동적 시간 왜곡(DTW)을 이용한 오토인코더 기반 청각치환 방법의 사용자 학습 가능성 검증

이수비*, 정치윤**, 문경덕**, 김채규***
*부경대학교 물리학과
**한국전자통신연구원 인공지능연구소
***부경대학교 IT융합응용공학부
e-mail : kyu0707@pknu.ac.kr

A Study on Applicability of Autoencoder-based Sensory Substitution Method using Dynamic Time Warping

Soo-Bee Lee*, Chi Yoon Jeong**, KyeongDeok Moon**, Chae-Kyu Kim***
*Dept of Physics, Pukyong National University
**Artificial Intelligence Research Laboratory, ETRI
***Dept of IT Convergence and Application Engineering, Pukyong National University

요 약

감각치환 기술은 손상되거나 상실된 감각을 신체의 다른 감각을 통하여 전달함으로써 사용자가 손실된 감각을 느낄 수 있도록 하는 기술이다. 그 중 신체에서 가장 많은 정보를 처리하는 시각이 손실된 사람들을 위하여 시각 정보를 청각 정보로 변환하여 전달하는 청각치환 기술에 대한 연구가 활발히 이루어지고 있다. 사용자가 청각치환 정보를 학습할 때 입력 시각 정보와 변환된 소리 정보간의 일관성이 필수적이지만, 오토인코더 기반의 청각치환 방법은 시각 정보와 변환된 소리 정보의 매핑이 블랙박스처럼 동작하여 사용자가 변환 규칙을 이해하기 어려울 수 있다. 따라서 본 논문에서는 동적 시간 왜곡을 이용하여 오토인코더 기반 청각치환 방법의 입력 정보 변화에 따른 생성 소리 신호의 유사성을 비교하여 사용자 학습 가능성을 분석하는 연구를 수행하였다. 또한 오토인코더 기반 청각치환 방법의 학습 가능성을 높이기 위한 향후 연구 방향을 제시하였다.

Keywords : Sensory substitution, Machine Learning, Deep Learning, Autoencoder

1. 서론

최근 HCI (Human-Computer Interaction) 분야에서는 사람이 다양한 형태의 입출력 정보와 상호작용하면서 실제처럼 느끼며 컴퓨터와 커뮤니케이션할 수 있는 가상현실 및 증강현실 기술에 대한 연구가 활발히 진행되고 있다. 또 한편으로는 사람의 뇌가소성 (Neuro-plasticity)을 이용하여 현실세계의 감각정보를 뇌로 전달함으로써 감각을 지각할 수 있게 하는 감각 대체(치환) 및 증강기술도 꾸준히 개발되고 있다.

감각치환 기술은 손상되거나 상실된 감각을 신체의 다른 감각을 통하여 전달함으로써 사용자가 손실된 감각을 느낄 수 있도록 하는 기술이다 [1]. 고령화 사회에 진입하고 장애인이 증가하면서 감각치환 기술에 대한 관심이 증가하고 있으며, 그 중 신체에서 가장 많은 정보를 처리하는 시각이 손실된 사람들을 위하여 시각 정보를 청각 정보로 변환하여 전달하는 청각치환 기술에 대한 연구가 활발히 이루어지고 있다.

청각치환 기술은 명시적 방법과 묵시적 방법으로 구분된다. 명시적 방법은 영상의 색상, 위치, 에지 등의 정보를 소리의 주파수, 크기, 리듬 등으로 변환하는 규칙을 정의한 후, 규칙에 따라서 영상 정보를 소리 정보로 변환하는 방법이다. 대표적인 방법으로 vOICe [2], Vibe [3]등의 방법이 있다. vOICe는 1990년대 피터 메이어르 박사에 의해서 제안된 감각치환 장치로써, 영상 정보를 소리로 변환하여 전달 할 수 있다는 가능성을 확인시켜주었다. vOICe에서는 영상의 가로 축은 소리의 발생 시간, 세로 축은 소리의 주파수, 밝기는 소리의 크기에 매핑하여 시각 정보를 소리 정보로 변환하였다. 사람의 눈은 짧은 순간 특정 영역에 집중하여 정보를 습득 한 후 다른 영역으로 이동하는 특징이 있다. Vibe에서는 이러한 특징을 사용하여 영상 곳곳에 존재하는 수용체 (RF: Receptive Field)를 정의한 후 수용체의 주변 정보를 사용하여 소리 신호를 생성

* 본 연구는 한국전자통신연구원(ETRI) 연구운영비지원사업의 일환으로 수행되었음[19ZS1500, 인간의 감각·지각 능력을 증강하는 다중 감각 융합 기술 개발 사업]

하였다. 수용체의 가로 축 위치는 스테레오 음향 생성을 위한 소리의 레벨 차이 정보로 사용되며, 세로 축 위치는 소리의 높이, 그리고 수용체 주변의 밝기 정보를 소리의 크기로 매핑하였다. 이와 같이 명시적 방법들은 시각 정보를 소리 정보로 변환하기 위한 명확한 규칙이 있어 사용자가 학습을 통하여 변환 규칙을 이해하고 소리 정보를 통하여 시각 정보를 이해할 수 있음이 증명되었다.

묵시적 방법은 전문가가 영상 정보의 소리 정보 변환 규칙을 사전에 정의하지 않고 기계가 학습을 통하여 자동으로 입력 영상에 맞는 최적의 소리 정보를 생성하는 방법이다. 최근 오토인코더 기반의 청각치환 기술 [4]이 제안되어 청각치환 기술의 새로운 가능성을 제시하고 있다.

명시적 방법의 청각치환 기술은 사용자가 학습을 통하여 시각 정보를 이해할 수 있다고 증명된 반면, 묵시적 방법의 경우 사용자의 학습 가능성에 대한 연구가 진행되지 않았다. 오토인코더 기반의 청각치환 방법은 블랙박스 형태로 동작하기 때문에 입력 영상의 작은 변화에 전혀 다른 소리가 생성될 수 있으며, 이는 사용자가 소리 정보를 통하여 시각 정보를 이해하기 어렵게 만들 수 있다. 따라서 본 논문에서는 오토인코더 기반 청각치환 방법의 사용자 학습 가능성을 분석하는 연구를 수행하였다.

본 논문의 구성은 다음과 같다. 2절에서는 오토인코더 기반 청각치환 방법에 대해서 소개하고, 3절에서는 동적 시간 왜곡 (DTW: Dynamic Time Warping)을 사용하여 사용자 학습 가능성을 분석한다. 마지막으로 4절에서는 결론 및 향후 연구 방향을 제시한다.

2. 오토인코더 기반 청각치환 방법

오토인코더는 대표적인 비지도 학습 신경망으로써 차원 축소를 통한 데이터셋의 특징 추출 및 재구성을 목표로 한다 [4]. 오토인코더는 인코더와 디코더로 구성되며 인코더에서는 중요 정보를 최대한 보존하면서 입력 정보의 차원을 축소시키고, 디코더에서는 저차원의 은닉층 정보를 활용하여 원본 데이터를 복원한다. 오토인코더는 입력 값과 최종 출력 값이 같아지는 것을 목표로 하며 역전파 알고리즘을 사용하여 재구성 에러를 최소화 하도록 학습한다.

VAE (Variational Autoencoder) [5]는 오토인코더의 구조를 가지지만 잠재변수가 확률분포로 표현되는 차이가 있다. 인코더에서 잠재변수 z가 확률분포를 따르도록 하고, 디코더는 확률분포 z로부터 원본데이터 x를 가장 잘 얻을 수 있도록 샘플링한다. VAE의 손실함수는 재구성 손실함수와 학습된 잠재변수의 분포와 기존 분포 사이의 쿨백 라이블러 발산 (Kullback-Leibler divergence) 함수의 합으로 구성된다. VAE는 일반 오토인코더보다 군집도가 높고 생성모델의 손실 값을 계산할 수 있다는 장점을 가지지만 결과물의 해상도가 낮다는 단점이 있다.

오토인코더 기반 청각치환 방법인 AEV2A [6]는 2019년 빅토르 토스에 의해서 제안되었다. AEV2A는 DRAW (Deep Recurrent Attentive Writer) 네트워크 [7]와 음성합성 모델로 구성된다. AEV2A는 DRAW의 기본 구조에 잠재변수로부터 음성 합성 모델을 사용하여 소리 신호를 생성하고, 청각 모델을 거쳐 청각적 특징을 추출하는 과정을 추가하였으며, 그 구조는 그림 1과 같다. DRAW는 VAE를 기반으로 사람의 시각 인지 메커니즘을 모방하여 순차적으로 영상을 생성하는 네트워크로써 공간적 어텐션 메커니즘과 LSTM (Long Short-Term Memory) 구조를 가지고 있다. DRAW에서는 동적 어텐션 메커니즘을 통해 각 타임스탬프 마다 어떤 부분을 집중해서 읽고 쓸 것인지 결정하며, LSTM 네트워크를 사용하여 순차적으로 어텐션 영역의 정보를 갱신하게 된다.

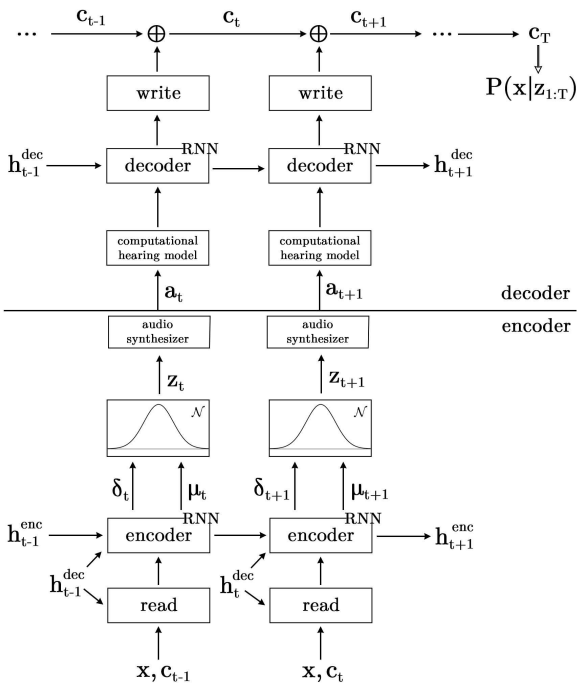

그림 1. AEV2A 구조 [6]

AEV2A에서의 각 구조의 기능과 데이터 처리 과정은 다음과 같다. 먼저 입력 영상 정보를 샘플링하는 Read 연산자는 이전 시간의 디코더 출력 값과 결과 값을 반영하여 현재 어텐션 영역의 크기와 위치를 결정하며, 생성된 어텐션 영역의 정보는 Write 연산자와 공유한다. 인코더에서는 Read 연산자가 결정한 어텐션 영역으로부터 정보를 추출하여 잠재변수 z로 정보를 축약한다. 어텐션은 N x N의 필터뱅크로 구성되며, 각 필터는 가우시안 분포를 따른다. 음성합성 모델은 잠재변수 z를 사용하여 소리 신호를 생성하며, 청각 모델은 음성합성 모델이 생성한 소리 신호로부터 청각 특징을 추출한다. 이 후 디코더에서는 청각 모델에서 추출한 청각 특징과 이전 시간에서의 디코더 출력 값을 사용하여 원본 영상을 재현하며, Write 연산자는 어텐션 영역에 그림을 그리는 기능을 수행한다.

AEV2A의 학습을 위한 손실 함수는 기존 VAE의 손실

함수와 일치비용으로 구성된다. 일치비용은 발생된 소리의 높이와 영상의 수직위치 간의 거리, 소리의 공간적 위치와 영상의 수평위치 간의 거리, 그리고 소리의 진폭과 영상의 휘도 간의 차이로 이루어져 있다.

3. DTW를 이용한 사용자 학습 가능성 분석

오토인코더 기반 청각치환 방법은 시각 정보를 분석하여 데이터를 기반으로 최적의 소리 신호를 생성한다. 따라서 입력 시각 정보의 변화에 따른 생성 소리 신호를 예측할 수 없기 때문에 사용자의 학습 가능 여부를 판단하기에 어려움이 있다. 따라서 본 논문에서는 규칙성을 가지고 변화하는 시각 정보와 변환된 소리들 간의 유사도를 분석함으로써 소리 신호 변화의 일관성을 분석하고 사용자의 학습 가능성 여부를 분석하였다.

오토인코더의 기반의 청각치환 모델은 기존 논문에서 제시된 최적의 파라미터를 사용하여 학습하였으며, 학습 데이터 역시 기존 논문에서 사용한 “simple_hand” 데이터셋을 사용하였다. 테스트 데이터는 기존 데이터셋의 테스트 데이터 중 손 모양이 다른 4장의 영상을 그림 2와 같이 선정하였다. 원본 테스트 영상의 경우 손목과 팔 부분을 포함하고 있지만, 영상의 위치 이동을 적용할 때 영상의 형태가 달라 질 수 있기 때문에 손 모양만을 크롭하여 사용하였다.

그림 2. 시각 정보 변환 위한 테스트 데이터

테스트 영상에 수평 이동, 수직 이동, 회전 등의 영상 변화를 그림 3과 같이 적용하였다. 수평 이동의 경우 X축 방향으로 -20 픽셀에서 2픽셀 단위로 20 픽셀까지 이동하였으며, 수직 이동의 경우 Y축 방향으로 -20 픽셀에서 2픽셀 단위로 20 픽셀까지 이동하였다. 영상 회전의 경우 시계방향으로 -20도에서 2도 단위로 20도까지 회전을 적용하였다.

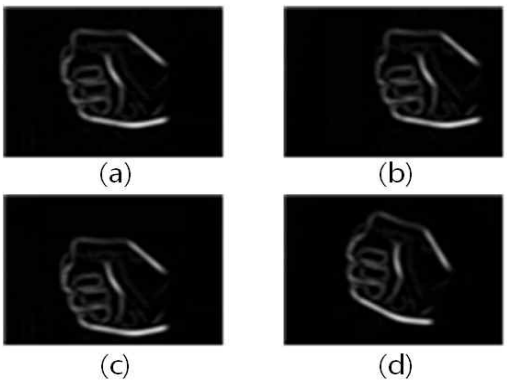

그림 3. 시각 정보의 변환 예 (a) 입력 영상 (b) 수평 이동 (c) 수직 이동 (d) 회전

사람의 청각은 소리의 주파수 변화에 로그스케일로 반응하며, 이러한 특성을 활용한 대표적인 소리 신호 특징으로는 MFCC (Mel Frequency Cepstral Coefficient)가 있다. 따라서 본 논문에서는 소리 신호로부터 MFCC를 추출하여 유사도를 비교하였다. 소리 신호와 같은 시계열 데이터의 경우 유클리디언 거리를 사용하면 동일한 신호의 형태를 가지더라도 시간 축이 달라지는 경우 유사도가 낮아지게 된다. 따라서 신호의 시간 정보를 배제하고 유사도를 분석하기 위하여 본 논문에서는 DTW [8]기반의 거리를 유사도 기준으로 사용하였다.

테스트 영상에 수평 이동, 수직 이동, 회전 등의 영상 변화를 적용하였을 때 변환된 소리 신호에 대한 MFCC 특징 기반의 유사도는 그림 5, 그림 6, 그림 7에 각각 표시하였다. 그림에서 유사도 (Similarity)는 영상 변화가 적용된 영상과 원본 영상의 변환된 소리신호에 대한 DTW 기반 거리를 의미하며, 영상 간 유사도 (Pair similarity)는 영상 변화가 적용된 현재 영상과 이전 영상의 변환된 소리 신호에 대한 DTW 기반 거리를 의미한다. DTW 기반 거리는 신호 간 차이를 나타낸 지표이므로 DTW 기반 거리가 클수록 유사하지 않음을 의미하며, 이는 소리 신호의 형태가 다르다는 것을 의미한다.

사람은 영상의 작은 변화에 강건하게 반응하기 때문에 오토인코더 기반 청각치환 방법도 영상의 작은 변화에 대해서 유사한 소리를 생성해야 사용자의 학습 가능성이 높다고 생각할 수 있다. 따라서 오토인코더 기반 청각치환 방법의 생성한 소리 신호가 영상의 변화가 클수록 DTW 기반 거리가 증가하고 영상 간 유사도는 비슷한 값을 가질 때 사용자의 학습 가능성이 높다고 판단할 수 있다.

영상 변화를 적용했을 때 유사도의 변화를 살펴보면 2픽셀의 수직 및 수평 이동, 2도의 회전만 적용하여도 DTW 기반 거리는 급격하게 증가함을 확인할 수 있다. 이는 영상의 작은 변화에도 오토인코더 기반의 청각치환 방법이 생성하는 소리가 급격하게 달라진다는 것을 의미한다. 또한 영상 변화에 따른 유사도의 분포가 일관되게 증가하지 않고 증가와 감소가 반복됨을 확인 할 수 있다. 영상 간 유사도의 경우에도 일정한 차이를 보여주지 않고 유사도 값의 변화가 큰 것을 확인 할 수 있다.

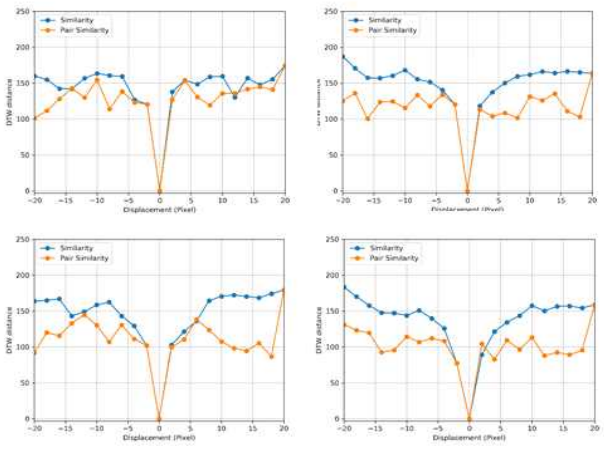

그림 4. 테스트 영상의 수평 이동에 따른 소리 신호의 유사도 비교

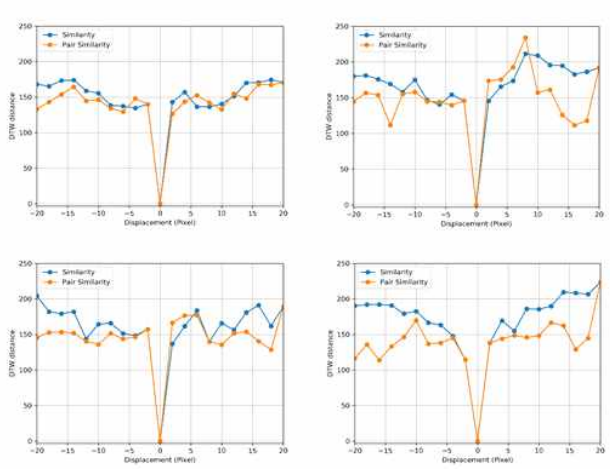

그림 5. 테스트 영상의 수직 이동에 따른 소리 신호의 유사도 비교

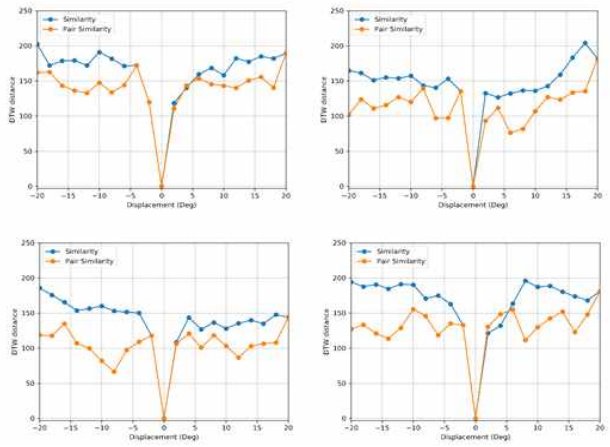

그림 6. 테스트 영상의 회전에 따른 소리 신호의 유사도 비교

테스트 영상에 대한 수평 및 수직 이동의 경우 원본 영상의 손 모양이 변화하지 않지만 영상 회전을 적용하는 경우 손 모양이 변화하게 된다. 실험 결과를 보면 손 모양이 변화하는 영상 회전에 대한 유사도 변화와 그렇지 않은 수직 및 수평 이동에 대한 유사도 변화가 같은 경향성을 보여주는 것을 확인 할 수 있다. 이는 영상 변화에 따라서 생성되는 소리가 일관성 있게 변화하지 않는다는 것을 의미하여, 오토인코더 기반 청각치환 방법의 사용자 학습 가능성은 낮을 것이라고 판단할 수 있다.

본 실험을 통해 오토인코더 기반 청각치환 방법은 사용자가 시각 정보와 변환된 소리 정보를 학습하기에 한계가 있음을 확인하였다. 이는 현재 오토인코더 기반 청각치환 방법이 흑백의 에지 정보만을 사용하여 네트워크에 충분한 정보가 전달되지 않기 때문으로 판단된다. 따라서 사용자 학습 가능성을 높이기 위해서는 색상, 형태 등의 모든 정보를 포함하는 가공되지 않은 영상 정보를 입력으로 사용하여 네트워크를 학습시키는 것이 필요하다. 또한 현재 오토인코더 기반의 청각치환 방법의 손실 함수는 입력 시각 정보와 변환된 소리 정보와의 일관성을 고려하지 않고 있기 때문에 입력 시각 정보와 변환된 소리 정보의 일관성을 고려할 수 있는 손실 함수를 설계하면 사용자의 학습 가능성이 높아질 것으로 판단된다. 이를 위하여 최근 다양한 분야에 적용되어 높은 성능을 보여주고 있는 GAN (Generative Adversarial Network) [8]이라는 새로운 네트워크 모델을 쓰는 것도 방법이 될 것이다. GAN은 가상의 영상을 만드는 생성 네트워크와 영상이 가상으로 생성되었는지 여부를 판단하는 판별 네트워크로 구성된다. 생성 네트워크와 판별 네트워크는 서로 경쟁하면서 성능이 향상되며 생성 네트워크에서 만든 영상이 판별 네트워크에서 실제 영상으로 판단될 때 까지 발전시킨다. GAN을 이용한다면 기존의 방법보다 생성 영상이 원본 영상과 더 유사해질 수 있기 때문에 생성되는 소리 신호 역시 일관성을 가질 수 있을 것으로 예상된다.

4. 결론 및 향후 연구

본 논문에서는 오토인코더 기반의 청각치환 방법의 사용자 학습 가능성을 분석하기 위하여 테스트 영상의 변화에 따른 생성 소리 신호의 유사성을 DTW를 사용하여 비교하였다. 영상 변화는 수평 이동, 수직 이동, 회전 등을 적용하였으며 유사도 및 영상간 유사도를 분석하여 영상의 작은 이동과 회전이 생성된 소리의 큰 변화를 야기한다는 것을 확인하였다. 이는 영상의 변화에 따른 생성소리의 일관성이 낮다는 것을 의미하므로, 사용자에게 오토인코더 기반의 청각치환 방법을 사용하여 시각 정보와 변환된 소리 정보를 학습시키기에는 어려움이 있을 것으로 예상된다. 오토인코더 기반 청각치환 방법의 사용자 학습 가능성을 높이기 위해서는 단순한 에지 정보보다는 많은 정보를 가진 원본 영상을 입력으로 사용하여 학습할 수 있는 프레임워크에 대한 연구가 필요할 것으로 생각된다. 또한 입력 시각 정보와 변환된 소리 정보의 일관성을 고려할 수 있는 손실 함수에 대한 연구가 필요하며, GAN을 적용하는 경우 이를 해결 할 수 있을 것으로 예상된다.

참고문헌

[1] 문경덕 외 7인, “감각치환 기술 동향,” 전자통신동향분석, 제34권, 4호, pp. 65-75, 2019.

[2] P. Meijer, “An experimental system for auditory image representations,” *IEEE Transactions on Biomedical Engineering*. Vol. 39, pp. 112-121, 1992.

[3] S. Hanneton, M. Auvray, and B. Durette. “The Vibe: A Versatile Vision-to-Audition Sensory Substitution Device,” *Applied Bionics and Biomechanics*, Vol. 7, pp. 269-276, 2010.

[4] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096-1103, 2008.

[5] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” arXiv:1312.6114, 2013.
[6] V. Tóth and L. Parkkonen, “Autoencoding sensory substitution,” arXiv:1907.06286, 2019.
[7] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, “DRAW: A Recurrent Neural Network For Image Generation,” *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1462-1471, 2015.
[8] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 359-370, 1994
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al, “Generative adversarial nets,” *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS),* pp. 2672-2680, 2014.