환경 불일치 문제를 극복하는 연구를 수행할 예정이다.

감사의 글

이 논문은 지식경제부 지원으로 수행되는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술 개발 사업)의 일환으로 수행되었음.

참고문헌

[1] Gales, M. J. F., Young, S. J. (1996). “Robust continuous speech

recognition using parallel model combination", IEEE Trans. On Speech and Audio Proc., vol. 5, no. 5, pp. 352-359, Sep

[2] Moreno, P. J., Raj, B., Stern, R. M. (1996). "A vector Taylor series approach for environment-independent speech recognition", Proc. of ICASSP, vol. 2, pp. 733-736

[3] Boll, S. F. (1979). "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust. Speech Signal Process. vol. 27, no. 2, 113-120.

[4] Ephraim, Y., Malah, D. (1984). "Speech enhancement using a minimum mean square error short time spectral amplitude estimator", IEEE Trans. Acoust. Speech Signal Process. vol. 32, no. 6, pp. 1109-1121.

[5] Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M. (2009). "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction", IEEE Trans. Audio Speech Language Process. vol. 17, no. 4, pp. 534-545.

[6] Kanedera, N., Arai, T., Hermansky, H., Pavel, M. (1999). "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Comm. vol. 28, no. 1, pp. 43-55.

[7] Chen, C. P., Bilmes, J. (2007). "MVA processing of speech features", IEEE Trans. Audio Speech Language Process. vol. 15, no. 1, pp. 257-270.

[8] Lu, X., Matsuda, S., Unoki, M., Nakamura, S. (2010). "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition", Speech Comm. Vol. 52, no. 1, pp. 1-11.

[9] Viikki, O., Laurila, K. (1998). "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Comm., vol. 25, pp. 133-147

[10] McGovern, S. (2003). A Model for Room Acoustics, [On-line]. Available: http://2pi.us/rir.html

[11] 반성민, 김형순, (2010). "강인한 음성인식을 위한 Temporal Modulation 필터 비교", 제23회 신호처리합동 학술대회 논문집, pp. 595-596.

[12] ETSI standard doc. "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm", ETSI ES 202 050 Ver.1.1.1 (2002-10)

• **반성민 (Ban, Sung Min)**
부산대학교 전자전기공학부
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-1704 Fax: 051-510-4279
Email: bansungmin@pusan.ac.kr
관심분야: 음성인식, 음성전처리
현재 전자전기공학부 대학원 박사과정 재학중

• **김형순 (Kim, Hyung Soon)**
부산대학교 전자전기공학부
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-2452
Email: kimhs@pusan.ac.kr
관심분야: 음성인식, 음성합성, 음성전처리
1992~현재 전자전기공학부 교수