확률적 스펙트럼 차감법을 이용한 잡음 환경에서의 음성인식

Noisy Speech Recognition using Probabilistic Spectral Subtraction

지 상 문*, 오 영 환*
(Sang-Mun Chi*, Yung-Hwan Oh*)

요 약

본 논문에서는 잡음환경에서의 음성인식을 위하여 잡음의 확률적 특성과 음성모델을 이용하는 확률적 스펙트럼 차감법을 제안한다. 기존의 스펙트럼 차감법은 음성이 존재하지 않는 구간에서 추정한 잡음을 잡음음성에서 차감하여 잡음을 제거하므로, 추정한 잡음의 형태가 음성인식기에 입력되는 잡음음성에 포함된 잡음과 상이한 특성을 나타낼 경우에는 효과적인 잡음의 제거가 불가능하다. 이러한 단점을 보완하기 위해서 여러 가지 형태를 가지는 잡음의 원형을 사용하여, 잡음음성에서 잡음을 제거하는 방법을 사용하였다. 잡음의 확률적인 특성을 여러 개의 잡음원형으로 나타내므로, 스펙트럼 차감법은 입력음성에 대해서 확률적으로 수행되어 잡음이 제거된 다중의 스펙트럼을 출력하게 되고, 인식시에는 조용한 환경의 음성으로 학습된 음성모델에 따른 최적의 스펙트럼을 이용하여 인식을 수행한다. 또한 정적인 파라미터와 동적인 특징파라미터를 동시에 고려하여 잡음을 영향을 최소화하므로 보다 효과적인 잡음처리가 가능하다.

제안한 방법의 타당성을 실험적으로 검증하기 위해서, 잡음환경하의 음성인식에 적용하였다. SNR 10 dB인 50개의 고립단어에 대한 실험결과, 잡음처리를 하지 않았을 경우 72.75%, 스펙트럼 차감법은 80.25%, 제안한 방법을 사용하였을 경우는 86.25%의 인식률을 얻음으로써, 효과적인 잡음처리 방법임을 확인할 수 있었다.

ABSTRACT

This paper describes a technique of probabilistic spectral subtraction which uses the knowledge of both noise and speech so as to reduce automatic speech recognition errors in noisy environments. Spectral subtraction method estimates a noise prototype in non-speech intervals and the spectrum of clean speech is obtained from the spectrum of noisy speech by subtracting this noise prototype. Thus noise can not be suppressed effectively using a single noise prototype in case the characteristics of the noise prototype are different from those of the noise contained in input noisy speech. To modify such a drawback, multiple noise prototypes are used in probabilistic subtraction method. In this paper, the probabilistic characteristics of noise and the knowledge of speech which is embedded in hidden Markov models trained in clean environments are used to suppress noise. Futhermore, dynamic feature parameters are considered as well as static feature parameters for effective noise suppression.

The proposed method reduced error rates in the recognition of 50 Korean words. The recognition rate was 86.25% with the probabilistic subtraction, 72.75% without any noise suppression method and 80.25% with spectral subtraction at SNR (Signal-to-Noise Ratio) 10 dB.

I. 서 론

현재의 음성인식기술은 조용한 환경에서는 이미 높은 성능을 나타내고 있지만, 실제 현장의 여러 가지 요인에 의해 성능이 크게 저하된다. 잡음의 첨가에 의한 환경의 불일치는 음성인식기의 성능저하의 요인으로서, 음성인식시스템을 제작할 때 사용한 학습음성과 실제현장에서 입력되는 음성의 특성의 변이 때문에 발생한다. 따라서, 잡음을 제거하여 환경의 변화에 민감하지 않은 음성인식기를 개발하려는 연구가 음성인식기의 실용화를 위한 기반기술로 많은 관심의 대상이 되고 있다.

잡음환경에 강인한 음성인식을 위해 여러 접근방법이 사용되고 있는데, 인식모델에 포함된 음성의 특성을 이용하는가에 따라 두 가지로 분류할 수 있다.

첫 번째 접근방법은 음성인식의 전 단계에서 잡음을 제거하는 방법으로, 잡음에 강인한 특징추출과 거리척도인 SMC(short-time modified coherence)[1], RASTA(RelAtive

*한국과학기술원 전산학과 및 인공지능 연구센터
접수일자 : 1997년 6월 20일

SpecTrAl) 처리와 동적인 특징파라미터[2, 3], 켑스트럼 식상척도[4] 등이 있고, 음성신호에 포함된 잡음을 제거하는 방법인 스펙트럼 차감법[5, 6], 청각기관의 특성을 이용하는 방법[7]이 있다. 이러한 방법들은 음성인식기와 독립적인 처리가 가능하고 비교적 계산량이 적다는 장점이 있으나, 음성을 모델링 할 때에 잡음대칭 음성의 특성을 이용하지 못하고 비화하는 잡음을 적절히 처리할 수 없다는 단점이 있다.

두 번째 접근방법은 음성모델의 파라미터를 이용하여 잡음을 제거하는 방법으로, 베이지안 추정으로 음질을 개선하는 방법[8]은 음성과 잡음을 ARHMM(autoregressive hidden Markov model)으로 모델링한 후, MMS(minimum mean square) 추정은 음성신호와 잡음신호로 이루어진 은닉마르코프 모델의 상태에서 잡음음성을 위너필터링 한 결과의 가중합으로 깨끗한 음성을 추정하고, MAP (maximum a posteriori) 추정은 음성과 잡음의 상태들에서 위너필터링한 값의 조화평균으로 깨끗한 음성을 추정하는데, 은닉마르코프 모델의 상태를 Viterbi 알고리즘을 사용해서 찾는 과정과 찾아진 상태에서의 위너필터링을 하는 두 가지 과정이 반복적으로 수행한다. 이방법은 음성인식을 위한 처리가 아니고, 잡음음성에서 잡음을 제거하여 음질을 개선하는 방법이다. 음성인식을 위해서 인식모델의 파라미터를 잡음환경에 적응하는 방법[9]은 잡음과 조용한 환경의 음성으로 각각 학습된 HMM의 각 상태에서의 켑스트럼 계수의 평균과 분산을 IDCT(inverse discrete cosine transform)하여 로그 스펙트럼 영역으로 바꾸고, 로그 스펙트럼 영역에서의 평균, 분산과 스펙트럼 영역에서의 평균, 분산과의 변환관계를 이용하여 스펙트럼 영역에서의 잡음의 첨가를 로그 스펙트럼영역으로 변환하고 DCT하여 잡음환경의 HMM의 파라미터를 구한다. 두 번째 방법은 잡음과 음성의 특성을 모두 이용하므로 첫 번째 접근방법에 비해서 효과적이나, 음성인식기에 종속적인 처리가 필요하고 비교적 계산량이 많다는 단점이 있다.

본 논문에서는 음성모델의 파라미터를 잡음의 제거에 이용할 수 있도록, 특징추출단계에서 잡음의 확률적인 특성을 고려하여 하나의 음성신호에 대해서 다중의 특징파라미터를 추출하고, 인식단계에서 음성모델에 따른 특징파라미터를 이용할 수 있게 한다. 본 연구에서는 잡음의 특징을 이용하여 다중의 특징파라미터를 추출하기 위해서 확률적 스펙트럼 차감법을 사용하였다. 일반적인 스펙트럼 차감법은 음성이 존재하지 않는 구간에서 추정한 잡음을 잡음음성에서 차감하여 잡음을 제거하므로, 추정한 잡음의 형태가 음성인식기에 입력되는 잡음음성에 포함된 잡음과 상이한 특성을 나타낼 경우에는 효과적인 잡음의 제거가 불가능하다. 이러한 문제를 피하기 위해서 확률적 스펙트럼 차감법에서는 음성이 존재하지 않는 구간에서 벡터양자화를 사용하여 잡음의 스펙트럼을 군집화하여 여러개의 잡음의 원형을 구한 후, 잡음의 원형이 잡음구간에서 나타난 확률에 따라 입력되는 잡음음성에서 차감하여 잡음이 제거된 다중의 스펙트럼을 생성한다. 잡음이 제거된 스펙트럼은 음성인식을 위해서 인식대상의 단어에 해당하는 여러 개의 DHMM(discrete HMM)에 입력되는 관찰열이 되며, 각각의 DHMM은 각 모델의 상태에서의 출력분포에 의적으로 대응되는 스펙트럼을 이용할 수 있으므로, 스펙트럼 차감법에서와 같이 하나의 스펙트럼만을 사용하는 것에 반하여 모델의 특성을 인식에 반영할 수 있다. 또한 인식확률의 계산에는 잡음이 제거된 정적인 파라미터열만을 이용하는 것이 아니라, 잡음이 제거된 정적인 파라미터열에서 구한 동적인 특징파라미터열도 동시에 고려하므로 보다 효과적인 잡음처리가 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 확률적 스펙트럼 차감법을 설명하고, 3장에서는 확률적 스펙트럼 차감법에 의해서 구해진 스펙트럼열을 입력으로 하여 DHMM을 사용한 음성인식을 설명한다. 4장에서는 음성인식 실험을 통해 제안한 방법의 타당성을 검토하고, 5장에서 결론과 추후연구에 대해서 알아본다.

Ⅱ. 확률적 스펙트럼 차감법

잡음환경에서는 화자의 발성에너지가 증가하고 피치나 스펙트럼 구조가 변이된 음성이 발성되므로, 조용한 환경의 음성과는 특성이 상이하지만[10, 13], 본 연구에서는 가산잡음의 첨가에 따른 음성인식기의 성능저하에 대해서만 고려하기로 한다. 잡음음성 신호는 다음과 같이 나타낼 수 있다.

$$z(k) = x(k) + n(k) \tag{1}$$

여기서 $z(k)$, $x(k)$, $n(k)$는 각각 잡음음성, 깨끗한 음성, 잡음이고, k는 시간을 나타낸다. 주파수 영역에서 식 1은 다음과 같이 표현된다.

$$Z(\omega) = X(\omega) + N(\omega) \tag{2}$$

여기서 $Z(\omega)$, $X(\omega)$, $N(\omega)$은 각각 $z(k)$, $x(k)$, $n(k)$의 푸리어 변환을 통해 얻은 값이고, ω는 주파수를 나타낸다.

본 장에서는 식 2에서 잡음의 스펙트럼 성분인 $N(\omega)$를 제거하는 방법을 설명하는데, 이에 앞서 스펙트럼 영역에서의 처리를 위해서 사용하는 인간의 청각특성을 반영하여 스펙트럼을 추출하는 방법인 주대역 스펙트럼(critical-band spectrum) 분석을 설명한다[11].

2.1 스펙트럼 분석을 위한 필터뱅크의 구성

스펙트럼 차감법등의 스펙트럼 영역에서의 처리는 푸리어 변환의 값을 그대로 사용하는 것보다 필터뱅크의 출력값을 사용하는 것이 주파수상의 평활화 효과로 인해

서 잡음환경에 강인한 장점이 있다. 주대역 스펙트럼은 다음의 과정을 통해서 구한다.

먼저 푸리에 변환을 통해서 얻은 파워 스펙트럼을 $P(\omega)$라고 하면, $P(\omega)$를 인간의 청각기 인지 단위인 바크(Bark)단위 주파수축 Ω로 변환한다.

$$\Omega(\omega) = 6\ln\{f/600 + [(f/600)^2 + 1]^{1/2}\} \tag{3}$$

여기서 f는 주파수이다. 이렇게 주파수에서 바크주파수로 변환된 파워 스펙트럼은 청각필터를 근사한 함수인 주대역 매스킹 곡선 $\Psi(\Omega)$와 컨볼루션된다.

$$\Psi(\Omega) = \begin{cases} 0, & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)}, & -1.3 \le \Omega \le -0.5 \\ 1, & -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5)}, & 0.5 \le \Omega \le 2.5 \\ 0, & \Omega > 2.5 \end{cases} \tag{4}$$

$\Omega(\Omega)$를 이용하여 바크 주파수 단위의 파워 스펙트럼을 얻는다.

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega_i - \Omega)\Psi(\Omega) \tag{5}$$

본 연구에서는 첫 번째 필터의 중심주파수 Ω_0는 bark단위로 대략 0.985이고 첫 번째 필터의 중심주파수로부터 bark단위로 약 0.985씩 중심주파수의 간격을 이루고 마지막 중심주파수 Ω_{18}는 bark 단위로 18.723이고 6735Hz에 해당하는 19차원의 파워스펙트럼을 사용한다.

2.2 확률적 스펙트럼 차감법

스펙트럼 차감법은 가산잡음이 첨가된 잡음음성의 신호에서 잡음 스펙트럼의 크기 성분을 제거하는 방법으로서, 잡음과 음성신호사이에 상관관계가 없다는 가정과 음성을 인지하는 인간의 청각특성은 음성의 주파수 성분별 위상 정보보다는 크기 정보에 더 많이 영향을 받는다는 연구결과에 기초한다. 스펙트럼 차감법에서는 잡음을 제거하기 위해서, 잡음의 평균 스펙트럼을 잡음음성에서 제거한다.

$$|\hat{X}(\omega)|^b = |Z(\omega)|^b - |\overline{N(\omega)}|^b \tag{6}$$

여기에서 b는 잡음제거 정도에 가변성을 주는 파라미터이며, b=1인 경우는 스펙트럼의 크기를 차감하는 것이고 b=62인 경우는 파워스펙트럼 차감법(power spectral subtraction) 또는 자기상관 차감법(autocorrelation subtraction)이다. 본 논문에서는 b=1을 사용한다. 잡음의 평균 스펙트럼 $|\overline{N(\omega)}|$는 잡음만이 존재하는 구간에서 다음의 평균값으로 구한다.

$$|\overline{N(\omega)}| = \frac{1}{N}\sum_{i=1}^{N} |N_i(\omega)| \tag{7}$$

여기서 N_i는 잡음의 스펙트럼이고, 잡음의 길이는 N이다. 식 6으로 구한 스펙트럼은 잡음 스펙트럼의 크기의 변화에 따라 음수값이 되는 경우가 있는데, 이때에는 잡음음성에 1보다 작은 값 β를 곱하여 대치한다.

$$|\hat{X}(\omega)| = \beta\,|Z(\omega)| \tag{8}$$

스펙트럼 차감법은 잡음의 스펙트럼 형태를 미리 알고 있거나, 잡음의 스펙트럼을 추정하기에 충분한 잡음만이 존재하는 구간이 존재하고, 이 구간에서 안정적(stationary)인 특성을 갖고 있어야 잡음의 형태를 효과적으로 추정할 수 있다. 그러나 잡음은 시간에 따라 변이하므로 고정적인 하나의 잡음스펙트럼 평균값으로는 잡음을 효과적으로 제거할 수 없다. 본 연구에서는 이러한 문제를 해결하기 위해서 잡음의 평균 스펙트럼을 사용하는 대신에, 여러 개의 잡음원형을 가지는 확률적인 스펙트럼 차감법을 사용한다. 확률적인 스펙트럼 차감법에서는 다음과 같은 M개의 잡음원형만이 잡음음성에서 나타난다고 가정한다.

$$NP = \{N_i; i = 0, 1, \cdots, M-1\},\ \sum_{i=0}^{M} p_N(N_i) = 1 \tag{9}$$

여기서 NP는 잡음원형 N_i들의 집합이고, N_i가 잡음음성에서 나타날 확률은 $p_N(N_i)$이다. NP와 $p_N(N_i)$는 정규화된 에너지를 사용하는 간단한 끝점검출 방법을 사용하여 잡음음성에서 잡음구간을 분리하여, 잡음만이 존재하는 구간에서 잡음의 크기 스펙트럼을 벡터양자화를 통하여 M개의 잡음원형을 구하고, N_i가 잡음구간에서 나타나는 빈도에 따라 $p_N(N_i)$를 결정한다.

z_0^{L-1}를 시간길이가 L인 잡음음성의 스펙트럼 열이라 하면

$$z_0^{L-1} = (z_0, z_1, \cdots, z_{L-1}),\ z_t = [z_t(\omega_0), z_t(\omega_0), \cdots, z_t(\omega_{D-1})]^T \tag{10}$$

여기서 D는 필터뱅크의 개수이고, T는 벡터의 전치를 나타낸다. 스펙트럼 차감법은 잡음의 스펙트럼 평균 $\bar{n} = [n(\omega_0), n(\omega_1), \cdots, n(\omega_{D-1})]^T$가 고정되어 있기 때문에 깨끗한 음성의 스펙트럼열 $\hat{x}_0^{L-1}$은 다음 식과 같이 하나만 존재하지만,

$$\hat{x}_0^{L-1} = (\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_{L-1}),$$
$$|\hat{x}_i(\omega_j)| = |z_i(\omega_j)| - |\bar{n}(\omega_j)|,\ i = 0, 1, \cdots, D-1 \tag{11}$$

확률적 스펙트럼 차감법에서는 잡음음성에 포함된 잡음의 스펙트럼열 n_0^{L-1}은 다음과 같다.

$$n_0^{L-1} = (n_0, n_1, \cdots, n_{L-1}),$$

$$n_t=[n_t(\omega_0), n_t(\omega_1),\cdots,n_t(\omega_{D-1})]^T,\ |n_t|\in NP \tag{12}$$

여기서 각 시간 t에서 나타나는 잡음의 형태는 NP 중의 하나이고 이들은 확률적으로 독립이므로, n_0^{L-1}의 확률은 $\prod_{t=0}^{L-1} p_N(|n_t|)$이다. 따라서, 잡음음성의 스펙트럼열 z_0^{L-1}이 주어졌을 때, 이로부터 얻어지는 깨끗한 음성의 스펙트럼 열 $\hat{x}_0^{L-1}$은 확률 $\prod_{t=0}^{L-1} p_N(|n_t|)$로서 나타난다.

$$\hat{x}_0^{L-1}=(\hat{x}_0,\hat{x}_1,\cdots,\hat{x}_{L-1}),\ |\hat{x}_t(\omega_i)|=|z_t(\omega_i)|-|n_t(\omega_i)| \tag{13}$$

즉, 스펙트럼 차감법은 고정적인 잡음원형으로부터 하나의 깨끗한 음성의 스펙트럼열을 얻는 반면에, 확률적 스펙트럼 차감법에서는 확률적인 잡음원형으로부터 여러 개의 깨끗한 음성의 스펙트럼열을 얻는다.

Ⅲ. 확률적 스펙트럼 차감법과 HMM을 이용한 음성인식

본 장에서는 확률적 스펙트럼 차감법을 이용하여 구한 다중의 스펙트럼열과 이산형 HMM(discrete hidden Markov model:DHMM)을 이용해서 음성을 인식하는 과정을 설명한다.

HMM은 파라미터 집합 $\lambda=\{\pi_i, a_{ij}, b_i(x), i, j=0,\cdots,N-1\}$로 구성된다. 여기서 N은 상태의 수, π_i는 상태 i에서의 초기확률, a_{ij}는 상태 i에서 상태 j로 천이 하는 확률, $b_i(x)$는 상태 i의 출력 확률 밀도 함수이다. DHMM은 출력 확률 밀도 함수가 이산형 분포로서 연속분포를 가지는 연속형 HMM보다 양자화 오류가 있다는 단점이 있으나, 계산량이 작고 연속형 HMM에서 처럼 확률분포에 대한 가정이 없으므로, 임의의 분포를 모델링할 수 있는 장점이 있다. 본 연구에서 DHMM을 음성인식에 사용하여 미지의 관측 벡터열에 $x_0^{L-1}=(x_0, x_1,\cdots,x_{L-1})$에 대해 음성인식을 수행하였는데, 인식대상의 모든 어휘의 모델에 대해서 다음의 확률을 최대화하는 모델 λ를 선택한다.

$$P_1(X|\lambda)=\sum_{all\,S}\prod_{t=0}^{L-1} a_{s_t,s_{t+1}}b_{s_t}(x_t)) \tag{14}$$

여기서 확률은 모든 가능한 상태열 $S=(s_0, s_1,\cdots,s_{L-1})$에서 계산한다.

확률적 스펙트럼 차감법에서 얻은 스펙트럼 열들을 이용하여 음성인식을 하는 과정은 첫 번째로 스펙트럼열을 캡스트럼열로 변환한 후에, 두 번째로 전 단계에서 얻어진 특징파라미터를 이용하여 각 어휘의 모델에 대해서 최대확률을 갖는 모델을 선택한다. 첫 번째 단계에서는 식 13의 스펙트럼 $\hat{x}_0^{L-1}$로부터 다음의 식을 사용하여 캡스트럼 파라미터 $c_0^{L-1}=(c_0, c_1,\cdots,c_{L-1})$, $c_t=(c_t(0), c_t(1),\cdots, c_t(M-1))^T$로 변환한다.

$$c_t(i)=\sum_{k=0}^{D-1}\log(\hat{x}_t(\omega_k))\cos(i(k+0.5)\pi/D),\ i=0,1,\cdots M-1 \tag{15}$$

여기서 M은 캡스트럼 파라미터의 차원으로 본 연구에서는 14를 사용한다. 첫 번째 단계에서는 잡음처리 과정이 없는 캡스트럼 파라미터가 나타날 확률 $p_N(c_0^{L-1})=\sum_{t=0}^{L-1} p_N(|n_t|)$과 이러한 특징파라미터가 모델에서 발생할 확률을 모든 가능한 잡음의 스펙트럼 열 $n_0^{L-1}=(n_0, n_1,\cdots,n_{L-1})$에 대해서 계산한다.

$$\begin{aligned}
&\sum_{all\,n_0^{L-1}} p(c_0^{L-1}|\lambda)p_N(c_0^{L-1})\\
&=\sum_{all\,n_0^{L-1}}\{\sum_{all\,S}\prod_{t=0}^{L-1} a_{s_t,s_{t+1}}b_{s_t}(c_t)\}\prod_{t=0}^{L-1} p_N(|n_t|)\\
&=\sum_{all\,S}\sum_{all\,n_0^{L-1}}\prod_{t=0}^{L-1} a_{s_t,s_{t+1}}b_{s_t}(c_t)p_N(|n_t|)\\
&=\sum_{all\,S}\prod_{t=0}^{L-1} a_{s_{t-1},s_t}[\sum_{n_t\in NP} b_{s_t}(c_t)p_N(|n_t|)]
\end{aligned} \tag{16}$$

Ⅳ. 실험 및 검토

4.1 실험자료와 실험조건

효과적인 잡음처리 방법의 개발을 위해서는 실제 잡음환경에서 발성된 음성자료가 필요하나, 실제 잡음환경에서의 음성자료의 수집은 많은 시간과 노력을 필요로 한다. 본 논문에서는 인파가 많은 거리에서 발생한 자동차, 사람들의 음성과 발자국 소리 등의 잡음을, 헤드폰을 통하여 발성자에게 들려줌으로써 잡음환경을 모의하고, 모의된 잡음환경에서 발성된 음성을 수집하여 실험하였다[12, 13]. 실험에 사용한 잡음자료는 전자통신연구소에서 제공한 JEIDA(Japan electronic industry development association)에서 수집한 자료의 일부이다.

음성인식자료로서 HMM의 학습자료로서, 남자 5명, 여자 5명이 조용한 환경에서 50단어를 2회 반복 발성한 음성을 사용하였고, 평가를 위해서는 학습에 참여하지 않은 남자 2명과 여자 2명이 잡음환경에서 2회씩 발성한 음성자료를 사용하였다. 음성은 16KHz 샘플링, 16bit 양자화 되었고, $1-0.95z^{-1}$로 전처리하였다. 해밍창을 씌운 32 msec구간을 분석하여 14차의 캡스트럼 파라미터를 구하였다. 음성인식을 위한 파라미터는 캡스트럼, 캡스트럼 파라미터의 차분 파라미터, 정규화된 에너지, 차분에너지, 2차 차분에너지의 3종류이며, 각각 256, 256, 32개의 코드워드를 갖는 코드북을 사용하여 양자화하였고, 인식모델은 15개의 상태를 갖는 이산분포 HMM을 사용하였다.

비교실험을 위한 특징파라미터로는 다음과 같다.

NoPro:잡음처리 과정이 없는 캡스트럼 파라미터

SS:스펙트럼 차감법을 사용하여 구한 캡스트럼 파라미터

PSS(M=k):확률적 스펙트럼 차감법으로서 구한 캡스

그림 파라미터, 잡음원형의 개수는 k개

4.2 음성인식 실험결과

그림 1은 캡스트럼 파라미터만을 사용하여 SNR 10dB에서 음성인식 실험을 수행한 결과이다. 그림에서 보듯이 확률적 스펙트럼 차감법은 잡음처리를 하지 않은 것과 스펙트럼차감법보다 효과적임을 알 수 있다. 그림 1에서의 잡음의 원형은 잡음만이 존재하는 구간을 M개의 동일한 길이로 나누고 각 구간의 평균으로 잡음원형을 초기화하여 벡터양자화를 수행하여 M개의 잡음원형을 구하였다. 각 구간에서의 평균을 잡음원형으로 사용하였을 경우에는 M = 4, 5, 6일 때에 69.5%, 67.5%, 69.5%로 벡터양자화를 사용했을 경우보다는 인식률이 낮았다.

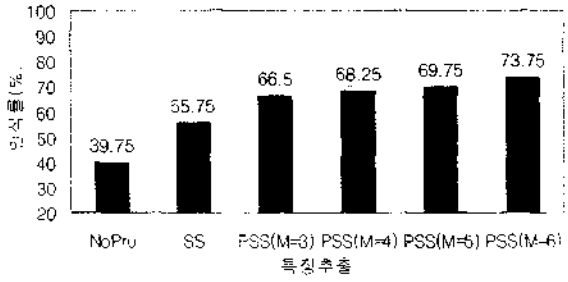

그림 1. 캡스트럼을 이용한 때의 음성인식률

그림 2는 캡스트럼, 캡스트럼 파라미터의 차분 파라미터, 정규화된 에너지, 차분에너지, 2차 차분에너지의 3종류의 파라미터를 사용하여 음성인식에 적용하였다. 차분 파라미터는 차분되는 두 개의 캡스트럼 파라미터의 확률의 곱으로 확률이 결정되고, M개의 잡음원형을 사용한 경우에는 M x M개의 동적인 파라미터가 각 프레임마다 생성된다.

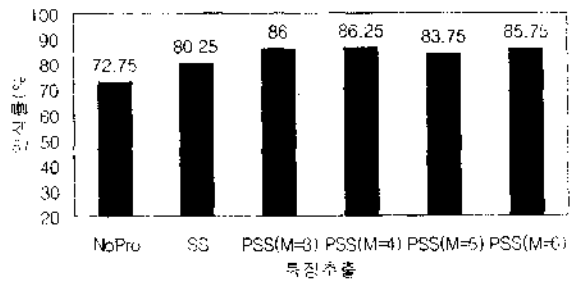

그림 2. 캡스트럼과 동적파라미터를 이용할 때의 음성인식률

그림 1과 2에서 보듯이 확률적 스펙트럼 차감법은 특징추출 파라미터에서는 잡음의 원형이 나타날 확률에 따라 다중의 파라미터열을 추출하고 인식단계에서 각 모델에 적합한 파라미터열이 선택되어 인식에 이용되므로 잡음에 의한 영향을 스펙트럼 차감법에 비해서 덜 받는 장점이 있다.

V. 결 론

본 논문에서는 잡음환경에서 음성인식기의 성능저하를 방지하기 위해 잡음의 확률적인 특성을 특징파라미터의 추출에 이용하는 확률적 스펙트럼 차감법을 제안하였다.

확률적 스펙트럼 차감법에서는 음성이 존재하지 않는 구간에서 벡터양자화를 사용하여 잡음의 스펙트럼을 군집화하여 여러 개의 잡음의 원형을 구한 후, 잡음의 원형이 잡음구간에서 나타날 확률에 따라 입력되는 잡음음성에서 차감하여 잡음이 제거된 다중의 스펙트럼을 생성한다. 이러한 다중의 스펙트럼은 불확실한 잡음의 형태를 스펙트럼 차감에서처럼 하나의 형태로 고정시키지 않고, 불확실한 잡음에 대한 정보를 특징추출에 확률적으로 반영한다. 인식단계에서는 각 모델의 상태에서의 출력분포에 최적으로 대응되는 스펙트럼을 이용하므로, 조용한 환경에서 학습된 은닉마르코프 모델에 내재된 음성의 특성을 이용할 수 있는 장점이 있다. 또한 인식확률의 계산에는 확률적 스펙트럼 차감법을 통하여 얻은 정적인 파라미터와 동시에 동적인 특징파라미터도 사용하므로 보다 효과적인 잡음처리가 가능하였다.

화자독립 음성인식실험을 통하여, 제안한 잡음처리방법의 유효성을 실험적으로 확인할 수 있었고, 향후 연구로는 스펙트럼 차감법이외에도 확률적 잡음원형을 이용한 위너필터링, 효과적인 잡음원형의 구성방법 등에 대한 연구가 필요하다.

참 고 문 헌

1. D. Mansour and B. J. Juang, "The short-time modified coherence representation and its application for noisy speech recognition," *IEEE Trans. on ASSP*, Vol. 37, No. 6, pp. 795-804, 1989.
2. H. Hermansky, N. Morgan, and H. G. Hirsh, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *proc. of ICASSP*, pp. 83-86, 1993.
3. T. H. Applebaum and B. A. Hanson, "Regression feature for recognition of speech in quiet and in noise," *proc. of ICASSP*, pp. 985-988, 1991.
4. D. Mansour and B. J. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on ASSP*, Vol. 37, No. 11, pp. 1659-1671, 1989.
5. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, Vol. 27, No. 2, pp. 113-120, 1979.
6. P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtracter (nss), hidden markov models and the projection, for robust speech recognition in cars," *Speech communication*, Vol 11, pp. 215-228, 1992.

7. O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer, Speech and Language*, Vol 1, pp. 109-130, 1986.
8. Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. on ASSP*, Vol 40, No. 4, pp. 725-735, 1992.
9. M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, Vol. 12, pp. 231-239, 1993.
10. J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizer," *J. Acoust. Soc. Amer.*, Vol 93, No. 1, pp. 510-524, Jan. 1993.
11. J. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, Vol 87, No. 4, pp. 1738-1752, April. 1990.
12. S. M. Chi and Y. H. Oh, "Lombard effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition," *proc. of ICSLP*, pp. 2013-2016, 1996.
13. 지상문, 오영환, "롬바드 효과의 보정을 위한 스펙트럼 크기의 정규화와 켑스트럼 변환," 한국음향학회지, 제 15권 4호, 83-92, 1996.

▲지 상 문(Sang-Mun Chi)
한국음향학회지 제15권 4호 참조

▲오 영 환(Yung-Hwan Oh)
1972년 2월 : 서울대학교 공과대학(학사)
1974년 2월 : 서울대학교 교육대학원(석사)
1980년 3월 : Tokyo Institute of Technology 정보공학전공 (박사)
1981년 4월~1985년 6월 : 충북대학교 컴퓨터 공학과 조교수
1983년 12월~1984년 11월 : University of California(Davis) 연구교수
1995년 9월~1996년 8월 : Carnegie-Mellon university 연구교수
1985년 7월~현재 : 한국과학기술원 전산학과 교수
▲관심분야 : 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가 시스템