Attention 기반 Cross-Modal 통합 방법을 적용한 Lipreading 활용 무음 영상 음성 생성모델 연구

이건우 [1], 한요섭 [1,2]

[1] 숭실대학교 정보통신공학과
[2] 숭실대학교 지능형반도체학과

gunwoo@soongsil.ac.kr, yoseob.han@ssu.ac.kr

Silent Video Speech Generation Using Lipreading with Attention-Based Cross-Modal Integration

Gunwoo Lee[1], Yoseob Han[1,2]

[1]Department of Information and Telecommunication Engineering, Soongsil University
Department of Intelligence Semiconductors, Soongsil University

요 약

본 논문은 무음 영상으로부터 음성을 생성하는 과정에서, 입력 영상에서 얻을 수 있는 다양한 정보를 어텐션 메커니즘(Attention Mechanism)을 활용해 인공신경망으로 효과적으로 통합하는 방법의 우수성을 실험적으로 검증한다. 특히, 비디오 프레임 정보와 립 리딩(Lipreading)을 통해 얻은 텍스트 정보를 어텐션 메커니즘으로 통합하여 음성을 생성하였으며, 이를 통해 기존 모델 대비 음성의 자연스러움과 텍스트 정확도 측면에서 더욱 향상된 성능을 확인하였다

Ⅰ. 서 론

최근 Video to Speech 분야는 소리가 없는 비디오로부터 화자가 어떤 말을 하고 있는지를 예측하여 음성을 생성하는 기술로 많은 주목을 받고 있다. 특히, 사람의 입 모양, 얼굴 표정, 움직임과 같은 시각적 자료를 활용하여 실제로 들리지 않는 음성을 복원하는 것을 목표로 한다. 이러한 Video to Speech 분야는 영상 콘텐츠의 후처리, 영화 더빙, 실시간 영상 통화, 그리고 청각 장애인을 위한 보조 기술 등 다양한 분야에서 응용 가능성이 높다.

그러나 Video to Speech 분야에서는 여전히 두가지 핵심적인 어려움이 존재한다. 첫째, 입술 움직임만으로 정확한 문장을 예측하기 어려운 입술의 모호함(Lip ambiguity)문제가 있으며, 둘째, 생성된 음성의 억양, 속도, 자연스러움이 실제 화자의 음성과 괴리되는 문제가 있다. 기존 연구들은 문장의 명확성을 우선하면 음성의 자연스러움이 떨어지고, 반대로 음성의 자연스러움을 높이면 문장의 명확성이 저하되는 트레이드오프(Trade-off)관계에 있다.[1][2]

본 연구에서는 이러한 한계를 극복하기 위해 최근 음성 및 이미지 생성 분야에서 주목받고 있는 확산 모델(Diffusion Model)[3]을 기반으로, Figure 1 을 통해서 확산 모델을 통한 음성 생성 과정을 알 수 있다. 본 연구에서는 그 중 조건부(Condition)입력 과정의 설계에 주목하였다. 특히 Video to Speech 분야에서

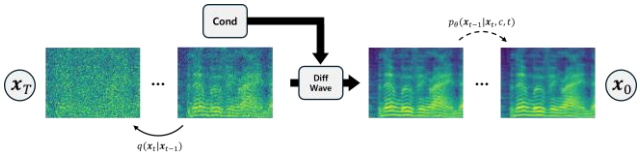

Figure 1. Diffusion Model을 통한 음성 생성 과정

비디오 프레임과 텍스트의 자연스러운 융합이 생성 음성의 품질에 결정적인 영향을 미친다는 점을 착안하여, 비디오 프레임에서 추출한 임베딩(Embedding)과 립 리딩을 통해서 얻은 텍스트 임베딩 간의 크로스 어텐션(Cross-Attention)과 셀프 어텐션(Self-Attention)을 결합한 통합 Attention 구조를 제안한다.[4]

제안하는 모델은 텍스트를 학습 과정에 직접 활용하여 비디오 프레임과 텍스트 간의 상관관계를 강화하고, 이를 통해 문장의 명확성과 음성의 자연스러움을 동시에 확보하였다.

실험 결과, 제안하는 모델은 다른 Video to Speech 모델과 비교하여 음성의 자연스러움과 문장 인식 정확도(Word Error Rate)모두에서 우수한 성능을 나타냈으며, 다양한 상황에서 안정적인 결과를 도출하였다.

Ⅱ. 본론

해당 논문에서 제시하는 모델은 기본적으로 확산 모델을 기본으로 한다. 확산 모델의 경우 학습 과정에서

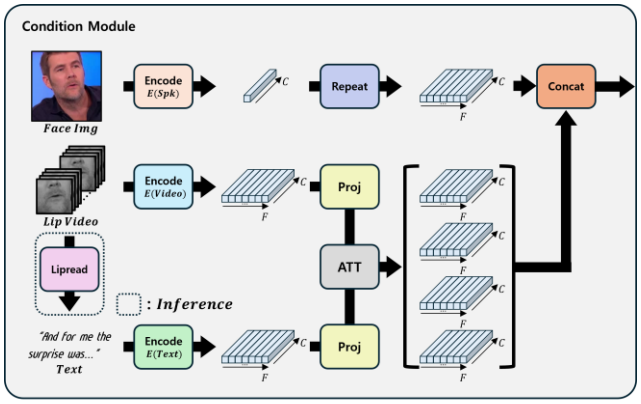

Figure 2. Attention 기반 음성 생성 인공신경망 모식도

이미지에 노이즈를 입히며 이를 다시 복원하는 과정에서 노이즈를 예측하는 모델을 학습한다. 이를 통해서 평가 과정에서는 노이즈 데이터에서 노이즈를 제거하면서 이미지를 생성해간다. 음성 생성에서는 멜 스펙토그램을 생성하여 이를 보코더(Vocoder)를 통해 음원으로 만들게 된다. 손실 함수의 경우 Equation 1 에 의해 계산되며[2], 평가 과정에서는 조건부와 함께 Equation 2 에 의해 계산된다. [5]

$$\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\|^2\right]$$

Equation 1. Loss Function about Diffusion Model

$$\hat{\boldsymbol{\epsilon}}=\boldsymbol{\epsilon}_{mg}-\omega\,\gamma_t\sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{x}_t}\log p(t_{LR}\mid\mathbf{x}_t)$$

Equation 2. Guided noise prediction for Condition Speech Generation

Figure 2 는 어텐션 메커니즘을 이용한 음성 생성 인공신경망의 전체 모식도를 보여주고 있다. 해당 모델에서는 립 리딩을 통해 얻어낸 실제 텍스트를 기반으로 기존 다른 모델에서는 없는 직접적인 텍스트 정보를 얻어낼 수 있다. 이를 기반으로 최종 입력으로 총 세가지의 입력이 사용된다. 첫째, 얼굴 이미지가 사용된다. 얼굴 이미지는 입력되는 프레임 중 랜덤하게 선택된 한 개의 얼굴 이미지 전체를 사용하게 된다. 둘째, 입력된 무음 영상에서 25 프레임을 랜덤하게 선택하게 된다. 이때, 입술 주변 부분을 Crop 하여 사용하게 되며, 출력의 형태는 프레임 정보가 살아있는 형태로 출력되게 된다. 마지막으로, 텍스트가 입력으로 들어가게 되며, 이는 학습 과정 중에서는 실제 정답 텍스트를 사용하고, 평가 과정 중에서는 립 리딩을 통해서 구한 텍스트를 사용한다.

각각의 인코더(Encoder)를 통해서 구해진 임베딩에서 비디오 임베딩과 텍스트 임베딩을 어텐션 메커니즘을 사용하게 된다. 각각 임베딩을 이용하여 두개의 셀프 어텐션과 두개의 크로스 어텐션을 사용하게 되며, 결과적으로 총 4 개의 임베딩 값들을 얻게 된다. 앞서 얼굴 이미지를 통해 얻어진 임베딩에서 프레임 개수에 맞추어 반복하게 되어 총 5 개의 값들을 채널 방향으로 합치게 된다. 이를 확산 모델의 조건부 입력으로 넣어주어 음성 생성에서 비디오와 텍스트 간의 상관관계를 더 강화하여 자연스러운 음성이 나오도록 유도한다.

Ⅲ. 결론

Table 1 은 본 연구에서 제안하는 모델과 기존에 존재하는 모델과의 정량적 비교 결과를 보여준다. Video to Speech 분야의 정량적 지표 중 가장 중요한 지표는 Word Error Rate(WER)이다. 이는 문장의 정확성을 측정하는 지표로 음성 인식 결과와 정답 텍스트 간의 차이를 계산하는 지표이다. 해당 지표에서 다른 모델에 비해 뛰어난 성능을 보여주는 것을 확인할 수 있다. 그 외에 Short – Time Objective Intelligibility Network(STOI – Net)은 음성의 명료도를 평가하는 지표이고, Deep Noise Suppression Mean Opinion Score(DNSMOS)는 음성의 자연스러움, Lip Sync Error-Distance, Confidence(LSE-D, LSE-C)는 입술과 음성 표현 사이의 거리를 표현한 지표이며, Distance 가 낮을수록 음성과 입술 동기화가 된 의미이며, Confidence 가 높을수록 상관관계가 높다. 이러한 지표에서 마찬가지로 다른 모델에 비해 뛰어난 성능을 보여주는 것을 확인할 수 있다.

결과적으로 본 논문에서는 비디오 임베딩과 텍스트 임베딩간의 어텐션을 통해서 상관관계를 강화하고, 이를 통해 음성의 자연스러움과 문장의 명확도를 상승시키는 것을 실험적으로 확인하였다.

Table 1. Performance comparison on LRS2

Method	WER↓	STOI-Net↑	DNSMOS↑	LSE-C↑	LSE-D↓
GT	1.5%	0.91	3.14	6.840	7.194
VCA-GAN	100.78%	0.51	2.26	2.396	11.763
DiffV2S	50.78%	0.89	2.92	6.432	7.654
Intelligible	44.23%	0.86	2.70	7.128	7.021
Intelligible + Avhubert	34.66%	0.86	2.67	7.018	7.111
V2Sflow-A	35.49%	**0.93**	**3.14**	7.029	7.329
V2Sflow-V	36.13%	**0.93**	3.12	7.189	7.264
Proposed	**25.91%**	0.91	3.03	**7.301**	**6.985**

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2020-II201602)

참 고 문 헌

[1] Choi, Jeongsoo, Joanna Hong, and Yong Man Ro. "Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[2] Choi, Jeongsoo, Minsu Kim, and Yong Man Ro. "Intelligible lip-to-speech synthesis with speech units." *arXiv preprint arXiv:2305.19603* (2023).

[3] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

[4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[5] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." *arXiv preprint arXiv:2207.12598* (2022).