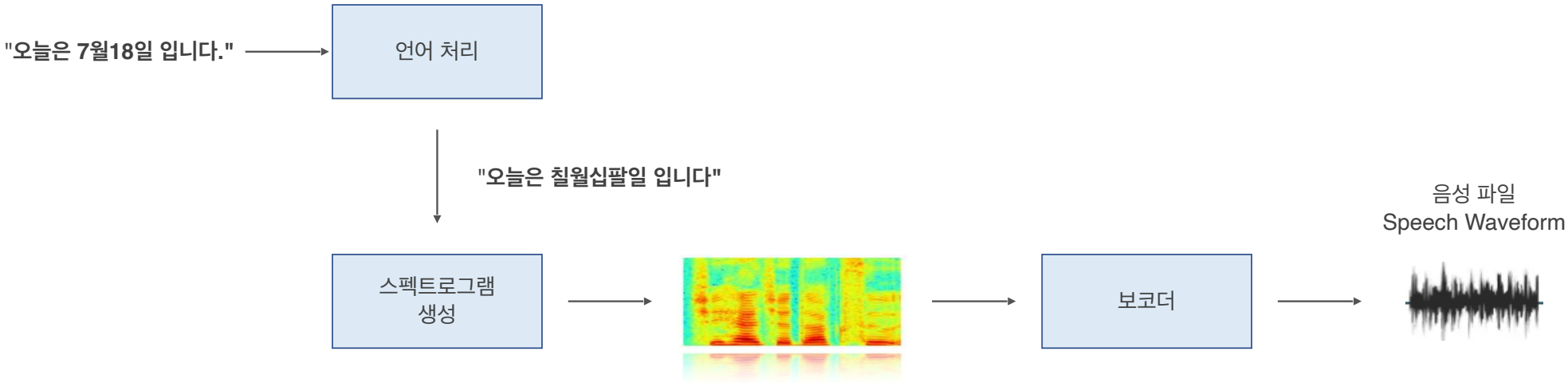
음성 합성

- **음성합성 구조도**

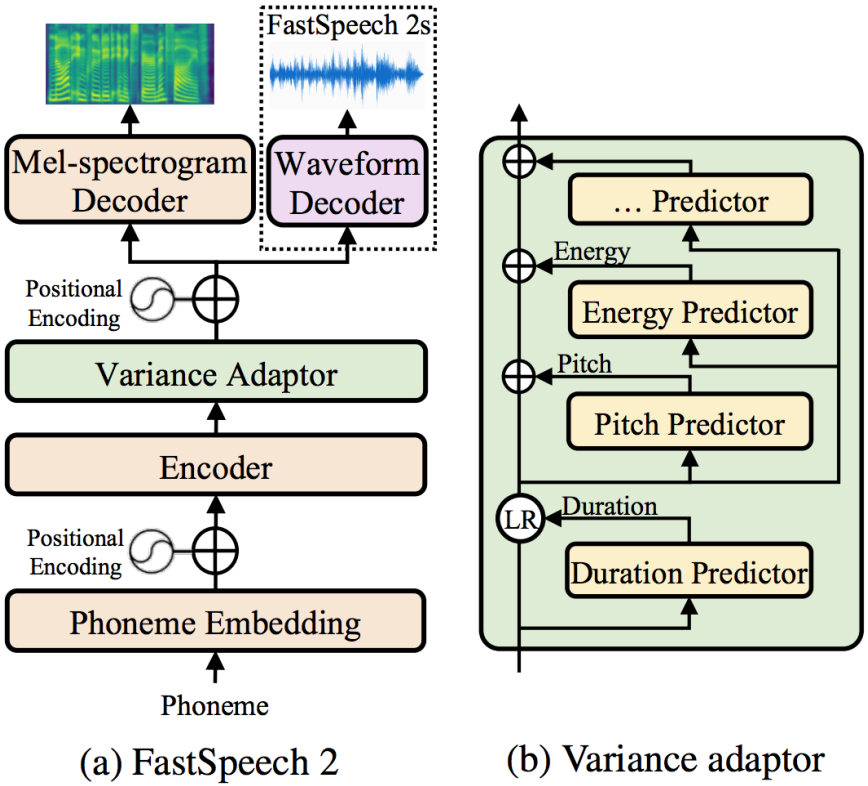


음성 합성

- **언어 처리**
  - 텍스트 정규화 (Text Normalization) 기법 적용
    - 모든 텍스트에 대한 발성 기준 한글화
    - 문장 부호, 숫자, 그 외 기회를 읽기 형태로 변환
    - 예) "119 불러줘" --> "일일구 불러줘"
  - 띄어 읽기에 대한 텍스트 처리 적용
    - 발화에서 띄어 읽기는 구절을 분리하는 기능
    - 쉼표(,)를 띄어 읽기 특수 키워드로 활용하여 적용
  - 발음 변환 (G2P, Grapheme to Phoneme) 수행
    - 읽기 형태로 변환된 문자열을 미리 저장된 음소로 또는 합성 단위열로 변환

음성 합성

- **FastSpeech2 (텍스트로부터 스펙트로그램 생성)**
  - Phoneme 입력 생성
    - 텍스트로부터 G2P를 활용하여 Phoneme 생성
    - G2P: Grapheme to Phoneme
  - Variance Adaptor
    - Variance 정보: duration, pitch, energy 등
  - Encoder 및 Mel-spectrogram Decoder
    - Transformer 구조 활용
    - MSE (Mean Squared Error) Loss 정의



(a) FastSpeech 2 (b) Variance adaptor

Yi et al, FastSpeech 2: Fast and High Quality End-to-End Text to Speech, Microsoft Research, 2020.06

음성 합성

- **Multi-band MelGAN 보코더 (스펙트로그램으로부터 소리로 변환)**
  - Neural 보코더를 통한 오디오 파일 생성
  - Mel-Spectrogram으로부터 4개 Multi-band 소리 생성
    - Filter bank를 적용하여 Multi-band 소리 생성
  - 4개 Multi-band 소리를 하나의 소리로 합성하여 real/fake 식별
    - Discriminator 에서 real/fake 식별 분석

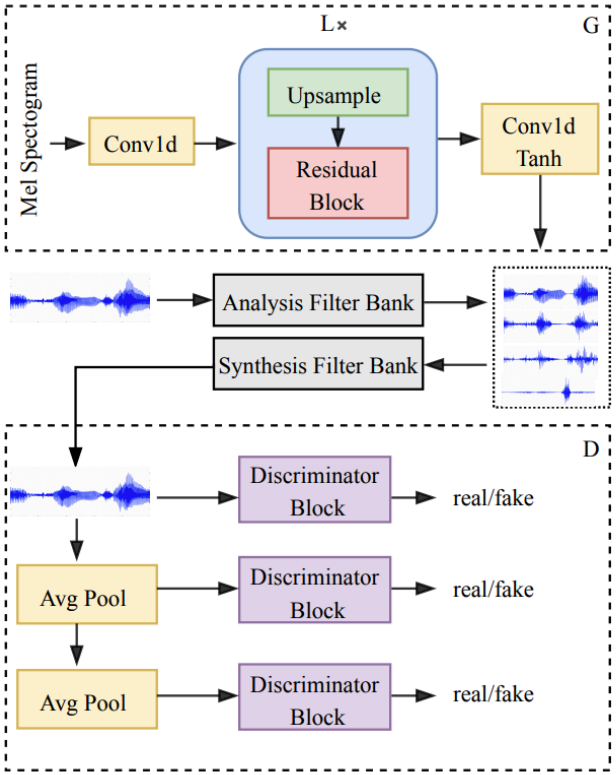


Figure 1: *Multi-band MelGAN Architecture.*

Geng et al, Multi-band MelGAN: Faster Waveform Generation for High Quality Text to Speech, 2020.05

음성 합성

- **성능 지표**
  - 음성합성 분야의 MOS (Mean Opinion Score) 성능 지표 적용
  - AI허브의 대화 발화 텍스트 100문장에 대해서 합성음을 생성하고,
  - 2명의 검증자가 합성음에 대해서 1-5점으로 평가 수행
  - 목표 : 국내 한국어 기준 3.8 수준 제시 **(@TODO 목표치 설정)**
    - 해외 영어 기준 4.2 (구글 자체 발표)) : (출처: https://cloud.google.com/text-to-speech/docs/wavenet?hl=ko)

- **데이터 확보**
  - 공개되어 있는 KSS 한글 TTS 데이터셋
    - 출처: https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset
    - 전문 여자 성우 1명 (12시간, 12853개 발성 문장)
  - 인공지능 학습용 데이터 구축사업(2차) 에서 세부과제15의 '다화자 음성 합성 데이터' 활용
    - 전문성우 6명 (성우별 40시간 제공)