MPSV 방법을 이용한 음성에서의 잔향 추출

김래훈, 성굉모
서울대학교 전기공학부

Extracting room reverberation from speech using the minimum phase space volume technique (MPSV)

Lae-Hoon Kim and Koeng-Mo Sung
Seoul National University

E-mail:laihong1@acoustics.snu.ac.kr

요약

음장의 공간 음향적인 특성에 영향을 받은 음성신호를 원래 신호로 복원하기 위해서 본 논문에서는 MPSV (Minimum Phase Space Volume) 방법을 도입한다. MPSV 방법은 신호를 복원하기 위해 원래 신호의 어떠한 사전 정보나 가정을 필요로 하지 않고 그 신호의 비선형적인 동적 특이성만을 이용하는 블라인드 디콘볼루션 (Blind deconvolution) 방법이다. 또한, 이 방법을 이용하여 원래 신호를 복원하는 동시에 음장의 충격응답과 같은 시스템 특성 까지도 유추가 가능 하다.

1. 서론

현재 음성 신호를 복원 하는 것은 잔류 잡음과 같은 본 신호에 연관성 없이 더해져 있는 랜덤 노이즈 같은 신호를 제거 하는 쪽으로 그 초점이 맞춰져 있다. 그러나, 근래 들어 핸즈 프리 시스템의 사용이 빈번해지고, 음성 인식 시스템이 발달함에 따라 음장의 잔향과 같은 콘볼루션날 노이즈의 제거가 중요한 문제로 부각되고 있다. 특히 이러한 잔향 제거의 경우 특별한 방법을 통해 시스템 특성을 먼저 파악하고 이를 이용해 인버스 필터링 하는 일반적인 방법이 적용되기 힘들기 때문에, 본 논문에서 소개하는 MPSV 방법과 같은 블라인드 디콘볼루션 방법이 이용되어야만 한다.

기존의 일반적으로 알려진 고차 통계수치(HOS : Higher-Order Statistics)를 이용하는 블라인드 디콘볼루션 방법 [5]의 경우에는 왜곡 되기 전 신호의 확률 분포에 대한 정보를 알아야만 가능한 반면, 이 MPSV 는 원 신호의 일반적으로 알려져 있는 비선형적인 동적 특이성만을 이용하여 간단한 구조로 인버스 필터링을 수행하는 장점을 가진다.

또한 이 방법을 이용하면 동시에 왜곡의 원인이 되는 시스템의 충격 응답의 예측이 가능하므로 또 다른 응용분야를 생각해 볼 수도 있다. 예를 들어 음장 공간의 음향 특성을 대변해 주는 충격응답 패턴을 현재 사용하고 있는 MLS 와 같은 듣기 싫은 소리대신에 음성 신호를 이용해 구할 수 있을 것이다. 이 경우 오히려 MLS 를 사용할 때 사람이 없는 상황에서 실험을 해야 하는 모순 같은 것을 해결 할 수 있다. 즉, 콘서트 홀의 경우라면, 관객의 수에 따라 잔향 패턴이 달라지는 것까지도 음성 혹은 음악 신호를 가지고 그 자리에서 파악 할 수 있는 장점을 가질 수 있는 것이다. 만약, 그 공간이 음성을 전용으로 사용하는 강의실이라면 오히려 음성에 대한 잔향 패턴을 파악하는 것이 그에 적합한 정보를 파악하게 되는 것이라는 장점을 갖기도 한다.

앞에서 전술한 바와 같이 음성 신호에 잔향이 부가 되어 있는 경우 주관심사가 음성 신호 복원인가 아니면 잔향 패턴 파악인가에 따라 그 응용이 달라 질 수 있다. 본 논문에서는 무향 녹음 되어 있는 음성 신호에 특성을 알고 있는 임의의 시스템을 통과 시키고, 이를 MPSV 방법을 이용하여 원래 신호 복원과 그 임의의 시스템을 추정해 내는 결과를 시험해 보았다. 또한 이를 임의의 잔향 응답에 콘볼루션 시키고 같은 실험을 행해 보았다.

2 장에서는 MPSV 방법이라는 것이 무엇인가 소개하고, 3 장에서는 이를 이용하여 인버스 필터링 하는 과정에 대하여 설명한 후 4 장에서 실제 음성 신호를 가지고 실험과 결과와 이에 따른 고찰을 서술 하였다. 마지막으로 5 장에서 본 논문의 결론과 향후 과제에 대해서 언급하였다.

2. MPSV 방법

이 부분에서는 본 논문에서 잔향 추출을 위해 사용된 MPSV 방법에 대해서 소개 하겠다 [1]. 이 방법은 음장 공간의 충격 응답과 같은 콘볼루션 되어있는 왜곡으로부터 원래 신호를 복원시키기 위해 원래 신호의 동적 특이성(specific dynamic property)을 이용하는 블라인드 디콘볼루션 방법이다. 즉 이 MPSV 방법을 이용하면 왜곡되어 녹음되어 있는 신호로부터 그것의 동적 특이성을 제외한 어떠한 사전 정보나 확률적인 가정이 없어도 왜곡되기 전 신호를 복원 할 수 있게 된다. MPSV 방법은 Phase Space Volume (PSV)을

최소화 시키는 방법이다. 여기서 ε-클리디안 공간 (Euclidean space) R^m에 숨겨져 있는 신호 $\{u_n\}$의 PSV 의 정의는 아래와 같다.

$$V_\varepsilon^m(u_n) = V_\varepsilon^m(A) = \inf\left\{\sum_{i=1}^{\infty}|U_i|^m\right\}$$

여기서 $A = \left\{\mathbf{u}_n = (u_n, u_{n+1}, \cdots, u_{n+m-1}) \middle| n = 1,2,\cdots,N\right\}$

이고 $\{U_i\}$는 A의 ε -cover이다.즉

$$A \subset \cup_{i=1}^{\infty} U_i \text{ with } 0 \langle |U_i| \le \varepsilon$$

$$\text{The diameter } |U_i| = \sup\left\{\|\mathbf{u}_{nj} - \mathbf{u}_{nk}\| : \mathbf{u}_{nj}, \mathbf{u}_{nk} \in U_i\right\}$$

예를 들어 우리가 본 논문에서 다루고자 하는 음성 신호와 같이 정해진 m 개의 시간 값 안에서 그 동적 특이성을 확인할 수 있는 일정한 Phase Space를 갖는 신호는 PSV 값 역시 정해진 경계 값 안쪽의 값을 갖게 된다. 그러나 일반적인 랜덤 노이즈와 같이 동적 특이성을 얻을 수 없는 경우에는 PSV 값이 무한대로 커지게 된다. 이로부터 다음과 같은 사실을 알 수 있다. "왜곡되어져 있는 신호의 PSV 을 최소화 하는 방향으로 인버스 필터링(inverse filtering)을 수행하여, 왜곡되기 전의 원래 신호의 PSV 값에 근사 시키면, 이로부터 원래 신호를 복원 할 수 있게 된다." 또한 "신호를 왜곡 시키는 음장 충격 응답과 같은 시스템의 충격 응답을 예측 할 수 있게 된다." 첫번째 표현은 원래 신호의 복원에 주안점을 두고 MPSV 방법을 서술한 것이라면, 두 번째 표현은 왜곡의 원인이 되는 시스템의 규명에 주안점을 둔 것이라고 생각해 볼 수 있다. 어디에 주안점을 두는 가에 따라서 그 응용 방법도 차이가 날 수 있다. 첫번째의 경우 음성신호의 복원을 통해 음성 인식률 향상에 적용 하는 예를 생각해 볼 수 있고, 두 번째의 경우 음성 신호를 통해 음장 공간의 잔향 패턴을 파악하고 더 나아가서는 잔향시간과 같은 그 공간의 음향 특성 지표를 얻어 내는 기법으로의 연구도 가능할 것이다.

PSV $V_\varepsilon^m(A)$ 을 계산하기 위해서는 위에서 정의 되어 있는 ε 값을 무한히 0 값에 근사시켜야 한다. 그러나 작은 ε 값을 사용할수록 계산량이 증가하게 되는 단점을 가지기 때문에 ε 값을 결정하는 일을 피하고 만족할 만한 결과를 얻기 위해서 본 논문에는 아래와 같이 $V_\varepsilon^m(A)$ 를 근사화 할 수 있는 알고리즘을 사용하였다.

1) 우선 왜곡된 신호를 그 동적 특성이 확인 될 수 있는 m 차원 Phase Space 로 그룹핑 한다. 즉 앞 정의의 $A = \left\{\mathbf{u}_n = (u_n, u_{n+1}, \cdots, u_{n+m-1}) \middle| n = 1,2,\cdots,N\right\}$ 를 수행하는 것이다. 물론, N+m-1 값이 입력 신호의 마지막 시간 값이 될 것이다. 음성 신호의 경우 그 동적 특성은 attractor dimension 값이 4 이상의 값에서 확인 되어 지며 [2] 이때 m 값은 Takens 이론에 따라 4×2+1=9 값으로 취할 수 있다 [3].

2) $\mathbf{u}_i$ 를 선택한다. A 안에서 $\mathbf{u}_j$ 값을 $\prod_{k=0}^{m-1}|u_{i+k} - u_{j+k}|$ 값을 최소화 할 수 있는 값으로 선택한다. 즉 $\mathbf{u}_j$ 는 $\mathbf{u}_i$의 가장 가까운 이웃 값이 되는 것이다.

3) 전체 PSV 값에 위에서 얻은 값을 추가한다. 그러나 전에 위의 값을 더해준 적이 있을 경우에는 추가하지 않는다. 이것은 $\mathbf{u}_j$가 $\mathbf{u}_i$의 가장 가까운 이웃이면서 동시에 $\mathbf{u}_i$ 가 $\mathbf{u}_j$ 의 가장 가까운 이웃이 되는 경우에 같은 PSV 가 전체 PSV 에 두 번 더해지는 것을 방지하기 위한 조건이다.

4) i 값이 N 값에 도달 할 때 까지 2),3)번 과정을 반복한다. 마지막으로 얻어지는 전체 PSV 값이 우리가 얻고자 하는 PSV 값의 근사값이 된다.

3. MPSV 방법을 이용한 inverse filtering

콘볼루션날 노이즈는 선형 또는 비선형 모델로 근사화가 가능하다. 본 논문에서 다루는 음장 충격 응답과 같은 시스템의 경우에는 선형 모델로 생각 할 수 있고, 이런 경우 본 논문에서는 적절한 차수 p 의 AR(autoregressive)로 모델링이 가능하다. AR 모델은 여러 분야에서 선형 시스템의 모델로 많이 사용되며, 또한 낮은 차수의 ARMA(autoregressive moving average)모델 까지도 이를 이용해 근사가 가능하다. 이 때, 왜곡되어 관찰되는 신호 x_n 를 다음과 같이 표현 할 수 있다면,

$$x_n = \sum_{i=1}^{p} a_i x_{n-1} + s_n$$

인버스 필터링 과정을 다음과 같이 하여 원래 신호에 대한 추정이 가능 해 진다 [4].

$$u_n = x_n - \sum_{i=1}^{q} b_i x_{n-1}$$

위 식에서 원래 신호를 추정하는 과정 또는 선형 시스템의 계수를 구하는 과정은 무작위로 얻어진 계수 값들 중 $V_\varepsilon^m(u_n)$ 을 최소화 하는 방법으로 이루어 진다. 이렇게 얻어진 u_n 값은 원래 신호 s_n 의 복원 신호이고, 각 계수 b_i 는 그에 관여하는 선형 시스템의 계수 a_i의 추정 값이 된다.

그러나 이러한 AR 모델링의 경우 최소 위상(minimum phase) 시스템의 경우에만 적용되어 진다. 따라서 일반적으로 최소 위상 조건을 만족 시키지 못하는 시스템(nonminimum phase system)인 음장 충격 응답의 경우에는 이러한 방법을 이용 하면 최소 위상 부분만 추출이 가능하게 된다. 그러나 음성의 경우 위상왜곡에 많은 영향을 받지 않는 다는 사실이 알려져 있고, 최소 위상 부분에 대부분의 음장 특성이 포함되어 있기 때문에 이렇게

얻어낸 원래 신호와 시스템 응답은 주목할 만한 의미를 갖는다고 할 수 있다.

4. 실험 및 토의

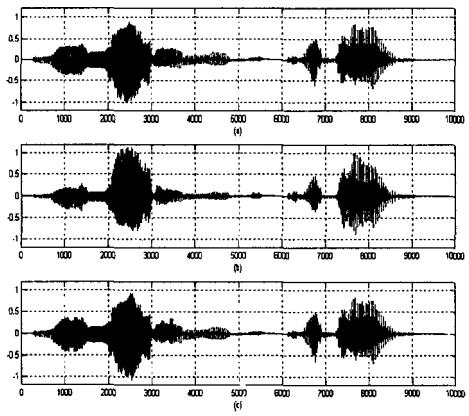
본 실험을 위해서는 무향 녹음되어있는 음성 신호를 이용하였다. 음성 신호는 다음과 같은 문장을 여성의 목소리를 이용하여 얻었다.

"내 놈이 무식하다."

샘플링레이트(sampling rate)는 8kHz 를 이용하였다. 또한 MPSV 방법을 이용하기 위해서는 원래 신호가 stationary 해야 한다는 조건이 필요하므로, 이를 만족시켜 주기 위해서 12.5ms씩 블록을 나누고 각각을 따로 취급하여 원래 신호를 복원하고 시스템 계수를 추출 하였다.

(a) 2 차 AR 시스템에 의한 왜곡 복원 실험

먼저 최소 위상을 갖도록 시스템의 계수를 각각 a_1=-0.1, a_2=-0.8 으로 설정하고 음성 신호를 이 시스템에 통과 시켰다. 이 왜곡된 신호를 가지고 MPSV 방법을 이용하여 원래 신호 복원과 시스템 계수를 구해 보았다. 충분히 다양한 b_1 , b_2 계수를 무작위로 추출한 후 이 중에서 u_n 의 PSV 를 최소로 하는 값을 선택하여 인버스 필터링을 수행 하였다. 이 결과를 그림 1 에 나타내었다.



`그림 1. 2 차 AR 시스템의 복원 실험

그림 1. (a)는 원 신호를, 그림 2. (b)는 AR 시스템 통과 후 왜곡된 신호를, 그리고 그림 3. (c)는 다시 이를 MPSV 방법을 이용해서 복원해낸 신호를 나타낸다. 예상했던 대로 훌륭하게 복원됨을 확인 할 수 있다. 이 신호를 들어봐도 귀로 원 신호와 거의 차이가 없음을 확인 할 수 있었다. 또한 추정해낸 시스템 계수 b_1 , b_2 의 경우도 90 여 개의 블록 중에서 75 개 이상의 블록에서 각각 -0.1, -0.8 로 가정했던 AR 시스템 계수를 훌륭하게 찾아 내고 있었다. 이중 다른 계수를 답한 부분은 묵음 구간으로서 음성의 동적 특성을 확인 할 수 없는 부분이고, 따라서 복원에 별 영향을 미치지 않음을 확인 할 수 있다. 즉, 원 신호 복원과 시스템 파악을 충실하게 해내고 있음을 확인 할 수 있었다.

(b) 3 차 AR 시스템에 의한 왜곡 복원 실험

이번에는 시스템의 계수를 각각 -0.1,-0.3,-0.8 로 설정하였다. 방법은 (a)의 경우와 같다. 결과를 그림 2 에 나타내었다. 이번 경우 눈에 보기에도 심각한 왜곡을 확인 할 수 있고, 귀로 들어도 왜곡의 정도가 상당함을 확인 할 수 있었다. 그러나 놀랍게도 실험 결과 복원 계수가 모든 프레임에서 실제 시스템 계수와 같게 구해졌다. 또한 복원된 신호는 원래 신호와 차이를 발견하기 힘들 정도였고, 들어서도 차이를 식별할 수 없었다.

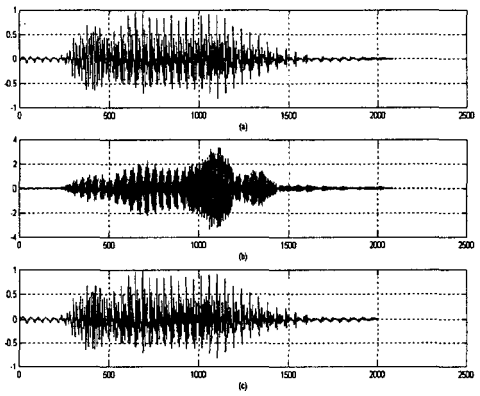


그림 2. 3 차 AR 시스템의 복원 실험

(c) 실제 잔향기를 통과 시킨 후의 음원을 가지고 복원 실험

잘 알려져 있는 음향 편집기인 CoolEdit 를 이용해서 0.5 초 정도의 잔향시간을 갖도록 잔향을 설정하여 잔향이 부가된 신호를 얻는다. 이렇게 왜곡된 신호를 3 장에서 언급한 인버스 필터링 과정을 거쳐서 신호를 복원 하였다. 10000 번의 무작위 계수 선택을 통해서 가장 작은 PSV 값을 갖도록 했다. 물론 적절한 차수의 선택이 중요한 문제였다. 여기서는 우선 가장 간단한 2 차 시스템이라고 가정하고 복원을 시도 하였고, 그 결과를 그림 3 에 나타내었다. 이 그림은 사용된 문장 중 '다' 부분에 해당된다.

그림에서 보이는 원래 신호와의 차이는 복원 신호에 시스템의 전체 통과(all pass) 부분은 제거 되지 않았기 때문이고 잔향 시스템을 간단한 2 차 AR 모델로 근사했기 때문이다.

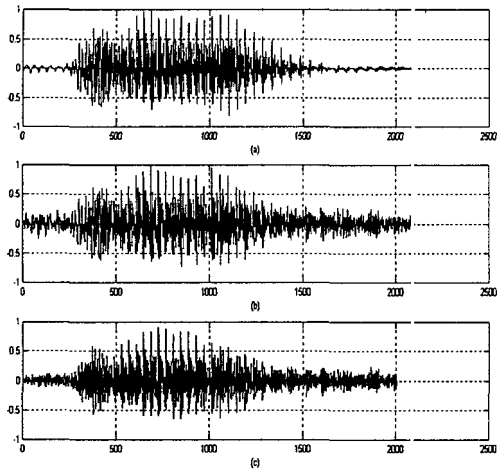


그림 3. 실제 잔향 신호의 복원 실험

5. 결론 및 향후과제

본 논문에서는 MPSV 방법을 이용해서 음장의 충격응답과 같은 콘볼루션날 노이즈가 첨가된 신호를 복원하고, 그 시스템을 추정해 내는 방법에 대해서 다루었다. 이 방법은 원 신호의 동적 특이성 외에는 다른 가정이나 정보를 필요로 하지 않기 때문에 그 응용범위가 다양 할 수 있을 것이다.

시스템의 차수를 적합하게 추정해 내는 부분에 대해서 좀 더 연구가 필요하지만, 작은 차수의 시스템을 가정해도 경향은 파악할 수 있는 결과를 얻을 수 있었다. 좋은 결과를 얻어 내기 위해서는 많은 수의 무작위 변수의 사용이 필수적이고 이는 많은 계산 시간의 원인이 되므로, 실시간 응용을 위해서는 앞으로는 최적화에 대한 연구가 수행되어야 할 것이다.

잔향 시스템의 패턴에 대한 실제 응용을 위해서는 음장 충격 응답 중 최소 위상 부분(minimum phase)부분이 가지고 있는 정보와 전체 응답과의 연관성을 면밀하게 분석 해야 하는 것 역시 앞으로의 향후 과제라 하겠다.

References

1. H. Leung and X. Huang, "Parameter estimation in chaotic noise," IEEE Trans. Signal Processing, vol. 44, pp. 2456-2463, Oct. 1996.
2. A. Kumar and S. Mullick, "Nonlinear dynamical analysis of speech," J. Acoust. Soc. Amer., vol. 100, pp. 615-629, July 1996.
3. F. Takens, "Detecting strange attractor in turbulence," in Dynamical Systems and Turbulence, Lecture Notes in Mathematics, D. A. Rands and L.S. Youngs, Eds. Berlin, Germany: Springer-Verlag, 1981, vol. 898, pp. 366-381
4. A. M. Chan and H. Leung, "Equalization of Speech and Audio Signals Using a Nonlinear Dynamical Approach," IEEE Trans. Speech and Audio Processing, vol. 7, pp. 356-360, May, 1999
5. J. A. Cadzow, "Blind deconvolution via cumulant extrema" IEEE Signal Processing Magazine, vol. 13, pp. 24-42, May, 1996