Journal of The Korea Society of Computer and Information
Vol. 17 No. 9, September 2012
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci/2012.17.9.075

운율경계정보를 이용한 HMM기반 한국어 TTS 자연성 향상 연구

임 기 정*, 이 정 철*

Improvement of Naturalness for a HMM-based Korean TTS using the prosodic boundary information

Gi-Jeong Lim*, Jung-Chul Lee*

요 약

HMM 기반 음성합성시스템은 성능향상을 위해 일반적으로 대용량 음성 DB로부터 생성된 문맥의존 tri-phone을 이용한다. 그리고 대용량 DB의 경량화를 위해서 문맥의존정보를 이용하여 결정트리 방식으로 발화특성이 유사한 문맥의존음소들을 군집화한다. 군집화에 사용하는 문맥의존정보는 음소열 뿐만 아니라 운율정보도 포함하는데 이는 합성음의 자연성이 끊어 읽기, 억양패턴, 음의 장단과 같은 운율에 의해 크게 좌우되기 때문이다. 그러나 복잡한 운율정보를 사용할 경우 훈련과정에 포함되지 않은 문맥의존음소는 하나의 대표값으로 평활화되며 이로 인해 합성음의 자연성이 크게 저하된다.

본 논문에서는 합성음의 자연성을 향상시키기 위해 복잡한 운율정보 대신 억양 변화를 상승, 평탄, 하강으로 구분함으로써 운율정보표현을 간소화시킨 운율경계정보를 포함하는 문맥의존정보에 대한 문맥질의, 그리고 해당 질의의 패턴을 정의하는 방법을 제안하였다.

본 논문에서 제안하는 세 가지 운율경계정보를 포함한 문맥의존정보를 이용하여 합성음을 생성하고 MOS평가를 수행한 결과 운율경계정보를 이용한 HMM기반 한국어 TTS 합성음의 자연성이 향상됨을 확인하였다.

▸ Keywords : HMM 기반 음성합성, 은닉 마코프 모델, 트라이폰, 결정트리 기반 군집화

Abstract

HMM-based Text-to-Speech systems generally utilize context dependent tri-phone units from a large corpus speech DB to enhance the synthetic speech. To downsize a large corpus speech DB, acoustically similar tri-phone units are clustered based on the decision tree using context dependent information. Context dependent information includes phoneme sequence as well as

• 제1저자 : 임기정 • 교신저자 : 이정철
• 투고일 : 2012. 07. 31, 심사일 : 2012. 08. 21, 게재확정일 : 2012. 08. 30.
* 울산대학교 전기공학부(School of Electrical Engineering, University of Ulsan)

prosodic information because the naturalness of synthetic speech highly depends on the prosody such as pause, intonation pattern, and segmental duration. However, if the prosodic information was complicated, many context dependent phonemes would have no examples in the training data, and clustering would provide a smoothed feature which will generate unnatural synthetic speech.

In this paper, instead of complicate prosodic information we propose a simple three prosodic boundary types and decision tree questions that use rising tone, falling tone, and monotonic tone to improve naturalness. Experimental results show that our proposed method can improve naturalness of a HMM-based Korean TTS and get high MOS in the perception test.

▸ Keywords : HTS, HMM, tri-phone, decision tree-based clustering

I. 서 론

최근 들어 모바일 기기를 중심으로 음성합성기술이 많이 사용되고 있다. 확률모델 기반의 음성합성은 코퍼스기반의 음성합성보다 합성에 필요한 DB용량을 경량화 할 수 있으며 동일한 합성 DB로 다양한 합성음을 생성할 수 있는 장점이 있다. 확률모델 기반 음성합성은 HMM (Hidden Makov Model)을 이용하여 구현할 수 있다[1-3].

캠브리지대학 Speech Vision and Robotics Group의 Steve Young은 HMM의 훈련과 인식에 필요한 알고리즘들을 라이브러리화한 HTK (Hidden markov model ToolKit)을 1989년 발표하였다. 이후 HTK는 많은 개선과정을 거쳐 현재까지 HMM기반의 음성인식관련 연구에 효과적으로 사용되고 있다[4]. 음성합성 분야에서도 K. Tokuda는 HMM을 이용한 음성합성시스템 (HMM-based speech synthesis system, HTS)을 제안하였다[5].

Source-filter 모델을 사용하는 HTS는 음성합성에 필요한 성도특징, 기본주파수, 지속시간 파라미터에 대한 HMM 모델과 훈련이 필요하다. 이를 위해서 기존 음성인식용 HTK를 수정하여 훈련에 사용할 수 있도록 패치 파일이 제공되고 있으며 이를 이용하여 음성합성과 관련된 HMM들을 문맥의존 방식으로 훈련한다[5-6]. 합성단계에서는 파라미터생성 알고리즘을 이용하여 훈련된 HMM 모델로부터 합성음 생성에 필요한 파라미터열을 생성하고 MLSA (Mel-Log Spectrum Approximation) 필터를 이용하여 합성음을 생성한다. 국내에서도 HTS를 이용하여 HMM기반 한국어 음성합성이 연구되었다[7-8].

일반적으로 HMM기반의 음성합성은 합성 성능향상을 위해 tri-phone형태의 문맥의존정보를 이용하여 훈련 및 합성을 하게 된다. 합성음의 자연성은 문장 내에서 끊어 읽기, 억양패턴, 음의 장단과 같은 운율에 의해 크게 좌우된다. 이를 위해 기존 한국어 음성합성시스템에서는 운율정보를 표현하기 위해 훈련과 합성에 사용되는 모든 전사데이터에 음절에서의 음소위치, 어절(단어)에서의 음절의 수, 어절에서의 음절위치, 음절 앞 또는 뒤에서의 끊어 읽기 정보, 어절 앞 또는 뒤에서의 끊어 읽기 정보 등을 포함시킨다. 하지만 이러한 정보들을 모두 포함시킬 경우 요구되는 훈련데이터의 크기가 지수적으로 증가하는 문제점이 있다.

또한 모든 문맥의존 정보를 포함하는 학습용 음성DB를 구성하는 것이 현실적으로 불가능하기 때문에 결정트리에 기반하여 유사한 발화특성의 문맥의존음소를 군집화한다. 그러나 군집화의 영향으로 훈련과정에 포함되지 않은 문맥의존음소는 유사한 문맥의존음소들로 구성되는 하나의 그룹으로 분류되어 해당 그룹의 대표값으로 평활화된다. 이로 인해 합성음의 자연성이 저하되는 문제가 있다[8].

본 논문에서는 운율정보를 단순화하면서 효과적으로 자연성을 향상시킬 수 있는 방법으로써 먼저 운율경계정보, 운율경계정보를 포함하는 문맥의존정보에 대한 결정트리의 질의, 그리고 해당 질의의 패턴을 정의하고 이를 이용한 HMM 학습과 합성음을 생성하는 HMM기반 합성음 자연성 향상 방법을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 베이스라인 시스템으로 사용한 HMM기반 음성 분석/합성 시스템의 구성을 살펴보고 3장에서는 제안하는 방법을 사용한 문맥의존정보 구성 방법과 결정트리의 질의문 그리고 질의문에 해당하는 응답패턴의 구성방법을 설명한다. 4장에서는 실험과 평가를 통해 제안된 방법의 성능을 확인하고 5장에서 결론을 서술하였다.

II. HMM기반 음성 분석/합성 시스템 구성

1. 음성합성을 위한 HMM 모델

음성인식에서 사용되는 기존의 HMM 모델은 관측벡터인 성도특성 파라미터와 성도특성 파라미터의 delta, delta-delta를 사용한다. 그리고 tri-phone의 문맥의존모델을 구성하고 결정트리에 기반하여 tri-phone의 중심음소에 대해 군집화함으로써 문맥의존 HMM을 훈련한다.

그러나 음성합성을 위해서는 성도특성 파라미터뿐만 아니라 기본주파수 (fundamental frequency, F0) 정보와 음소 지속시간에 대한 정보가 추가로 요구되며 문맥의존정보에서도 추가적으로 많은 문맥의존 요소들이 요구된다. 따라서 HTS에서는 관측벡터를 성도특성 파라미터, 기본주파수 파라미터 그리고 음소지속시간 파라미터로 구성하며 결정트리에 기반한 스트림의존 문맥 군집화방법을 사용한다. 이때 성도특성 파라미터는 0차 계수를 포함하는 mel-cepstrum 계수(MCC)와 MCC의 delta, delta-delta 계수를 사용하고 기본주파수 파라미터는 log 크기의 기본주파수 (logF0)와 logF0의 delta, delta-delta 계수가 사용된다[4].

1.1 성도특징 파라미터 모델

음성합성용 성도특징 파라미터는 HMM의 관측벡터로부터 정의될 수 있어야 한다. MFCC는 음성인식에서 널리 사용하는 성도특징 모델이지만 안정된 합성필터로는 사용되지 않는다. 따라서 MFCC와 유사한 특성을 가지면서 안정된 합성필터가 존재하는 mel-cepstrum 계수를 가우시안 확률 분포로 모델링하고 MCC를 MLSA 합성필터 파라미터로 사용하여 합성음을 생성한다[9-10].

1.2 F0 모델

F0의 패턴은 일차원의 연속 가우시안 분포로 표현되는 유성음과 이산적 기호로 표현되는 무성음으로 구성된다. 그러므로 기존 음성인식에서 사용되는 연속 HMM 또는 이산 HMM에 F0 패턴을 적용할 수 없다. 따라서 F0 패턴을 모델링하기 위한 수학적 모델인 MSD (Multi-Space probability Distribution) HMM이 사용된다[11-13].

1.3 상태지속시간 모델

음소에 대한 합성파라미터를 생성하기 위해서는 합성할 음소의 지속시간을 결정해야 한다. 음소에 대한 HMM 모델은 상태열로 구성되어 있으므로 음소의 지속시간은 각 상태의 지속시간에 의해 결정된다. HMM기반 음성합성시스템에서 상태지속시간 모델은 다변량 가우시안 분포로 모델링된다. 이때 HMM의 상태지속시간 밀도의 차원 수는 HMM을 구성하는 상태의 수와 같다[12]. 상태지속시간을 따로 모델링함으로써 음성합성시 다양한 발화속도를 간단하게 조절할 수 있으며 음소 HMM들의 훈련 과정에서 자동적으로 추정되므로 전사정보에서 음소경계정보가 불필요하다는 장점이 있다.

2. HMM 모델 훈련

훈련과정은 그림 1에 보인 바와 같이 크게 훈련용 음성DB 분석과정과 HMM 훈련과정으로 구성된다. 훈련용 음성DB 분석과정은 음성DB로부터 음성신호의 특징벡터인 기본주파수와 성도특징벡터를 의미하는 MCC를 추출한다.

HMM 훈련과정은 그림 2에 보인 바와 같이 음성인식에서 사용되는 문맥의존음소 HMM의 훈련과정과 유사한 방법으로 각 tri-phone의 state에 대한 HMM 모델 파라미터를 학습한다[4]. Boot-strap 훈련 음성DB의 전사정보로부터 훈련 문장이 정의되면 훈련 문장으로부터 전체평균과 전체분산을 구한다. 그리고 mHTK의 HCompV 모듈을 이용하여 초기 HMM을 구성하기 위한 초기값을 계산하여 훈련할 모든 단음소 HMM에 계산된 초기값을 할당한다. 초기값이 할당된 HMM모델들은 boot-strap 훈련 음성DB의 전사정보에 포함된 음소분할 정보와 mHTK의 HInit과 HRest모듈을 이용하여 각 음소별로 HMM의 초기값을 훈련한다. 초기값 훈련이 수행된 HMM들은 다시 mHTK의 HERest 모듈을 이용하여 전향-후향 알고리즘으로 훈련된다.

단음소별 훈련이 충분히 이루어지면 mHTK의 HHEd 모듈을 통해 단음소 HMM들을 tri-phone형태의 문맥의존음소 HMM으로 변환한다. 변환된 문맥의존음소 HMM들은 다시 HERest모듈을 통해 훈련된다. 훈련 음성DB에 포함된 문맥의존음소의 훈련이 충분히 이루어지면 훈련DB에 포함되지 않은 문맥의존음소를 위한 결정트리기반 군집화과정이 수행된다[1][5].

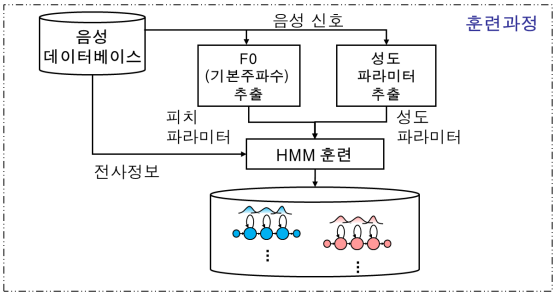

그림 1. 훈련과정의 블록도
Fig. 1. Block diagram of HMM training

그림 2. HMM 훈련의 흐름도
Fig. 2. Flow chart of the HMM training

최종적으로 문맥의존음소에 대해 훈련된 성도특징, F0, 그리고 상태지속시간 각각에 대한 pdf와 결정트리가 생성된다.

2.1 결정트리 기반 HMM 군집화

결정트리에 기반한 HMM군집화는 크게 결정트리 생성과정과 생성된 결정트리를 이용한 군집화과정으로 구성된다. 결정트리 생성과정은 노드를 나누는 과정과 트리생성을 중지하는 과정으로 구분할 수 있다.

음성인식에서의 노드분할은 노드집합들의 likelihood를 최대화 하도록 구성하고 노드생성 중지조건으로 노드분할에 의한 likelihood의 증가치에 대해 문턱값을 적용하여 트리생성을 중지한다. 단순히 likelihood만 고려할 경우 likelihood의 증가치가 문턱치에 도달하기까지의 계산비용이 데이터에 따라 변화가 심한 문제가 있다[14].

이를 보완하기 위해서 HTS에서는 likelihood 뿐만 아니라 노드분할에 따른 노드 집합의 복잡도를 description length를 통해 측정하고 description length를 최소로 하는 방향으로 노드분할을 수행하는 MDL (Minimum Description Length) 방법을 사용한다[15-16]. mHTK의 HHEd모듈을 이용하여 주어진 문맥질의로부터 MDL방법에 기반한 성도특징, F0, 상태지속시간 HMM의 결정트리를 생성한다[5].

HTS에서는 성도특징, F0, 그리고 상태지속시간 모델 각각이 서로 다른 문맥적 요소에 영향을 받기 때문에 각 모델별로 결정트리를 구성하는 stream-dependent 군집화 방법을 사용한다. 그리고 음성인식의 경우 tri-phone의 중심음소가 동일한 HMM을 묶어 중심음소 HMM별로 군집화를 수행하지만, HTS에서는 HMM의 동일한 상태를 묶어 상태별로 군집화 한다. HTS의 결정트리기반 군집화 과정에서는 생성된 성도특징, F0, 상태지속시간 HMM의 결정트리를 이용하여 군집화된 초기 HMM들을 얻고, 초기 HMM들은 mHTK의 HERest 모듈을 이용하여 훈련한다.

3. HMM 기반 음성 합성 시스템

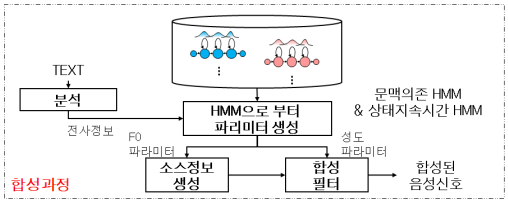

그림 3. HMM기반 음성합성 과정의 블록도
Fig. 3. Block diagram of HMM-based speech synthesis

합성과정은 그림 3에 보인 바와 같이 입력된 문장으로부터 언어처리를 통해 tri-phone형태의 문맥의존 음소열을 생성한다. 생성된 문맥의존 음소열로부터 문맥의존 HMM으로 구성된 문장 HMM을 구성하고 HMM기반의 지속시간 모델을 이용하여 각 tri-phone 모델의 상태 지속시간을 결정한다. Log 크기로 학습된 상태별 피치 모델을 이용하여 프레임 단위로 피치 값을 결정한다. 0차를 포함한 34차의 MCC를 활용한 스펙트럼 파라미터 모델을 이용하여 프레임별로 합성에 사용할 MCC 벡터열을 생성한다. 생성된 파라미터 열은 MLSA 필터를 사용하여 합성음으로 변환된다[9].

HMM으로부터 합성음을 생성하기 위해서는 HMM 모델 λ와 전체 지속시간 T가 주어졌을 때, 식 (1)에서의 확률값 P

를 최대로 하는 상태열 Q와 관측벡터열 O를 결정하여 합성에 필요한 합성파라미터를 생성한다[10-13].

$$\bar{\mathbf{O}} = \max_{\mathbf{Q}} \arg\max_{\mathbf{O}} P(\mathbf{O},\mathbf{Q}\,|\,\lambda) \qquad (1)$$

3.1 상태열 결정

상태지속시간 모델의 결정트리기반 문맥질의를 이용하여 문장 HMM의 각 상태별 지속시간을 결정한다. 상태지속시간 모델이 가우시안 확률분포로 모델링된 경우 해당 결정트리에 의해 선택된 확률밀도함수의 평균값과 분산값을 이용하여 식 (2)와 같이 결정할 수 있다.

$$d_k = m_k + \rho \bullet \sigma_k^2, \quad 1 \le k \le K \qquad (2)$$

$$\text{단, } \rho = \left(T - \sum_{k=1}^{K} m_k\right) \Big/ \sum_{k=1}^{K} \sigma_k^2$$

여기서 dk는 k번째 상태의 상태지속시간, mk는 k번째 상태의 평균지속시간, σk는 k번째 상태의 표준편차, K는 문장 전체의 상태수를 의미한다. 식 (2)를 통해 k번째 상태 지속시간을 결정할 수 있고 각 상태별 상태지속시간이 결정되면 문장 HMM을 구성하는 전체 상태열 Q가 결정된다.

3.2 관측벡터열 결정

식 (1)은 Bayes' rule에 의해 식 (3)과 같이 전개된다.

$$P(O,Q|\lambda,T) = P(Q|\lambda,T) \bullet P(O|Q,\lambda,T) \qquad (3)$$

상태열 Q가 주어졌을 때 식 (3)의 확률값을 최대로 하는 관측벡터열 O를 결정한다. 관측벡터열 O는 식 (4)와 같이 동적 특성윈도우벡터 W와 성도특징벡터 C로 구성할 수 있다.

$$O = WC \qquad (4)$$

식 (3)에 식 (4)를 대입하고 양변에 log를 취한 뒤 성도특징벡터 C에 대해 편미분한 결과가 0이 되도록 하는 최적의 C를 구하면 식 (5)가 된다.

$$RC = r \qquad (5)$$

$$\text{단, } R = W^T U^{-1} W, \; r = W^T U^{-1} \mu$$

여기서 U는 공분산행렬, μ는 평균값 벡터를 나타낸다. 식 (5)로부터 최적의 C를 구하고 식 (4)을 이용하여 관측벡터열 O를 결정할 수 있으므로, 합성에 필요한 F0, 성도특성 파라미터 MCC 벡터열을 생성할 수 있다.

3.3 합성음 생성

HMM 모델을 이용하여 생성된 합성필터 파라미터인 MCC를 이용하여 MLSA필터를 합성한다[9].

MLSA필터는 mel-log응답을 직접 근사화한 필터이기 때문에 지수함수 형태의 필터응답을 가진다. 지수함수 형태의 응답을 가지는 digital 필터를 설계하는 것은 불가능하기 때문에 일반적으로 식(7)과 같이 Padé근사화를 통해 지수함수를 유리함수로 근사화하여 구현한다.

$$H(z) = \exp F(z) \cong R_L(F(z)) \qquad (7)$$

$$= \frac{1 + \sum_{l=0}^{L} A_{L,l}\{F(z)\}^l}{1 + \sum_{l=0}^{L} A_{L,l}\{-F(z)\}^l}$$

$$\text{단, } F(z) = \sum_{m=0}^{M} c(m) z^{-m}$$

이때 c(m)은 MCC, AL,l은 L차 Padé근사화의 l번째 Padé근사화 계수를 의미한다. Padé근사화를 이용한 MLSA 필터는 그 구조는 그림 4와 같다. 그러나 Padé근사화를 이용한 필터는 delay-free loop를 포함하기 때문에 구현이 불가능하다.

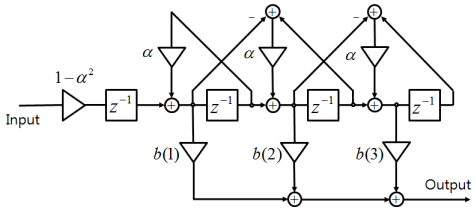

그림 4. MLSA필터의 기본필터의 블록도
Fig. 4. Block diagram of MLSA basic filter

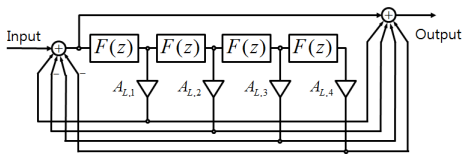

그림 5. MLSA필터의 블록선도
Fig. 5. Block diagram of MLSA filter

따라서 그림 5와 같이 delay free loop를 제거한 필터구조로 변경하고 MCC계수를 선형 변환하여 필터파라미터 벡터를 구한다. HMM 모델을 이용하여 생성된 기본주파수값으로 여기신호를 생성하고, 생성된 여기신호를 필터파라미터벡터가 적용된 MLSA필터에 입력시켜 프레임별로 음성신호를 합성한다.

III. 운율경계정보를 이용한 HMM 상태 군집화

1. 군집화를 위한 한국어 문맥정보 분류

유사한 음향특성을 가지는 음소들을 군집화 하기위해서는 분류 기준이 요구된다. 본 연구에서는 한국어 자음과 모음의 변이음 분류 기준을 표 1, 표 2와 같이 정의하였다.

표 1. 한국어 모음의 조음방법에 따른 분류
Table 1. Vowel clustering according to the articulation

분류			모음 음소(음소기호)
단모음	평순 저모음		ㅏ(a), ㅓ(v)
	후설 원순모음		ㅗ(o), ㅜ(u)
	전설 평순 고모음		ㅣ(i)
	후설 고모음		ㅡ(U)
	전설 평순 준고모음		ㅔ(E), ㅐ(e)
이중모음	앞모음	전설 평순 고모음	ㅑ(ja), ㅕ(jv) ㅛ(jo), ㅠ(ju), ㅒ(jE), ㅖ(je)
		원순모음	ㅘ(wa), ㅝ(wv) ㅙ(wE), ㅞ(we) ㅚ(we), ㅟ(wi)
		후설 평순 고모음	ㅢ(Wi)
	뒷모음	평순 저모음	ㅑ(ja), ㅕ(jv), ㅘ(wa), ㅝ(wv)
		후설 원순모음	ㅛ(jo), ㅠ(ju)
		전설 평순 고모음	ㅟ(wi), ㅢ(Wi)
		전설 평순 준고모음	ㅒ(jE), ㅖ(je), ㅙ(wE), ㅞ(we), ㅚ(we)

표 2. 한국어 자음의 조음방법에 따른 분류
Table 2. Consonant clustering according to the articulation

분류			자음 음소(음소기호)
초성	유성음	비음	ㄴ(n), ㅁ(m)
		설측음	ㄹ(r)
	무성음	경음	ㄲ(G), ㄸ(D), ㅃ(B), ㅆ(S), ㅉ(Z)
		격음	ㅍ(p), ㅌ(t), ㅋ(k), ㅊ(c)
		파열음	ㄱ(g), ㄷ(d), ㅂ(b)
		마찰음	ㅅ(s), ㅈ(z)
		모음성 마찰음	ㅎ(h)
종성	유성음	비음	ㄴ(N), ㅁ(M), ㅇ(O)
		설측음	ㄹ(L)
	무성음	폐쇄음	ㄱ(K), ㄷ(T), ㅂ(P)
묵음			#e, #l, #s

그리고 표 1과 표 2의 분류 기준을 따르는 결정트리를 위한 질의와 그 질의에 해당하는 문맥의존패턴이 정의되어야 한다. 결정트리의 질의는 분류요소 각각에 대하여 정의되고 정의된 질의에 해당하는 문맥패턴은 해당 분류에 속한 음소 기호로써 정의하였다. 묵음의 경우 합성의 시작과 끝, 끊어읽기 유무를 모델링하기 위해 세 가지 휴지구간 기호를 정의하였다.

2. 운율경계정보 구성

운율경계현상을 모델링하기 위해서 먼저 운율경계정보의 유형을 구분해야 한다. 본 연구에서는 운율경계현상을 크게 운율경계위치에서의 억양의 상승, 하강, 평탄 3가지 유형으로 분류하였다. 억양의 상승, 하강, 평탄 3가지 운율경계정보를 기호에 추가로 정의하고 표 3과 같이 표기하였다. 표 3과 같이 세 가지 유형의 운율경계정보 표현을 가지는 음소기호들을 HMM으로 모델링하고, 각 모델들을 tri-phone 형태의 문맥의존 HMM으로 변환하였다. 그리고 문맥의존 HMM으로 변환된 각 모델들을 훈련하고 결정트리기반 군집화를 수행하였다.

표 3. 운율경계유형별 음소기호 정의
Table 3. phone symbol according to the prosodic boundary type

억양상승 어절경계음소	운율하강 어절경계음소	그 외 어절경계음소
〔음소기호〕2	〔음소기호〕1	〔음소기호〕0
〔음소기호〕 : (초성) g, n, d, r, m, ... (중성) a, ja, v, ... (종성) K, N, T, L, ...		

3. 운율경계정보를 이용한 HMM군집화

운율경계정보를 문맥의존정보에 추가하기 때문에 기존 tri-phone보다 문맥의존정보의 조합이 크게 증가한다. 따라서 군집화과정은 필수적인 요소이며 가장 중요하다.

운율경계정보의 경우 운율경계가 아닌 중성과 종성의 발생빈도에 비해 그 발생빈도가 낮기 때문에 군집화의 영향을 많이 받는다. 또한 운율경계정보는 군집화로 인한 평활화의 영향으로 인해 합성음의 자연성 변화가 민감하게 나타나기 때문에 군집화로 인한 HMM의 평활화 현상을 최소화 할 필요가 있다. 그리고 MDL을 이용하는 결정트리 구성방식에서 트리의 depth가 커질수록 복잡도는 증가하게 된다. 따라서 트리의 depth가 불필요하게 커지지 않도록 운율경계음소에 대한 질의와 응답패턴들을 운율경계가 아닌 문맥의존음소의 질의에 비해 세분화시키는 방법으로 문제를 해결하였다.

노드 분할에 의한 복잡도의 증가를 최소화하기 위해서 운율경계정보를 포함하는 질의의 경우 모음전체를 하나의 질의 응답패턴으로 그룹화하는 질의를 구성하지 않고 단모음과 이중모음으로 세분화하여 그룹화하였다. 또 그 하위 질의 항목은 운율경계정보를 포함하지 않은 문맥의존음소에 비해 단모음과 이중모음을 혀의 위치 (전설, 중설, 후설), 입술 모양 (평순, 원순), 그리고 혀의 높이 (고모음, 저모음, 반고모음, 반저모음) 등과 같이 세분화하여 질의의 응답패턴을 정의하였다. 표 4에 문맥질의와 응답패턴 구성 예를 나타내었다.

표 4. 문맥질의문의 예
Table 4. Example of context question

문맥 질의	응답패턴
L_Monothong-Flat_Open	{a-*,v-*}
L_Back_Rounded	{o-*,u-*,jo-*,ju-*}
L_Back_closed	{U-*,Wi-*}
L_Front_Rounded	{o-*,u-*,wa-*,wv-*,we-*,wE-*,wi-*}
L2_Monothong-Back_Rounded	{o2-*,u2-*}
L2_Monothong-Front_flat_closed	{i2-*}
L2_Monothong-Back_closed	{U2-*}
L2_Monothong-Front_Flat_SemiClosed	{e2-*,E2-*}
L2_Diphthong-Front_Part-Front_Flat_Closed	{ja2-*,jv2-*,jo2-*,ju2-*,jE2-*,je2-*}
...	

IV. 실험 및 결과

실험을 위해 남성화자가 낭독체로 발성한 2,000문장의 한국어 ETRI 음성 DB를 사용하였다. 문장은 4~12 어절로 구성되는 방송뉴스, 신문기사 코퍼스에서 발췌되었으며 46,242 음절, 13,264 단어로 구성되고 음성데이터 용량은 283MB이다. 이 중 1,800문장은 훈련과정에 사용하고 200문장은 합성음 평가에 사용하였다. ETRI 음성데이터는 16bit, 16kHz로 표본화되어 있다. HMM훈련에는 HTS용으로 수정된 HTK를 사용하였으며 스펙트럼 파라미터는 0차를 포함한 34차의 MGC를 사용하였다. 전사정보는 ETRI에서 제공한 전사정보를 기반으로 본 논문에서 제안한 운율경계정보를 추가하여 사용하였다. 결정트리기반 군집화 과정에서도 운율경계정보를 포함한 tri-phone들과 질의를 적용하여 훈련하였다.

결정트리를 위한 질의의 세분화와 관련된 실험에서 운율경계 비적용 문맥의존음소에 대해서 큰 범주로 그룹화하는 질의들을 제외시키면 합성음의 자연성이 낮아지는 결과를 보였다. 그리고 운율경계 적용 문맥의존음소의 경우 질의구성을 세분화시킬수록 합성음의 자연성이 향상됨을 확인하였다. 따라서 운율경계 비적용 문맥의존음소는 질의와 응답패턴을 큰 범주의 그룹화에서 작은 범주의 그룹화로 이어지도록 구성하였고, 운율경계 적용 문맥의존음소의 경우 질의문 구성을 트리의 depth가 커지지 않도록 세분화된 그룹으로 정의하였다.

결정트리 기반 군집화의 결과로 생성된 결정트리와 확률밀도함수의 정보를 저장하는 파일크기 변화를 표 5에 나타내었다. 결정트리의 질의가 증가하여도 운율경계요소의 발생빈도가 낮기 때문에 결정트리와 확률밀도함수의 파일사이즈의 변화가 크지 않음을 확인할 수 있다.

표 5. 운율경계정보 포함 여부에 따른 결정트리와 확률밀도함수 정보파일의 크기 비교
Table 5. Comparison of information file sizes between with and without the prosodic boundary information

구분	tree(Kbyte)	pdf(Kbyte)
문맥의존정보만 이용한 군집화	202.3	781.5
운율경계정보가 적용된 군집화	211.0	806.5

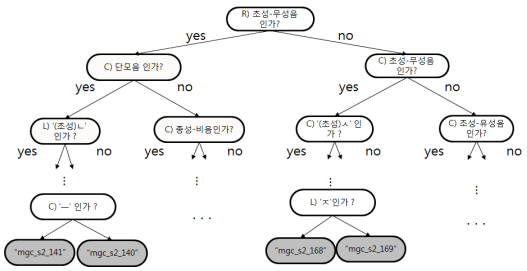

그림 6. 결정트리구조의 예
Fig. 6. Example of the decision tree structure

표 6. 생성된 결정트리의 구성
Table 6. Structure of the generated decision trees

구분	문맥질의 수	노드 수	leaf노드 수
상태지속시간 모델	111개	166	168
F0 모델	254개	1267	1277개
성도특성모델	220개	869개	879개

본 논문에서 제안한 문맥질의로부터 생성된 성도특성, F0, 상태지속시간 모델 결정트리의 예는 그림 6과 같고 생성된 모델별 결정트리의 문맥질의 수, 노드 수, 그리고 leaf 노드 수를 표 6에 나타내었다.

그림 7은 "내가 뭘 잘못했다고 이름을 대냐구요"로 녹음된 문장의 원음 파형과 피치 곡선을 나타낸다. 그리고 합성음 평가에 사용된 음성파형과 동일한 문장의 전사정보를 이용하여 운율정보를 추가하지 않은 합성음과 본 논문에서 제안하는 운율경계정보를 추가한 합성음의 음성파형과 피치 곡선을 그림 8과 그림 9에 각각 나타내었다.

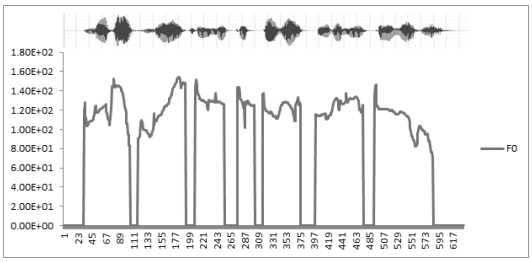

그림 7. 원음 신호의 파형과 피치 곡선
Fig. 7. The waveform and pitch contour of original speech.

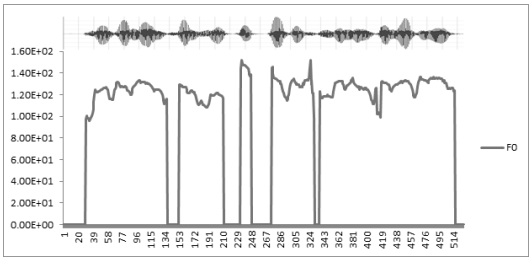

그림 8. 운율정보를 사용하지 않은 합성음의 파형과 피치 곡선
Fig. 8. The waveform and pitch contour of synthetic speech not using the prosodic boundary information.

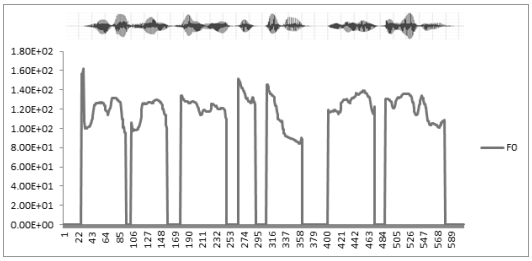

그림 9. 운율경계정보를 사용한 합성음의 파형과 피치 곡선
Fig. 9. The waveform and pitch contour of synthetic speech using the prosodic boundary information.

표 7. MOS 평가 결과
Table 7. The results of the MOS about the synthetic speech

구분	명료도	자연성
문맥의존정보만 이용한 합성 방법	3.20	2.72
운율경계정보를 이용한 합성 방법	3.44	3.10

지속시간 모델의 영향으로 각 음소들의 지속시간은 다소 차이가 있지만, 본 논문에서 제안한 방식이 원음의 파형과 피치 곡선에 가까운 합성음을 생성할 수 있음을 확인할 수 있다. 운율정보를 포함하지 않는 데이터를 이용하여 학습된 HMM

을 사용한 합성음과 제안된 방법을 사용한 HMM의 합성음의 명료도와 자연성을 MOS (Mean Opinion Score)를 통해 평가한 결과를 표 7에 나타내었다.

MOS 평가 결과에서 볼 수 있듯이 어절경계에 대한 표현이 운율경계정보를 사용하지 않는 경우보다 정확해지면서 명료도가 상승하였고, 특히 자연성 부분에서 큰 향상을 보임을 확인할 수 있다. 또한 어절경계정보가 많이 사용된 문장이거나 문장 내에 운율변화가 많을수록 자연성이 크게 향상되는 것을 확인할 수 있었다.

V. 결 론

본 논문에서는 HMM기반으로 한국어 텍스트를 음성으로 변환하는데 있어서 합성음의 자연성을 향상시키기 위해 운율경계정보를 세 가지 유형으로 정의하고, 운율경계정보를 포함한 tri-phone들의 결정트리기반 군집화 방법을 제안하였다. 그리고 운율경계정보를 포함하는 문맥의존정보에 대한 문맥질의, 그리고 해당 질의의 패턴을 정의하는 방법을 제안하였다.

본 논문에서 제안한 방법을 이용하면 전사정보의 운율정보를 표현하기 위한 구성이 간결해졌다. 그리고 운율경계정보를 포함한 문맥의존정보를 이용하여 합성음을 생성하고 MOS평가를 수행한 결과 운율경계정보를 이용한 HMM기반 한국어 TTS 합성음의 자연성이 향상됨을 확인하였다.

VI. 감사의 글

이 논문은 울산대학교 연구비에 의하여 연구되었음.

참고문헌

[1] K. Tokuda, H. Zen, and A.W. Black, "An HMM based approach to multilingual speech synthesis," Text to speech synthesis: New paradigms and advances, S. Narayanan, A. Alwan (Eds.), Prentice Hall, pp.135-153, Aug. 2004.

[2] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," Proc. ICASSP 2007, vol. 4, pp. 1229-1232, Apr. 2007.

[3] H.C. Lee, and J.M. Seo, " A study of Implementing An Embedded System for Conversion from Text to Speech ," Journal of the Korea Society of Computer and Information, v.13, no.3, pp.77-83, May 2008.

[4] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The Hidden Markov Model Toolkit (HTK)," http://htk.eng.cam.ac.uk/

[5] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, and T. Nose, "The HMM based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/

[6] A.W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," http://www.festvox.org/ festival/

[7] S. Kim, J. Kim, and M. Hahn, "HMM-Based Korean Speech Synthesis System for Hand-held Devices," IEEE Trans. Consumer Electronics, vol. 52, no. 4, pp.1384-1390, Nov. 2006.

[8] J. Lee, "A Tree-based Reduction of Speech DB in a Large Corpus-based Korean TTS," Journal of the Korea Society of Computer and Information, v.15, no.7, pp.91-98, Jul. 2010.

[9] S. Imai, "Cepstral analysis synthesison the mel-frequency scale," Proc. ICASSP, vol. 1, pp. 93-96, Apr. 1983.

[10] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," Proc. of EUROSPEECH,

vol. 1, pp. 757-760, Sep. 1995.
〔11〕 J. Latorre, and et. al., "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?," Proc. ICASSP, pp. 4724-4727, May 2011.
〔12〕 Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V /U decision in HMM-based TTS," Proc. ICASSP, pp. 4606-4609, Mar. 2010.
〔13〕 K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM (Invited paper)," IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp.455-464, Mar. 2002
〔14〕 S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Human Language Technology Workshop, pp. 307 - 312, Mar. 1994.
〔15〕 K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn.(E), vol.21, no.2, pp. 79-86, Feb. 2000.
〔16〕 K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," Proc. Eurospeech, vol. 1, pp. 99-102, Sep. 1997.

저 자 소 개

임 기 정

2010 : 울산대학교
컴퓨터정보통신공학부 공학사.
현 재 : 울산대학교
컴퓨터정보통신공학부 석사졸업
관심분야 : 디지털신호처리, 음성합성
Email : gadama2@gmail.com

이 정 철

1984년 : 서울대학교
전자공학과 공학사
1988년 : 서울대학교
전자공학과 공학석사
1998년 : 서울대학교
전자공학과 공학박사
1985년~2000년 : ETRI 책임연구원
2000년 : L&H Korea 전문위원
2001년 : (주)보이스텍 전문위원
2002년 : (주)코난테크놀로지 책임연구원
2002년 9월~현재 :
울산대학교 전기공학부 부교수
관심분야 : 디지털신호처리,
음성신호처리, 음성합성
Email : jungclee@ulsan.ac.kr