Neighborhood 관계를 이용한 DUET Generalization

*우성민, 정 홍
포항공과대학교 전자전기공학과
e-mail : innosm@postech.ac.kr, hjeong@postech.ac.kr

Generalization of DUET using neighborhood relationship

*Sung-Min Woo, Hong Jeong
Electrical and Electronic Engineering
Pohang University of Science and Technology

Abstract

In this paper, we propose a method that makes use of neighborhood relationship in 2D spectrogram of separated sources toward the generalization of the binary mask in Degenerate Unmixing Estimation Technique (DUET). A new generalized mask can be consist of five to ten mask. According to the new mask, the original power of the spectrogram in each frequency-time point is assigned. The result showed a smooth and tender wave-form, indicating a high speech separation performance compared to the original method.

I. 서론

원거리 음성인식의 배경잡음과 잔향문제를 해결하기 위해 다양한 암묵신호 분리방법(Blind Source Separation)들이 제시 되었다. 두 개 이상의 배열 마이크를 이용한 Independent Component Analysis(ICA), Degenerate Unmixing Estimation Technique(DUET) 등이 대표적이다[1][2]. DUET은 ICA를 이용한 암묵신호분리법과 달리 스테레오 마이크를 이용해서 두 개 이상의 원본 소스를 분리 해 내는 것이 가능하다. 이것은 ICA처럼 통계적 정보(Statistical Information)를 이용하지 않고 두 개 마이크의 간격을 샘플링 주파수(Sampling Frequency)보다 작게 하여 두 개의 스펙트로그램을 얻은 후, 같은 시간, 주파수 별 파워를 고려한 히스토그램을 그려 Clustering하는 방법이다. 본 논문에서는 binary masking을 이용한 원문과 달리 Belief Propagation을 이용하여 주변 시간, 주파수 슬롯의 정보를 함께 고려하여 일반화시키는 방법을 제안한다.

II. 본론

2.1 DUET

$s_0(t), s_1(t), \ldots, s_N(t)$와 같이 N개의 음성 신호mixing 모델을 다음과 같이 나타낼 수 있다.

$$x_1(t) = \sum_{i=0}^{N} s_i(t) + n_1(t)$$

$$x_2(t) = \sum_{i=0}^{N} \alpha_i s_i(t-\delta_i) + n_2(t)$$

$x_1(t)$와 $x_2(t)$는 각각 첫 번째 마이크와 두 번째 마이크에서 관측된 신호이다. $n_1(t)$와 $n_2(t)$는 채널에 삽입된 잡음이며 α_i는 relative attenuation, δ_i은 relative delay를 나타낸다. 여기서 입력신호가 두 개라고 가정하면 Hamminng Window를 씌워 Fourier Transform을 하여 다음과 같이 표현할 수 있다.

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_i e^{-jw\delta_i} \end{bmatrix} S_i(\tau,\omega)$$

따라서 relative attenuation과 relative delay를 다음과 같이 구할 수 있다.

$$(\alpha_i,\delta_i) = \left(\left| \frac{X_2(\tau,w)}{X_1(\tau,w)} \right|, Img\left(\log\left(\frac{X_1(\tau,w)}{X_2(\tau,w)} \right) \right) \Big/ w \right)$$

이렇게 얻어진 α_i와 δ_i에 따라 히스토그램을 그려 Clustering한 다음 binary masking을 적용하여 다음과 같이 특정 시간, 주파수의 j번째 신호 성분을 분리해 낼 수 있다.

$$M_j(\tau,w) = \begin{bmatrix} 1 & ,(\tau,w) \in \Lambda_j \\ 0 & ,otherwise \end{bmatrix}$$

마지막으로 Inverse FFT를 통하여 주파수 영역에서 시간영역으로 변환이 가능하다.

2.2 Belief Propagation

실제 음향 신호의 경우 DUET의 W-Disjoint Orthogonal 가정을 완전히 만족시키지 않기 때문에 binary mask를 쓰는 위의 방법은 신호의 겹침(Overlapping)이 일어나는 시간-주파수 슬롯에서 큰 오차를 발생시킨다[3]. 그림1과 2를 비교해보면 알 수 있듯이 DUET으로 분리된 음성 스펙트로그램은 Binary Mask로 인해 군데군데 구멍이 나 있는 형태를 취하고 있다.

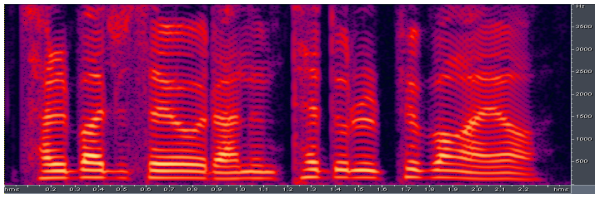

그림 1. 원본 음성 Spectrogram

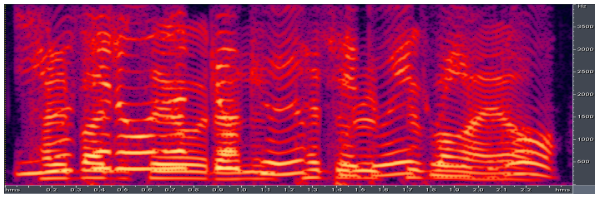

그림 2. 혼합된 음성 Spectrogram

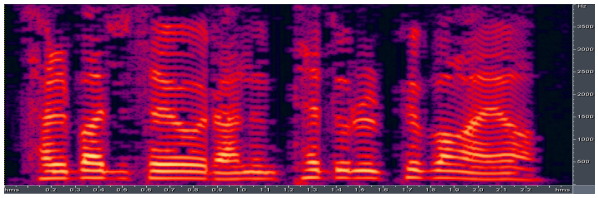

그림 3. DUET으로 분리된 음성 Spectrogram

이것을 보완하기 위한 방법으로 영상처리에서 쓰이는 Belief Propagation 알고리즘을 이용하여 이웃하는 시간-주파수 슬롯의 값을 고려하여 그 슬롯이 위치한 값을 보정하여 준다. 우리는 max-product probability distribution 모델 대신 min-sum negative log probabilities를 이용하였다. "t"반복후의 메시지 업데이트와 belief는 각각 다음과 같이 계산된다[].

$$m^{t_{p\to q}}(f_q) = \min_{f_p}\left(V(f_p - f_q) + D_p(f_p) + \sum_{s\in N(p)\setminus q} m^{t-1}_{s\to p}(f_p) \right).$$

$$b_q(f_q) = D_q(f_q) + \sum_{p\in N(q)} m^{T}_{p\to q}(f_q).$$

f_p는 "p" 상태(state)를 할당해주는 레이블링(labeling)을 의미하여 $V(f_p - f_q)$는 p상태와 q상태의 Compatibility, $D_p(f_p)$는 픽셀p에 상태p를 할당했을때의 Cost function이다. DUET에서 peak와 각각의 시간-주파수 슬롯상의 떨어진 거리에 따라 Cost function을 주고 한 번에 두 개 이상의 상태를 옮길 수 없도록 Compatibility를 조절하고 10개의 state를 써서 다음과 같이 Spectrogram을 얻었다. 그림2의 혼합된 음성에서 특별히 겹쳐지는 부분은 그림3에서 분리가 되어도 구멍이 난 것처럼 error가 생기는 데 이 부분들이 보정되어 부드럽게 보정이 되었다.

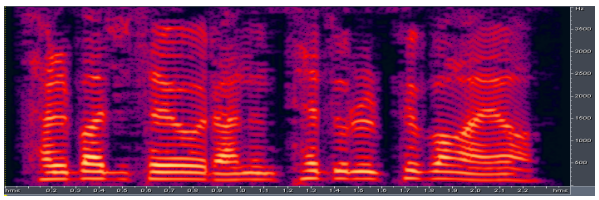

그림 4. Belief Propagation을 적용한 음성 Spectrogram

Ⅳ. 결론 및 향후 연구 방향

본 논문에서 Belief Propagation을 이용하여 DUET의 binary mask를 보완하여 좀 더 원본과 유사한 음성 스펙트로그램을 얻도록 분리 성능을 개선하였다.

참고문헌

[1] Bingham, E. Hyvarinen, A., ICA of complex valued signals: a fast and robust deflationary algorithm, Neural Networks, Proceedings of the IEEE conference, 2000, pp.357-362.

[2] Yilmaz, O. Rickard, S., Blind separation of speech mixtures via time-frequency masking, Acoustics, Speech, and signal Processing, IEEE Tranc. 52, 2004, pp.1830- 1847.

[3] Sungmin, W. Hong, J., Performance Evaluation of Blind Source Separation Schemes in Anechoic and Echoic Environments, Proceedings of the WSEAS, Signal Proc. 251-255.