1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

음소인식 기반의 립싱크 구현을 위한 한국어 음운학적 Viseme의 제안

주희열* 강선미** 고한석*

*고려대학교 전자공학과
**서경대학교 컴퓨터과학과

Korean Phonological Viseme for Lip Synch Based on Phoneme Recognition

Heeyeol Joo* Sunmee Kang** Hanseok Ko*

*Department of Electronics Engineering, Korea University,
5ka-1 Anam-dong, Sungbuk-ku, Seoul, Korea
**Department of Computer Science, Seokyeong University,
16-1 Jeongreung-4dong, Sungbuk-ku, Seoul, Korea
E-mail: hyjoo@ispl.korea.ac.kr

요약

본 논문에서는 한국어에 대한 실시간 음소 인식을 통한 Lip Synch 구현에 필수요소인 Viseme(Visual Phoneme)을 한국어의 음운학적 접근 방법을 통해 제시하고, Lip Synch에서 입술의 모양에 결정적인 영향을 미치는 모음에 대한 모음 인식 실험 및 결과 분석을 한다. 모음인식 실험에서는 한국어 음소 51개 각각에 대해 3개의 State로 이루어진 CHMM (Continuous Hidden Makov Model)으로 모델링하고, 각각의 음소가 병렬로 연결되어진 음소네트워크를 사용한다. 입력된 음성은 12차 MFCC로 특징을 추출하고, Viterbi 알고리즘을 인식 알고리즘으로 사용했으며, 인식과정에서 Bigram 문법과 유사한 구조의 음소배열 규칙을 사용해서 인식률과 인식속도를 향상시켰다.

I. 서론

최근 정보통신 분야의 눈부신 발전과 더불어 휴먼 인터페이스에 관련된 연구가 활발히 진행되고 있다. 음성인식은 가장 편리한 휴먼 인터페이스로써 그 응용분야가 무궁무진하다. Lip Synch를 음성인식을 이용해서 구현한다면 가장 쉽고, 편리하게 구현할 수 있을 것이다. 국내의 경우에 영상 및 음성 합성 분야[1]에서 텍스트가 주어졌을 때 그 텍스트에 맞는 영상을 매칭 함으로써 입술모양이나 얼굴표정을 결정하는 Lip Synch 연구활동은 있어왔으나, 음성인식을 통한 Lip Synch 연구는 전무한 실정이다. 국외의 경우엔 음성인식과 음성합성분야에서 다양하게 Lip Synch연구 활동이 있어왔다.[2]~[6] 음성인식을 통한 Lip Synch구현은 음성인식 결과에 상당한 영향을 받게된다. 음성인식 분야는 60년대 이후부터 발전해서 현재는 대용량 음성 인식기 구현과 실용화를 위한 연구, 연속적인 음성의 인식과 자유발화를 인식하는 연구가 활발히 진행되고 있다. 본 논문의 구성은 2장에서 Lip Synch에 사용되는 Viseme을 한국어의 음운학적인 특성을 고려하여 제시하고, 3장에서 음성인식을 통한 Lip Synch에 결정적인 요소인 모음에 대한 인식실험 및 결과, 마지막으로 4장 결론에서는 향후 연구방향에 대해서 언급한다.

II. Viseme의 설정

언어의 음운학적인 특성을 고려한 영상 매칭을 위해서는 음성학이나 음운론에 입각한 입술모양, 얼굴표정이 요구되는데, 음성의 기본단위가 음소라면 그에 상응하는 영상에서의 기본단위 Viseme이 정의된다. 주로 청각장애자에게 음성을 가르칠 때 화자의 입술모양만을 보고도 그 뜻을 파악할 수 있도록 만든 것인데, 영어에 대한 Viseme은 많은 연구 그룹에 의해 만들어져 사용되고 있으나, 아직 그 표준은 정해진 것이 없다. 한국어에 대한 Viseme 역시 그 표준이 정해진 것이 없기 때문에 본 논문에서 한국어에 대한 Viseme을 한국어 음운학적인 특성을 고려해서 모음과 자음 각각에 대하여 제시하고 향후 음성인식을 통한 Lip Synch 구현에 사용할 것이다.

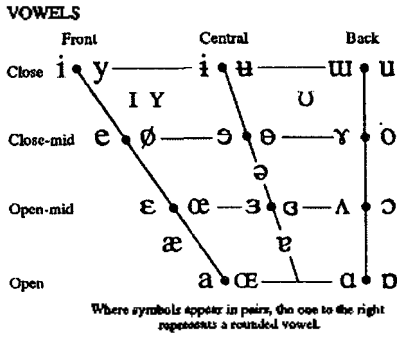

그림 1 IPA Vowel Chart

한국어에 대한 Viseme 설정을 위하여 한국어의 조음학적인 특성을 살펴보면, 모음은 후두에서 진동 또는 수축된 폐기류가 성대와 성도를 통하여 방출될 때 만들어지고, 후두에서 만들어진 음성은 성도의 인두강, 구강, 비강 등에서 변화된다. 성대의 크기와 형태는 원칙적으로 혀와 입술의 위치에 의하여 결정된다.[7]-[9] 따라서, 모음을 만드는 가장 기본적인 두 가지 조음방식은 혀의 형태와 위치, 그리고 입술의 형태와 돌출 정도이다. 그림 1은 IPA에서 제정한 Vowel Chart로써 가로축은 혀의 위치를 세로축은 입술의 형태를 정의하고 있다. 자음의 경우에는 일반적으로 모음보다 성도의 큰 수축을 보여주고, 돋들림(prominence)이 덜하다. 또한, 한국어에서 자음의 구조적 조직은 모음들이 음절의 중심 혹은 핵이 되고 자음은 음절의 가장자리가 된다. 그림 2는 IPA에서 제정한 Consonant Chart로써, 가로축은 조음위치를 세로축은 조음방식을 나타낸다. 모음과 자음의 조음위치와 방식에 대한 학계의 의견은 여러 가지가 있지만 국제 표준이 되고있는 IPA chart를 기본으로 해서 한국어에 대한 Viseme설정의 근거로 삼았다.

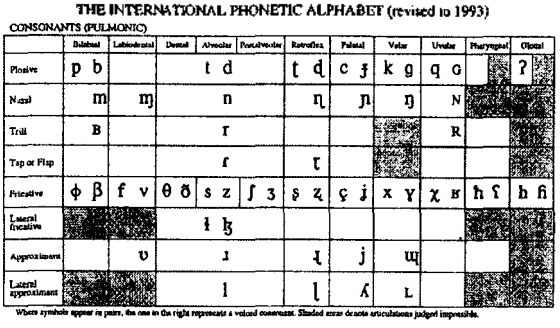

그림 2 IPA Consonant Chart

앞에서 말한 바와 같이 모음의 음색을 결정하는 중요한 두 가지 요인은 첫째, 입술의 모양(둥글고, 안 둥글음) 둘째, 혀의 높낮이라고 했다. 모음을 표기함에 있어 혀의 위치가 비교적 고정될 수 있는 소리를 미리 결정하여, 실제의 말소리를 이와 상호 대조하여 기술하는 방법을 취할 수 있는데, 이러한 모음 기술 방법을 위하여 고안된 것이 D. Jones의 기본모음(Cardinal Vowel)이다. 기본모음은 1차 기본모음 8개(i, e, ɛ, a, u, o, ɔ, ɑ)과 2차 기본모음 8개(y, ø, œ, Œ, ɯ, ɤ, ʌ, ɒ)로 구성되어지며 그림 1에서 기본모음의 조음위치를 알 수 있다. 한국어 단모음을 D. Jones의 기본모음으로 표기하면 표 1과 같고, 그림 1과 같은 모음사각도상에 표시하면 그림 3과 같이 나타낼 수 있다.

표 1 국어의 단모음 표기

단모음	표기	단모음	표기
「일, 이렇다」 이	[i]	「우리, 운다」 우	[u]
「네, 녜」 에	[e]	「그, 흐르다」 으	[ɨ]
「내, 대」 애	[ɛ]	「없다, 멀다」 어	[ə]
「강, 아래」 아	[a] or [ɑ]	「뒤, 쉬, 쥐」 위	[y]
「먹어, 업다」 어	[ʌ]	「쇠, 되」 외	[ø]
「오리, 오이」 오	[o]		

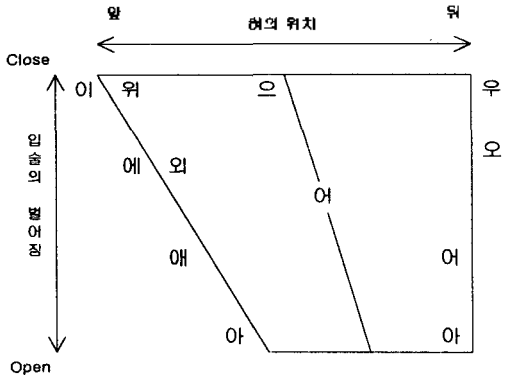

그림 3 모음 사각도상에서의 한국어 단모음

지금까지 살펴본 바와 같이 모음은 혀의 위치와 입술의 모양에 의해 그 음색이 결정됨을 알 수 있다. 따라서, 본 논문에서는 그림 3에 표시한 모음의 위치 결정 요인을 가지고 Lip Synch에 사용될 모음 Viseme을 표 2와 같이 제시한다.

표 2 한국어 모음에 대한 Viseme

모음	모음 Viseme	모음	모음 Viseme
아	/a/	야	/j/+/a/
어	/v/	여	/j/+/v/
우	/u/	유	/j/+/u/
오	/o/	요	/j/+/o/
에 , 애	/e/	예 , 애	/j/+/e/
이	/i/	와	/w/+/a/
으 , 의	/U/	워	/w/+/v/
과도음 이	/j/	위	/w/+/i/
과도음 우	/w/	웨 , 왜 , 외	/w/+/e/

표 2에서 「야, 여, 유, 요, 예, 애」 나 「와, 워, 위, 웨, 왜, 외」와 같은 소리는, 그 소리가 나는 동안에 입술이나 혀가 한 지점에서 다른 지점으로 움직이는데, 이러한 소리를 중모음이라 한다. 예를 들면, 「야」의 경우에

「이」에서 「아」로 움직인다. 이때, 혀가 「이」가까운 위치에서 다른 모음의 위치로 옮아가는 과정에서 나는 과도음(Gliding Sound)을 [j]로 표기하고, 「우」가까운 위치에서 다른 모음의 위치로 옮아가는 과정에서 나는 과도음을 [w]로 표기한다. 따라서, 이러한 과도음에 대한 Viseme을 정의하면 중모음의 Viseme 표기를 표 2와 같이 할 수 있다. 따라서, 한국어 모음은 표 2에서 보는 바와 같이 9개의 Viseme으로 mapping할 수 있다.
자음의 경우에는 두 입술이 붙어야 발음이 되는 「ㅁ, ㅂ,ㅍ」과 그렇지 않은 나머지 자음으로 분류해서 두 개의 Viseme으로 설정(표 3)했다.

표 3 한국어 자음에 대한 Viseme

자음	자음 Viseme
ㅁ, ㅂ, ㅍ	/b/
그외	/g/

III. 모음인식 실험 및 결과분석

Lip Synch구현을 위한 모음인식 실험에 사용한 음성데이터는 고려대 공대 교수님 100명의 이름을 10명의 20대 남자가 발음한 데이터 베이스를 사용해서 MFCC로 특징추출을 하고, 음소 CHMM로 학습했으며, 그림 4와 같은 네트워크를 사용해서 인식 실험을 수행했다.

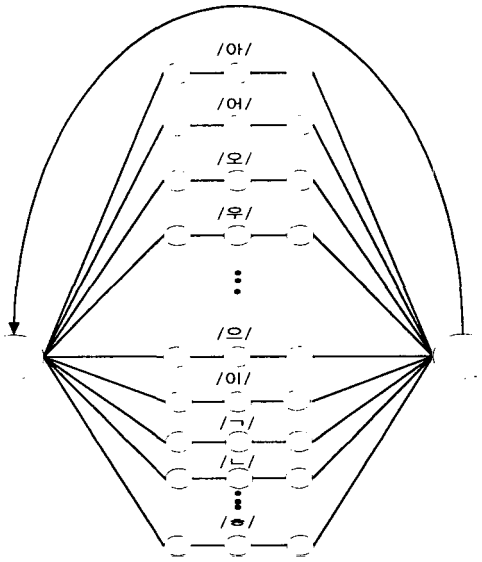

그림 4 음소 네트워크

인식 알고리즘으로 Viterbi Beam Search 알고리즘을 사용했으며, 100명의 교수님 이름에서 나타난 음소배열 규칙을 사용했다.

(1)-(4)과정은 Viterbi 알고리즘에 의해 최적의 State Sequence를 찾아내는 과정으로써, N은 HMM모델의 States 수를 나타내고, T는 관찰 값의 길이를 나타낸다.

(1)Initialization

$$\delta_1(i) = \pi_i b_i(o_1),\ 1 \le i \le N$$

$$\psi_1(i) = 0$$

(2)Recursion

$$\delta_t(j) = \max_{1 \le i \le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t),\ 2 \le t \le T\ \ 1 \le j \le N$$

$$\psi_t(j) = \arg\max_{1 \le i \le N}[\delta_{t-1}(i)a_{ij}],\ 2 \le t \le T\ \ 1 \le j \le N$$

(3)Termination

$$P^* = \max_{1 \le i \le N}[\delta_T(i)]$$

$$q_T^* = \arg\max_{1 \le i \le N}[\delta_T(i)]$$

(4)Path(state sequence)Backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, \ldots, 1$$

본 실험에서 사용된 음소배열 규칙은 언어 모델의 Bigram 문법과 유사한 구조를 가지고 있으며, 한 음소 뒤에 올 수 있는 후보 음소의 개수를 제한하여 인식과정에서 Search 공간을 줄이고, 인식률 향상을 도모했다. 그림 5가 본 실험에 사용된 음소배열 규칙이다. 첫줄에서 0은 음소번호를 !sil는 묵음을 나타내는 음소기호이고, [ ]안의 숫자는 묵음 뒤에 올 수 있는 후보음소의 개수를 나타낸다. 콜론(:) 다음의 숫자 1은 후보음소의 음소번호를 쉼표(,) 다음의 1.00은 후보음소가 올 수 있다는 뜻의 확률 값 1.00 이다.
즉, 첫줄에서 0번 음소 묵음(!sil)은 그 뒤에 28개의 후보음소가 올 수 있는데, 그 후보음소의 번호는 1, 2, 3, 4, 5, 6, ... ,28 이라는 뜻이다.

```
0 !sil+[ 28 ]: 1,1.00; 2,1.00; 3,1.00; 4,1.00; 5,1.00; 6,1.00;
          7,1.00 ; 8,1.00; 9,1.00; 10,1.00; 11,1.00; 12,1.00;
         13,1.00; 14,1.00; 15,1.00;16,1.00;17,1.00;18,1.00;
         19,1.00;20,1.00; 21,1.00; 22,1.00; 23,1.00; 24,1.00;
         25,1.00; 26,1.00; 27,1.00; 28,1.00;
1 g+[ 4 ]: 18,1.00; 20,1.00; 22,1.00; 29,1.00;
2 gv+[ 2 ]: 25,1.00; 29,1.00;
3 gs+[ 3 ]: 13,1.00; 17,1.00; 29,1.00;
4 n+[ 2 ]: 21,1.00; 29,1.00;
5 ns+[ 3 ]: 11,1.00; 16,1.00; 29,1.00;
6 d+ [ 2 ]: 19,1.00; 29,1.00;
7 r+ [ 2 ]: 22,1.00; 29,1.00;
8 l+ [ 2 ]: 11,1.00; 29,1.00;
9 ms+ [ 3 ]: 6,1.00; 11,1.00; 29,1.00;
10 b+ [ 2 ]: 18,1.00; 29,1.00;
11 s+ [ 5 ]: 18,1.00; 19,1.00; 21,1.00; 22,1.00; 29,1.00;
12 ng+ [ 6 ]: 11,1.00;14,1.00;15 ,1.00;16,1.00;21,1.00; 29,1.00;
13 j+ [ 2 ]: 19,1.00; 29,1.00;
14 jv+ [ 2 ]: 23,1.00; 29,1.00;
15 c+ [ 3 ]: 18,1.00; 19,1.00; 29,1.00;
16 h+ [ 6 ]: 18,1.00; 20,1.00; 21,1.00; 24,1.00; 26,1.00; 29,1.00
17 C+ [ 2 ]: 22,1.00; 29,1.00;
18 a+ [ 5 ]: 2,1.00; 3,1.00; 5,1.00; 12,1.00; 29,1.00;
19 v+ [ 5 ]: 3,1.00; 7,1.00; 8,1.00; 12,1.00; 29,1.00;
20 o+ [ 3 ]: 11,1.00; 16,1.00; 29,1.00;
21 u+ [ 3 ]: 5,1.00; 27,1.00; 29,1.00;
22 i+ [ 4 ]: 4,1.00; 5,1.00; 9,1.00; 29,1.00;
23 e+ [ 1 ]: 29,1.00;
24 yv+ [ 2 ]: 5,1.00; 29,1.00;
25 yu+ [ 2 ]: 5,1.00; 29,1.00;
26 wa+ [ 2 ]: 12,1.00; 29,1.00;
27 wv+ [ 2 ]: 5,1.00; 29,1.00;
28 silb+ [ 7 ]: 1,1.00; 10,1.00; 11,1.00; 13,1.00; 15,1.00; 16,1.0
29 sile+[ 0 ];
30 sil!+ [ 0 ];
```

그림 5 실험에 사용된 음소배열 규칙

그림 6은 실험결과로써 학습에 참가하지 않은 사람이 발성한 10명의 교수님 이름을 인식한 결과이다.

```
0 (98)
7 : g a ng c v r i (강철희)
!sil(0) silb(9) a(15) gs(9) j(14) v(15) r(6) i(21)

1 (98)
7 : g o s v ng jv e (고성제)
!sil(0) silb(5) h(7) o(6) h(4) o(5) s(10) v(8) ng(20) jv(16) e(13)

2 (97)
8 : g o h a ns s v gs (고한석)
!sil(0) silb(9) h(15) o(11) h(11) a(7) ns(12) s(15) v(8)

3 (110)
9 : g i ms d v gs C i ns (김덕진)
!sil(0) silb(4) g(6) i(4) ms(21) d(8) v(4) gs(23) C(8) i(16) ns(11)

4 (116)
7 : g i ms s u wv ns (김수현)
!sil(0) silb(2) g(12) i(4) ms(13) s(17) u(32) wv(20) ns(16)

5 (111)
8 : b a gs j v ng h o (박정호)
!sil(0) b(10) a(5) gs(18) j(12) v(10) ng(17) u(4) ns(10) h(8) o(17)

6 (110)
7 : b a gs C i n u (박진우)
!sil(0) silb(2) b(6) a(20) ns(5) s(10) i(13) ns(17) s(11) u(26)

7 (116)
8 : a ns s u ns s i ns (안순산)
!sil(0) a(13) a(10) ng(11) s(16) v(7) l(16) s(16) i(12) ms(14)

8 (82)
9 : j v ng c a ng s v ng (정창성)
!sil(0) silb(2) j(6) v(5) ng(10) s(10) v(8) ng(11) s(9) v(7) ng(16)

9 (64)
8 : c a gv yu ns h yv ns (차균현)
!sil(0) c(14) a(11) gv(11) yu(8) ns(11) h(4) yv(13) ns(13)
```

그림 6 인식 결과

그림 6의 인식결과를 보면 비교적 모음부분의 인식은 정확한 반면 자음의 인식에 있어서 에러가 나는 것을 볼 수 있는데, 이것은 자음의 지속시간이 모음의 지속시간보다 훨씬 짧기 때문이다.[10] 인식된 음소는 음소의 지속시간에 대한 정보를 가지고 있기 때문에 표 2와 표 3에 설정한 Viseme으로 mapping해서 Lip Synch를 할 수 있게된다.

예를 들면, 그림 6에서 첫 번째 인식 결과인

7 : g a ng c v r i (강철희)
!sil(0) silb(9) a(15) gs(9) j(14) v(15) r(6) i(21)

에서 a(15)는 음소 [a]가 15개의 프레임 동안 지속된다는 것을 나타낸다. 따라서, Lip Synch를 구현할 때 음소 [a]에 해당되는 Viseme /a/를 15프레임에 해당하는 지속 시간 동안 출력해준다.

IV. 결론

본 논문에서는 Lip Synch 구현을 위한 Viseme을 한국어 음운학에 근거해서 제시함으로써 음소 인식 기반의 Lip Synch 구현을 위한 기초적인 발판을 마련하였다. 연속분포 HMM기반의 음소인식기를 설계하고, 입술 모양에 결정적인 영향을 미치는 한국어 모음에 대한 인식 실험을 통해 자음에 비해 모음이 비교적 정확한 인식이 되는 것을 볼 수 있었다.

그러나, 자음에서 발생되는 인식 에러는 Lip Synch에 치명적인 영향을 미치므로 이 에러를 줄이기 위한 노력이 필요하다. 향후에는 음소 인식 성능의 향상을 위하여 음향 모델링, 음소 배열 규칙, 인식 알고리즘 등을 개선하기 위한 연구를 진행한다. 또한, 입력되는 음성을 자음과 모음으로 구분해서 정확한 음소의 위치를 파악할 수 있게 해주는 알고리즘을 전처리 부분에 추가하는 방법도 고려하고자 한다.

참고문헌

[1] 송경준, 이기영, 최창식, 민병의, "표정 짓고 말하는 가상얼굴의 실시간 합성," 한국음향학회지 제 17권 제 8호, pp.3-11, 1998

[2] Tsuhan Chen, Hans Peter Graf, Kuansan Wang, "Lip Synchronization Using Speech-Assisted Video Processing," in IEEE Signal Processing Letters. Vol.2 No.4 April. 1995

[3] Koster, B. Rodman, R. and Bitzer, D. "Automated Lip-Sync: Direct Translation of Speech-Sound to Mouth Shape," Proceedings of the 28th Annual Asilomar Conference on Signals, Systems and Computers, IEEE:1994, pp.36-46

[4] Koster, Barret E. Automatic Lip-Sync: Direct Translation of Speech-Sound to Mouth-Animation. Ph.D. Dissertation. Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, 1995

[5] Matthew Brand and Ken Shan, "Voice-driven animation," Mitsubishi Electric Information Technology Center America, 1998

[6] F. Parke. "A model for human faces that allows speech synchronized animation," Journal of Computers and Graphics, 1(1):1-4, 1975

[7] 허웅, "국어 음운학," 정음사, 1982

[8] 梅田博之, "한국어의 음성학적 연구," 형설출판사, 1982

[9] 고도흥, 구희산, 김기호, 양병곤 공역,"음성언어의 이해," 한신문화사, 1995

[10] 김범국, 정현열,"가변장 음소모델을 이용한 음소인식," 한국음향학회지 제 16권 제8호, pp.112-118, 1997