음성 신호처리를 위한 군중잡음 제거 모델

안용운[0], 김중환, 김상철
한국외국어대학교 컴퓨터 공학과, 멀티미디어 정보통신 연구실
coolcld@hotmail.com[0], jhkim@hufs.ac.kr, kimsa@hufs.ac.kr

A Crowd Noise Reduction Model for Speech Signal processing.

Yong-Woon Ahn[0], Joong-Hwan Kim, Sang-Chul Kim
MIT Lab, Dept. of Computer Science & Engineering, Hankuk University of Foreign Studies

요 약

군중잡음(crowd noise)이 발생하는 환경에서 음성 통화 및 화자 인식을 할 때에는 음성에 파열음이나 마찰음과 같은 유색잡음(colored noise)이 부가되어 원래 음성이 왜곡된다. 이와 같이 왜곡된 음성 신호를 처리할 때에는 군중잡음을 제거하는 과정이 반드시 필요하다. 본 논문에서는 군중잡음의 특성을 분석하고, 그 결과를 이용하여 음성 신호처리 시에 효과적으로 군중잡음만을 제거할 수 있는 모델을 제안한다. 제안된 모델은 시간 영역에서는 침묵 구간을 검출하여 마찰음과 파열음을 제거하는 과정과 주파수 영역에서는 잡음 평균을 생성하고 이를 이용한 스펙트럼 차감법(spectral subtraction)으로 군중 잡음을 제거하는 과정으로 이루어진다.

1. 서론

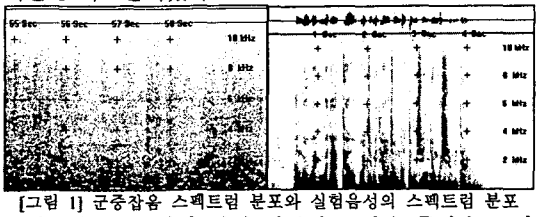
큰 쇼핑몰이나 백화점 또는 운동 경기장 같이 많은 사람이 모이는 곳에서 음성 통화 및 화자 인식을 할 때에는 군중잡음(crowd noise)이 음성 신호에 부가되어 신호가 왜곡된다. 이와 같이 왜곡된 음성 신호는 음성 통화나 화자 인식의 품질을 저하시키게 된다. 따라서 품질 향상을 위해서는 왜곡된 음성 신호에서 군중잡음을 제거하는 과정이 반드시 필요하다. 지금까지 음성 신호에 부가된 잡음을 제거하기 위한 연구는 특성이 뚜렷한 자동차 잡음에 대한 연구[1]와 갑자기 커다란 에너지를 포함하는 문닫는 소리 또는 재채기 소리와 같은 잡음을 다룬 연구[2]가 대부분이다. 그러나 군중잡음은 다른 잡음 보다는 비교적 일정한 패턴이나 주파수의 특성을 규명하기가 어렵다. 따라서 본 연구에서는 군중잡음이 발생하는 쇼핑몰과 경기장에서 군중들의 소음을 채집하여 군중잡음의 특성을 분석하고, 그 결과를 이용하여 음성 신호처리 시에 효과적으로 군중잡음만을 제거할 수 있는 모델을 제안한다. 효과적인 모델 설정을 위해 실제 군중이 모인 쇼핑몰과 경기장에서 군중들의 소음을 채집하고, 이것을 음성과 합성하여 잡음이 부가된 샘플을 작성하여 이용한다.

음성 신호는 기본적으로 유성음(voiced), 무성음(unvoiced) 및 침묵(silence) 구간으로 구성되어진다. 이중 실제 필요한 음성에너지는 유성 및 무성 구간이다. 전화 통화의 경우에는 침묵구간이 전체 통화 시간의 40%에서 50%를 차지하고 있다고 한다.[1] 본 연구에서는 군중잡음이 부가된 샘플에서 이와 같은 침묵 구간을 군중잡음 구간으로 규정하여 제거하며, 이것을 이용하여 음성 구간에 스펙트럼 차감법을 적용 한다. 스펙트럼 차감법은 비정체성(nonstationarity), 유색잡음(colored noise)의 제거 방법으로 널리 쓰이고 있는 방법이다.[3]

본 논문에서는 군중잡음의 특성에 기인하여 스펙트럼 차감법(spectral subtraction)을 사용한다.

2. 군중잡음 분석

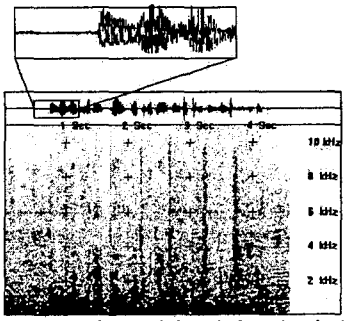
군중잡음이란 많은 수의 사람이 모여 발생하는 잡음으로서 일정한 패턴이나 주파수적 특성을 규정하긴 어렵다. 사람이 모인 수 혹은 주변 환경, 사람들의 연령, 성별, 주변상황 등에 따라 군중 잡음의 특성이 달라진다. 하지만 이러한 군중 잡음에도 일정한 에너지 분포를 찾을 수 있다. 일반적으로 사람이 음성을 마이크를 통하여 녹음할 경우 마이크에 전달되는 사람의 음성은 주위 잡음보다 높은 에너지를 갖게 된다. 물론 이때 주위의 잡음 또한 마이크를 통하여 녹음이 되지만, 일반적으로 가까운 거리에서 녹음된 음성이 더 큰 에너지를 갖게 된다. 또한 군중잡음은 넓은 주파수 대역에 에너지가 분포 되어 있다. [그림 1]은 군중잡음의 전형적인 모습을 보여준다. 다음 샘플은 분당에 위치한 할인마트에서 채집하였으며 채집시간은 약 1분이었다.



[그림 1] 군중잡음 스펙트럼 분포와 실험음성의 스펙트럼 분포

샘플에서 보는 바와 같이 에너지는 낮은 주파수로 가면서 점차 많이 지는 것을 알 수 있다. 이는 일반적인 비정체성 유색잡음과 비슷한 특성을 가진다. 하지만 문닫는 소리, 재채기 소리와 같은 순간적으로

발생 하는 잡음보다는 에너지가 적으며, 널리 분포되어 있고, 자동차 잡음과 비슷한 구조를 갖지만 비슷한 자동차 잡음 보다는 넓은 주파수 영역(frequency domain)에 걸쳐 에너지가 분포되어 있다. 자동차 잡음은 2kHz이하에 에너지가 집중되어 분포한다.[1] 또한 군중잡음은 여러 가지 잡음이 복합되어 있는 잡음이다. 군중잡음 안에는 위에서 언급한 문닫는 소리, 재채기 소리 등의 관련된 잡음, 혹은 여러 기구에서 발생하는 잡음 등과 같은 파열음이나 마찰음이 포함될 수 있다. 여기서 자동차 잡음과 구별할 수 있는 가장 큰 차이점을 찾을 수 있다. 자동차 잡음의 경우 자동차 안이라는 비교적 폐쇄된 공간이므로 엔진소리, 바람소리를 이외의 외부 잡음의 효과가 적기 때문이다. [그림 2]는 군중잡음이 첨가된 음성 샘플이다. 음성은 "한국 외국어 대학교 정보통신 연구실입니다."이다. 샘플은 SNR(Signal-to-Noise Ratio)을 조절하기 위하여 각각을 녹음 후 합성하였다.

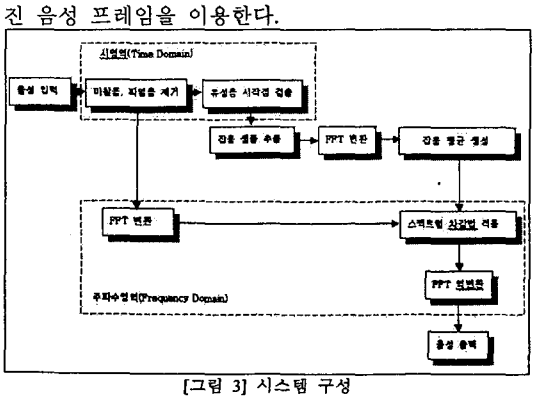


[그림 2] 군중잡음이 첨가된 음성 스펙트럼 분포

[그림 2]에서 보는 바와 같이 잡음구간과 음성구간은 확연히 차이가 난다. 음성구간은 에너지가 2kHz이하 부분에 많은 에너지가 분포되어 있는걸 볼 수 있다. [그림 2]에서는 시영역(time domain) 신호도 함께 보여주고 있는데, 시영역 신호 또한 그림에서와 같이 음성 신호 구간보다 작은 에너지를 가지고 있다.

3. 군중잡음 제거 모델

군중잡음 제거를 위한 시스템의 구성은 [그림 3]과 같다. 다음 시스템에서 사용하는 데이터는 512개의 샘플로 이루어진 음성 프레임을 이용한다.



[그림 3] 시스템 구성

3.1 시영역 및 주파수 영역의 동시 처리

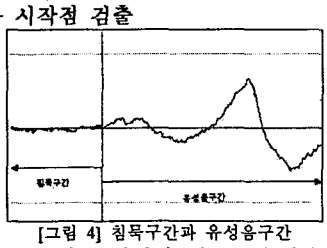
본 시스템은 다른 잡음제거 시스템과 달리 군중잡음에 특성화된 성격을 보여준다. 그러기 위해서 사용한 것이 시영역상에서의 마찰음, 파열음 제거와 주파수 영역에서의 스펙트럼 차감법을 동시에 처리하는 것이다. 일단 음성이 입력되고 시영역상에서 마찰음, 파열음이 제거되면 바로 주파수 영역에서 FFT(Fast Fourier Transform) 변환이 시작되며, 시영역 구간에선 유성음 시작점 검출이 동시에 이루어진다. 이 두개의 과정은 별개로 진행되어지다가 잡음 평균이 생성되는 순간 동기화 된다.

3.2 마찰음, 파열음 제거

군중잡음의 특성 중에 마찰음 및 파열음이 존재한다. 처음 잡음이 포함된 음성이 입력되었을 때 시영역 상에서 마찰음 및 파열음을 미리 제거한다. 마찰음, 파열음의 제거는 피치 변화량(PD: Pitch Different)을 계산하여 제거한다. 수식 (1)에서 T_1은 주기성을 알아보기 위한 자기상관함수를 구했을 때 시작점부터 첫 번째 피크 사이의 간격인 주기가 되고 T_2는 첫번째 피크부터 다음 피크 사이의 간격이 된다. PD가 5이하이면 유성음으로 간주하며[2], 마찰음, 파열음에 해당하는 피크 부분의 양 끝 0점 사이의 에너지는 0으로 바꾸어 삭제한다.

$$PD = |T_1 - T_2| \quad (1)$$

3.3 유성음 시작점 검출



[그림 4] 침묵구간과 유성음구간

[그림 4]는 시영역 상에서의 침묵구간(잡음)과 유성음 구간의 경계를 보여준다. 유성음구간을 검출하는 방법으로 크게 임계값을 사용하는 방법과 패턴인식 기법을 사용하는 방법으로 나뉘어진다.[1] 본 시스템에서는 임계값을 사용한다. 파열음, 마찰음의 비교적 잡음보다 큰 에너지를 갖는 신호는 이미 앞 과정에서 제거하였으므로 시영역 상의 피크를 검출하며 피크의 값이 항상 최대값을 갖도록 업데이트한다. 다음 피크값이 임계값보다 크게 변화하면 유성음구간으로 판단하며, 현재 피크값을 저장한다. 유성음구간의 시작은 임계값보다 큰 파형의 좌측 0점이다. 이 값은 시스템 종료시까지 유지하며 피크 임계값보다 작은 피크가 도착하면 다시 침묵구간으로 생각한다. 다시 침묵구간에 들어서면 처음과 같이 저장되어 있던 피크 최대값을 업데이트 해나간다. 초기 임계값은 음성입력 처음 1초의 피크높이 평균의 두 배로 한다. 실생활에서 초기 1초 동안의 입력은 잡음이라고 가정한다.[1] 실험에서의 임계값은 8bit, 22kHz 실험 샘플을 이용하여 실험 하였을 경우 파열음 및 마찰음이 제거된 잡음파형의 피크가 음성파형의 1/2를 넘지않았으므로 결정되었다.

3.4 침묵구간 추출 및 잡음 평균 생성

침묵구간, 즉 잡음구간을 검출하였다면, 침묵구간을

FFT변환하여 잡음 평균을 생성한다. 이를 이용하여 마찰음 파열음이 제거된 상태의 잡음이 포함된 음성을 FFT 변환한 것에 스펙트럼 차감법을 적용한다.

3.5 스펙트럼 차감법의 적용

본 시스템의 초기 모델은 유성음이 저주파 성분쪽에 집중되어 있음에 착안하여 로우패스필터(Low Pass Filter)를 사용하였으나 로우패스필터를 사용할 경우 유성음의 고주파 성분까지 삭제 되므로 음질의 향상을 기대할 수 없었다. 하지만 스펙트럼 차감법은 잡음의 효과를 최소화하는 방법으로 주변 잡음에 의해 손상된 음성 스펙트럼에서 잡음 스펙트럼의 크기성분만을 제거하는 방법이다.[3] 스펙트럼 차감법을 사용하면 유성음의 고주파성분이 모두 삭제되는 경우를 방지 할 수 있으므로, 로우패스필터를 사용하는 것보다 음질의 향상을 기대할 수 있다. 스펙트럼 차감법을 사용할 수 있는 조건은 첫째로 배경잡음의 스펙트럼 형태를 미리 알고 있거나, 잡음의 스펙트럼을 추정하기에 충분한 묵음 구간(약 300ms)가 주어져야 한다. 두번째로 배경잡음은 최소한 부분적으로 안정한 특성을 가져야 하며, 통계적특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 잡음만이 존재하는 구간을 검출 할 수 있어야 한다. 마찰음, 파열음이 제거된 군중잡음의 경우 잡음의 스펙트럼을 추정하기에 충분한 초기 잡음 입력 시간을 가지고 있으므로 첫번째 조건을 만족하며, 잡음구간 만을 검출 할 수 있는 모듈을 본 시스템은 가지고 있으므로 두번째 조건 또한 만족하므로 스펙트럼 차감법을 사용할 수 있다. 잡음신호($n(k)$)가 음성신호($x(k)$)에 더해졌을 때, 잡음과 음성이 합쳐진 음성신호는 수식 (2)와 같이 나타낼 수 있다고 가정한다.

$$x(k) = s(k) + n(k) \tag{2}$$

음성신호를 FFT하여 수식 (3)을 구한다.

$$X(e^{jw}) = S(e^{jw}) + N(e^{jw}) \tag{3}$$

전체 신호 $|X(e^{jw})|$에서 음성신호가 없는 침묵 구간의 데이터를 추출하여 잡음의 평균을 구한다.

$$\mu(w) = \frac{1}{M}\sum_{j=0}^{M} |N_j(w)| \tag{4}$$

전체 신호 $|X(e^{jw})|$에서 잡음평균을 차감하여 잡음이 제거된 음성 신호를 구한다.

$$\hat{S}(e^{jw}) = [|X(e^{jw})| - \mu(e^{jw})]\, e^{j\theta_x(e^{jw})} \tag{5}$$

3.6 FFT 역변환, 음성 출력

스펙트럼 차감법이 완료되면 이것을 FFT 역변환하여 음성을 출력한다. 음성 출력 후에는 바로 다음 프레임 처리를 시작한다.

4. 결론

본 논문에서는 군중잡음의 특성과 이를 처리하기 위한 시스템의 모델을 제안 하였다. 군중잡음은 마찰음, 파열음이 포함된 비정체성 유색 잡음이다. 군중잡음은 여러 가지 잡음요인이 복합된 구성을 가지고 있다. 전체적으로는 자동차 잡음과 비슷하게 저주파대에 비교적 많은 에너지를 가지고 있다. 하지만 마찰음, 파열음의 포함은 자동차 잡음과 구별되는 특성이다. 군중 잡음에서 마찰음, 파열음의 검출은 시영역상에서 자기상관함수를 이용한 피치 변화량을 계산하여 시영역상에서 에너지를 0으로 바꾸는 방법으로 제거 하였다. 본 실험 샘플에서 마찰음, 파열음이 제거된 후 시영역에서의 잡음 에너지는 유성음구간보다 1/2이하의 피크 높이를 가졌다. 여기서 피크 임계값을 이용하여 유성음 시작점, 즉 잡음 구간의 종료점을 검출 하였다. 잡음 구간이 정해졌다면, 이를 이용하여 잡음 평균을 구하였으며, 이는 음성신호에서 군중잡음을 제거하기 위한 스펙트럼 차감법에 사용되어진다. 또한 본 시스템은 시영역 처리와 주파수 영역 처리를 동시에 수행하며, 잡음 평균이 산출되면 이 두 과정은 동기화 되어진다. 잡음 평균의 산출과정과 마찰음, 파열음이 제거된 음성에 대한 FFT 과정을 같이 수행하므로 순차적으로 수행하는 것보다 잡음 제거에 걸리는 시간을 줄일 수 있는 이익을 가지고 있다. 앞으로 군중 잡음뿐 아니라 여러가지 다양한 잡음 요소가 포함된 복합잡음에 대해 강력한 제거 기능을 갖는 시스템의 개발이 연구과제라 하겠다.

5. 참고 문헌

[1] 구본웅, "음성신호에 중첩된 유색잡음 감쇠를 위한 음성검출기에 관한 연구", Inst. of Engin. Journal, 경기대학교. Vol. 10., 1994.

[2] 김치수, 배건성, " 비음성 식별을 이용한 잡음음성의 끝점검출 성능 개선" , 전자기술지, 20, 1, 1999.

[3] 오영환, " 음성언어 정보처리" , 홍릉출판사, ISBN 89-7283-056-9, 1998.

[4] Ann M. Rollins, "Speech Recognition and Manner of Speaking in Noise and in Quiet", CHI '85 PROCEEDING April., 1985.

[5] 윤태성, 심재성, "인간의 청각특성을 이용한 잡음혼입시의 음성인식에 관한 연구", 창원대학교 산업기술연구소 논문집 제6권., 1992.

[6] 조기량, 조의주, 박영창, "잡음 섞인 신호의 해석과 잡음제거", 여수대학교 논문집 제14권., 1999.

[7] 권영욱, 허만탁, "부가잡음 환경에서의 음성인식을 위한 전처리 방식 연구", 홍익대학교 논문집 제2권 2호.

[8]문현택, 이병수, "잡음환경에서의 음성인식", 순천향산업기술연구소 논문집 제3권 2호., 1997.

[9] 권영욱, 김형순, "잡음환경에서의 음성인식을 위한 전처리의 성능비교", 부산대학교 논문집 제37권.