롬바드 효과의 보정을 위한 스펙트럼 크기의 정규화와 켑스트럼 변환

Normalization of Spectral Magnitude and Cepstral Transformation for Compensation of Lombard Effect

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

※본 연구는 전자통신연구소 95년도 위탁과제 결과물의 일부입니다.

요 약

본 연구에서는 음성인식기의 성능이 잡음환경하에서 급격히 저하되는 것을 완화하기 위해, 성능저하의 원인인 롬바드 효과의 보정과 잡음의 제거방법을 제안하였다. 롬바드 효과는 조용한 환경에서 발성된 음성에 비해, 스펙트럼 포락과 발성음의 세기를 변이 시키는 것으로 모델링하였고, 변이의 제거를 위해 스펙트럼 크기의 정규화와 켑스트럼 변환을 사용하였다. 주변 잡음의 첨가에 의한 음성신호의 왜곡은 스펙트럼 차감법을 사용하여 완화하였고, 음성의 동적인 특성을 강조하기 위해 대역통과 필터링을 하였다. 잡음환경에서 발성된 롬바드 음성의 분석 및 잡음처리 기술의 개발과 평가를 위해, 음성인식 기술의 적용이 예상되는 자동차, 전시장, 시내 공중전화 부스, 거리, 전산실 잡음을 이용하여 롬바드 음성을 수집하여 실험하였다. 제안한 방법을 여러 가지 잡음환경하에서 음성인식에 적용한 결과, 효과적인 잡음처리 방법임을 확인할 수 있었다.

ABSTRACT

This paper describes Lombard effect compensation and noise suppression so as to reduce speech recognition error in noisy environments. Lombard effect is represented by the variation of spectral envelope of energy normalized word and the variation of overall vocal intensity. The variation of spectral envelope can be compensated by linear transformation in cepstral domain. The variation of vocal intensity is canceled by spectral magnitude normalization. Spectral subtraction is used to suppress noise contamination, and band-pass filtering is used to emphasize dynamic features. To understand Lombard effect and verify the effectiveness of the proposed method, speech data are collected in simulated noisy environments. Recognition experiments were conducted with contamination by noise from automobile cabins, an exhibition hall, telephone booths in down town, crowded streets, and computer rooms. From the experiments, the effectiveness of the proposed method has been confirmed.

Ⅰ. 서 론

조용한 환경에서의 음성인식기는 이미 높은 성능을 보이고 있으나, 실제 환경에 존재하는 여러 요인에 의해 성능이 저하된다. 잡음에 의한 음성인식기의 성능저하는 음성인식기술의 광범위한 활용을 가로막는 요인의 하나로서, 이에 대한 많은 연구가 진행되고 있다. 잡음환경에서 발성된 음성은 그림 1과 같이 발성된 음성에 잡음이 첨가될 뿐만 아니라, 화자가 의사 전달을 명확하게 하기 위해 조용한 환경에서와 다르게 발성하는 롬바드 효과에 의해 음성이 왜곡된다[11, 14]. 이밖에도 음성이 전화선을 경유할 때의 채널잡음과, 사람의 감정상태에 따른 음성신호의 변이가 음성인식기의 성능저하 요인이 될 수 있으나, 본 연구에서는 연구대상으로 하지 않는다.

잡음환경에 강인한 음성인식을 위해 여러 가지 접근방법이 연구되고 있다. 잡음의 첨가에 강인한 특징추출과 거리척도로서, SMC(short-time modified coherence)[8], RASTA (RelAtive SpectrAl) 처리와 동적인 특징파라미터[10, 16], 켑스트럼 사상척도 등이 사용된다[9]. 음성신호에 포함된 잡음을 제거하는 방법으로, 스펙트럼 차감법[17, 18], 다중선형회귀분석이나 신경망[5, 19], 베이지안 추정으로 음

*한국과학기술원 전산학과 인공지능 연구실
접수일자 : 1996년 5월 29일

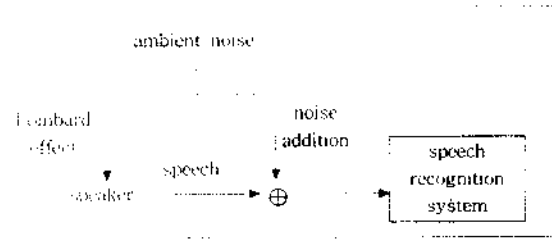

그림 1. 잡음환경에서 발생된 음성의 왜곡과정

질을 개선하는 방법이 사용되고 있다[12]. 이외에 청각기관의 특성을 이용하는 방법[15], 인식모델의 파라미터를 잡음환경에 적응되도록 변환하는 방법이 사용된다[1].

롬바드 효과는 화자나 잡음의 종류에 따른 비선형적인 왜곡이므로, 이에 대한 분석과 왜곡의 보정방법의 개발이 어렵지만, 여러 가지 실험적인 방법이 사용되고 있다. 켑스트럼 영역에서 지수함수의 가산적인 항이나, 곱해지는 상수 또는, 더해지는 상수로 롬바드 효과를 모델링하여 제거하는 방법[21, 13, 14], 롬바드 음성의 일부를 학습자료에 포함시키는 multi-style 학습[6], 동적인 파라미터를 적용하는 방법[4], 학습환경에서 작성된 코드북을 롬바드 음성의 특징벡터의 평균으로 변환시키는 방법이 사용되고 있다[7].

잡음환경에서 발성된 음성에는 상이한 특성을 가진 여러 가지 왜곡이 존재하므로, 각각의 특성에 맞는 잡음처리 방법이 요구된다. 본 연구에서는 잡음하에서 음성의 왜곡을 처리 가능한 구체적인 형태로 주파수 영역에서 모델링하여, 각각의 왜곡에 대응되는 처리방법을 사용하여 잡음에 강인한 음성인식 방법을 개발하고자 한다. 롬바드 효과에 의한 왜곡을 정규화된 에너지를 갖는 단어의 스펙트럼 포락(spectral envelope)의 변이와 발성음의 세기 변이로 모델링하였다. 스펙트럼 포락의 변이는 조용한 음성의 켑스트럼과 잡음음성의 켑스트럼이 선형관계로 나타나는 것을 이용하여 변이를 보정하고, 발성음의 세기 변이는 스펙트럼 크기의 정규화를 통해 제거하였다. 주변 잡음의 첨가에 의한 음성신호의 왜곡은 가산잡음의 제거에 널리 쓰이는 스펙트럼 차감법과, 대역통과 필터링으로 음성의 동적인 특성을 강조함으로서 제거하였다. 잡음처리 기술의 개발과 평가를 위해, 실제 환경의 잡음을 사용하여 모의된 잡음환경에서 롬바드 음성을 수집하고, 이를 사용하여 잡음처리 방법을 비교 실험하였다.

논문의 구성은 다음과 같다. 2장에서는 실험에 사용한 음성자료의 수집과 분석에 대해 기술하고, 3장에서는 잡음에 의한 음성의 왜곡모델과 왜곡모델에 따른 가산잡음의 제거와 롬바드 효과에 의한 왜곡의 보정방법을 설명한다. 4장에서는 음성인식 실험을 통해 제안한 방법의 타당성을 검토하고, 5장에서 결론을 맺는다.

Ⅱ. 잡음과 음성 자료

2.1 잡음의 종류

효과적인 잡음처리 방법의 개발을 위해서는 실제 잡음환경에서 발성된 음성자료가 필요하며, 이러한 자료는 연구대상의 모든 왜곡을 포함하고 있어야 한다. 그러나 다양한 실제 잡음환경에서의 음성자료의 수집은 많은 시간과 노력을 필요로 한다. 본 논문에서는 음성인식기의 응용이 기대되는 자동차, 전시장, 시내 공중전화 부스, 거리, 전산실에서 발생한 잡음을, 헤드폰을 통하여 발성자에게 들려줌으로써 잡음환경을 모의하고, 모의된 잡음환경에서 롬바드 효과에 의해 변형된 음성을 수집하여 실험하였다. 실험에 사용한 잡음자료는 JEIDA(Japan electronic industry development association)에서 수집한 자료의 일부이다[22].

표 1. 실험에 사용한 잡음의 분류

잡음번호	잡 음 종 류	잡음번호	잡 음 종 류
0	자동차 (고속도로)	7	도로변 (동네 1)
1	자동차 (시가지)	8	도로변 (동네 2)
2	전시회장	9	전산실 1
3	전시회장	10	전산실 2
4	주택가 공중전화부스	11	전산실 3
5	역북측 공중전화부스	12	전산실 4
6	역남측 공중전화부스	13부터 21	0부터 8과 동일한 잡음

표 1은 실험에 사용한 잡음목록으로, 분류를 위해서 번호를 붙였다. 여기서 13번부터 21번 잡음은 0번부터 8번과 동일한 잡음으로, 에너지를 크게 하여 실험에 사용하였다. 0번은 고속도로를 주행중인 차안에서 수집되었고, 1번은 시내를 주행중인 차안에서 수집되었으며, 잡음의 크기는 0번보다 작지만 변이는 크다. 2번과 3번은 전시회장의 여러사람의 음성과 발자국 소리이고, 4번은 공중전화부스에서 녹음된 자동차 소리와 사람의 음성이다. 5번은 기차역 근처의 공중전화부스에서 수집된 소음으로, 기차, 자동차, 사람의 음성이다. 6번도 기차역 근처의 공중전화부스에서 수집된 소음으로 스피커에서 커다란 음악과 여성의 목소리가 계속 흘러나오며, 간헐적으로 자동차가 지나가는 소음이 포함되어있다. 7번과 8번은 인파가 많은 거리의 소음으로 자동차, 사람들의 음성과 발자국 소리이다. 9번, 10번, 11번, 12번은 전산실에서 발생한 소음이다.

표 1의 잡음은 종류에 따라 상이한 음향학적인 특성을 가진다. 그림 2부터 그림 5까지 가로축은 0-8kHz를 19개의 bark-scale로 나눈 것이고, 세로축은 각대역의 에너지를 표시하였다. 그림에서 보듯이 잡음의 종류에 따라 각기 다른 스펙트럼 포락을 보인다.

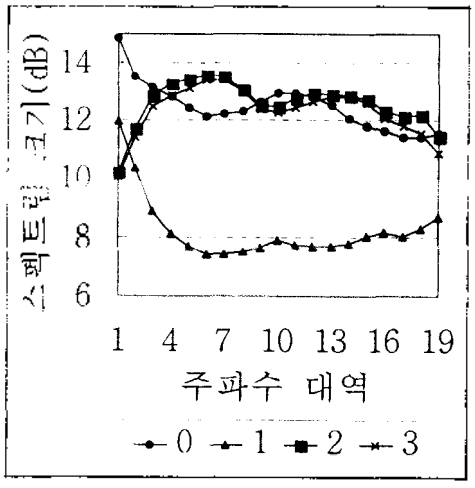

그림 2. 잡음 0, 1, 2, 3의 스펙트럼 포락

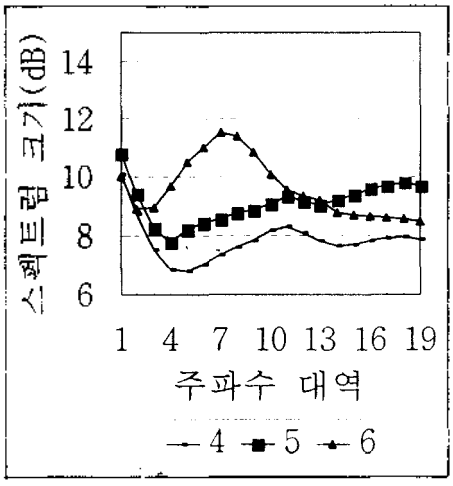

그림 3. 잡음 4, 5, 6의 스펙트럼 포락

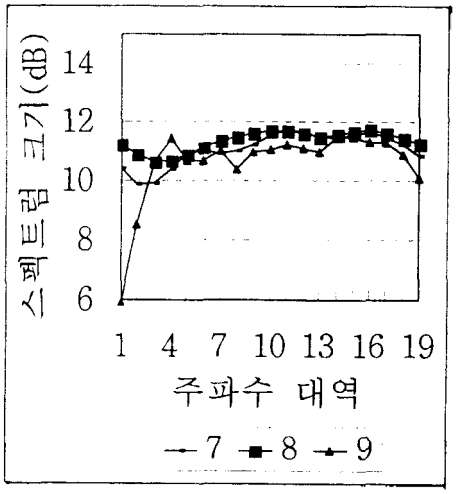

그림 4. 잡음 7, 8, 9의 스펙트럼 포락

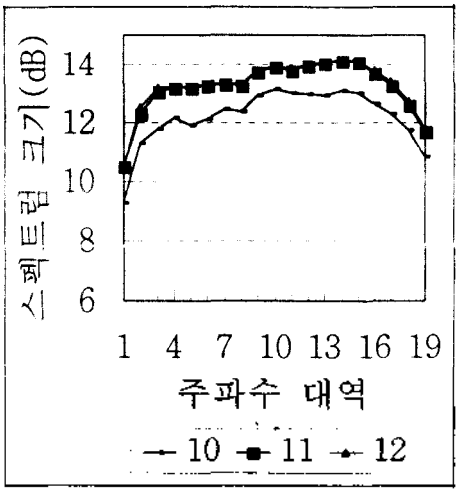

그림 5. 잡음 10, 11, 12의 스펙트럼 포락

2.2 롬바드 음성의 수집

잡음환경에서의 롬바드 효과에 의해 왜곡된 음성을 수집하기 위하여, 헤드폰을 통하여 표 1의 22개의 잡음을 발성자에게 들려주어 잡음환경을 모의하고, 모의된 22개의 잡음환경에서 발성된 음성을 녹음하였다. 자연스러운 발성을 유도하기 위해서 발성자의 70-100cm 전방에 인물사진을 놓고, 그 사람에게 말하는 기분으로 발성할 것을 요구하였다. 20대의 남자 10명과 여자 10명의 음성을 수집하였고, 특정내용의 단어집합에 종속적이지 않은 잡음처리 방법을 개발하기 위해 음운학적으로 균형을 이룬 표 2의 단어집합을 사용하였다.

표 2. 녹음 단어 집합

의무	깨끗이	뉴욕	더위	벽화	예보	철회	원양	왜구	의례
몸집	햇볕	좌표	뜀틀	앞서	위쪽	의원	세제	금융	야구
되찾다	폐대	필요	왜병	돌쇠	법원	좌우	혜택	교류	신의
끝점	뒤다	의제	태양계	열쇠	규모	뵘다	펜치	화약	꽃밭
원유	쿵더쿵	돼지	으뜸	재료	계획표	벗짚	세월	초과	윗변

수집된 단어의 SPL(sound pressure level)과 헤드폰을 통해 들려준 잡음의 SPL을, 각 22개의 환경별로 그림 6에 나타내었다. 가로축은 잡음번호이고, 세로축은 SPL이다.

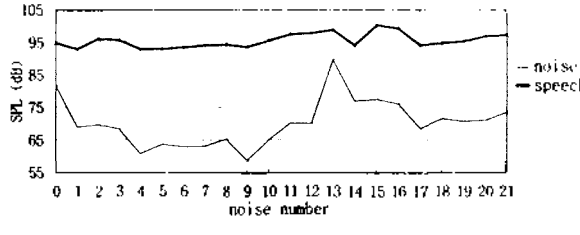

그림 6. 수집된 화자 20인의 음성의 각 환경별 SNR

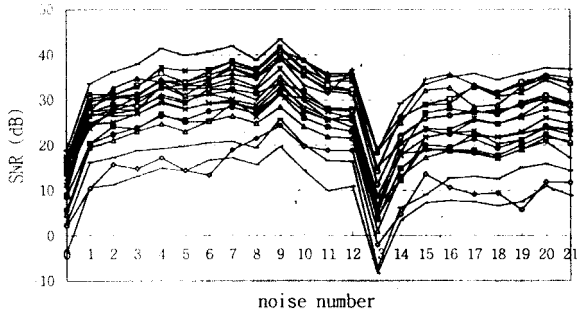

그림 7. 수집된 화자 20인의 음성의 각 환경별 SNR

그림 6에서 보듯이, 발성자는 롬바드 효과에 의해서 주변 잡음의 SPL에 따라 발성음의 세기를 비례적으로 증가시키는 경향을 볼 수 있다. 그러나, 발성음의 크기는 화자, 잡음의 종류와 크기에 따라 변이가 크다. 그림 7에 보듯이 20명의 화자가 발성한 음성의 SNR(signal-to-noise ratio)은 −10dB에서 45dB사이로 크게 변이하며, 음성인식기의 성능저하의 한 요인이 된다.

Ⅲ. 잡음처리 방법

3.1 잡음에 의한 음성의 왜곡 모델링

잡음환경에서 발성된 음성은 롬바드 효과에 의한 발성방식의 변이와 잡음의 첨가에 의해, 학습환경의 음성과는 상이한 특성을 가지므로 음성인식기의 성능을 저하시킨다. 본 연구에서는 잡음의 영향에 의한 음성의 변이과정을 나타내는 왜곡모델을 제안하고, 제안한 모델에 따라 변이를 제거하여 잡음환경에 강인한 특징을 추출한다.

롬바드 효과에 의한 변이는 평균에너지가 일정한 값을 갖도록 정규화된 단어에서의 스펙트럼 포락을 변이 시키는 왜곡과, 전체적인 에너지를 변이 시키는 부분으로 나누어 모델링하였다. 첫째로, 정규화된 단어에서의 스펙트럼 포락의 변이는, 롬바드 효과에 의해 포만트의 위치, 포만트의 대역폭(bandwidth), 기본주파수의 변이, 각 주파수 대역의 에너지 등이 변하므로 발생한다[11]. 이러한 왜곡요인은 비선형적인 주파수 변환(frequency warping) 함수 $F(\cdot)$와, 주파수 대역별 스펙트럼의 크기변이 함수 $A(\cdot)$로 모델링하였다. 즉, 깨끗한 음성의 스펙트럼 $S(\omega)$가 롬바드 효과에 의한 주파수 변환과 주파수 대역별 스펙트럼의 크기변이에 따라 스펙트럼 $Y_1(\omega)$로 변환된다.

$$Y_1(\omega) = A(\omega)S(F(\omega)) \tag{1}$$

둘째, 화자는 그림 6에서 보는 바와 같이 주변잡음이 크면, 효과적인 의사소통을 위해서 발성음의 세기를 잡음의 크기에 비례하여 증가시킨다. 그러나, 이러한 발성음의 세기의 변화는 그림 7에서 보는 바와 같이, 잡음의 종류와 세기, 음소의 종류, 화자의 특성에 따라서 상이한 변이를 나타내는 왜곡요인이다. 롬바드 효과에 의한 전체적인 발성음의 크기변화는 에너지 증가 G가 $Y_1(\omega)$에 곱해져서 $Y_2(\omega)$로 변환되는 것으로 모델링하였다.

$$Y_2(\omega) = G \cdot Y_1(\omega) = G \cdot A(\omega)S(F(\omega)) \tag{2}$$

마지막으로 주변잡음이 음성신호에 더해짐으로 생기는 음성신호의 왜곡은 주파수 영역에서 가산적인 항으로 나타낼 수 있다. 잡음의 스펙트럼을 $N(\omega)$라 하면, 롬바드 효과에 의해 왜곡된 음성의 스펙트럼 $Y_2(\omega)$는 $N(\omega)$가 첨가되어 롬바드 음성에 잡음이 첨가된 스펙트럼 $Y_3(\omega)$가 된다.

$$Y_3(\omega) = Y_2(\omega) + N(\omega) = G \cdot Y_1(\omega) + N(\omega) = G \cdot A(\omega)S(F(\omega)) + N(\omega) \tag{3}$$

3.2 잡음 영향 제거

롬바드 효과와 가산잡음이 첨가된 음성에서 왜곡요인을 제거하기 위해서, 잡음에 의한 음성의 왜곡과정의 역과정인 그림 8의 과정을 사용하였다.

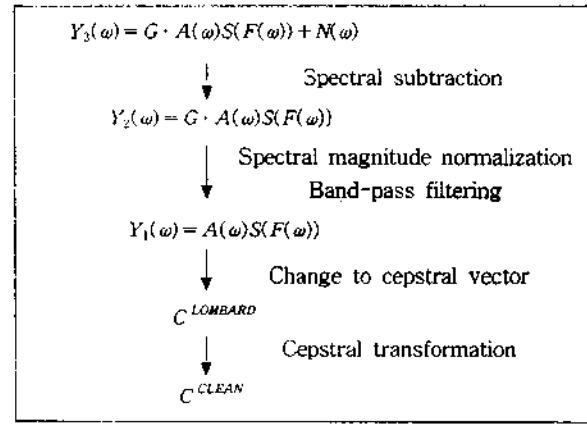

그림 8. 음성의 개선 과정

3.2.1 가산잡음의 제거

잡음하에서 음성인식기에 입력된 음성의 스펙트럼 $Y_3(\omega)$로부터 잡음의 스펙트럼 $N(\omega)$를 제거하여, 롬바드 영향만을 받은 음성의 스펙트럼 $Y_2(\omega)$를 얻기 위해 스펙트럼 차감법을 사용하였다[18]. 스펙트럼 차감법은 잡음의 성질이 음성에 비해 완만하게 변화하므로, 잡음을 묵음구간에서 추정하여 입력음성에서 제거하는 방법이다. 본 연구에서는 잡음의 평균 스펙트럼 크기 $|N(\omega)|$를 묵음구간의 스펙트럼 크기의 평균으로 구하여, 이를 시간 t에서의 입력음성의 스펙트럼 크기 $|Y_{3,t}(\omega)|$에서 제거하여 $|Y_{2,t}(\omega)|$를 추정하였다. 음성이 시간축에서 변화하는 특성을 반영하기 위해, 가중치 평균을 사용하여 $|\overline{Y_{3,t}(\omega)}|$를 구하였다.

$$|\overline{Y_{3,t}(\omega)}| = \frac{|Y_{3,t-1}(\omega)| + 2\cdot|Y_{3,t}(\omega)| + |Y_{3,t+1}(\omega)|}{4} \quad (4)$$

추정된 스펙트럼이 음수가 되는 것을 방지하기 위해 식 5의 flooring을 사용하였다.

$$|Y_{3,t}(\omega)| = |\overline{Y_{3,t}(\omega)}| - |N(\omega)|$$
$$\text{if } |\overline{Y_{3,t}(\omega)}| - |N(\omega)| > 0.1\cdot|N(\omega)| \quad (5)$$
$$= 0.1\cdot|\overline{Y_{3,t}(\omega)}| \text{ otherwise}$$

3.2.2 에너지의 정규화와 청각특성의 반영

앞절에서 스펙트럼 차감법을 사용하여 얻은 스펙트럼 $Y_2(\omega)=G\cdot Y_1(\omega)=G\cdot L(\omega)S(F(\omega))$로부터 발성에너지의 변이 G를 제거하기 위해, 스펙트럼 크기의 정규화를 한 후, 음성인식에 보다 유효한 특징으로의 변환을 위한 대역통과 필터링을 하였다.

잡음환경에서 화자는 의사전달을 분명히 하기 위해 발성음의 세기를 잡음의 크기에 맞추어서 변화시키는 특성이 있으나, 화자와 잡음의 종류에 따라 큰 변이가 존재한다. 발성음의 세기 변이를 제거하기 위해, 변이요인 G는 음성의 평균 스펙트럼 크기를 기준 스펙트럼 크기로 하여, 입력 음성의 스펙트럼 크기의 증가량으로 정의하였다.

$$G = \frac{\text{입력신호에서 음성부분만의 스펙트럼 크기의 평균}}{\text{기준 스펙트럼 크기}} \quad (6)$$

정의된 발성음의 세기변이는 입력신호에 따라 결정되므로, 화자, 잡음의 종류, 단어내용에 종속적인 특징을 갖는다. $Y_2(\omega)$를 변이요인으로 나누어, 세기 변이가 제거된 $Y_1(\omega)$는 음성부분의 스펙트럼 크기는 일정한 평균을 갖는다. 이러한 스펙트럼 크기의 정규화는 발성음성의 변이를 제거할 뿐만 아니라, 이후의 처리인 Lin-Log RASTA에서 사용하는 대역통과 필터링의 입력을 안정시킨다. Lin-Log RASTA에서 대역통과 필터링은 가산잡음과 채널잡음의 제거를 위해서 작은 스펙트럼 값에서는 선형적이고, 큰 값에서는 지수적인 스펙트럼 영역에서 수행된다[10]. 이러한 스펙트럼 영역으로 최적의 변환을 위해서는 입력신호에 종속적인 파라미터 J값이 결정 되야 하므로, 음성인식률에 민감한 영향을 끼친다. 이러한 문제는 여러 가지 J 값으로 인식기를 학습시키거나, 주파수 사상(spectral mapping)을 사용하여 해결 할 수 있지만, 본 연구에서는 스펙트럼 크기의 정규화로 인해 입력신호의 특성이 정규화되므로 이러한 처리를 사용하지 않았다.

대역통과 필터링이외에 고역통과 필터나 마스크(mask) 필터를 스펙트럼 영역이나 캡스트럼 영역에서 사용하여도, 잡음의 제거와 롬바드 음성의 인식에도 유효하다는 결과에서 보듯이[3,4,10], 동적인 특성을 강조하는 대역통과 필터링은 잡음음성의 처리에 필요하다. Lin-Log RASTA 처리를 위해, $Y_1(\omega)$를 $Y(\omega)=\ln(1+J\cdot Y_1(\omega))$으로 변환하여 필터 $0.1\cdot\frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.94z^{-1}}$를 통과시킨 후, 스펙트럼의 크기가 항상 양수가 되기위해 근사된 역변환 $Y_1(\omega)=e^{Y(\omega)}/J$를 사용하여 원래의 스펙트럼으로 변환한다.

3.2.3 캡스트럼 변환에 의한 스펙트럼 구조의 왜곡 보정

앞절의 처리를 통해서 얻은 $Y_1(\omega)=A(\omega)S(F(\omega))$에서 왜곡요인인 $F(\cdot)$와 $A(\cdot)$를 캡스트럼 공간에서 제거하기 위해, $Y_1(\omega)$에서 유도한 캡스트럼의 k번째 차원의 계수를 $C_k^{LOMBARD}$, 조용한 환경에서 발성한 음성의 스펙트럼 $S(\omega)$에서 유도한 캡스트럼의 n번째 차원의 계수를 C_n^{CLEAN}이라 하면, 캡스트럼과 스펙트럼의 관계로부터 식 7과 8이 성립한다. $F(\cdot)$의 역함수를 $F^{-1}(\cdot)$라하고, 식 8-b를 식 7에 대입하면 식 9의 결과를 얻는다.

$$C_n^{CLEAN} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|S(\omega)|\,e^{j\omega n}d\omega \quad (7)$$

$$\log|A(\omega)S(F(\omega))| = \sum_{k=-\infty}^{\infty} C_k^{LOMBARD} e^{-j\omega k} \quad (8\text{-a})$$

$$\log|A(\omega)S(F(\omega))| = \log|A(\omega)||S(F(\omega))| = \log|A(\omega)| + \log|S(F(\omega))|$$
$$= \sum_{k=-\infty}^{\infty} C_k^{LOMBARD} e^{-j\omega k} \quad (8\text{-b})$$

$$C_n^{CLEAN} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|S(F(F^{-1}(\omega)))|\,e^{j\omega n}d\omega$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{k=-\infty}^{\infty} C_k^{LOMBARD} e^{-jF^{-1}(\omega)k} - \log|A(F^{-1}(\omega))|\right]e^{j\omega n}d\omega$$
$$= \sum_{k=-\infty}^{\infty} A(n,k)\cdot C_k^{LOMBARD} + B(n) \quad (9)$$

단, $A(n,k)$는 $\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jF^{-1}(\omega)k}e^{j\omega n}d\omega$, $B(n)$은 $-\frac{1}{2\pi}\int_{-\pi}^{\pi}\log|A(F^{-1}(\omega))|\,e^{j\omega n}d\omega$이다. 식 9에서 롬바드 효과에 의한 왜곡요인인 $F(\cdot)$와 $A(\cdot)$는 화자, 음소, 잡음환경에 종속적인 행렬 $A(n,k)$과 벡터 $B(n)$으로 표시된다. 본 연구에서는 이러한 종속성을 벡터양자화를 이용하여 캡스트럼 공간을 분할하여 특징파라미터 영역에서 근사하였다. $A(n,k)$과 $B(n)$을 구하기 위해서, 첫째, 롬바드 효과에 의해 왜곡된 단어의 캡스트럼과 조용한 환경에서 발성된 단어의 캡스트럼을, 해당 단어의 HMM(hidden Markov model)과 Viterbi 정합하여 같은 상태에 머무는 캡스트럼의 쌍을 만든다. 둘째, 롬바드 음성의 캡스트럼을 벡터양자화를 통해, 각각 속하는 코드워드 별로 모은다. 셋째, 각 코드워드 별로 롬바드 음성의 캡스트럼을 식 9와 같이 선형변환할 경우에, 대응되는 쌍의 조용한 환경에서 발성한 음성

의 캡스트럼과 가장 작은 평균제곱에러를 갖도록 $A(n, k)$과 $B(n)$을 추정한다. 분할된 영역에서 M개의 N-1차원의 롬바드 캡스트럼의 행렬을 C^L, M개의 N-1차원의 조용한 음성의 캡스트럼의 행렬을 C^C, i번째의 롬바드 캡스트럼의 j번째 차원을 $C^L_{i,j}$, 대응되는 조용한 음성의 캡스트럼을 $C^C_{i,j}$이라고 하면, 식 9는 식 10과 같이 쓸 수 있다.

$$\begin{bmatrix} C^L_{1,1} & C^L_{1,2} & \cdots & C^L_{1,N-1} & 1 \\ C^L_{2,1} & C^L_{2,2} & \cdots & C^L_{2,N-1} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C^L_{M,1} & C^L_{M,2} & \cdots & C^L_{M,N-1} & 1 \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{N-1,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{N-1,2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1,N-1} & a_{2,N-1} & \cdots & a_{N-1,N-1} \\ b_1 & b_2 & \cdots & b_{N-1} \end{bmatrix}$$

$$\approx \begin{bmatrix} C^C_{1,1} & C^C_{1,2} & \cdots & C^C_{1,N-1} \\ C^C_{2,1} & C^C_{2,2} & \cdots & C^C_{2,N-1} \\ \vdots & \vdots & \vdots & \vdots \\ C^C_{M,1} & C^C_{M,2} & \cdots & C^C_{M,N-1} \end{bmatrix} \tag{10}$$

식 10에서 C^L을 singular value decomposition하여 $C^L = UWV^T$로 나타냈을때, U는 M×N, V는 N×N행렬이며, 두 개의 행렬은 모두 열벡터가 정규직교(orthonormal)하고 W는 영보다 같거나 큰 양수로 이루어진 대각행렬이다[20]. 식 10의 양변에 $V\{diag(1/w_j)\}U^T$을 곱하여 $A(n, k)$과 $B(n)$을 추정한다. 단, $diag(1/w_j)$는 W의 j행과 j열의 원소의 역수로서 w_j가 영이면 영으로 정의된다. 스펙트럼 포락의 변이가 제거된 캡스트럼은 그림 9에서 처럼, 롬바드 음성의 캡스트럼을 양자화하여, 해당되는 $A(n, k)$과 $B(n)$을 사용하여 변환함으로써 얻는다. 학습자료에 지나치게 종속적이지 않기 위해, 최종 변환된 캡스트럼 파라미터는 변환된 캡스트럼과 변환되기 전의 캡스트럼의 평균값으로 한다.

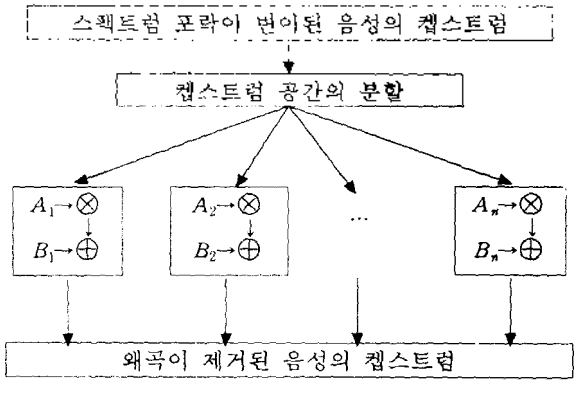

그림 9. 캡스트럼 변환을 통한 스펙트럼 포락의 변이제거

Ⅳ. 음성인식 실험 및 분석

제안한 잡음처리 방법의 유효성을 알아보기 위해 22가지 잡음환경에서 발성된 롬바드 음성에 잡음을 첨가하여 음성인식 실험을 수행하였다. 음성인식 실험은 음성을 입력받아 특징파라미터로 변환하는 음향분석 및 특징추출 단계, 벡터 양자화를 통한 코드북의 작성 및 HMM 단어모델을 학습하는 단계, 저장된 모델로부터 단어를 인식해 내는 인식 단계로 구성되어 있다.

4.1 실험조건

대부분의 음성인식기가 응용되는 상황은 학습환경과 평가환경이 상이한 것이 일반적이다. 이러한 점을 고려하여 HMM의 학습자료로서, 남자 5명, 여자 5명이 조용한 환경에서 50단어를 2회 반복 발성한 음성을 사용하였고, 평가를 위해서는 학습에 참여하지 않은 남자 5명과 여자 5명이 표 1의 잡음환경에서 발성된 11000개의 단어를 사용하였다. 음성은 16kHz, 16bit 샘플링 되었고, $1-0.95z^{-1}$로 전처리 하였다. 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 특징파라미터는 다음과 같다.

(1) PROP : 본 연구에서 제안한 특징추출 방법으로 FFT (fast Fourier transformation)를 사용하여 19개의 bark-scale filter bank를 구성하고 스펙트럼 차감법, 스펙트럼 크기의 정규화, Lin-Log 스펙트럼영역에서의 대역통과 필터링, DCT(discrete cosine transformation)를 사용하여 캡스트럼 벡터로 변환한 후, 캡스트럼 변환을 하였다.

(2) BARK-CEP : 기본적인 비교대상의 파라미터로 bark-scale filter bank에서 유도한 캡스트럼 벡터로 잡음처리를 하지 않은 특징벡터이다.

(3) SPEC-SUB : 스펙트럼 차감법을 수행한 후 캡스트럼을 구한다.

(4) LIN-LOG RASTA : LIN-LOG 스펙트럼 영역에서 대역통과 필터링을 한다[10].

(5) LPC-CEP : LPC(linear predictive coding) 계수로부터 유도한 멜캡스트럼 계수이다.

(6) PROJ : LPC-CEP에 거리척도로 projection 거리의 사용하였다[9].

(7) SMC : SMC(short-time modified coherence)방법으로 구한 LPC 계수로부터 유도한 멜캡스트럼 계수이다[8].

LPC로 얻은 캡스트럼은 식 11의 bilinear 변환을 통해서 저주파 대역의 해상도를 고주파 대역보다 크게한다

[2]. 캡스트럼의 n번째 차원을 f_n이라 하면 $g_{k,0}$는 멜캡스트럼의 k번째 차원의 계수이다. 여기서, a는 0.64를 사용하였다.

$$g_{0,n}=ag_{0,n-1}+f_n$$
$$g_{1,n}=ag_{0,n-1}+(1-a^2)g_{0,n-1} \quad (11)$$
$$g_{k,n}=a\{g_{k,n-1}-g_{k-1,n-1}\}+g_{k-1,n-1},\ k=2,3,4,\cdots$$

4.2 실험 및 결과

여러 가지 특징추출 방법을 음성인식 실험을 통하여 비교하기에 앞서, 제안한 방법인 PROP에 의한 특징벡터와 잡음처리를 하지 않은 특징추출 방법인 BARK-CEP에 의한 특징벡터를 비교하였다. 그림 10과 그림 11은 각각 조용한 환경과 잡음 환경 21번에서 "금융"을 발성한 음성신호를 BARK-CEP으로 구한 캡스트럼이다. 그림

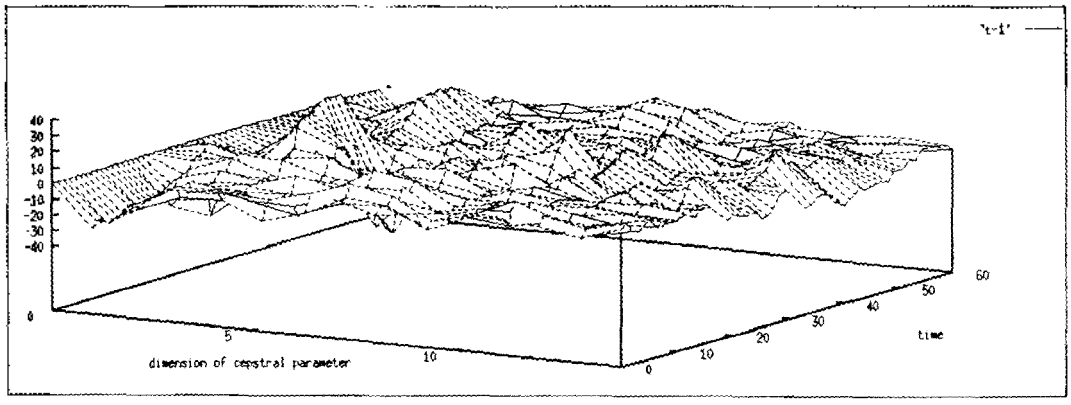

그림 10. 조용한 환경에서 발성한 단어의 캡스트럼.
잡음처리 방법을 사용하지 않는 BARK-CEP을 사용한 분석.

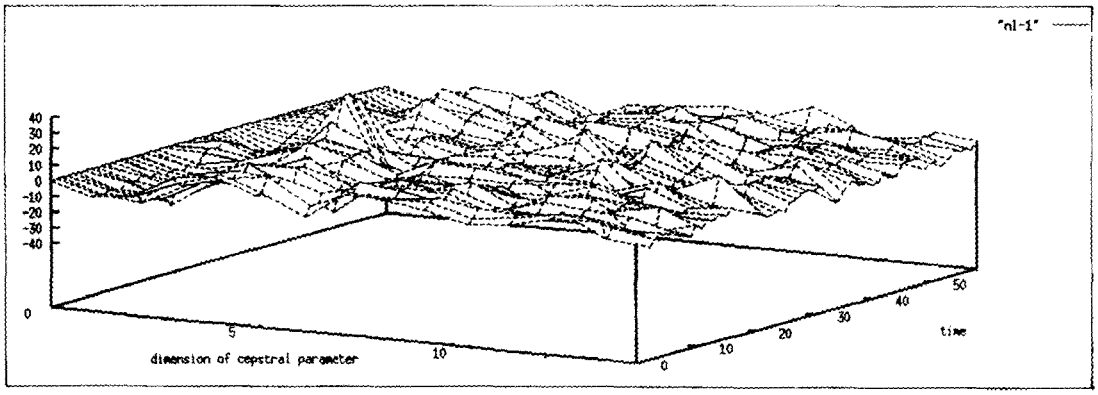

그림 11. 잡음 환경에서 발성한 단어의 캡스트럼.
잡음처리 방법을 사용하지 않는 BARK-CEP을 사용한 분석.

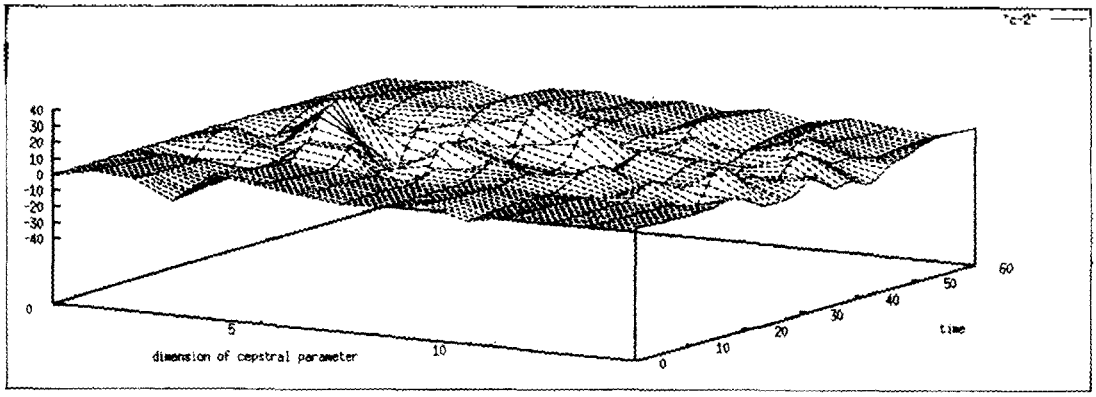

그림 12. 조용한 환경에서 발성한 단어의 캡스트럼.
제안한 잡음처리 방법을 사용한 분석.

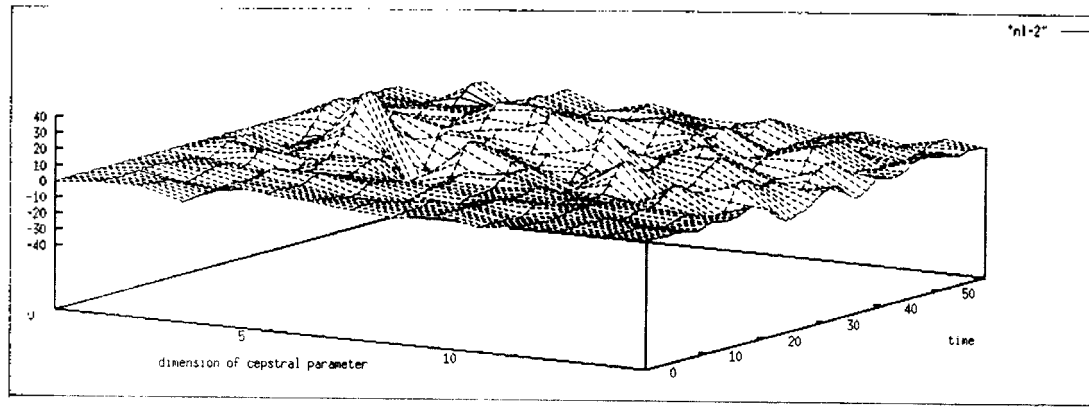

그림 13. 잡음환경에서 발성한 단어의 켑스트럼.
제안한 잡음처리 방법을 사용한 분석.

12와 그림 13은 조용한 환경과 잡음 환경에서 발성된 음성을 제안한 방법인 PROP로 구한 켑스트럼 벡터로, BARK-CEP으로 구한 켑스트럼인 그림 10과 그림 11에 비해 잡음의 제거와 롬바드 효과의 보정을 통해, 서로 유사한 켑스트럼 구조를 가짐을 볼 수 있다.

잡음환경하의 음성인식에 제안한 방법의 유효성을 알아보기 위해서, 먼저 표 1의 잡음의 일부인 7번부터 12번, 20번, 21번의 잡음환경에서 발성한 음성에 10dB SNR로 잡음을 첨가하여 잡음처리 기술의 개발과 평가를 하였고, 인식률은 표 3에 나타내었다.

표 3. 예비 실험의 음성 인식률 (SNR 10dB)

처리 방법	제안방법 PROP	BARK-CEP	SPEC-SUB	LIN-LOG RASTA	LPC-CEP	PROJ	SMC
인식률(%)	95.13	72.75	78.50	87.93	60.68	69.58	71.35

표 3의 인식률에서 보듯이 제안한 방법인 PROP는 롬바드 효과의 보정과 잡음의 제거에 효과적임을 알수 있다. LPC기반의 특징추출 방법인 LPC-CEP, PROJ, SMC는 음성을 전극모델로 가정하므로 잡음이 첨가된 음성신호에 대해서는 부적절한 가정이 되므로, 비모수적인 방법인 FFT 기반의 특징추출인 PROP, BARK-CEP, SPEC-SUB, LIN-LOG RASTA보다 인식률이 저하되었다. 표 3의 실험에서 LIN-LOG RASTA 처리는 J값으로 10^{-7}을 사용하였다. J값이 10^{-6}일 때는 87.50%의 인식률을 얻었고 10^{-5}나 10^{-8}을 사용하였을 경우에는 인식률이 크게 저하되었다. 제안한 방법에서 켑스트럼의 변환에 사용되는 행렬과 벡터는 학습자료에 포함된 화자들이 7번부터 12번까지와 20번, 21번의 잡음환경에서 발성한 롬바드 음성과 조용한 환경에서 발성한 음성을 이용하여 추정하였고, 이후의 실험에도 동일한 행렬과 벡터를 사용하였다.

표 4는 표 1의 모든 22개 잡음환경과 여러 가지 SNR에서의 음성 인식률의 평균을 나타낸 것이다. SNR CLEAN은 잡음을 첨가하지 않은 음성이고, REAL은 헤드폰을 통해 들려온 잡음을 들려온 크기대로 첨가한 음성이다. 20dB, 10dB, 0dB는 한 화자가 각 잡음환경에서 발성한 50단어의 평균 SNR이 20dB, 10dB, 0dB이 되도록 잡음을 음성에 첨가한 것이다. SMC 방법은 LPC 계열의 파라미터 중 예비실험에서 가장 좋은 인식률을 보였으나, 인식률의 향상이 적으며 계산량이 많고, LIN-LOG RASTA는 J값의 선택에 어려우므로 실험에서 제외하였다. 제안방법인 PROP는 조용한 환경에서 발성한 음성을 학습자료로 사용한 것이고, PROP-LOM도 제안한 방법이지만, 7번부터 12번까지와 20번, 21번 잡음환경에서 발성한 음성에 SNR REAL이 되도록 잡음을 첨가하여 학습자료로 사용하였으므로, 롬바드 효과와 잡음에 의한 왜곡이 포함하고 있으므로, 켑스트럼 변환은 사용하지 않았다.

표 4. 여러 SNR과 특징추출 방법에서의 인식률

특징추출 SNR	PROP	PROP-LOM	BARK-CEP	SPEC-SUB	LPC-CEP	PROJ
CLEAN	**98.60**	**99.17**	96.15	95.58	92.75	92.12
REAL	**97.90**	**99.02**	91.16	91.17	86.03	87.30
20dB	**98.15**	**99.01**	92.45	91.65	85.92	87.45
10dB	**95.91**	**97.71**	79.68	82.26	72.91	75.98
0dB	**79.75**	**83.80**	43.74	55.03	42.56	46.42

잡음이 첨가되지 않고 롬바드 효과에 의해 왜곡된 SNR CLEAN 음성에 대한 실험에서, 잡음영향에 대한 처리를 하지 않는 BARK-CEP의 인식률 96.15%이고, 제안방법인 PROP는 98.60%로서 에러율을 64% 감소시킴으로써, 제안한 방법이 롬바드 효과를 완화하는데 효과적임을 알 수 있다. 또한, 실제 잡음환경의 발성된 음성의 SNR인 REAL 환경에서는 에러율을 76%, 20dB에서는 75%, 10dB에서는 80%, 0dB에서는 64% 감소시킴으로써, 제안한 방법은 첨가된 잡음의 제거와 롬바드 효과에 모두 효과적임을 실험적으로 확인할 수 있었다. 롬바드 효과와 잡음에 의한 왜곡을 학습자료가 포함하고 있으므로, 가산잡음과 발성

음의 세기의 변이만을 처리하는 PROP-LOM은 학습환경이 평가환경의 왜곡을 포함하고 있어서, 가장 높은 음성인식률을 보였다.

표 4에서도 LPC 기반의 특징추출 방법인 LPC-CEP, PROJ는 FFT기반의 특징추출에 비해서, 잡음에 의해 인식률이 더욱 저하되는 경향을 보였고, 잡음처리 방법인 SPEC-SUB과 PROJ는 잡음처리를 하지 않는 BARK-CEP과 LPC-CEP에 비해, SNR CLEAN환경에서는 오히려 인식률이 저하되었고, SNR이 작을 때 잡음처리의 효과가 나타났다.

V. 결론 및 검토

본 논문에서는 잡음환경에서 음성인식기의 성능저하를 방지하기 위해 롬바드 효과의 보정과 잡음제거 방법을 연구하였다. 롬바드 효과와 잡음의 첨가에 의한 음성의 왜곡과정을 나타내기 위해서, 잡음환경에서의 음성의 왜곡모델을 제안하였고, 왜곡모델을 이용하여 잡음음성에 포함된 여러 가지 왜곡요인을 제거하였다. 롬바드 효과에 의해 음성의 포만트의 위치, 포만트 대역폭, 기본주파수, 스펙트럼 틸트, 각 주파수 대역의 에너지, 전체 발성에너지 등이 변하므로, 스펙트럼 포락의 모양이 조용한 환경에서 발성한 음성과는 상이하다. 본 논문에서는 롬바드 효과에 의한 왜곡을 비선형적인 주파수 변환과 주파수 대역별 스펙트럼의 크기변이로서, 평균에너지가 일정한 값을 갖도록 정규화된 단어의 스펙트럼 포락의 변이를 모델링하였고, 전체적인 에너지를 변이시키는 왜곡은 입력단어의 스펙트럼 크기에 종속적인 상수의 곱으로 모델링하였다. 이러한 모델에 따라서 롬바드 효과에 의한 정규화된 단어의 스펙트럼 포락의 변이는 켑스트럼 영역에서 추정하여 보정하였고, 발성음의 세기의 변이는 스펙트럼 크기의 정규화를 통해 제거하였다. 잡음의 첨가는 스펙트럼 차감법과 대역통과 필터링을 사용하여 제거하였다.

HMM을 사용한 화자독립 음성인식실험을 하여, 음성의 왜곡모델에 따른 잡음의 제거와 롬바드 음성의 처리방법의 유효성을 실험적으로 확인할 수 있었다. 실험에 사용한 음성자료는 실제현장의 잡음인 자동차, 전시장, 시내 공중전화 부스, 거리, 전산실의 잡음을 이용하여 모의된 잡음환경에서 발성된 음성에 잡음을 첨가하여 실험하였다.

향후 연구로는 가산잡음의 제거를 위해 효과적인 스펙트럼 차감법의 개발과, 특정한 잡음환경에서 인식률을 높이기 위해서는 잡음의 특성을 이용하는 방법이 필요하며, 본 연구에서는 다루지 않았지만 전화선을 경유할 때 발생하는 채널잡음의 제거기술에 대한 연구가 필요하다.

참 고 문 헌

1. A. Nadas, D. Nahamoo, and M. A. Picheby, "Speech recognition using noise-adaptive prototypes," IEEE Trans on ASSP, Vol. 37, No. 10, pp. 1495-1502, 1989.
2. A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," proc. of IEEE, Vol. 60, No. 6, pp. 681-691, 1972.
3. B. A. Hanson and T. H. Applehaum, "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech," proc. of ICASSP, pp. 79-82, 1993.
4. B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech," proc. of ICASSP, pp. 857-890, 1990.
5. C. Mokbel and G. Chollet, "Speech recognition in adverse environment: speech enhancement and spectral transformations", proc. of ICASSP, pp. 925-928, 1991.
6. D. B. Paul, "A speaker-stress resistant hmm isolated word recognizer," proc. of ICASSP, pp. 713-716, 1987.
7. D. B. Roe, "Speech recognition with a noise-adapting codebook," proc. of ICASSP, pp. 1139-1142, 1987.
8. 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.
9. 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.
10. 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.
11. 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.
12. J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," proc. of ICASSP, pp. 18A2. 1-2.4, 1990.
13. J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACE) for speech recognition in noise and Lombard effect," IEEE Trans. on SAP, Vol. 2, No. 4, pp. 598-614, Oct. 1994.
14. 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.
15. 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.
16. 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.
17. 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.
18. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on ASSP, Vol. 27, No. 2, pp. 113-120, 1979.
19. S. Tamura and A. Waibel, "Noise reduction using connectionist models," proc. of ICASSP, pp. 553-556, 1988.
20. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical recipes in C, Cambridge univ. Press, New York, 1992, pp. 59-70.
21. Y. Chen, "Cepstral domain stress compensation for robust speech recognition," proc. of ICASSP, pp. 717-720, 1987.
22. 소음 데이터베이스, 일본전자공업진흥협회, 1990.

▲지 상 문(Sang-Mun Chi)

1991년 2월 : 서울대학교 수학교육과 (학사)
1993년 2월 : 한국과학기술원 수학과 (석사)
1993년 3월~현재 : 한국과학기술원 전산학과 박사과정 재학중
※주관심분야 : 음성 인식, 잡음 처리

▲오 영 환(Yung-Hwan Oh) : 제 15권 1호 참조