論文98-35S-12-12

음성인식을 위한 잡음하의 음성왜곡제거

(The suppression of noise-induced speech distortions for speech recognition)

池 相 文 *, 吳 永 煥 **

(Sang-Mun Chi and Yung-Hwan Oh)

요 약

본 논문에서는 잡음에 의해 기인된 음성의 왜곡을 제거하여 음성인식기의 성능을 향상시키는 방법을 기술한다. 잡음 환경에서는 음성의 발성 방식이 변이하고(롬바드 효과), 잡음이 음성신호에 첨가되므로 음성인식기의 성능을 저하시킨다. 롬바드 효과는 주변 잡음의 크기나 종류, 화자의 특성과 음소 등에 종속적인 비선형적인 변환이므로 측정방법이 알려져 있지 않았다. 본 연구에서는 롬바드 효과의 크기를 측정하는 방법을 제시하고, 롬바드 효과의 크기에 따른 롬바드 효과의 보정 방법을 제안한다. 잡음에 의한 음성의 왜곡은 다음의 과정을 통해서 제거한다. 우선, 스펙트럼 차감법을 사용하여 음성에 포함된 잡음을 제거하고, 음성의 동적인 특성을 강조하기 위해 대역 통과 필터링을 한다. 두 번째로 에너지 정규화 과정을 통해서 롬바드 효과에 의한 음성의 발성 강도의 변이를 제거한다. 마지막으로 제안한 롬바드 효과의 크기 척도는 롬바드 음성의 켑스트럼에 존재하는 왜곡을 제거하는 변환에 이용한다. 제안한 방법을 음성인식에 적용한 결과, SNR(signal-to-noise ratio) 0, 10, 20 dB에서 46.3%, 75.5%, 87.4%의 인식률을 82.6%, 95.7%, 97.6%로 향상시켰다.

Abstract

In noisy environments, human speech productions are influenced by noises(Lombard effect), and speech signals are contaminated. These distortions dramatically reduce the performance of speech recognition systems. This paper proposes a method of the Lombard effect compensation and noise suppression in order to improve speech recognition performance in noisy environments. To estimate the intensity of the Lombard effect which is a nonlinear distortion depending on the ambient noise levels, speakers, and phonetic units, we formulate the measure of the Lombard effect level based on the acoustic speech signal, and the measure is used to compensate the Lombard effect. The distortions of speech under noisy environments are cancelled out as follows. First, spectral subtraction and band-pass filtering are used to cancel out noise. Second, energy normalization is proposed to cancel out the variation of vocal intensity by the Lombard effect. Finally, the Lombard effect level controls the transform which converts Lombard speech cepstrum to clean speech cepstrum. The proposed method was validated on 50 korean word recognition. Average recognition rates were 82.6%, 95.7%, 97.6% with the proposed method, while 46.3%, 75.5%, 87.4% without any compensation at SNR 0, 10, 20 dB, respectively.

* 正會員, 三星電子 無線開發팀
(Samsung Electronics, Wireless Terminals Division)
** 正會員, 韓國科學技術院 電算學科
(Dept. of Computer Science, Korea Advanced Institute of Science and Technology)
接受日字: 1998年5月4日, 수정완료일: 1998年10月13日

Ⅰ. 서 론

음성인식기는 실제 현장, 특히 잡음 환경하에서 성능이 크게 저하된다. 예를 들면, 조용한 환경에서 100%의 인식률을 갖는 음성인식기가 시속 90 Km로 주행하는 차안에서는 30%의 인식률을 가지며[1], 롬바드 효과에 의해서 95.7%에서 65.7%로 감소한다[2]. 음성인식기의 성능에 영향을 미치는 요인들은 가산잡음, 롬바드 효과, 심리적 스트레쓰와 채널잡음에 의한 음성의 변이를 포함하지만, 본 논문은 주변잡음에 의해 발생하는 왜곡인 발성방식의 변이(롬바드 효과)와 잡음에 의한 음성신호의 왜곡의 제거에 중점을 둔다.

잡음환경에 강인한 음성인식을 위해서 다양한 방법이 연구되어 왔다. 잡음의 영향에 둔감한 특징추출법이나 거리척도는 잡음의 크기나 종류에 민감하지 않은 음성신호의 표현을 구하는 것으로서, SMC(short-time modified coherence), RASTA(RelAtive SpecTAl) 처리, 사영척도와 청각기반의 특징추출방법이 이 범주에 속한다[3, 4, 5, 6]. 스펙트럼 차감법, 중회귀분석이나 신경회로망을 이용하여 잡음음성을 잡음이 제거된 음성으로 변환하는 음질개선방법과[7, 8], 인식모델인 HMM(hidden Markov model)에 포함된 잡음이 없는 음성에 관한 지식을 이용하여 잡음을 제거하는 방법인 모델기반의 위너필터링과 HMM의 파라미터를 잡음환경에 적응시키는 방법[9, 10]도 잡음환경에 강인한 음성인식을 위해서 사용되고 있다.

롬바드 효과는 화자, 성별, 잡음의 강도와 종류에 종속적인 비선형적인 왜곡으로서, 이론적인 분석이 존재하지 않고 여러 가지 실험적인 방법이 사용되고 있다. 롬바드 효과를 켑스트럼 파라미터 영역에서 가산항이나 승산항으로 모델링하는 방법이 있고[11, 12], 롬바드 음성을 학습자료에 포함시키는 방법과 파라미터의 동적인 특성을 강조하는 필터링도 롬바드 효과에 강인한 것으로 알려져 있다.

본 연구에서는 잡음의 영향과 롬바드 효과에 동시에 강인한 특징파라미터를 추출하려 한다. 롬바드 효과의 크기에 민감하게 변이하는 특징파라미터를 찾아내어 롬바드 효과에 의한 음성의 발성방식의 변이를 측정하기 위해, 통계적 분석을 사용하여 객관적인 척도를 개발한다. 이 척도는 켑스트럼 영역에서의 변환과 대역통과 필터링과 함께 사용되어 잡음에 강인한 특징파라미터를 추출하는데 이용한다.

본 논문의 구성은 2장에서 실험에 사용된 음성자료와 롬바드 효과의 크기를 측정하기 위한 방법을 설명하고, 3장에서 구체적인 잡음 영향의 제거 방법을 설명한다. 4장에서는 음성인식실험의 결과를 보이고, 5장에서 결론을 맺는다.

Ⅱ. 롬바드 효과의 분석과 크기 척도

롬바드 효과는 화자가 잡음환경에서 음성을 보다 명확하게 전달하기 위해서 발성방식을 변화시키는 것을 말하며, 발성음의 에너지 증가, 모음의 지속시간 증가, 저대역에서 중간대역이나 고대역으로 에너지분포 이동, 모음의 첫 번째 포만트 위치 증가, 스펙트럼 기울기 등이 변하는 것으로 보고되고 있다[13, 14, 15]. 그러나 롬바드 효과는 화자의 특성, 문맥, 주변환경, 잡음의 종류와 강도, 성별 등에 종속적인 현상이므로 일관성을 찾기가 어렵다. 예를 들면, 음향학적인 변이인 피치 등은 남성과 여성간에 매우 큰 차이를 보인다[15].

본 연구에서는 위에서 기술한 바와 같이 여러 가지 요인에 동시에 영향을 받는 롬바드 효과의 특성에도 불구하고, 비교적 일관성 있게 나타나는 특성을 조사하여, 이를 롬바드 효과의 크기를 측정할 수 있는 척도로 사용하고자 한다.

표 1. 실험에 사용된 잡음의 종류와 크기의 평균과 분산 (데시벨)

Table 1. Noise type, mean and variance of noise sound pressure level (dB).

Number	Noise type	Mean	Variance
1	Automobile cabin (highway)	89.4	15.2
2	Automobile cabin (down town)	76.8	52.4
3	Exhibition hall 1	77.3	5.7
4	Exhibition hall 2	76.0	4.9
5	Telephone booth 1	68.5	22.2
6	Telephone booth 2	71.4	7.0
7	Telephone booth 3	70.6	10.4
8	A crowded street 1	71.0	7.9
9	A crowded street 2	73.4	8.1

1. 잡음과 음성자료

롬바드 음성의 음향학적 특징에 대한 실험적인 조사를 위해서, 자동차, 전시장, 시내 공중전화 부스, 거리

에서 발생한 9종류의 잡음을 사용하여 롬바드 음성을 수집하였다. 표 1은 이러한 잡음의 특성을 보여주는 것으로, 잡음의 크기는 70dB에서 90dB사이이고, 다양한 분산을 가졌다. 잡음을 헤드폰을 통해서 발성자에게 들려줌으로써 잡음환경을 모의하고, 모의된 환경에서 발성된 음성을 수집하였다. 50단어를 20명(남 10명, 여 10명)이 조용한 환경에서 두 번 발성하였고, 9종류의 잡음환경에서 각각 한번 발성하였다.

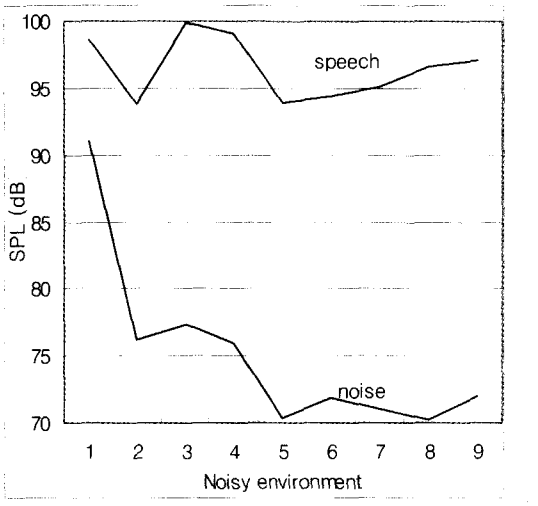

그림 1. 9개의 잡음환경에서 20명이 발성한 음성과 잡음의 평균에너지

Fig. 1. Average SPL for noises and 20 speakers' utterances in 9 noisy environments.

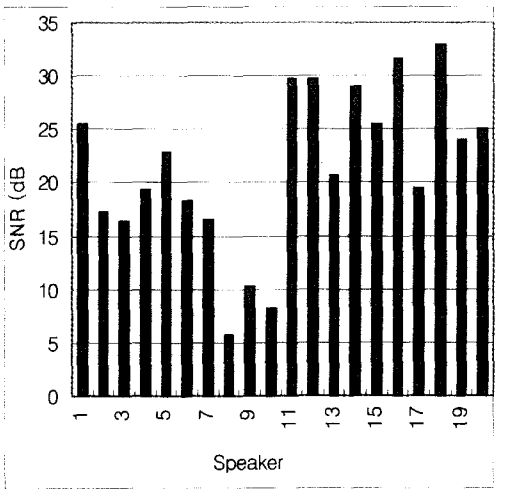

그림 2. 9개의 잡음환경에서 20명이 발성한 음성의 평균 SNR

Fig. 2. Average SNR for 20 speakers' utterances in 9 noisy environments.

그림 1은 수집된 음성의 각 잡음환경에서 에너지의 평균을 나타낸다. 그림에서 보듯이 음성의 에너지는 잡음의 에너지와 비례적으로 증가한다. 하지만, 그림 2에서 보듯이 각 화자가 발성한 음성과 잡음과의 에너지 비(SNR)는 화자에 따라 크게 변하는 것을 알 수 있다.

2. 롬바드 음성의 통계적 분석

롬바드 효과를 정량화하기 위해서, 롬바드 효과에 대해서 표 2의 특징들을 조사하였다[15, 16, 17]. E500과 CG의 변이는 화자나 성별에 대해서 비교적 영향을 적게 받고[15], LST의 변이도 큰 동질성을 나타내는 것으로 알려져 있다[16]. 표 2에 나타난 여러 파라미터의 변이는 무성음 부분보다 유성음 부분에서 더욱 확연히 나타나므로 본 연구에서는 유성음 부분에 대해서만 표 2의 파라미터의 값을 계산하여 조사하였다. 유무성음의 구분은 음성 프레임의 스펙트럼 기울기를 조사하여 결정하였다.

조용한 환경에서 발성한 음성과 롬바드 음성의 차이를 명확히 나타내는 특징파라미터를 찾아내기 위해서 통계적 테스트과, 잡음에 대한 강인성을 평가하였다.

(1) 통계적 테스트 : 롬바드 효과에 의해서만 음성이 왜곡된 정도를 측정하기 위해서, 잡음의 첨가에 의한 영향을 스펙트럼 차감법을 사용하여 제거하고 발성음의 크기도 정규화한 음성을 테스트에 사용한다. 롬바드 효과에 의해 변이가 큰 유성음의 스펙트럼을 구해 벡터양자화를 사용하여 분류한다. 벡터양자화를 위한 코드북은 조용한 환경에서 발성한 음성을 사용하여 미리 학습하였다. 벡터양자화를 사용하는 이유는 표 2의 특징은 롬바드 효과의 강도에도 비례하지만 음소자체의 특성에도 영향을 받으므로, 벡터양자화를 이용하여 입력음성의 각 프레임을 음향학적 특성에 따라 분류하여 분류된 각 코드워드별로 특징들의 평균을 구하여 롬바드 효과에 의한 왜곡을 측정한다. $y_1, y_2, \cdots, y_n$를 이러한 값이라 할 때, 코드워드 k에 속하는 롬바드 음성의 특징에서 구한 y를 사용하여 구한 평균을 $\overline{y_k}$, 분산을 $s^2_{2,k}$, 코드워드 k에 속하는 y들의 개수를 n_k라 하고, $\overline{x_k}$, $s^2_{1,k}$, m_k를 조용한 환경에서 발성한 음성의 평균, 분산, 개수라 하면 다음의 식을 테스트에 이용하였다.

$$z_k=(\overline{y_k}-\overline{x_k})/\sqrt{\frac{s_{2,k}^2}{n_k}+\frac{s_{1,k}^2}{m_k}} \quad (1)$$

즉, $\overline{x_k}$와 $\overline{y_k}$를 계산하기 위한 특징들은 서로 독립이고, 특징값들의 개수가 통계적으로 볼 때 충분히 큰 값이므로 large-sample test statistic인 z를 사용하여 조용한 환경의 음성과 롬바드 음성의 차이에 대한 통계적인 유의성 (statistical significance)을 평가하였다. 본 연구에서는 잡음환경에서 발성한 음성은 롬바드 효과에 의해서 영향을 받았다고 가정하고, 표2의 여러 특징중 $|z|\geq z_{\alpha/2}$ 이면, 롬바드 효과에 의한 왜곡을 측정하는데 유효하다고 판단하였다.

표 2. 롬바드 효과에 따른 특징 파라미터들의 변이

Table 2. The variations of features by the Lombard effect.

특징 파라미터	롬바드 효과에 의한 변이
E500	0에서 500 Hz사이의 에너지로 모음의 경우에 17%에서 37%정도 감소한다
CG	무게중심으로 다음과 같이 계산된다. $CG=\frac{1}{M}\sum_{i=1}^{D} i\cdot M_i,\ M=\sum_{i=1}^{D} M_i$ M_i는 i번째 필터 뱅크의 출력값이고 D는 필터 뱅크의 개수로 19를 사용하였다. 무게중심은 모든 화자와 음소에 대해(특히, 모음) 증가한다.
LST	저대역 스펙트럼 기울기는 처음 14개의 필터뱅크(0-3kHz) 에너지값을 선형회귀분석으로 기울기를 구하여 사용하였다. LST는 증가한다.
에너지	발성된 음성의 에너지는 크게 증가(특히, 모음과 유성음구간)한다.
켑스트럼 C_i	롬바드 효과에 의한 성도구조의 변이는 스펙트럼 대역간의 에너지의 이동을 발생시킬 수 있으므로, 음성신호의 스펙트럼 변이를 나타내는 켑스트럼계수를 변이시킨다. C_i 는 켑스트럼의 i번째 계수를 나타낸다.
XC	켑스트럼 계수간의 cross-correlation은 스펙트럼구조의 대역구조와 국소구조사이의 상대적인 변화를 나타낸다. $XC_{i,j}^{(0)}(k)=\sum_{m=k}^{k+L}\frac{C_i(m)C_j(m+l)}{L}$, 여기서 L과 l은 4와 3이 쓰였다.
AC	i번째 켑스트럼 계수간의 자기상관계수는 스펙트럼 에너지의 상관정도를 나타낸다. $AC_i^{(0)}(k)=\sum_{m=k}^{k+L}\frac{C_i(m)C_i(m+l)}{L}$
Delta	차분파라미터는 파라미터의 속도를 나타내고, 켑스트럼 파라미터, cross-correlation, 자기상관계수 등의 차분파라미터, DC_i, DXC_i, DAC_i를 실험에 사용하였다.

(2) 잡음에 대한 강인성 평가 : 롬바드 음성은 잡음환경에서 발생하고 롬바드 효과의 척도도 다양한 잡음환경에 적용하여야 하므로, 롬바드 효과의 크기를 나타나는 값은 잡음에 의해 변화가 적어야 한다. 다음식은 잡음에 의해 z값이 변이하는 정도를 나타낸다. 분자는 롬바드 음성과 잡음이 섞인 롬바드 음성의 롬바드 효과의 크기의 차이에 대한 제곱이고, 분모는 정규화를 위해 사용했다.

$$n-robust=\frac{1}{3N}\sum_{dB\in\{0,10,20dB\}}\sum_{i=0}^{N}\frac{(z_{\infty,i}-z_{dB,i})^2}{z_{\infty,i}\cdot z_{\infty,i}} \quad (2)$$

단, $z_{\infty,i}$ 는 잡음환경 i에서 발성된 롬바드 음성과 조용한 환경에서 발성된 음성간의 z값이고, $z_{dB,i}$ 는 잡음환경 i에서 발성된 롬바드 음성에 잡음을 dB SNR이 되도록 혼합한 잡음음성과 조용한 환경에서 발성된 음성간의 z값이다. N 은 잡음환경의 개수이다.

표 3. 롬바드 효과에 의한 왜곡의 측정에 유효한 특징파라미터의 z값의 평균과 n-robust값

Table 3. Average z-value and n-robust.

Feature	codeword	z	*n-robust*
E500	0	-15.57	0.029
LST	0	9.84	0.013
CG	0	11.52	0.002
E500	1	-14.08	0.043
LST	1	8.60	0.027
CG	1	9.84	0.004
E500	2	-16.32	0.022
LST	2	14.17	0.021
CG	2	10.68	0.028
E500	3	-11.13	0.019
LST	3	11.05	0.038

표 2의 여러 특징중에서 9개의 잡음환경과 SNR 0, 10, 20dB에서의 z값의 평균이 $|z|\geq z_{\alpha/2}$ 인 것은 너무 많아서 표시하기가 어려우므로, $|z|\geq 8$이고 $n-robust<0.05$ 인 것을 표 3에 나타내었다. 표 2의 값들은 유성음구간을 4개의 코드북을 사용하여 분류하였을 경우이다. 표 3의 특징은 롬바드 효과에 의한 왜곡을 잘 나타내므로 이를 이용하여 다음장에서 롬바드 효과의 크기 척도를 정의한다.

3. 롬바드 효과의 크기 척도

표 3의 통계적인 평가는 여러 잡음환경에서 발성된

모든 음성의 평균값을 사용하였으므로, z값은 각각의 특징파라미터가 롬바드 효과를 얼마나 잘 반영하는지에 대한 것만을 나타낸 것이지, 입력음성 자체의 롬바드 효과를 나타낸 것은 아니다.

본 논문에서는 입력음성으로부터 롬바드 효과의 크기를 직접적으로 측정하기 위해 유성음구간의 스펙트럼을 벡터양자화를 이용하여 분류하고, 같은 코드워드로 분류된 특징값의 평균을 이용하여 롬바드 효과의 크기를 정의한다. 즉, 입력음성에서 추출한 특징열의 각 코드워드별 평균을 $y_1, y_2, \cdots y_N$ 이라 하고, y_k를 코드워드 k로 분류된 특징값의 평균, m_k를 코드워드 k로 분류된 특징값의 개수, μ_k와 σ_k^2는 조용한 환경에서 발성한 음성을 사용하여 코드북을 학습할 때 코드워드 k로 분류된 파라미터의 평균과 분산을 나타낸다. 롬바드 효과의 크기(Lombard effect level: LEL)는 코드워드 별로 표준화된 변수(standardized variable)의 평균으로서 정의하였다. 즉, $LEL = \sum_{k=1}^{N} p_k \frac{y_k - \mu_k}{\sigma_{p,k}}$, N은 코드워드의 개수, p_k는 코드워드 k로 특징이 분류될 확률로 $p_k = m_k/T$ 이다. 또한, 안정적인 LEL의 계산을 위해서 양자화 과정에서 가장 많이 발생한 코드워드의 인덱스 k에 대해 $p_k = 1$로 하고 나머지는 0으로 하였다. 제안한 롬바드 효과의 크기는 3장에서 설명할 롬바드 효과의 보정에 사용된다.

Ⅲ. 잡음의 제거와 롬바드 효과의 보정

1. 잡음에 의한 음성의 왜곡과정

잡음에 의한 음성의 왜곡을 모델링하기 위해서 조용한 환경의 음성은 다음과 같은 세 가지의 왜곡을 거친다고 가정한다. 음성은 롬바드 효과에 의해서 에너지가 정규화된 음성의 스펙트럼의 구조가 변하는 단계 1과 전체적인 음성의 에너지가 변하는 단계 2의 왜곡을 거친 후에, 잡음이 첨가되는 단계 3의 왜곡을 거친다.

단계 1: 포만트의 위치와 대역폭의 변이, 피치, 스펙트럼 기울기, 각 대역의 에너지의 크기 변화 등은 비선형적인 주파수 변환(frequency warping) $F_{LEL}(\cdot)$와 주파수 대역별 스펙트럼 크기변이 $A_{LEL}(\cdot)$로 모델링하였다. 이 왜곡함수들은 롬바드 효과의 크기에 따라 각기 다른 모양을 가지며, 조용한 환경에서 발성된 음성의 스펙트럼 $S(\omega)$를 왜곡시킨다,

$$Y_1(\omega) = A_{LEL}(\omega) S(F_{LEL}(\omega)). \quad (3)$$

단계 2: 화자는 그림 1에서 보는 바와 같이, 주변잡음의 크기에 따라 효과적인 의사전달을 위해서 발성음의 크기를 조절하지만, 그림 2에 나타난 바와 같이 그 변이는 화자나 잡음의 종류나 크기에 따라 다양하다. 이러한 발성음의 에너지의 변이도 왜곡요인의 하나이며, 이를 G_{LEL}로 모델링한다,

$$Y_2(\omega) = G_{LEL} \cdot Y_1(\omega) = G_{LEL} \cdot A_{LEL}(\omega) S(F_{LEL}(\omega)). \quad (4)$$

단계 3: 가산잡음의 첨가는 주파수 영역에서 다음과 같은 잡음의 스펙트럼 $N(\omega)$가 더해지는 것으로 나타난다,

$$Y_3(\omega) = Y_2(\omega) + N(\omega) = G \cdot A_{LEL}(\omega) S(F_{LEL}(\omega)) + N(\omega). \quad (5)$$

2. 잡음에 의한 왜곡의 제거

잡음이 첨가된 롬바드 음성의 왜곡은 3.1절에서 설명된 왜곡과정의 역과정을 통해서 제거할 수 있다. 그림 3은 왜곡의 제거과정을 보여준다.

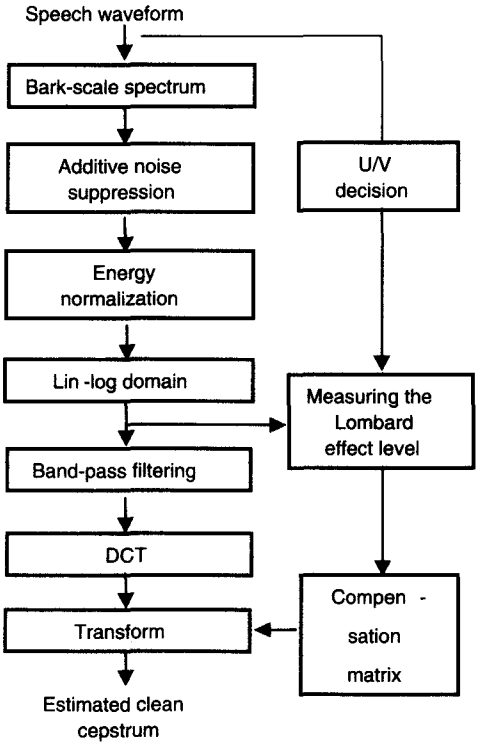

그림 3. 잡음과 롬바드 효과의 처리과정
Fig. 3. The procedure of noise suppression and the Lombard effect compensation.

첫째, 잡음과 롬바드 효과에 의해 왜곡된 음성으로부터 바크단위의 스펙트럼을 얻고[18], 스펙트럼 차감법[7]을 사용하여 잡음을 제거한다. 즉, 롬바드 음성의 스펙트럼 $Y_2(\omega)$ 는 잡음이 첨가된 롬바드 음성의 스펙트럼 $Y_3(\omega)$ 로부터 음성이 존재하지 않은 구간에서 추정된 잡음의 스펙트럼 $N(\omega)$를 차감함으로서 얻어진다.

둘째, $Y_2(\omega)$ 에 포함된 에너지의 변이 요인인 G_{LEL}는 다음과 같이 정의한다,

$$G_{LEL}=\frac{\text{롬바드음성의 평균에너지}}{\text{조용한 환경에서 발성한 음성의 평균에너지}} \quad (6)$$

단, 롬바드 음성의 평균에너지는 $Y_2(\omega)$를 이용하여 구하고, 조용한 환경에서 발성한 음성의 평균에너지는 학습자료로 사용한 음성의 평균에너지이다. $Y_1(\omega)$는 $Y_2(\omega)$를 G_{LEL}로 나눔으로서 추정된다. 따라서 G_{LEL}을 입력음성에서 제거함으로서 모든 음성은 동일한 평균에너지를 갖지만, 입력음성안에서의 에너지의 시간적인 변화는 보존된다. 이러한 방법은 화자나 잡음의 종류와 세기에 영향을 받는 음성의 에너지를 정규화하고, 다음의 처리인 lin-log RASTA필터링의 입력을 안정적으로 만드는 역할을 한다.

셋째, 스펙트럼 $Y_1(\omega)$ 은 lin-log 스펙트럼 영역으로 변환한다,

$$LY(\omega)=\ln(1+J\cdot Y_1(\omega)) \quad . \quad (7)$$

이 변환에서 최적의 J값은 잡음의 크기와 입력신호의 크기에 달려있으므로, 음성인식에 있어서 하나의 변이요인으로 작용한다. 이러한 점을 보완하기 위해 스펙트럼 변환이나 다중의 J값을 사용하는 방법[4]이 있으나, 본 연구에서는 에너지를 정규화한 스펙트럼을 입력으로 사용하므로, 고정된 J값을 사용하였다. lin-log스펙트럼 영역에서의 스펙트럼은 LST, CG를 구할 때 이용되었다. lin-log스펙트럼 영역에서의 필터링은 가산잡음과 채널잡음을 제거하고, 동적인 특성을 강조하는 필터링은 롬바드 효과를 제거하는 데에도 유효하다고 알려져 있으므로[4, 19], 다음의 대역통과 필터를 통과시켰다,

$$0.1\cdot\frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.94z^{-1}} \quad . \quad (8)$$

필터링된 스펙트럼 값은 다음의 근사된 역변환을 사용하여 스펙트럼 영역으로 되돌려진다,

$$Y_1(\omega)=e^{LY(\omega)}/J \quad . \quad (9)$$

마지막으로, 이산 코사인 변환으로 스펙트럼으로부터 켑스트럼을 유도하고, 선형변환과 롬바드 효과의 크기를 이용하여 $Y_1(\omega)$에 포함된 왜곡요인인 $F_{LEL}(\cdot)$와 $A_{LEL}(\cdot)$를 제거한다. 조용한 환경에서 발성한 음성의 스펙트럼을 $S(\omega)$, 켑스트럼을 C_n^{normal}라하고, 롬바드 음성의 켑스트럼 $C_k^{Lombard}$, 이때의 스펙트럼 $Y_1(\omega)=A_{LEL}(\omega)S(F_{LEL}(\omega))$라 하자. 이산코사인 변환의 정의에 의해서 식 (10)-(11)을 얻고,

$$C_n^{normal}=\alpha(n)\sum_{\omega=0}^{N-1}\log S(\omega)\cdot\cos\left[\frac{(2\omega+1)n\pi}{2N}\right] \quad , \quad (10)$$

$$\log(A_{LEL}(\omega)S(F_{LEL}(\omega)))=\sum_{k=0}^{N-1}\alpha(k)C_k^{Lombard}\cos\left[\frac{(2\omega+1)k\pi}{2N}\right] \quad , \quad (11)$$

여기서 N은 19이고 $\alpha(0)=\sqrt{1/N}, \alpha(n)=\sqrt{2/N}$, n=0, 1, ..., N-1이다. 식 (11)을 식(12)와 같이 바꾸고, ω 대신에 $F_{LEL}^{-1}(\omega)$를 사용하여 식 (13)과 같이 나타낸다.

$$\log(S(F_{LEL}(\omega)))=\sum_{k=0}^{N-1}\alpha(k)C_k^{Lombard}\cos\left[\frac{(2\omega+1)k\pi}{2N}\right]-\log(A_{LEL}(\omega)), \quad (12)$$

$$\log S(\omega)=\sum_{\omega=0}^{N-1}\alpha(k)C_k^{Lombard}\cos\left[\frac{(2F_{LEL}^{-1}(\omega)+1)k\pi}{2N}\right]-\log(A_{LEL}(F_{LEL}^{-1}(\omega))). \quad (13)$$

식 (13)을 식 (10)에 대입하면 식 (14)를 얻는다.

$$C_n^{normal}=\sum_{k=0}^{N-1}M_{LEL}(n,k)\cdot C_k^{ag} \quad , \quad (14)$$

여기서, $C_N^{ag}=1, C_k^{ag}=C_k^{Lombard}$, k = 0, 1, ..., N-1,

$$M_{LEL}(n,k)=\sum_{\omega=0}^{N-1}\alpha(n)\alpha(k)\cos\left[\frac{(2F_{LEL}^{-1}(\omega)+1)k\pi}{2N}\right]\cos\left[\frac{(2\omega+1)n\pi}{2N}\right] \quad , \quad (15)$$

$$M_{LEL}(n,N)=-\sum_{\omega=0}^{N-1}\alpha(n)\log A(F_{LEL}^{-1}(\omega))\cos\left[\frac{(2\omega+1)n\pi}{2N}\right] \quad . \quad (16)$$

조용한 환경에서 발성한 음성의 켑스트럼과 롬바드 음성의 켑스트럼이 식 (14)와 같은 관계가 있으므로, 변환행렬이 존재하면 롬바드 음성의 켑스트럼에서 왜곡을 제거하는 것이 가능하다. 왜곡함수 $F_{LEL}(\cdot)$와 $A_{LEL}(\cdot)$ 은 음소, 화자, 롬바드 효과의 크기에 따라

각기 다른 특성을 가지므로, 변환행렬은 이들 요소에 따라 구해져야 한다. 그러나 이들 여러 요소에 따라 변환행렬을 구하기는 어려우므로, 롬바드 효과의 크기를 몇 개로 나누고, 각각의 롬바드 효과의 크기에 해당하는 롬바드 음성의 켑스트럼 파라미터들을 벡터양자화를 통하여 군집화한다. 화자나 음소 등에 따라 다른 코드워드로 군집화된다고 가정하여, 이들 각각의 군집에 대해서 변환행렬을 구한다. 이 과정은 다음과 같다.

(1) 롬바드 음성의 켑스트럼 벡터와 조용한 환경에서 발성된 음성의 켑스트럼 벡터를 동적시간 정합법으로 매칭시켜서, 켑스트럼 쌍을 만든다.
(2) 롬바드 음성의 켑스트럼을 롬바드 효과의 크기와 코드워드에 따라 분류한다.
(3) 분류된 롬바드 음성의 켑스트럼 벡터와 이에 대응되는 조용한 환경에서 발성한 음성의 켑스트럼을 사용하여, 중회귀분석을 사용하여 변환행렬을 추정한다.
(4) 입력된 음성의 롬바드 효과의 크기와 대응되는 코드워드에 따라 이미 추정되어 있는 변환행렬로 식 (14)에 따라 변환함으로써 조용한 환경에서 발성한 음성의 켑스트럼을 추정한다. 변환이 학습자료에 너무 종속적이 되지 않기 위해서 변환된 켑스트럼과 롬바드 음성의 켑스트럼을 평균하여, 최종적인 켑스트럼을 얻는다.

Ⅳ. 실험 및 분석

1. 실험조건

제안한 방법을 음성인식 실험을 통하여 검증하였다. 대부분의 음성인식기가 사용되는 환경의 특성은 학습환경의 특성과 다르므로, 이러한 점을 고려하여 음성인식기의 학습자료는 남자 5명과 여자 5명이 조용한 환경에서 2회 발성한 50단어를 사용하였고, 평가를 위해서는 또다른 남자 5명과 여자 5명이 표 1의 각각의 모의된 잡음환경에서 발성한 음성을 사용하였다. 음성은 16 kHz, 16 비트로 샘플링되었고, $1-0.95z^{-1}$ 의 필터를 사용하여 전처리를 하였다. 헤밍창을 씌운 32ms구간을 분석하여 16ms마다 14차의 켑스트럼 계수를 구하였다. 켑스트럼 계수는 19개의 바아크 스케일의 필터뱅크의 출력을 이산 코사인 변환을 사용하여 계산하였다. 14차의 켑스트럼, 켑스트럼의 차분파라미터, 정규화된 에너지와 에너지의 일차, 이차 차분값을 음성인식에 이용하였고, 세 종류의 파라미터를 각각 256, 256, 16개의 크기를 갖는 코드북을 사용하여 양자화하였다. 이산형 HMM(hidden Markov model)을 패턴분류기로 사용하였고, HMM의 상태수는 각 단어의 (음소기호의 숫자x3 + 2)를 사용하였다. 모델의 구조는 점프천이를 갖는 left-to-right 구조이다.

2. 음성인식 실험

표 4는 SNR 0, 10, 20dB에서 특징 파라미터들의 인식률을 비교한 것이다. CEP은 잡음처리가 되지 않은 기본적인 켑스트럼으로서 바아크 스케일의 필터뱅크의 출력으로부터 얻은 특징파라미터이고, SS는 필터뱅크의 출력에 스펙트럼 차감법을 수행한 후에 얻은 켑스트럼이다. LPC-CEP은 선형예측계수로부터 얻은 멜스케일의 켑스트럼이고, PROJ는 LPC-CEP의 거리측정에 사영척도를 사용한 것이다. RASTA는 lin-log 영역에서 대역통과 필터링을 수행한 스펙트럼으로 구한 켑스트럼으로서, 표 4에 나타난 인식률은 최적의 J 값은 실험적으로 선택했을 때이다. ENOR는 스펙트럼 차감법, 에너지 정규화방법과 lin-log영역에서의 대역통과 필터링을 사용하여 얻은 스펙트럼으로부터 구한 켑스트럼이다.

표 4에서 보듯이, SS는 잡음을 제거하므로 기본적인 특징파라미터 CEP보다 SNR 0, 10dB에서 향상된 인식률을 얻었지만, 잡음이 적은 SNR 20dB에서는 약간 인식률이 저하되었다. 또한, 필터뱅크에 기반한 특징파라미터 CEP과 SS가 선형예측에 기반한 방법 LPC-CEP과 PROJ보다 잡음환경에 강인함을 알 수 있다. 대역통과 필터링은 잡음과 롬바드 효과를 제거하므로 RASTA와 ENOR는 인식률을 향상시켰다. RASTA는 SNR 0, 10, 20dB에서 각기 다른 최적의 J 값인 $1.5\times10^{-7}, 1\times10^{-6}, 3\times10^{-6}$ 을 가졌고, 최적의 값과 다를 때에는 성능이 크게 저하되었다. ENOR은 식 (4)에서 발성음의 에너지 변이를 정규화할 뿐만 아니라, SNR에 따라 최적의 J값이 종속적이지 않으므로 $J=1.8\times10^{-7}$을 모든 실험에 사용하였다. ENOR은 기본적인 특징파라미터인 CEP에 비해 SNR 0, 10, 20dB에서 각각 35.3%, 19.7%, 10%의 인식률을 향상시켰다.

표 4. 여러 특징파라미터의 인식률(%)
Table 4. Recognition rates(%) using several features.

Feature / SNR (dB)	CEP	SS	LPC-CEP	PROJ	RASTA	ENOR
0	46.3	57.1	49.1	53.8	72.6	81.6
10	75.5	76.8	74.3	77.5	91.4	95.2
20	87.4	84.6	84.9	85.8	96.6	97.4

표 5는 롬바드 효과의 크기 *LEL*에 따른 인식률을 보여준다. 롬바드 효과의 크기는 표 3에서 커다란 z값을 갖는 E500, LST, CG를 사용하여 구하였다. 유성음구간은 4개의 코드북을 사용하여 분류하였고, 양자화 과정에서 가장 많이 발생한 코드워드에 해당하는 파라미터만을 이용하여 *LEL*을 계산하였고, ENOR를 사용하여 인식실험을 하였다. 표에서 보듯이 CG로부터 구한 롬바드 효과의 절대값이 커짐에 비례하여 오인식률이 증가함을 볼 수 있다. 이로부터 제안한 롬바드 효과의 크기가 음성인식률의 기준으로는 롬바드 효과에 의한 왜곡을 잘 표현한다고 할 수 있다. E500과 SNR 10dB에서의 LST는 롬바드 효과를 정확하게 표현하지 않았지만 비교적 오류가 적었다. 표 6의 실험에는 CG를 사용하였다.

표 5. *LEL*에 따른 음성인식률(괄호 안의 숫자는 각 *LEL*에 속하는 자료의 빈도)
Table 5. Recognition rates(%)(the numbers in parenthesis mean the relative frequency of data which are contained in each LEL).

	$LEL \le 1$	$1 < LEL \le 2$	$2 < LEL$
LST(0 dB)	85.4(0.38)	82.0(0.31)	76.4(0.31)
LST(10 dB)	95.8(0.44)	94.8(0.34)	94.9(0.22)
LST(20 dB)	97.5(0.50)	97.4(0.34)	97.0(0.16)
CG (0 dB)	83.5(0.36)	81.1(0.29)	80.0(0.31)
CG (10 dB)	96.4(0.42)	95.3(0.33)	93.2(0.25)
CG (20 dB)	98.2(0.49)	97.2(0.34)	95.3(0.17)
	$LEL \le -2$	$-2 < LEL \le -$	$-1 \le LEL$
E500(0 dB)	75.6(0.37)	85.1(0.36)	84.8(0.27)
E500(10 dB)	94.8(0.25)	94.7(0.43)	96.3(0.32)
E500(20 dB)	97.2(0.17)	97.0(0.42)	97.8(0.41)

표 6. 변환의 구성과 인식률
Table 6. Configuration of linear transform and recognition.

	학 습	평 가	인식률 (%)
실험 1	M: $LEL > 0.5$	M, if $LEL > 0.5$	82.5 (0dB) 95.6 (10dB) 97.6 (20dB)
실험 2	$M1$: $0 < LEL \le 2$ $M2$: $1 < LEL \le 3$	$M1$, if $0.5 < LEL \le 1.4$ $(M1+M2)/2$, if $1.4 < LEL \le 1.6$ $M2$, if $LEL > 1.6$	82.6 (0dB) 95.7 (10dB) 97.6 (20dB)

ENOR로부터 얻은 켑스트럼에서 왜곡함수 $F_{LEL}(\cdot)$와 $A_{LEL}(\cdot)$의 영향을 제거하기 위해서 선형변환을 사용하였다. 표 6은 이러한 변환의 구성과 인식률을 나타낸다. 예를 들어, 실험 2의 경우에는 $0.0 < LEL \le 2.0$ 에 속하는 롬바드 음성을 벡터양자화를 이용하여 512개의 군집으로 분류하고, 각각의 군집에 대해서 행렬들의 집합 M1을 얻었다. 마찬가지로 512개 행렬의 집합 M2는 $1.0 < LEL \le 3.0$ 에 속하는 음성을 이용하여 학습한다. 행렬은 조용한 환경에서 발성한 음성과 남녀 각각 5명이 9개의 잡음환경에 발성한 음성을 SNR 10, 20dB로 만들어서 사용하였다. 선형변환은 $LEL < 0.5$ 일 때는 롬바드 효과에 의한 왜곡이 적으므로, 변환을 하지 않았고, M1을 학습시킨 음성자료들의 *LEL*은 0에서 2사이이므로, $LEL > 1.5$ 인 경우는 변환이 부정확하고, 마찬가지고 M2의 경우에는 $LEL < 1.5$인 경우에는 부정확한 변환이 발생하므로, $1.4 < LEL \le 1.6$ 일 때는 각각 M1과 M2에 속한 행렬로 변환하여 이들의 평균값을 사용하였고, $0.5 < LEL \le 1.4$ 일 때는 M1에 속하는 행렬로, $LEL > 1.6$일 때는 M2에 속하는 행렬을 이용하여 변환하였다. 실험 2는 기본적인 특징파라미터인 CEP에 비해 SNR 0, 10, 20dB에서 에러율을 68%, 82%, 81% 감소 시켰다.

V. 결 론

본 논문은 다양한 잡음환경에서도 성능이 저하되지 않도록, 잡음에 의한 음성의 왜곡을 제거하는 방법을 설명하였다. 잡음환경에서 음성의 왜곡과정을 모델링하였고, 이들 왜곡을 제거하기 위한 방법을 개발하였다.

롬바드 효과의 크기를 추정하기 위해서, 음성신호로부터 직접 롬바드 효과의 크기를 측정할 수 있는 척도를 제안하였고, 이를 이용하여 롬바드 효과를 제거하

는데 이용하였다. 잡음에 의한 음성의 왜곡을 제거하는 과정은 다음과 같다. 첫째, 스펙트럼 차감법을 사용하여, 음성에 포함된 잡음을 제거하였고, 대역통과 필터링으로 동적인 특성을 강화하였다. 둘째, 에너지 정규화를 통해서 롬바드 효과에 의해서 발생하는 에너지의 변이를 제거하였고, 마지막으로 제안한 롬바드 효과의 크기와 선형변환을 이용하여 롬바드 효과에 의한 스펙트럼 구조의 변이를 제거하였다. 제안한 방법을 50단어의 인식에 적용한 결과, SNR 0, 10, 20dB에서 46.3%, 75.5%, 87.4%의 인식률을 82.6%, 95.7%, 97.6%로 향상시켰다.

참 고 문 헌

[1] 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.

[2] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACE) for speech recognition in noise and Lombard effect", *IEEE Trans. SAP*, Vol. 2, No. 4, pp. 598-614, Oct. 1994.

[3] D. Mansour and B. J. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 37, No. 6, pp. 795-804, 1989.

[4] H. Hermansky, N. Morgan, and H. G. Hirsh, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing", Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 83-86, 1993.

[5] D. Mansour and B. J. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 37, No. 11, pp. 1659-1671, 1989.

[6] 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.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.

[8] C. Mokbel and G. Chollet, "Speech recognition in adverse environment: speech enhancement and spectral transformations", Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 925-928, 1991.

[9] Y. Ephraim, "A Bayean estimation approach for speech enhancement using hidden Markov models", *IEEE Trans. Acoust. Speech Signal Processing*, Vol 40, No. 4, pp. 725-735, 1992.

[10] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise", *Speech Communication*, Vol 12, pp. 231-239, 1993.

[11] Y. Chen, "Cepstral domain stress compensation for robust speech recognition", Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 717-720, 1987.

[12] J. H. L. Hansen and D. A. Cairns, "ICARUS: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments", *Speech Communication*, Vol. 16, No. 4, pp. 598-614, Oct. 1994.

[13] B. Stanton, L. Jamieson and G. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions", Proc. Internat. Conf. Acoust. Speech Signal Process., Vol. 1, pp. 331-334, 1988.

[14] V. Summers D. Pisoni, R. Bernacki, R. Pedlow and M. Stokes, "Effect of noise on speech production: Acoustic and perceptual analysis", *J. Acoust. Soc. Amer.*, Vol. 84, pp. 917-928, 1988.

[15] J. C. Junqua, "The Lombard reflex an its role on human listeners and automatic

speech recognizers", *J. Acoust. Soc. Amer.*, Vol. 93, pp. 510-524, 1993.

[16] A. Castellanos, J. M. Benei and F. Casacuberta, "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect", *Speech Communication*, Vol. 20, pp. 23-35, 1996.

[17] B. D. Womac k and J. H. Hansen, "Classification of speech under stress using target dirven features", *Speech Communication*, Vol. 20, pp. 131-151, 1996.

[18] H. Hermansky, "Perceptual linear preditive (PLP) analysis of speech", *J. Acoust. Soc. Amer.*, Vol.. 87, No. 4, pp. 1738-1752, 1990.

[19] B. A. Hanson and T. H. Applehaum, "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech," Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 79-82, 1993.

저 자 소 개

池 相 文(正會員)

1991년 2월 : 서울대학교 수학교육과(학사). 1993년 2월 : 한국과학기술원 수학과 (석사). 1998년 8월 : 한국과학기술원 전산학과 (박사). 1998년 3월 ~ 현재 : 삼성전자 무선개발팀. 주관심분야 : 음성처리, 이동통신

吳 永 煥(正會員)

1972년 서울대학교 공과대학(학사). 1974년 서울대학교 교육대학원(석사). 1980년 Tokyo Institute of Technology 정보공학전공(박사). 1981년 ~ 1985년 충북대학교 컴퓨터공학과 조교수. 1983년 ~ 1984년 University of California, Davis 연구교수. 1995년 ~ 1996년 Carnegie-Mellon University 연구교수. 1985년 ~ 현재 한국과학기술원 전산학과 교수. 관심분야는 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가 시스템.