한국음향학회지 제25권 제7호 pp. 312~318 (2006)

클래스 종속 반연속 HMM을 이용한 립싱크 시스템 최적화

Lip-Synch System Optimization Using Class Dependent SCHMM

이 성 희*, 박 준 호*, 고 한 석*
(Sunghee Lee*, Junho Park*, Hanseok Ko*)

*고려대학교 전자컴퓨터 공학과

(접수일자: 2006년 8월 25일; 채택일자: 2006년 9월 26일)

기존의 립싱크 시스템은 음소 분할 후, 각각의 음소를 인식하는 2단계의 과정을 거쳤다. 하지만, 정확한 음소 분할의 부재와 음성이 끊긴 분할 된 음소로 이루어진 훈련 데이터들은 시스템의 전체 성능을 크게 떨어뜨렸다. 이런 문제를 해결하기 위해 Head-Body-Tail (HBT) 모델을 이용한 단모음 연속어 인식 기술을 제안한다. 주로 소규모 어휘를 다루는데 적합한 HBT 모델은 Head 와 Tail 부분에 문맥 종속 정보를 포함하여 앞 뒤 문맥에 따른 조음효과를 최대한 반영한다. 또한, 7개의 단모음을 입모양이 비슷한 세 개의 클래스로 분류하여, 클래스에 종속적인 코드북 3개를 가진 반연속HMM (Hidden Markov Model)을 적용하여 시스템을 최적화하고, 변이 부분이 큰 단어의 처음과 끝은 연속HMM의 8 믹스쳐 가우시안 구조를 사용하여 모델링하였다. 제안한 방법은 HBT구조의 연속HMM과 대등한 성능을 보이지만, 파라미터 수는 33.92% 감소하였다. 파라미터 감소는 계산 양을 줄여주므로, 시스템이 실시간으로 동작 가능하게 한다.

핵심용어: Head-Bod-Tail (HBT), 문맥 종속, 단모음 연속어 인식, 반연속 HMM, 립싱크

투고분야: 음성처리 분야 (2.5)

The conventional lip-synch system has a two-step process, speech segmentation and recognition. However, the difficulty of speech segmentation procedure and the inaccuracy of training data set due to the segmentation lead to a significant performance degradation in the system. To cope with that, the connected vowel recognition method using Head-Body-Tail (HBT) model is proposed. The HBT model which is appropriate for handling relatively small sized vocabulary tasks reflects co-articulation effect efficiently. Moreover, the 7 vowels are merged into 3 classes having similar lip shape while the system is optimized by employing a class dependent SCHMM structure. Additionally, in both end sides of each word which has large variations, 8 components Gaussian mixture model is directly used to improve the ability of representation. Though the proposed method reveals similar performance with respect to the CHMM based on the HBT structure, the number of parameters is reduced by 33.92%. This reduction makes it a computationally efficient method enabling real time operation.

Key words: Head-Bod-Tail (HBT), Context dependent, Connected vowel recognition, Lip-synch

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

현재 국내의 립싱크 시스템은 오프라인 상에서 음성 합성을 통해 아바타를 움직인다. 아바타는 3D를 이용하여 입모양이나 얼굴 근육 등을 자연스럽게 움직여 [1] 실물과 비슷하도록 시스템을 구축하고 있지만, 오프라인 상에서만 작동하는 한계를 가진다. 연속적으로 음성이 입력 될 때, 각각의 음소를 아바타의 입모양에 동기화 시켜 실시간으로 영상을 보여주는 시스템을 실시간 립싱크 시스템이라고 하며, 이는 인터넷을 통한 실시간 아바타 채팅, 전화 통화 시 청각 장애인을 위한 그래픽 스크린, 또는 지능 로봇의 입모양 구현 등 여러 분야에 널리 응용될 수 있다. 현재 실시간 립싱크 기술의 방법으로,

책임저자: 고 한 석 (hsko@korea.ac.kr)
서울시 성북구 안암5가-1 고려대학교 전자컴퓨터 공학과
(전화: 02-3290-3239; 팩스: 02-3291-2450)

음성 특징과 영상 특징 (입술 파라미터)을 1:1 또는 다대일 매칭 시켜 [2][4] 입술 좌표를 추정하는 방법을 사용하며, 음성신호와 영상 프레임을 joint하기 위해 VQ (Vector Quantization), NN (Neural Network) 또는 GMM (Gaussian Mixture Model)[2][3][4] 등의 훈련 알고리즘을 사용한다. 하지만, 이런 방법들은 많은 양의 음성, 영상 데이터와 정확한 입술 파라미터의 좌표를 요구하므로 정확한 실험에 어려움이 생긴다. 본 논문은 정확한 인식과 실시간 립싱크 시스템 구현에 초점을 두었으며, 입력된 음성은 음소분할과정 없이 고정 된 프레임마다 특징을 추출하여 모음정보만을 인식하였다. 기존 립싱크 시스템의 음소 인식은 음소 분할의 정확성에 따라 성능이 좌우 되었지만 [5], 음소 분할 과정을 없앰으로써 전체적인 성능 향상을 도모하였다. 기존 훈련 데이터의 경우, 모음 (/아/, /에/, /이/, /오/, /우/, /으/, /어/)들을 수동으로 직접 분할하여 얻은 음소 파일을 사용하였지만, 분할 된 데이터는 음성이 끊겨 있으므로, 연속적으로 들어오는 음성에 대한 DB의 신뢰성을 떨어뜨리고 인식률 측정을 부정확하게 했다. 이를 해결하기 위해 우리는 음소 분할을 생략하고 고정된 프레임단위로 들어오는 음성 정보들을 해당 모음 정보로 그룹 화하여 인식함으로써, 연속적으로 들어오는 발성에 대한 DB의 신뢰성을 확보했다. 또한, 각각의 음소 인식 율을 높이기 위해 Chou [6]가 제안한 Head-Body-Tail (HBT) 구조 [6][7]을 사용함으로써 연속으로 발화되는 음성의 모음 인식성능을 크게 높였다. 하지만, HBT 구조는 파라미터의 수를 크게 증가시켜 계산 처리 양을 증가시키는 단점을 가진다. 본 논문에서는 시스템의 처리시간을 줄이기 위해 클래스 종속의 반연속 HMM (Hidden Markov Model)을 사용하였고, 변이 부분이 큰 단어의 처음과 끝 스테이트는 8 mixture 가우시안의 연속 HMM을 사용하여 인식 율을 높였다. 시스템 구현은 실시간 처리를 위하여 5프레임 (1/20s)마다 단어 가능성이 가장 높은 인식 결과를 영상으로 보여준다.

II. 전체 립싱크 시스템 개요도

다음 그림 1 은 전체 립싱크 시스템을 나타낸 개요도이다. 시스템은 크게 음성 분석 모듈과 음소를 해당 영상으로 보여주기 위한 음성 영상 동기화 모듈로 나눠진다. 우선, 마이크로폰을 통해 들어간 음성은 음성 분석

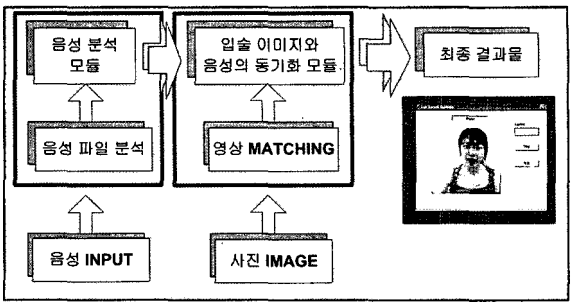


그림 1. 립싱크 시스템 전체 개요도
Fig. 1. Block diagram of Lip-synch system.

모듈을 통해 7개의 단모음 중 하나로 인식되며, 인식된 단모음은 정해진 Viseme 클래스 테이블을 참고하여 [8] 해당 이미지에 매핑 하고, 최종 결과물로 동기화된 이미지를 실시간으로 보여준다. Viseme 이란 우리가 눈으로 구분 할 수 있는 영상의 최소 단위로서 음성의 최소 단위인 phoneme과 비견될 수 있다. 다음 3장과 4장에서는 정확한 단모음 인식을 위한 알고리즘을 제안하고 5장에서는 아바타 이미지의 생성과 처리 방법에 대해 설명할 것이다.

III. HBT 구조의 Hidden Markov Model

HBT 방법은 Chou [6][7]에 의해 제안된 방법으로, 주로 연결숫자인식에 적용되었다. HBT란 Head-Body-Tail 구조로 인식하는 단어에 대해 Head, Body, Tail의 세부분으로 나눈다. Body부분은 문맥 독립형으로, 문맥에 따라 변하지 않는 특성이 있고, Head와 Tail부분은 문맥 종속형으로 앞뒤의 단어에 따라 변한다. 음운 변화를 고려한 문맥 종속 모델의 사용은 문맥 독립형 모델만 사용했을 경우와 비교하여 성능이 크게 향상되었다.

3.1. 음운 변화를 고려한 문맥 종속 및 독립 구조

문장이 발화 될 때, 문장을 구성하는 단어와 단어사이의 음운 변화는 크다. 즉, 단어의 처음부분과 마지막부분의 음운 변화는 다음단어, 또는 이전 단어에 의해 크게 영향을 받는다. 본 논문에서는 단어의 처음 부분과 마지막 부분을 Head 부분, Tail부분이라고 정의하고, 변화가 없는 단어의 중간부분을 Body 부분으로 정의한다. 문맥 독립형 구조는 앞뒤에 오는 단어에 영향을 받지 않고 간단하게 표현된다. 이 모델은 (/아/, /이/, /우/, /에/, /오/, /으/, /어/) 7개의 단모음을 기준으로, 7개의 Body로 구성되며, 각각의 Body는 3 스테이트 8

mixture 가우시안 set을 가진다. 문맥 종속형 구조는 문맥 독립형 Body를 중심으로 cloning에 의해 바이 폰 구조로 생성된다. 다양한 가우시안 믹스쳐는 음성 데이터의 특성을 잘 반영 하는 장점이 있지만, 파라미터의 수를 증가시키므로, 적당한 믹스쳐 수의 선택은 중요하다.

3.2. HBT 구조 기반의 모음 네트워크 구성

그림 2 는 고립단어 '한국 [hankuk]' 을 HBT 네트워크 구조로 나타내었다. 사용된 단어는 자음과 모음으로 이루어져 있으며, 모음정보로만 이루어진 HMM을 생성하기 위해 단어 사전은 새로운 단어 사전으로 재정의 하였다. 새로운 단어 사전은 자음을 제거하고 이중모음은 가장 비슷한 소리로 발음되는 단모음으로 매칭 시켜 만들어졌으며, 오직 7개의 단모음 정보로만 구성하였다. 모음-자음모음 연결에서 자음을 제거 하고 모음간의 문맥만을 고려한 것은 모음-모음 동시조음 현상의 '순행동시조음' 과 '역행동시조음' 현상 [9]을 반영한 것이다. 표1 은 기존 단어 목록에 있는 44개의 음소들 중, 자음을 제외시키고 이중모음과 단모음을 7종류의 단모음으로 영상에 매칭 시킨 Viseme 클래스 테이블이다. 이중모음은 7개의 단모음 중 가장 비슷한 발음으로 발성되는 Viseme 클래스 에 각각 그룹화 시켰으며, 자음은 제거 하고 음성 발화의 시작과 끝은 묵음 처리 하였다. 영상은 모두 묵음 포함 8개의 Viseme 클래스로 나뉐으며, 각각 그룹 화된 음소들의 대표되는 모음은 대문자 알파벳으로 표기하였다. 고립단어 '한국' 을 구성하는 단모음 /아/, /우/는 각각 HBT모델로 확장되며, Head와 Tail부분은 문맥 종속형 모델의 바이 폰 구조를 형성한다. 문맥 종속형과 문맥 독립형으로 이루어진 HBT 모델은 트라이 폰 모델보다 sub-word의 개수가 훨씬 적고, 단어의 처음과 끝의 변화를 반영하기 때문에 높은 인식

표 1. 각 음소에 해당하는 Viseme 클래스
Table 1. Viseme class of each phoneme.

Viseme 클래스	음소종류	영문 표현	한국어 표현
1	/a/, /wa/, /ya/	A	/아/
2	/e/, /yae/, /ye/, /we/	E	/에/
3	/o/, /yo/	O	/오/
4	/u/, /yu/	U	/우/
5	/eo/	V	/어/
6	/eu/	EU	/으/
7	/i/, /wi/, /eui/	I	/이/
8	묵음	SIL	/. /

률을 얻을 수 있다. 그림 3은 HBT 모델 /아/의 네트워크 구조를 나타낸 블록도이다. 이 그림에서 문맥 종속형 구조의 Head/Tail 부분은 각각 8개의 Head, Tail sub-word를 가진다. 구체적으로, 'HEAD A' 는 이전에 발음될 가능성 있는 8종류 (7개의 모음, Silence)의 Tail 모델과 결합되어 /아/의 Head부분을 나타내며 'TAIL A'도 역시, 다음에 발음될 후보 모음 8종류의 Head 모델과 결합되어 Tail부분을 확장시킨다. 즉, 단모음 /아/에 대해 8개의 Head, 8개의 Tail 그리고 1개의 Body를 생성한다. 각각의 단모음 마다 총 17개의 sub-word가 생성되므로, 7개의 단모음 경우, 총 119개의 sub-word를 생성한다. 결국, 각각의 단모음에 대해 생성된 HBT 구조는 한국어의 동시조음 효과를 Head/Tail 부분에 반영함으로써, 들어오는 연속음성을 효과적으로 모델링한다.

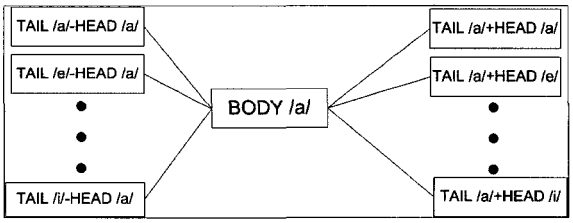


그림 3. 모음 /아/에 대한 HBT 구조
Fig. 3. HBT Structure of vowel /a/.

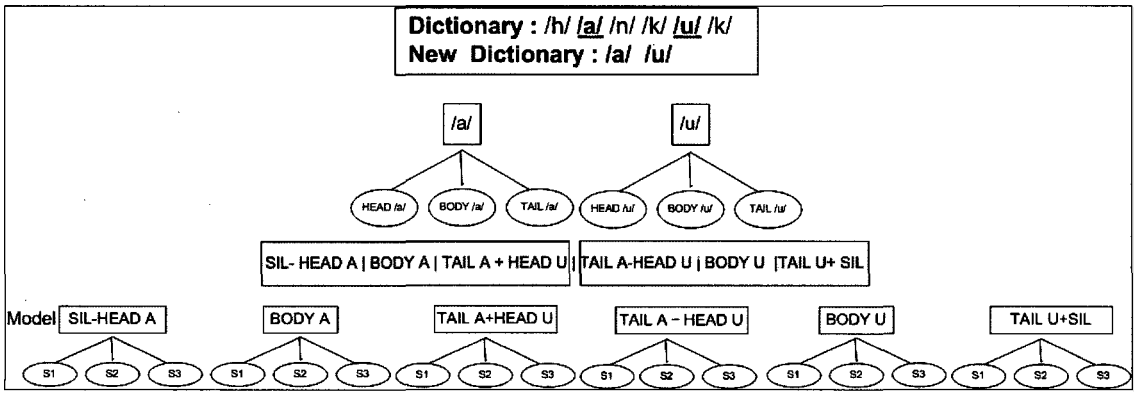


그림 2. 한국 (hankuk)의 HBT 구조
Fig. 2. HBT structure of 'Hankuk' .

IV. 시스템 최적화를 위한 반연속 HMM

본 논문에서는 연속적으로 들어오는 문장 속의 단어와 단어 사이에서 일어나는 다양한 변이를 적용하기 위해 HBT구조의 연속 HMM을 사용하였다. 하지만, 연속 HMM은 늘어나는 파라미터로 인해 실시간 립싱크 시스템의 동작에 어려움을 준다. 이를 해결하기 위해, 반연속 HMM에 기반을 둔 HBT모델을 사용하였다. 연속 HMM에 기반을 둔 HBT 모델은 인식률을 높이기 위해 수천 개의 스테이트와 수백만 개의 가우시안 파라미터를 생성하지만, 과도한 파라미터 수의 증가와 파라미터의 중복으로 인해 메모리의 비효율을 가져온다. 즉, 시스템의 실시간 처리를 위해 모델 크기를 줄이는 것이 필요하다.

4.1. 반연속 HMM

반연속 HMM [10][11]은 학습 과정에서 모든 모델의 상태에서 공유되는 M개의 가우시안 확률 밀도와 각 가우시안 확률 밀도들의 가중치를 결정하는 혼합 밀도 계수에 의해 확률을 얻는다. 정확한 음성 모델링을 위한 HMM 개선은 궁극적으로 파라미터수를 늘려 시스템의 복잡성을 증가시키지만, 한정된 DB에 과도하게 세분화된 모델은 통계적인 신뢰도를 저하시켜, 인식률을 떨어뜨린다. 본 논문은 제한된 데이터로 신뢰성 있는 통계치를 얻기 위해 반연속 HMM을 제안한다.

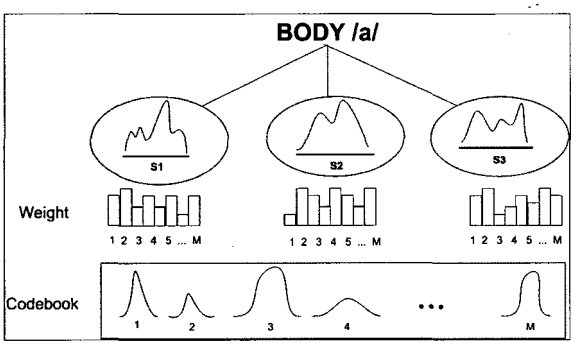


그림 4. Body /아/의 반연속 HMM 구조
Fig. 4. SCHMM structure of 'BODY /a/.

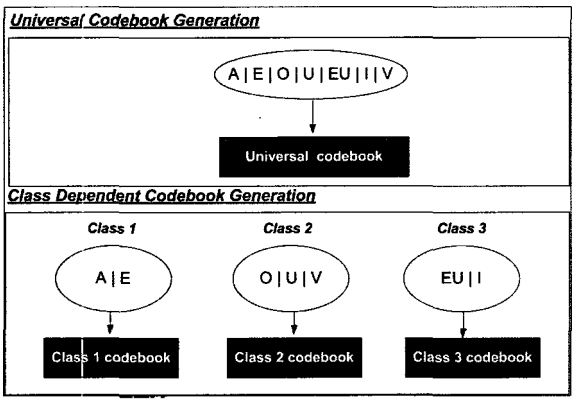


그림 5. 클래스 종속의 코드북 생성
Fig. 5. Block diagram of class dependent codebook generation.

4.2. HBT 구조를 가진 클래스 종속 반연속 HMM

HBT 구조를 가지는 연속 HMM의 경우, 파라미터 수는 급격히 증가하여, 시스템에 부하를 주게 된다. 이런 문제를 해결하기위해 256개 가우시안 공통 코드북을 가지는 반연속 HMM을 사용하였으며, 각각의 스테이트 출력 확률은 코드북에 생성된 가우시안들의 weight 정보 합들로 이루어진다. 그림 4는 Body /아/의 반연속 HMM 구조를 나타낸 것이다. Body/아/는 3개의 스테이트를 가지며, 각각의 스테이트 출력확률은 코드북에 있는 가우시안들의 weight 합으로 표현된다. M은 코드북 안의 가우시안 개수를 의미하며, 각 S1, S2, S3은 각각 첫 번째 스테이트, 두 번째 스테이트, 세 번째 스테이트를 의미한다. 각 스테이트는 M개의 서로 다른 weight 정보를 가진다. 하지만, 공통의 코드북을 가지는 반연속 HMM은 연속 HMM보다 파라미터의 이산화로 인해 성능을 떨어뜨리므로, 이를 보완하기 위해 예외적인 연속 HMM을 적용한 클래스 종속의 반연속 HMM을 제안한다. 그림 5는 제안한 클래스 종속의 코드북과 공통의 코드북을 가진 반연속 HMM을 비교하였다. 공통의 코드북을 가진 반연속 HMM은 256개의 가우시안을 가지며, 각 스테이트는 256개의 weight 정보로 구성된다. 반면 클래스 종속의 반연속 HMM은 클래스 종속의 3개 코드북을 생성하며, 각각의 스테이트는 32개의 weight 정보를 가진다. 클래스 종속이란 7개의 단모음을 입모양이 비슷한 모음끼리 3개의 클래스로 나눈 것으로 각 단모음의 코드북은 32개의 가우시안을 가진다. 음향 모델링 최적화를 위해 변이부분이 큰 Head의 첫 번째 스테이트와 Tail의 마지막 스테이트는 8 mixture를 가지는 예외적인 연속 HMM을 사용하여 변이를 잘 반영하도록 모델링하고, 나머지 부분은 클래스 종속 반연속 HMM을 사용하였다. 이는 기존의 연속 HMM이 가진 메모리 문제를 해결하고, 동시에 반연속 HMM이 가진 인식 성능의 저하를 최소로 한다.

V. 실시간 립싱크 시스템 구현

실시간 립싱크 구현을 위해서는 음성의 각 프레임을 좀 더 빠르고 정확하게 인식해야 한다. 최근 립싱크 기술로, 음성신호의 Mel Frequency Cepstrum Coefficient (MFCC) 과 영상의 특징 파라미터 (가로, 세로길이)를 joint하여 훈련하는 방법을 많이 사용하는데, 훈련을 위해 많은 양의 음성, 영상 데이터와 정확한 입술 파라미터의 좌표를 요구하는 단점을 가진다. 또한, 훈련에 사용되는 동영상 데이터와 신뢰성 있는 입술 좌표를 얻는 것은 현실적으로 어려우므로, 본 논문에서는 음성정보만을 이용하여 해당 음소를 인식한 후, 영상에 매핑 시키는 방법을 제안한다. 제안된 시스템은 끝점 검출 후, 인식결과를 표현하지 않고, 프레임 별로 가능성이 가장 큰 인식 단어 결과를 Viseme 클래스 테이블에 매핑 시켜 실시간 처리 가능하게 한다.

5.1. 아바타 이미지

아바타 이미지는 1.3메가 픽셀의 QuickCam을 이용하여 화자가 발성한 /아/, /이/, /우/, /에/, /오/, /으/, /어/ 7 단어와 아무것도 발성 하지 않은 상태의 영상을 캡쳐하였다. 영상의 크기는 320 X 240 pixel이며, 다음 그림 6은 각각의 모음에 해당하는 영상을 나타낸다.

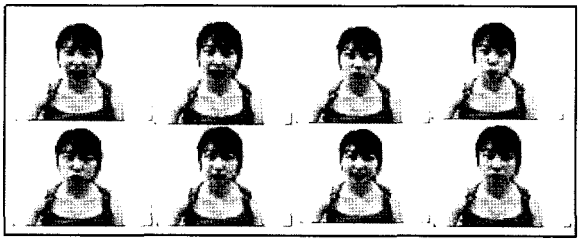


그림 6. 아바타로 사용한 /아/, /에/, /이/, /오/, /우/, /으/, /어/, /묵음/
Fig. 6 Avatar of /a/, /e/, /i/, /o/, /u/, /eu/, /eo/, /silence/.

5.2. 음성과 영상의 매핑

실시간 립싱크 시스템은 마이크로폰을 통해 입력 받은 음성을 Viseme 클래스 테이블을 통해 해당하는 입술모양으로 출력한다. 즉, 입력 신호인 음성을 출력신호인 영상으로 변환해주기 위해 44개의 음소를 Viseme 클래스 테이블을 이용하여 해당하는 8개의 Viseme 클래스로 다대일 매칭 시킨다. Viseme 클래스 테이블은 입모양이 비슷한 음소끼리 묶어서 대표되는 영상을 표현한 테이블로써, 각각의 음소인식을 하는 음성인식과 비교될 수 있다. 우선, 립싱크 시스템은 44개의 음소 클래스를 8개의 영상 클래스로 매칭 시키므로 각 음소를 인식하는 음성인식기보다 인식 성능이 높아지는 장점이 있다.

5.3. 립싱크 시스템 구현

연결어 인식기의 경우, 음성이 끝난 후, 끝점 검출을 통해 인식 결과 값을 보여준다. 이는 실시간으로 입술모양을 보여주는 립싱크 시스템 적용에 한계가 있다. 본 논문에서는 5프레임 (1/20초)마다 가능성이 가장 큰 인식 단어결과를 아바타 이미지로 보여준다. 매 프레임마다 결과 영상을 보여주면, 불안정한 확률 값으로 오 인식 되는 경우가 생기는데, 이는 영상의 떨림이나 끊김 현상을 보여준다. 좀 더 자연스러운 아바타 영상의 interpolation을 위해 본 논문에서는 확률적으로 가능성이 가장 큰 단어에 해당하는 영상을 1초에 20프레임씩 보여주게 된다. 그림 7은 Visual C++의 Dialog 기반의 MFC를 이용하여 립싱크 시스템을 구현한 것으로, 시스템의 초기화면 (왼쪽)과 인식 시작버튼을 누른 후, /에/를 발음했을 때 (오른쪽)의 동작을 캡처하였다.

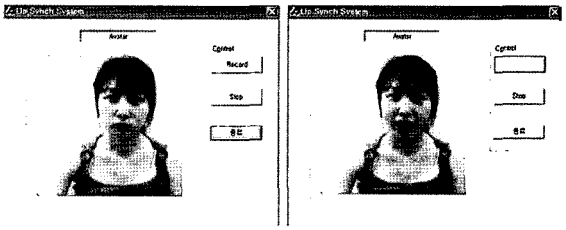


그림 7. 립싱크 초기 화면 (왼쪽)과 /에/를 발음 했을 때 (오른쪽)의 데모시스템
Fig. 7. Lip synch Demo system of initializing (left) and /e/ pronunciation image (right).

VI. 실험 결과

본 실험은 일반적으로 사용하는 8 믹스쳐 연속 HMM의 CI 모델을 기본으로 사용하여 HBT 구조의 연속 HMM, 반연속 HMM 구조의 인식성능을 각각 비교분석하였다. HBT 구조의 반연속 HMM 알고리즘은, 모델 크기를 줄이고 계산 양을 크게 감소시키는 효과가 있지만, 인식 성능이 떨어진다. 이를 보완하기 위해 예외적인 연속 HMM을 사용한 클래스 종속 반연속 HMM을 제안하였다. 실험은 연속적으로 들어오는 음성 신호에 대해 고정된 프레임마다 특징벡터를 추출하면서, 단모음의 인식 성능을 살펴본다. 모음 인식 태스크는 44개의 음소로 이루어진 고립단어 452 개를 훈련하여 사용하였다.

6.1. 실험 환경

음성데이터는 SiTEC 에서 제작한 phonetic balanced 452 단어로, 11Khz로 샘플링 된 남성 20명의 데이터베이스를 이용하였다. 훈련 15명, 인식 5명의 음성 데이터가 사용되었고, 음성 특징은 25ms의 프레임 단위로 10ms씩 이동면서 분석하였다. 특징은 로그 에너지와 1차 미분 값을 적용한 12차의 Mel Frequency Cepstral Coefficients (MFCC)를 사용하였으며 훈련은 dummy 스테이트를 제외하고 3 스테이트, 2 stream 을 사용하였다.

6.2. 실험 결과

실험은 고립단어 인식 시 사용하는 문맥 독립적 (CI)인 음소 모델과 문맥종속, 문맥 독립적인 두 가지 형태를 포함하는 HBT모델의 오인식율로 성능을 비교하였다. 각각의 HMM 모델은 3 스테이트를 가지며 각 스테이트는 3개의 전이확률을 가진다. 기본 실험으로 사용된 CI 모델은 연속 HMM과 반연속 HMM을 이용하여 성능을 측정하였고, HBT구조를 가진 연속 HMM과 반연속 HMM의 성능을 각각 비교한다. 반연속 HMM은 256개의 공통 코드워드를 가지고 있으며, 연속 HMM은 8개의 가우시안 믹스쳐를 가진다. 표2 는 기본 실험인 CI모델과 HBT 구조를 가진 연속 HMM을 비교 실험한 결과이다. HBT구조는 기본 CI모델의 인식율과 비교하여 음소 오인식 감소율은 45.28%, 문장 오인식 감소율은 38.04% 이다. 반연속 HMM의 경우, 연속 HMM보다 인식 율이 떨어졌지만 가우시안 개수는 연속 CHMM의 경우, 2,904개에서 256개로 크게 감소하였다. 표3 은 HBT구조의 연속HMM과 제안한 HBT구조의 클래스 종속 반연속 HMM 성능을 비교하여 나타내었다. 제안한 방법과 연속 HMM의 인식성능에는 거의 차이가 없었지만, 파라미터 수는 33.92% 감소되었다. 표4 는 각 모음별 오인식율을 나타낸다. HBT구조의 클래스 종속 반연속 HMM은 HBT구조가 아닌 기본 실험 연속 HMM과 비교하여, /어/, /우/, /어/에서 인식성능이 크게 향상되었다. 문맥 독립형만 가지는 기본 연속 HMM에서는 /오/, /우/, /어/의 성능이 다른 모음과 비교하여 현저히 떨어지는데, 이는 훈련 데이터의 부족에서 기인한 것이라 할 수 있다.

표 2. 기본 연속 HMM, HBT구조를 가진 연속 HMM과반연속 HMM의 음소 오인식율 (PER), 문장 오인식율 (SER) 비교
Table 2. Average phone error rate (PER) and string error rate (SER) comparison between CI model and HBT model based on the CHMM and SCHMM.

		CI model	HBT model	Error Reduction (%)
연속 HMM	PER (%)	22.95	12.53	45.28
	SER (%)	76.99	47.70	38.04
반연속 HMM	PER (%)	24.19	16.00	33.86
	SER (%)	80.39	57.78	28.12

표 3. HBT구조의 연속 HMM과 제안된 방법의 클래스 종속 HMM의 파라미터 수와 음소 오인식율 (PER), 문장 오인식율 (SER) 비교
Table 3. Average error rate and the number of parameters comparison between HBT models based on the CHMM and proposed method.

	HBT Model (연속 HMM)	클래스 종속 반연속HMM과 예외적인 연속 HMM
파라미터 수	151,008	99,776
PER	12.53	12.47
SER	47.70	48.53

표 4. Viseme 클래스 별 오인식율
Table 4. Each Viseme class error rate.

Viseme 클래스	훈련 데이터 수	CI 연속 HMM	HBT구조의 클래스 종속 반연속 HMM
		PER(%)	PER(%)
1(아)	8,460	19.3	10.5
2(에)	8,639	19.1	7.9
3(오)	5,249	33.0	18.4
4(우)	4,080	43.9	23.1
5(어)	8,219	35.3	16.5
6(으)	3,690	29.6	18.2
7(이)	7,650	17.4	9.0

VII. 결 론

본 논문에서는 제안된 HBT구조의 HMM을 이용하여 인식 율을 높이고 계산 양을 줄여 실시간으로 동작 가능한 립싱크 시스템을 구현한다. HBT 구조의 HMM은 문맥 종속형과 문맥 독립형을 결합한 모델링의 형태로 연속적인 소규모 어휘를 다룰 때 주로 사용하지만, 많은 가우시안 파라미터가 발생하여 메모리의 비효율을 가져온다. 반연속 HMM은 연속HMM이 가진 메모리와 계산상의 문제를 해결해주는 것으로, 본 논문에서는 단어의 시작 스테이트와 마지막 스테이트부분에 예외적인 연속 HMM을 사용하고 나머지 부분은 클래스 종속의 32개 가우시안을 갖는 3개의 코드북을 사용하였다. 이는 HBT구조의 연속 HMM을 사용하였을 때와 비교하여 비슷한 성능을 보여주면서, 파라미터 수는 크게 줄어들었다. 기본 연속 HMM 모델과 비교하여 HBT모델은 음소 오인식율이 45.28%, 문장 오인식율이 38.04% 감소하였고, 클래스 종속의 코드북을 사용함으로써, HBT구조의 연속

HMM보다 파라미터수가 33.92% 감소하였다. 마지막으로, 시스템의 구현단계에서는 음성의 끝점 검출을 통한 연결어 인식이 아니라, 프레임 단위로 가능성이 가장 큰 단어를 출력하여 시스템이 실시간으로 동작 가능하게 하였다. 향후, 우리는 입모양에 영향을 주는 유성 자음 ㅁ, ㅂ, ㅍ 등의 음소를 추가시켜 좀 더 자연스러운 아바타 립싱크 시스템을 구축 할 것이다.

감사의 글

본 논문은 정보 통신연구진흥원의 TT분야 해외교수초빙 지원 사업 국제공동연구' (No. C1012 0601 0005 01) 를 통해 수행된 연구결과의 일부입니다.

참 고 문 헌

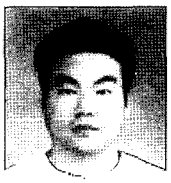
1. 이혜정, 정석태 "아바타 기반 교육용 멀티미디어 컨텐츠 저작시스템의 설계 및 구현", 한국해양정보통신학회논문지 8 (5) 1042-1049, 2004
2. F.J. Huang, T. Chen, "Real-Time Lip-Synch Face Animation Driven By Human Voice" Proc. IEEE Workshop on Multimedia Signal Processing, 352-357, 1998.
3. M. Brand, "Voice Puppetry" Proceedings of SIGGRAPH' 99, 21-28, 1999
4. T.Chen and R.Rao, "Audio-visual integration in multimodal communication", Proceedings of IEEE, Special Issue on Multimedia Signal Processing, 837-852, 1998
5. T. Kim, Y. Kang, H. Ko, "Achieving Real -Time Lip Synch via SVM-Based Phoneme Classification and Lip Shape Refinement," ICMI, Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02), 299-304, 2002
6. W. Chou, C. -H. Lee, B. -H. Huang, "Minimum Error Rate Training of Inter-Word Context-Dependent Acoustic Model Units in Speech Recognition", Proceeding ICSLP, 439-442, 1994
7. M. B. Gandhi, J. Jacob, "Natural Number Recognition using MCE Trained Inter-Word Context-Dependent Acoustic Models," Proceedings ICASSP, pp. 457-460, 1998
8. 주회열, 강선미, 고한석, "음소인식 기반의 립싱크 구현을 위한 한국어 음운학적 Viseme의 제안", 한국음향학회, 70-73, 1999
9. 신지영, "모음-자음-모음 연결에서 자음의 조음특성과 모음-모음 동시조음" 음성과학, 1226-5276, 1 55-81, 1997
10. J. R. Bellegarda, D.Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," IEEE Trans. Acoustic Speech Signal Processing, 38 2033-2045, 1990
11. X. D. Huang, "Phoneme Classification using Semi continuous hidden Markov Models" IEEE Trans. Acoustic Speech Signal Processing, 40 1062-1067, 1992

저자 약력

• 이 성 희 (Sunghee Lee)

2005년 2월: 이화여자대학교 공과대학 정보통신학과 (공학사)
2005년 3월~현재: 고려대학교 전자컴퓨터학과 석사과정 재학중
※주관심 분야: 신호처리, 음성 인식, 립싱크 시스템

• 박 준 호 (Junho Park)



2000년 2월: 고려대학교 전기전자전파 공학부 (공학사)
2002년 2월: 고려대학교 전자컴퓨터학과 (공학석사)
2007년 8월: 고려대학교 전자컴퓨터학과 (공학박사)
2007년 8월~현재: 고려대학교 전자컴퓨터학과 연구교수
※주관심 분야: 신호처리, 대어휘 연속어 음성 인식, 음향 모델링

• 고 한 석 (Hanseok Ko)

1982년 5월: 미국 카네기 멜론 대학교 전기공학 (공학사)
1986년 5월: 미국 메릴랜드 대학교 시스템 공학 (공학석사)
1988년 5월: 미국 존스 홉킨스 대학교 전기공학 (공학석사)
1992년 5월: 미국 카톨릭 대학교 전기공 (공학박사)
1995년 3월~현재: 고려대학교 전자컴퓨터공학과 교수
※주관심 분야: 영상 및 음성 신호처리, 패턴 인식, 데이터 융합