STOI 메트릭 기반의 다중 레벨 음성 향상 및 데이터 증강 기법

신도경†, 김영대

LIG 넥스원

dokyung.shin@lignex1.com†, youngdae.kim@lignex1.com

STOI Metric-Based Multi-Level Speech Enhancement and Data Augmentation

Do Kyuing Shin†, Young Dae Kim

LIG Nex1

요 약

최근 음성 신호 처리 분야에서 데이터 부족 문제 해결을 위한 음성 데이터 증강 기술이 주목받고 있다. 기존 GAN 기반 음성 향상 기법은 전통적인 손실 함수를 사용하여 인간의 지각적 품질과의 상관성이 낮고, 고품질 데이터에만 초점을 맞춰 일반화 능력이 부족하다는 한계를 갖는다. 본 논문에서는 STOI(Short-Time Objective Intelligibility) 메트릭을 활용한 다중 레벨 음성 향상 및 데이터 증강 모델을 제안한다. 제안된 모델은 MetricGAN 기반으로 5개의 품질 레벨(STOI: 0.55~0.95)에서 순차 학습을 수행하며, 생성자는 L1 손실과 메트릭 손실을 결합한 손실 함수로, 판별자는 STOI 점수를 예측하는 회귀 모델로 동작한다. Emo 데이터셋 실험 결과, 각 품질 레벨별로 목표 STOI에 근사한 음성 생성이 가능함을 확인하였으며, 스펙트로그램 분석을 통해 고품질 레벨에서 원본과 유사한 패턴 생성을 입증하였다. 제안된 방법은 다양한 품질의 데이터를 의도적으로 생성하여 모델의 견고성과 일반화 능력을 향상시키는 새로운 패러다임을 제시한다.

Ⅰ. 서 론

최근 음성 데이터 증강(Speech Data Augmentation)은 음성인식, 화자인식, 음성 합성 등 다양한 음성 신호 처리 분야에서 데이터 부족 문제를 해결하기 위한 핵심 기술로 주목받고 있다. 음성 데이터 증강 연구는 크게 신호 변형 기반(signal-level based), 특징 변환 기반(feature-level based), 생성 모델 기반(generative model-based)의 세 가지 접근 방식으로 분류된다. 생성적 적대 신경망(GAN, Generative Adversarial Networks)은 오디오 신호 처리 분야에서 광범위하게 활용되고 있으며, 특히 음성 신호 처리에서 상당한 성과를 달성하고 있다. 그러나 기존의 GAN 기반 음성 향상 기법들은 평균 제곱 오차(MSE, Mean Squared Error), STFT 차이, SNR과 같은 전통적인 신호 기반 손실 함수를 사용함으로써 실제 인간의 지각적 품질(perceptual quality)과의 상관성이 낮다는 근본적인 한계를 갖는다. 이러한 한계점을 해결하기 위해 Fu 등[1]은 MetricGAN을 제안하였다. MetricGAN은 음성 처리를 위한 새로운 GAN 기반 패러다임으로, 지각적으로 의미 있는 음성 품질 메트릭을 직접 최적화하는 방법을 적용한다. 기존 GAN의 판별자가 실제 샘플과 생성된 샘플을 구분하는 이진 분류를 수행하는 것과 달리, MetricGAN은 PESQ나 STOI와 같은 비미분 가능한 품질 지표를 근사하는 대리 모델(surrogate model)을 학습하여 생성자 훈련에 활용한다. 본 논문에서는 MetricGAN의 핵심 메커니즘을 기반으로 하여 품질 메트릭을 직접 손실 함수로 활용하는 음성 데이터 증강 모델을 제안한다. 품질 메트릭으로는 STOI(Short-Time Objective Intelligibility)를 사용하였으며, 이는 단시간 단위에서 깨끗한 음성과 향상된 음성 간의 상관관계를 통해 음성의 명료도를 측정하는 지표이다. STOI 점수는 0과 1 사이의 값을 가지며, 높을수록 더 나은 품질을 의미한다. 본 연구의 핵심 목표는 품질 메트릭을 활용하여 5단계 품질별 데이터 증강을 통해 다양한 품질의 데이터를 생성하는 것이다. 일반적으로 고품질 데이터만으로 학습된 모델은 과적합(overfitting)에 취약하며, 실제 환경의 다양성을 충분히 반영하지 못한다. 기존 연구들이 주로 잡음 제거 및 품질 향상에 초점을 맞춘 반면, 본 연구에서는 의도적으로 다양한 품질의 데이터를 생성하여 모델의 견고성과 일반화 능력을 향상시키는 새로운 패러다임을 제안한다.

Ⅱ. 제안한 방법

본 논문에서 제안하는 다중 레벨 품질 기반 오디오 데이터 증강 모델은 생성된 오디오의 품질 레벨에 따라 음성 향상 모델을 단계적으로 훈련하는 프레임워크를 채택한다. 학습 과정은 Algorithm 1에 제시된 바와 같이, N개의 서로 다른 STOI(Short-Time Objective Intelligibility) 목표값에 대해 순차적으로 모델을 훈련한다. 본 실험에서는 표 1과 같이 5개의 품질 레벨을 설정하였으며, 각 레벨의 STOI 목표값은 0.55, 0.65, 0.75, 0.85, 0.95로 구성된다. 제안된 모델의 생성자(Generator)는 잡음이 포함된 음성 신호를 입력으로 받아 향상된 음성을 생성한다. 생성자의 손실 함수는 수식 (1)과 같이 L_1 손실과 메트릭 손실의 결합으로 구성된다.

$$L_G = L_1\_loss + \lambda \cdot MSE(stoi\_pred, target\_stoi) \quad (1)$$

수식 (1)에서 L_1 손실은 생성된 음성과 실제 깨끗한 음성 간의 차이를 최소화하며, 메트릭 손실은 판별자가 예측한 STOI 점수와 현재 레벨의 목표 STOI 값 간의 평균제곱오차(MSE)를 나타낸다. 이러한 결합 손실 함수를 통해 생성자는 단순한 음성 복원을 넘어서 특정 품질 수준에 맞는 음성을 생성하도록 학습된다. 제안 모델의 판별자(Discriminator)는 기존 GAN의 이진 분류 방식과 달리 STOI 점수를 예측하는 회귀 모델로 동작한다. 판별자는 입력된 음성 신호에 대해 STOI를 예측하며, 다음 손실 함수를 통해 훈련된다: 수식 (2)에서 $D(x)$는 판별자의 예측값이고, stoi_real은 깨끗한 음성과 향상된 음성 간에 계산된 실제 STOI 값이다. 이를 통해 판별자는 음성의 객관적 품질을 정확하게 평가할 수 있는 능력을 획득하며, 동시에 생성자에게 품질 개선을 위한 지도 신호를 제공한다.

표 1 Emo 데이터 셋의 각 클래스 별 5단계 품질 레벨에 따른 목표 STOI 대비 증강 데이터의 평균 STOI 비교

Lv.	Target STOI	Aug. STOI	STOI of class-specific augmented data						
			A	E	F	L	N	T	W
1	**0.55**	0.65	0.67	0.67	0.62	0.66	0.76	0.57	0.60
2	**0.65**	0.70	0.74	0.74	0.64	0.74	0.80	0.60	0.63
3	**0.75**	0.75	0.77	0.79	0.75	0.76	0.82	0.63	0.70
4	**0.85**	0.80	0.80	0.81	0.83	0.80	0.84	0.70	0.83
5	**0.95**	0.86	0.88	0.87	0.88	0.85	0.89	0.75	0.89

Algorithm 1. Multi-Level MetricGAN Training

```
For each target_STOI in [0.30, 0.46, 0.63, 0.79, 0.95]:
    Initialize G, D networks
    While not converged:
        # Generator training
        x_enhanced = G(x_noisy)
        stoi_pred = D(x_enhanced)
        stoi_real = STOI(x_clean, x_enhanced)
        L_G = L1_loss + λ * MSE(stoi_pred, target_STOI)

        # Discriminator training
        L_D = MSE(D(x), stoi_real)

        Update G, D networks
    Save model for current quality level
```

$$L_D = MSE(D(x), stoi\_real) \quad (2)$$

두 네트워크는 상호보완적 학습을 통해 성능을 향상시킨다. 생성자는 판별자의 피드백을 활용하여 목표 품질 수준에 부합하는 음성 생성 능력을 학습하고, 판별자는 생성된 오디오의 품질을 평가함으로써 보다 정확하고 다양한 레벨의 품질 예측 능력을 발전시킨다. 각 품질 레벨에서 목표 품질에 수렴할 때까지 이 과정을 반복하여, 최종적으로 다양한 품질 수준의 음성 데이터를 생성할 수 있는 모델을 구축한다.

Ⅲ. 실험 결과

본 연구에서는 음성 발화 증강 실험을 위해 Emo 데이터셋을 사용하였다. 해당 데이터셋은 불안(A), 역겨움(E), 행복(F), 지루함(L), 중립(N), 슬픔(T), 분노(W)의 7가지 감정 클래스로 구성되어 있으며, 총 353개의 발화를 포함한다. 모든 입력 오디오는 16kHz 샘플링 레이트로 표준화하고 3초 단위로 분할하여 학습에 사용하였다. 증강된 데이터 또한 3초 길이로 생성하였으며, 각 원본 클래스 데이터 개수의 3배수로 증강을 수행하였다. 표 1은 각 클래스별 5개 레벨에 대한 증강 데이터의 STOI 평균값을 비교한 결과를 보여준다. 목표 STOI 품질 대비 증강된 데이터의 평균 STOI를 분석한 결과, 다음과 같은 특성을 확인하였다. 레벨 1과 레벨 2에서는 목표값 대비 과향상(over-enhancement) 경향이 나타났으며, 이는 저품질 목표에서 모델이 보다 적극적인 품질 개선을 수행하기 때문으로 분석된다. 레벨 3에서는 목표값과 증강 결과가 정확히 일치하여 가장 안정적인 수렴을 보였다. 반면 레벨 4와 레벨 5에서는 목표 대비 미달성(under-achievement) 결과를 나타냈으며, 이는 고품질 목표값에 대한 학습의 한계를 시사한다. 이러한 결과를 바탕으로 향후 고품질 구간에서의 성능 개선을 위해 적응적 손실 가중치 조정 및 계층적 학습 전략의 도입이 필요할 것으로 판단된다. 그림 1은 원본 발화 오디오와 5개 품질 레벨로 생성된 오디오의 스펙트로그램을 비교한 결과이다. 그림 1(a)는 원본 오디오, 그림 1(b)는 잡음 데이터를 나타낸다. 그림 1(c)-(g)는 각각 레벨 1부터 레벨 5까지의 증강 데이터 결과를 보여준다. 스펙트로그램 분석 결과, 고품질 레벨로 갈수록 증강된 데이터가 원본 데이터의 스펙트럼 패턴과 유사하게 생성되는 것을 확인할 수 있다. 이는 제안된 모델이 목표 품질 수준에 따라 적절한 품질의 음성 데이터를 생성할 수 있음을 시각적으로 입증한다.

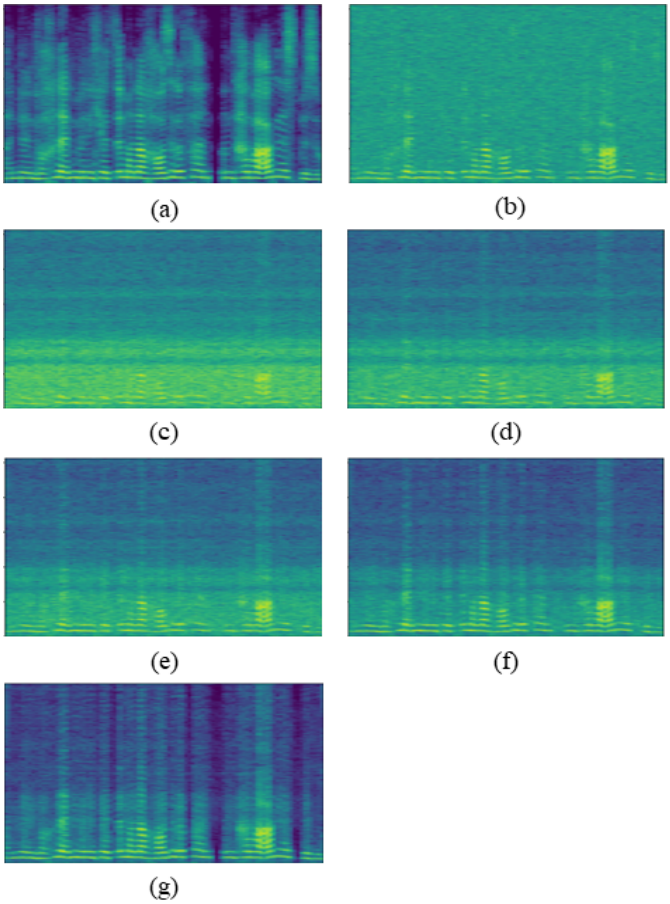

그림 1 'A 클래스의 원본 음성과 다중 레벨 품질로 생성된 증강 데이터의 스펙트로그램 비교: (a) 원본 음성, (b) 잡음 음성, (c) 레벨 1 (STOI=0.538), (d) 레벨 2 (STOI=0.642), (e) 레벨 3 (STOI=0.751), (f) 레벨 4 (STOI=0.839), (g) 레벨 5 (STOI=0.890)

참 고 문 헌

[1] Fu S. W., Liao C. F., Tsao Y., Lin S. D.,"MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," International Conference on Machine Learning (ICML), arXiv:1905.04874, 2019.