음성 합성 기술의 현황 및 과제

오영환

한국과학기술원

전산학과

차 례

1. 문서-음성 변환 시스템

2. 언어처리부

3. 운율생성부

4. 음성 합성부

   4.1 음성 DB 및 단위음의 선택

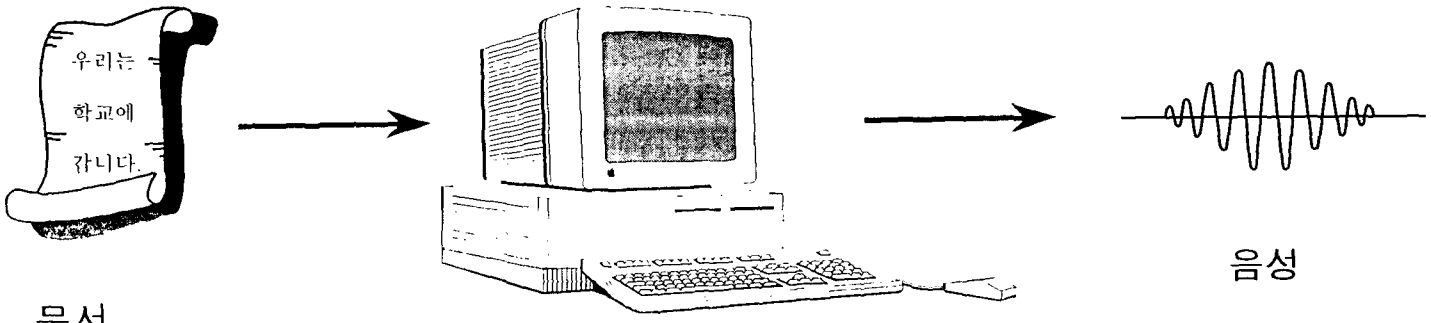
   4.2 음성 신호 생성

5. 연구 사례

6. 앞으로의 과제

문서-음성 변환 시스템 (Text-to-Speech System)

- 임의의 문장을 입력 받아 해당하는 음성신호로 변환하는 장치

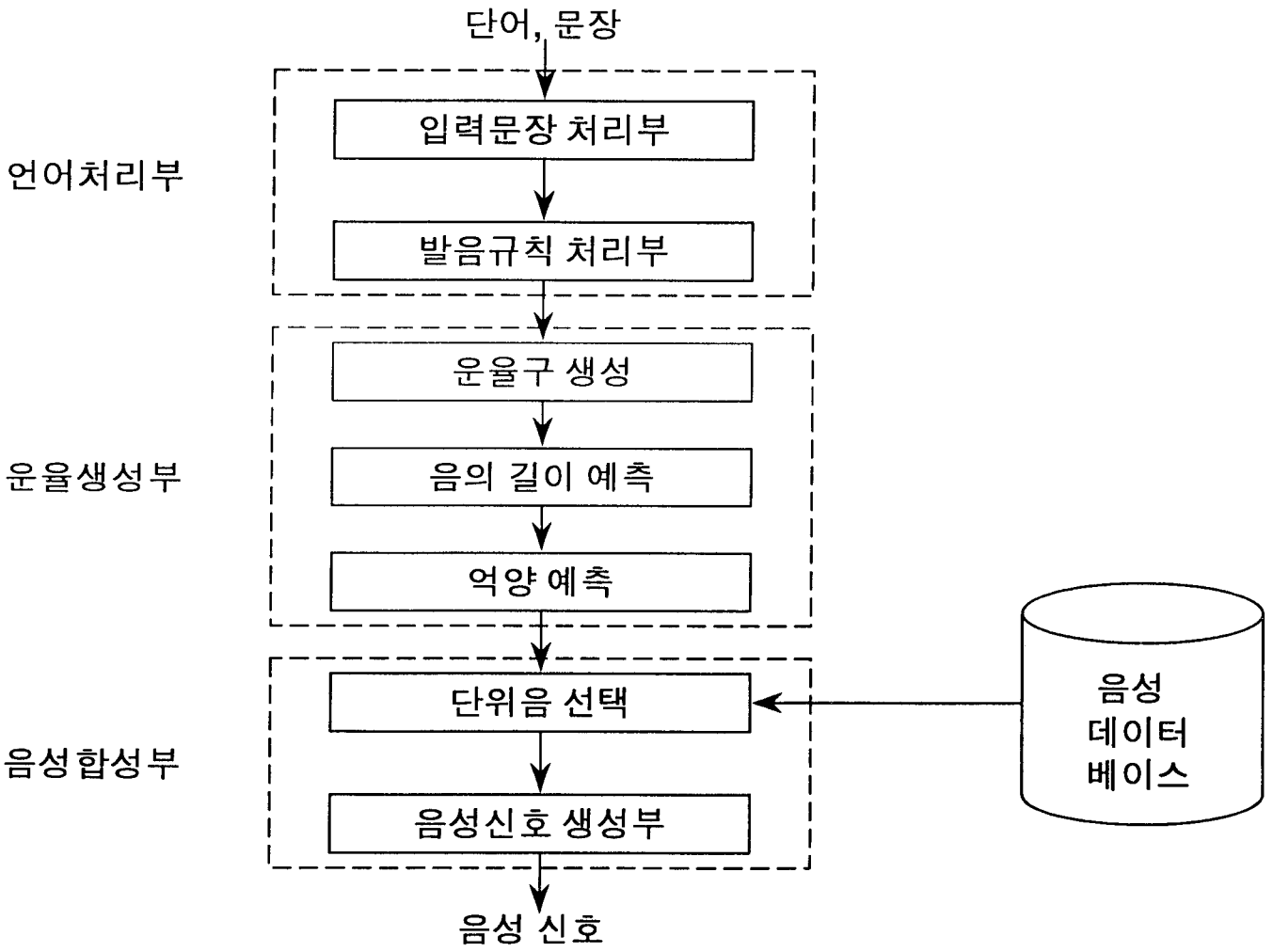


- 응용 분야
  - 맹인용 독서기
  - 전자 메일의 음성화 (전화기를 통한 전자우편 check)
  - Web browser의 음성 서비스 지원
  - Man-Machine Interface의 음성화
  - 음성 안내 시스템 (내용의 변경이 빈번한 경우)

문서-음성 변환 시스템 (Text-to-Speech System)

- 필요 기술
  - 단위음성 데이터베이스의 구축 (음성 DB)
  - 문서로부터 언어 정보 추출 (언어 처리부)
  - 언어 정보에 운율 정보 부여 (운율 생성부)
  - 단위 음성에 운율을 실어 합성하는 기술 (음성 합성부)

문서-음성 변환 시스템의 구성



언어 처리부

- 전처리
  - 숫자, 기호, 영어 등 비한글의 한글화
- 형태소 분석
- 품사 태깅
- 구문 분석
- 발음 표기 변환
  - 규칙 기반
  - 예외 발음 사전

운율 생성부

- 운율 (prosody)
  - 띄어읽기 (운율구)
  - 음소별 지속시간 (duration)
  - 억양 (pitch contour)
  - 강세 (에너지)

- 방법론
  - 규칙 기반
  - 통계적 방법
    - 신경회로망, HMM, CART 등

음성 데이터베이스 및 단위음의 선택

- 단위음 연결 방식 **(unit concatenation approach)**
  - 인위적으로 발성된 단위음
  - 운율 제어 과정을 거쳐 연결
  - 명료도 우수, 자연성 부족

- 대용량 음성데이터 기반 방식 **(corpus-based approach)**
  - 문장으로부터 얻어진 불균일 단위음
  - **unit selection** 과정을 통해 최적의 **segment** 선택
  - 제한된 **domain** 내 합성음의 자연성 매우 우수

단위음 연결 방식

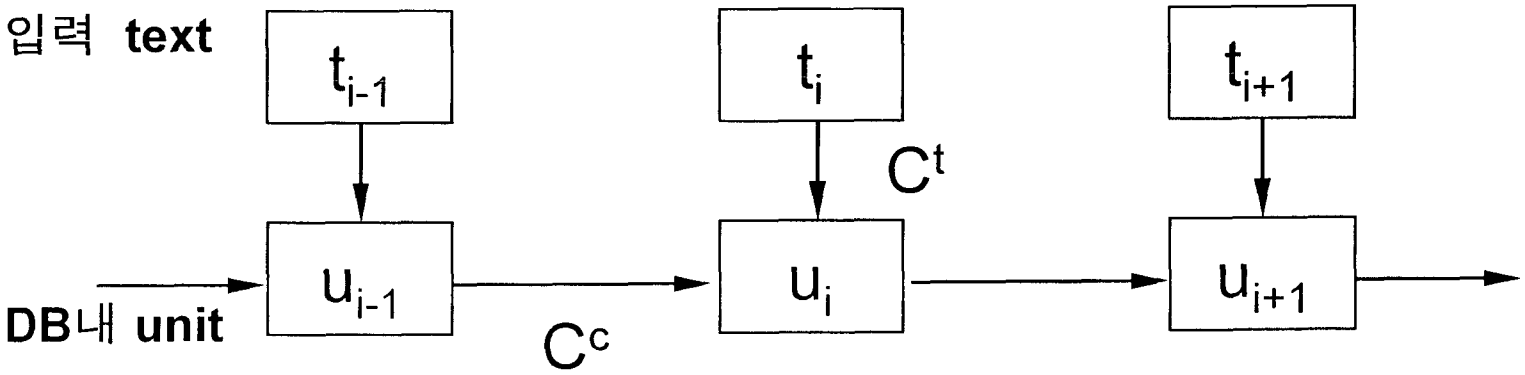
- 단위음 데이터베이스
  - 음소, 이음소(**diphone**), **VCV** 연쇄, 음절 등
  - **COC (context oriented clustering)**에 의한 복수개의 후보
- 단위음 합성
  - 예측된 운율값을 갖도록 신호처리
  - **PSOLA, sinusoidal** 기반 합성기 이용
- 문제점
  - 제한된 수의 단위음
  - 운율제어 과정에서의 왜곡현상

⇒ 합성음의 자연성 부족

대용량 음성데이터 기반 방식

- 연구배경
  - **lower memory costs**
  - **faster processors**
  - 많은 음성자료의 처리가 용이해짐.
  - **Naturalness cannot be created, it can only be preserved (from CHATR)**
- 음성 데이터의 구성
  - 음성 데이터와 **transcription**의 **aligning**
  - **unit selection**을 위한 특징값 생성
    - **phone label**
    - **f0, duration, power**
    - **stress, accent type** 등

Unit Selection



- target cost (C^t) : unit의 phonetic context 고려
- concatenation cost (C^c) : 인접 unit 간의 차이 고려
- choose best matched segments by viterbi algorithm

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i), \qquad \bar{u}_1^n = \min_{u_1,\ldots,u_n} C(t_1^n, u_1^n)$$

⇒ larger segment를 선호

대용량 음성데이터 기반 방식의 문제

- 문장 선정
  - **phone balanced, VCV chain balanced** 등
  - 입력될 문장에 대한 **coverage**가 높아야 한다.
- **phone segmentation**의 자동화
  - 많은 자료를 대상으로 함으로 자동화 되어야 함
  - **HMM** 등의 음성 인식 기법을 이용
- **unit selection**
  - **cost function**의 정의
  - **weighting value**의 학습
  - **fast algorithm**

⇒ 현재 제한된 **domain**내의 시스템으로 개발되고 있다.

음성 신호 생성

- **PSOLA (Pitch Synchronous OverLap and Add)**
  - **pitch, duration** 조절 용이
  - **pitch marking** 필요
  - 접합 부분에서의 왜곡
- 정현파 모델 기반의 분석 합성
  - 단위음 사이의 보간 용이 : 왜곡을 줄일 수 있다.
  - 무성음 생성에 취약
- **concatenation**
  - 대용량 음성데이터 기반의 방법
  - 신호 처리 배제

연구 사례 (1)

- **CHATR (일본 , ATR)**
  - **45분 분량의 음성데이터**
  - 합성 단위 : **phone**
  - 선택된 단위음의 재배열(**resequencing**) 통한 합성음 생성
  - 한국어 버전 구현

- **Next-Generation TTS (미국, AT&T)**
  - **80분 분량의 음성데이터**
  - 합성 단위 : **half-phone**
  - 단위음 합성 : **sinusoidal model 기반, no processing**

연구 사례 (2)

- **ETRI (한국)**
  - **2092** 문장
  - 합성 단위 : **triphone**
  - **no signal processing**

- **Whistler (Microsoft)**
  - **trainable TTS**
  - **HMM** 기반의 단위음 **model**과 운율 **template model**의 분리
- **Lucent Technology**
  - **multilingual TTS**
  - 현재 **9**개 언어에 대한 시스템 완료

앞으로의 과제

- 합성음의 음질평가
  - 명료성 **(intelligibility)**, 자연성 **(naturalness)**
  - 일반적으로 **MOS (mean opinion score)** 사용
  - 보다 객관적인 척도가 필요
  - 표준 문장 집합의 선정

- **Text-to-Speech**의 새로운 경향
  - 대용량 음성 데이터로부터 생성된 불균일 단위 사용
  - 시스템 구성의 자동화
  - **language independent methods**
  - **audio-visual approach**