AI 목소리 - 음성합성기술 소개

2019. 11. 22.

엔씨소프트 AI센터 - Speech Lab 음성합성팀

김 영 익 (youngik@ncsoft.com)

NCSOFT

목차

음성신호에 포함된 정보

무슨 말을 하였는가 (텍스트 정보)

어른인가, 아이인가 (나이/연령대 정보)

어느 나라 말인가 (언어 정보)

누가 말하였는가 (화자정보)

어떤 감정 상태인가 (감정/스트레스 정보)

어느 지역 출신인가 (출신지 정보)

지하철역, 자동차 실내 혹은 집 안인가 (음향 환경 정보)

남성인가, 여성인가 (성별 정보)

Speech AI 기술의 세부 분야

① 사용자가 누구인지 인식

화자식별/화자검증/언어/성별인식 기술

② 말의 내용과 감정을 인식

음성인식/감정인식 기술

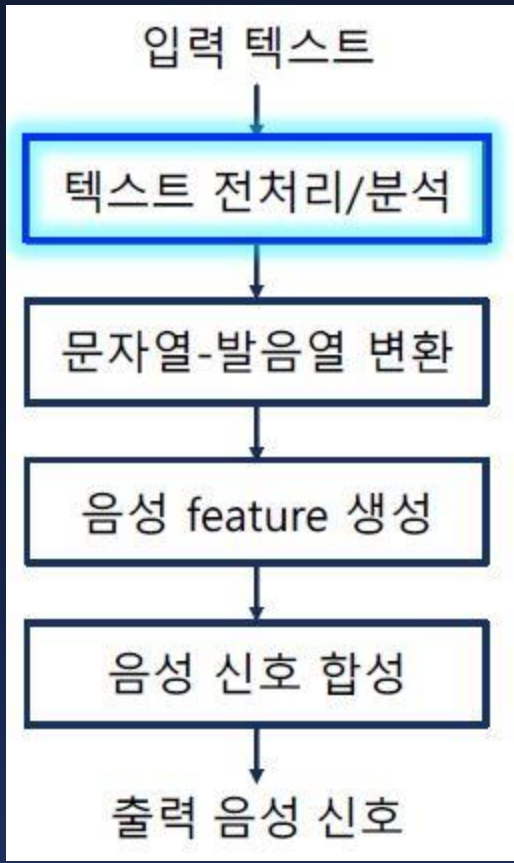
③ 자연스럽고 다양한 톤의 음성으로 응답

대화체/감정표현 TTS 기술

④ 사용자 주변 음향 환경을 이해

음향 신호 분류/잡음 처리/원거리 인식

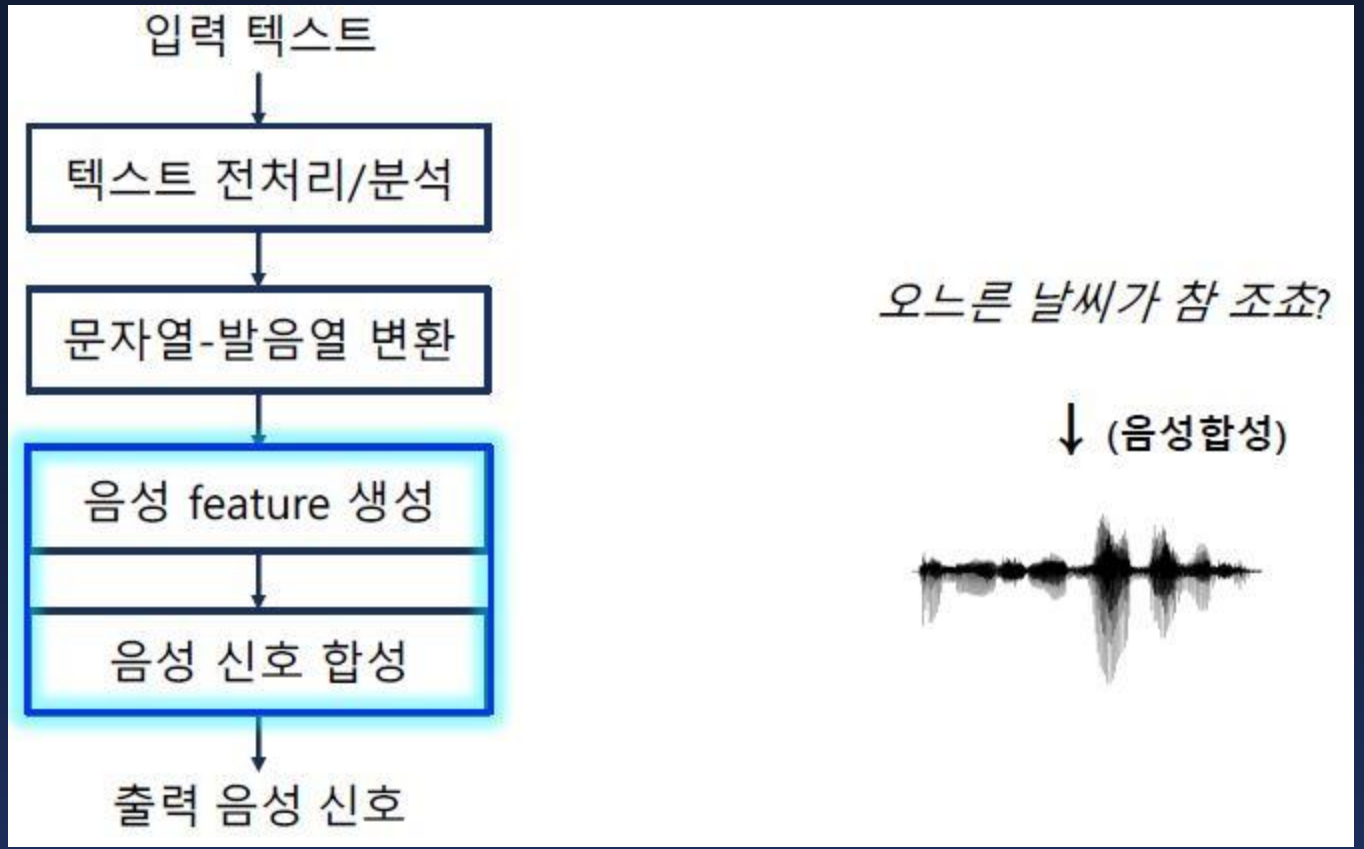
음성합성기술 이해 - Flow of TTS System



"환경부와 서울시·경기도는 "내일(26일) 오전 6시부터 오후 9시까지 미세먼지 비상저감조치를 시행한다"고 25일 밝혔다. 이날 일평균 초미세먼지 PM-2.5 농도는 서울 103㎍/㎥, 경기 110㎍/㎥ 등으로 '나쁨'(51~100㎍/㎥) 단계에 들었다."

Flow of TTS System

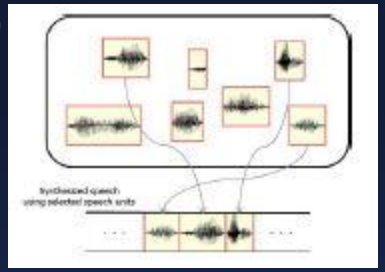
Flow of TTS System



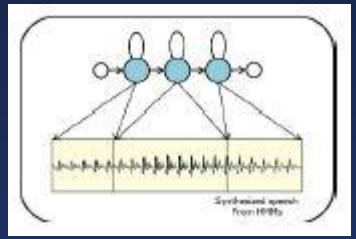
Brief History of Speech Synthesis

1) 규칙 기반의 포먼트 합성(Rule-based formant synthesis)
   - 전문가가 음성 생성 규칙을 공식화
   - 성대 (vocal tract)의 움직임을 디지털 필터로 모델링

2) 데이터베이스 기반의 연결 합성(Corpus-based concatenative synthesis)
   - 음성 조각 검색 & 연결
   - 대용량 음성 DB, 고품질(고음질, 자연스러운 운율)
   - 연결 부분 및 DB에 없는 단어의 부자연스러움



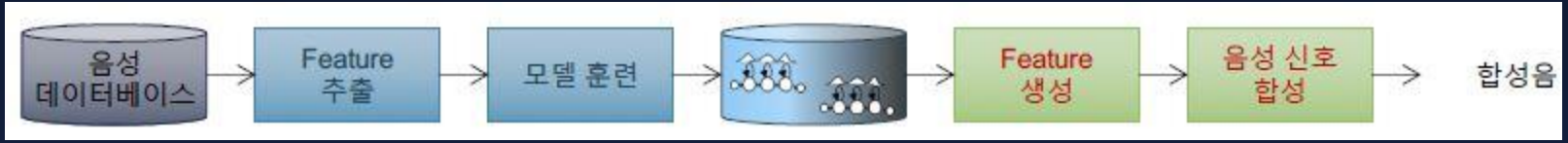
3) 데이터베이스 기반의 통계적 파라메트릭 합성(Corpus-based statistical parametric synthesis)
   - 음성 신호의 파라미터화
   - 통계적 모델링
   - 음성 특징 조절 및 변환 용이



Statistical Parametric Speech Synthesis (SPSS)

통계적 파라메트릭 음성 합성

- 통계적 : linguistic feature 에서 acoustic feature 를 통계적 음성 모델 기반으로 예측
- 파라메트릭 : 음성 신호에서 acoustic feature 추출 및 음성 신호 복원(합성)



- 훈련 단계 : Linquistic/acoustic feature 추출, 모델 훈련
- 합성 단계 : Acoustic feature 생성, 음성 신호 합성

Major Issues in Speech Synthesis

1) Vocoding, Vocoder Design, Acoustic Feature
   - 고품질(녹음 음성과 동일)의 음성 신호 합성을 위한 Vocoder 필요
   - 통계적 모델링에 적합한 Vocoder

2) Acoustic Modeling
   - Acoustic feature 와 linguistic feature 간의 관계를 학습할 수 있는 모델
   - 예) HMM, DNN

3) Post-processing
   - 모델링의 부족한 부분(예: 훈련 모델의 부정확성)을 보완하는 기술
   - 합성음의 음질을 향상시킬 수 있는 기술

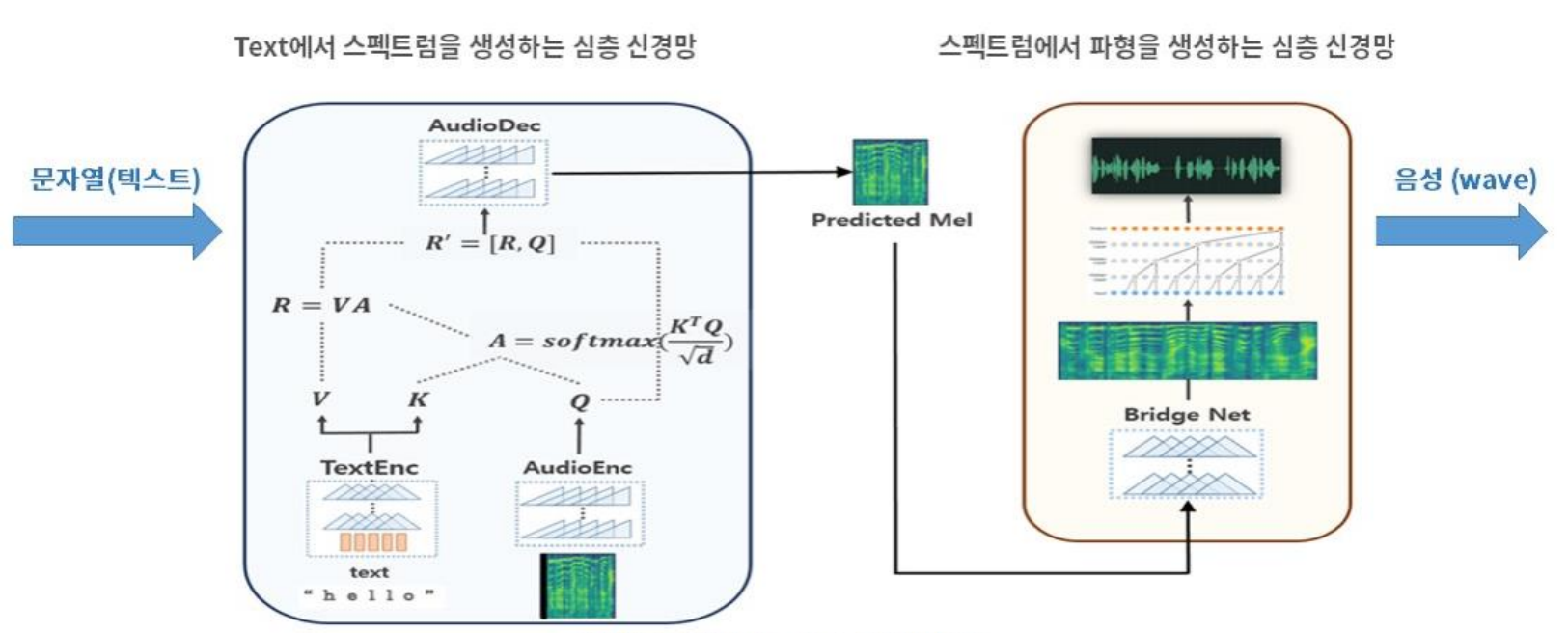
최근의 음성합성기술 연구 : End-to-End TTS 모델

- 입력과 출력 사이의 함수관계를 모델 스스로 학습
- 세부 단계에 필요한 기존 지식이나 리소스가 필요 없음

End-to-End 음성합성 기술의 개념

NCSOFT의 End-to-End 음성합성기 구조

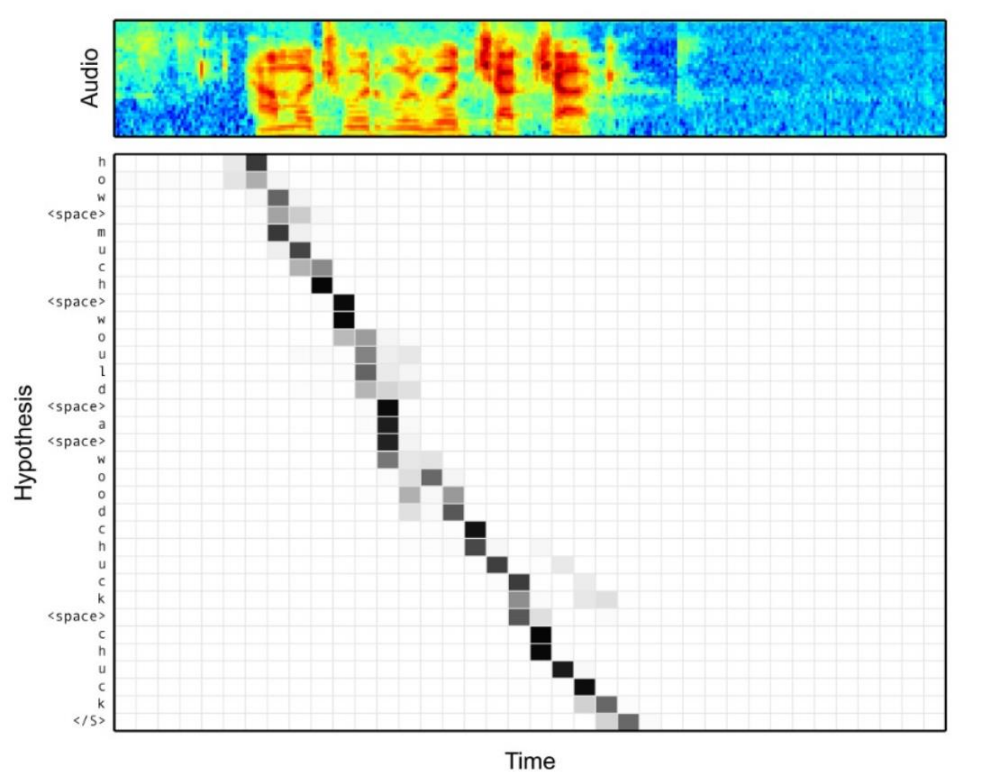
- Text에서 스펙트럼을 생성하는 모델 = Sequence to Sequence 네트워크 + Attention 메커니즘
- 스펙트럼에서 음성 파형을 생성하는 모델 = WaveGlow 기반 뉴럴보코더



엔씨소프트의 End-to-End 음성합성 방식

음성합성기의 Attention 메커니즘

- 텍스트와 음성 사이의 길이 정보를 예측



최근의 연구 주제

Neural vocoder 와 End-to-End TTS 로 인해 합성음의 자연스러움이 크게 향상

- 고사양 컴퓨팅 능력과 대용량 음성 DB 등이 뒷받침 되지 않는다면 재현/사용이 어려움
- 해결해야 할 다양한 이슈들 존재

안정성 향상을 위한 연구가 활발히 진행

- Neural vocoder : 생성 속도 단축
- End-to-End TTS : 대용량 음성 DB / 모델 훈련 / 안정적인 attention 생성

음성합성의 응용 분야에 대한 연구

- 감정 음성합성기술
- Voice Conversion 기술 등

음성합성기술의 응용

음성합성의 대표적 응용 분야

Text-To-Speech (TTS)

네비게이션 음성

오디오북

로봇

게임

* 기존에 사람이 직접 발성하여 정보를 전달하던 모든 곳에 적용 가능

음성합성 중점 연구 주제

다화자 End-to-End 합성 엔진

다수의 목소리로 합성하기 위한 프레임워크 개발

- 목표
  - 하나의 합성 모델에서 다수의 목소리를 합성
- 기대효과
  - 극소량의 데이터로 다양한 음색/운율 표현 가능

감정 스타일 표현 기술

감정 스타일 표현 기술 개발

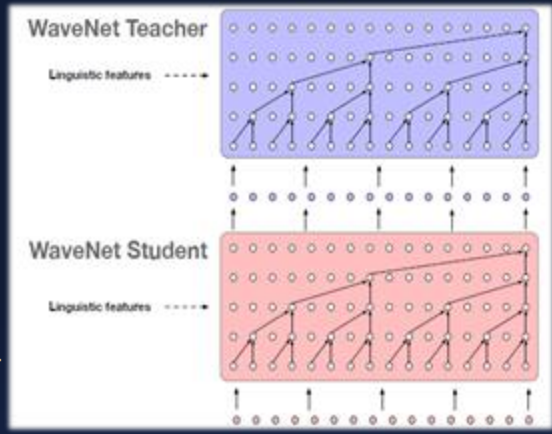
- 목표
  - 기쁨, 슬픔, 화남, 공포 등 주요 감정 표현
  - 다양한 발성 스타일 합성기술 연구
- 기대효과
  - 자연스러운 대화체 음성합성이 가능

음성합성 중점 연구 주제

뉴럴 보코더 개발

고품질 음성 생성을 위한 뉴럴 보코더 기술 개발

- 목표
  - 1초 분량의 합성음 생성 시간을 1초 이내로 단축
- 기대효과
  - 기존 보코더 대비 고품질 합성음 확보
  - Parallel WaveNet 기술 기반의 실시간 합성 서비스



다화자/다채널 합성 DB

다화자 합성을 위한 다양한 채널의 음성 DB 수집 및 구축

- 개발 내용
  - 사운드센터, 인터넷, DB 전문기관 등을 통해 음성 DB 수집
  - 전문 성우 섭외, 녹음
  - TTS 적용을 위한 음성/텍스트 전처리 도구 개발

Q&A