# ETRI 소용량 대화체 음성합성시스템

*김종진, *김정세, *김상훈, *박준, *이윤근, **한민수
*한국전자통신연구원 음성언어정보연구센터
**한국정보통신대학교

# ETRI small-sized dialog style TTS system

*Jong-Jin Kim, *Jeong Se Kim, *Sanghun Kim, *Jun Park, *Yunkeun lee, **Minsoo Hahn
*Speech/Language information processing Center, ETRI
**Information and Communication University
E-mail : {kimjj, jungskim, ksh, junpark, yklee}@etri.re.kr, mshahn@icu.ac.kr

## Abstract

This study outlines a small-sized dialog style ETRI Korean TTS system which applies a HMM based speech synthesis techniques. In order to build the VoiceFont, dialog-style 500 sentences were used in training HMM. And the context information about phonemes, syllables, words, phrases and sentence were extracted fully automatically to build context-dependent HMM. In training the acoustic model, acoustic features such as Mel-cepstrums, logF0 and its delta, delta-delta were used. The size of the VoiceFont which was built through the training is 0.93Mb. The developed HMM-based TTS system were installed on the ARM720T processor which operates 60MHz clocks/second. To reduce computation time, the MLSA inverse filtering module is implemented with Assembly language. The speed of the fully implemented system is the 1.73 times faster than real time.

## I. 서론

최근 텔레매틱스, 모바일 환경, 지능형 로봇 등 임베디드 환경에서 사용할 수 있는 대화체 음성합성 기술의 요구가 증대되고 있다. 이에 본 연구팀에서는 임베디드 응용 분야에서 사용 가능한 고품질의 소용량 대화체 음성합성 알고리듬 및 시스템을 개발하고자 한다.

이를 위해 본 연구에서는 음편접합 방식에 비해, 상대적으로 적은 량의 데이터를 사용하고 정교한 수동 레이블링이 반드시 필요하지는 않으며, 음색변환 및 스타일변환 등이 비교적 용이한 HMM모델로부터 직접 합성음을 생성하는 HMM기반 음성합성 방법을 이용한 고품질의 소용량 대화체 음성합성 알고리듬 및 시스템을 개발하고자 한다.

HMM기반 음성합성 기술은 음성신호에 대한 스펙트럼 정보, 피치 정보, 지속시간 정보에 대해 음성인식 시스템 개발에서 음향모델 HMM을 훈련시키는 방식과 유사한 방식으로 보이스폰트를 생성하며[1,2,3,4], 훈련된 HMM 모델 파라미터로부터 합성음 생성을 위한 음성 특징 파라미터를 생성한다[5,6,7]. HMM기반 음성합성 방식은 파형접합방식을 이용한 소용량 합성시스템에서 나타나는 접점에서의 불연속 등이 없다는 장점이 있으나, 과도하게 스무딩된 음성 특징 파라미터 궤적과 VOCODING 방식의 합성음 생성 과정에서 나타나는 둔탁한 음질 등 전반적으로 파형접합방식에 비해 명료성이 떨어진다는 것이 단점으로 지적되고 있다[8,9]. 그러나 최근 이러한 문제점을 해결하기 위해, 정교한 음향모델링[10], Global variance를 이용한 다이나믹한 음성파라미터 궤적의 생성[11], Mixed excitation을 이용한 VOCODER의 명료성 개선[12,13] 등을 통해 상당부분 음질개선이 이루어지고 있다.

본 연구에서는 [9]에서 개발된 HMM기반 한국어 음성합성 베이스라인 시스템을 기반으로, 대화체 훈련DB

및 대화체 문맥정보를 이용한 대화체 보이스폰트를 개발하였으며, 합성엔진은 ARM720T 60MHz 급 프로세서가 장착된 ITS(Intelligent Traffic System)단말용 OBE(On board Equipment)보드에 탑재시켜 임베디드 환경에서 사용가능한 HMM기반 소용량 대화체 음성합성 시스템 베이스라인을 구축하였다.

본 논문의 구성은 다음과 같다. 2장에서는 HMM기반 음성합성 방식에 대하여 개략적으로 기술하고, 3장에서는 대화체 보이스폰트 개발 과정에 대하여 기술한다. 4장에서는 ITS단말기에 포팅된 HMM기반 합성엔진에 대해서 기술하고, 5장에서는 결론과 향후 연구계획을 기술한다.

## II. HMM기반 음성합성 방식

HMM기반 음성합성 시스템은 오프라인에서 HMM모델들과 문맥결정트리로 구성된 보이스폰트를 생성하는 훈련부와 생성된 보이스폰트를 이용한 합성음 생성을 수행하는 음성신호생성부로 구성된다.

훈련부에서는 먼저 합성 DB로부터 문맥종속 HMM 훈련을 위한 다양한 언어정보를 생성하고, 음성신호로부터 스펙트럼정보, 피치정보, 지속시간 정보를 추출하여 훈련을 준비한다. 훈련과정에서는 일반적으로 left-to-right 상태천이 구조를 가지는 HMM모델을 구성하여 음성인식시스템의 훈련과정과 유사한 훈련과정을 거쳐서 HMM 보이스폰트를 생성한다.

합성부에서는 입력문장에 대한 다양한 언어정보의 생성 및 예측을 수행하고, 이를 이용하여 적합한 HMM을 검색하여 연결시킨 후 음성 파라미터 생성 알고리듬을 통하여 부드러운 피치 및 스펙트럼 궤적을 생성한다. 생성된 음성 특징파라미터는 프레임단위로 피치정보와 스펙트럼 정보를 적합한 음성신호 생성 필터에 통과시켜 합성음을 생성한다.

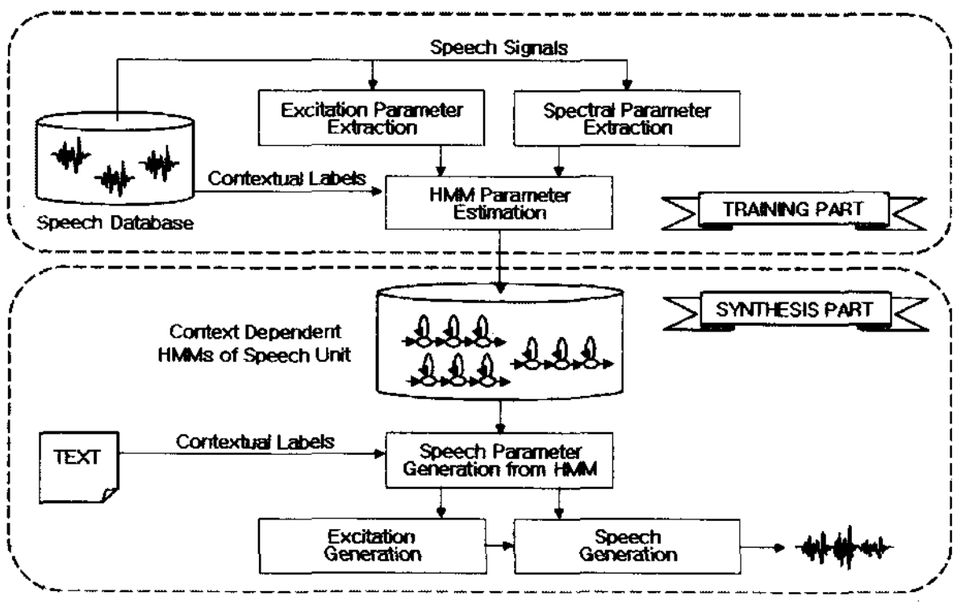


그림 1. HMM기반 음성합성시스템의 구성

## III. 대화체 보이스폰트 개발

### 3.1 데이터 전처리 단계

본 연구에서 음향모델 훈련을 위해 [14]에서 개발된 대화체 합성DB 여성 1인의 500문장을 이용하였으며, 문맥종속 HMM훈련을 위한 문맥정보는 다음과 같은 과정을 거쳐 생성하였다.

1) 훈련DB로 사용된 500 문장의 발성목록에 대해 한국어 음운규칙과 예외발음사전으로 구성된 합성엔진에서 사용되는 철자-발음변환 모듈을 이용하여 발성목록에 대한 발음열을 생성하였다.

2) 훈련DB의 자동 레이블링을 수행하였다. 자동 레이블링을 HTK 툴킷을 이용하였으며, 음향모델의 특징 파라미터는 MFCC를 이용하였고, 전문성우 4인의 수동 레이블링된 16,000문장(남녀 각 2인, 각 화자별 약 4000문장)으로부터 훈련된 모노폰 HMM모델을 이용하여 자동 레이블링을 수행하였다.

3) 어절단위 자동 끊어읽기 경계강도 태깅을 수행하였다. 끊어읽기 경계강도 태깅은 어절단위로 수행하였으며, 사용된 특징 파라미터는 자동 레이블링 정보에서 추출한 음소 및 음절의 지속시간 정보, 휴지구간 정보 등을 이용한 규칙기반 시스템을 이용하였다. 끊어읽기 레벨은 끊어읽지 않은 어절의 경계, 마지막 음절이 장음화된 어절의 경계, 짧은 무음구간을 가진 어절의 경계, 긴 무음구간을 가진 어절 경계를 규칙기반으로 자동 태깅하였다. 그 외 음절 경계, 형태소 경계, 어절경계, 단어경계 정보 등은 특별한 판별규칙 없이 추가적으로 자동 태깅되었다.

4) 대화체 문장의 문말 운율을 보다 자연스럽게 모델링하기 위해 문장 끝 어미 유형을 이용한 규칙기반 문장 의미 태깅을 수행하였다. 문장 끝 종결어미의 의미 태깅은 문장의 종결어미를 기 정의된 의미 부류(semantic class)로 태깅하는 것으로, 종결어미의 의미 태깅 결과는 문장의 양태를 결정하고 나아가 문장의 운율, 특히 문장 끝 운율 패턴 모델링에 이용되도록 하였다. 종결어미의 의미태깅은 형태소 태깅된 결과에서 분리된 종결어미 형태에 대해 기정의/분류된 44개의 의미태그셋과 이에 해당하는 종결어미 형태(556개)로 구성된 사전에 의해 수행된다.

표-1. 문장의 의미 부류 예

| 의미부류 | 문말 어미(예) |
|---|---|
| proposal(권유) | ㅂ시다. 시다. 세, ... |
| request(요청) | 십시오. ㅂ시오. 쇼, ... |
| allowance(허락) | 렴. 려무나. |
| commit(약속) | ㄹ게, ㄹ게요. |

본 연구에서 사용한 어미패턴 기반 의미 부류 판별 정보를 이용한 문장 끝 운율 모델링은 톤 타입에 대한 정교한 수동 레이블링을 사용하지 않으므로 보이스폰트 개발비용을 감축할 수 있을 것으로 예상한다.

이상에서 기술한 데이터 전처리 과정에서 발생한 오류에 대해서 수작업 오류 정정없이 사용되었다.

## 3.2 훈련단계

훈련단계에서는 전처리 단계에서 구축된 문맥정보를 이용하여 문맥종속 HMM 모델을 훈련시키는 과정으로, 음성인식시스템의 훈련과정과 유사하게 진행된다[15].

훈련에 사용된 HMM의 구조는 left-to-right 5 state HMM을 사용하였다. 음향모델 훈련을 위한 특징 파라미터는 프레임 크기 20ms, 프레임 중첩 5ms, Blackman windowing을 통한 프레임단위로 25차 Mel-cepstrum을 추출하고, 이들의 delta, delta-delta 총 75차를 구하였으며, 여기신호모델 훈련을 위해 Praat를 이용해 F0를 구하여 log를 취한 후 이들의 delta, delta-delta 총 3차를 구하여 총 78차의 다중 스트림 특징파라미터 벡터를 구성하였다. 스펙트럼 정보는 diagonal covariance matrice를 가지는 Gaussian 분포로 모델링 하였으며, F0 정보는 유성음에 대해서는 diagonal covariance matrice 를 가지는 Gaussian 분포로 모델링하고, 무성음 구간은 이산분포로 모델링 하는 MSD(Multi-space distribution)[3,9] 으로 모델링 하였다.

# IV. 합성엔진의 ITS단말 포팅

현재까지 개발된 HMM기반 합성엔진은 저사양 프로세스와 스피커를 가진 임베디드 환경에서 문제점 및 성능개선 부분 등을 고찰하기 위하여 60MHz clocks/second 처리속도의 ARM720T 프로세서, 16MB NAND메모리, 32MB SDRAM으로 구성된 ITS단말용 OBE보드에 포팅하였다.

## 4.1 메모리 사용량

현재 개발된 단말형 대화체 합성시스템을 동작시키기 위한 메모리 요구량은 언어처리사전 및 보이스폰트는 저장하는 SDRAM 메모리와 실행시간에 사용하는 NAND 메모리로 구성된다. 아래 표-2 는 ITS단말기에 탑재한 소용량 합성시스템의 메모리 사용량을 보여준다.

표-2. 메모리 사용량. (단위: MB)

| | SDRAM | NAND |
|---|---|---|
| Dictionaries | 1.70 | 2.1 |
| VoiceFont | 0.93 | 1.3 |
| TTS Engine | 0.47 | 2.2 |
| Total | 3.19 | 5.6 |

표-2의 합성엔진의 NAND메모리 요구량은 1채널 기준 실행시간 메모리 요구량이다..

## 4.2 속도 개선

HMM기반 합성부에서 MLSA(Mel Log Spectrum Approximation) 필터를 통한 음성신호 생성 부분이 가장 많은 연산량을 사용하는 것으로 나타났다. 속도 개선을 위해 “안녕하세요?“ 문장을 이용하여 테스트 벡터를 구성하였다. 구성된 테스트 벡터는 최종 출력 16kHz, 16bits, mono형식의 오디오 출력 형식을 기준으로 총 350 프레임이며, 프레임당 80 샘플을 생성하여 총 28000개의 음성샘플(지속시간: 1.75초)을 생성하는 것으로 하였다.

속도 개선을 위해 베이스라인 실수연산 MLSA 필터링 모듈을 정수형 연산 버전과 어셈블러어로 구현된 버전의 수행속도를 비교하였다. 아래 표-3은 각 구현 방법별 연산 소요시간을 비교한 것이다.

표-3. MLSA 필터 구현 방법별 수행속도

| 구현방법 | 실행시간(초) | 실시간 |
|---|---|---|
| 실수형 연산 버전 | 12.0 | 6.85 |
| 정수형 연산 버전 | 2.40 | 1.38 |
| 어셈블리어 버전 | 0.92 | 0.52 |

# V. 결론

본 연구에서는 소용량 음성합성 기술 개발의 일환으로 ETRI 음성언어정보연구센터에서 개발한 ITS단말용 HMM기반 소용량 대화체 음성합성 시스템에 대하여

기술하였다. 현재 개발된 버전은 대화체 합성 DB 500 문장을 이용하여 개발되었으며, 데이터 전처리 과정은 수작업없이 이루어졌으며 최종 개발된 보이스폰트의 크기는 0.83Mb 이다. HMM기반 합성엔진은 ARM7급 프로세서를 이용하는 ITS단말용 보드에 탑재되었으며 총 3.19Mb의 SDRAM과 실행시간에는 5.6Mb의 NAND 메모리를 사용한다.

향후 연구에서는 보코딩 스타일의 음질의 개선을 위하여 혼합여기모델을 이용한 명료성 개선[12,13], Global variance를 이용한 보다 자연스러운 스펙트럼 및 F0 특징 파라미터 생성[11], MGE(Minimum Generation Error)기반 정교한 음향모델 훈련[10] 방법을 도입하여 합성음의 음질을 개선하고자 한다.

## 참고문헌


[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Proc. of Eurospeech, pp.2347-2350, Sept. 1999.

[2] K. Tokuda, T. Mausko, N. Miyazaki, T. Kobayashi, Multi-space probability distribution HMM, IEICE Trans. Inf. & Syst., vol.E85-D, no.3, pp.455-464, March 2002.

[3] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In Proc. ICASSP-99, pages 229-232, March 1999.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, Duration modeling for HMM-based speech synthesis., In Proc. ICSLP-98, pages 29-32, December 1998.

[5] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation From HMM using dynamic features," Proc. ICASSP-95, pp.660--663, May 1995.

[6] Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, Satoshi Imai, "An Algorithm For Speech Parameter Generation From Continuous Mixture HMMs With Dynamic Features," Proc of EUROSPEECH, vol.1, pp.757-760, 1995

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp.1315-1318, June 2000.

[8] K. Tokuda, H. Zen, A.W. Black, An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sept. 2002.

[9] S.-J. Kim, J.-J. Kim, M.-S. Hahn, Implementation and evaluation of an HMM-based Korean speech synthesis system, IEICE Trans. Inf. & Syst., vol. E89-D, no.3, pp.1116-1119, 2006.

[10] Y.-J. Wu, R.-H. Wang, Minimum generation error training for HMM-based speech synthesis, Proc. of ICASSP, pp.89-92, 2006.

[11] T. Toda, K. Tokuda, ''Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis,'' Proc. INTERSPEECH2005-EUROSPEECH, pp. 2801-2804, Lisbon, Portugal, Sep. 2005.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Mixed excitation for HMM-based speech Synthesis, Proc. of Eurospeech, pp.2259-2262, Sept. 2001.

[13] S.-J. Kim, M.-S. Hahn, Two-band excitation for HMM-based speech synthesis, IEICE Trans. Inf. & Syst., vol.E90-D, no.1, pp.378-381, Jan. 2007.

[14] 김종진, 오승신, 최문옥, 김상훈, 박준, 이영직, ETRI 대화체 음성합성시스템 소개, 2003 대한음성학회 봄 학술대회

[15] Keiichi Tokuda, Heiga Zen, Junichi Yamagishi, Takashi Masuko, Shinji Sako, Alan W. Black, Takashi Nose, http://hts.sp.nitech.ac.jp/