신경 음성 합성에 대한 조사

Xu Tan*, Tao Qin, Frank Soong, Tie-Yan Liu
{xuta,taoqin,frankkps,tyliu}@microsoft.com Microsoft
Research Asia

추상적인

TTS(Text to Speech) 또는 음성 합성은 주어진 텍스트를 통해 이해하기 쉽고 자연스러운 음성을 합성하는 것을 목표로 하며 음성, 언어 및 기계 학습 커뮤니티에서 뜨거운 연구 주제이며 업계에서 광범위하게 응용되고 있습니다. 딥러닝과 인공지능 의 발전으로 신경망 기반 의 TTS는 최근 합성 음성의 품질을 획기적으로 향상시켰다.

이 논문에서는 현재 연구와 미래 동향에 대한 좋은 이해를 제공하는 것을 목표로 신경 TTS에 대한 포괄적인 조사를 수행합니다. 우리는 텍스트 분석, 음향 모델 및 보코더를 포함한 신경 TTS의 핵심 구성 요소와 빠른 TTS, 저자원 TTS, 강력한 TTS, 표현력이 풍부한 TTS 및 적응형 TTS 등을 포함한 몇 가지 고급 주제에 중점을 둡니다. 리소스를 추가로 요약 합니다 . TTS 관련(예: 데이터 세트, 오픈 소스 구현) 및 향후 연구 방향에 대해 논의합니다. 이 설문 조사는 TTS를 연구하는 학술 연구원과 업계 실무자 모두에게 도움이 될 수 있습니다.

1. 소개

음성 합성으로도 알려진 TTS(Text to Speech)는 텍스트 [346] 에서 이해할 수 있고 자연스러운 음성을 합성하는 것을 목표로 하며 인간 의사소통 [1] 에 광범위하게 적용되며 오랫동안 인공 지능, 자연 언어 및 음성 처리 [296, 228, 147].

TTS 시스템을 개발하려면 언어와 인간의 음성 생성에 대한 지식이 필요하며 언어학 [63], 음향학 [170], 디지털 신호 처리 [320] 및 기계 학습[25, 146]을 포함한 여러 분야가 관련됩니다.

딥 러닝 [183, 89]의 발전으로 신경망 기반 TTS가 번창했고, 신경 TTS의 다양한 측면에 초점을 맞춘 많은 연구가 나왔다 [426, 254, 382, 303, 150, 270, 192] , 290]. 결과적으로 합성 음성의 품질은 최근 몇 년 동안 크게 향상되었습니다 . 현재 연구 현황을 파악하고 미해결 연구 문제를 파악하는 것은 TTS를 작업하는 사람들에게 큰 도움이 됩니다. 통계적 파라메트릭 음성 합성 [27, 425, 357, 422] 및 신경 TTS [331, 226, 306, 248, 118, 260, 242]에 대한 여러 조사 논문이 있지만 , 이 분야의 주제는 다양하고 빠르게 발전하기 때문에 신경 TTS는 여전히 필요 합니다. 이 백서에서는 신경 TTS2에 대한 심층적이고 포괄적인 조사를 수행합니다. 3 다음 하위 섹션에서는 먼저 TTS 기술의 역사를 간략하게 검토한 다음 신경 TTS에 대한 기본 지식을 소개하고 마지막으로 이 조사의 개요를 설명합니다.

.

교신저자: Xu Tan, xuta@microsoft.com 2이 설문지는
ISCSLP 2021의 TTS 튜토리얼을 포함한 TTS 튜토리얼 (https: //tts-tutorial.github.io/iscslp2021/) 에서 가져온 것입니다. 및 IJCAI 2021의 TTS 튜토리얼 (https://tts-tutorial.g ithub.io/ijcai2021/).

3Readers는 이 Github 페이지 (https://github.com/tts-tutorial/survey) 를 사용할 수 있습니다. 이 설문 조사 보고서에 대한 업데이트를 확인하고 토론을 시작하십시오.

1.1 TTS 기술의 역사

사람들은 12세기로 거슬러 올라가 인간의 말을 합성하는 기계를 만들려고 시도했습니다 [388].
18세기 후반에 헝가리 과학자 볼프강 폰 켐펠렌(Wolfgang von Kempelen)은 몇 가지 간단한 단어와 짧은 문장을 생성하기 위해 풀무, 용수철, 백파이프 및 공명 상자로 구성된 음성 기계를 제작했습니다 [72]. 컴퓨터를 기반으로 한 최초의 음성 합성 시스템은 20세기 후반에 나왔다 [388]. 초기 컴퓨터 기반 음성 합성 방법에는 조음 합성 [53, 300], 포만트 합성 [299, 5, 171, 172] 및 연결 합성 [253, 241, 297, 127, 26]이 포함됩니다. 이후 통계 기계 학습의 발전으로 음성 합성 을 위한 스펙트럼, 기본 주파수, 지속 시간 등의 매개변수를 예측하는 SPSS(Statistical Parametric Speech Synthesis)가 제안되었다 [416, 356, 425, 357] .

2010년대부터 신경망 기반 음성 합성 [426, 284, 78, 424, 375, 191, 254, 382] 이 점차 지배적인 방법이 되었고 훨씬 더 나은 음성 품질을 달성했습니다.

조음 합성 조음 합성 [53, 300] 은 입술, 혀, 성문 및 움직이는 성대와 같은 인간 조음기의 동작을 시뮬레이션하여 음성을 생성합니다 . 이상적으로 조음 합성은 인간이 음성을 생성하는 방식이기 때문에 음성 합성에 가장 효과적인 방법이 될 수 있습니다 . 그러나 실제로 이러한 교합기 동작을 모델링하는 것은 매우 어렵습니다. 예를 들어, 교합기 시뮬레이션을 위한 데이터 수집이 어렵습니다. 따라서 조음 합성에 의한 음성 품질 은 일반적으로 후기 포만트 합성 및 연결 합성에 의한 것보다 나쁩니다.

포먼트 합성 포먼트 합성 [299, 5, 171] 은 단순화된 소스-필터 모델 을 제어하는 일련의 규칙에 따라 음성을 생성합니다 . 이러한 규칙은 일반적으로 음성의 형식 구조 및 기타 스펙트럼 속성을 가능한 한 가깝게 모방하기 위해 언어학자가 개발합니다. 음성은 추가 합성 모듈과 기본 주파수, 보이싱 및 소음 수준과 같은 다양한 매개 변수를 사용하여 음향 모델에 의해 합성됩니다. 포만트 합성은 임베디드 시스템에 적합 하고 연결 합성에서와 같이 대규모 인간 음성 말뭉치에 의존하지 않는 적당한 계산 리소스로 매우 이해하기 쉬운 음성을 생성할 수 있습니다. 그러나 합성된 음성은 덜 자연스럽고 아티팩트가 있습니다. 또한 합성에 대한 규칙을 지정하기가 어렵습니다.

연결 합성 연결 합성 [253, 241, 297, 127, 26] 은 데이터베이스에 저장된 음성 조각의 연결 에 의존 합니다. 일반적으로 데이터베이스는 성우가 녹음한 전체 문장에서 음절에 이르는 음성 단위로 구성됩니다. 추론에서 연결 TTS 시스템은 주어진 입력 텍스트와 일치하도록 음성 단위를 검색하고 이러한 단위를 함께 연결하여 음성 파형을 생성합니다. 일반적으로 연결 TTS는 명료도가 높고 원래 성우에 가까운 진정한 음색을 가진 오디오를 생성할 수 있습니다. 그러나 연결 TTS는 구어에 대한 가능한 모든 음성 단위 조합을 다루기 위해 거대한 녹음 데이터베이스가 필요합니다 . 또 다른 단점은 생성된 음성이 덜 자연스럽고 감정적이라는 것입니다. 연결하면 강세, 감정, 운율 등에서 부드러움이 떨어질 수 있기 때문입니다.

Statistical Parametric Synthesis 연결 TTS의 단점을 해결하기 위해 SPSS(Statistical Parametric Speech Synthesis)가 제안됩니다 [416, 356, 415, 425, 357]. 기본 아이디어는 연결을 통해 파형을 직접 생성 하는 대신 음성을 생성하는 데 필요한 음향 매개 변수 [82, 355, 156] 를 먼저 생성한 다음 일부 알고리즘을 사용하여 생성된 음향 매개변수에서 음성을 복구할 수 있다는 것입니다 [132, 131]. , 155, 238]. SPSS는 일반적으로 텍스트 분석 모듈, 매개변수 예측 모듈(음향 모델) 및 보코더 분석/합성 모듈(보코더)의 세 가지 구성 요소로 구성됩니다. 텍스트 분석 모듈은 먼저 텍스트 정규화 [317], 자소에서 음소로의 변환 [24], 단어 분할 등을 포함하여 텍스트를 처리한 다음 음소, 기간 및 POS 태그와 같은 언어적 특징을 다른 세분성에서 추출합니다.

음향 모델(예: 은닉 마르코프 모델(HMM) 기반)은 짝을 이룬 언어 특징 및 매개변수(음향 특징)로 훈련되며, 여기서 음향 특징에는 기본 주파수, 스펙트럼 또는 켑스트럼 [82, 355] 등이 포함되며 보코더 분석 을 통해 음성에서 추출 [132, 155, 238]. 보코더는 예측된 음향 특징에서 음성을 합성합니다. SPSS 는 이전 TTS 시스템에 비해 몇 가지 장점이 있습니다. 1) 자연스러움, 오디오가 더 자연스럽습니다. 2) 유연성, 음성 생성을 제어하기 위해 파라미터를 수정하는 것이 편리하다; 3) 낮은 데이터 비용, 연결 합성보다 적은 녹음이 필요합니다. 그러나 SPSS에는 다음과 같은 단점도 있습니다. 1) 생성된 음성은 잡음, 윙윙거리는 소리 또는 시끄러운 오디오와 같은 아티팩트로 인해 명료도가 낮습니다. 2) 생성된 음성은 여전히 로봇 음성이며 사람이 녹음하는 음성과 쉽게 구분할 수 있습니다.

2010년대에 들어서 신경망과 딥러닝이 급속히 발전함에 따라 DNN(Deep Neural Network) 기반 [426, 284] 및 RNN(Recurrent Neural Network) 기반 [78, 422, 424]. 그러나 이러한 모델은 HMM을 신경망으로 대체하고 여전히 SPSS의 패러다임을 따르는 언어적 특징에서 음향 특징을 예측합니다. 나중에 Wang et al. [375] 는 end-to-end4 음성 합성 을 위한 첫 번째 탐색으로 간주할 수 있는 언어적 특징 대신 음소 시퀀스 에서 음향 특징을 직접 생성할 것을 제안합니다 . 이 설문 조사에서는 신경 기반 음성 합성과 주로 종단 간 모델에 중점을 둡니다. 나중에 SPSS도 신경망을 음향 모델로 사용하기 때문에 이러한 모델에 대해 간략하게 설명하지만 세부 사항에 대해서는 자세히 다루지 않습니다.

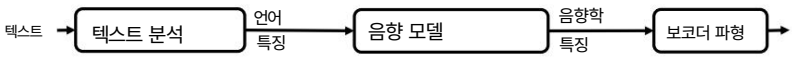


그림 1: 신경 TTS의 세 가지 주요 구성 요소.

신경 음성 합성 딥 러닝의 발전으로 음성 합성을 위한 모델 백본으로 (심층) 신경망을 채택한 신경망 기반 TTS(줄여서 신경 TTS)가 제안 됩니다. 일부 초기 신경 모델은 음향 모델링을 위해 HMM을 대체하기 위해 SPSS에서 채택되었습니다.

나중에 WaveNet [254] 은 언어적 특징에서 파형을 직접 생성하기 위해 제안되었으며, 이는 최초의 현대적인 신경 TTS 모델로 간주될 수 있습니다. DeepVoice 1/2 [8, 87] 와 같은 다른 모델은 여전히 통계적 파라메트릭 합성에서 세 가지 구성 요소를 따르지만 해당 신경망 기반 모델로 업그레이드합니다. 또한 텍스트 분석 모듈 을 단순화 하고 직접 _ _ 문자/음소 시퀀스를 입력으로 사용하고 mel-스펙트로그램으로 음향 기능을 단순화합니다. 나중에 ClariNet [269], FastSpeech 2s [292] 및 EATS [69] 와 같은 텍스트에서 파형 을 직접 생성하기 위해 완전한 종단 간 TTS 시스템이 개발되었습니다 . 연결 합성 및 통계 매개변수 합성을 기반으로 하는 이전 TTS 시스템과 비교 하여 신경망 기반 음성 합성의 장점은 명료도와 자연성 측면에서 높은 음성 품질과 인간의 전처리 및 기능 개발에 대한 요구 사항이 적다는 것입니다.

1.2 본 조사의 구성

본 논문에서는 그림 2 와 같이 두 부분으로 구성된 신경 TTS에 대한 연구를 주로 검토 한다.

TTS의 주요 구성 요소 최신 TTS 시스템은 텍스트 분석 모듈, 음향 모델 및 보코더의 세 가지 기본 구성 요소5로 구성됩니다. 그림 1과 같이 텍스트 분석 모듈 은 텍스트 시퀀스를 언어적 특징으로 변환하고 음향 모델은 언어적 특징에서 음향적 특징을 생성 한 다음 보코더에서 음향적 특징에서 파형을 합성합니다. 2절에서는 신경 TTS의 세 가지 구성요소에 대한 연구를 검토한다. 구체적으로 2.1절에서는 신경 TTS의 기본 구성요소에 대한 주요 분류법을 먼저 소개하고, 2절에서는 텍스트 분석, 음향 모델 및 보코더에 대한 연구를 소개한다. 2.2절, 2.3절 및 2.4절 . 섹션 2.5에서 완전 종단 간 TTS에 대한 연구를 더 소개합니다. 우리는 주로 신경 TTS의 주요 구성 요소의 분류법에 따라 연구 작업을 검토 하지만 시퀀스 생성 방법(자동 회귀 또는 비자동 회귀), 다양한 생성 모델 및 다양한 네트워크 구조를 포함한 몇 가지 다른 분류법도 섹션 2.6에서 설명합니다. .

게다가, 우리는 섹션 2.6에서 일부 대표적인 TTS 작업의 시간 변화를 설명합니다.

4TTS에서 "end-to-end"라는 용어는 모호한 의미를 가지고 있습니다. 초기 연구에서 "엔드 투 엔드"는 텍스트 투 스펙트로그램 모델이 엔드 투 엔드이지만 여전히 별도의 파형 합성기(보코더)를 사용하는 것을 의미합니다. 또한 복잡한 언어 또는 음향 특성을 사용하지 않는 신경 기반 TTS 모델을 광범위하게 참조할 수 있습니다. 예를 들어 WaveNet [254] 은 음향적 특징을 사용하지 않고 언어적 특징에서 파형을 직접 생성 하고, Tacotron [382] 은 언어적 특징을 사용하지 않고 문자나 음소에서 스펙트로그램을 직접 생성한다.
그러나 엄격한 종단 간 모델은 텍스트에서 직접 파형을 생성하는 것을 말합니다. 따라서 본 논문 에서는 TTS 모델 의 end-to-end 정도를 구분하기 위해 “end-to-end”, “more end-to-end”, “full end-to-end”를 사용한다.

5 일부 end-to-end 모델은 텍스트 분석(예: Tacotron 2 [303]), 음향 모델(예: WaveNet [254]) 또는 보코더(예: Tacotron [382])를 명시적으로 사용하지 않고 일부 시스템은 이러한 구성 요소를 사용 하는 단일 엔드-투-엔드 모델(예: FastSpeech 2s [292]) 은 현재 TTS 연구 및 제품에서 여전히 인기가 있습니다.

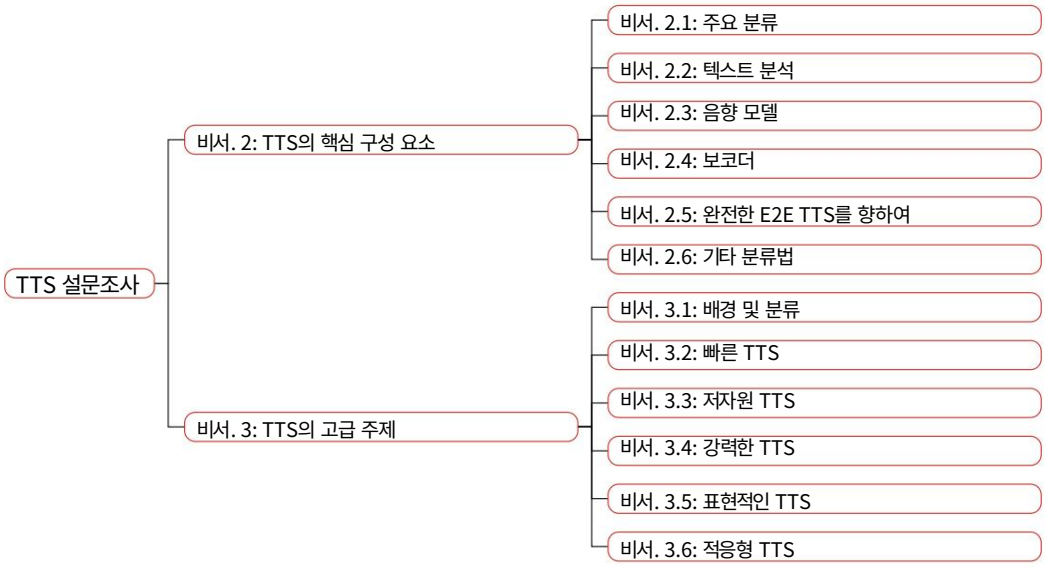


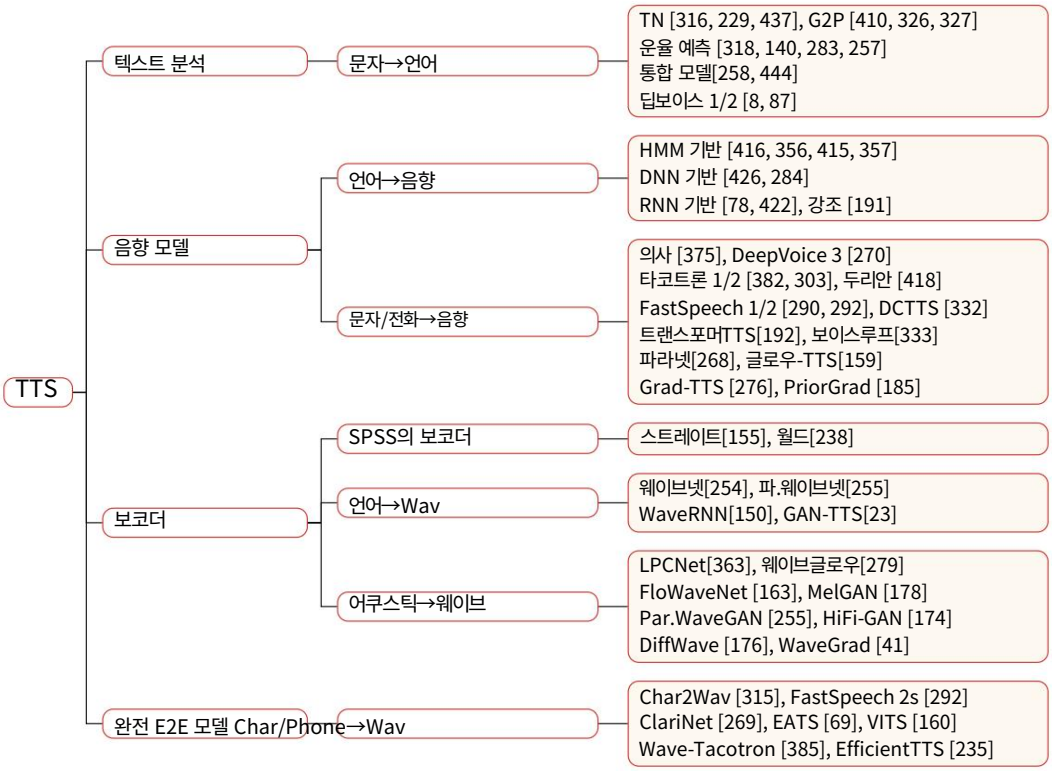
그림 2: 이 조사 보고서의 구성.

TTS의 고급 주제 신경 TTS의 핵심 구성 요소 외에도 TTS 연구의 최전선을 밀고 TTS 제품의 실제 문제를 해결하는 신경 TTS 의 몇 가지 고급 주제를 검토합니다. 예를 들어, TTS는 시퀀스 생성 작업에 대한 일반적인 시퀀스 이고 출력 시퀀스는 일반적으로 매우 길기 때문에 자동 회귀 생성 속도를 높이고 빠른 음성 합성을 위해 모델 크기를 줄이는 방법이 뜨거운 연구 주제입니다 (섹션 3.2). 좋은 TTS 시스템은 자연스럽고 이해하기 쉬운 음성을 모두 생성해야 하며 많은 TTS 연구 작업은 음성 합성의 명료도와 자연스러움을 개선하는 것을 목표로 합니다. 예를 들어 TTS 모델을 교육하는 데 필요한 데이터가 부족한 리소스가 적은 시나리오에서 합성된 음성은 명료도와 자연스러움이 모두 낮을 수 있습니다. 따라서 많은 작업이 낮은 리소스 설정에서 데이터 효율적인 TTS 모델을 구축하는 것을 목표로 합니다 (섹션 3.3). TTS 모델은 생성된 음성의 단어 건너뛰기 및 반복 문제가 음성 품질에 영향을 미치는 견고성 문제에 직면하고 있기 때문에 많은 작업 이 음성 합성의 견고성을 개선하는 것을 목표로 합니다(섹션 3.4). 자연스러움과 표현력을 향상시키기 위해 많은 작품들이 표현적인 말을 생성하기 위해 말의 스타일/운율을 모델링, 제어 및 전달합니다 (3.5절). 모든 대상 화자의 음성을 지원하도록 TTS 모델을 조정하면 TTS를 광범위하게 사용하는 데 매우 유용합니다. 따라서 제한된 적응 데이터 및 매개변수를 사용한 효율적인 음성 적응은 실제 TTS 응용 프로그램에 매우 중요합니다(섹션 3.6).

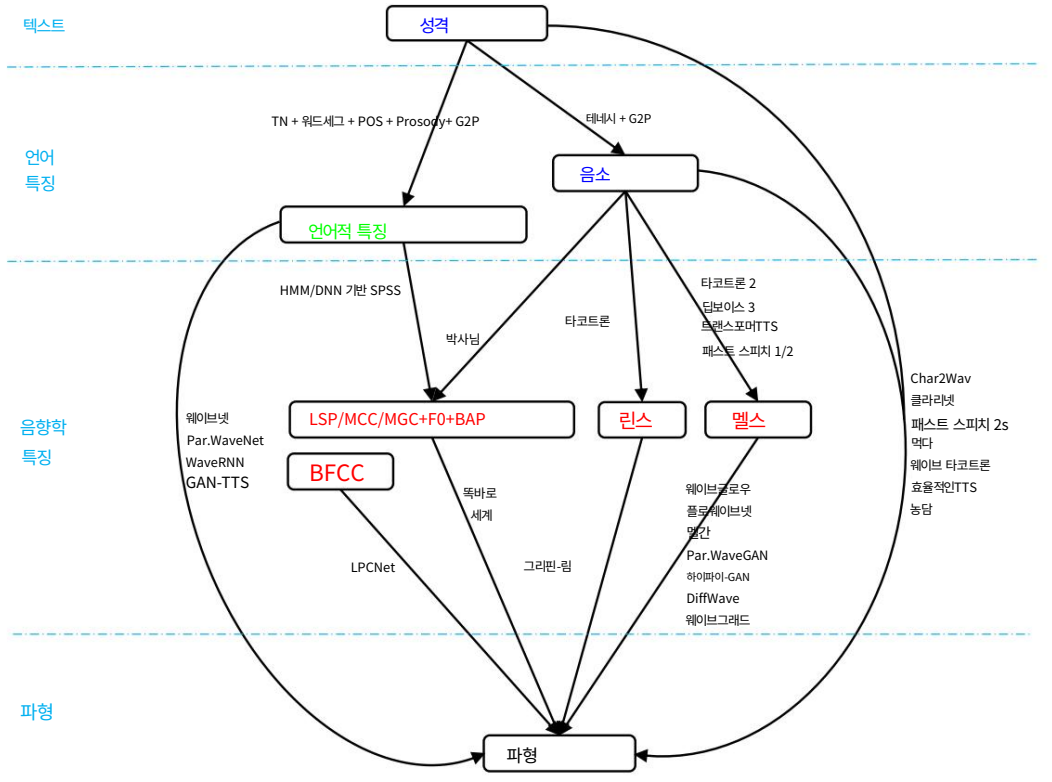
이 설문 조사를 더욱 풍부하게 하기 위해 섹션 4에 오픈 소스 구현, 말뭉치 및 기타 유용한 리소스를 포함한 TTS 관련 리소스를 요약합니다 . 섹션 5에서 이 설문 조사를 요약하고 향후 연구 방향에 대해 논의합니다.

2 TTS의 핵심 구성 요소

이 섹션에서는 신경 TTS의 핵심 구성 요소(텍스트 분석, 음향 모델 및 보코더)의 관점에서 연구 작업을 검토합니다. 먼저 섹션 2.1에서 이 관점에서 주요 분류법 을 소개한 다음 섹션 2.2, 섹션 2.3 및 섹션 2.4에서 각각 세 가지 TTS 구성 요소를 소개합니다. 또한 섹션 2.5에서 완전한 종단 간 TTS를 향한 작업을 검토합니다 . 주요 분류법 외에도 autoregressive/non-autoregressive 시퀀스 생성, 생성 모델, 네트워크 구조, TTS에 대한 대표적인 연구 작업의 타임라인과 같은 더 많은 분류법을 섹션 2.6에서 소개합니다.



(a) 신경 TTS의 분류.



(b) 데이터는 텍스트에서 파형으로 흐릅니다.

그림 3: 주요 구성 요소 및 데이터 흐름의 관점에서 본 신경 TTS의 분류.

2.1 주요 분류

그림 3a와 같이 텍스트 분석, 음향 모델, 보코더 6 및 완전 종단 간 모델과 같은 기본 TTS 구성 요소의 관점에서 주로 신경 TTS에 대한 작업을 분류합니다 . 이 분류법은 텍스트 분석 모듈에서 음소 또는 언어적 특징을 추출하고 음향 모델에서 음향 특징을 예측하는 일반적인 파이프라인을 따릅니다. 2) 음향 모델은 언어적 특징 또는 문자/음소로부터 음향적 특징을 생성합니다. 3) 보코더는 언어적 특징이나 음향적 특징에서 파형을 생성합니다. 4) 완전 종단 간 모델은 문자/음소를 파형으로 직접 변환합니다.

그림 3b와 같이 텍스트에서 파형으로의 데이터 흐름에 따라 TTS 작업을 재구성합니다. 텍스트를 음성으로 변환하는 과정에는 몇 가지 데이터 표현이 있습니다. 1)
텍스트의 원시 형식인 문자. 2) 텍스트 분석을 통해 얻은 언어적 특징으로 발음 및 운율에 대한 풍부한 컨텍스트 정보를 포함합니다. 음소는 언어 기능에서 가장 중요한 요소 중 하나이며 일반적으로 신경 기반 TTS 모델에서 텍스트를 나타내는 데 단독으로 사용됩니다. 3) 음성 파형 의 추상적인 표현인 음향 특징 . 통계적 파라메트릭 음성 합성 [416, 356, 415, 425, 357], LSP(line spectral pairs) [135], MCC(mel-cepstral coefficients) [82], MGC(mel-generalized coefficients) [355], F0 그리고 BAP(band aperiodicities) [156, 157] 는 음향적 특징으로 사용 되며 STRAIGHT [155] , WORLD [238] 와 같은 보코더를 통해 쉽게 파형으로 변환할 수 있다. 신경 기반 종단 간 TTS 모델에서 mel-스펙트로그램 또는 선형 스펙트로그램은 일반적으로 음향 특징으로 사용되며 신경 기반 보코더를 사용하여 파형으로 변환됩니다. 4) 음성 의 최종 형식인 파형 . 그림 3b에서 볼 수 있듯이 텍스트에서 파형까지 다양한 데이터 흐름이 있을 수 있습니다. 1) 문자 → 언어 특징 → 음향 특징 → 파형; 2) 문자 → 음소 → 음향 특징 → 파형; 3) 문자 → 언어적 특징 → 파형; 4) 문자 → 음소 → 음향 특징 → 파형; 5) 문자 → 음소 → 파형 또는 문자 → 파형.

2.2 텍스트 분석

TTS의 프런트엔드라고도 하는 텍스트 분석은 입력 텍스트를 발음 및 운율에 대한 풍부한 정보를 포함하는 언어적 특징으로 변환하여 음성 합성을 용이하게 합니다. 통계적 파라메트릭 합성에서 텍스트 분석은 일련의 언어 특징 벡터 [357] 를 추출하는 데 사용되며 텍스트 정규화 [316, 439], 단어 분할 [400], 품사 (POS) 태깅 과 같은 여러 기능 을 포함 합니다. [298], 운율 예측 [51], 자소 대 음소 변환 [410]. 종단간 신경 TTS 에서는 신경 기반 모델의 큰 모델링 용량으로 인해 문자 또는 음소 시퀀스를 합성을 위한 입력으로 직접 취하므로 텍스트 분석 모듈이 크게 단순화됩니다. 이 시나리오에서 문자 입력에서 표준 단어 형식을 얻으려면 텍스트 정규화가 여전히 필요하고 표준 단어 형식에서 음소를 얻으려면 자소에서 음소로의 변환이 더 필요합니다. 일부 TTS 모델은 텍스트에서 파형을 직접 생성하는 완전한 종단 간 합성을 요구 하지만 실제 사용을 위해 가능한 비표준 형식으로 원시 텍스트를 처리하려면 텍스트 정규화가 여전히 필요 합니다. 게다가 일부 종단 간 TTS 모델은 기존의 텍스트 분석 기능을 통합합니다. 예를 들어 Char2Wav [315] 및 DeepVoice 1/2 [8, 87] 는 문자에서 언어로의 기능 변환을 파이프라인으로 구현하고 순전히 신경망을 기반으로 하며 일부 작업[321] 은 텍스트 인코더로 운율 기능을 명시적으로 예측합니다. 이 하위 섹션의 나머지 부분에서는 먼저 통계적 파라메트릭 합성에서 텍스트 분석을 위한 일반적인 작업을 소개한 다음 종단 간 TTS 모델에서 텍스트 분석 개발에 대해 논의합니다.

텍스트 분석에서 대표적인 몇 가지 작업을 Table 1에 정리하고, 각 작업에 대한 대표적인 작업을 몇 가지 소개 하면 다음과 같다.

- 텍스트 정규화. 원시 서면 텍스트(비표준 단어)는 TTS 모델에서 단어를 쉽게 발음할 수 있도록 텍스트 정규화를 통해 음성 형태의 단어로 변환되어야 합니다. 예를 들어, "1989"는 "1989", "Jan. 24"는 " 1월 24일"로 정규화됩니다. 텍스트 정규화에 대한 초기 작업은 규칙 기반 [317],

6WaveNet [254] 및 WaveRNN [150] 과 같은 일부 신경 TTS 모델 은 언어 기능에서 파형을 직접 생성하기 위해 처음 도입되었습니다. 이러한 관점에서 WaveNet은 음향 모델과 보코더의 결합으로 볼 수 있습니다. 다음 작업은 일반적으로 파형을 생성하기 위해 mel-spectrograms를 입력으로 사용하여 WaveNet 및 Wav eRNN을 보코더로 활용합니다. 따라서 WaveNet/WaveRNN을 보코더로 분류하여 2.4절에서 소개한다.

표 1: 텍스트 분석의 일반적인 작업(즉, TTS 프런트엔드, 문자→언어).

일	연구 작업
텍스트 정규화	규칙 기반[317], 신경 기반[316, 229, 413, 437], 하이브리드[439]
단어 분할 [400, 451, 267]	
POS 태깅 [298, 329, 227, 451, 138]	
운율 예측 [51, 412, 318, 190, 140, 328, 283, 64, 447, 216, 218, 3]	
음소에서 음소로	N-gram [42, 24], 신경 기반 [410, 289, 33, 326]
- - 폴리폰 명확화 [448, 398, 230, 301, 327, 29, 263]	

그런 다음 신경망은 원본 및 대상 시퀀스가 각각 비표준 단어 및 구어체 단어인 시퀀스 대 시퀀스 작업으로 텍스트 정규화를 모델링하는 데 활용됩니다 [316, 229, 437]. 최근 일부 작업 [439] 은 텍스트 정규화의 성능을 더욱 향상시키기 위해 규칙 기반 및 신경 기반 모델 의 장점을 결합할 것을 제안합니다 . • 단어 분할. 중국어와 같은 문자 기반 언어 의 경우 원시 텍스트에서 단어 경계를 감지하기 위해 단어 분할 [400, 451, 267] 이 필요합니다. 전환 과정. • 품사 태깅. 명사, 동사, 전치사 등 각 단어의 품사(POS) 도 TTS에서 자소-음소 변환 및 운율 예측에 중요합니다. 여러 작업 에서 음성 합성에서 POS 태깅을 조사했습니다[298, 329, 227, 451, 138].

• 운율 예측. 음성의 리듬, 강세, 억양과 같은 운율 정보는 음절 길이, 음량, 음높이의 변화에 해당하며, 이는 인간의 음성 의사소통에서 중요한 지각적 역할을 합니다. 운율 예측은 태깅 시스템을 사용하여 각 종류의 운율에 레이블을 지정합니다. 언어마다 운율 태깅 시스템과 도구 가 다릅니다 [307, 294, 345, 112, 249]. 영어의 경우 ToBI(tones and break indices) [307, 294] 는 인기 있는 태깅 시스템으로 성조(예: 피치 악센트, 구문 악센트 및 경계 음) 및 중단(단어 사이의 중단이 얼마나 강한지) 에 대한 태그를 설명합니다.). 예를 들어, “Mary 는 가게에 갔다 ?”라는 문장에서 “Mary”와 “store”를 강조할 수 있으며 이 문장은 어조를 높입니다. 많은 작업 [318, 190, 140, 283] 은 ToBI를 기반으로 운율 태그를 예측하기 위해 다양한 모델과 기능을 조사합니다 . 중국어 음성 합성의 경우 일반적인 운율 경계 레이블은 운율 단어(PW), 운율 구(PPH) 및 억양 구(IPH)로 구성 되며 3계층 계층 운율 트리 를 구성할 수 있습니다 [51, 328, 64]. 일부 작업 [51, 3, 64, 328, 216, 218] 은 CRF [180], RNN [114], self-attention [368] 과 같은 다양한 모델 구조를 조사 하여 중국어 운율 예측을 수행합니다. • 자소-음소(G2P) 변환. 문자(문자소)를 발음 (음소)으로 변환하면 음성 합성이 훨씬 쉬워집니다. 예를 들어, "speech"라는 단어는 "s p iy ch"로 변환됩니다. 수동으로 수집된 자소에서 음소로의 어휘집은 일반적으로 변환에 활용됩니다.

그러나 영어와 같은 알파벳 언어의 경우 어휘집이 모든 단어의 발음을 다룰 수 없습니다. 따라서 영어의 G2P 변환은 주로 어휘에 없는 단어의 발음을 생성하는 역할을 합니다 [42, 24, 410, 289, 33, 326]. 중국어와 같은 언어의 경우 어휘가 거의 모든 문자를 다룰 수 있지만 문자 의 문맥에 따라 결정될 수 있는 다음어가 많이 있습니다7 . 따라서 이러한 종류의 언어에서 G2P 변환은 현재 단어 문맥을 기반으로 적절한 발음을 결정하는 다성어 명확화를 주로 담당합니다[448, 398, 230, 301, 327, 29, 263].

위의 텍스트 분석 후 언어 기능을 추가로 구성하고 이를 SPSS 또는 보코더의 음향 모델과 같은 TTS 파이프라인의 후반부에 대한 입력으로 사용할 수 있습니다 [254]. 일반적으로 우리는 음소, 음절, 단어, 구 및 문장 수준을 포함한 다양한 수준의 텍스트 분석 결과를 집계하여 언어적 특징을 구성 할 수 있습니다[357].

토론 텍스트 분석은 SPSS에 비해 신경 TTS에서 덜 주목받는 것처럼 보이지만 다양한 방식으로 신경 TTS에 통합되었습니다. 1) 다중 작업 및 통합 프런트엔드 모델. 최근 Pan et al. [258], 장 외. 다중 작업 패러다임에서 텍스트 분석의 모든 작업을 포괄하고 좋은 결과를 얻을 수 있도록 통합 모델을 설계 합니다. 2) 운율 예측. 작시법

7영어를 포함한 많은 언어에는 폴리폰이 있습니다. 예를 들어, 영어에서 "resume"은 "ri zju:m"(중단 후 계속하거나 계속한다는 의미) 또는 "rezjumei"(커리큘럼 이력을 의미)로 발음될 수 있습니다.

음성 합성의 자연스러움에 중요합니다. 신경 TTS 모델은 텍스트 분석 모듈을 단순화하지만 피치 [292], 지속 시간 [290], 구 나누기 [206], 호흡 또는 채워진 일시 중지 [404] 예측 과 같은 운율 예측을 위한 일부 기능은 텍스트 인코더에 통합됩니다 . TTS 모델의 텍스트(문자 또는 음소) 인코더 위에 구축 됩니다. 운율 기능 을 통합하는 다른 방법 에는 1) 참조 음성에서 운율 표현을 학습하는 참조 인코더; 2) 자기 지도 사전 훈련을 통해 암묵적인 운율 정보로 좋은 텍스트 표현을 학습하는 텍스트 사전 훈련 [104, 98]; 3) 그래프 네트워크[208]와 같은 전용 모델링 방법 을 통해 구문 정보를 통합합니다 .

2.3 음향 모델

이 섹션에서는 언어적 특징 또는 음소 또는 문자에서 직접 음향 특징을 생성하는 음향 모델에 대한 작업을 검토합니다. TTS의 발전과 함께 SPSS(Statistical Parametric Speech Synthesis) 의 초기 HMM 및 DNN 기반 모델 [416, 356, 426, 284, 78, 424]을 포함하여 다양한 종류의 음향 모델이 채택되었습니다 . 인코더-어텐션-디코더 프레임워크(LSTM, CNN 및 셀프 어텐션 포함) [382, 303, 270, 192] 및 최신 피드포워드 네트워크(CNN 또는 셀프 어텐션) [290, 268]에 기반한 시퀀스 모델 세대.

음향 모델은 보코더를 사용하여 파형으로 추가 변환되는 음향 특징을 생성하는 것을 목표로 합니다. 음향 기능의 선택은 주로 TTS 파이프라인의 유형을 결정합니다. mel-cepstral coefficients(MCC) [82], mel generalized coefficients(MGC) [355], BAP(band aperiodicity) [156, 157], 기본 주파수 (F0), 유성음 /무성(V/UV), BFCC(bark-frequency cepstral coefficients) 및 가장 널리 사용되는 mel-스펙트로그램. 따라서 음향 모델을 두 가지 기간으로 나눌 수 있습니다. 1) 일반적으로 언어적 특징에서 MGC, BAP 및 F0와 같은 음향 특징을 예측하는 SPSS의 음향 모델 과 2) 신경 기반 종단간 TTS의 음향 모델, 음소 또는 문자에서 mel-스펙트로그램과 같은 음향 특징을 예측 합니다.

2.3.1 SPSS의 음향 모델

SPSS [425, 357]에서 HMM [416, 356], DNN [426, 284] 또는 RNN [78, 424] 과 같은 통계 모델 을 활용하여 언어적 특징에서 음향 특징(음성 매개변수)을 생성합니다. 파라미터는 STRAIGHT [155] 및 WORLD [238] 와 같은 보코더를 사용하여 음성 파형으로 변환됩니다 . 이러한 음향 모델의 개발은 몇 가지 고려 사항에 의해 추진됩니다. 1) 더 많은 컨텍스트 정보를 입력으로 사용합니다. 2) 출력 프레임 간의 상관 관계를 모델링합니다 . 3) 언어 특징에서 음향 특징으로 의 매핑이 일대다이기 때문에 과도하게 평활화되는 예측 문제 [425]에 더 잘 대처합니다. 다음과 같이 몇 가지 작품을 간략하게 검토 합니다.

HMM [286] 은 Yoshimura et al.에서 음성 매개변수를 생성하는 데 활용됩니다. [416], Tokuda et al. [356] 여기서 HMM의 관측 벡터는 MCC(mel cepstral coefficients) 및 F0 와 같은 스펙트럼 매개변수 벡터로 구성됩니다 . 이전의 연결 음성 합성에 비해 HMM 기반 파라메트릭 합성은 화자의 정체성, 감정 및 말하는 스타일 을 변경하는 데 더 유연합니다 [356]. 독자는 Zen [422], Zen et al. [425], Tokuda et al. [357] HMM 기반 SPSS의 장점과 단점에 대한 몇 가지 분석 . HMM 기반 SPSS의 주요 단점 중 하나 는 합성된 음성의 품질이 충분하지 않다는 것 [425, 357], 주로 두 가지 이유 때문입니다. 1) 음향 모델의 정확도가 좋지 않고 예측된 음향 특성이 스무딩 및 디테일 부족, 2) 보코딩 기술이 충분하지 않습니다. 첫 번째 이유 는 주로 HMM의 모델링 역량이 부족하기 때문입니다. 따라서 SPSS에서는 HMM 기반 모델의 합성 품질을 향상시키는 DNN 기반 음향 모델 [426] 이 제안된다. 나중에 음성 발화에서 긴 시간 범위의 컨텍스트 효과를 더 잘 모델링하기 위해 LSTM 기반 반복 신경망 [78] 을 활용하여 컨텍스트 종속성을 더 잘 모델링합니다. 딥 러닝의 발전으로 CBHG [382] 와 같은 일부 고급 네트워크 구조 를 활용하여 음향 특성을 더 잘 예측합니다 [191]. VoiceLoop [333] 는 음운 루프라는 작업 메모리를 채택하여 음소 시퀀스에서 음향 특징(예: F0, MGC, BAP)을 생성한 다음 WORLD [238] 보코더를 사용하여 이 음향 특징에서 파형을 합성합니다. Yang et al. GAN [90] 을 활용 하여 음향 특징의 생성 품질을 개선합니다. Wang et al. 프레임별 정렬을 피할 수 있는 음소 시퀀스에서 음향 특징을 직접 생성 하기 위해 주의 기반 반복 시퀀스 트랜스듀서 모델을 활용하는 보다 종단 간 방식을 탐색합니다.

이전 신경망 기반 음향 모델에서 필요했습니다. Wang et al. 다양한 음향 모델에 대한 철저한 실험적 연구를 수행합니다. SPSS의 일부 음향 모델은 표 2에 요약되어 있습니다.

표 2: 음향 모델 및 해당 특성 목록. "Ling"은 언어 특징, "Ch"는 문자, "Ph"는 음소, "MCC"는 mel-cepstral 계수 [82], "MGC"는 mel-generalized 계수 [355], " BAP"는 밴드 비 주기 [156, 157], "LSP"는 라인 스펙트럼 쌍 [135], "LinS"는 선형 스펙트로그램, "MelS"는 멜 스펙트로그램을 나타냅니다. "NAR*"은 모델 이 비자동회귀 구조에서 자동회귀 구조를 사용하고 완전히 병렬이 아님을 의미합니다.

음향 모델	입력→출력	AR/NAR 모델링 구조		
HMM 기반 [416, 356]	링→MCC+F0	/	/	흠 /
DNN 기반 [426]	성별→MCC+BAP+F0 NAR			DNN /
LSTM 기반 [78]	링→LSP+F0	와 함께		RNN /
강조 [191]	링→린스+모자+F0 AR			잡종
의사 [375]	Ph→LSP+BAP+F0 AR		Seq2Seq RNN / 하이브리	
보이스루프 [333]	Ph→MGC+BAP+F0 AR		드	
타코트론 [382]	Ch→LinS	와 함께	Seq2Seq 하이브리드/RNN	
타코트론 2 [303]	Ch→MelS	와 함께	Seq2Seq RNN	
두리안 [418]	Ph→MelS	와 함께	Seq2Seq RNN /	
무무 타코트론 [304]	Ph→MelS	와 함께		하이브리드/CNN/RNN /
에게. 타코트론 1/2 [74, 75]	Ph→MelS	NAR		하이브리드/셀프-Att/CNN /
멜넷 [367]	Ch→MelS	와 함께		RNN
딥보이스 [8]	Ch/Ph→MelS	와 함께	/	CNN /
딥보이스 2 [87]	Ch/Ph→MelS	와 함께		CNN
딥보이스 3 [270]	Ch/Ph→MelS	와 함께	Seq2Seq CNN	
파라넷 [268]	Ph→MelS	NAR	Seq2Seq CNN	
DCTTS [332]	Ch→MelS	와 함께	Seq2Seq CNN /	
스피디스피치 [361]	Ph→MelS	NAR		CNN /
토크넷 1/2 [19, 18]	Ch→MelS	NAR		CNN
트랜스포머TTS [192]	Ph→MelS	와 함께	Seq2Seq 자가 공격	
멀티스피치 [39]	Ph→MelS	와 함께	Seq2Seq 자가 공격	
패스트스피치 1/2 [290, 292]	Ph→MelS	NAR	Seq2Seq 자가 공격	
정렬TTS [429]	Ch/Ph→MelS	NAR	Seq2Seq 자가 공격	
JDI-T [197]	Ph→MelS	NAR	Seq2Seq 자가 공격	
패스트피치 [181]	Ph→MelS	NAR	Seq2Seq 자가 공격	
AdaSpeech 1/2/3 [40, 403, 404]	Ph→MelS	NAR	Seq2Seq 자가 공격	
데노이스피치 [434]	Ph→MelS	NAR	Seq2Seq 자가 공격 /	
장치TTS [126]	Ph→MelS	NAR		하이브리드/DNN/RNN /
라이트스피치 [220]	Ph→MelS	NAR		하이브리드/셀프 공격/CNN
흐름-TTS [234]	Ch/Ph→MelS	NAR* 흐름		하이브리드/CNN/RNN
글로우-TTS [159]	Ph→MelS	NAR	흐름	하이브리드/셀프 공격/CNN
플로트론 [366]	Ph→MelS	와 함께	흐름	하이브리드/RNN
효율적인TTS [235]	Ch→MelS	NAR	흐름	하이브리드/CNN
GMVAE-타코트론 [119]	Ph→MelS	와 함께	피트	하이브리드/RNN
VAE-TTS [443]	Ph→MelS	와 함께	피트	하이브리드/RNN
BVAE-TTS [187]	Ph→MelS	NAR	피트	CNN
GAN 노출 [99]	Ph→MelS	와 함께	하지만	하이브리드/RNN
TTS-스타일화 [224]	Ch→MelS	와 함께	하지만	하이브리드/RNN
다중 SpectroGAN [186]	Ph→MelS	NAR	하지만	하이브리드/셀프 공격/CNN
차이-TTS [141]	Ph→MelS	NAR* 확산 하이브리드/CNN		
대학원-TTS [276]	Ph→MelS	NAR	확산 하이브리드/Self-Att/CNN	
프라이어그라드 [185]	Ph→MelS	NAR	확산 하이브리드/Self-Att/CNN	

2.3.2 종단 간 TTS의 음향 모델

신경 기반 엔드-투-엔드 TTS의 음향 모델은 SPSS에 비해 몇 가지 장점이 있습니다. 1) 기존의 음향 모델은 언어와 음향 특징 사이의 정렬이 필요한 반면 시퀀스 대 시퀀스 기반 신경 모델은 주의를 통해 정렬을 암묵적으로 학습하거나 예측할 수 있습니다. 지속 시간을 합치면 보다 종단 간이며 사전 처리가 덜 필요합니다. 2) 신경망의 모델링 능력이 증가함에 따라 언어적 특징은 문자 또는 음소 시퀀스만으로 단순화되고 음향적 특징은 저차원 및 압축된 켑스트럼(예: MGC)에서 고차원 멜 스펙트로그램 또는 심지어 더 높은 차원의 선형 스펙트로그램. 다음 단락에서는 신경 TTS8 의 대표적인 음향 모델을 소개합니다.

, 표 2에 포괄적인 음향 모델 목록을 제공합니다.

RNN 기반 모델(예: Tacotron 시리즈) Tacotron [382] 은 인코더 주의 디코더 프레임워크를 활용하고 문자를 입력9으로 취하고 선형 스펙트로그램을 출력하며 Griffin Lim 알고리즘 [95] 을 사용하여 파형을 생성합니다. Tacotron 2 [303] 는 추가 WaveNet [254] 모델 을 사용하여 mel-스펙트로 그램을 생성하고 mel-스펙트로그램을 파형으로 변환하기 위해 제안되었습니다 . Tacotron 2 는 Concatenative TTS, Parametric TTS, Tacotron과 같은 신경 TTS 를 포함한 이전 방법보다 음성 품질을 크게 향상시킵니다 .

나중에 많은 작업이 Tacotron을 다양한 측면에서 개선했습니다. 1) GST-Tacotron [383] 및 Ref-Tacotron [309] 과 같은 음성 합성의 표현력을 향상시키기 위해 참조 인코더 및 스타일 토큰을 사용합니다 . 2) Tacotron에서 주의 메커니즘을 제거하고 대신 DurIAN [418] 및 Non-attentative Tacotron [304] 과 같은 자동 회귀 예측을 위한 기간 예측기를 사용합니다. 삼)

Parallel Tacotron 1/2 [74, 75] 와 같이 Tacotron의 자동회귀 생성을 비자동회귀 생성으로 변경합니다 . 4) Wave-Tacotron[385] 과 같은 Tacotron을 기반으로 종단 간 텍스트-파형 모델 구축 .

CNN 기반 모델(예: DeepVoice 시리즈) DeepVoice [8] 는 실제로 컨볼루션 신경망으로 강화된 SPSS 시스템입니다. 신경망을 통해 언어적 특징을 얻은 후 DeepVoice는 WaveNet [254] 기반 보코더를 활용하여 파형을 생성합니다. DeepVoice 2 [87] 는 DeepVoice의 기본 데이터 변환 흐름을 따르며 향상된 네트워크 구조 및 다중 화자 모델링 으로 DeepVoice를 향상시킵니다 . 또한 DeepVoice 2는 먼저 Tacotron을 사용하여 선형 스펙트로그램을 생성한 다음 WaveNet을 사용하여 파형을 생성하는 Tacotron + WaveNet 모델 파이프라인을 채택합니다. DeepVoice 3 [270] 는 문자로부터 mel-스펙트로그램을 생성하고 실제 다중 화자 데이터 세트 로 확장할 수 있는 음성 합성 을 위한 완전 컨벌루션 네트워크 구조를 활용합니다 . DeepVoice 3는 보다 간결한 sequence-to-sequence 모델을 사용하고 복잡한 언어 기능 대신 mel-스펙트로그램을 직접 예측함으로써 이전 DeepVoice 1/2 시스템보다 개선되었습니다.

나중에 ClariNet [269] 은 완전히 종단 간 방식으로 텍스트에서 파형을 생성하도록 제안되었습니다.
ParaNet [268] 은 멜 스펙트로그램 생성 속도를 높이고 합리적으로 우수한 음성 품질을 얻을 수 있는 완전 컨벌루션 기반 비자동회귀 모델입니다. DCTTS [332] 는 Tacotron 과 유사한 데이터 변환 파이프라인을 공유하고 문자 시퀀스에서 mel-스펙트로그램을 생성하기 위해 완전 컨벌루션 기반 인코더-어텐션 디코더 네트워크를 활용합니다. 그런 다음 스펙트로그램 초해상도 네트워크를 사용하여 선형 스펙트로그램을 얻고 Griffin Lim[95]을 사용하여 파형을 합성합니다.

트랜스포머 기반 모델(예: FastSpeech 시리즈) TransformerTTS [192] 는 트랜스 포머 [368] 기반 인코더-어텐션-디코더 아키텍처를 활용하여 멜 스펙트로그램을 생성합니다.

8RNN, CNN, Transformer(self-attention) 등 네트워크 구조에 따른 음향 모델을 주로 검토하고, autoregressive-based, flow-based, GAN-based, Diffusion-based 등의 generative model에 따라 vocoder를 검토한다. , 섹션 2.4에 나와 있습니다. 그러나 이것이 유일한 관점은 아닙니다. 음향 모델도 다른 생성 모델을 다루는 반면 보코더도 다른 네트워크 구조를 다루기 때문입니다.

9뉴럴 TTS에서는 문자나 음소 중 하나를 입력으로 사용하지만 주로 다음 두 가지 고려 사항으로 인해 명시적으로 구분하지 않습니다. 1) 제품 사용에 대한 높은 발음 정확도를 보장하기 위해 특히 자소와 음소는 큰 차이가 있습니다.
2) 문자를 직접 입력하는 모델의 경우 문자 입력을 위한 특별한 설계가 없기 때문에 문자를 음소로 쉽게 변환할 수 있다. 데이터 희소성 문제를 해결하기 위해 문자와 음소의 혼합 표현을 입력으로 사용 하는 일부 작업 [270, 268, 154] 이 있음을 언급할 가치가 있습니다.

음소. 그들은 Tacotron 2와 같은 RNN 기반 인코더-어텐션-디코더 모델이 다음 두 가지 문제를 겪고 있다고 주장 합니다 . 추론에서 병렬일 수 없으며, 이는 훈련과 추론 모두에서 효율성에 영향을 미칩니다. 2) 일반적으로 텍스트 및 음성 시퀀스가 매우 길기 때문에 RNN은 이러한 시퀀스의 긴 종속성을 모델링하는 데 적합하지 않습니다. TransformerTTS 는 Transformer 의 기본 모델 구조를 채택하고 디코더 pre-net/post-net 및 정지 토큰 예측과 같은 Tacotron 2의 일부 설계를 흡수합니다 . Tacotron 2와 비슷한 음성 품질을 달성하지만 훈련 시간이 더 빠릅니다. 그러나 위치에 민감한 어텐션과 같은 안정적인 어텐션 메커니즘을 활용하는 Tacotron과 같은 RNN 기반 모델과 비교할 때 Transformer 의 인코더-디코더 어텐션은 병렬 계산으로 인해 강력하지 않습니다. 따라서 일부 연구 에서는 Transformer 기반 음향 모델의 견고성 을 향상시킬 것을 제안합니다 . 예를 들어, MultiSpeech [39] 는 인코더 정규화, 디코더 병목 현상 및 대각선 주의 제약을 통해 주의 메커니즘의 견고성을 증명하고 RobuTrans [194] 는 지속 시간 예측을 활용 하여 자동 회귀 생성의 견고성을 향상시킵니다.

Tacotron 1/2 [382, 303], DeepVoice 3 [270] 및 TransformerTTS [192] 와 같은 이전의 신경 기반 음향 모델은 모두 자동 회귀 생성을 채택하며 몇 가지 문제가 있습니다. 1)
느린 추론 속도. 자동 회귀 mel-스펙트로그램 생성은 특히 긴 음성 시퀀스의 경우 느립니다(예: 1초 음성의 경우 홉 크기가 10ms인 경우 거의 500프레임의 mel-스펙트로그램이 있으며, 이는 긴 시퀀스임). 2) 강력한 문제. 생성된 음성에는 일반적으로 많은 단어 건너뛰기 및 반복 문제가 있으며, 이는 주로 인코더-어텐션-디코더 기반 자동 회귀 생성에서 텍스트와 mel-스펙트로그램 간의 부정확한 어텐션 정렬 로 인해 발생합니다. 따라서 이러한 문제를 해결하기 위해 FastSpeech [290] 가 제안됩니다. 1) 병렬로 멜 스펙트로그램을 생성하기 위해 피드 포워드 Transformer 네트워크를 채택하여 추론 속도를 크게 높일 수 있습니다.

2) 단어 건너뛰기 및 반복 문제를 방지하고 견고성을 향상시키기 위해 텍스트와 음성 사이의 주의 메커니즘을 제거합니다. 대신 길이 조절기를 사용하여 음소와 mel-스펙트로그램 시퀀스 사이의 길이 불일치를 연결합니다. 길이 조절기는 길이 예측기를 활용하여 각 음소의 길이를 예측하고 음소 길이에 따라 음소 숨겨진 시퀀스를 확장합니다. 여기서 확장된 음소 숨겨진 시퀀스는 멜 스펙트로그램 시퀀스의 길이와 일치하고 병렬 생성을 용이하게 할 수 있습니다. FastSpeech는 다음과 같은 몇 가지 장점 을 가지고 있습니다 [290]. 1) 매우 빠른 추론 속도(예: 멜 스펙트로그램 생성 시 270x 추론 속도 향상, 파형 생성 시 38x 속도 향상); 2) 단어 건너뛰기 및 반복 문제 가 없는 강력한 음성 합성 ; 3) 이전의 자동 회귀 모델과 동등한 음성 품질. FastSpeech 는 Azure TTS11 의 모든 언어와 로케일 을 지원하기 위해 Microsoft Azure Text to Speech Service10에 배포되었습니다 .

FastSpeech 는 mel-spectrograms의 길이와 일치하도록 숨겨진 음소 시퀀스를 확장하기 위해 명시적 기간 예측자를 활용합니다. 기간 예측기를 훈련하기 위해 기간 레이블을 얻는 방법은 생성된 음성의 운율 및 품질에 중요합니다. 섹션 3.4.2에서 기간 예측이 포함된 TTS 모델을 간략하게 검토합니다 . 다음에는 FastSpeech 를 기반으로 하는 몇 가지 다른 개선 사항을 소개합니다 . FastSpeech 2 [292] 는 주로 두 가지 측면에서 FastSpeech를 더욱 향상시키기 위해 제안 되었습니다 . 이는 FastSpeech의 2단계 교사-학생 증류 파이프라인 을 단순화하고 증류 후 대상 mel-스펙트로그램에서 정보 손실을 방지합니다. 2)

디코더 입력으로 피치, 지속 시간 및 에너지와 같은 더 많은 분산 정보를 제공 하여 텍스트에서 음성 으로의 일대다 매핑 문제 [139, 84, 382, 456] 를 완화합니다12. FastSpeech 2는 FastSpeech 보다 더 나은 음성 품질을 달성하고 FastSpeech13에서 빠르고 강력하며 제어 가능한 음성 합성의 이점을 유지합니다. FastPitch [181] 는 피치 정보 를 디코더 입력으로 사용하여 FastSpeech를 개선하며 FastSpeech 2의 분산 예측기와 유사한 아이디어를 공유합니다.

기타 음향 모델(예: Flow, GAN, VAE, Diffusion) 위의 음향 모델 외에도 표 2와 같이 많은 다른 음향 모델 [367, 22, 126, 187, 55]이 있습니다. 흐름 기반 모델 오랫동안 신경 TTS에서 사용되었습니다. 보코더(예:

10https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/ 11https://techcommunity.microsoft.com/t5/azure-ai/neural-text-to-speech- extends-support-to-15-more languages-with/ba-p/1505911 12TTS의 일대다 매핑은 하나의 언어에 해당하는 가능한 음성 시퀀스가 여러 개 있음을 나타냅니다.
음높이, 길이, 음량, 운율 등 음성의 변화로 인한 텍스트 시퀀스

13FastSpeech 2s [292] 는 FastSpeech 2와 함께 제안된다. 완전한 end-to-end text-to-waveform 모델이므로 2.5절에서 소개한다.

Parallel WaveNet [255], WaveGlow [279], FloWaveNet [163]), 흐름 기반 모델은 자동 회귀 흐름 기반 멜 스펙트로그램 생성 모델인 Flowtron [366] , Flow-TTS 와 같은 음향 모델에도 적용됩니다. [234] 및 비 자동 회귀 mel-스펙트로그램 생성 을 위한 생성 흐름을 활용하는 Glow-TTS [159] . 흐름 기반 모델 외에도 다른 생성 모델 도 음향 모델에 활용되었습니다. 예를 들어, 1) GMVAE-Tacotron [119], VAE-TTS [443] 및 BVAE-TTS [187] 는 VAE [168]를 기반으로 합니다. 2) GAN 노출 [99], TTS-Stylization [224] 및 Multi-SpectroGAN [186] 은 GAN [90]을 기반으로 합니다. 3) Diff-TTS [141], Grad-TTS [276], PriorGrad [185]는 확산 모델[310, 113]을 기반으로 한다.

표 3: 보코더 및 해당 특성 목록.

보코더	입력	AR/NAR 모델링 아키텍처		
웨이브넷 [254]	언어 기능 AR /			CNN
샘플RNN [233]		와 함께		RNN
웨이브RNN [150]	언어 기능 AR			RNN
LPCNet [363]	BFCC	와 함께		RNN
대학 WaveRNN [215] Mel-Spectrogram AR SC-WaveRNN [265]				RNN
	Mel-Spectrogram AR MB			RNN
WaveRNN [418]	Mel-Spectrogram AR FFTNet			RNN
[145]	AR 센터		////////	CNN
평가. 웨이브넷 [255]	언어적 특징 NAR		흐름	CNN
웨이브글로우 [279]	멜 스펙트로그램 NAR		흐름	하이브리드/CNN
플로웨이브넷 [163]	멜 스펙트로그램 NAR		흐름	하이브리드/CNN
웨이브플로우 [271]	멜 스펙트로그램 AR		흐름	하이브리드/CNN
스퀴즈웨이브 [433]	멜 스펙트로그램 NAR		흐름	CNN
웨이브GAN [68]	/	NAR	하지만	CNN
겔프 [149]	멜 스펙트로그램 NAR		하지만	CNN
GAN-TTS [23]	언어적 특징 NAR		하지만	CNN
멜간 [178]	멜 스펙트로그램 NAR		하지만	CNN
평가. 웨이브GAN [402]	멜 스펙트로그램 NAR		하지만	CNN
하이파이-GAN [174]	멜 스펙트로그램 NAR		하지만	하이브리드/CNN
복건 [408]	멜 스펙트로그램 NAR		하지만	CNN
검정고시 [96]	언어적 특징 NAR		하지만	CNN
프레간 [161]	멜 스펙트로그램 NAR		하지만	CNN
웨이브 다리 [268]	멜 스펙트로그램 NAR		피트	CNN
웨이브그래드 [41]	멜 스펙트로그램 NAR		확산 하이브리드/CNN	
디프웨이브 [176]	멜 스펙트로그램 NAR		확산 하이브리드/CNN	
프라이어그라드 [185]	멜 스펙트로그램 NAR		확산 하이브리드/CNN	

2.4 보코더

보코더의 발전은 대략적으로 SPSS(Statistical Parametric Speech Synthesis)에 사용되는 보코더 [155, 238, 3]와 신경망 기반 보코더 [254, 315, 150, 279, 163]의 두 단계로 분류할 수 있다.]. SPSS에서 인기 있는 일부 보코더로는 STRAIGHT [155] 및 WORLD [238]가 있습니다. 보코더 분석 및 보코더 합성 단계 로 구성된 WORLD 보코더를 예로 들어 보겠습니다 . 보코더 분석에서는 음성을 분석하고 mel-cepstral 계수 [82], 대역 비주기성 [156, 157] 및 F0 과 같은 음향 특성을 얻습니다. 보코더 합성에서는 이러한 음향 기능에서 음성 파형을 생성합니다. 이 섹션에서는 음성 품질이 높은 신경 기반 보코더에 대한 작업을 주로 검토합니다 .

WaveNet [254, 255], Char2Wav [315], WaveRNN [150] 과 같은 초기 신경 보코더 는 언어 특징을 직접 입력으로 받아 파형을 생성합니다. 나중에 Prenger et al. [279], Kim et al. [163], Kumar et al. [178], Yamamoto et al. mel- 스펙트로그램을 입력으로 취하고 파형을 생성한다. 음성 파형이 매우 길기 때문에 자동 회귀 파형 생성에는 많은 추론 시간이 걸립니다 . 따라서 Flow [65, 169, 167], GAN [90], VAE [168] 및 DDPM(Denoising Diffusion Probabilistic Model, 줄여서 Diffusion) [310, 113] 과 같은 생성 모델이 파형 생성에 사용됩니다.

따라서 신경 보코더를 여러 범주로 나눕니다. 1) 자동 회귀 보코더,

2) 흐름 기반 보코더, 3) GAN 기반 보코더, 4) VAE 기반 보코더, 5) 확산 기반 보코더. 대표적인 보코더를 표 3에 나열하고 다음과 같이 설명합니다.

Autoregressive Vocoders WaveNet [254] 은 최초의 신경 기반 보코더로 확장된 컨벌루션을 활용하여 파형 포인트를 자동 회귀적으로 생성합니다. SPSS [82, 355, 156, 135, 155, 238] 의 보코더 분석 및 합성과 달리 WaveNet은 오디오 신호에 대한 사전 지식을 거의 통합하지 않으며 전적으로 종단 간 학습에 의존합니다. 원래의 WaveNet과 WaveNet을 보코더 [8, 87]로 활용하는 일부 후속 작업은 언어 기능에 따라 음성 파형을 생성 하는 반면 WaveNet은 선형 스펙트로그램 [87] 및 멜 스펙트로그램 [336] 에 대한 조건에 쉽게 적응할 수 있습니다. , 270, 303]. WaveNet은 좋은 음성 품질을 달성하지만 추론 속도가 느립니다. 따라서 많은 작업 [256, 117, 233] 이 가볍고 빠른 보코더를 조사합니다. SampleRNN [233] 은 무조건적인 파형 생성을 위해 계층적 순환 신경망을 활용하고 Char2Wav [315] 에 통합되어 음향 특징에 따라 파형 조건을 생성합니다. 또한 WaveRNN [448] 은 효율적인 오디오 합성을 위해 개발되었으며 순환 신경망을 사용하고 이중 소프트맥스 레이어, 가중치 가지치기 및 하위 스케일링 기술을 포함한 여러 디자인을 활용하여 계산을 줄입니다. Lorenzo-Trueba et al. [215], 폴 외. [265], Jiao et al. 보코더 의 견고성과 보편성을 더욱 향상시킵니다. LPC Net [363, 364] 은 기존의 디지털 신호 처리를 신경망에 도입하고 선형 예측 계수를 사용하여 다음 파형 포인트를 계산하는 동시에 경량 RNN을 활용하여 잔차를 계산합니다. LPCNet은 BFCC(bark-frequency cepstral coefficients) 기능에 따라 조정된 음성 파형을 생성하며 mel-spectrograms에 대한 조건에 쉽게 적응할 수 있습니다. 다음 작업 은 속도 향상을 위한 복잡성 감소[370, 275, 151] 및 더 나은 품질을 위한 안정성 향상[129] 과 같은 다양한 관점에서 LPCNet을 더욱 개선 합니다.

흐름 기반 보코더 흐름 정규화 [65, 66, 293, 169, 167] 는 일종의 생성 모델입니다. 가역 매핑 시퀀스로 확률 밀도를 변환합니다 [293]. 변수의 변화 규칙에 기반한 일련의 가역 매핑을 통해 표준/정규화된 확률 분포(예: Gaussion)를 얻을 수 있으므로 이러한 종류의 흐름 기반 생성 모델 을 정규화 흐름이라고 합니다. 샘플링하는 동안 이러한 변환의 역을 통해 표준 확률 분포에서 데이터를 생성합니다 . 신경 TTS에 사용되는 흐름 기반 모델 은 두 가지 기술 [262] 에 따라 두 가지 범주 로 나눌 수 있습니다. 1) 자동 회귀 변환 [169] (예: Parallel WaveNet [255] 에서 사용되는 역 자동 회귀 흐름) 이분 변환(예: WaveGlow [279] 에서 사용되는 Glow [167] 및 FloWaveNet [163] 에서 사용되는 RealNVP [66]), 표 4 에 나와 있습니다.

표 4: 몇 가지 대표적인 흐름 기반 모델 및 공식[271]. (엑스)

흐름		평가 $z = f^{-1}$	합성 $x = f(z)$
와 함께	OF [261]	$z_t = x_t \cdot \sigma_t(x<t;\theta) + \mu_t(x<t;\theta)$	$x_t = \frac{z_t - u_t(x<t;\theta)}{\sigma_t(x<t;\theta)}$
	IAF [169]	$\sigma_t(z<t;\theta)\frac{x_t - \mu_t(z<t;\theta)}{z_t =}$	$x_t = z_t \cdot \sigma_t(z<t;\theta) + \mu_t(z<t;\theta)$
이분법	RealNVP [66]	$z_a = x_a,$	$x_a =$ 의지
	글로우 [167]	$z_b = x_b \cdot \sigma_b(x_a;\theta) + \mu_b(x_a;\theta)$	$= \sigma_b(x_a;\theta)\frac{z_b - \mu_b(x_a;\theta)}{x_b}$

- 자기회귀 변환, 예: 역 자기회귀 흐름(IAF) [169]. IAF는 autoregressive flow(AF) 의 이중 공식으로 간주될 수 있습니다 [261, 124]. AF 훈련은 병렬이고 샘플링은 순차적입니다. 대조적으로, IAF의 샘플링은 병렬이지만 우도 추정에 대한 추론은 순차적입니다. Parallel WaveNet [255] 은 IAF의 효율적인 샘플링과 AR 모델링의 효율적인 훈련을 결합하기 위해 확률 밀도 증류를 활용 합니다. 자동 회귀 WaveNet을 교사 네트워크로 사용하여 학생 네트워크(Parallel WaveNet)의 교육을 안내하여 데이터 우도를 근사화합니다. 마찬가지로 ClariNet [269] 은 IAF 및 교사 증류를 사용하고 폐쇄형 KL 분기를 활용하여 증류 프로세스를 단순화하고 안정화합니다.

  Parallel Wavenet 및 ClariNet은 병렬로 음성을 생성할 수 있지만 정교한 교사-학생 교육에 의존 하며 여전히 많은 계산이 필요합니다.

- 예를 들어 글로우 [167] 또는 RealNVP [66]와 같은 이분 변환. 변환이 가역적이 되도록 하기 위해 이분 변환은 출력이 계산될 수 있도록 하는 아핀 결합 계층을 활용합니다.

입력과 그 반대. 이분 변환을 기반으로 하는 일부 보코더에는 WaveGlow [279] 및 FloWaveNet [163]이 포함되어 있어 높은 음성 품질과 빠른 추론 속도를 달성합니다.

자기회귀 변환과 이분 변환 모두 장단점이 있습니다 [271]: 1)
자동회귀 변환은 데이터 분포 x 와 표준 확률 분포 z 사이의 종속성을 모델링하여 이분 변환보다 표현력이 뛰어나지만 학습 시 복잡한 교사 증류가 필요합니다 . 2) 이분 변환은 훈련 파이프라인이 훨씬 간단하지만 일반적 으로 자동 회귀 모델과 비슷한 용량에 도달하려면 더 많은 매개변수(예: 더 깊은 레이어, 더 큰 은닉 크기)가 필요 합니다. autoregressive 및 bipartite 변환의 장점을 결합하기 위해 WaveFlow [271] 는 모델 용량에 대한 추론 병렬 처리를 명시적으로 교환하기 위해 오디오 데이터에 대한 우도 기반 모델의 통합 보기를 제공합니다. 이와 같이 WaveNet, WaveGlow, FloWaveNet 은 WaveFlow의 특수한 경우라고 볼 수 있습니다.

GAN 기반 보코더 GAN (Generative adversarial networks) [90] 은 이미지 생성 [90, 455], 텍스트 생성 [419] 및 오디오 생성 [68] 과 같은 데이터 생성 작업에 널리 사용되었습니다 .
GAN은 데이터 생성을 위한 생성기와 생성된 데이터의 진위 여부를 판단하는 판별기로 구성된다. WaveGAN [68], GAN-TTS [23], MelGAN [178], Parallel WaveGAN [402], HiFi-GAN [174] 및 기타 GAN 기반 보코더 를 포함하여 많은 보코더가 GAN을 활용하여 오디오 생성 품질을 보장합니다 . [401, 391, 312, 417, 372, 137].

표 5: 몇 가지 대표적인 GAN 기반 보코더 및 그 특성.

GAN 생성기		판별기	상실
웨이브GAN [68]	디씨건 [287]	/	WGAN-GP [97]
GAN-TTS [23]	/	임의 창 D 힌지 손실 GAN[198]	
멜간 [178]	/	멀티 스케일 D	LS-GAN [231] 특징 매칭 손실[182]
Par.WaveGAN [402] 웨이브넷 [254]		/	LS-GAN, 다중 STFT 손실
하이파이-GAN [174]	다중 수용 필드 퓨전	다주기 D, 멀티 스케일 D	LS-GAN, STFT 손실, 기능 일치 손실
복건 [408]	멀티 스케일 G	계층적 D	LS-GAN, 다중 STFT 손실, 기능 일치 손실
검정고시 [96]	/	랜덤 윈도우 D	힌지 손실 GAN, 반발 손실

각 보코더에서 사용되는 Generator, Discriminator, Loss에 따른 특성을 Table 5에 정리 하였다.

• 발전기. 대부분의 GAN 기반 보코더는 확장된 컨벌루션을 사용하여 수용 필드를 증가시켜 파형 시퀀스의 긴 종속성을 모델링하고 트랜스포즈된 컨벌루션을 사용하여 파형 시퀀스의 길이와 일치하도록 조건 정보(예: 언어 특징 또는 멜 스펙트로그램)를 업샘플링합니다. Yamamoto et al. 조건부 정보를 한 번 업샘플링하도록 선택한 다음 모델 용량을 보장하기 위해 확장된 컨볼루션을 수행합니다. 그러나 이러한 종류의 업 샘플링은 시퀀스 길이를 너무 일찍 증가시켜 계산 비용이 더 많이 듭니다. 따라서 일부 보코더 [178, 174] 는 조건 정보를 반복적으로 업샘플링하고 확장 컨볼루션을 수행하여 하위 계층에서 너무 긴 시퀀스를 피할 수 있습니다. 구체적으로, VocGAN [408] 은 거친 입자에서 미세 입자까지 다양한 스케일에서 파형 시퀀스를 점진적으로 출력할 수 있는 다중 스케일 생성기를 제안합니다 . HiFi-GAN [174] 은 다중 수용 필드 퓨전 모듈을 통해 다양한 길이의 서로 다른 패턴을 병렬로 처리하고 합성 효율성과 샘플 품질 사이에서 균형을 이룰 수 있는 유연성도 가지고 있습니다.
• 판별자. 판별기에 대한 연구 노력 [23, 178, 174, 408] 은 생성기 에 더 나은 안내 신호를 제공하기 위해 파형의 특성을 캡처하는 모델을 설계하는 방법에 중점을 둡니다 . 우리는 이러한 노력을 다음과 같이 검토합니다. 1) GAN-TTS [23] 에서 제안된 무작위 창 판별기, 여러 판별기를 사용하며 각각 조건부 정보가 있거나 없는 파형의 서로 다른 무작위 창을 공급 합니다.
임의 창 판별자

다양한 보완적 방식으로 오디오를 평가하고, 전체 오디오와 비교하여 참/거짓 판단을 단순화하고, 데이터 증대 효과로 작용하는 등 여러 가지 이점이 있습니다.
2) MelGAN [178] 에서 제안한 다중 스케일 판별기, 여러 판별기를 사용 하여 다양한 스케일(원본 오디오와 비교하여 다른 다운샘플링 비율)의 오디오를 판단합니다. 다중 척도 판별기 의 장점은 각 척도의 판별기가 서로 다른 주파수 범위의 특성에 집중할 수 있다는 것입니다. 3) HiFi GAN [174] 에서 제안한 다중 주기 판별기. 다중 판별기를 활용하며 각 판별기는 마침표가 있는 입력 오디오의 균일한 간격 샘플을 받아들입니다. 구체적으로, 길이가 T 인 1D 파형 시퀀스는 p 가 주기인 2D 데이터 [p, T /p] 로 재구성 되고 2D 컨볼루션에 의해 처리됩니다. 다중 기간 판별자는 서로 다른 기간에 있는 입력 오디오의 서로 다른 부분을 살펴봄으로써 서로 다른 암시적 구조를 캡처할 수 있습니다 . 4) VocGAN [408] 에서 활용되는 계층적 판별자는 거친 입자에서 미세 입자까지 다양한 해상도로 생성된 파형 을 판단하여 생성기가 저주파 및 고주파 모두에서 음향 특징과 파형 사이의 매핑을 학습하도록 안내할 수 있습니다.

- 손실. WGAN-GP [97], 힌지 손실 GAN [198], LS -GAN [231] 과 같은 일반 GAN 손실을 제외하고 STFT 손실 [10, 401] 및 특징 매칭 손실 [182] 과 같은 기타 특정 손실 또한 활용됩니다. 이러한 추가 손실은 적대적 훈련 의 안정성과 효율성을 개선하고 [402] 지각 오디오 품질을 개선할 수 있습니다. Gritsenkoet al. [96] 다중 모달 파형 분포를 더 잘 캡처하기 위해 반발 항이 있는 일반화된 에너지 거리를 제안합니다 .

확산 기반 보코더 최근 DiffWave [176], WaveGrad [41] 및 PriorGrad [185] 와 같은 보코더에 노이즈 제거 확산 확률 모델(DDPM 또는 Diffusion) [113] 을 활용하는 작업이 있습니다. 기본 아이디어는 확산 과정과 역과정을 통해 데이터와 잠재 분포 사이의 매핑을 공식화하는 것 입니다. 확산 과정에서 파형 데이터 샘플은 점차 임의의 노이즈와 함께 추가되고 최종적으로 가우션 노이즈가 됩니다. 역 과정에서 임의의 가우션 노이즈는 단계적으로 파형 데이터 샘플로 점진적으로 노이즈가 제거됩니다. 확산 기반 보코더는 매우 높은 음성 품질로 음성을 생성할 수 있지만 긴 반복 프로세스로 인해 추론 속도가 느려집니다. 따라서 확산 모델에 대한 많은 작업 [313, 185, 384, 175] 은 생성 품질을 유지하면서 추론 시간을 줄이는 방법을 연구하고 있습니다.

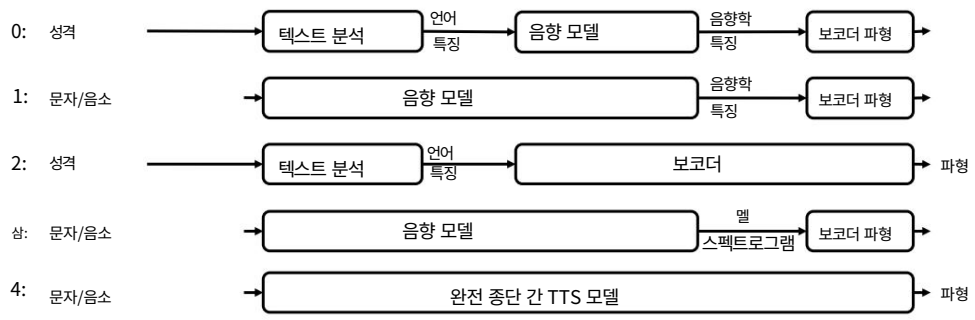
기타 보코더 일부 작품은 제어 가능한 음성 생성을 유지 하면서 높은 음성 품질을 달성하는 것을 목표로 하는 파형 생성 [381, 380, 377, 213, 149, 148, 77, 311, 414]에 신경 기반 소스 필터 모델을 활용합니다 . Govalkaret al. 다양한 종류의 보코더 에 대한 포괄적인 연구를 수행합니다 . Hsu et al. 포괄적인 실험을 통해 몇 가지 일반적인 보코더 를 평가하여 보코더 의 견고성을 연구합니다 .

토론 표 6과 같이 보코더에 사용되는 여러 종류의 생성 모델의 특성을 요약합니다. 1) 수학적 단순성 측면에서 자기회귀(AR) 기반 모델은 VAE, Flow, Diffusion, 간. 2) AR을 제외한 모든 생성 모델은 병렬 음성 생성을 지원할 수 있습니다. 3) AR 모델을 제외한 모든 생성 모델은 잠재 조작을 어느 정도 지원할 수 있습니다. 4)

GAN 기반 모델은 데이터 샘플의 가능성을 추정할 수 없지만 다른 모델은 이러한 이점을 누릴 수 있습니다.

표 6: 보코더에 사용되는 여러 대표적인 생성 모델의 일부 특성.

생성 모델	와 함께	피트	흐름/AR	유동/이분 확산 GAN		
보코더(예)	WaveNet WaveVAE Par.WaveNet WaveGlow DiffWave MelGAN					
단순한	와이	N	N	N	N	N
평행한	N	와이	와이	와이	와이	와이
잠재적인 조작	N	와이	와이	와이	와이	와이*
가능성 추정	와이	와이	와이	와이	와이	N



무대 모델	
0	SPSS [416, 356, 415, 425, 357]
1 의사	[375] 2 3 4
	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
	DeepVoice 3[270], Tacotron 2[303], FastSpeech 1/2[290, 292], WaveGlow[279], FloWaveNet[163]
	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

그림 4: TTS 모델에 대한 점진적인 종단간 프로세스.

2.5 완전한 종단 간 TTS를 향하여

완전한 엔드-투-엔드 TTS 모델은 문자 또는 음소 시퀀스에서 직접 음성 파형을 생성할 수 있으며 다음과 같은 이점이 있습니다. 2) 공동 및 종단 간 최적화는 계단식 모델(예: 텍스트 분석 + 음향 모델 + 보코더) 에서 오류 전파를 피할 수 있습니다 . 3) 또한 교육, 개발 및 배포 비용을 줄일 수 있습니다.

그러나 TTS 모델을 종단간 방식으로 교육하는 데는 큰 어려움이 있습니다. 주로 텍스트와 음성 파형 간의 양식이 다르고 문자/음소 시퀀스와 파형 시퀀스 간의 길이 불일치가 크기 때문입니다. 예를 들어, 길이가 5 초이고 약 20단어인 음성의 경우 음소 시퀀스의 길이는 약 100인 반면 파형 시퀀스의 길이는 80k입니다(샘플 속도가 16kHz인 경우). 메모리의 한계로 인해 발화 전체 의 파형 포인트를 모델 훈련에 넣기가 어렵습니다 . 종단 간 교육을 위해 짧은 오디오 클립만 사용하는 경우 컨텍스트 표현 을 캡처하기 어렵습니다 .

완전한 end-to-end 훈련의 어려움으로 인해 신경 TTS의 개발은 완전한 end-to-end 모델을 향한 점진적 프로세스를 따릅니다. 그림 4는 초기 통계적 파라메트릭 합성 [416, 356, 415, 425, 357]에서 시작하는 이 점진적 프로세스를 보여줍니다. 완전한 종단 간 모델을 향한 프로세스에는 일반적으로 다음과 같은 업그레이드가 포함됩니다. 1) 텍스트 분석 모듈 및 언어 기능 단순화. SPSS에서 텍스트 분석 모듈에는 텍스트 정규화, 구/단어/음절 분할, POS 태깅, 운율 예측, 자소에서 음소로의 변환(다음성 명확화 포함) 과 같은 다양한 기능이 포함되어 있습니다 . 종단간 모델에서는 문자를 음소로 변환하기 위해 텍스트 정규화와 자소-음소 변환만 유지 하거나 문자를 직접 입력받아 전체 텍스트 분석 모듈을 제거한다. 2) 음향 특성 단순화, SPSS에서 사용되는 MGC, BAP 및 F0과 같은 복잡한 음향 특성 을 mel-spectrogram으로 단순화합니다. 3) 2개 또는 3개의 모듈을 단일 엔드-투-엔드 모델로 교체합니다. 예를 들어 음향 모델과 보코더 를 WaveNet과 같은 단일 보코더 모델로 교체할 수 있습니다. 따라서 그림 4의 진행 과정을 설명하고 다음과 같이 설명합니다.

• Stage 0. 통계적 파라메트릭 합성 [416, 356, 415, 425, 357] 은 텍스트 분석 모듈이 문자를 언어적 특징으로 변환하고 음향 모델이 언어적 특징에서 음향적 특징을 생성 하는 세 가지 기본 모듈을 사용 합니다(여기서 대상 음향 특징을 얻음). 보코더 분석 을 통해), 보코더는 파라메트릭 계산을 통해 음향 특징에서 음성 파형을 합성합니다.

표 7: 완전한 종단 간 TTS 모델 목록.

모델	1단계 교육 AR/NAR 모델링			건축학
Char2Wav [315]	N	와 함께	Seq2Seq RNN	
클라리넷 [269]	N	와 함께	흐름	CNN
패스트스피치 2s [292]	와이	NAR	하지만	자체 공격/CNN
먹다 [69]	와이	NAR	하지만	CNN
웨이브 타코트론 [385] Y		와 함께	흐름	CNN/RNN/하이브리드
EfficientTTS-Wav [235] Y		NAR	하지만	CNN
농담 [160]	와이	NAR	VAE+Flow CNN/Self-Att/하이브리드	

• 1기. Wang et al. 통계 파라메트릭 합성에서 텍스트 분석과 음향 모델을 음소 시퀀스에서 음향 특징을 직접 생성하는 엔드 투 엔드 음향 모델로 결합한 다음 SPSS에서 보코더를 사용하여 파형을 생성하는 방법 을 탐색합니다.

• Stage 2. WaveNet [254] 은 어쿠스틱 모델과 보코더의 조합으로 볼 수 있는 언어적 특징에서 음성 파형을 직접 생성하기 위해 처음으로 제안됩니다 . 이러한 종류의 모델[254, 255, 150, 23]은 언어적 특징을 생성하기 위해 여전히 텍스트 분석 모듈이 필요합니다.

• 3단계. Tacotron [382] 은 인코더-어텐션-디코더 모델을 사용하여 문자/음소에서 선형 스펙트로그램을 직접 예측 하고 Griffin-Lim [95] 을 사용하여 선형 스펙트로그램을 파형으로 변환 하는 언어 및 음향 기능을 단순화하기 위해 추가로 제안됩니다 .]. DeepVoice 3 [270], Tacotron 2 [303], TransformerTTS [192], FastSpeech 1/2 [290, 292] 와 같은 다음 작품 은 문자/음소로부터 mel-스펙트로그램을 예측하고 추가 로 WaveNet과 같은 신경 보코더를 사용합니다 . 254], WaveRNN [150], WaveGlow [279], FloWaveNet [163] 및 Parallel WaveGAN[402]을 사용하여 파형을 생성합니다.

• 단계 4. 표 7에 나열된 것과 같이 일부 완전한 종단 간 TTS 모델은 직접 텍스트-파형 합성을 위해 개발되었습니다 . Char2Wav [315] 는 RNN 기반 인코더-주의-디코더 모델을 활용하여 문자에서 음향 특징을 생성합니다. 그런 다음 SampleRNN [233] 을 사용 하여 파형을 생성합니다.
두 모델은 직접 음성 합성을 위해 공동으로 조정됩니다. 마찬가지로 ClariNet [269] 은 직접 파형 생성 을 위해 자동회귀 음향 모델과 비자동회귀 보코더를 공동으로 조정합니다 . FastSpeech 2s [292] 는 완전히 병렬 구조로 텍스트에서 음성을 직접 생성 하여 추론 속도를 크게 높일 수 있습니다. 공동 텍스트-파형 훈련의 어려움을 완화하기 위해 보조 mel-스펙트로그램 디코더를 활용하여 음소 시퀀스의 컨텍스트 표현을 학습하는 데 도움을 줍니다. EATS [69] 라는 동시 작업은 또한 문자/음소에서 파형을 직접 생성 하며, 이는 종단 간 정렬 학습을 위해 기간 보간 및 소프트 동적 시간 래핑 손실 을 활용합니다. Wave-Tacotron [385] 은 흐름 부분에서 병렬 파형 생성을 사용하지만 Tacotron 부분에서는 여전히 자동 회귀 생성 을 사용하는 파형을 직접 생성하기 위해 Tacotron에 흐름 기반 디코더를 구축합니다 .

2.6 기타 분류법

그림 3 에 표시된 주요 구성 요소 및 데이터 흐름의 관점에서 주요 분류법 외에도 그림 5에 표시된 것처럼 여러 가지 분류법에서 TTS 작업을 분류할 수 있습니다. 1) 자동 회귀 또는 비자동 회귀. 이러한 작업을 자동 회귀 및 비 자동 회귀 생성 모델 로 나눌 수 있습니다 . 2) 생성 모델. TTS는 일반적인 시퀀스 생성 작업이며 일반적인 생성 모델을 통해 모델링할 수 있으므로 일반 시퀀스 생성 모델, 흐름, GAN, VAE 및 확산 모델과 같은 다양한 생성 모델로 분류할 수 있습니다. 삼)

네트워크 구조. CNN, RNN, self-attention 및 하이브리드 구조(CNN+RNN, CNN+self-attention 과 같은 하나 이상의 구조 유형 포함) 와 같은 네트워크 구조에 따라 작업을 나눌 수 있습니다.

신경 TTS 모델의 진화 신경 TTS 에 대한 다양한 연구 개발과 그 관계를 더 잘 이해하기 위해 그림 6과 같이 신경 TTS 모델의 진화를 설명합니다 . 논문은 대중에게 공개되지만(예: arXiv에 게시) 나중에 공식적으로 게시되지는 않습니다. 연구자들이 지식 공유를 장려하기 위해 자신의 논문을 일찍 공개하는 것에 감사하기 때문에 우리는 이른 시간을 선택합니다.

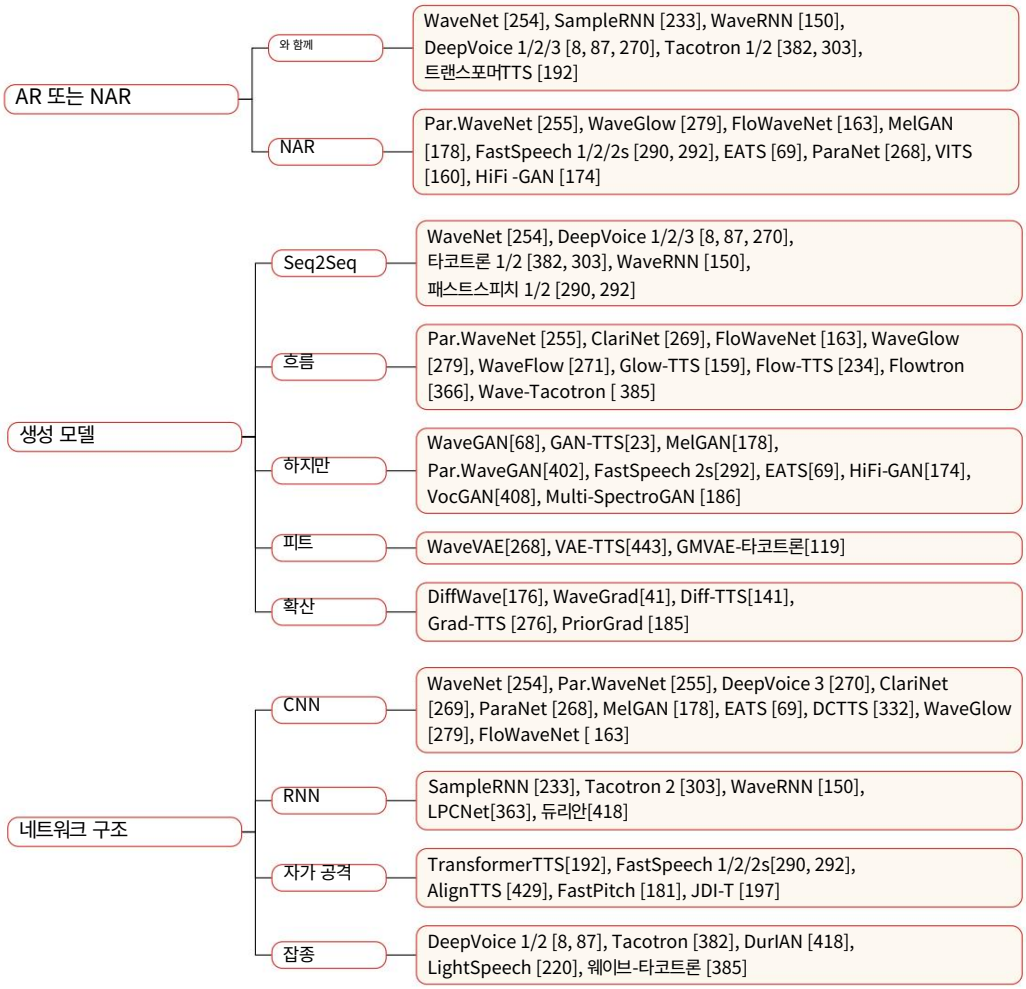


그림 5: AR/NAR, 생성 모델 및 네트워크 구조 의 관점에서 본 신경 TTS의 몇 가지 다른 분류법 .

신경 TTS에 대한 연구는 매우 풍부하므로 그림 6에서 대표적인 연구를 일부만 선택하고 표 18에 더 많은 연구를 나열합니다.

3 TTS의 고급 주제

3.1 배경 및 분류

이전 섹션에서는 기본 모델 구성 요소 측면에서 신경 TTS를 소개했습니다. 이 섹션에서는 새로운 영역을 개척하고 보다 실용적인 제품 사용을 다루는 것을 목표로 하는 신경 TTS의 일부 고급 주제를 검토합니다. 구체적으로, TTS는 느린 자기회귀 생성을 갖는 시퀀스 생성 작업에 대한 전형적인 시퀀스이기 때문에 어떻게 자기회귀 생성 속도를 높이거나 빠른 음성 합성을 위한 모델 크기를 줄이는지가 뜨거운 연구 주제입니다(섹션 3.2). 좋은 TTS 시스템은 자연스럽고 이해하기 쉬운 음성을 모두 생성해야 하며 많은 TTS 연구 작업 은 음성 합성의 명료도와 자연스러움을 개선하는 것을 목표로 합니다. 예를 들어 TTS 모델을 교육하는 데이터가 충분하지 않은 리소스가 적은 시나리오에서 합성된 음성은 낮은 명료도와 자연스러움을 모두 가질 수 있습니다. 따라서 많은 작업이 저자원 설정에서 데이터 효율적인 TTS 모델을 구축하는 것을 목표로 합니다 (섹션 3.3). TTS 모델은 일반적으로 생성된 음성에 명료도에 영향을 미치는 단어 건너뛰기 및 반복 문제가 있는 강력한 문제가 발생하기 쉽기 때문에 많은 작업 이 음성 합성의 견고성을 개선하는 것을 목표로 합니다(섹션 3.4). 자연스러움을 향상시키기 위해 많은 작업들이 표현력을 생성하기 위해 말의 스타일/운율을 모델링, 제어 및 전달하는 것을 목표로 합니다.

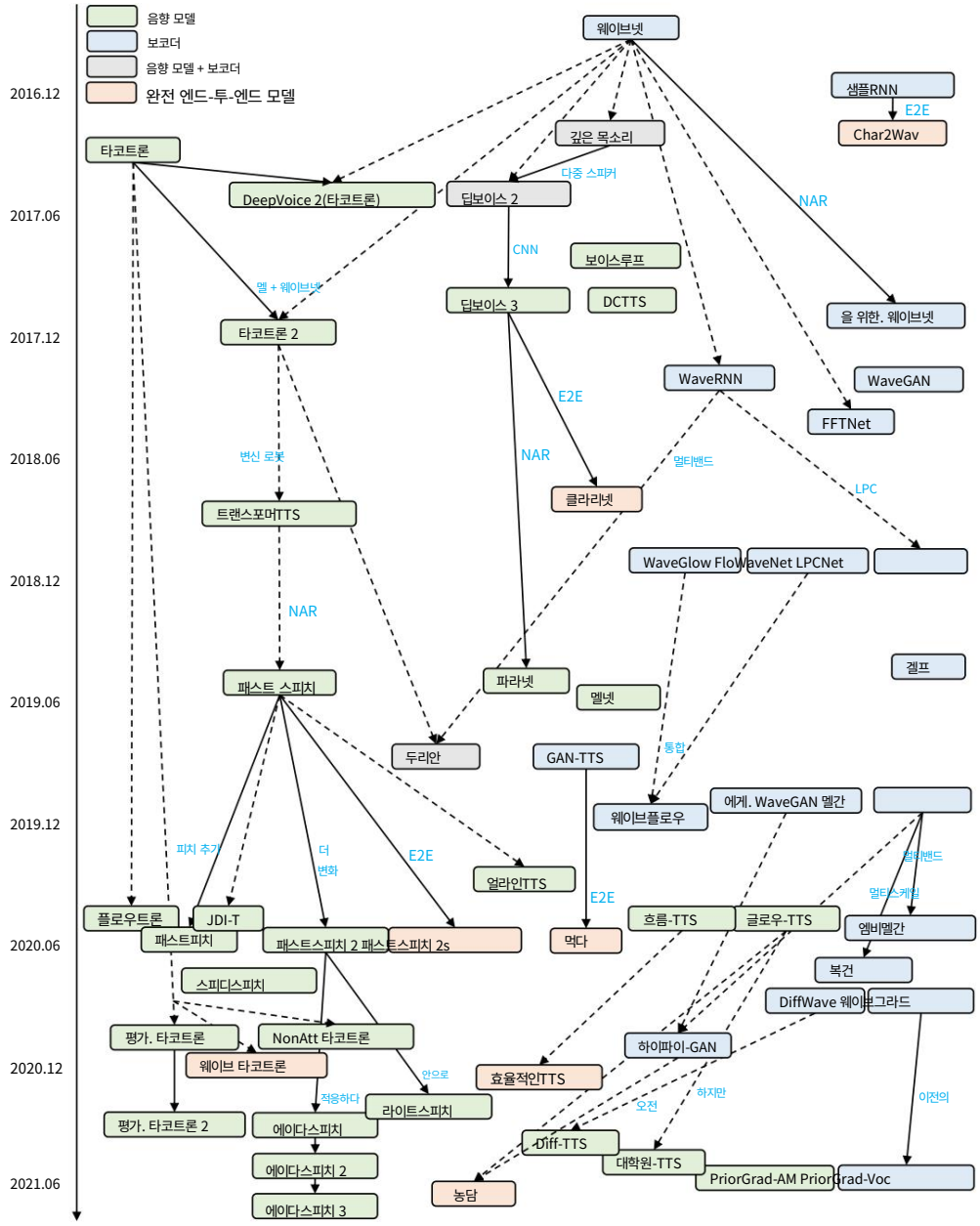


그림 6: 신경 TTS 모델의 진화.

음성(섹션 3.5). 대상 화자의 음성을 지원하도록 TTS 모델을 조정하면 TTS 를 광범위하게 사용하는 데 매우 유용합니다. 따라서 제한된 적응 데이터 및 매개 변수와 고품질 음성으로 효율적인 음성 적응이 실제 사용에 중요합니다(섹션 3.6). 이러한 고급 항목 의 분류는 그림 7에 나와 있습니다.

3.2 빠른 TTS

텍스트 음성 변환 시스템은 일반적으로 빠른 합성 속도 가 필요한 클라우드 서버 또는 임베디드 장치에 배포됩니다 . 그러나 초기 신경 TTS 모델은 일반적으로 긴 음성 시퀀스를 고려할 때 매우 느린 자동 회귀 멜 스펙트로그램 및 파형 생성을 채택합니다 (예: 1초 음성은 일반적으로 홉 크기가 10ms인 경우 500 mel-스페트로그램을 가지며 샘플링 속도인 경우 24k 파형 포인트 를 갖습니다. 24kHz). 이 문제를 해결하기 위해 1) mel을 생성하는 비자동회귀 생성을 포함하여 TTS 모델의 추론 속도 를 높이기 위해 다양한 기술이 활용되었습니다.

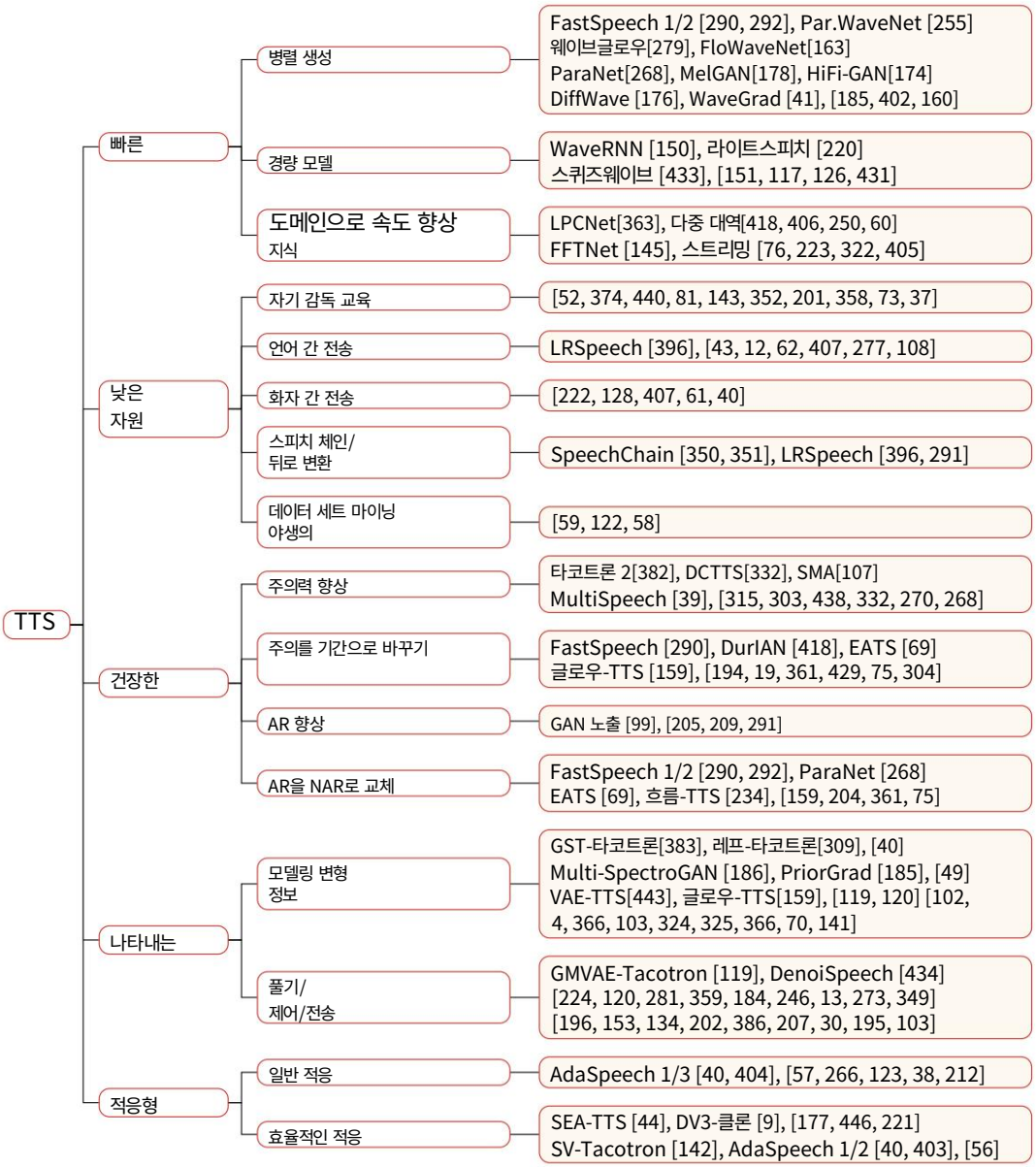


그림 7: 섹션 3에 설명된 신경 TTS의 고급 항목 개요.

스펙트로그램 및 파형 병렬; 2) 가볍고 효율적인 모델 구조; 3) 빠른 음성 합성을 위해 음성의 도메인 지식을 활용하는 기술. 이러한 기술 을 다음과 같이 소개합니다.

표 8: 시퀀스 길이 N에 대한 교육 및 추론에서 다양한 TTS 모델의 시간 복잡도 . T는 흐름/확산 기반 모델의 단계/반복 수입니다.

모델링 패러다임	TTS 모델	훈련	추론	O(N)
아르곤(RNN)	Tacotron 1/2, SampleRNN, LPCNet			켜짐)
AR(CNN/셀프 공격)	DeepVoice 3, TransformerTTS, WaveNet	O(1)		켜짐)
NAR(CNN/Self-Att)	FastSpeech 1/2, ParaNet	MelGAN,	O(1)	오(1)
NAR(GAN/VAE)	HiFi-GAN, FastSpeech 2s, EATS		O(1)	오(1)
흐름(AR)	평가. 웨이브넷, 클라리넷, 플로우트론	O(T)	오(1)	오(1)
흐름(이분)	WaveGlow, FloWaveNet, Glow-TTS			오(티)
확산	DiffWave, WaveGrad, Grad-TTS, PriorGrad	O(T)		오(티)

병렬 생성 표 8은 일반적인 모델링 패러다임, 해당 TTS 모델, 교육 및 추론의 시간 복잡도를 요약합니다. 알 수 있듯이 RNN 기반 자동 회귀 모델 [382, 303, 233, 363] 을 사용하는 TTS 모델은 O(N) 계산(여기서 N 은 시퀀스 길이 임) 을 사용하여 교육 및 추론 모두에서 느립니다 . RNN 구조 로 인한 느린 훈련 시간을 피하기 위해 DeepVoice 3 [270] 및 TransformerTTS [192] 는 병렬 훈련을 지원할 수 있지만 여전히 자동 회귀 추론이 필요한 CNN 또는 self-attention 기반 구조를 활용 합니다. 추론 속도를 높이기 위해 FastSpeech 1/2 [290, 292] 는 계산이 O(1) 로 감소되는 병렬 교육 및 추론 모두에 대해 self-attention 구조를 활용하는 피드포워드 Transformer를 설계합니다 . mel-spectrogram 및 파형 생성을 위한 대부분의 GAN 기반 모델 [178, 174, 292, 69] 은 학습 및 추론 모두에서 O(1) 계산을 사용하여 비 자동 회귀 입니다. Parallel WaveNet [255] 및 ClariNet [269] 은 역 자기회귀 흐름 [169] 을 활용 하여 병렬 추론을 가능하게 하지만 병렬 훈련을 위해서는 Teacher Distillation이 필요합니다. WaveGlow [279] 및 FloWaveNet [163] 은 병렬 교육 및 추론을 위해 생성 흐름을 활용 합니다. 그러나 일반적으로 데이터와 이전 분포 간의 매핑 품질을 보장하기 위해 여러 흐름 반복 T 를 쌓아야 합니다. 흐름 기반 모델과 유사하게 확산 기반 모델 [41, 176, 185, 141, 276] 도 정방향 및 역방향 프로세스 에서 여러 확산 단계 T 가 필요 하므로 계산이 증가합니다.

경량 모델 비자동회귀 생성은 추론 속도 향상을 위해 병렬 계산을 완전히 활용할 수 있지만 모델 매개변수의 수와 총 계산 비용은 줄어들지 않습니다 . 이러한 장치는 충분히 강력하지 않습니다. 따라서 자동 회귀 생성 을 사용하더라도 추론 속도 향상을 위해 더 적은 계산 비용으로 가볍고 효율적인 모델을 설계 해야 합니다. 경량 모델을 설계하는 데 널리 사용되는 기술에는 가지치기, 양자화, 지식 추출 [111], 신경 구조 검색 [220, 397] 등 이 있습니다.

WaveRNN [150] 은 추론 속도를 높이기 위해 이중 소프트맥스, 가중치 가지치기, 하위 규모 예측과 같은 기술을 사용합니다 . LightSpeech [220] 는 신경 아키텍처 검색 [457, 219] 을 활용하여 음성 품질을 유지하면서 FastSpeech 2 [292] 의 추론 속도를 6.5배 더 높이는 경량 아키텍처를 찾습니다 . SqueezeWave [433] 는 파형 재구성을 활용하여 시간 길이를 줄이고 1D 컨볼루션을 깊이 분리 가능한 컨볼루션으로 대체하여 유사한 오디오 품질을 달성하면서 계산 비용을 줄입니다. Kanagawa와 Ijima [151] 는 텐서 분해 로 LPCNet의 모델 매개변수를 압축합니다 . Hsu와 Lee [117] 는 계산 자원 을 줄이기 위해 압축된 흐름 기반 모델 과 오디오 품질을 유지하기 위해 WaveNet 기반 사후 필터를 제안합니다. DeviceTTS [126] 는 DFSMN [441] 의 모델 구조 와 혼합 해상도 디코더를 활용하여 하나의 디코딩 단계에서 여러 프레임을 예측하여 추론 속도를 높입니다. LVCNet [431] 은 서로 다른 파형 간격에 대해 위치 변수 컨볼루션을 채택하며 여기서 컨볼루션 계수는 mel-스펙트로그램에서 예측됩니다.

음질 저하 없이 Parallel WaveGAN 보코더의 속도를 4배 향상시킵니다. Wang et al. [373] 은 mel-스펙트로그램 생성을 위한 semi-autoregressive 모드를 제안합니다. 여기서 mel 스펙트로그램은 개별 음소에 대한 자동 회귀 모드와 다른 음소에 대한 비자동 회귀 모드로 생성됩니다 .

도메인 지식으로 속도 향상 음성의 도메인 지식은 선형 예측 [363], 다중 대역 모델링 [418, 406, 60], 하위 규모 예측과 같은 추론 속도를 높이는 데 활용될 수 있습니다 .

[150], 다중 프레임 예측 [427, 382, 373, 126, 210], 스트리밍 합성 [76] 등. LPC Net [363] 은 선형 예측 계수를 사용하여 다음 파형 및 경량 모델을 사용하여 잔차 값을 예측하여 자동 회귀 파형 생성의 추론 속도를 높일 수 있습니다. 보코더의 추론 속도를 높이는 데 널리 사용되는 또 다른 기술은 빠른 추론을 위해 파형을 여러 하위 대역으로 나누는 하위 대역 모델링입니다. 대표적인 모델로는 DurIAN [418], 다중 대역 Mel GAN [406], 하위 대역 WaveNet [250] 및 다중 대역 LPCNet [348, 60]이 있습니다. Bunched LPCNet [370] 은 샘플 번칭 및 비트 번칭으로 LPCNet의 계산 복잡성을 줄여 2배 이상의 속도 향상을 달성합니다. 스트리밍 TTS [76, 223, 322, 405, 323, 237] 는 일부 입력 토큰이 오면 전체 입력 문장을 기다리지 않고 음성을 합성 하여 추론 속도를 높일 수도 있습니다. FFTNet [145] 은 간단한 아키텍처를 사용하여 오디오 샘플을 실시간으로 생성할 수 있는 고속 푸리에 변환(FFT)을 모방합니다 . Okamotoet al. 노이즈 쉐이핑 및 서브밴드 기술로 FFTNet을 더욱 강화하여 작은 모델 크기를 유지하면서 음성 품질을 향상시킵니다. Popovet al. [274] 는 프레임 분할 및 교차 페이딩을 제안하여 파형의 일부를 병렬로 합성한 다음 합성된 파형을 함께 연결하여 저가형 장치에서 빠른 합성을 보장합니다. Kang et al. 단일 CPU 스레드 로 음성을 실시간으로 합성할 수 있는 그룹 하이웨이 활성화와 같은 네트워크 축소 및 충실도 개선 기술로 DCTTS를 가속화 합니다 .

3.3 저자원 TTS

고품질 TTS 시스템을 구축하려면 일반적으로 대량의 고품질 쌍 텍스트 및 음성 데이터가 필요합니다. 그러나 전 세계에는 7,000개 이상의 언어가 있으며14 대부분의 언어는 TTS 시스템 개발을 위한 학습 데이터가 부족합니다. 그 결과 인기 있는 상용화된 음성 서비스 15는 TTS에 대해 수십 가지 언어만 지원할 수 있습니다. 저자원 언어에 대한 TTS 지원 은 비즈니스 가치를 가질 수 있을 뿐만 아니라 사회적 이익에도 도움이 됩니다. 따라서 많은 연구 작업 이 낮은 데이터 리소스 시나리오에서 TTS 시스템을 구축합니다. 저자원 TTS의 대표적인 기법을 표 9에 정리하고 , 이러한 기법을 다음과 같이 소개한다.

표 9: 저자원 TTS를 위한 몇 가지 대표적인 기술.

기법	데이터 일하다
자기 감독 교육	페어링되지 않은 텍스트 또는 음성 [52, 374, 440, 81, 143, 352, 201, 358, 73]
다국어 전송	쌍으로 된 텍스트 및 음성 [43, 396, 12, 407, 62, 277, 108]
화자 간 전송	쌍으로 된 텍스트 및 음성 [222, 128, 61, 407, 40]
음성 체인/역 변환	페어링되지 않은 텍스트 또는 음성 [291, 396, 350, 351]
야생에서의 데이터 세트 마이닝	쌍으로 된 텍스트 및 음성 [59, 122, 58]

- 자기 감독 교육. 페어링된 텍스트 및 음성 데이터는 수집하기 어렵지만 페어링되지 않은 음성 및 텍스트 데이터(특히 텍스트 데이터)는 비교적 쉽게 얻을 수 있습니다. 자기 감독 사전 훈련 방법은 언어 이해 또는 음성 생성 기능을 향상시키기 위해 활용될 수 있습니다 [52, 374, 440, 81]. 예를 들어, TTS의 텍스트 인코더는 사전 훈련된 BERT 모델 [52, 81, 143]에 의해 향상될 수 있으며 TTS의 음성 디코더는 자동 회귀 멜 스펙트로그램 예측 [52] 을 통해 사전 훈련 되거나 음성 변환 작업 [440].

  게다가, 음성은 음소 또는 문자 시퀀스와 유사하도록 이산 토큰 시퀀스로 양자화될 수 있습니다 [352]. 이러한 방식으로 양자화된 개별 토큰과 음성은 TTS 모델을 사전 훈련하기 위한 의사 쌍 데이터로 간주될 수 있으며, 그런 다음 몇 가지 진정으로 쌍을 이루는 텍스트 및 음성 데이터[201, 358, 436]에서 미세 조정됩니다.

- 다국어 전송. 짝을 이룬 텍스트 및 음성 데이터는 자원이 적은 언어에서는 드물지만 자원 이 풍부한 언어에서는 풍부합니다. 인간의 언어는 유사한 발성 기관, 발음 [389] 및 의미 구조 [341]를 공유하므로 풍부한 자원 언어에 대한 TTS 모델을 사전 훈련하면 자원 이 적은 언어에서 텍스트와 음성 간의 매핑을 도울 수 있습니다 [43, 396, 12, 62, 101, 342, 442, 247, 407, 435]. 일반적으로 풍부하고 자원이 적은 언어 사이에는 서로 다른 음소 세트가 있습니다. 따라서 Chen et al. [43] 서로 다른 언어 의 음소 집합 사이에 임베딩 매핑을 제안 하고 LRSpeech [396] 는 사전 훈련된 음소를 폐기합니다.

14https://www.ethnologue.com/
browse 15예: Microsoft Azure, Google Cloud 및 Amazon AWS.

자원이 적은 언어에 대해 음소 임베딩을 처음부터 포함하고 초기화합니다.
IPA(International Phonetic Alphabet) [109] 또는 바이트 표현 [108] 은 여러 언어로 된 임의의 텍스트를 지원하기 위해 채택되었습니다. 또한 언어 유사성 [341] 도 언어 간 전이를 수행할 때 고려할 수 있다 .

- 화자 간 전송. 특정 화자의 음성 데이터가 제한된 경우 다른 화자의 데이터를 활용하여 이 화자의 합성 품질을 향상시킬 수 있습니다. 이는 음성 변환을 통해 다른 화자의 음성을 이 대상 음성으로 변환 하여 훈련 데이터 를 늘리거나 [128], 음성 적응 또는 음성 복제 를 통해 다른 음성으로 훈련된 TTS 모델을 이 대상 음성에 적응[44, 40] 섹션 3.6에서 소개합니다.

- 스피치 체인/백 변환. TTS(텍스트 음성 변환) 및 ASR(자동 음성 인식) 은 두 가지 이중 작업 [285] 이며 함께 활용하여 서로를 개선 할 수 있습니다. 음성 체인 [350, 351] 및 역 변환 [291, 396] 과 같은 기술은 짝을 이루지 않은 추가 텍스트 및 음성 데이터를 활용 하여 TTS 및 ASR의 성능을 향상시킵니다.

- 야생에서의 데이터 세트 마이닝. 일부 시나리오에서는 웹에 낮은 품질의 쌍을 이루는 텍스트 및 음성 데이터 가 일부 존재할 수 있습니 다 . Cooper [59], Hu et al. [122] 는 이러한 종류의 데이터를 마이닝 하고 TTS 모델을 교육하기 위한 정교한 기술을 개발할 것 을 제안합니다 . 음성 향상 [362], 잡음 제거 [434] 및 얽힘 해제 [383, 120] 와 같은 일부 기술을 활용 하여 야생에서 채굴된 음성 데이터의 품질을 향상시킬 수 있습니다 .

3.4 강력한 TTS

우수한 TTS 시스템은 코너 케이스가 발생 하더라도 텍스트에 따라 항상 "올바른" 음성을 생성하도록 견고해야 합니다 . 신경 TTS 에서 문자/음소 시퀀스에서 mel- 스펙트로그램 시퀀스를 생성할 때 음향 모델17에서 단어 건너뛰기, 반복 및 주의 붕괴 와 같은 강력한 문제가 종종 발생합니다. 기본적으로 말해서, 이러한 강력한 문제의 원인은 두 가지 범주에 있습니다 . 2) 자기 회귀 생성에서 발생하는 노출 편향 및 오류 전파 문제 . 보코더는 음향 특성과 파형이 이미 프레임별로 정렬되어 있기 때문에 (즉, 음향 특성의 각 프레임은 파형 포인트의 특정 수 (홉 크기)에 해당함) 심각하게 강력한 문제에 직면하지 않습니다 . 따라 서 강력한 TTS에 대한 기존 작업은 위의 두 가지 문제를 각각 해결 합니다.

- 문자/음소와 mel-spectrogram 간의 정렬 학습을 위해 작업은 1) Attention 메커니즘의 견고성 향상 [382, 315, 303, 438, 332, 107, 39] 및 2) 두 가지 측면으로 나눌 수 있습니다. 주의를 제거하고 대신 지속 시간을 명시적으로 예측하여 텍스트와 음성 사 이의 길이 불일치를 연결합니다[290, 418, 69, 75].

- 자기회귀 생성의 노출편향 및 오차전파 문제에 대해서도 1) 자기회귀 생성을 개선하여 노출 편향 및 오차전파 문제 를 완화 [99, 205, 209, 291], 2) 작업으로 나눌 수 있다. 자동회귀 생성을 제거 하고 대신 비자동회귀 생성을 사용합니다[290, 292, 268, 69].

견고성을 개선하기 위해 이러한 범주에서 인기 있는 기술을 요약하면 표 10에 나와 있습니다. 두 가지 문제를 해결하는 작업 은 겹칠 수 있습니다. 예를 들어 일부 작업은 AR 또는 NAR 생성에서 주의 메커니즘을 향상시킬 수 있으며 유사하게 지속 시 간 예측 은 AR 및 NAR 생성 모두에 적용됩니다. 우리는 이러한 범주를 다음과 같이 검토합니다.

16주의 붕괴는 생성된 음성에 이해할 수 없는 횡설수설이 있음을 의미하며, 이는 일반적으로 단일 입력 토큰에 주의를 집중하지 않습니다[107].

17강력한 문제는 생성된 파형에 쉰 목소리, 금속성 노이즈, 지터 또는 피치 깨짐과 같은 약간의 결함이 있을 수 있는 신경 보코더에서도 발생할 수 있습니다. 그러나 어쿠스틱 모델에서와 같이 심각하지 않으며 이러한 문제를 일으키는 원인이 명확하지 않으며 범용 보코더 모델링 [215, 265, 137, 144] 또는 정교한 설계 [35]로 수리할 가능성이 높습니다. 따라서 본 조사에서는 음향모델의 강건한 문제를 해결 한 작업을 주로 소개한다 .

18테스트 도메인이 교육 도메인 에 의해 잘 다루어지지 않는 것과 같이 강력한 문제를 일으킬 수 있는 몇 가지 다른 이유가 있습니다 . 보이지 않는 영역으로 확장되는 연구 작업 은 훈련 데이터의 양과 다양성 증가 [130], 훈련에서 보이지 않는 긴 시퀀스를 지원하기 위해 상대 위치 인코딩 채택 [17, 430] 등과 같이 이 문제를 완화할 수 있습니다.

표 10: 견고한 TTS를 위한 방법의 분류.

범주	기술	일하다
주의력 향상	콘텐츠 기반 관심	[382, 192]
	위치 기반 관심	[315, 333, 367, 17]
	콘텐츠/위치 하이브리드 주의	[303] [438, 107,
	단조로운 관심	411] [332, 438, 270,
	윈도우잉 또는 비대각선 페널티	39] [382, 303, 270,
	enc-dec 연결 개선	203, 39] [268, 234, 204]]
	위치 주의	
주의를 다음으로 바꾸기 기간 예측	인코더-디코더 주의 레이블	[290, 361, 197, 181]
	CTC 정렬의 레이블	[19] [292, 418, 194,
	HMM 정렬의 레이블	252, 74, 304] [429, 193, 235]
	동적 프로그래밍	[159]
	단조로운 정렬 검색	
	소프트 DTW를 사용한 단조 보간 [69, 75]	
AR 향상	교수 강요	[99, 205]
	교육/추론 격차 줄이기	[361]
	지식 증류	[209]
	양방향 정규화	[291, 452]
AR을 NAR 병렬 생성으로 대체		[290, 292, 268, 69]

3.4.1 주의력 향상

자동 회귀 음향 모델 에서 인코더-디코더 어텐션에서 학습된 잘못된 어텐션 정렬로 인해 많은 단어 건너뛰기/반복 및 어텐션 붕괴 문제가 발생합니다. 이 문제를 완화하기 위해 텍스트(문자/음소) 시퀀스와 멜 스펙트로그램 시퀀스 간의 정렬의 일부 속성 이 고려됩니다 [107]. 1) 로컬: 하나의 문자/음소 토큰 을 하나 이상의 연속 멜 스펙트로그램 프레임에 정렬할 수 있습니다. 하나의 mel-스펙트로그램 프레임은 단일 문자/음소 토큰에만 정렬될 수 있으므로 흐릿한 주의 및 주의 붕괴를 피할 수 있습니다. 2) 단조: 문자 A가 문자 B 뒤에 있으면 A에 해당하는 멜 스펙트로그램도 B에 해당하는 것 뒤에 있으므로 단어 반복을 피할 수 있습니다. 3) 완료: 각 문자/음소 토큰은 단어 건너뛰기 를 방지할 수 있는 최소 하나의 멜 스펙트로그램 프레임으로 덮여 있어야 합니다 . 위의 세 가지 속성을 만족 하는지 여부에 따라 주의력을 향상시키는 기술(표 10에서)을 분석하고 표 11에 나열합니다. 이러한 기술을 다음과 같이 설명합니다.

표 11: 주의력 향상 기술 및 세 가지 속성 (국소/단조/완전)을 충족하는지 여부.

기법	로컬	단조	완료
콘텐츠 기반 관심	×	×	×
위치 기반 관심	×		×
콘텐츠/위치 하이브리드 주의	×		×
단조로운 관심			×
단계적 단조주의			
윈도우잉 또는 비대각선 페널티	×	×	×
enc-dec 연결 개선	×	×	×
위치 주의	×	×	×
기간 예측			

- 콘텐츠 기반 관심. TTS에서 채택된 초기 어텐션 메커니즘(예: Tacotron [382]) 은 콘텐츠 기반 [14]이며 , 어텐션 분포는 인코더와 디코더의 숨겨진 표현 간의 일치 정도에 따라 결정됩니다. 콘텐츠 기반 어텐션은 소스 토큰과 대상 토큰 간의 정렬 이 순전히 의미론적 의미(콘텐츠) 기반인 신경 기계 번역 [14, 368] 과 같은 작업에 적합합니다 . 그러나 자동 음성 인식 [50, 34, 48] 및 텍스트 음성 합성 [382]과 같은 작업 의 경우 정렬이

텍스트와 음성 사이에는 몇 가지 특정 속성이 있습니다. 예를 들어, TTS [107] 에서 어텐션 정렬은 국소적이고 단조적이며 완전해야 합니다. 따라서 고급 어텐션 메커니즘 은 이러한 속성을 더 잘 활용하도록 설계되어야 합니다.

- 위치 기반 관심. 텍스트와 음성 간의 정렬이 위치에 따라 다르다는 점을 고려하여 정렬을 위해 위치 정보를 활용하기 위해 위치 기반 주의 [93, 17] 가 제안됩니다 . Char2Wav [315], VoiceLoop [333] 및 MelNet [367] 과 같은 여러 TTS 모델은 위치 기반 어텐션을 채택합니다. 표 11에 요약된 것처럼 위치 기반 주의는 적절하게 처리되면 단조성 속성을 보장할 수 있습니다.

- 콘텐츠/위치 기반 하이브리드 주의. 콘텐츠와 위치 기반 관심의 장점을 결합하기 위해 Chorowski et al. [50], Shen et al. 위치 감지 주의 도입: 현재 주의 정렬을 계산할 때 이전 주의 정렬이 사용됩니다. 이런 식 으로 단조로운 정렬로 인해 주의가 더 안정적일 것입니다.

- 단조로운 관심. 단조로운 주의 [288, 47, 107, 411, 347]의 경우 주의 위치는 단조롭게 증가하며, 이는 텍스트와 음성 사이의 정렬이 단조롭다는 사전도 활용합니다. 이러한 방식으로 건너뛰기 및 반복 문제를 피할 수 있습니다. 그러나 위의 단조주의에서는 완전성 속성을 보장할 수 없습니다. 그러므로 He et al. [107] 은 각 디코딩 단계에서 주의 정렬 위치가 최대 한 단계 앞으로 이동하고 입력 단위를 건너뛸 수 없는 단계적 단조 주의를 제안합니다.

- 윈도우잉 또는 비대각선 페널티. 주의 정렬은 단조롭고 대각선이기 때문에 Chorowski et al. [50], Tachibana et al. [332], 장 외. [438], Ping et al. [270], Chen et al. [39] 소스 시퀀스에 대한 주의를 창 하위 집합으로 제한할 것을 제안합니다. 이러한 방식으로 학습 유연성과 난이도가 감소합니다. Chenet al. [39] 밴드 마스크를 구성하고 어텐션 웨이트가 대각선 밴드에 분산 되도록 장려하여 오프-대각선 어텐션 웨이트에 대한 페널티 손실을 사용합니다 .

- 인코더-디코더 연결 강화. 음성은 인접한 프레임 간의 상관관계가 더 높기 때문에 디코더 자체에 다음 프레임을 예측하기에 충분한 정보가 포함되어 있으므로 인코더의 텍스트 정보 를 무시하는 경향이 있습니다. 따라서 일부 작업에서는 인코더와 디코더 간의 연결을 향상시켜 주의 정렬을 향상시킬 수 있다고 제안합니다 . Wang et al. [382], Shen et al. 각 디코더 단계 에서 다중 비중첩 출력 프레임을 생성하는 다중 프레임 예측을 사용합니다 . 이러한 방식으로 연속 프레임을 예측하기 위해 디코더는 정렬 학습을 향상시킬 수 있는 인코더 측의 정보를 활용해야 합니다. 다른 작업들도 디코더 이전에 프리넷에서 큰 드롭아웃을 사용하거나 [382, 303, 39] 프리넷에서 작은 숨겨진 크기를 병목 현상으로 사용하여 [39], 현재 음성을 예측할 때 이전 음성 프레임을 단순히 복사하는 것을 방지할 수 있습니다 . 액자. 디코더는 인코더 측에서 더 많은 정보를 얻을 수 있으므로 정렬 학습에 도움이 됩니다. 핑 외. [270], Chen et al. [39] 어텐션 얼라인먼트 학습 에 도움이 되는 소스와 타겟 시퀀스 사이의 위치 정보 연결 을 강화할 것을 제안합니다 . Liu et al. CTC [94] 기반 ASR을 주기 손실로 활용하여 생성된 mel- 스펙트로그램이 텍스트 정보를 포함하도록 장려하고, 이는 또한 더 나은 주의 정렬을 위해 인코더-디코더 연결을 향상시킬 수 있습니다.

- 위치 주의. 일부 비자동회귀 생성 모델 [268, 234] 은 위치 정보를 쿼리로 활용하여 인코더의 키와 값을 처리합니다. 이는 병렬 생성을 위해 인코더와 디코더 간의 연결을 구축하는 또 다른 방법입니다.

3.4.2 기간 예측으로 주의 대체

텍스트와 음성 간의 주의 정렬을 개선하면 강력한 문제를 어느 정도 완화할 수 있지만 완전히 피할 수는 없습니다. 따라서 일부 작업 [290, 418, 159, 69] 은 인코더-디코더 주의 를 완전히 제거하고 각 문자/음소의 지속 시간을 명시적으로 예측하고 지속 시간에 따라 텍스트 숨겨진 시퀀스를 확장하여 멜의 길이와 일치하도록 제안합니다. 스펙트로그램 시퀀스. 그 후 모델은 autoregressive 또는 non-autoregressive 방식으로 mel-spectrogram 시퀀스를 생성할 수 있습니다. 초기 SPSS는 정렬을 위해 duration을 사용 하고, sequence-to-sequence 모델은 duration을 제거하지만 대신 어텐션을 사용하고, 후기 TTS 모델은 어텐션을 버리고 듀레이션을 다시 사용하는 일종의 기술 르네상스라는 점이 매우 흥미롭습니다.

신경 TTS에서 기간 예측을 조사하기 위한 기존 작업은 두 가지 관점에서 분류할 수 있습니다. 1) 기간 레이블을 얻기 위해 외부 정렬 도구를 사용하거나 공동으로 훈련합니다. 2)
엔드-투-엔드 방식으로 지속 시간 예측 최적화 또는 교육에서 실제 지속 시간 사용

추론에서 예측된 기간. 표 12 의 두 가지 관점에 따라 작업을 요약하고 다음과 같이 기술한다.

표 12: 기간 예측에 대한 신경 TTS의 범주.

관점	범주	일하다
외부 내부	외부	FastSpeech 1/2 [290, 292], DurIAN [418], TalkNet [19], [361, 74, 304]
	내부	AlignTTS [429], Glow-TTS [159], EATS [69], [235, 75]
E2E 최적화	E2E 아님	[290, 361, 19, 292, 418, 194, 74, 304, 429, 197, 159]
	E2E	EATS [69], 병렬 타코트론 2 [75]

• 외부 정렬. 외부 정렬 도구[387, 94, 232, 193]를 활용하는 작업 은 사용 된 정렬 도구 에 따라 여러 범주로 나눌 수 있습니다 . 음향 모델. SpeedySpeech [361] 는 FastSpeech와 유사한 파이프라인을 따라 autoregressive teacher 모델에서 기간을 추출하지만 전체 네트워크 구조를 순전히 CNN으로 대체합니다.

2) CTC 정렬. Beliaevet al. [19] 는 CTC [94] 기반 ASR 모델을 활용하여 음소와 mel-스펙트로그램 시퀀스 간의 정렬을 제공합니다. 3) HMM 정렬: FastSpeech 2 [292] 는 지속 시간을 얻기 위해 HMM 기반 몬트리올 강제 정렬(MFA) [232] 을 활용합니다.
DurIAN [418], RobuTrans [194], Parallel Tacotron [74], Non-Attentive Tacotron [304]과 같은 다른 작업은 강제 정렬 또는 음성 인식 도구를 사용하여 정렬을 얻습니다. • 내부 정렬. AlignTTS [429] 는 FastSpeech의 기본 모델 구조를 따르지만 동적 프로그래밍 기반 방법을 활용 하여 다단계 교육을 통해 텍스트와 멜 스펙트로그램 시퀀스 간의 정렬을 학습합니다 . JDI-T [197] 는 FastSpeech를 따라 자동회귀 교사 모델에서 기간을 추출 하지만 자동회귀 모델과 비자동회귀 모델을 함께 훈련하므로 2단계 훈련이 필요하지 않습니다. Glow-TTS [159] 는 지속 시간을 추출하기 위해 새로운 단조 정렬 검색을 활용합니다 . EATS [69] 는 보간 및 소프트 동적 시간 왜곡 (DTW) 손실을 활용하여 완전히 종단 간 방식으로 기간 예측을 최적화합니다. • 비종단 간 최적화. 일반적인 지속 시간 예측 방법 [290, 361, 19, 292, 418, 194, 74, 304, 429, 197, 159] 은 일반적으로 훈련을 위해 외부/내부 정렬 도구에서 얻은 지속 시간을 사용하고 추론을 위해 예측된 지속 시간을 사용합니다. 예측된 지속 시간은 멜 스펙트로그램 손실에서 안내 신호(그라디언트)를 수신 하여 종단 간 최적화되지 않습니다 .

• 종단 간 최적화. 더 나은 운율을 달성하기 위해 기간을 공동으로 최적화하기 위해 EATS [69] 는 내부 모듈을 사용하여 기간을 예측하고 기간 보간 및 소프트 DTW 손실의 도움으로 종단 간 기간을 최적화합니다 . Parallel Tacotron 2 [75] 는 차별화 가능한 기간 예측을 보장하기 위해 EATS의 관행을 따릅니다. Non-Attentive Tacotron [304] 은 기간 예측을 위한 준지도 학습을 제안합니다. 여기서 예측 기간은 사용 가능한 기간 레이블이 없는 경우 업샘플링에 사용할 수 있습니다.

3.4.3 AR 생성 강화

자동 회귀 시퀀스 생성은 일반적으로 노출 편향 및 오류 전파로 인해 어려움을 겪습니다 [20, 390]. 노출 편향은 시퀀스 생성 모델이 일반적으로 이전 ground-truth 값을 입력(즉, teacher-forcing)하여 학습하지만 추론에서는 이전 예측 값을 입력으로 사용하여 자동 회귀적으로 시퀀스를 생성하는 것을 말합니다. 훈련과 추론 사이의 불일치는 추론에서 오류 전파를 유발할 수 있으며, 여기서 예측 오류는 생성된 시퀀스를 따라 빠르게 누적될 수 있습니다.

일부 작업에서는 노출 편향 및 오류 전파 문제를 완화하기 위한 다양한 방법을 조사했습니다. Guoet al. [99] 실제 데이터와 예측 데이터 의 서로 다른 분포 사이의 불일치를 완화하기 위해 교수 강제 [92] 를 활용합니다. Liu et al. [209] 교사-학생 증류 [111, 164, 343] 를 수행하여 노출 편향 문제를 줄입니다. 여기서 교사는 교사-강제 모드로 훈련되고 학생은 이전에 예측된 값을 입력으로 사용하고 거리를 줄이기 위해 최적화됩니다. 교사 모델과 학생 모델 사이의 숨겨진 상태. 생성된 mel-spectrogram 시퀀스의 오른쪽 부분이 일반적으로 오류 전파로 인해 왼쪽 부분보다 나쁘다는 점을 고려하면 일부 작업은 데이터 증대 [291] 를 위해 왼쪽에서 오른쪽 및 오른쪽에서 왼쪽 생성 [344] 을 모두 활용합니다.] 및 정규화 [452]. Vainer와 Du ek [361] 은 각각에 임의의 가우시안 노이즈를 추가하여 노출 바이어스 및 오류 전파 문제를 완화하기 위해 일부 데이터 확대를 활용합니다.

입력 스펙트로그램 픽셀을 사용하여 예측 오류를 시뮬레이션하고 모델이 시간적으로 더 먼 프레임 을 사용하도록 장려하기 위해 여러 프레임을 무작위 프레임으로 무작위로 교체 하여 입력 스펙트로그램을 저하시킵니다.

3.4.4 AR 생성을 NAR 생성으로 교체

AR 생성의 노출 편향 및 오류 전파 문제 는 위의 방법을 통해 완화할 수 있지만 문제를 완전히 해결할 수는 없습니다. 따라서 일부 작업에서는 이러한 문제를 피하기 위해 비자동회귀 생성을 직접 채택합니다. 주의 사용 또는 기간 예측에 따라 두 가지 범주로 나눌 수 있습니다 . ParaNet [268] 및 Flow-TTS [234] 와 같은 일부 작업 은 병렬 생성에서 텍스트 및 음성 정렬을 위해 위치 주의 [270] 를 사용합니다. FastSpeech [290, 292] 및 EATS [69] 와 같은 나머지 작업은 지속 시간 예측을 사용하여 텍스트와 음성 시퀀스 간의 길이 불일치를 연결합니다.

위 하위 섹션의 소개를 기반으로 표 13과 같이 정렬 학습 및 AR/NAR 생성에 따라 TTS의 새로운 범주 가 있습니다 . [270] 및 TransformerTTS [192]. 2) AR + Non-Attention (Duration), DurIAN [418], RobuTrans [194], Non-Attentive Tacotron [304] 등. 3) ParaNet [268], Flow-TTS [234] 및 VARA-TTS [204] 와 같은 Non AR + Attention . 4) FastSpeech 1/2[290, 292], Glow-TTS[159] 및 EATS[69]와 같은 Non-AR + Non-Attention.

표 13: 정렬 학습 및 AR/NAR 생성에 따른 TTS의 새로운 범주.

와 함께? / 주목?	와 함께	비 AR
주목	타코트론 2[303], 딥보이스 3[270]	ParaNet[268], Flow-TTS[234]
비주의	DurIAN [418], 비 Att Tacotron [304]	FastSpeech [290, 292], EATS [69]

3.5 표현적인 TTS

텍스트 음성 변환의 목표는 이해하기 쉽고 자연스러운 음성을 합성하는 것입니다. 자연성은 합성된 음성의 표현력에 크게 좌우되며, 이는 내용, 음색, 운율, 감정, 스타일 등 여러 특성에 의해 결정됩니다 . 표현력 TTS에 대한 연구는 모델링, 분리, 제어 및 전달을 포함한 광범위한 주제를 다룹니다. 내용, 음색, 운율, 스타일 및 감정 등. 이 하위 섹션에서 이러한 주제를 검토합니다.

표현 음성 합성의 핵심은 일대다 매핑 문제를 처리하는 것입니다. 즉, 동일한 텍스트에 해당하는 지속 시간, 피치, 음량, 화자 스타일, 감정 등 의 측면에서 여러 음성 변형이 있음을 의미합니다. 충분한 입력 정보 없이 일반 L1 손실 [86, 360] 에서 일대다 매핑을 모델링 하면 과도한 스무딩 멜 스펙트로그램 예측 [353, 334]이 발생 합니다. 모든 음성 발화의 표현력 을 캡처하는 대신 데이터 세트로 인해 낮은 품질과 표현력이 떨어지는 음성이 생성됩니다. 따라서 이러한 변화 정보를 입력으로 제공하고 이러한 변화 정보를 더 잘 모델링하는 것이 이 문제를 완화하고 합성 음성의 표현력을 향상시키는 데 중요합니다. 또한, 변형 정보를 입력으로 제공함으로써 변형 정보를 풀고, 제어하고, 전송할 수 있습니다. 합성어; 2) 다른 스타일에 해당하는 변주 정보를 제공함으로써 음성을 이 스타일로 옮길 수 있습니다. 3) 세밀한 음성 제어 및 전송을 위해서는 콘텐츠 및 운율, 음색 및 노이즈 등과 같은 다양한 변형 정보를 분리해야 합니다.

이 하위 섹션의 나머지 부분에서는 먼저 이러한 변형 정보에 대한 포괄적인 분석을 수행한 다음 이러한 변형 정보를 모델링, 분리, 제어 및 전송하기 위한 몇 가지 고급 기술을 소개합니다.

3.5.1 변동 정보의 분류

먼저 음성을 합성하는 데 필요한 정보를 네 가지 측면으로 분류합니다.

- 문자 또는 음소일 수 있는 텍스트 정보는 합성된 음성의 내용(즉, 할 말)을 나타냅니다. 일부 작업 은 합성된 음성 의 품질 과 표현력 을 향상시키는 것을 목표로 향상된 단어 임베딩 또는 텍스트 사전 훈련 을 통해 텍스트의 표현 학습을 개선합니다 [81, 104, 393, 143] .

- 화자의 특성을 나타내는 화자 또는 음색 정보(즉, 말할 사람).
  일부 다중 화자 TTS 시스템은 화자 조회 테이블 또는 화자 인코더를 통해 화자 표현을 명시적으로 모델링합니다 [87, 270, 142, 240, 39].
- 억양, 강세, 말 의 리듬을 다루고 텍스트를 말하는 방법을 나타내는 운율, 스타일 및 감정 정보 [371, 179]. 운율/스타일/감정은 말의 표현력을 향상시키기 위한 핵심 정보 이며, 표현적 TTS에 대한 대부분의 작업은 말 의 운율/스타일/감정을 개선하는 데 초점을 맞추고 있다[309, 383, 321, 85, 359, 324].

- 음성을 전달하는 채널이며 음성 의 내용/화자/운율과는 관련이 없지만 음성 품질에 영향을 미치는 녹음 장치 또는 소음 환경. 이 분야 의 연구 작업은 깨끗한 음성 합성을 위한 풀림, 제어 및 노이즈 제거에 중점을 둡니다 [120, 40, 434].

3.5.2 변동 정보 모델링

표 14에 나와 있는 것처럼 서로 다른 세분성으로 서로 다른 유형의 변형 정보를 모델링하기 위해 많은 방법이 제안되었습니다 .

표 14: 표현 음성 합성을 위한 변형 정보 모델링의 일부 관점.

관점 범주 설명			일하다
정보 유형	명백한	언어/스타일/화자 ID [445, 247, 195, 162, 39]	
		피치/지속 시간/에너지	[290, 292, 181, 158, 239, 365]
	절대적인	레퍼런스 엔코더	[309, 383, 224, 142, 9, 49, 37, 40]
		피트	[119, 4, 443, 120, 324, 325, 74]
		GAN/유동/확산	[224, 186, 366, 234, 159, 141]
		텍스트 사전 훈련	[81, 104, 393, 143]
정보 세분성	언어/화자 수준	다국어/화자 TTS [445, 247, 39]	
	단락 수준	긴 형식의 읽기	[11, 395, 376]
	발화 수준	음색/음향/소음	[309, 383, 142, 321, 207, 40]
	단어/음절 수준	세분화된 정보	[325, 116, 45, 335]
	문자/음소 수준		[188, 324, 430, 325, 45, 40, 189]
	프레임 수준		[188, 158, 49, 434]

정보 유형 모델링 되는 정보의 유형에 따라 작업을 분류할 수 있습니다 . 1) 이러한 변형 정보의 레이블을 명시적으로 얻을 수 있는 명시적 정보 와 2) 이러한 변형 정보를 암묵적으로만 얻을 수 있는 암시적 정보입니다.

명시적 정보의 경우 표현 합성을 위한 모델을 향상시키기 위한 입력으로 직접 사용합니다.
다양한 방법으로 이러한 정보를 얻을 수 있습니다. 1) 라벨링 데이터에서 언어 ID, 화자 ID, 스타일 및 운율을 가져옵니다 [445, 247, 195, 39]. 예를 들어, Prosody 정보는 ToBI [307], AuToBI [294], Tilt [345], INTSINT [112] 및 SLAM [249] 과 같은 일부 주석 스키마에 따라 레이블이 지정될 수 있습니다 . 2) 음성에서 피치 및 에너지 정보를 추출하고 쌍을 이루는 텍스트 및 음성 데이터 [290, 292, 181, 158, 239, 365]에서 길이를 추출합니다.

경우에 따라 사용 가능한 명시적 레이블이 없거나 명시적 레이블 지정은 일반적으로 많은 인력을 필요로 하며 특정 또는 세분화된 변형 정보를 포함할 수 없습니다. 따라서 데이터에서 암시적으로 변형 정보를 모델링할 수 있습니다. 일반적인 암시적 모델링 방법은 다음과 같습니다.

- 참조 인코더 [309, 383, 224, 142, 9, 49, 40, 102]. Skerry-Ryan et al. [309] 운율 을 텍스트 콘텐츠로 인한 변형을 제거한 후 남아 있는 음성 신호의 변형으로 정의하고, 화자

명시적인 주석 이 필요하지 않은 참조 인코더를 통한 음색, 채널 효과 및 모델 운율 . 구체적으로 참조 오디오에서 운율 임베딩을 추출하여 디코더의 입력으로 사용합니다. 학습 중에는 실측 기준 오디오가 사용되고, 추론 중에는 또 다른 기준 오디오가 유사한 운율로 음성을 합성하는 데 사용됩니다. Wang et al. 참조 오디오에서 임베딩을 추출하고 이를 질의로 사용하여(Q/K/V 기반 어텐션[368]을 통해) 스타일 토큰 뱅크에 어텐션 결과 를 TTS 모델의 운율 조건으로 사용합니다. 표현 음성 합성. 스타일 토큰은 TTS 모델의 용량 과 변형을 증가시켜 다양한 종류의 스타일을 학습 하고 데이터 세트의 데이터 샘플 간에 지식 공유를 활성화할 수 있습니다. 스타일 토큰 뱅크의 각 토큰은 다른 말하기 속도 및 감정과 같은 다른 운율 표현을 학습할 수 있습니다. 추론하는 동안 참조 오디오를 사용하여 운율 표현을 참석하고 추출하거나 단순히 하나 이상의 스타일 토큰을 선택하여 음성을 합성할 수 있습니다.

- 변이 자동 인코더 [119, 4, 443, 120, 103, 324, 325, 74]. 장 외. VAE 를 활용 하여 정규화로 사전에 가우스를 사용하여 잠재 공간의 분산 정보를 모델링하고 합성된 스타일에 대한 표현 모델링 및 제어를 가능하게 할 수 있습니다. 일부 연구 [4, 120, 2, 74] 는 VAE 프레임워크를 활용하여 표현 합성을 위한 분산 정보를 더 잘 모델링합니다. • 고급 생성 모델 [224, 186, 366, 234, 159, 70, 141, 185]. 일대다 매핑 문제 를 완화하고 과도한 스무딩 예측과 싸우는 한 가지 방법은 고급 생성 모델을 사용하여 변형 정보를 암묵적으로 학습하는 것입니다. 그러면 다중 모달 분포를 더 잘 모델링할 수 있습니다.

- Text pre-training [81, 104, 393, 143, 98, 454]: 다음과 같이 더 나은 텍스트 표현을 제공할 수 있습니다.
  사전 훈련된 단어 임베딩 또는 모델 매개변수를 사용합니다.

정보 세분성 변형 정보는 여러 세분성으로 모델링할 수 있습니다. 1) 다국어 및 다중 화자 TTS 시스템이 언어 ID 또는 화자 ID 를 사용 하여 언어와 화자를 구별 하는 언어 수준 및 화자 수준 [445, 247, 39]. 2) 문단 수준 [11, 395, 376], 여기서 TTS 모델 은 긴 형식의 읽기를 위해 발화/문장 간의 연결을 고려해야 합니다. 3) 발화 수준 [309, 383, 142, 321, 207, 40], 이 발화의 목재/스타일/운율을 나타내기 위해 참조 음성에서 단일 숨겨진 벡터가 추출됩니다 . 4) 단어/음절 수준 [325, 116, 45, 335], 발화 수준 정보로 커버할 수 없는 세분화된 스타일/운율 정보를 모델링할 수 있습니다. 5) 기간, 피치 또는 운율 정보 와 같은 문자/음소 수준 [188, 324, 430, 325, 45, 40, 189] . 6) 가장 세분화된 정보인 프레임 레벨 [188, 158, 49, 434] . 다른 세분성에 대한 일부 해당 작업은 표 14에서 찾을 수 있습니다.

또한 다양한 세분성을 포함하는 계층 구조로 분산 정보를 모델링하면 표현 합성에 도움이 됩니다. Suniet al. 운율 의 계층 구조가 구어에 본질적으로 존재한다는 것을 보여줍니다. Kenteret al. 프레임 및 음소 수준에서 음절 수준까지 운율 특징을 예측하고 단어 및 문장 수준 특징과 연결 합니다 . Honoet al. 다단계 VAE 를 활용하여 다양한 시간 해상도 잠재 변수를 얻고 거친 수준의 잠재 변수에서 더 미세한 수준의 잠재 변수를 샘플링합니다(예: 발화 수준에서 구 수준으로, 그 다음 단어 수준으로). Sun et al. VAE 를 사용 하여 음소 및 단어 수준 모두에 대한 분산 정보를 모델링 하고 이들을 함께 결합하여 디코더에 공급합니다. Chien과 Lee [45] 는 운율 예측에 대해 연구하고 운율 예측을 개선하기 위해 단어에서 음소 수준까지의 계층적 구조를 제안합니다.

3.5.3 풀기, 제어 및 이송

이 하위 섹션에서는 그림과 같이 [224, 120, 281] 풀기 , [359, 184, 246, 13, 273, 349, 196] 제어 및 [153, 134, 399, 6] 변형 정보 전송 에 대한 기술을 검토합니다. 표 15에서.

Adversarial Training과 Disentangleling with Adversarial Training 여러 스타일이나 운율 정보가 함께 얽혀 있는 경우 더 나은 표현 음성 합성 및 제어를 위해 훈련 중에 이를 풀어야 합니다. Ma et al. 적대적 이고 협력적인 게임 을 통해 콘텐츠 스타일의 풀림 능력과 제어 가능성을 향상시킵니다 . Hsu et al. VAE 프레임워크를 적대적 훈련과 함께 활용하여 화자 정보에서 잡음을 분리합니다. Qianet al. [281] 3개의 병목 재구성을 사용하여 리듬, 피치, 콘텐츠 및 음색 을 분리하기 위한 음성 흐름을 제안 합니다. 장 외. [434] 는 프레임 레벨 노이즈 모델링 및 적대적 훈련을 통해 스피커에서 노이즈를 분리할 것을 제안합니다.

표 15: 표현 음성 합성 에서 풀기, 제어 및 전달을 위한 몇 가지 대표적인 기술 .

기술	설명	일하다
적대적 훈련으로 풀기	제어를 위한 분리 [224, 120, 281, 434]	
제어를 위한 주기 일관성/피드백	스타일/음색 생성 향상 [202, 386, 207, 30, 195]	
제어를 위한 준지도 학습	VAE 및 적대적 훈련 사용 [103, 119, 120, 434, 302]	
추론에서 다른 정보 전송을 위한 분산 정보 변경 [309, 383, 142, 443, 40]		

Cycle Consistency/Feedback Loss for Control 스타일 태그 와 같은 분산 정보 를 입력으로 제공할 때 TTS 모델은 해당 스타일로 음성을 합성해야 합니다.
그러나 제약 조건이 추가되지 않으면 TTS 모델은 분산 정보와 스타일을 따르지 않는 합성 음성을 무시하는 경향이 있습니다. TTS 모델의 제어 가능성을 향상시키기 위해 일부 작업에서는 합성된 음성 이 입력에 분산 정보 를 포함하도록 장려하기 위해 주기 일관성 또는 피드백 손실을 사용할 것을 제안합니다 . Li et al. 피드백 주기가 있는 감정 스타일 분류기를 추가하여 제어 가능한 감정 전달을 수행 합니다. 여기서 분류기는 TTS 모델이 음성을 특정 감정과 합성하도록 권장합니다. Whitehillet al. 스타일 분류기를 사용하여 주어진 스타일의 음성 합성을 장려하기 위해 피드백 손실을 제공 합니다 . 한편, 여러 참조 오디오에서 서로 다른 스타일을 보존하기 위해 서로 다른 스타일 분류기 간의 적대적 학습을 통합합니다. Liu et al. 무작위 로 선택된 오디오가 추론의 기준으로 사용 되기 때문에 훈련과 추론 사이의 불일치를 줄이는 것을 목표로 하는 일치하지 않는 텍스트와 음성 을 훈련하기 위해 ASR을 사용하여 피드백 손실을 제공합니다 . 다른 작업 [244, 207, 30, 305, 399, 6] 은 피드백 손실을 활용하여 스타일 및 스피커 임베딩 등에 대한 제어 가능성을 보장합니다.

제어를 위한 반지도 학습 음성을 제어하는 데 사용되는 일부 속성에는 피치, 지속 시간, 에너지, 운율, 감정, 화자, 소음 등이 있습니다. 각 속성에 대한 레이블이 있으면 합성된 음성을 쉽게 제어할 수 있습니다. 태그를 모델 훈련을 위한 입력으로 사용하고 해당 태그를 사용하여 추론에서 합성된 음성을 제어합니다. 그러나 사용할 수 있는 태그/라벨이 없거나 일부만 사용할 수 있는 경우 이러한 속성을 풀고 제어하는 방법이 어렵습니다.

부분 레이블을 사용할 수 있는 경우 Habib et al. [103] VAE 모델의 잠재성(latent)을 학습하여 정동이나 말하기 속도와 같은 속성을 제어하기 위한 준지도 학습 방법을 제안한다 . 사용 가능한 라벨이 없으면 Hsu et al. [119] 서로 다른 속성을 분리하기 위해 가우시안 혼합 VAE 모델을 제안 하고 Hsu et al. [120], 장 외. 시끄러운 화자에 대해 깨끗한 음성을 합성하기 위해 잡음에서 화자 음색 을 분리하기 위해 그래디언트 반전 또는 적대적 훈련을 활용합니다 .

변환을 위한 변형 정보 변경 변형 정보를 다른 스타일 로 변경하여 합성된 음성의 스타일을 변환할 수 있습니다 . 레이블이 지정된 태그 에 변형 정보가 제공되면 교육에서 음성과 해당 태그를 사용하고 추론에서 해당 태그로 스타일을 전달할 수 있습니다 [445, 247, 195, 39]. 또는 변형 정보에 대한 레이블이 지정된 태그가 없는 경우 위에서 소개한 명시적 또는 암시적 모델링을 통해 교육 중에 음성에서 변형 정보를 얻을 수 있습니다. 피치, 지속 시간 및 에너지는 음성에서 명시적으로 추출할 수 있으며 일부는 잠재적 표현은 참조 인코더 또는 VAE 에 의해 암시적으로 추출될 수 있습니다 . 이러한 방식으로 추론에서 스타일 전달을 달성하기 위해 세 가지 방법으로 변형 정보를 얻을 수 있습니다 . 1) 참조 음성 [309, 383, 142, 443, 49, 40, 399, 6] 에서 추출 2) 텍스트 [321, 290, 324, 430, 292, 40]에서 예측; 3) 잠재 공간에서 샘플링하여 얻음 [383, 443, 119].

3.6 적응형 TTS

적응형 TTS19 는 모든 사용자의 음성을 합성할 수 있는 TTS의 중요한 기능입니다. 음성 적응 [44], 음성 복제 [9], 사용자 정의 음성 [40] 등과 같이 학계와 산업계에서 다른 용어로 알려져 있습니다 . 적응형 TTS는 뜨거운 연구 주제였습니다. 통계적 파라메트릭 음성 합성의 작업은 음성 적응을 연구했으며 [79, 392, 450, 80, 67, 125], 최근의 음성 복제 도전도 많은 참가자를 끌어들입니다 [394, 121, 337, 46]. 적응형 TTS 시나리오에서 소스 TTS 모델(일반적으로 다중 화자 음성 데이터 세트에서 훈련됨)은 일반적으로 각 대상 음성에 대한 적응 데이터가 거의 없는 적응형입니다.

19여기서는 언어, 스타일, 도메인 등이 아닌 다양한 음성에 대한 적응형 TTS에 대해 주로 논의합니다.

우리는 두 가지 관점에서 적응형 TTS에 대한 작업을 검토합니다. 1) 새로운 화자를 지원하기 위해 소스 TTS 모델의 일반화 개선과 다른 도메인에 대한 적응을 다루는 일반 적응 설정. 2) 각 대상 화자에 대한 적응 데이터 및 적응 매개변수 의 감소를 포함하는 효율적인 적응 설정 . 두 가지 관점에서 작업을 Table 16에 정리하여 다음과 같이 소개한다.

표 16: 이 연구는 두 가지 관점에서 적응형 TTS에서 작동합니다.

범주	주제	일하다
일반 적응	모델링 변형 정보[40]	
	데이터 적용 범위 증가[57, 407]	
	교차 음향 적응	[40, 54]
	교차 스타일 적응	[404, 266, 123]
	언어 간 적응	[445, 38, 212]
효율적인 적응	적은 데이터 적응 [44, 9, 177, 240, 446, 49, 40, 236]	
	전사되지 않은 데이터 적응 [403, 133, 221]	
	소수 매개변수 적응 [9, 44, 40]	
	제로 샷 적응 [9, 44, 142, 56]	

3.6.1 일반 적응

소스 모델 일반화 이 범주의 작업은 소스 TTS 모델의 일반화를 개선하는 것을 목표로 합니다. 소스 모델 학습에서 소스 텍스트에는 운율, 화자 음색 및 녹음 환경과 같은 음향 정보가 대상 음성을 생성하기에 충분하지 않습니다.

결과적으로 TTS 모델은 훈련 데이터에 과대적합되기 쉽고 적응 시 새로운 화자에 대한 일반화가 좋지 않습니다. Chenet al. [40] 은 암기 대신 더 나은 일반화로 텍스트-음성 매핑을 학습 하기 위해 필요한 음향 정보를 모델 입력으로 제공하는 음향 조건 모델링을 제안합니다 . 원본 TTS 모델의 일반화를 개선하는 또 다른 방법 은 교육 데이터의 양과 다양성을 늘리는 것입니다. Cooperet al. [57] 소스 TTS 모델을 교육할 때 화자 수를 늘리기 위해 화자 확대를 활용 합니다. 이는 적응에서 보이지 않는 화자로 잘 일반화할 수 있습니다. Yang과 He [407] 는 50개 언어 로케일의 여러 화자로 범용 TTS 모델을 훈련 하여 새 화자에 적응할 때 일반화를 높입니다.

Cross-Domain Adaptation 적응 적 TTS에서 중요한 요소는 적응 음성이 소스 TTS 모델을 훈련시키는 데 사용되는 음성 데이터와 음향 조건이나 스타일이 다르다는 것입니다. 이와 같이 소스 TTS 모델의 일반화를 개선하고 타겟 스피커의 스타일을 지원 하기 위해 특별한 설계를 고려해야 합니다. AdaSpeech [40] 는 녹음 장치, 환경 소음, 악센트, 화자 속도, 화자 음색 등과 같은 음향 조건을 더 잘 모델링하기 위해 음향 조건 모델링을 설계합니다 . 음향 조건이 다른 음성 데이터에 잘 적응해야 합니다.

AdaSpeech 3 [404] 는 특정 채워진 일시 중지 적응, 리듬 적응 및 음색 적응을 설계하여 읽기 스타일의 TTS 모델을 즉흥적인 스타일에 적응시킵니다. 일부 다른 작업 [266, 123] 은 Lombard [266] 또는 속삭임 [123] 과 같은 다양한 말하기 스타일에 걸친 적응을 고려합니다 .
일부 작업 [445, 38, 212, 449, 110, 319, 225, 453, 109] 은 언어 간 음성 전송을 제안합니다 .

3.6.2 효율적인 적응

대략적으로 말하면 적응 데이터가 많을수록 음성 품질은 좋아지지만 데이터 수집 비용이 많이 듭니다. 적응 매개변수의 경우 전체 TTS 모델 [44, 177] 또는 모델의 일부(예: 디코더) [240, 446] 또는 스피커 임베딩 [9, 44, 40] 만 미세 조정할 수 있습니다. 마찬가지로 더 많은 매개 변수를 미세 조정하면 음성 품질이 좋아지지만 메모리 및 배포 비용이 증가합니다. 실제로 우리는 높은 적응 음성 품질을 달성하면서 가능한 적은 데이터와 매개변수를 적응시키는 것을 목표로 합니다. 우리는 이 범주의 작업을 여러 측면으로 나눕니다. 1) 데이터 적응이 거의 없습니다. 2) 소수의 매개변수 적응; 3) 전사되지 않은 데이터 적응; 4) 제로 샷 적응. 이 작품들을 아래와 같이 소개 합니다.

• 데이터 적응이 거의 없습니다. 일부 작품 [44, 9, 177, 240, 446, 49, 46, 40, 236] 은 몇 분에서 몇 초 까지 다양한 텍스트 및 음성 데이터 쌍을 사용 하는 소수의 샷 적응을 수행 합니다. Chienet al. [46] 퓨샷 적응을 위한 다양한 스피커 임베딩을 탐색합니다. Yue et al. 퓨 -샷 적응을 위해 음성 체인을 활용한다 [350] . Chenet al. [40], Ar k et al. [9] 적응 데이터의 양을 달리하여 음성 품질을 비교 한 결과, 데이터 크기가 작을 때(20문장 미만) 적응 데이터가 증가할수록 음성 품질이 빠르게 향상되고 수십 개의 적응 문장에서는 느리게 향상됨을 발견했습니다.

• 매개변수 적응이 거의 없습니다. 많은 사용자/고객을 지원하려면 높은 음성 품질을 유지하면서 메모리 사용량을 줄이기 위해 각 대상 화자에 대해 적응 매개변수가 충분히 작아야 합니다.
예를 들어 각 사용자/음성이 100MB 매개변수를 사용하는 경우 총 메모리 스토리지는 1백만 사용자에 대해 100PB와 동일하며 이는 엄청난 메모리 비용입니다. 일부 작업 은 적응 품질을 유지하면서 적응 매개변수를 가능한 한 적게 줄이도록 제안 합니다. AdaSpeech [40] 는 컨텍스트 매개변수 생성 [272] 을 기반으로 스피커 임베딩 에서 레이어 정규화의 스케일 및 바이어스 매개 변수를 생성 하고 조건부 레이어 정규화 및 스피커 임베딩과 관련된 매개변수 만 미세 조정 하여 우수한 적응 을 달성하는 조건부 레이어 정규화를 제안 합니다. 품질. Mosset al. [240] 은 소수의 음성 샘플만으로 특정 화자의 음성을 합성하는 목표 를 달성하기 위해 베이지안 최적화를 기반으로 화자 마다 다른 모델 하이퍼파라미터를 선택하는 미세 조정 방법을 제안합니다 . • 전사되지 않은 데이터 적응. 많은 시나리오에서 변환 또는 온라인 회의와 같이 해당 대화 내용 없이 음성 데이터만 수집할 수 있습니다. AdaSpeech 2 [403] 는 음성 재구성 및 잠재 정렬 [221] 의 도움으로 음성 적응을 위해 전사되지 않은 음성 데이터를 활용합니다 . Inoueet al. ASR 모델을 사용하여 음성 데이터를 전사하고 전사된 쌍 데이터를 음성 적응에 사용합니다.

• 제로 샷 적응. 일부 작업 [9, 44, 142, 56, 32] 은 기준 오디오가 주어진 스피커 임베딩을 추출하기 위해 스피커 인코더를 활용하는 제로 샷 적응을 수행합니다 . 이 시나리오는 적응 데이터와 매개변수가 필요하지 않기 때문에 매우 매력적입니다. 그러나 특히 대상 화자가 소스 화자와 매우 다른 경우 적응 품질이 충분하지 않습니다.

4 자원

표 17과 같이 오픈 소스 구현, TTS 자습서 및 기조 연설, TTS 과제 및 TTS 코퍼스를 포함하여 TTS의 일부 리소스를 수집합니다.

표 17: TTS 리소스.

오픈 소스 구현	
ESPnet-TTS [105]	https://github.com/espnet/espnet https://
모질라-TTS	github.com/mozilla/TTS https://github.com/
TensorflowTTS	TensorSpeech/TensorflowTTS https://github.com/coqui-ai/TTS
쿡-TTS	https:// github.com/PaddlePaddle/Parakeet https://github.com/
잉꼬	NVIDIA/NeMo https://github.com/ibab/tensorflow-wavenet https://
니모	github.com/r9y9/wavenet_vocoder https://github.com/ basveeling/
웨이브넷	wavenet https://github.com/soroushmehr/sampleRNN_ICLR2017
웨이브넷	https://github.com/sotelo/parrot https://github.com/keithito/
웨이브넷	tacotron https://github.com/Kyubyong/tacotron https:/ /github.com/
샘플RNN	Rayhane-mamah/Tacotron-2 https://github.com/NVIDIA/tacotron2 https://
Char2Wav	github.com/r9y9/deepvoice3_pytorch https://github.com/as-ideas/
타코트론	TransformerTTS https:/ /github.com/xcmyz/FastSpeech https://github.com/
타코트론	ming024/FastSpeech2 https://github.com/descriptinc/melgan-neurips
타코트론 2	https://github.com/seungwonpark/melgan https://github.com /fatchord/
타코트론 2	WaveRNN https://github.com/mozilla/LPCNet https://github.com/NVIDIA/
딥보이스 3	WaveGlow https://github.com/ksw0306/FloWaveNet https://github.com/
트랜스포머TTS	chrisdonahue/wavegan https://github.com/r9y9/gantts https://github.com/
패스트 스피치	kan-bayashi/ParallelWaveGAN https://github.com/jik876/hifi-gan
패스트스피치 2	
멜간	
멜간	
WaveRNN	
LPCNet	
웨이브글로우	
플로웨이브넷	
WaveGAN	
GAN-TTS	
병렬 WaveGAN	
하이파이-GAN	

글로우-TTS	https://github.com/jaywalnut310/glow-tts https://github.com/
플로우트론	NVIDIA/flowtron https://github.com/lmnt-com/diffwave
DiffWave	https://github.com/ivanvovk/WaveGrad https: //github.com/
웨이브그래드	jaywalnut310/vits https://github.com/seungwonpark/
농담	awesome-tts-samples https://github.com/faroit/awesome-
TTS 샘플	python-scientific-audio
오디오용 소프트웨어/도구	

TTS 자습서 및 기조 연설

ISCSLP 2014의 TTS 자습서 [282] https://www.superlectures.com/iscslp2014/tutorial-4-deep-learning-for-sp eech-generation-and-synthesis

ISCSLP 2016의 TTS 자습서 [200] http://staff.ustc.edu.cn/~zhling/download/ISCSLP16_tutorial_DLSPSS.pdf IEICE의 TTS 자습서 [378] https://www.slideshare.net/jyamagis/tutorial-on-endtoend-texttospeech-synthesis-part-1-neural-waveform-modeling 음성 생성 모델 [21] https://www.youtube.com/watch ?v=vEAq_sBf1CA 생성 모델 기반 TTS [423] https://static.googleusercontent.com/media/research.google.com/en//pubs/

archive/45882.pdf http://

INTERSPEECH 기조연설 [354]	www.sp.nitech.ac.jp/~tokuda/INTERSPEECH2019.pdf
ISCSLP 2021의 TTS 자습서 [339]	https://www.microsoft.com/en-us/research/uploads/prod/2021/02/ISCSLP2021
	-TTS-Tutorial.pdf https://
TTS 웨비나 [338]	www.youtube.com/watch?v=MA8PCvmr8B0 https://tts-
IJCAI 2021의 TTS 튜토리얼 [340]	tutorial.github.io/ijcai2021/

TTS 챌린지

눈보라 도전	http://www.festvox.org/blizzard/ https://
제로 리소스 스피치 챌린지	www.zerospeech.com/ http://
ICASSP2021 M2VoC	challenge.ai.iqiyi.com/detail?raceId=5fb2688224954e0b48431fe0 http://www.vc-challenge.org /
음성 변환 챌린지	

TTS Corpora

신체	#Hours	#Speakers	샘플링 속도(kHz)	언어
북극 [173]	7	7	16	영어
VCTK [369]	44	109	48	영어
블리자드-2011 [165]	16.6	1 1	16	영어
블리자드-2013 [166]	319	1	44.1	영어
엘제이스피치 [136]	25	2484	22.05	영어
리브리스피치 [259]	982	2456	16 24	영어
책TTS [428]	586	12 11	22.05	영어
VCC 2018 [214]	1	666	44.1 /	영어
하이파이 TTS [16]	300	120 1	8 44.1	영어
테드리움 [295]	118		48 8	영어
콜홈 [31]	60 10		16 44.1	영어
라이언스피치 [421]	12			영어
CSSMSC [15]		1		만다린 오렌지
HKUST [211]	200	2100		만다린 오렌지
AISHELL-1 [28]	170	400		만다린 오렌지
AISHELL-2 [71]	1000	1991		만다린 오렌지
AISHELL-3 [305]	85	218	44.1	만다린 오렌지
DiDiSpeech-1 [100]	572	4500	48	만다린 오렌지
DiDiSpeech-2 [100]	227	1500	48	만다린 오렌지
JSUT [314]	10	1	48	일본어
카자흐어TTS [243]	93	2	44.1/48	카자흐어
루슬란 [83]	31		44.1 44.1	러시아인
HUI-오디오-코퍼스 [280]	326	1	48 16 16	독일 사람
인디아 바디 [106]	39	122	22.05 48	다국어
M-AILABS [88]	1000	253 /		다국어
MLS [278]	51K	6K		다국어
CSS10 [264]	140	1		다국어
커먼보이스 [7]	2.5K	50K		다국어

5가지 향후 방향

본 논문에서는 신경 텍스트 음성 변환에 대한 설문 조사를 수행했으며 주로 (1) 텍스트 분석, 음향 모델, 보코더 및 완전 종단 간 모델을 포함하는 TTS의 기본 모델과 (2) 몇 가지 고급 주제 에 중점을 두었습니다. 빠른 TTS, 저자원 TTS, 강력한 TTS, 표현력이 풍부한 TTS 및 적응형 TTS를 포함합니다. 간단한 요약으로 표 18에 대표적인 TTS 알고리즘을 나열했습니다. 페이지 제한으로 인해 TTS의 핵심 알고리즘만 검토했습니다. 독자는 음성 변환 [308], 노래하는 음성 합성 [115, 217, 35], 말하는 얼굴 합성[36] 등과 같은 TTS 관련 문제 및 응용에 대해 다른 논문을 참조할 수 있습니다.

우리는 TTS 의 최종 목표에 따라 주로 두 가지 범주로 신경 TTS에 대한 몇 가지 향후 연구 방향을 지적합니다 .

고품질 음성 합성 TTS의 가장 중요한 목표는 고품질 음성 합성입니다. 음성 의 품질은 명료성, 자연성, 표현력, 운율, 감정, 스타일, 견고성, 제어 가능성 등 음성 인식에 영향을 미치는 많은 측면에 의해 결정됩니다 . 추가 개선을 위한 큰 방.

- 강력한 생성 모델. TTS는 강력한 생성 모델로 더 잘 처리할 수 있는 파형 및/또는 음향 기능의 생성을 포함하는 생성 작업 입니다. VAE, GAN, 흐름 또는 확산에 기반한 고급 생성 모델이 음향 모델, 보코더 및 완전 종단 간 모델에 채택되었지만 더 강력하고 효율적인 생성 모델에 대한 연구 노력은 합성 음성의 품질을 더욱 향상시키기 위해 매력적입니다. .

- 더 나은 표현 학습. 텍스트와 음성을 잘 표현 하면 합성된 음성의 품질을 개선할 수 있는 신경 TTS 모델에 도움이 됩니다. 텍스트 사전 훈련에 대한 일부 초기 탐색 은 더 나은 텍스트 표현이 실제로 음성 운율을 향상시킬 수 있음을 나타냅니다. 비지도/자기 지도 학습 및 사전 훈련을 통해 텍스트/음소 시퀀스, 특히 음성 시퀀스에 대한 강력한 표현을 학습 하는 방법 은 도전적이며 추가 탐색할 가치가 있습니다.

- 강력한 음성 합성. 현재 TTS 모델은 잘못된 어텐션 정렬로 인한 단어 건너뛰기 및 반복 문제를 제거하지만 더 긴 텍스트 길이, 다른 텍스트 도메인 등과 같이 훈련 세트에서 다루지 않는 코너 케이스를 만날 때 여전히 견고성 문제가 있습니다. TTS 모델을 서로 다른 도메인으로 변환하는 것은 강력한 합성을 위해 매우 중요합니다.

- 표현/제어/전송 가능한 음성 합성. TTS 모델의 표현력, 제어 가능성 및 전달 가능성은 더 나은 변형 정보 모델링에 의존합니다. 기존 방법은 추론에서 제어 가능성과 전달 가능성이 좋은 변형 모델링을 위해 참조 인코더 또는 명시적 운율 기능(예: 피치, 지속 시간, 에너지)을 활용하지만 교육에 사용되는 실측 기준 음성 또는 운율 기능으로 인해 교육/추론 불일치가 발생합니다. 일반적으로 추론할 수 없습니다. 고급 TTS 모델은 변형 정보를 암묵적으로 캡처합니다. 이 정보는 합성된 음성에서 좋은 표현력을 보이지만 잠재 공간에서 샘플링하여 각 운율 특징(예: 피치, 스타일)을 명시적이고 정확하게 제어하고 전달할 수 없기 때문에 제어 및 전달에서는 좋지 않습니다. 표현/제어/전송 가능한 음성 합성 을 위한 더 나은 방법을 설계하는 방법 도 매력적입니다. • 더 인간과 유사한 음성 합성. TTS 교육에 사용되는 현재 음성 녹음은 일반적으로 일시 중지, 반복, 속도 변경, 다양한 감정 및 오류 가 허용 되지 않는 공식적인 읽기 스타일 입니다. 그러나 일상적인 대화나 대화에서 인간은 거의 표준 읽기처럼 말하지 않습니다. 따라서 캐주얼하고 감성적이며 즉흥적인 스타일을 더 잘 모델링하는 것이 합성어의 자연스러움을 향상시키는 데 중요합니다.

효율적인 음성 합성 고품질 음성을 합성할 수 있게 되면 다음으로 가장 중요한 작업은 효율적인 합성입니다. 즉, 훈련 데이터 수집 및 라벨링 비용, TTS 모델 훈련 및 제공 비용 등 음성 합성 비용을 줄이는 방법입니다.

- 데이터 효율적인 TTS. 많은 저자원 언어는 학습 데이터가 부족합니다. 자원이 적은 언어를 돕기 위해 비지도/반지도 학습 및 교차 언어 전이 학습을 활용하는 방법 은 흥미로운 방향입니다. 예를 들어, ZeroSpeech Challenge [432] 는 텍스트나 언어 지식 없이 음성에서만 학습하는 기술을 탐구하는 좋은 이니셔티브 입니다. 게다가, 음성 적응에서 대상 화자는 일반적으로 적응 데이터가 거의 없으며 이는 데이터 효율적인 TTS를 위한 또 다른 응용 시나리오입니다.

- 파라미터 효율적인 TTS. 오늘날의 신경 TTS 시스템은 일반적으로 수천만 개의 매개 변수가 있는 대규모 신경망을 사용하여 고품질 음성을 합성하므로 제한된 메모리 및 전력 소비로 인해 모바일, IoT 및 기타 저가형 장치의 응용 프로그램을 차단합니다.
  메모리 공간, 전력 소비 및 대기 시간이 적은 작고 가벼운 모델을 설계 하는 것은 이러한 애플리케이션 시나리오에 매우 중요합니다.
- 에너지 효율적인 TTS. 고품질 TTS 모델을 교육하고 제공하는 것은 많은 에너지를 소비하고 많은 탄소를 배출합니다. 에너지 효율 개선(예: TTS 교육 및

추론은 환경을 보호하기 위해 탄소 배출량을 줄이면서 고급 TTS 기술의 혜택을 더 많은 인구에게 제공하는 데 중요 합니다.

표 18: TTS 모델 개요. "AM"은 음향 모델, "Voc"는 보코더, "E2E"는 완전한 종단간 모델, "ling"은 언어적 특징, "ch"는 문자, "ph"는 음소, "ceps"는 켑스트럼을 나타냅니다. , "linS"는 선형 스펙트로그램, "melS"는 mel-스펙트로그램, "wav"는 파형, "FF"는 피드포워드, "AR"은 자동회귀, "∅" 는 조건부 정보 없음, "IS"는 INTERSPEECH를 나타냅니다. .

모델	AM/Voc 데이터 흐름	출판 시간	
웨이브넷 [254]	보크 링 AR →wav	SSW16	2016.09
샘플RNN [233]	Voc ∅ →wav (와 잠재)	ICLR17	2016.12
깊은 목소리 [8]	AM+Voc ch→ ph→ling →wav (와 잠재)	ICML17	2017.02
Char2Wav [315]	E2E ch →ceps AR →wav (와 잠재)	ICLR17 WS 2017.02	
타코트론 [382]	오전 ch/ph →linS →wav (와 잠재)	IS17	2017.03
깊은 목소리 2 [87]	AM+Voc ch→ph →ling FF →wav (와 잠재)	NIPS17	2017.05
DV2-타코트론 [87]	AM+Voc ch →linS →wav (와 잠재)	NIPS17	2017.05
보이스루프 [333]	오전 ph→ceps→wav	ICLR18	2017.07
깊은 목소리 3 [270]	오전 ch/ph →melS →wav (와 잠재)	ICLR18	2017.10
DCTTS [332]	오전 AR ch→melS →wav	ICASSP18 2017.10	
Par.WaveNet [255]	보크 링 FF →wav	ICML18	2017.11
타코트론 2 [303]	오전 ch/ph →melS →wav (와 잠재)	ICASSP18 2017.12	
웨이브GAN [68]	보크 ∅ FF →wav	ICLR19	2018.02
웨이브RNN [150]	보크 링 AR →wav	ICML18	2018.02
DV3-클론 [9]	오전 ch/ph →linS→wav (와 잠재)	신경 IPS18	2018.02
GST-타코트론 [383]	오전 ph →melS→wav (와 잠재)	ICML18	2018.03
레프-타코트론 [309]	오전 ph →melS→wav (와 잠재)	ICML18	2018.03
FFT넷 [145]	당신은 ceps AR →wav	ICASSP18 2018.04 IS18	
VAE 루프 [4]	오전 ph→ceps→wav	2018.04	
SV-타코트론 [142]	오전 ch/ph →melS →wav (와 잠재)	NeurIPS18	2018.06
클라리넷 [269]	E2E 채널/ AR →wav	ICLR19	2018.07
전방공격 [438]	오전 ph →linS→wav (와 잠재)	ICASSP18 2018.07	
MCNN [10]	You linS →wav FF	SPL18	2018.08
트랜스포머TTS [192] AM	ph →melS AR →wav (와 잠재)	AAAI19	2018.09
바다-TTS [44]	보크 링 AR →wav	ICLR19	2018.09
GMVAE-타코트론 [119] AM	ph →melS AR →wav (와 잠재)	ICLR19	2018.10
LPCNet [363]	당신은 ceps AR →wav	ICASSP19 2018.10	
웨이브글로우 [279]	Voc melS →wav FF	ICASSP19 2018.10	
플로웨이브넷 [163]	Voc melS →wav FF	ICML19	2018.11
대학 웨이브RNN [215]	Voc melS →wav (와 잠재)	IS19	2018.11
VAE-TTS [443]	오전 ph →melS AR →wav (와 잠재)	ICASSP19 2018.12	
TTS-스타일화 [224] AM	ch →melS→wav AR	ICLR19	2018.12
AdVoc [245]	Voc melS →linS→wav FF	IS19	2019.04
GAN 노출 [99]	오전 ph →melS AR →wav (와 잠재)	IS19	2019.04
겔프 [149]	Voc melS →wav FF	IS19	2019.04
거의 풀려나 [291]	오전 ph →melS→wav (와 잠재)	ICML19	2019.05
패스트 스피치 [290]	오전 FF ph→melS →wav	NeurIPS19	2019.05
파라넷 [268]	오전 FF ph→melS →wav	ICML20	2019.05
웨이브VAE [268]	Voc melS →wav FF	ICML20	2019.05
멜넷 [367]	오전 AR ch→melS →wav	arXiv19	2019.06
스텝와이즈MA [107]	오전 ph →melS AR →wav (와 잠재)	IS19	2019.06

GAN-TTS [23]		보크 링 →wav	ICLR20	2019.09
두리안 [418]	오전	ph →wav	IS20	2019.09
MB WaveRNN [418]		Voc melS →wav	IS20	2019.09
멜간 [178]		Voc melS →wav	NeurIPS19	2019.10
을 위한. 웨이브GAN [402]		Voc melS →wav	ICASSP20 2019.10	
DCA-타코트론 [17]	오전	ph →wav	ICASSP20 2019.10	
웨이브플로우 [271]		Voc melS →wav	ICML20	2019.12
스퀴즈웨이브 [433]		Voc melS →wav	arXiv20	2020.01
정렬TTS [429]	오전	ch/ph →melS →wav	ICASSP20 2020.03	
로부트랜스 [194]	오전	ph →wav	AAAI20	2020.04
흐름-TTS [234]	오전	ch/ph →melS →wav	ICASSP20 2020.05	
플로트론 [366]	오전	FF ph wav melS	ICLR21	2020.05
글로우-TTS [159]	오전	FF ph wav melS	NeurIPS20	2020.05
JDI-T [197]	오전	FF ph wav melS→melS	IS20	2020.05
토크넷 [19]	오전	→wav	arXiv20	2020.05
MB 멜간 [406]		Voc melS →wav	에스엘티21	2020.05
멀티스피치 [39]	오전	FF ph wav melS	IS20	2020.06
패스트스피치 2 [292]	오전	FF ph wav melS	ICLR21	2020.06
패스트스피치 2s [292]		E2E ph →wav	ICLR21	2020.06
먹다 [69]		E2E 채널/ →wav	ICLR21	2020.06
패스트피치 [181]	오전	FF ph wav melS	ICASSP21 2020.06	
복건 [408]		Voc melS →wav	IS20	2020.07
LR스피치 [396]	오전	FF 채널 wav melS	KDD20	2020.08
스피디스피치 [361]	오전	FF ph wav melS	IS20	2020.08
검정고시 [96]		보크 링 →wav	NeurIPS20	2020.08
SC-WaveRNN [265]		Voc melS →wav	IS20	2020.08
웨이브그래드 [41]		Voc melS →wav	ICLR21	2020.09
디프웨이브 [176]		Voc melS →wav	ICLR21	2020.09
하이파이-GAN [174]		Voc melS →wav	NeurIPS20	2020.10
NonAtt Tacotron [304] AM Para. 타코트		ph →wav	arXiv20	2020.10
론 [74]	오전	ph →wav	arXiv20	2020.10
장치TTS [126]	오전	ph →Ceps→wav	arXiv20	2020.10
웨이브-타코트론 [385]		E2E 채널/ →wav	ICASSP21 2020.11	
데노이스피치 [434]	오전	FF ph wav melS	ICASSP21 2020.12	
효율적인TTS [235]	오전	FF 채널 wav melS	ICML21	2020.12
효율적인TTS-Wav [235]		E2E FF 뵈 →wav	ICML21	2020.12
Multi-SpectroGAN [186] AM		FF ph wav melS	AAAI21	2020.12
LightSpeech [220]	오전	FF ph wav melS	ICASSP21 2021.02	
에게. 타코트론 2 [75]	오전	ph →wav	arXiv21	2021.03
에이다스피치 [40]	오전	FF ph wav melS	ICLR21	2021.03
BVAE-TTS [187]	오전	FF ph wav melS	ICLR21	2021.03
PnG 버트 [143]	오전	ph →wav	IS21	2021.03
빠른 DCTTS [152]	오전	ch →melS →wav	ICASSP21 2021.04	
AdaSpeech 2 [403]	오전	ph →melS →wav ch	ICASSP21 2021.04	
토크넷 2 [18]	오전	→melS →wav	arXiv21	2021.04
트리플엠 [199]		AM+Voc ch →melS →wav	arXiv21	2021.04
차이-TTS [141]	오전	FF ph wav melS	arXiv21	2021.04
대학원-TTS [276]	오전	FF ph wav melS	ICML21	2021.05
프레간 [161]		Voc melS →wav	IS21	2021.06

농담 [160]	E2E	ph FF →wav	ICML21	2021.06
AdaSpeech 3 [404]	오전	FF ph wav melS	IS21	2021.06
PriorGrad-AM [185]	오전	FF ph wav melS	arXiv21	2021.06
PriorGrad-Voc [185]	Voc melS	→wav FF	arXiv21	2021.06
Meta-StyleSpeech [236] AM		FF ph wav melS	ICML21	2021.06

참조

[1] Ronald Brian Adler, George R Rodman, Alexandre Sévigny. 인간 커뮤니케이션의 이해. Holt, Rinehart 및 Winston Chicago, 1991.

[2] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba 및 Roberto Barra Chicote. 표현적인 음성의 원샷 텍스트-음성 합성을 위해 vaes 및 정규화 흐름을 사용합니다 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6179–6183. IEEE, 2020.

[3] Yang Ai와 Zhen-Hua Ling. 통계적 파라메트릭 음성 합성을 위한 진폭 및 위상 스펙트럼의 계층적 생성이 있는 신경 보코더 . 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션 , 28:839–851, 2020.

[4] 아쿠자와 케이, 이와사와 유스케, 마츠오 유타카. Variational Autoencoder 를 사용한 모델링 표현을 통한 표현 음성 합성 . 절차 Interspeech 2018, 페이지 3067–3071, 2018.

[5] Jonathan Allen, Sharon Hunnicutt, Rolf Carlson, Bjorn Granstrom. Mitalk-79: 1979년 mit 텍스트 음성 변환 시스템. 미국 음향학회 저널, 65(S1): S130–S130, 1979.

[6] Xiaochun An, Frank K Soong, Lei Xie. 엔드-투-엔드 신경 tts에서 보이는 음성 스타일 전송 및 보이지 않는 음성 스타일 전송의 성능을 개선합니다 . arXiv 프리프린트 arXiv:2106.10003, 2021.

[7] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers 및 Gregor Weber. 일반 음성: 대규모 다국어 음성 코퍼스. 제12차 언어 자원 및 평가 회의 진행, 4218~4222페이지, 2020.

[8] Sercan Ö Ar k, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman 등. 딥 보이스: 실시간 신경 텍스트 음성 변환. 기계 학습에 관한 국제 회의, 195–204페이지. PMLR, 2017.

[9] Sercan Ö Ar k, Jitong Chen, Kainan Peng, Wei Ping 및 Yanqi Zhou. 몇 가지 샘플을 사용한 신경 음성 복제 . 신경 정보 처리 시스템 에 관한 제32회 국제 회의 진행 , 10040-10050페이지, 2018.

[10] Sercan Ö Ar k, Heewoo Jun, Gregory Diamos. 멀티 헤드를 사용한 빠른 스펙트로그램 반전 컨벌루션 신경망. IEEE 신호 처리 서한, 26(1):94–98, 2018.

[11] Adele Aubin, Alessandra Cervone, Oliver Watts 및 Simon King. 음성 합성 개선 담화 관계로. INTERSPEECH, 페이지 4470–4474, 2019.

[12] Kurniawati Azizah, Mirna Adriani 및 Wisnu Jatmiko. 자원이 적은 언어에서 다국어, 다중 화자 및 스타일 전송 dnn 기반 tts를 위한 계층적 전송 학습 . IEEE 액세스, 8:179798–179812, 2020.

[13] 배재성, 배한빈, 주영선, 이준모, 이경훈, 조훈영 . 문장 수준 조건화를 이용한 종단 간 음성 합성의 말하기 속도 제어.

절차 Interspeech 2020, 페이지 4402–4406, 2020.

[14] Dzmitry Bahdanau, 조경현, Yoshua Bengio. 정렬 및 번역 을 공동으로 학습하여 신경 기계 번역. arXiv 사전 인쇄 arXiv:1409.0473, 2014.

[15] 데이터 베이커. 중국어 표준 북경어 말뭉치. https://www.data-baker.com/open_source.html , 2017.

[16] Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, Yang Zhang. Hi-fi 다중 화자 영어 tts 데이터 세트. arXiv 프리프린트 arXiv:2104.01497, 2021.

[17] Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, Tom Bagby. 강력한 장문 음성 합성을 위한 위치 관련 주의 메커니즘. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6194–6198. IEEE, 2020.

[18] Stanislav Beliaev와 Boris Ginsburg. Talknet 2: 명시적 피치 및 기간 예측을 통한 음성 합성을 위한 비자동회귀 깊 이별 분리 가능 컨볼루션 모델. arXiv 프리프린트 arXiv:2104.08189, 2021.

[19] Stanislav Beliaev, Yurii Rebryk, Boris Ginsburg. Talknet: Fully-convolutional non autoregressive 음 성 합성 모델. arXiv 프리프린트 arXiv:2005.05514, 2020.

[20] Samy Bengio, Oriol Vinyals, Navdeep Jaitly 및 Noam Shazeer. 순환 신경망을 사용한 시퀀스 예측을 위한 예 약된 샘플링 . 신경 정보 처리 시스템에 관한 제28회 국제 회의 절차 - 1권, 1171-1179페이지, 2015년.

[21] 요슈아 벤지오. 음성 및 이미지에 대한 심층 생성 모델. https://www.youtube.co m/watch?v=vEAq_sBf1CA, 2017.

[22] Mengxiao Bi, Heng Lu, Shiliang Zhang, Ming Lei 및 Zhijie Yan. 음성 합성을 위한 심층 피드포워드 순차 메모 리 네트워크. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4794–4798. IEEE, 2018.

[23] Miko aj Binkówski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo 및 Karen Simonyan. 적대적 네트워크 를 사용한 충실도 높은 음성 합성 . 학습 표현에 관한 국제 회의 에서, 2019.

[24] 막시밀리안 비사니와 헤르만 네이. 문자소에서 음소로의 변환을 위한 결합 시퀀스 모델. 스피치 커뮤니케이션, 50(5):434–451, 2008.

[25] 크리스토퍼 M 비숍. 패턴 인식 및 기계 학습. 스프링거, 2006.

[26] Alan Black, Paul Taylor, Richard Caley, Rob Clark. 축제 음성 합성 시스템, 1998.

[27] Alan W Black, Heiga Zen, Keiichi Tokuda. 통계적 파라메트릭 음성 합성. 2007 년 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, 4권, 페이지 IV–1229에서. IEEE, 2007.

[28] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, Hao Zheng. Aishell-1: 오픈 소스 북경어 음성 말뭉치 및 음성 인식 기준선입니다. 2017년 제20회 O-COCOSDA(International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment) 동양 지부 컨퍼런스, 1~5페이지. IEEE, 2017.

[29] Zexin Cai, Yaogen Yang, Chuxiong Zhang, Xiaoyi Qin, Ming Li. 다단계 임베딩 기능 이 있는 조건부 신경망을 사용하여 북경어에 대한 폴리폰 disam biguation . 절차 Interspeech 2019, 페이지 2110–2114, 2019.

[30] Zexin Cai, Chuxiong Zhang, Ming Li. 화자 확인에서 다중 화자 음성 합성, 피드백 제약 조건이 있는 심층 전송까 지. 절차 Interspeech 2020, 페이지 3974–3978, 2020.

[31] Alexandra Canavan, Graff David, 그리고 Zipperlen George. Callhome 미국 영어 연설. https://catalog.ldc.upenn.edu/LDC97S42, 2021.

[32] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Maria Aluisio, Moacir Antonelli Ponti. Sc-glowtts: 효율적인 제로 샷 다중 스피커 텍스트 음성 변환 모델입니다. arXiv 프리프린트 arXiv:2104.05557, 2021.

[33] 채문정, 박규병, 방진현, 서수빈, 박종혁, 김남주, 박 롱훈. 문자소 에서 음소로의 변환을 위한 비순차적 그리디 디코딩을 사용하는 시퀀스 모델에 대한 컨벌루션 시퀀스. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 2486–2490. IEEE, 2018.

[34] William Chan, Navdeep Jaitly, Quoc Le 및 Oriol Vinyals. 듣기, 참석 및 철자: 대규모 어휘 대화 음성 인식을 위한 신경망. ICASSP(음향, 음성 및 신호 처리), 2016 IEEE 국제 회의, 4960–4964페이지. IEEE, 2016.

[35] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin 및 Tie-Yan Liu. Hifisinger: 고성능 신경 노래 음성 합성을 향하여. arXiv preprint arXiv:2009.01776, 2020.

[36] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, Chenliang Xu. 좋은 화두 비디오 세대를 구성하는 요소: 설문 조사 및 벤치마크. arXiv preprint arXiv:2005.03201, 2020.

[37] Liping Chen, Yan Deng, Xi Wang, Frank K Soong, Lei He. 신경 tts에서 운율을 개선 하기 위한 음성 버트 임베딩 . ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6563–6567. IEEE, 2021.

[38] Mengnan Chen, Minchuan Chen, Shuang Liang, Jun Ma, Lei Chen, Shaojun Wang 및 Jing Xiao 신경 화자 임베딩을 사용한 교차 언어, 다중 화자 텍스트 음성 합성.
절차 Interspeech 2019, 페이지 2105–2109, 2019.

[39] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin. Multispeech: 변환기를 사용한 다중 화자 텍스트 음성 변환. INTERSPEECH, 페이지 4024–4028, 2020.

[40] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, sheng zhao 및 Tie-Yan Liu.
Adaspeech: 맞춤형 음성을 위한 적응형 텍스트 음성 변환. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=Drynvt7gg4L.

[41] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, William Chan.
Wavegrad: 파형 생성을 위한 기울기 추정. 2021년 ICLR에서.

[42] 스탠리 F 첸. 문자소에서 음소로의 변환을 위한 조건부 및 결합 모델. 음성 통신 및 기술에 관한 제8차 유럽 회의, 2003년.

[43] Yuan-Jui Chen, Tao Tu, Cheng-chieh Yeh, Hung-Yi Lee. 언어 간 전이 학습을 통한 저자원 언어를 위한 종단 간 텍스트 음성 변환 . 절차 Interspeech 2019, 페이지 2075–2079, 2019.

[44] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie 등. 효율적인 적응형 텍스트 음성 변환을 샘플링합니다. 학습 표현에 관한 국제 회의에서, 2018.

[45] 치엔충밍과 이흥이. 비자동 회귀 음성 합성을 위한 계층적 운율 모델링. 2021년 IEEE SLT(Spoken Language Technology Workshop), 446–453페이지.
IEEE, 2021.

[46] Chien Chung-Ming, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, Hungyi Lee. 다중 화자 다중 스타일 텍스트 음성 변환을 위해 사전 학습되고 학습 가능한 화자 표현을 통합하는 방법을 조사 합니다. arXiv 프리프린트 arXiv:2103.04088, 2021.

[47] Chung-Cheng Chiu와 Colin Raffel. 단조로운 청크 와이즈 어텐션. 국제에서
학습 표현에 관한 회의, 2018.

[48] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina 등. sequence-to-sequence 모델을 사용한 최첨단 음성 인식. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 4774–4778. IEEE, 2018.

[49] 최승우, 한승주, 김동영, 하성주. Attentron: 어텐션 기반 가변 길이 임베딩을 활용한 몇 번의 텍스트 음성 변환. 절차 Interspeech 2020, 페이지 2007–2011, 2020.

[50] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, 조경현, Yoshua Bengio.
음성 인식을 위한 주의 기반 모델. 신경 정보 처리 시스템에 관한 28차 국제 회의 절차 - 1권, 577–585페이지, 2015년.

[51] 민추와 야오첸. 제한되지 않은 북경어 텍스트 에서 운율 구성요소의 경계 찾기 . In International Journal of Computational Linguistics & Chinese Language Processing, 6권, 1호, 2001년 2월: MSRA의 자연어 처리 연구 특별호, 61-82페이지, 2001년.

[52] Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang 및 RJ Skerry-Ryan. 엔드투엔드 음성 합성에서 데이터 효율성을 향상시키기 위한 준지도 교육입니다. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6940–6944. IEEE, 2019.

[53] 세실 H 코커. 조음 역학 및 제어 모델. IEEE 절차, 64(4):
452–460, 1976.

[54] Jian Cong, Shan Yang, Lei Xie, Guoqiao Yu, Guanglu Wan. 도메인 적대적 훈련을 통해 시끄러운 샘플에서 데이터 효율적인 음성 복제 . 절차 Interspeech 2020, 페이지 811–815, 2020.

[55] Jian Cong, Shan Yang, Lei Xie 및 Dan Su. Glow-wavegan: 고 충실도 흐름 기반 음성 합성을 위해 gan 기반 변형 자동 인코더에서 음성 표현을 학습합니다. arXiv 프리프린트 arXiv:2106.10831, 2021.

[56] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, Junichi Yamagishi. 최첨단 신경 화자 임베딩 이 포함 된 Zero-shot 다중 화자 텍스트 음성 변환 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6184–6188. IEEE, 2020.

[57] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Junichi Yamagishi. 화자 증강 이 다중 화자 종단 간 tts를 개선할 수 있습니까? 절차 Interspeech 2020, 페이지 3979–3983, 2020.

[58] Erica Cooper, Xin Wang, Yi Zhao, Yusuke Yasuda, Junichi Yamagishi. 다중 스피커 종단 간 음성 합성 을 위한 사전 학습 전략, 파형 모델 선택 및 음향 구성. arXiv 프리프린트 arXiv:2011.04839, 2020.

[59] 에리카 린제이 쿠퍼. 자원이 적은 언어에 대해 발견된 데이터를 사용한 텍스트 음성 합성.
컬럼비아 대학교 박사 논문, 2019.

[60] Yang Cui, Xi Wang, Lei He, Frank K Soong. lpcnet 기반 신경 합성 을 위한 효율적인 부대역 선형 예측 .
INTERSPEECH, 페이지 3555–3559, 2020.

[61] Dongyang Dai, Li Chen, Yuping Wang, Mu Wang, Rui Xia, Xuchen Song, Zhiyong Wu 및 Yuxuan Wang 사전 훈련된 모델 및 음성 향상을 사용하여 저자원 화자를 위한 잡음 강인 tts arXiv preprint arXiv:2005.12531, 2020.

[62] Marcel de Korte, 김재복, Esther Klabbers. 다국어 모델링을 통한 저자원 언어 에 대한 효율적인 신경 음성 합성 절차 Interspeech 2020, 페이지 2967–2971, 2020.

[63] 페르디낭 드 소쉬르. 일반 언어학 코스. 컬럼비아 대학교 출판부, 2011.

[64] Chuang Ding, Lei Xie, Jie Yan, Weini Zhang, Yang Liu. blstm-rnn 및 임베딩 기능을 사용한 중국어 음성 합성을 위한 자동 운율 예측. 자동 음성 인식 및 이해(ASRU) 에 관한 2015 IEEE 워크샵 , 98–102페이지. IEEE, 2015.

[65] 로랑 딘, 데이비드 크루거, 요슈아 벤지오. 좋음: 비선형 독립 구성 요소
견적. arXiv 사전 인쇄 arXiv:1410.8516, 2014.

[66] Laurent Dinh, Jascha Sohl-Dickstein, Samy Bengio. 실제 nvp를 사용한 밀도 추정. arXiv 사전 인쇄 arXiv:1605.08803, 2016.

[67] Rama Doddipatla, Norbert Braunschweiler, Ranniery Maia. d-벡터를 사용한 dnn 기반 음성 합성 의 화자 적응 . INTERSPEECH, 페이지 3404–3408, 2017.

[68] Chris Donahue, Julian McAuley, Miller Puckette. 적대적인 오디오 합성. ~ 안에
학습 표현에 관한 국제 회의, 2018.

[69] Jeff Donahue, Sander Dieleman, Miko aj Binkowsḱi, Erich Elsen, Karen Simonyan.
엔드투엔드 적대적 텍스트 음성 변환. 2021년 ICLR에서.

[70] Chenpeng Du와 Kai Yu. 음성 합성 에서 전화 수준 운율 모델링을 위한 혼합 밀도 네트워크 . arXiv 프리프린트
arXiv:2102.00851, 2021.

[71] Jiayu Du, Xingyu Na, Xuechen Liu 및 Hui Bu. Aishell-2: 만다린 asr 변형
산업 규모에 대한 연구. arXiv 사전 인쇄 arXiv:1808.10583, 2018.

[72] 호머 더들리와 토마스 H 타노치. 볼프강 폰 켐펠렌의 말하는 기계.
The Journal of the Acoustical Society of America, 22(2):151–166, 1950.

[73] Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao,
Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W Black 등. 제로 리소스 스피치 챌린지 2019: Tts
without t. 절차 Interspeech 2019, 페이지 1088–1092, 2019.

[74] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron Weiss 및 Yonghui Wu. 병렬 타코트론: 비자
동 회귀 및 제어 가능한 tts. arXiv preprint arXiv:2010.11439, 2020.

[75] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Jia Ye, RJ Ryan, Yonghui Wu. Parallel tacotron 2:
미분 가능한 기간 모델링이 있는 비자동회귀 신경 tts 모델입니다. arXiv 프리프린트 arXiv:2103.14574, 2021.

[76] Nikolaos Ellinas, Georgios Vamvoukakis, Konstantinos Markopoulos, Aimilios Chalamandaris,
Georgia Maniati, Panos Kakoulidis, Spyros Raptis, June Sig Sung, Hyoungmin Park, Pirros
Tsiakoulis. 낮은 문장 길이 독립적 대기 시간 으로 고품질 스트리밍 음성 합성 . 절차 Interspeech 2020, 페이지
2022–2026, 2020.

[77] Jesse Engel, Chenjie Gu, Adam Roberts 등. Ddsp: 차별화 가능한 디지털 신호 처리.
학습 표현에 관한 국제 회의에서, 2019.

[78] Yuchen Fan, Yao Qian, Feng-Long Xie, Frank K Soong. 양방향 lstm 기반 순환 신경망을 사용한 Tts 합성.
2014년 제15회 국제 음성커뮤니케이션 학회 연례학술대회에서

[79] Yuchen Fan, Yao Qian, Frank K Soong, Lei He. dnn 기반 tts 합성을 위한 다중 화자 모델링 및 화자 적응. 2015
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4475–
4479. IEEE, 2015.

[80] Yuchen Fan, Yao Qian, Frank K Soong, Lei He. dnn 기반 tts 합성에서 화자 및 언어 분해. 2016 IEEE
International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 5540–5544.
IEEE, 2016.

[81] Wei Fang, Yu-An Chung, James Glass. 심층 사전 훈련된 언어 모델에서 종단 간 음성 합성을 위한 전이 학습을 향
하여 . arXiv 프리프린트 arXiv:1906.07307, 2019.

[82] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptive algorithm for
mel-cepstral analysis of speech. In Proc. ICASSP, volume 1, pages 137–140, 1992.

[83] Lenar Gabdrakhmanov, Rustem Garaev, Evgenii Razinkov. Ruslan: 음성 합성을 위한 러시아어 구어체. 음
성 및 컴퓨터에 관한 국제 회의, 113–121페이지. 스프링거, 2019.

[84] Michael Gadermayr, Maximilian Tschuchnig, Laxmi Gupta, Nils Krämer, Daniel Truhn, D Merhof 및
Burkhard Gess. 이미지 번역에서 다대일 매핑을 처리하기 위한 비대칭 주기 일관성 손실: 허벅지 mr 스캔에 대한 연
구. 2021년 IEEE 18th International Symposium on Biomedical Imaging(ISBI), 페이지 1182–1186.
IEEE, 2021.

[85] Yang Gao, Weiyi Zheng, Zhaojun Yang, Thilo Kohler, Christian Fuegen 및 Qing He.
준지도 방식 전이 학습을 통한 대화형 텍스트 음성 변환. arXiv 프리프린트 arXiv:2002.06758, 2020.

[86] 사이드 가조르와 웨이 장. 음성 확률 분포. IEEE 신호 처리 문자,
10(7):204–207, 2003.

[87] Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman 및 Yanqi Zhou. 딥 보이스 2: 다중 화자 신경 텍스트 음성 변환. 2017년 국정원에서.

[88] 뮌헨 인공 지능 연구소 GmbH. m-ailabs 음성 데이터 세트. https: //www.caito.de/2019/01/the-m-ailabs-speech-dataset/, 2019.

[89] Ian Goodfellow, Yoshua Bengio, Aaron Courville 및 Yoshua Bengio. 딥 러닝, 1권. MIT press Cambridge, 2016.

[90] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. 생성적 적대적 네트워크. 2014년 국정원에서.

[91] Prachi Govalkar, Johannes Fischer, Frank Zalkow 및 Christian Dittmar. 음성 신호 재구성을 위한 최신 신경 보코더 비교 . 프로세스에서 10회 ISCA 음성 합성 워크샵, 7-12페이지, 2019.

[92] Anirudh Goyal, Alex Lamb, Ying Zhang, Saizheng Zhang, Aaron Courville 및 Yoshua Bengio. 교수 강제: 순환 신경망 훈련을 위한 새로운 알고리즘. 신경 정보 처리 시스템에 관한 30차 국제 회의 절차 , 4608-4616페이지 , 2016년.

[93] 알렉스 그레이브스. 순환 신경망으로 시퀀스 생성. arXiv 사전 인쇄 arXiv:1308.0850, 2013.

[94] Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber. 연결 주의 시간적 분류: 순환 신경망으로 분할되지 않은 시퀀스 데이터에 레이블 지정.
기계 학습에 관한 23차 국제 회의 절차, 369–376페이지, 2006년.

[95] 다니엘 그리핀과 재림. 수정된 단시간 푸리에 변환에서 신호 추정.
음향, 음성 및 신호 처리에 관한 IEEE 트랜잭션, 32(2):236–243, 1984.

[96] Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, Nal Kalchbrenner.
병렬 음성 합성을 위한 스펙트럼 에너지 거리. 신경 정보 처리 시스템의 발전, 33, 2020.

[97] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville.
Wasserstein gans의 향상된 훈련. 신경 정보 처리 시스템에 관한 제31회 국제 회의 절차 , 5769–5779페이지, 2017.

[98] Haohan Guo, Frank K Soong, Lei He, Lei Xie. 종단 간 tts를 개선하기 위해 구문 분석된 트리 의 구문 기능 을 활용합니다. 절차 Interspeech 2019, 페이지 4460–4464, 2019.

[99] Haohan Guo, Frank K Soong, Lei He, Lei Xie. 새로운 gan 기반 end-to-end tts 훈련 알고리즘. 절차 Interspeech 2019, 페이지 1288–1292, 2019.

[100] Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han 등 Didispeech: A large scale mandarin speech corpus. arXiv preprint arXiv:2010.09275 , 2020.

[101] Weitong Guo, Hongwu Yang, Zhenye Gan. dnn 기반 북경어-티베트어 교차 언어 음성 합성. 2018년 아시아 태평양 신호 및 정보 처리 협회 연례 정상 회의 및 컨퍼런스(APSIPA ASC), 1702–1707페이지. IEEE, 2018.

[102] Siddharth Gururani, Kilol Gupta, Dhaval Shah, Zahra Shakeri 및 Jervis Pinto. 글로벌 피치 및 라우드니스 기능을 사용하여 신경 텍스트에서 음성으로 운율 전송. arXiv 프리프린트 arXiv:1911.09645, 2019.

[103] Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, Tom Bagby. 제어 가능한 음성 합성 을 위한 준지도 생성 모델링 . 학습 표현에 관한 국제 회의에서, 2019.

[104] 하야시 토모키, 와타나베 신지, 토다 토모키, 다케다 카즈야, Shubham Toshniwal, Karen Livescu. 향상된 텍스트-음성 합성을 위한 사전 훈련된 텍스트 임베딩. 절차
Interspeech 2019, 페이지 4430–4434, 2019.

[105] 하야시 토모키, 야마모토 류이치, 이노우에 카츠키, 요시무라 타케노리, 와타나베 신지, 토다 토모키, 다케다 카즈야, 장유, 쑤 탄. Espnet-tts: 통합되고 재현 가능 하며 통합 가능한 오픈 소스 종단 간 텍스트 음성 변환 툴킷. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 7654–7658 . IEEE, 2020.

[106] Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, et al. 구자라트어, 칸나다어, 말라얄람어, 마라티어, 타밀어 및 텔루구어 음성 합성 시스템 을 구축하기 위한 오픈 소스 다중 화자 음성 말뭉치입니다. 제12차 언어 자원 및 평가 회의 진행, 6494–6503페이지, 2020.

Mutian He, Yan Deng 및 Lei He. 신경 tts에 대한 단계적 단조 주의를 통한 강력한 시퀀스 간 음향 모델링 . 페이지 1293–1297, 2019.

[108] Mutian He, Jingzhou Yang, Lei He. 다국어 byte2speech 텍스트 음성 변환 모델은 소수 음성 언어 학습자입니다. arXiv 프리프린트 arXiv:2103.03541, 2021.

[109] 하메드 헤마티와 데미안 보스. 데이터 효율적인 교차 언어 화자 적응 및 발음 향상 을 위해 ipa 기반 tacotron을 사용합니다 . arXiv 프리프린트 arXiv:2011.06392, 2020.

[110] Ivan Himawan, Sandesh Aryal, Iris Ouyang, Sam Kang, Pierre Lanchantin, Simon King. 교차 언어 합성을 위한 다국어 음향 모델의 화자 적응. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 7629–7633. IEEE, 2020.

[111] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. 신경망에서 지식 추출. arXiv 사전 인쇄 arXiv:1503.02531, 2015.

[112] 다니엘 허스트. 다국어 음성 말뭉치에 대한 운율 자동 분석. 음성 합성의 개선 , 320-327페이지, 2001년.

[113] 조나단 호, 아제이 자인, 피터 아빌. 노이즈 제거 확산 확률 모델. arXiv 프리프린트 arXiv:2006.11239, 2020.

[114] Sepp Hochreiter와 Jürgen Schmidhuber. 장단기 기억. 신경계산, 9
(8):1735–1780, 1997.

[115] Yukiya Hono, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Generative adversarial network 기반의 노래 음성 합성. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6955–6959. IEEE, 2019.

[116] Yukiya Hono, Kazuna Tsuboi, Kei Sawada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Hierarchical multi-grained generative model for expressive speech synthesis. Proc. Interspeech 2020, page2 44.

[117] 쉬포춘과 이흥이. Wg-wavenet: 실시간 고음질 음성 합성
GPU. 절차 Interspeech 2020, 페이지 210–214, 2020.

[118] 수 포춘, 왕춘수안, Andy T Liu, 이흥이. 음성 생성을 위한 강력한 신경 보코딩을 향하여 : 설문 조사. arXiv 프리프린트 arXiv:1912.02461, 2019.

[119] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen 등 제어 가능한 음성 합성을 위한 계층적 생성 모델링 국제 회의에서 학습 표현, 2018.

[120] Hsu Wei-Ning, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, James Glass. 데이터 증대 및 적대적 인수분해 를 통해 음성 합성을 위한 상관 화자와 잡음을 분리합니다. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5901–5905. IEEE, 2019.

[121] Cheng-Hung Hu, Yi-Chiao Wu, Wen-Chin Huang, Yu-Huai Peng, Yu-Wen Chen, Pin-Jui Ku, Tomoki Toda, Yu Tsao 및 Hsin-Min Wang as-nu 시스템 m2voc 챌린지 arXiv preprint arXiv:2104.03009, 2021.

[122] Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik 및 Sachin Ka jarekar. 낮은 품질의 공개 녹음에서 신경 텍스트를 음성으로 변환합니다. 음성 합성 워크샵, 볼륨 10, 2019.

[123] Qiong Hu, Tobias Bleisch, Petko Petkov, Tuomo Raitio, Erik Marchi, Varun Lakshmi narasimhan. 속삭이고 롬바드 신경 음성 합성. 2021년 IEEE SLT(Spoken Language Technology Workshop), 454–461페이지. IEEE, 2021.

[124] Chin-Wei Huang, David Krueger, Alexandre Lacoste, Aaron Courville. 신경 autore gressive 흐름. 기계 학습에 관한 국제 회의, 2078–2087페이지. PMLR, 2018.

[125] Zhiying Huang, Heng Lu, Ming Lei 및 Zhijie Yan. 음성 합성을 위한 선형 네트워크 기반 스피커 적응 . 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5319–5323. IEEE, 2018.

[126] Zhiying Huang, Hao Li 및 Ming Lei. Deviceetts: 적은 공간을 차지하는 빠르고 안정적인 네트워크 온디바이스 텍스트 음성 변환. arXiv 프리프린트 arXiv:2010.15311, 2020.

[127] 앤드류 J 헌트와 앨런 W 블랙. 대규모 음성 데이터베이스를 사용하는 연결 음성 합성 시스템의 단위 선택 . 1996년 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1권, 373-376페이지. IEEE, 1996.

[128] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Rahil Shah, Jaime Lawrence-Trueba. 데이터 증대를 사용하는 저자원 표현 텍스트 음성 변환. arXiv 프리프린트 arXiv:2011.05707,

[129] 황민재, 송은우, Ryuichi Yamamoto, Frank Soong, 강홍구. 선형 예측 구조의 혼합 밀도 네트워크 작업 으로 lpcnet 기반 텍스트 음성 변환 개선 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7219–7223. IEEE, 2020.

[130] 황민재, 야마모토 류이치, 송은우, 김재민. Tts-by-tts: 빠르고 고품질 음성 합성을 위한 Tts 기반 데이터 증대. arXiv 프리프린트 arXiv:2010.13421, 2020.

[131] 이마이 사토시. 멜 주파수 스케일에서 켑스트럴 분석 합성. ICASSP'83에서. 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의, 8권, 93-96페이지. IEEE, 1983.

[132] 이마이 사토시, 스미타 카즈오, 후루이치 치에코. 음성 합성을 위한 Mel log 스펙트럼 근사화(mlsa) 필터. 일본의 전자 및 통신(1부: 통신), 66(2):10–18, 1983.

[133] 이노우에 카츠키, 하라 스나오, 아베 마사노부, 하야시 토모키, 야마모토 류이치, 와타나베 신지. 사전 학습된 모델 을 사용하여 종단 간 음성 합성을 위한 준지도 화자 적응 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7634–7638. IEEE, 2020.

[134] Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, and Yusuke Ijima. Model architectures to extrapolate emotional expressions in dnn-based text-to-speech. Speech Com munication, 126:35–43, 2021.

[135] 스모크 스톤. 음성의 선형 예측 계수의 라인 스펙트럼 표현 신호. 미국 음향학회 저널, 57(S1):S35–S35, 1975.

[136] 이토 키스. lj 음성 데이터 세트. https://keithito.com/LJ-Speech-Dataset/, 2017.

[137] 장원, 임단, 윤재삼. Universal melgan: 여러 도메인에서 충실도가 높은 파형 생성을 위한 강력한 뉴럴 보코더입니다. arXiv 프리프린트 arXiv:2011.09631, 2020.

[138] 아르투르 야니츠키. 광택을 위한 음성 합성에서 포스트 태깅 및 억양 제어를 위한 신경망의 적용 . 소프트 컴퓨팅 및 지능형 시스템(SCIS 2004), 7, 2004.

[139] 크리시나 제인, 안드레아스 라니티스, 크리스 크리스토둘루. 얼굴 이미지 합성을 위한 일대다 신경망 매핑 기술. 응용 프로그램이 있는 전문가 시스템, 39(10): 9778–9787, 2012.

[140] 전제훈과 류양. 음절 기반 음향 및 구문 기능을 사용하여 자동 운율 이벤트 감지. 2009년 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의, 4565–4568페이지. IEEE, 2009.

[141] 정명훈, 김형주, 천성준, 최병진, 김남수.
Diff-tts: 텍스트 음성 변환을 위한 노이즈 제거 확산 모델입니다. arXiv 프리프린트 arXiv:2104.01409, 2021.

[142] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno 등. 화자 확인에서 다중 화자 텍스트 음성 합성으로 학습을 전송 합니다. 신경 정보 처리 시스템에 관한 제32회 국제 회의 절차, 4485–4495페이지, 2018년.

[143] Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang 및 Yonghui Wu.Png bert: Augmented bert on
신경 tts에 대한 음소 및 자소. arXiv 프리프린트 arXiv:2103.15060, 2021.

[144] Yunlong Jiao, Adam Gabrys, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa 및 Viacheslav Klimkov.
병렬 웨이브넷을 사용한 범용 신경 보코딩. arXiv 프리프린트 arXiv:2102.01106, 2021.

[145] Zeyu Jin, Adam Finkelstein, Gautham J Mysore 및 Jingwan Lu. Fftnet: 실시간 화자 의존 신경 보코더. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 2251–2255. IEEE, 2018.

[146] 마이클 I 조던과 톰 M 미첼. 기계 학습: 추세, 관점 및 전망.
사이언스, 349(6245):255–260, 2015.

[147] 댄 주라프스키. 음성 및 언어 처리. 피어슨 교육 인도, 2000.

[148] Lauri Jewel, Bajibabu Bollepalli, Vassilis Tsiaras 및 Paavo Alku. Glotnet - 통계적 파라메트릭 음성 합성에서 성문 여기를 위한 원시 파형 모델입니다. 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 27(6):1019–1030,

[149] Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi 및 Paavo Alku. Gelp: mel-spectrogram에서 음성 합성을 위한 Gan-excited 선형 예측. 절차 Interspeech 2019, 페이지 694–698, 2019.

[150] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman 및 Koray Kavukcuoglu. 효율적인 신경 음성 합성. 기계 학습에 관한 국제 회의, 2410-2419페이지.
PMLR, 2018.

[151] 카나가와 히로키와 이지마 유스케. 텐서가 있는 경량 lpcnet 기반 뉴럴 보코더
분해. 절차 Interspeech 2020, 페이지 205–209, 2020.

[152] 강민수, 이지현, 김시민, 김인정. 빠른 dctts: 효율적인 심층 컨벌루션 텍스트 음성 변환. arXiv 프리프린트 arXiv:2104.00624, 2021.

[153] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, Thomas Drugman. Copycat: 신경 텍스트 음성 변환 을 위한 다대다 세분화된 운율 전송 . 절차 Interspeech 2020, 페이지 4387–4391, 2020.

Kyle Kastner, João Felipe Santos, Yoshua Bengio, Aaron Courville. tts 합성을 위한 표현 혼합 . ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5906–5910. IEEE, 2019.

[155] 카와하라 히데키. 똑바로, 보코더의 다른 측면의 착취: 음성 소리의 지각 동형 분해. 음향과학과 기술, 27(6):349–353, 2006.

[156] 가와하라 히데키, 마스다-가츠세 이쿠요, 알랭 드 슈베뉴. 피치 적응형 시간-주파수 평활화 및 순간 주파수 기반 f0 추출을 사용하여 음성 표현 재구조화: 소리에서 반복 구조의 가능한 역할. 스피치 커뮤니케이션, 27(3-4):187–207, 1999.

[157] 카와하라 히데키, 조 에스틸, 후지무라 오사무. 고품질 음성 분석, 수정 및 합성 시스템 스트레이트 를 위한 혼합 모드 여기 및 그룹 지연 조작을 사용하는 비주기성 추출 및 제어 . 2001년 생의학 응용을 위한 음성 방출의 모델 및 분석에 관한 두 번째 국제 워크숍에서 .

[158] Tom Kenter, Vincent Wan, Chun-An Chan, Rob Clark, Jakub Vit. 골파: 언어적으로 구동되는 동적 계층적 조건부 변형 네트워크 를 사용한 음성 합성의 다양한 운율. 기계 학습에 관한 국제 회의, 3331–3340페이지. PMLR, 2019.

[159] 김재현, 김성원, 공정일, 윤성로. Glow-tts: 단조로운 정렬 검색을 통한 텍스트 음성 변환을 위한 생성 흐름입니다. 신경 정보 처리 시스템의 발전, 33, 2020.

[160] 김재현, 공정일, 손주희. 엔드투엔드 텍스트 음성 변환을 위한 적대적 sarial 학습이 포함된 조건부 변형 오토인코더 . arXiv 프리프린트 arXiv:2106.06103, 2021.

[161] 김지훈, 이상훈, 이지현, 이성환. Fre-gan: 적대적 주파수 일치 오디오 합성. arXiv 프리프린트 arXiv:2106.02297, 2021.

[162] 김민찬, 천성준, 최병진, 김종진, 김남수. 스타일 태그를 사용하여 표현 적인 텍스트 음성 변환. arXiv 프리프린트 arXiv:2104.00436, 2021.

[163] 김성원, 이상길, 송종윤, 김재현, 윤성로. Flowavenet: 원시 오디오를 위한 생성 흐름입니다. 기계 학습에 관한 국제 회의, 3370-3378페이지. PMLR, 2019.

[164] 김윤과 알렉산더 M 러쉬. 시퀀스 수준의 지식 증류. 자연어 처리의 경험적 방법에 관한 2016 회의 절차 , 1317-1327페이지 , 2016.

[165] S. 킹 및 V. 카라이스코스. 블리자드 챌린지 2011. 블리자드 챌린지 워크숍에서, 2011.

[166] S. 킹과 V. 카라이스코스. 블리자드 챌린지 2013. 블리자드 챌린지 워크숍에서, 2013.

[167] Diederik P Kingma 및 Prafulla Dhariwal. 글로우: 반전 가능한 1× 1 컨볼루션 을 사용한 생성 흐름 . 신경 정보 처리 시스템에 관한 제32회 국제 회의 진행 , 10236–10245페이지, 2018.

[168] 디데릭 P 킹마와 맥스 웰링. 자동 인코딩 변형 베이. arXiv 사전 인쇄 arXiv:1312.6114, 2013.

[169] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever 및 Max Welling. 역 자기회귀 흐름으로 변동 추론이 개선되었습니다. 신경 정보 처리 시스템의 발전 , 29:4743–4751, 2016.

[170] 로렌스 E 킨슬러, 오스틴 R 프레이, 앨런 B 코펜스, 제임스 V 샌더스. 음향학 의 기초 . 존 와일리 & 아들들, 1999.

[171] 데니스 H 클래트. 캐스케이드/병렬 포먼트 신디사이저용 소프트웨어. 저널 오브 더 미국 음향학회, 67(3):971–995, 1980.

[172] 데니스 H 클래트. 영어에 대한 텍스트 음성 변환 검토. 미국 음향 학회 저널, 82(3):737–793, 1987.

[173] 존 코미넥, 앨런 W 블랙, Ver Ver. 음성 합성을 위한 Cmu arctic 데이터베이스. 2003.

[174] 공정일, 김재현, 배재경. Hifi-gan: 효율적이고 충실도가 높은 음성 합성을 위한 생성적 적대 네트워크. 신경 정보 처리 시스템의 발전, 33, 2020.

[175] Zhifeng Kong과 Wei Ping. 확산 확률 모델의 빠른 샘플링. arXiv 프리프린트 arXiv:2106.00132, 2021.

[176] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao 및 Bryan Catanzaro. Diffwave: 다목적 오디오 합성을 위한 확산 모델. 2021년 ICLR에서.

Zvi Kons, Slava Shechtman, Alex Sorin, Carmel Rabinovitz, Ron Hoory . lpcnet을 사용하는 고품질, 경량 및 적응형 tts. 절차 Interspeech 2019, 페이지 176–180, 2019.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo , Alexandre de Brébisson, Yoshua Bengio 및 Aaron Courville. Melgan: 조건부 파형 합성을 위한 생성적 적대 신경망. NeurIPS에서, 2019.

[179] D 로버트 래드. 국제 음운론. 캠브리지 대학 출판부, 2008.

[180] John Lafferty, Andrew McCallum, Fernando CN Pereira. 조건부 임의 필드: 시퀀스 데이터를 분할하고 레이블을 지정하기 위한 확률적 모델입니다. 2001.

[181] 애드리안 ancucki. Fastpitch: 피치 예측이 포함된 병렬 텍스트 음성 변환. arXiv:2006.06873, arXiv 프리프린트 2020.

[182] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle 및 Ole Winther. 학습된 유사성 메트릭을 사용하여 픽셀 이상으로 자동 인코딩합니다. 기계 학습 에 관한 국제 회의 , 1558–1566 페이지. PMLR, 2016.

[183] Yann LeCun, 요슈아 벤지오, 제프리 힌튼. 딥 러닝. 자연, 521(7553): 436–444, 2015.

[184] 이건, 박규민, 김대영. Styler: 표현력이 풍부하고 제어 가능한 신경 텍스트 음성 변환을 위한 음성 분해를 통해 신속하고 견고하게 스타일 모델링. arXiv 프리프린트 arXiv:2103.09474, 2021.

[185] 이상길, 김희승, 신채훈, 쉬탄, 창류, 기맹, 타오친, 첸웨이, 윤성로, 류태연. Priorgrad: 데이터 기반 적응형 사전으로 조건부 잡음 제거 확산 모델 개선. arXiv 프리프린트 arXiv:2106.06406, 2021.

[186] 이상훈, 윤현욱, 노형래, 김지훈, 이성환. 다중 스펙트로건: 음성 합성을 위한 적대적 스타일 조합을 사용한 고다양성 및 고충실도 스펙트로그램 생성. arXiv 프리프린트 arXiv:2012.07267, 2020.

[187] 이윤형, 신중보, 정교민. 비자동 회귀 텍스트 음성 변환을 위한 양방향 변이 추론 . 학습 표현에 관한 국제 회의에서, 2020.

[188] 이영건, 김태수. 종단 간 음성 합성 의 강력하고 세밀한 운율 제어 . ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5911–5915. IEEE, 2019.

[189] Yi Lei, Shan Yang, Lei Xie. 감정 음성 합성을 위한 세밀한 감정 강도 전달, 제어 및 예측 . 2021년 IEEE SLT(Spoken Language Technology Workshop), 423–430페이지. IEEE, 2021.

[190] 지나 앤 레보우. 조건부 임의 필드와 풍부한 음향 기능을 사용한 자동 운율 레이블 지정. 자연어 처리에 관한 제3차 국제 합동 회의 진행 : 제1권, 2008.

[191] 리하오리, 강용궈, 왕젠위. 강조: 음성 합성 시스템을 위한 감정 음소 기반 음향 모델. 절차 Interspeech 2018, 페이지 3077–3081, 2018.

[192] 나이한 리, 슈지에 류, 옌칭 류, 셍 자오, 밍 류. 변압기 네트워크를 사용한 신경 음성 합성 . 인공 지능에 관한 AAAI 회의 절차, 33권, 6706-6713페이지, 2019년.

[193] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu 및 Ming Zhou Moboaligner: 단조 경계 검색을 사용하는 비자동회귀 tts에 대한 신경 정렬 모델 Proc.
Interspeech 2020, 페이지 3999–4003, 2020.

[194] Naihan Li, Yanqing Liu, Yu Wu, Shujie Liu, Sheng Zhao 및 Ming Liu Robutrans: 강력한 변환기 기반 텍스트 음성 변환 모델 AAAI Conference on Artificial Intelligence, 34권, 페이지에서 8228–8235, 2020.

[195] Tao Li, Shan Yang, Liumeng Xue 및 Lei Xie. 종단 간 음성 합성 을 위한 제어 가능한 감정 전달 . 2021년 제12회 중국어 구어 처리 국제 심포지엄(ISCSLP), 1~5페이지. IEEE, 2021.

[196] Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia 및 Helen Meng. 표현 음성 합성을 위한 다중 스케일 스타일 제어를 향하여. arXiv preprint arXiv:2104.03521, 2021.

[197] 임단, 장원, 오경환, 박해영, 김봉완, 윤재삼. Jdi-t: 명시적 정렬 없이 텍스트 음성 변환을 위한 공동으로 훈련된 기간 정보 변환기.
절차 Interspeech 2020, 페이지 4004–4008, 2020.

[198] 임재현, 예종철. 매우 기하학적입니다. arXiv 프리프린트 arXiv:1705.02894, 2017.

[199] Shilun Lin, Fenglong Xie, Li Meng, Xinhui Li 및 Li Lu. Triple m: 다중 안내 주의 및 다중 대역 다중 시간 lpcnet을 갖춘 실용적인 텍스트-음성 합성 시스템입니다. arXiv 프리프린트 arXiv:2102.00247, 2021.

[200] 젠화 링. 통계적 파라메트릭 음성 합성을 위한 딥 러닝. 2016.

[201] Alexander H Liu, Tao Tu, Hung-yi Lee, Lin-shan Lee. 양자화된 음성 표현 학습을 통한 감독되지 않은 음성 인식 및 합성을 향하여 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7259–7263. IEEE, 2020.

[202] Da-Rong Liu, Chi-Yu Yang, Szu-Lin Wu, Hung-Yi Lee. 종단 간 음성 인식을 사용하여 종단 간 음성 합성에서 감독되지 않은 스타일 전송을 개선합니다. 2018년 IEEE SLT(Spoken Language Technology Workshop), 640–647페이지. IEEE, 2018.

[203] Peng Liu, Xixin Wu, Shiyin Kang, Guangzhi Li, Dan Su 및 Dong Yu 상호 극대화 타코트론 정보입니다. arXiv 프리프린트 arXiv:1909.01145, 2019.

[204] Peng Liu, Yuewen Cao, Songxiang Liu, Na Hu, Guangzhi Li, Chao Weng 및 Dan Su. Vara tts: 잔류 주의력이 있는 매우 깊은 vae에 기반한 비자동 회귀 텍스트 음성 합성 arXiv preprint arXiv: 2102.06431, 2021.

[205] Renyuan Liu, Jian Yang 및 Mengyuan Liu. tacotron2 기반 의 새로운 엔드투엔드 장시간 음성 합성 시스템. 신호 처리 시스템에 관한 2019년 국제 심포지엄 절차 , 46–50페이지, 2019년.

[206] Rui Liu, Berrak Sisman, Feilong Bao, Guanglai Gao 및 Haizhou Li. 타코트론 기반 tts에서 멀티태스크 학습을 통한 운율적 표현 모델링 . IEEE 신호 처리 서한, 27: 1470–1474, 2020.

[207] Rui Liu, Berrak Sisman, Guanglai Gao 및 Haizhou Li. 프레임 및 스타일 재구성 손실을 사용한 표현적인 tts 교육. arXiv 프리프린트 arXiv:2008.01490, 2020.

[208] 루이 리우, 베락 시스만, 하이저우 리. Graphspeech: 신경 음성 합성을 위한 구문 인식 그래프 주의 네트워크 . arXiv 프리프린트 arXiv:2010.12423, 2020.

[209] Rui Liu, Berrak Sisman, Jingdong Li, Feilong Bao, Guanglai Gao 및 Haizhou Li. 강력한 타코트론 기반 tts를 위한 교사 학생 교육. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6274–6278. IEEE, 2020.

[210] Rui Liu, Berrak Sisman, Yixing Lin, Haizhou Li. Fasttalker: 얕은 그룹 자동 회귀를 사용하는 신경 텍스트 음성 변환 아키텍처입니다. 신경망, 141:306–314, 2021.

[211] Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, David Graff. Hkust/mts: 매우 큰 규모의 만다린 전화 음성 코퍼스. 중국어 구어 처리에 관한 국제 심포지엄 , 724-735페이지. 스프링거, 2006.

[212] Zhaoyu Liu와 Brian Mak. 보이지 않는 화자에 대해 병렬 코퍼스를 사용하지 않고 음성 복제를 위한 교차 언어 다중 화자 텍스트 음성 합성 . arXiv 사전 인쇄 arXiv:1911.11601, 2019.

[213] Zhijun Liu, Kuan Chen, Kai Yu 신경 동형 보코더 Proc. Interspeech 2020, 240~244페이지, 2020.

[214] Jaime Lawrence-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio , Tomi Kinnunen, Zhenhua Ling. 음성 변환 챌린지 2018: 병렬 및 비병렬 방법 개발 촉진. 프로세스에서 Odyssey 2018 The Speaker and Language Recognition Workshop, 페이지 195–202,

[215] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet 및 Vatsal Aggarwal. 강력한 범용 신경 보코딩을 달성하기 위해. 절차 Interspeech 2019, 페이지 181–185, 2019.

[216] 루춘후이, 장펑위안, 옌용홍. 중국어 음성 합성을 위한 Self-attention 기반 운율 경계 예측. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7035–7039. IEEE, 2019.

Peiling Lu, Jie Wu, Jian Luan, Xu Tan 및 Li Zhou. Xiaoicesing: 고품질 통합 노래 음성 합성 시스템. 절차 Interspeech 2020, 페이지 1306–1310, 2020.

[218] Yanfeng Lu, Minghui Dong 및 Ying Chen. 중국어 종단 간 음성 합성 에서 운율 표현을 구현 합니다. ICASSP 2019-2019 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing), 7050–7054페이지. IEEE, 2019.

[219] Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Enhong Chen 및 Tie-Yan Liu. gbdt를 사용한 신경 아키텍처 검색 arXiv preprint arXiv:2007.04785, 2020.

[220] Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Jinzhu Li, Sheng Zhao, Enhong Chen 및 Tie-Yan Liu. Lightpeech: 신경 구조 검색을 통한 가볍고 빠른 텍스트 음성 변환. 2021 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing)에서. IEEE, 2021.

[221] Hieu-Thi Luong 및 Junichi Yamagishi. Nautilus: 다목적 음성 복제 시스템. IEEE/ACM 오디오, 음성 및 언어 처리에 대한 트랜잭션, 28:2967–2981, 2020.

[222] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, Nobuyuki Nishizawa. 화자-불균형 음성 말뭉치를 사용하여 다중 화자 신경 텍스트 음성 변환 시스템을 교육 합니다. 절차 인터 스피치 2019, 페이지 1303–1307, 2019.

[223] Mingbo Ma, Baigong Zheng, Kaibo Liu, Renjie Zheng, Hairong Liu, Kainan Peng, Kenneth Church 및 Liang Huang. 접두사 대 접두사 프레임 작업을 사용한 증분 텍스트 음성 합성. 자연어 처리의 경험적 방법에 관한 2020년 회의 진행: 결과, 3886–3896, 2020페이지.

[224] 슈앙 마, 다니엘 맥더프, 예일 송. 적대적 및 협업 게임 을 사용한 신경 tts 스타일화 . 학습 표현에 관한 국제 회의에서, 2018.

[225] Soumi Maiti, Erik Marchi, Alistair Conkie. 이중 언어 화자 데이터를 기반으로 화자 공간 변환을 사용하여 다국어 음성을 생성 합니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7624–7628. IEEE, 2020.

[226] 프레라나스리 말리. 다국어의 텍스트 음성 번역에 대한 설문 조사입니다. Advanced Engineering Technologies ISSN 국제 연구 저널, 2347-2812페이지, 2014년.

Guljamal Mamateli, Askar Rozi, Gulnar Ali 및 Askar Hamdulla. 위구르어 음성 합성을 위한 형태학적 분석 기반 품사 태깅. 지식 엔지니어링 및 관리, 389–396페이지. 스프링거, 2011.

[228] 크리스토퍼 매닝과 힌리히 슈체. 통계적 자연어의 기초 처리. MIT 프레스, 1999.

[229] Courtney Mansfield, Ming Sun, Yuzong Liu, Ankur Gandhe 및 Björn Hoffmeister. 하위 단어 단위를 사용한 중립 텍스트 정규화. 전산 언어학 협회 북미 지부의 2019 회의 진행 : 인간 언어 기술, 2권(산업 논문), 190-196페이지, 2019.

[230] Xinnian Mao, Yuan Dong, Jinyu Han, Dezhi Huang, Haila Wang. 만다린 tts 시스템 에서 다음어 명확화를 위한 문자 기능이 있는 부등식 최대 엔트로피 분류기. 2007년 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의 ICASSP'07, 볼륨 4, 페이지 IV–705. IEEE, 2007.

[231] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, Stephen Paul Smolley. 최소자승 생성적 적대 신경망. 컴퓨터 비전에 관한 IEEE 국제 회의 절차, 2794-2802페이지, 2017.

[232] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner 및 Morgan Sondereg ger. 몬트리올 강제 정렬기: kaldi를 사용하여 훈련 가능한 텍스트-음성 정렬. In Interspeech, 2017년, 498-502페이지, 2017년.

[233] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, Yoshua Bengio. Samplernn: 무조건적인 종단 간 신경 오디오 생성 모델입니다. 2017년 ICLR에서.

[234] Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, Jing Xiao. 흐름 tts: 흐름을 기반으로 하는 텍스트 음성 변환을 위한 비자동회귀 네트워크입니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7209–7213. IEEE, 2020.

[235] Chenfeng Miao, Shuang Liang, Zhencheng Liu, Minchuan Chen, Jun Ma, Shaojun Wang, Jing Xiao Efficienttts: 효율적인 고품질 텍스트 음성 변환 아키텍처 arXiv preprint arXiv:2012.03500, 2020.

[236] 민동찬, 이동복, 양은호, 황성주. Meta-stylespeech: 다중 화자 적응형 텍스트 음성 변환 생성. arXiv 프리프린트 arXiv:2106.03153, 2021.

[237] Devang S Ram Mohan, Raphael Lenain, Lorenzo Foglianti, Tian Huey Teh, Marlene Staib, Alexandra Torresquintero 및 Jiameng Gao. 강화 학습을 사용하는 신경 시퀀스-시퀀스 모델을 위한 증분 텍스트 음성 변환 . 절차 Interspeech 2020, 페이지 3186–3190, 2020.

[238] 모리세 마사노리, 요코모리 후미야, 오자와 겐지. World: 실시간 애플리케이션을 위한 보코더 기반 고품질 음성 합성 시스템. 정보 및 시스템에 관한 IECICE 거래 , 99(7):1877–1884, 2016.

[239] Max Morrison, Zeyu Jin, Justin Salamon, Nicholas J Bryan 및 Gautham J Mysore. 제어 lable 신경 운율 합성. 절차 Interspeech 2020, 페이지 4437–4441, 2020.

[240] Henry B. Moss, Vatsal Aggarwal, Nishant Prateek, Javier González 및 Roberto Barra-Chicote. Boffin tts: 베이지안 최적화에 의한 퓨샷 화자 적응. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 7639–7643. IEEE, 2020.

[241] 에릭 물린과 프랜시스 샤팡티에. 디폰을 이용한 텍스트 음성 합성을 위한 피치 동기 파형 처리 기술 . 스피치 커뮤니케이션, 9(5-6):453–467, 1990.

[242] Zhaoxi Mu, Xinyu Yang, Yizhuo Dong 종단 간 음성 합성 기술 검토 딥러닝 기반. arXiv 프리프린트 arXiv:2104.09995, 2021.

[243] Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov 및 Huseyin Atakan Varol. Kazakhtts: 오픈 소스 카자흐어 텍스트 음성 합성 데이터 세트입니다. arXiv 프리프린트 arXiv:2104.08459, 2021.

[244] Eliya Nachmani, Adam Polyak, Yaniv Taigman 및 Lior Wolf. 전사되지 않은 짧은 샘플을 기반으로 새 스피커를 피팅 합니다. 기계 학습에 관한 국제 회의, 3683–3691페이지. PMLR, 2018.

[245] Paarth Neekhara, Chris Donahue, Miller Puckette, Shlomo Dubnov, Julian McAuley. 적대적인 보코딩으로 tts 합성을 촉진합니다. 절차 Interspeech 2019, 페이지 186–190, 2019.

[246] Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, Farinaz Koushanfar 및 Julian McAuley. 표현적인 신경 음성 복제. arXiv 프리프린트 arXiv:2102.00151, 2021.

[247] Tomá Nekvinda 및 Ond ˇ rej Du ek. 하나의 모델, 다양한 언어: 다국어 텍스트 음성 변환을 위한 메타 학습 . 절차 Interspeech 2020, 페이지 2972–2976, 2020.

[248] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, Liang-Jie Zhang. 심층 학습 기반 음성 합성 리뷰.Applied Sciences, 9(19):4050, 2019.

[249] Nicolas Obin, Julie Beliao, Christophe Veaux, Anne Lacheret. Slam: 음성 멜로디의 자동 스타일 지정 및 레이블 지정. Speech Prosody, 246페이지, 2014년.

[250] Takuma Okamoto, Kentaro Tachibana, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai. 제한된 음향 특성 으로 전체 가청 주파수 범위를 포괄하는 서브밴드 웨이브넷 보코더에 대한 조사 . 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5654–5658. IEEE, 2018.

[251] 오카모토 타쿠마, 토다 토모키, 시가 요시노리, 카와이 히사시. 노이즈 쉐이핑 및 서브밴드 접근 방식으로 fftnet 보코더를 개선 합니다. 2018년 IEEE SLT(Spoken Language Technology Workshop), 304–311페이지. IEEE, 2018.

[252] 오카모토 타쿠마, 토다 토모키, 시가 요시노리, 카와이 히사시. 실용적인 신경 텍스트 음성 변환 시스템을 위한 음소 정렬을 사용하는 Tacotron 기반 음향 모델. 2019년 IEEE 자동 음성 인식 및 이해 워크숍(ASRU), 214–221페이지. IEEE, 2019.

[253] 조셉 올리브. 2가 단위에서 음성의 규칙 합성. ICASSP'77에서. 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의, 2권, 568–570페이지. IEEE, 1977.

[254] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. Wavenet: 원시 오디오를 위한 생성 모델입니다. arXiv 프리프린트 arXiv:1609.03499, 2016.

[255] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg 등 병렬 웨이브넷: 빠른 고충실도 음성 합성. 기계 학습에 관한 국제 회의, 3918–3926페이지. PMLR, 2018.

[256] Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A Hasegawa-Johnson 및 Thomas S Huang. 빠른 웨이브넷 생성 알고리즘. arXiv 프리프린트 arXiv:1611.09482, 2016.

[257] Huashan Pan, Xiulin Li, Zhiqiang Huang 만다린 운율 경계 예측 다중 작업 학습을 기반으로 한 모델입니다. INTERSPEECH, 페이지 4485–4488, 2019.

[258] Junjie Pan, Xiang Yin, Zhiling Zhang, Shichao Liu, Yang Zhang, Zejun Ma, Yuxuan Wang. 북경어 텍스트 음성 합성을 위한 통합 시퀀스-시퀀스 프런트 엔드 모델입니다 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6689–6693. IEEE, 2020.

[259] Vassil Panayotov, Guoguo Chen, Daniel Povey 및 Sanjeev Khudanpur. Librispeech: 퍼블릭 도메인 오디오 북을 기반으로 하는 asr 코퍼스. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5206–5210. IEEE, 2015.

[260] Soumya Priyadarsini Panda, Ajit Kumar Nayak 및 Satyananda Champati Rai. 인도 언어의 음성 합성 기술에 대한 조사 . 멀티미디어 시스템, 26:453–478, 2020.

[261] George Papamakarios, Theo Pavlakou, Iain Murray. 밀도 추정 을 위한 마스킹된 자기회귀 흐름 . 신경 정보 처리 시스템에 관한 제31회 국제 회의 진행 , 2335–2344페이지, 2017.

[262] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, Balaji Lakshminarayanan. 확률적 모델링 및 추론을 위한 흐름 정규화. arXiv 프리프린트 arXiv:1912.02762, 2019.

[263] 박규병과 이세리. g2pm: 새로운 공개 벤치마크 데이터 세트를 기반으로 하는 북경어용 신경 문자소에서 음소로의 변환 패키지 입니다. 절차 Interspeech 2020, 페이지 1723–1727, 2020.

[264] 박규병과 Thomas Mulc. Css10: 단일 화자 음성 데이터 세트 모음
10개 언어. 절차 Interspeech 2019, 페이지 1566–1570, 2019.

[265] Dipjyoti Paul, Yannis Pantazis, Yannis Stylianou. Speaker conditional wavernn: 보이지 않는 스피커 및 녹음 조건에 대한 범용 신경 보코더를 향하여. 절차 Interspeech 2020, 페이지 235–239, 2020.

[266] Dipjyoti Paul, Muhammed PV Shifas, Yannis Pantazis, Yannis Stylianou. 말하기 스타일 변환을 사용하여 텍스트 음성 합성에서 음성 명료도를 향상 시킵니다. 절차 인터 스피치 2020, 페이지 1361–1365, 2020.

[267] Wenzhe Pei, Tao Ge 및 Baobao Chang. 중국어 단어 분할 을 위한 최대 여백 텐서 신경망 . 전산 언어학 협회 제52차 연례 회의 절차 (1권: 긴 논문), 293–303페이지, 2014년.

[268] Kainan Peng, Wei Ping, Zhao Song 및 Kexin Zhao 비자동회귀 신경 텍스트 음성 변환.
기계 학습에 관한 국제 회의, 7586–7598페이지. PMLR, 2020.

[269] Wei Ping, Kainan Peng, Jitong Chen. 클라리넷: 종단 간 병렬 웨이브 생성
텍스트 음성 변환. 학습 표현에 관한 국제 회의에서, 2018.

[270] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman 및 John Miller. Deep voice 3: 2000-speaker 신경 텍스트 음성 변환. 절차
ICLR, 페이지 214–217, 2018.

[271] Wei Ping, Kainan Peng, Kexin Zhao 및 Zhao Song. Waveflow: 원시 오디오를 위한 소형 흐름 기반 모델 입니다. 기계 학습에 관한 국제 회의, 7706–7716페이지. PMLR, 2020.

Emmanouil Antonios Platanios , Mrinmaya Sachan, Graham Neubig, Tom Mitchell. 범용 신경망 기계 번역을 위한 컨텍스트 매개변수 생성. 자연어 처리의 경험적 방법에 관한 2018 회의 절차 , 425–435페이지, 2018.

Adam Polyak , Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Ab delrahman Mohamed 및 Emmanuel Dupoux. 분리된 분리된 자기 감독 표현에서 음성 재합성. arXiv 프리프린트 arXiv:2104.00355, 2021.

[274] 바딤 포포프, 스타니슬라프 카메네프, 미하일 쿠디노프, 세르게이 레피옙스키, 타스니마 사데코바, 블라디미르 크리자노브스키 부샤예프, 데니스 파르호멘코. tacotron2 및 lpcnet을 사용 하는 빠르고 가벼운 온디바이스 tts. 절차 Interspeech 2020, 페이지 220–224, 2020.

[275] 바딤 포포프, 미하일 쿠디노프, 타스니마 사데코바. 다중 샘플 음성 합성 을 위한 Gaussian lpcnet . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6204–6208. IEEE, 2020.

[276] 바딤 포포프, 이반 보브크, 블라디미르 고고리안, 타스니마 사데코바, 미하일 쿠디노프.
Grad-tts: 텍스트 음성 변환을 위한 확산 확률 모델입니다. arXiv 프리프린트 arXiv:2105.06337, 2021.

[277] KR Prajwal 및 CV Jawahar. 신경 tts 시스템을 위한 데이터 효율적인 훈련 전략. 8일
ACM IKDD CODS 및 26번째 COMAD, 223–227페이지. 2021.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve 및 Ronan Collobert. Ms: 음성 연구를 위한 대규모 다국어 데이터 세트입니다. 절차 Interspeech 2020, 페이지 2757–2761, 2020.

Ryan Prenger , Rafael Valle 및 Bryan Catanzaro. Waveglow: 음성 합성을 위한 흐름 기반 생성 네트워크 입니다. ICASSP 2019-2019 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing), 3617–3621페이지. IEEE, 2019.

[280] Pascal Puchtler, Johannes Wirth, René Peinl. Hui-audio-corpus-german: 고품질
tts 데이터셋. arXiv 프리프린트 arXiv:2106.06309, 2021.

[281] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, David Cox. 삼중 정보 병목 현상을 통한 감독되지 않은 음성 분해. 기계 학습 에 관한 국제 회의 , 7836–7846페이지. PMLR, 2020.

[282] Yao Qian과 Frank K Soong. iscslp 2014의 Tts 튜토리얼. https://www.superlectures. com/iscslp2014/
tutorial-4-deep-learning-for-speech-generation-and- synthesis, 2014.

[283] Yao Qian, Zhizheng Wu, Xuezhe Ma 및 Frank Soong. 조건부 임의 필드(crf) 모델을 사용한 자동 운율 예측 및 감지. 2010년 중국어 구어 처리에 관한 제7회 국제 심포지엄, 135–138페이지. IEEE, 2010.

[284] Yao Qian, Yuchen Fan, Wenping Hu, Frank K Soong. 파라메트릭 tts 합성을 위한 심층 신경망(dnn) 의 훈련 측면 .
2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 3829–3833. IEEE, 2014.

[285] 도진. 이중 학습. 스프링거, 2020.

[286] Lawrence Rabiner와 Biinghwang Juang. 은닉 마르코프 모델을 소개합니다. 이야
asp 잡지, 3(1):4–16, 1986.

[287] 알렉 래드포드, 루크 메츠, 수미스 친탈라. 깊은 컨벌루션 생성 적대적 네트워크를 사용한 비지도 학습. arXiv 사전 인쇄
arXiv:1511.06434, 2015.

Colin Raffel , Minh-Thang Luong, Peter J Liu, Ron J Weiss, Douglas Eck. 단조로운 정렬을 적용하여 온라인 및 선형 시간 주의. 기계 학습 에 관한 국제 회의 , 2837–2846페이지. PMLR, 2017.

[289] Kanishka Rao, Fuchun Peng, Ha ¸sim Sak, Françoise Beaufays. 장단기 기억 순환 신경망을 이용한 자소-음소 변환. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 4225–4229.
IEEE, 2015.

[290] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao 및 Tie-Yan Liu.
빠르고 강력하며 제어 가능한 텍스트 음성 변환. NeurIPS에서, 2019.

[291] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao 및 Tie-Yan Liu. 거의 감독되지 않은 텍스트 음성 변환 및 자동 음성 인식. 기계 학습에 관한 국제 회의 , 페이지 5410–5419. PMLR, 2019.

[292] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao 및 Tie-Yan Liu Fastspeech 2: 빠르고 고품질 종단 간 텍스트 음성 변환 국제 학습 표현 회의에서 , 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

[293] 다닐로 레젠데와 샤키르 모하메드. 정규화 흐름을 사용한 변이 추론. ~ 안에
기계 학습에 관한 국제 회의, 1530-1538페이지. PMLR, 2015.

[294] 앤드류 로젠버그. 자동 tobi 주석을 위한 Autobi-a 도구입니다. 2010년 국제 음성 커뮤니케이션 협회 제11차 연례 회의에서.

[295] Anthony Rousseau, Paul Deléglise, Yannick Esteve. Ted-lium: 자동 연설 인식 전용 코퍼스. LREC, 페이지 125–129, 2012.

[296] 스튜어트 러셀과 피터 노빅. 인공 지능: 현대적인 접근 방식. 2002.

[297] Yoshinori Sagisaka, Nobuyoshi Kaiki, Naoto Iwahashi, and Katsuhiko Mimura. Atr μ-talk speech synthesis system. In Second International Conference on Spoken Language Processing, 1992.

[298] 게오르크 아이작 숄룬츠. 자원이 부족한 언어 에 대한 텍스트 음성 합성에 품사 태깅이 미치는 영향 . 2010년 노스웨스트 대학교 박사 논문.

[299] P Seeviour, J Holmes 및 M Judd. 병렬 포만트 음성 합성기에 대한 제어 신호의 자동 생성. ICASSP'76에서. 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의 , 1권, 690-693페이지. IEEE, 1976.

[300] Christine H Shadle과 Robert I Damper. 조음 합성에 대한 전망: 입장 문서. 2001년 음성 합성에 관한 제4회 ISCA 자습서 및 연구 워크숍(ITRW).

[301] Changhao Shan, Lei Xie 및 Kaisheng Yao. 북경어에서 다성어 명확화를 위한 양방향 lstm 접근 방식입니다. 2016년 제10회 중국어 구어 처리에 관한 국제 심포지엄(ISCSLP), 1-5페이지. IEEE, 2016.

[302] Slava Shechtman, Raul Fernandez, David Haws. sequence-to-sequence 음성 합성에서 좁은 어휘 초점을 제어하기 위한 감독 및 비감독 접근법 . 2021년 IEEE SLT(Spoken Language Technology Workshop), 431–437페이지. IEEE, 2021.

[303] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan 등. mel 스펙트로그램 예측에서 웨이브넷을 조건화하여 자연 tts 합성 . 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 4779–4783. IEEE, 2018.

[304] Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, Yonghui Wu. Non-attention tacotron: 비감독 기간 모델링을 포함하여 강력하고 제어 가능한 신경 tts 합성. arXiv 프리프린트 arXiv:2010.04301, 2020.

[305] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, Ming Li Aishell-3: 다중 화자 표준어 말뭉치 및 기준선 arXiv preprint arXiv:2010.11567, 2020.

[306] Desai Siddhi, Jashin M Verghese 및 Desai Bhavik. 텍스트 음성 합성 의 다양한 방법에 대한 조사 . 국제 컴퓨터 응용 저널, 165(6):26–30, 2017.

[307] Kim Silverman, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, Julia Hirschberg. 토비: 영어 운율 표기 기준. 1992 년 구어 처리에 관한 제2차 국제 회의에서.

[308] Berrak Sisman, Junichi Yamagishi, Simon King 및 Haizhou Li. 음성 변환 개요 및 해당 과제: 통계 모델링에서 딥러닝까지. 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 2020.

[309] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, Rif A Saurous. tacotron을 사용한 표현 음성 합성을 위한 종단 간 운율 전송을 향하여. 기계 학습에 관한 국제 회의, 4693–4702페이지. PMLR, 2018.

Jascha Sohl -Dickstein, Eric Weiss, Niru Maheswaranathan 및 Surya Ganguli. 비평형 열역학을 사용한 심층 비지 도 학습. 기계 학습 에 관한 국제 회의 , 2256–2265페이지. PMLR, 2015.

[311] 송은우, 황민재, 야마모토 류이치, 김진섭, 권오성, 김재민 . 세대별 모델링 여기 보코더를 사용한 신경 텍스트 음성 변환. 절차
Interspeech 2020, 페이지 3570–3574, 2020.

[312] 송은우, Ryuichi Yamamoto, 황민재, 김진섭, 권오성, 김재민 . 지각적으로 가중된 스펙트로그램 손실로 개선된 병렬 웨이브건 보코더.
2021년 IEEE SLT(Spoken Language Technology Workshop), 470–476페이지. IEEE, 2021.

[313] Jiaming Song, Chenlin Meng 및 Stefano Ermon. 확산 암시적 모델 노이즈 제거. arXiv preprint arXiv:2010.02502, 2020.

[314] 소노베 료스케, 다카미치 신노스케, 사루와타리 히로시. Jsut 코퍼스: 종단 간 음성 합성을 위한 무료 대규모 일본어 음성 코퍼스. arXiv 사전 인쇄 arXiv:1711.00354, 2017.

[315] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, Yoshua Bengio. Char2wav: 종단 간 음성 합성. 2017.

[316] Richard Sproat와 Navdeep Jaitly. Rnn은 텍스트 정규화에 접근합니다: 도전 과제입니다. arXiv 프리프린트 arXiv:1611.00068, 2016.

[317] Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf 및 Christopher Richards. 비표준 단어의 정규화. Computer speech & language, 15(3):287–333, 2001.

[318] Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore 및 Shrikanth Narayanan. 최대 엔트로피 프레임워크에서 운율 라벨링을 위한 음향 및 구문 기능을 활용합니다 . Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics; 2007년 주 회의 절차, 1-8페이지 .

[319] Marlene Staib, Tian Huey Teh, Alexandra Torresquintero, Devang S Ram Mohan, Lorenzo Foglianti, Raphael Lenain 및 Jiameng Gao. 제로샷 다국어 음성 합성을 위한 음운론적 특징. 절차 Interspeech 2020, 페이지 2942–2946, 2020.

[320] William D Stanley, Gary R Dougherty, Ray Dougherty 및 H Saunders. 디지털 신호 처리. 1988.

[321] Daisy Stanton, Yuxuan Wang, RJ Skerry-Ryan. 종단 간 음성 합성에서 텍스트 로부터 표현적인 말하기 스타일을 예측합니다 . 2018년 IEEE SLT(Spoken Language Technology Workshop), 595–602페이지. IEEE, 2018.

[322] Brooke Stephenson, Laurent Besacier, Laurent Girin, Thomas Hueber. 미래 가 가져오는 것: 증분 신경 tts에 대한 예측의 영향 조사. 절차 Interspeech 2020, 페이지 215–219, 2020.

[323] Brooke Stephenson, Thomas Hueber, Laurent Girin, Laurent Besacier. 대체 결말: 예측된 미래 텍스트 입력으로 증분 신경 tts에 대한 운율 개선. arXiv 프리프린트 arXiv:2102.09914, 2021.

[324] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran 및 Yonghui Wu. 양자화 된 미세 입자 vae 및 자동 회귀 운율 사전을 사용하여 다양하고 자연스러운 텍스트 음성 변환 샘플을 생성 합니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6699–6703. IEEE, 2020.

[325] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen 및 Yonghui Wu. 해석 가능한 음성 합성을 위한 완전 계층적 세분화된 운율 모델링. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 6264–6268. IEEE, 2020.

[326] Hao Sun, Xu Tan, Jun-Wei Gan, Hongzhi Liu, Sheng Zhao, Tao Qin 및 Tie-Yan Liu.
자소-음소 변환을 위한 토큰 수준 앙상블 증류. 2019 년 인터스피치에서 .

[327] Hao Sun, Xu Tan, Jun-Wei Gan, Sheng Zhao, Dongxu Han, Hongzhi Liu, Tao Qin 및 Tie-Yan Liu. 다성어 명확화 를 위한 사전 훈련 및 미세 조정에서 bert의 지식 추출 . 2019년 IEEE 자동 음성 인식 및 이해 워크숍(ASRU), 168–175페이지. IEEE, 2019.

[328] Jingwei Sun, Jing Yang, Jianping Zhang, Yonghong Yan. 조건부 랜덤 필드를 기반으로 한 중국어 운율 구조 예측. 2009년 자연 계산에 관한 제5차 국제 회의, 3권, 602-606페이지. IEEE, 2009.

[329] Ming Sun과 Jerome R Bellegarda. 텍스트-음성 합성을 위한 개선된 pos 태깅. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5384–5387. IEEE, 2011.

[330] Antti Suni, Juraj imko, Daniel Aalto 및 Martti Vainio. 연속 웨이블릿 변환을 이용한 운율의 계층적 표현 및 추정. 컴퓨터 음성 및 언어, 45: 123–136, 2017.

Youcef Tabet과 Mohamed Boughazi. 음성 합성 기술. 설문 조사. 시스템에 대한 국제 워크숍, 신호 처리 및 응용 프로그램, WOSSPA, 67–70페이지. IEEE, 2011.

[332] 타치바나 히데유키, 우에노야마 카츠야, 아이하라 슌스케. 주의 유도가 있는 깊은 컨볼루션 네트워크를 기반으로 효율적으로 훈련 가능한 텍스트 음성 변환 시스템입니다. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4784–4788. IEEE, 2018.

[333] Yaniv Taigman, Lior Wolf, Adam Polyak, Eliya Nachmani. Voiceloop: 음운 루프를 통한 음성 피팅 및 합성. 학습 표현에 관한 국제 회의에서, 2018.

[334] Shinnosuke Takamichi, Tomoki Toda, Alan W Black, Graham Neubig, Sakrani Sakti, Satoshi Nakamura. 통계적 파라메트릭 음성 합성을 위해 변조 스펙트럼을 수정하는 포스트 필터. 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 24 (4):755–767, 2016.

[335] Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, Martti Vainio, et al. 사전 훈련된 문맥화된 단어 표현을 사용하여 텍스트에서 운율적 중요성을 예측합니다. 전산 언어학(NoDaLiDa)에 관한 22차 북유럽 회의에서
회의 절차. Linköping 대학 전자 출판부, 2019.

[336] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. 스피커 종속 웨이브넷 보코더. In Interspeech, 2017권, 1118-1122페이지, 2017.

[337] Daxin Tan, Hingpang Huang, Guangyan Zhang 및 Tan Lee. icassp 2021 m2voc 챌린지를 위한 Cuhk-ee 음성 복제 시스템. arXiv preprint arXiv:2103.04699, 2021.

[338] 쑤 탄. Microsoft 연구 웨비나: 신경 텍스트를 음성으로 전환하기. https: //www.youtube.com/watch?v=MA8PCvmr8B0, 2021.

[339] 쑤 탄. iscslp 2021의 Tts 자습서. https://www.microsoft.com/en-us/research/uploads/prod/2021/02/ISCSLP2021-TTS-Tutorial.pdf , 2021.

[340] Xu Tan 및 Tao Qin. ijcai 2021의 Tts 튜토리얼. https://ijcai-21.org/tutorials/, 2021.

[341] Xu Tan, Jiale Chen, Di He, Yingce Xia, QIN Tao 및 Tie-Yan Liu. 언어 클러스터링을 사용한 다국어 신경 기계 번역. 자연어 처리의 경험적 방법에 관한 2019 회의 및 자연어 처리에 관한 제9차 국제 합동 회의 (EMNLP-IJCNLP), 962-972페이지, 2019년.

[342] Xu Tan, Yichong Leng, Jiale Chen, Yi Ren, Tao Qin, Tie-Yan Liu. 다국어 신경 기계 번역 연구 arXiv preprint arXiv:1912.11625, 2019.

Xu Tan, Yi Ren, Di He, Tao Qin 및 Tie-Yan Liu. 지식 추출을 통한 다국어 신경 기계 번역 . 학습 표현에 관한 국제 회의에서, 2019.
URL https://openreview.net/forum?id=S1gUsoR9YX.

[344] Xu Tan, Yingce Xia, Lijun Wu 및 Tao Qin. 효율적인 양방향 신경 기계 번역. arXiv preprint arXiv:1908.09329, 2019.

[345] 폴 테일러. 틸트 인토네이션 모델. 1998 년 제5차 음성 언어 처리 국제 회의에서.

[346] 폴 테일러. 텍스트 음성 합성. 캠브리지 대학 출판부, 2009.

[347] Qiao Tian, Zewang Zhang, Chao Liu, Heng Lu, Linghui Chen, Bin Wei, Pujiang He 및 Shan Liu. Feathertts: 강력하고 효율적인 주의 기반 신경 tts. arXiv preprint arXiv:2011.00935, 2020.

[348] Qiao Tian, Zewang Zhang, Heng Lu, Ling-Hui Chen, Shan Liu. Featherwave: 다중 대역 선형 예측 기능이 있는 효율적인 고성능 신경 보코더입니다. 절차 Interspeech 2020, 페이지 195–199, 2020.

[349] Noé Tits, Kevin El Haddad, Thierry Dutoit. 표현형 딥러닝 기반 tts 시스템 의 제어 가능성 분석 및 평가 . arXiv 프리프린트 arXiv:2103.04097, 2021.

[350] Andros Tjandra, Sakrani Sakti, Satoshi Nakamura. 말하는 동안 듣기: 딥 러닝에 의한 음성 체인. 2017년 IEEE 자동 음성 인식 및 이해 워크숍(ASRU), 301–308페이지. IEEE, 2017.

[351] Andros Tjandra, Sakrani Sakti, Satoshi Nakamura. 원샷 화자 적응이 있는 기계 음성 체인. 절차 Interspeech 2018, 페이지 887–891, 2018.

[352] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li 및 Satoshi Nakamura. zerospeech Challenge 2019를 위한 Vqvae 감독되지 않은 유닛 발견 및 멀티 스케일 code2spec 인버터 . Proc. Interspeech 2019, 페이지 1118–1122, 2019.

[353] 토다 토모키와 토쿠다 케이이치. hmm 기반 음성 합성을 위한 전역 분산을 고려한 음성 파라미터 생성 알고리즘 . 정보 및 시스템에 관한 IECICE 거래, 90(5):816–824, 2007.

[354] 도쿠다 케이이치. 음성 합성에 대한 통계적 접근: 과거, 현재, 미래. 인터에서
연설, 2019.

[355] 도쿠다 케이이치, 고바야시 타카오, 마스코 타카시, 이마이 사토시. Mel-generalized cepstral analysis - 음성 스펙트럼 추정에 대한 통합 접근 방식. 1994년 제3차 음성 언어 처리 국제 회의에서.

[356] 도쿠다 케이이치, 요시무라 다카요시, 마스코 다카시, 고바야시 다카오, 기타무라 타다시 . hmm 기반 음성 합성을 위한 음성 파라미터 생성 알고리즘. 음향, 음성 및 신호 처리에 관한 2000 IEEE 국제 회의에서. 절차 (Cat. No. 00CH37100), 3권, 1315–1318페이지. IEEE, 2000.

[357] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. Speech synthesis based on hidden markov models. Proceedings of the IEEE, 101(5):1234–1252, 2013.

[358] Tao Tu, Yuan-Jui Chen, Alexander H Liu 및 Hungyi Lee. 불연속 음성 표현을 사용하는 다중 화자 텍스트 음성 합성을 위한 준지도 학습 . 절차 Interspeech 2020, 페이지 3191–3195, 2020.

[359] 엄세윤, 오상신, 변경근, 장인선, 안충현, 강홍구 . 풍부하고 세분화된 제어를 통한 감정적인 음성 합성. ICASSP 2020-2020 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing), 7254–7258페이지. IEEE, 2020.

[360] Mohammed Usman, Mohammed Zubair, Mohammad Shiblee, Paul Rodrigues 및 Syed Jaffar.
최대 우도 추정을 사용하는 스펙트럼 영역에서 음성의 확률적 모델링.
대칭, 10(12):750, 2018.

[361] Jan Vainer와 Ond ˇ rej Du ek. Speedyspeech: 효율적인 신경 음성 합성. 절차
Interspeech 2020, 페이지 3575–3579, 2020.

[362] Cassia Valentini-Botinhao 및 Junichi Yamagishi. 텍스트 음성 변환을 위한 시끄럽고 잔향 음성의 음성 향상 . 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 26(8):1420–1433, 2018.

[363] Jean-Marc Valin과 Jan Skoglund. Lpcnet: 선형 예측 을 통한 신경 음성 합성 개선 . ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5891–5895. IEEE, 2019.

[364] Jean-Marc Valin과 Jan Skoglund. lpcnet 을 사용하는 1.6kb/s의 실시간 광대역 뉴럴 보코더 . 절차 Interspeech 2019, 페이지 3406–3410, 2019.

[365] 라파엘 발레, 제이슨 리, 라이언 프렌저, 브라이언 카탄자로. Mellotron: 리듬, 피치 및 글로벌 스타일 토큰을 조절하여 다중 화자가 표현하는 음성 합성. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6189–6193. IEEE, 2020.

[366] 라파엘 발레, 케빈 시, 라이언 프렌저, 브라이언 카탄자로. Flowtron: 텍스트-음성 합성을 위한 자동회귀 흐름 기반 생성 네트워크. arXiv 프리프린트 arXiv:2005.05957, 2020.

Sean Vasquez와 Mike Lewis . Melnet: 주파수 영역의 오디오 생성 모델입니다. arXiv 프리프린트 arXiv:1906.01083, 2019.

Ashish Vaswani, Noam Shazeer , Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser 및 Illia Polosukhin. 주의가 필요합니다. 신경 정보 처리 시스템의 발전, 페이지 5998–6008, 2017.

[369] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald 등. 대체된 cstr vctk 말뭉치: cstr 음성 복제 툴킷용 영어 다중 화자 말뭉치. 2016.

[370] Ravichander Vipperla, 박상준, 추기현, Samin Ishtiaq, Kyoungbo Min, Sourav Bhattacharya, Abhinav Mehrotra, Alberto Gil CP Ramos, Nicholas D Lane. Bunched lpcnet: 저비용 신경 텍스트 음성 변환 시스템용 보코더. 절차 Interspeech 2020, 페이지 3565–3569, 2020.

[371] Michael Wagner와 Duane G Watson. 운율의 실험적 및 이론적 발전: 리뷰. 언어 및 인지 과정, 25(7-9):905–945, 2010.

[372] Congyi Wang, Yu Chen, Bin Wang 및 Yi Shi 점별 상대론적 최소자승법을 사용하여 gan 기반 신경 보코더 개선 arXiv preprint arXiv:2103.14245, 2021.

[373] Disong Wang, Liqun Deng, Yang Zhang, Nianzu Zheng, Yu Ting Yeung, Xiao Chen, Xunying Liu 및 Helen Meng Fcl-taco2: 빠르고 제어 가능하며 가벼운 텍스트-음성 합성을 향하여.

[374] Peilu Wang, Yao Qian, Frank K Soong, Lei He 및 Hai Zhao. 순환 신경망 기반 tts 합성 을 위한 단어 임베딩 . 2015 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 4879–4883. IEEE, 2015.

[375] Wenfu Wang, Shuang Xu 및 Bo Xu. 종단 간 파라메트릭 tts 합성을 향한 첫 번째 단계: 신경주의로 스펙트럼 매개변수 생성. Interspeech, 페이지 2243–2247, 2016.

Xi Wang, Huaiping Ming, Lei He, Frank K Soong. s-transformer: 강력한 신경 음성 합성을 위한 세그먼트 변환기. arXiv 프리프린트 arXiv:2011.08480, 2020.

왕 신 (Xin Wang)과 야마기시 준이치(Junichi Yamagishi). 텍스트-음성 합성을 위해 훈련 가능한 최대 음성 주파수가 있는 신경 고조파 및 잡음 파형 모델 . 프로세스에서 10회 ISCA 음성 합성 워크숍, 1–6페이지.

[378] Xin Wang과 Yusuke Yasuda. ieice sp 워크샵에서 Tts 튜토리얼. https://www.slidesha re.net/jyamagis/tutorial-on-endtoend-texttospeech-synthesis-part-1-n eural-waveform-modeling, 2019.

[379] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela 및 Junichi Yamagishi. 신경망 기반 음성 합성 을 위한 최근의 파형 생성과 음향 모델링 방법 비교. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4804–4808. IEEE, 2018.

[380] Xin Wang, Shinji Takaki, Junichi Yamagishi. 통계적 파라메트릭 음성 합성을 위한 신경 소스 필터 파형 모델. 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 28:402–415, 2019.

왕 신, 타카키 신지, 야마기시 준이치. 통계적 파라메트릭 음성 합성을 위한 신경 소스 필터 기반 파형 모델. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5916–5920. IEEE, 2019.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio 등 Tacotron: 종단 간 음성 합성을 향하여. Proc. Interspeech 2017, 페이지 4006–4010, 2017.

[383] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren 및 Rif A Saurous. 스타일 토큰: 감독되지 않은 스타일 모델링, 종단 간 음성 합성에서 제어 및 전송. 기계 학습에 관한 국제 회의, 페이지 5180–5189. PMLR, 2018.

[384] Daniel Watson, Jonathan Ho, Mohammad Norouzi 및 William Chan. 확산 확률 모델에서 효율적으로 샘플링하는 방법을 학습합니다 . arXiv 프리프린트 arXiv:2106.03802, 2021.

Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad 및 Diederik P Kingma . Wave-tacotron: 스펙트로그램 없는 종단 간 텍스트 음성 합성. 2021년 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing). IEEE, 2021.

[386] Matt Whitehill, Shuang Ma, Daniel McDuff 및 Yale Song. 적대적 주기 일관성을 가진 다중 참조 신경 tts 양식화. 절차 Interspeech 2020, 페이지 4442–4446, 2020.

[387] Colin W Wightman과 David T Talkin. The aligner: markov 모델 을 사용한 텍스트-음성 정렬 . 음성 합성 진행 중, 313–323페이지. 스프링거, 1997.

[388] 위키백과. 음성 합성 — SpeedyLook 백과사전 http://en.wikipedia.org/w/index.php?title=Speech%20synthesis&age=1020857981 , _

[389] 얀 윈드. 인간 언어 기관의 진화 역사. 언어 기원 연구, 1:173–197, 1989.

[390] Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai 및 Tie-Yan Liu. 신경망 기계 번역의 오류 전파를 넘어서: 언어의 특성도 중요합니다. 자연어 처리의 경험적 방법에 관한 2018 회의 절차, 3602-3611페이지, 2018 .

[391] 우 이치아오, 하야시 토모키, 오카모토 타쿠마, 카와이 히사시, 토다 토모키. 준 주기적 병렬 웨이브건 보코더: 파라메트릭 음성 생성을 위한 비자동회귀 피치 종속 확장 컨볼루션 모델입니다. 절차 Interspeech 2020, 페이지 3535–3539, 2020.

Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, Steve Renals 및 Simon King. dnn 기반 음성 합성을 위한 화자 적응 연구 . 2015년 국제음성커뮤니케이션협회 제16차 연례회의에서 .

[393] Yujia Xiao, Lei He, Huaiping Ming, Frank K Soong. 다중 화자 기반 북경어 신경 tts에서 언어 및 bert 파생 기능을 사용하여 운율을 개선 합니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 6704–6708. IEEE, 2020.

[394] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong 등 멀티 스피커 멀티 스타일 음성 복제 챌린지 2021. arXiv 프리프린트 arXiv:2104.01818, 2021.

[395] Guanghui Xu, Wei Song, Zhengchen Zhang, Chao Zhang, Xiaodong He 및 Bowen Zhou. 엔드투엔드 음성 합성을 위한 상호 발화 버트 임베딩으로 운율 모델링을 개선합니다. arXiv 프리프린트 arXiv:2011.05161, 2020.

[396] Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao 및 Tie-Yan Liu. Lrspeech: 리소스가 매우 적은 음성 합성 및 인식. 지식 발견 및 데이터 마이닝에 관한 제26회 ACM SIGKDD 국제 회의 진행, 2802-2812페이지, 2020.

[397] Jin Xu, Xu Tan, Renqian Luo, Kaitao Song, Jian Li, Tao Qin 및 Tie-Yan Liu. Nas-bert: 신경 구조 검색을 통한 작업 불가지론 및 적응형 크기 bert 압축. 지식 발견 및 데이터 마이닝에 관한 제27회 ACM SIGKDD 국제 회의 진행, 2021 .

[398] Jun Xu, Guohong Fu 및 Haizhou Li. 중국어 텍스트 음성 변환을 위한 자소-음소 변환 . 2004년 제8차 음성 언어 처리 국제 회의에서.

[399] Liumeng Xue, Shifeng Pan, Lei He, Lei Xie 및 Frank K Soong. 엔드투엔드 스타일 전송 tts 교육을 위해 일관된 네트워크를 순환 합니다. 신경망, 140:223–236, 2021.

[400] Nianwen Xue. 문자 태깅으로 중국어 단어 세분화. International Journal of Computational Linguistics & Chinese Language Processing, 8권, 1호, 2003년 2월: 단어 형성 및 중국어 처리 특집, 29–48페이지, 2003년.

[401] 야마모토 류이치, 송은우, 김재민. 고품질 병렬 파형 생성을 위한 생성 적대 신경망 을 사용한 확률 밀도 증류 . 절차 Interspeech 2019, 페이지 699–703, 2019.

[402] 야마모토 류이치, 송은우, 김재민. 병렬 웨이브건(Parallel wavegan): 다중 해상도 스펙트로 그램을 사용하는 생성적 적대 신경망을 기반으로 하는 빠른 파형 생성 모델입니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6199–6203. IEEE, 2020.

[403] Yuzi Yan, Xu Tan, Bohan Li, Tao Qin, Sheng Zhao, Yuan Shen, Tie-Yan Liu. Adaspeech 2: 전사되지 않은 데이터가 있는 적응형 텍스트 음성 변환. 2021 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing)에서. IEEE, 2021.

[404] Yuzi Yan, Xu Tan, Bohan Li, Guangyan Zhang, Tao Qin, Sheng Zhao, Yuan Shen, Wei-Qiang Zhang 및 Tie-Yan Liu. Adaspeech 3: 즉흥적인 스타일을 위한 적응형 텍스트 음성 변환. INTERSPEECH, 2021 .

[405] 야나기타 토모야, 사크리아니 삭티, 나카무라 사토시. Neural itts: 종단간 신경 텍스트 음성 변환 프레임워크를 사용하여 음성을 실시간으로 합성합니다. 제10회 ISCA 음성 합성 워크숍 절차 , 183–188페이지, 2019년.

[406] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, Lei Xie 다중 밴드 멜간: 고품질 텍스트 음성 변환을 위한 빠른 파형 생성 arXiv preprint arXiv:2005.05106, 2020.

[407] Jingzhou Yang과 Lei He. 보편적인 텍스트 음성 변환을 향하여. INTERSPEECH, 페이지 3171–3175, 2020.

[408] 양진혁, 이준모, 김영익, 조훈영, 김인정. Vocgan: 계층적으로 중첩된 적대적 네트워크가 있는 충실도가 높은 실시간 보코더입니다. 절차 Interspeech 2020, 200-204페이지, 2020.

[409] Shan Yang, Lei Xie, Xiao Chen, Xiaoyan Lou, Xuan Zhu, Dongyan Huang 및 Haizhou Li. 다중 작업 학습 프레임워크 에서 생성적 적대 신경망을 사용한 통계적 파라메트릭 음성 합성 . 2017년 IEEE 자동 음성 인식 및 이해 워크숍(ASRU), 685–691페이지. IEEE, 2017.

[410] Kaisheng Yao와 Geoffrey Zweig. 문자소에서 음소로의 변환 을 위한 시퀀스에서 시퀀스로의 신경망 모델 . 2015년 국제 음성 커뮤니케이션 협회 제16차 연례 회의에서.

[411] 야스다 유스케, 왕 신, 야마기시 준이치. 단조로운 잠재 정렬의 주변화를 사용하는 인코더-디코더 종단 간 tts 프레임워크 의 초기 조사 . 2019.

[412] Zhiwei Ying과 Xiaohua Shi. 중국어 tts 에 대한 운율 구문을 감지하는 rnn 기반 알고리즘 입니다. 2001년 음향, 음성 및 신호 처리에 관한 IEEE 국제 회의에서.
절차(Cat. No. 01CH37221), 2권, 809–812페이지. IEEE, 2001.

[413] Sevinj Yolchuyeva, Géza Németh, Bálint Gyres-Tóth. 컨벌루션 신경망을 사용한 텍스트 정규화. 국제 음성 기술 저널, 21(3):589–600, 2018.

[414] 요네야마 레오, 우 이치아오, 토다 토모키. Unified source-filter gan: 준주기 병렬 웨이브건의 분해에 기반한 통합 소스 필터 네트워크. arXiv 프리프린트 arXiv:2104.04668,

[415] 요시무라 타카요시. 음성 및 운율 매개변수의 동시 모델링 및 hmm 기반 텍스트 음성 변환 시스템을 위한 특성 변환. 2002 년 나고야공업대학 박사학위 .

[416] 요시무라 타카요시, 도쿠다 케이이치, 마스코 타카시, 고바야시 타카오, 키타무라 타다시. hmm 기반 음성 합성에서 스펙트럼, 피치 및 기간의 동시 모델링. 음성 통신 및 기술에 관한 제6차 유럽 회의, 1999년.

[417] 유재성, 김달현, 남규현, 황금별, 채경수. Gan 보코더: 다중 해상도 판별기만 있으면 됩니다. arXiv 프리프린트 arXiv:2103.05236, 2021.

[418] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei 등 Durian: 음성 합성을 위한 기간 정보 주의 네트워크.
절차 Interspeech 2020, 페이지 2027–2031, 2020.

[419] Lantao Yu, Weinan Zhang, Jun Wang 및 Yong Yu. Seqgan: 정책 기울기를 사용하여 생성적 적대 신경망을 시퀀싱 합니다. 인공 지능에 관한 AAAI 회의 절차, 31권, 2017.

[420] Fengpeng Yue, Yan Deng, Lei He 및 Tom Ko. 도메인 적응 및 소수 화자 적응 을 위한 기계 음성 체인 탐색 . arXiv 프리프린트 arXiv:2104.03815, 2021.

[421] Rohola Zandie, Mohammad H. Mahoor, Julia Madse 및 Eshrat S. Emamian. Ryanspeech: 대화형 텍스트 음성 합성을 위한 말뭉치. arXiv 프리프린트 arXiv:2106.08468, 2021.

[422] 헤이가 젠. hmm에서 lstm-rnn까지 통계적 파라메트릭 음성 합성의 음향 모델링.
2015.

[423] 헤이가 젠. 생성 모델 기반 텍스트 음성 변환 합성. https://static.googleuse rcontent.com/media/
research.google.com/en//pubs/archive/45882.pdf, 2017.

[424] 헤이가 젠과 하심 삭. 대기 시간이 짧은 음성 합성을 위한 순환 출력 레이어가 있는 단방향 장단기 기억 순환 신경망 . 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4470–4474. IEEE, 2015.

[425] 헤이가 젠, 도쿠다 케이이치, 앨런 W 블랙. 통계적 파라메트릭 음성 합성. 스피치 커뮤니케이션, 51(11):1039–1064, 2009.

[426] 헤이가 젠, 앤드류 시니어, 마이크 슈스터. 심층 신경망 을 사용한 통계적 파라메트릭 음성 합성 . 2013년 음향, 음성 및 신호 처리에 관한 ieee 국제 회의, 페이지 7962–7966. IEEE, 2013.

[427] Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson 및 Przemys aw Szczepaniak. 모바일 장치용 빠르고 컴팩트한 고품질 lstm-rnn 기반 통계적 파라메트릭 음성 합성기. Interspeech 2016, 페이지 2273–2277, 2016.

[428] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen 및 Yonghui Wu Libritts: 텍스트 음성 변환을 위해 librispeech에서 파생된 말뭉치 Proc. Interspeech 2019, 페이지 1526 –1530, 2019.

[429] Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia 및 Jing Xiao. Aligntts: 명시적인 정렬 없이 효율적인 피드 포워드 텍스트 음성 변환 시스템입니다. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 6714–6718 . IEEE, 2020.

[430] Zhen Zeng, Jianzong Wang, Ning Cheng 및 Jing Xiao. 텍스트 길이 제한이 없는 음성 합성 시스템을 위한 Prosody 학습 메커니즘 . 절차 Interspeech 2020, 페이지 4422–4426, 2020.

[431] Zhen Zeng, Jianzong Wang, Ning Cheng 및 Jing Xiao. Lvcnet: 파형 생성을 위한 효율적인 조건 종속 모델링 네트워크. arXiv 프리프린트 arXiv:2102.10815, 2021.

제로스피치. 제로 리소스 스피치 챌린지. https://www.zerospeech.com/.

[433] Bohan Zhai, Tianren Gao, Flora Xue, Daniel Rothchild, Bichen Wu, Joseph E Gonzalez, Kurt Keutzer. Squeezewave: 온디바이스 음성 합성을 위한 초경량 보코더. arXiv 프리프린트 arXiv:2001.05685, 2020.

[434] Chen Zhang, Yi Ren, Xu Tan, Jinglin Liu, Kejun Zhang, Tao Qin, Sheng Zhao 및 Tie-Yan Liu. Denoispeech: 프레임 수준 노이즈 모델링을 사용하여 텍스트를 음성으로 노이즈 제거합니다. 2021 IEEE ICASSP(International Conference on Acoustics, Speech and Signal Processing)에서. IEEE, 2021.

[435] Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang 및 Tie-Yan Liu Uwspeech: 비문자 언어에 대한 음성 번역. AAAI, 2021.

[436] Haitong Zhang과 Yue Lin. 자원이 적은 언어를 위한 sequence-to-sequence text-to-speech를 위한 비지도 학습 . 절차 Interspeech 2020, 페이지 3161–3165, 2020.

[437] Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, Brian Roark. 음성 애플리케이션을 위한 텍스트 정규화의 신경 모델. 전산 언어학, 45(2):293–337, 2019.

[438] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai. 음성 합성을 위한 sequence-to sequence 음향 모델링에 주의를 기울이 십시오. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 4789–4793. IEEE, 2018.

[439] 장 준후이, 판 준지에, 샹 인, 첸 리, 시차오 류, 양 장, 위쉬안 왕, 저준 마. mandarin 을 위한 multi-head self-attention을 이용한 하이브리드 텍스트 정규화 시스템 . ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6694–6698. IEEE, 2020.

[440] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, Junichi Yamagishi. 다중 소스 타코트론 및 웨이브넷을 사용하여 텍스트 음성 변환 및 음성 변환을 위한 공동 교육 프레임워크입니다.
절차 Interspeech 2019, 페이지 1298–1302, 2019.

[441] Shiliang Zhang, Ming Lei, Zhijie Yan 및 Lirong Dai. 대규모 어휘 연속 음성 인식 을 위한 Deep-fsmn . 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 5869–5873. IEEE, 2018.

[442] Weizhao Zhang, Hongwu Yang, Xiaolong Bu 및 Lili Wang 북경어 티베트 언어 간 음성 합성을 위한 딥 러닝 IEEE Access, 7:167884–167894, 2019.

[443] Ya-Jie Zhang, Shifeng Pan, Lei He, Zhen-Hua Ling. 종단 간 음성 합성에서 스타일 제어 및 전송을 위한 잠재 표현을 학습합니다. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 페이지 6945–6949 . IEEE, 2019.

[444] Yang Zhang, Liqun Deng 및 Yasheng Wang. 통합 만다린 tts 프런트 엔드 기반 증류된 버트 모델. arXiv 프리프린트 arXiv:2012.15404, 2020.

[445] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg 및 Bhuvana Ramabhadran. 외국어로 유창하게 말하기 학습 : 다국어 음성 합성 및 교차 언어 음성 복제. 절차 Interspeech 2019, 페이지 2080–2084, 2019.

[446] Zewang Zhang, Qiao Tian, Heng Lu, Ling-Hui Chen, Shan Liu. Adadurian: 두리안을 사용한 신경 텍스트 음성 변환을 위한 퓨샷 적응. arXiv 프리프린트 arXiv:2005.05642, 2020.

[447] Zhengchen Zhang, Fuxiang Wu, Chenyu Yang, Minghui Dong 및 Fugen Zhou. 구문 트리를 기반으로 한 운율 구 예측. SSW, 160-165페이지, 2016.

[448] Zi-Rong Zhang, Min Chu, Eric Chang. 중국어에서 자소에서 음소 로 변환하는 규칙을 배우는 효율적인 방법 입니다. 2002년 중국어 구어 처리에 관한 국제 심포지엄에서.

[449] Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, Bin Ma. 단일 언어 녹음과 교차 언어 음성 변환 의 혼합을 기반으로 하는 자연스러운 이중 언어 및 코드 전환 음성 합성을 지향합니다. 절차 Interspeech 2020, 페이지 2927–2931, 2020.

[450] Yi Zhao, Daisuke Saito, and Nobuaki Minematsu. Speaker representations for speaker adapta 다중 화자 blstm-rnn 기반 음성 합성에서의 사용. 공간, 5(6):7, 2016.

[451] Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. 중국어 단어 세분화 및 pos 태깅 을 위한 딥 러닝 . 2013년 자연어 처리의 경험적 방법에 관한 회의 절차 , 647–657페이지, 2013년.

[452] Yibin Zheng, Jianhua Tao, Zhengqi Wen 및 Jiangyan Yi. 엔드-투-엔드 tts를 정규화하기 위한 포워드-백워드 디코딩 시퀀스 오디오, 음성 및 언어 처리에 관한 IEEE/ACM 트랜잭션, 27(12):2067– 2079년, 2019년.

[453] Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das 및 Haizhou Li. 다국어 언어 모델을 사용한 종단 간 코드 전환 tts. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 페이지 7614–7618 . IEEE, 2020.

[454] Yixuan Zhou, Changhe Song, Jingbei Li, Zhiyong Wu, Helen Meng. 텍스트 음성 변환의 표현력 향상을 위한 그래프 신경망을 이용한 종속성 파싱 기반 의미 표현 학습. arXiv 프리프린트 arXiv:2104.06835, 2021.

[455] Jun-Yan Zhu, 박태성, Phillip Isola, Alexei A Efros. 순환 일치 적대적 네트워크를 사용 하는 페어링되지 않은 이미지 대 이미지 변환 . 컴퓨터 비전에 관한 IEEE 국제 회의 절차, 2223–2232페이지, 2017.

[456] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei Efros, Oliver Wang, Eli Shechtman. 다중 모드 이미지 대 이미지 번역을 향하여. 신경 정보 처리 시스템의 발전, 2017.

Barret Zoph 및 Quoc V Le. 강화 학습을 통한 신경 구조 검색. arXiv 프리프린트 arXiv:1611.01578, 2016.