한국음향학회지 제38권 제5호 pp. 614~620 (2019)
The Journal of the Acoustical Society of Korea Vol.38, No.5 (2019)
https://doi.org/10.7776/ASK.2019.38.5.614

pISSN : 1225-4428
eISSN : 2287-3775

k-평균 알고리즘을 활용한 음성의 대표 감정 스타일 결정 방법

Determination of representative emotional style of speech based on k-means algorithm

오상신,[1] 엄세연,[1] 장인선,[2] 안충현,[2] 강홍구[1†]

(Sangshin Oh,[1] Se-Yun Um,[1] Inseon Jang,[2] Chung Hyun Ahn,[2] and Hong-Goo Kang[1†])

[1]연세대학교 전기전자공학부, [2]한국전자통신연구원 미디어연구본부

(Received July 16, 2019; accepted September 4, 2019)

초 록: 본 논문은 전역 스타일 토큰(Global Style Token, GST)을 사용하는 종단 간(end-to-end) 감정 음성 합성 시스템의 성능을 높이기 위해 각 감정의 스타일 벡터를 효과적으로 결정하는 방법을 제안한다. 기존 방법은 각 감정을 표현하기 위해 한 개의 대푯값만을 사용하므로 감정 표현의 풍부함 측면에서 크게 제한된다. 이를 해결하기 위해 본 논문에서는 k-평균 알고리즘을 사용하여 다수의 대표 스타일을 추출하는 방법을 제안한다. 청취 평가를 통해 제안 방법을 이용해 추출한 각 감정의 대표 스타일이 기존 방법에 비해 감정 표현 정도가 뛰어나며, 감정 간의 차이를 명확히 구별할 수 있음을 보였다.

핵심용어: 음성 합성, 종단 간 음성 합성, 감정 음성 합성, 스타일 토큰

ABSTRACT: In this paper, we propose a method to effectively determine the representative style embedding of each emotion class to improve the global style token-based end-to-end speech synthesis system. The emotion expressiveness of conventional approach was limited because it utilized only one style representative per each emotion. We overcome the problem by extracting multiple number of representatives per each emotion using a k-means clustering algorithm. Through the results of listening tests, it is proved that the proposed method clearly express each emotion while distinguishing one emotion from others.

Keywords: Speech synthesis, End-to-end speech synthesis, Emotional speech synthesis, Style token

PACS numbers: 43.72.Ja, 43.70.Ep

I. 서 론

음성 합성(speech synthesis or Text-To-Speech, TTS) 시스템은 주어진 텍스트 입력에 알맞은 음성을 합성해내는 기술로, 내비게이션, 모바일 인공 지능 비서 서비스, 인공 지능 스피커 등 다양한 음성 인터페이스 시스템에 탑재되어 널리 이용되고 있다. 급격히 발달하고 있는 딥러닝 기술이 활용됨에 따라 합성 음성의 품질이 매우 향상되고 있으며,[1,2,3] 특히 최근에 제안된 종단 간 음성 합성 시스템[4,5,6,7,8] 중 타코트론 모델(Tacotron)[4,5]을 사용하면 이전 모델에 비해 간단한 과정을 통해 음성을 합성할 수 있을 뿐만 아니라, 실제 녹음한 음성과 크게 다르지 않은 합성음 품질을 얻을 수 있다. 하지만, 합성음의 자연스러움이나 생동감은 아직까지는 일상 생활에서 흔히 들을 수 있는 음성에 미치지 못하는 상황으로, 음성을 이용한 인간-컴퓨터 인터페이스 시스템의 사용자 만족도를 저하시키는 요인으로 작용한다. 특히 대화형 서비스나, 음성 더빙 등의 서비스에서는 음성에 포함된 감정이 합성되는 문장의 의미를 파악하는데 중요하게

†**Corresponding author:** Hong-Goo Kang (hgkang@yonsei.ac.kr)
School of Electrical and Electronic Engineering 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
(Tel: 82-2-2123-4534)

작용하는 요소로, 감정을 반영하여 음성을 합성하는 것이 매우 중요하다고 할 수 있다.

감정을 반영하여 음성을 합성하기 위하여 많은 방법들이 제안되고 있다[9,10,11,12,13]. Lee *et al.*[9]은 종단 간 음성 합성 시스템에 감정을 조건 벡터로 입력하여 감정 음성을 합성하는 방법을 제안하였다. 이 논문에서 제안한 모델은 간단하게 감정 음성을 합성할 수 있다는 장점이 있지만, 단순한 구조로 인해 감정을 조절할 수 있는 범위가 크게 제한되기 때문에 감정 표현의 세밀한 조절에는 한계가 존재한다.

한편, Kwon *et al.*[10]은 다양한 스타일의 합성음을 생성할 수 있는 전역 스타일 토큰 기반 타코트론(GST-Tacotron) 시스템[14]을 기본 시스템으로 하여, 각 감정의 대표 스타일을 정하여 감정 음성을 합성하는 방법을 제안하였다. 이 방법은 Lee *et al.* [9]에 비하여 합성 과정이 더 복잡하지 않으면서 감정 표현을 세밀하게 조절할 가능성이 있다는 측면에서 큰 의의가 있다. 하지만 세밀한 조절을 위해서는 각 토큰이 음성에 미치는 영향을 사전에 확인하여야 하고, 조절 과정에서 다른 감정처럼 들리게 될 우려가 있어 사용의 용이성 및 안정성 측면에서 개선의 여지가 있다. 이에 본 연구에서는 각 감정을 다양하게 표현하기 위해 여러 개의 대표 스타일을 추출하는 방법을 제안하고, 이를 통해 제안 방법이 기존 방법보다 안정적이고 세밀한 조절이 가능한 감정 음성 합성 방법임을 보인다.

본 연구에서 제시한 방법은 k-평균(k-means) 알고리즘을 통해 각 감정의 대표 스타일 벡터를 계산하는 방식으로, 레이블 된 학습 데이터로부터 각 감정을 표현하는 복수의 대표 스타일을 추출함으로써 합성 시 다양한 스타일의 감정을 표현할 수 있다. 이 방법을 통해 추출된 대표 스타일 벡터는 해당 감정을 다양한 형태로 표현하고 있으며, 다른 감정과는 확실히 구분할 수 있는 특성을 가지고 있다. 청취 실험을 통해 제안 방법이 기존 방법보다 성능이 우수하고, 세밀한 감정 표현이 가능하다는 것을 보였다.

II. 배경 지식

타코트론과 전역 스타일 토큰 모듈로 구성된 GST-Tacotron[14]은 합성음의 스타일을 참조 오디오와 유사한 특성을 갖도록 유동적으로 변형 시킬 수 있다. 합성음의 스타일을 결정하는 스타일 벡터가 생성되는 과정은 다음과 같다. 먼저, Fig. 1의 윗 부분에 묘사된 것과 같이, 구해진 참조 오디오의 멜-스펙트로그램은 참조 인코더에 입력되어 정해진 길이의 운율 임베딩을 생성한다[15]. 그 다음 다중-헤드 어텐션 모듈을 통해 스타일 토큰과 운율 임베딩 사이의 유사성을 측정하여 스타일 토큰의 가중치를 계산한다. 마지막으로 스타일 토큰들의 가중 합으로 생성된 스타일 벡터는 텍스트 인코더에서 생성된 은닉 특징들과 함께 타코트론의 어텐션 모듈로 입력된다. 이 방법을 활용하여 감정이 포함된 음성신호를 생성하기 위해서는 학습된 스타일 벡터의 각 차원이 음성의 어떤 특성에 영향을 끼치는지 확인하는 수작업이 필요하다.

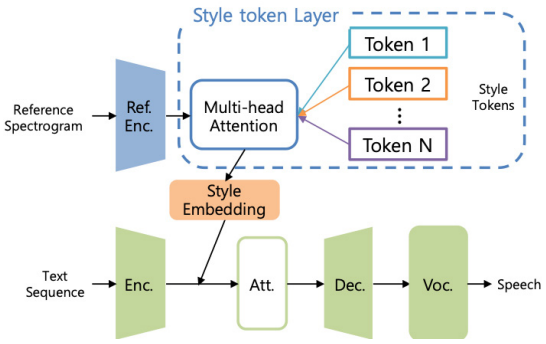

Fig. 1. Structure of GST-Tacotron. Style token layer for style representation is added to Tacotron model. Reference spectrogram is fed to the reference encoder (Ref. Enc.), and the text sequence is fed to encoder (Enc.), attention module (Att.), decoder (Dec.), and neural vocoder (Voc.) sequentially.

또 다른 방법으로 Kwon *et al.*[10]에서는 GST-Tacotron을 기반으로 각 감정에 따라 여러 학습 데이터로부터 토큰의 가중치를 구하고, 그 평균값을 사용하여 감정 음성을 표현하였다. 학습 단계에서는 행복, 슬픔, 화남 그리고 중립적인 감정과 같은 각각의 감정 데이터 세트로부터 가중치를 얻는다. 각 샘플 데이터에서 얻어진 가중치는 감정 별로 분류할 수 있으므로, 각 집단의 평균은 그 감정의 대표적인 가중치로 간주 할 수 있다. 그러나, 그렇게 계산된 대표 가중치가 각 감정 범주를 모두 잘 반영한다는 보장은 없다.

합성음에 나타나는 각 감정의 표현을 자유롭게 조

절하기 위해, 본 연구에서는 각 감정의 대푯값을 결정하기 위한 다른 전략을 조사한다. 특히, 동일한 감정도 다양하게 표현할 수 있다는 점에 착안하여 각 감정 데이터를 여러 개의 클러스터로 나누어 모델링하고 그 대푯값들을 해당 감정의 대표 스타일로 사용한다.

III. 제안 방법

3.1 기본 시스템

본 연구에서는 Kwon *et al.*[10]과 같이 GST-Tacotron을 기본 시스템으로 사용한다. 앞서 설명한 바와 같이 타코트론은 종단 간 음성 합성 시스템으로, 별도의 특징 추출 과정 없이 주어진 텍스트 입력으로부터 합성 음성을 생성한다. GST-Tacotron 시스템은 합성음의 스타일을 조절하기 위해 추가적인 스타일 토큰 층이 더해진 형태로, 추가된 층을 통해 음성의 스타일을 추정 및 학습한다. 훈련 과정이 종료된 후, 학습된 스타일 토큰 값들을 이용하여 특정 감정 음성을 표현하는 스타일 벡터를 추출하게 된다. 이렇게 추출된 스타일 벡터는 음성을 합성할 때 텍스트 입력과 함께 조건 벡터의 형태로 입력되어 합성음의 스타일을 조절한다.

참조 오디오를 사용하여 매번 원하는 스타일을 추출하는 대신 간결한 형태로 감정 음성을 합성하기 위해서는 각 감정을 대표할 수 있는 스타일 벡터가 결정되어야 한다. 즉, 각 감정의 대표 스타일 벡터를 정하면, 그 대표 스타일 벡터를 시스템에 텍스트 입력과 함께 주어줌으로써 간단하게 감정이 표현된 음성을 합성할 수 있다. 하지만, 해당 감정의 스타일이 대표 스타일 벡터로 고정되게 되므로, 감정의 대푯값을 정하는 일은 감정 음성 합성 시스템에서 매우 중요한 요소라고 할 수 있다. 다음 절에서는 각 감정의 대표 스타일을 결정하는 방법을 제안한다.

3.2 대표 감정 스타일 추정 방법

본 연구에서는 각 감정 음성 데이터에서 한 개의 대표 스타일 벡터를 추출하는 대신, 다양한 대표 스타일 벡터를 추출하기 위하여 k-평균 알고리즘[16]을

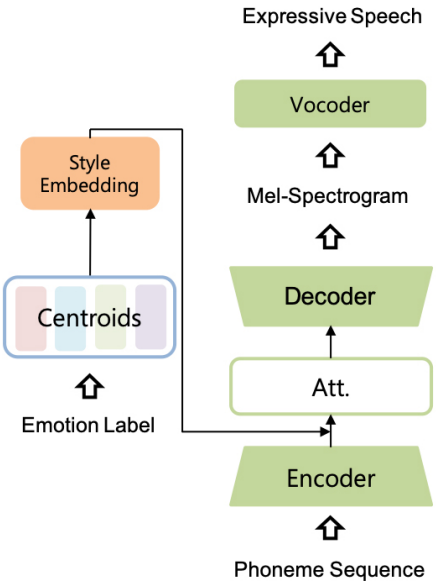

Fig. 2. Structure of the proposed system. The emotional expression is controlled by the emotional label and the control vector. The centroids from k-means algorithm are used for representative style embeddings.

사용한다. k-평균 알고리즘은 대표적인 비 지도학습 방식 중 하나로, 주어진 벡터들을 k개의 클러스터로 나누기 위하여 사용되는 알고리즘이다. 이 알고리즘을 사용하면 각 벡터는 특정 클러스터에 할당되게 되며, 각 클러스터의 대푯값은 그 클러스터에 속하는 벡터들의 평균으로 결정된다. k-평균 알고리즘의 k 값은 사전에 설정할 수 있는 하이퍼파라미터 값 중 하나로 사용자의 편의에 맞추어 설정할 수 있다. 따라서, k-평균 알고리즘을 사용하면 감정 음성 데이터로부터 각 감정마다 k 개의 대표 스타일을 추출할 수 있다. Fig. 2는 k-평균 알고리즘을 사용하여 추출한 대푯값을 통해 감정 음성을 합성할 수 있는 시스템의 구조도이다.

일반적인 k-평균 알고리즘에서와 같이, 대표 감정 스타일을 얻기 위해서 각 벡터들을 가까운 클러스터에 배당하는 할당 과정과 대푯값을 재설정하는 업데이트 과정이라는 두 개의 작업을 번갈아 수행한다. 할당 과정은 스타일 벡터를 각 클러스터의 대푯값까지의 유클리드 거리를 계산하여 가장 가까운 클러스터에 할당하는 과정으로서 다음과 같은 수식으로 표현될 수 있다.

$$C_i = \{\boldsymbol{x_p}: \| \boldsymbol{x_p} - \boldsymbol{r_i} \|^2 \leq \min_j \| \boldsymbol{x_p} - \boldsymbol{r_j} \|^2 \}, \quad (1)$$

이 때, 집합 C_i는 벡터 $\boldsymbol{r_i}$를 대푯값으로 가지는 스타일에 대한 클러스터를 의미하며 스타일 벡터를 원소로 갖는다. 한편, 업데이트 과정에서는 클러스터에 해당하는 스타일 벡터의 무게중심을 계산하여 클러스터의 새로운 대푯값으로 재설정하며 다음과 같은 수식으로 표현될 수 있다.

$$\boldsymbol{r_i} = \frac{1}{|C_i|} \sum_{\boldsymbol{x_p} \in C_i} \boldsymbol{x_p}, \quad (2)$$

위에서 설명한 두 작업은 클러스터와 대푯값이 더 이상 업데이트 되지 않을 때까지 번갈아 반복되며, 수렴한 클러스터의 대푯값을 대표 감정 스타일 벡터로 사용하여 감정 음성을 합성한다.

본 논문에서 제안하는 방법은 감정마다 하나의 대푯값만을 사용하는 기존 방법과 달리 여러 개의 대푯값을 얻을 수 있다. 또한, 분류 작업이 각 감정마다 독립적으로 적용될 수 있으므로, 감정마다 서로 다른 개수의 클러스터로 모델링하여 대표 스타일 벡터를 얻을 수도 있다. 따라서, 제안 방법은 기존 방법에 비해 보다 다양한 스타일의 감정 음성, 그리고 세밀한 감정 표현이 가능하다고 할 수 있다.

IV. 실험 및 결과

본 절에서는 본 논문에서 제안하는 감정 음성 합성 방법의 유용성과 성능을 보이기 위한 실험의 내용을 설명하고 그 결과에 대해 분석한다. 실험 기존 방법과 비교가 가능한 경우 Kwon *et al.*[10]에서 제안된 시스템과 비교하여 진행되었다. 기존 방법과 제안 방법 모두 내부 데이터베이스를 이용해 학습되었으며, 각 감정 별 1시간으로 총 4시간의 남성 화자 데이터베이스이다.

4.1 청취 평가

음성 합성 시스템의 성능을 확인하는 가장 대표적인 방법은 사람이 직접 듣고 평가하는 것이다. 본 연구에서는 청취 평가의 신뢰성을 높이기 위해 다음과 같이 세 가지 항목, 즉 (1) 합성음의 음질, (2) 합성음이 감정을 잘 표현하는가, (3) 서로 다른 클러스터 간에 구분이 잘 되는가를 확인하기 위해 설계되었으며, 총 9명을 대상으로 진행하였다.

Table 1은 기존 방법과 제안하는 방법의 합성음의 품질에 대한 MOS(Mean Opinion Score) 결과이다. 점수는 감정에 따라서 따로 측정되었으며, 각 평가자는 합성음의 음질을 1점부터 5점 사이의 점수로 평가하였다. 음질 평가 결과는 각 감정에 따라 다른 양상을 보인다. 행복(happiness)과 중립(neutral)의 경우 기존 방법에 비해 조금 낮은 결과를 보이지만, 슬픔(sadness)의 경우에는 기존 방법보다 높은 음질 점수를 보이며 전반적으로 비슷한 성능을 보인다는 것을 알 수 있다. 또한, 감정 간 음질 차이가 큰 기존 방법과 달리 제안하는 방법은 그 차이가 줄어들어 더 안정적인 합성음 품질을 기대할 수 있다.

Table 2는 기존 방법과 제안하는 방법이 감정을 얼마나 명확하게 표현하는지를 보여준다. 각각의 방법으로 생성된 합성음이 어떤 감정으로 들리는지에 대하여 실험하였으며, 실제 분류와 비교하여 정확도를 측정하였다. 전반적으로 두 방법 모두 높은 정확성을 보여주지만, 모든 감정에서 제안하는 방법을 통해 합성된 음성이 감정을 더 명확하게 표현하는 것을 확인할 수 있다. 명확한 감정 표현은 의사 전달을

Table 1. Mean opinion score (MOS) of synthesized speech according to the emotion.

	Conventional	Proposed
Anger	3.56	3.56
Happiness	3.89	3.67
Neutral	3.73	3.71
Sadness	3.42	3.56
Average	3.65	3.63

Table 2. Expression accuracy according to the emotion.

	Conventional	Proposed
Anger	100 %	100 %
Happiness	76 %	78 %
Neutral	71 %	100 %
Sadness	96 %	100 %

Table 3. Discrimination accuracy between the different style in same emotion.

	TPR	TNR	Accuracy
Anger	95 %	65 %	75 %
Happiness	85 %	58 %	66 %
Neutral	90 %	40 %	56 %
Sadness	100 %	80 %	87 %

보다 효과적으로 수행할 수 있으므로, 제안하는 방법은 감정 음성 합성 방법으로서 중요한 장점을 지닌다고 할 수 있다.

마지막으로 진행한 청취 실험의 결과는 Table 3에 나타나 있다. 두 개의 음성 샘플이 같은 스타일인지 혹은 다른 스타일인지 판단하는 평가로, 그 결과를 같은 스타일을 맞게 판단한 비율인 TPR(True Positive Rate), 다른 스타일을 맞게 판단한 비율인 TNR(True Negative Rate)과 둘 모두를 고려한 정확도로 나누어 표시하였다. TPR 은 모든 감정에서 높은 반면, TNR 의 경우에는 특히 행복과 중립 감정에서 비교적 낮은 결과를 얻었다. 이는 음성 데이터에 실제 존재하는 감정 표현의 종류보다 클러스터를 더 세밀하게 나누었기 때문으로, 클러스터의 수를 최적화하여 더욱 개선할 여지가 있다. 한편, 화남과 슬픔 감정의 경우 TNR 결과 역시 비교적 높은 결과를 얻었다. 이는 음성 데이터 내에 비교적 다양한 형태의 화남 및 슬픔 감정 표현이 있었기 때문으로 이해할 수 있다. 이를 통해 클러스터의 수가 적절히 설정되었을 때에는 서로 잘 구분된다는 것을 확인할 수 있다.

4.2 스타일 임베딩 분석

음성 스타일의 임베딩 공간에서 서로 먼 거리를 갖는 스타일 벡터들은 매우 특성이 다른 발화 스타일을 보이게 된다. 때문에, 특정 스타일을 대표하는 대푯값은 해당 스타일과 멀리 떨어지지 않아야 하며, 클러스터의 대푯값과 해당 클러스터에 속하는 스타일 벡터들 간의 거리는 벡터 양자화(vector quantization)에서의 양자화 오류(quantization error)와 비슷하게 이해될 수 있다. 따라서 이 거리가 작을수록 좋은 대푯값이라고 할 수 있다. 이러한 측면에서 봤을 때, k-평균 알고리즘을 사용할 경우 평균을 취했을

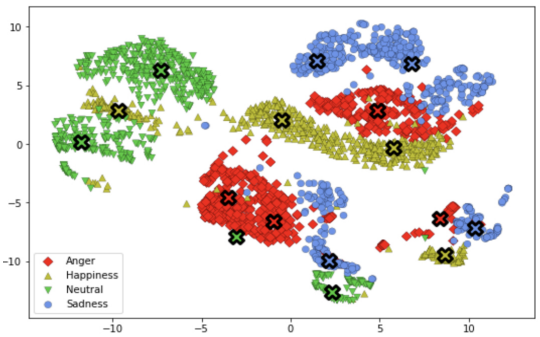

Fig. 3. Distribution of emotional style vector and the representative styles. The representatives are markers with 'x' sign.

때보다 거리가 작아지므로 원래의 음성 스타일을 잘 보존하는 좋은 대푯값을 생성한다고 할 수 있다.

한편, 고차원 벡터를 저차원으로 차원 축소하여 시각화함으로써 고차원에서의 분포를 추측해볼 수 있다. Fig. 3은 각 감정의 스타일 벡터와 클러스터들의 대푯값을 t-SNE 알고리즘을 이용하여 저차원 임베딩을 얻은 것이다. 보이는 바와 같이, 같은 감정도 여러 개의 클러스터로 구분되는 것을 확인할 수 있으며, 대푯값들이 각 클러스터를 잘 대표한다고 할 수 있다. t-SNE로 얻은 저차원 임베딩은 고차원 분포에서의 전체적인 모습을 보존하지는 못하지만 거리는 잘 보존되는 경향이 있으므로, 각 스타일 벡터들과 대푯값 사이의 거리가 작을 것이라는 추론이 가능하다. 따라서 위의 표에서 제시한 결과와 같이 제안 방법이 기존 방법보다 각 스타일 벡터까지의 거리가 작은 대푯값을 제시한다고 할 수 있다. 또한, 제안 방법은 각 감정을 더 작은 클러스터 단위로 나누어 모델링하므로 특정 감정의 스타일 벡터가 다른 감정 사이에 존재하는 경우에 대해서도 감정이 불분명해지거나 혼재되는 등의 문제를 방지할 수 있다.

V. 결 론

본 논문에서는 감정 음성 합성 시스템에서 효과적으로 감정을 표현하고, 그 다양성을 높이기 위해 k-평균 군집화를 도입하여 각 감정을 표현하는 대푯값을 여러 개로 추출하는 방법을 제안하였다. 제안 방법은 각 감정을 서로 다른 클러스터로 구분하여 모

델링하여 서로 확연히 다른 특성을 보인다. 또한, 주관적 청취 평가를 통해 제안 방법으로 감정 음성을 합성할 경우, 합성음의 품질을 유지하며 효과적으로 감정 음성을 합성할 수 있음을 보였다. 제안 방법을 통해 각 감정 데이터에서 얻은 다양한 대표 스타일 벡터들은 클러스터의 수가 적절히 설정되었을 경우 서로 잘 구분이 되며, 감정 표현을 더욱 분명히 하므로 세밀함 뿐만 아니라 명확함 또한 개선할 수 있음을 보였다.

감사의 글

본 연구 논문은 과학기술정보통신부 및 정보통신기획평가원의 출연금으로 수행하고 있는 한국전자통신연구원 시청각 장애인의 방송시청을 지원하는 감성표현 서비스 개발[2019-0-00447] 위탁연구과제의 연구결과입니다.

References

1. H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. IEEE ICASSP, 7962-7966 (2013).
2. Y. Qian, Y. Fan, W. Hu, and F. K Soong, “On the training aspects of deep neural network (dnn) for parametric tts synthesis,” Proc. IEEE ICASSP, 3829-3833 (2014).
3. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv: 1609.03499 (2016).
4. Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J Weiss, N. jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A Saurous, “Tacotron: Towards end-to-end speech synthesis,” Proc. Interspeech, 4006-4010 (2017).
5. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” Proc. IEEE ICASSP, 4779-4783 (2018).
6. J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2wav: End-to-end speech synthesis,” Proc. ICLR, 1-6 (2017).
7. A. Gibiansky, S. Arik, G. Diamos, J. Miler, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” Advances in NIPS, 2962-2970 (2017).
8. Y. Wang, R. J. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, “Uncovering latent style factors for expressive speech synthesis,” arXiv preprint arXiv:1711.00520 (2017).
9. Y. Lee, A. Rabiee, and S. -Y. Lee, “Emotional end-to-end neural speech synthesizer,” arXiv preprint arXiv: 1711.05447 (2017).
10. O. Kwon, I. Jang, C. H. Ahn, and H. -G. Kang, “Emotional speech synthesis based on style embedded Tacotron2 framework,” Proc. ITC-CSCC, 1-4 (2019).
11. J. Tao, Y. Kang, and A. Li, “Prosody conversion from neutral speech to emotional speech,” IEEE Trans. on Audio, Speech, and Lang. Process. **14**, 1145-1154 (2006).
12. Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice conversion with smoothed gmm and map adaptation,” Eighth European Conference on Speech Communication and Technology, 2413-2416 (2003).
13. Y. -J. Zhang, S. Pan, L. He, and Z. -H. Ling, “Learning latent representation for style control and transfer in end-to-end speech synthesis,” Proc. IEEE ICASSP, 6945-6949 (2019).
14. Y. Wang, D. Stanton, Y. Zhang, RJ. Skerry- Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to- end speech synthesis,” arXiv preprint arXiv:1803.09017 (2018).
15. RJ. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” arXiv preprit arXiv:1803.09047 (2018).
16. S. Lloyd, “Least squares quantization in PCM,” IEEE Trans. on information theory, **28**, 129-137 (1982).

저자 약력

▸ 오 상 신 (Sangshin Oh)

2019년 2월 : 연세대 전기전자공학과 학사
2019년 3월 ~ 현재 : 연세대 전기전자공학과 석사 과정

▸ 엄 세 연 (Se-Yun Um)

2017년 2월 : 숭실대 정보통신전자공학과 학사
2018년 9월 ~ 현재 : 연세대 전기전자공학과 통합 과정

▸ 장 인 선 (Inseon Jang)

2001년 2월 : 충북대학교 전기전자공학부 정보통신공학 학사
2004년 2월 : 포항공과대학교 컴퓨터공학과 석사
2018년 2월 : 충남대학교 전자전파정보통신공학과 박사
2004년 8월 ~ 현재 : 한국전자통신연구원 선임연구원

▸ 안 충 현 (Chung Hyun Ahn)

1985년 2월 : 인하대학교 해양학과 학사
1989년 8월 : 인하대학교 해양학과 석사
1986년 ~ 1991년 : 한국 해양연구소 연구원
1995년 3월 : 일본 치바대학교 환경원격탐사센터 박사
1995년 3월 ~ 12월 : 일본 치바대학교 정보공학과 연구조수
1996년 ~ 현재 : 한국전자통신연구원 책임연구원

▸ 강 홍 구 (Hong-Goo Kang)

1989년 2월 : 연세대 전자공학과 학사
1991년 2월 : 연세대 전자공학과 석사
1995년 8월 : 연세대 전자공학과 박사
1996년 4월 : AT&T Lab. Senior Technical Staff Member
2002년 9월 ~ 현재 : 연세대 전기전자공학과 교수