한국음향학회지 제20권 제3호 pp. 63~69 (2001)

시청각 코퍼스 기반의 립싱크 알고리듬 개발

Development of a Lipsync Algorithm Based on Audio-visual Corpus

김 진 영*, 하 영 민**, 이 화 숙***
(Jin Young Kim*, Young Min Ha*, Hwa Sook Lee**)

*전남대학교 공과대학 전자공학과, **전남대학교 공과대학 정보통신협동과정, ***광주여자대학교 예술디자인학부

(접수일자: 2001년 1월 18일; 채택일자: 2001년 4월 9일)

본 논문에서는 자연스러운 얼굴 합성을 위한 코퍼스 기반의 립싱크 알고리듬을 제안한다. 립싱크 알고리듬을 개발하기 위하여 여성 아나운서의 시청각 코퍼스를 구축하였다. 코퍼스 구축시, 입술파라미터 추출하기 위하여 여성화자의 얼굴에 스티커를 붙이고, 이의 위치를 영상처리기법에 의하여 얻었다. 그리고 길이, 세기 그리고 피치의 운율정보를 얻기 위하여 음성을 HTK (hidden Markov tool kit)를 사용하여 레이블 하였다. 립싱크의 기본단위로는 자음-모음-자음의 음절단위를 사용하였는데, 구축된 시청각 코퍼스는 입술의 정보 그리고 음운론적, 운율적 정보를 포함하는 음절들로 구성된다. 입술합성시에는 입력된 텍스트로부터 음절의 열을 만들고 각 음절에 적절한 대표들을 코퍼스로부터 N개씩 선정후, 최적의 열은 비터비탐색을 통하여 얻었다. 이를 위하여 음운론적 거리와 운율거리 함수가 정하였다. 컴퓨터 모의실험결과 제안된 알고리듬이 좋은 성능을 보임을 확인할 수 있었으며, 특히 립싱크에서는 길이정보뿐 아니라 길이와 피치의 정보도 유용함을 밝혔다.

핵심용어: 립싱크, 시청각 코퍼스
투고분야: 음성처리 분야 (2.4)

A corpus-based lip sync algorithm for synthesizing natural face animation is proposed in this paper. To get the lip parameters, some marks were attached some marks to the speaker's face, and the marks' positions were extracted with some image processing methods. Also, the spoken utterances were labeled with HTK and prosodic information (duration, pitch and intensity) were analyzed. An audio-visual corpus was constructed by combining the speech and image information. The basic unit used in our approach is syllable unit. Based on this Audio-visual corpus, lip information represented by mark's positions was synthesized. That is, the best syllable units are selected from the audio-visual corpus and each visual information of selected syllable units are concatenated. There are two processes to obtain the best units. One is to select the N-best candidates for each syllable. The other is to select the best smooth unit sequences, which is done by Viterbi decoding algorithm. For these process, the two distance proposed between syllable units. They are a phonetic environment distance measure and a prosody distance measure. Computer simulation results showed that our proposed algorithm had good performances. Especially, it was shown that pitch and intensity information is also important as like duration information in lip sync.

Keywords: *Lipsync, Corpus*
Ask subject classification: *Speech signal processing (2.4)*

책임저자: 김진영 (kimjin@dsp.chonnam.ac.kr)
500-757 광주광역시 북구 용봉동 300번지
전남대학교 전자공학과
(전화: 062-530-1757; 팩스: 062-530-1759)

I. 서 론

인간의 의사소통에 있어서 영상정보가 커다란 역할을 한다는 것은 널리 알려진 사실이다[1,2]. 특히 입술 영상은 대부분의 음성정보를 담고 있기 때문에 입술 영상과 음성을 동시에 전달하는 것은 의사소통에 큰 도움이 된다. 이러한 영상의 합성법으로는 해당되는 음성에 따른 비디오 프레임을 연결하거나 각각의 이미지를 연결하는 키-프레임 (key-frame) 방법[3], 영상 변수를 사용해서 입력된 텍스트나 음성으로부터 영상을 합성하는 변수 기반의 내삽 (interpolatiopn) 방법[4], 해부적 구조에 기반한 방법[5], 물리적 성질에 기반한 방법[6] 등이 있고, 합성을 위한 입력으로는 일반적으로 음성, 텍스트, 그리고 음성과 텍스트를 주로 사용한다.

본 논문에서는 녹음된 음성 DB를 기반으로 립싱크를 합성하는 알고리즘을 제안하고 영상과 음성의 동기화와 자연스러운 발화의 문제에 주안점을 두었다.

II. 시청각 코퍼스 구축

2.1. 데이터 수집

이 논문에서는 표준말을 사용하는 여성 아나운서가 텍스트를 보통 속도로 발음하는 정면영상을 디지털 카메라를 사용하여 저장하였고 녹화시 화자의 얼굴에 녹색의 원형 형광 스티커를 부착하였는데 이는 이미지의 분석을 용이하고 정확하게 한다. 약 2시간 분량의 음성데이터와 영상데이터를 8kHz와 초당 30프레임씩 저장하였다.

2.2. 영상처리

이 논문에서는 발화에 따른 화자의 얼굴 움직임 정보를 얻기 위해 얼굴에 형광 스티커를 부착하여 그 위치정보로 얼굴의 움직임을 제어하는 방법[7]을 사용하였다. 형광 스티커는 총 19개를 사용하였으며 부착위치는 그림 1과 같다. 본 논문에서는 형광의 스티커를 부착했기 때문에 밝기와 색상정보 두 가지를 사용해서 각 표시의 위치를 추적하였다. 영상은 음성에 비해 상대적으로 데이터가 크기 때문에 처리하는 시간이 길어지게 된다. 본 논문에서는 주성분 분석 (principal component analysis)을 이용하여 총 96%의 정보를 보유하는 상위 8개의 벡터를 선택하였다. 선택된 벡터는 합성시 원래의 값으로 복원과정을 거치게 된다. 그림 2는 누적된 고유치의 분포도이다.

그림 1. 얼굴 영상 샘플
Fig. 1. Sample of face image.

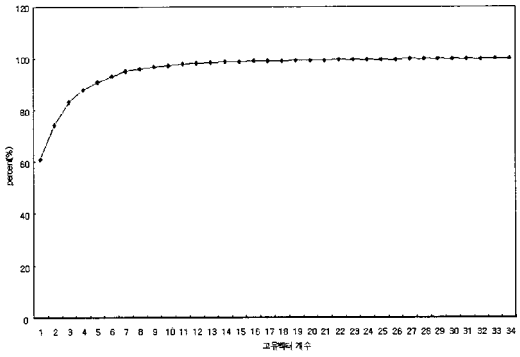


그림 2. 고유벡터 수에 따른 데이터 비율
Fig. 2. Data reduction rate with number of eigenvector.

2.3. 음성처리

이 논문에서는 입술 데이터가 발화된 음소와 관련이 있을 뿐만 아니라 음성의 다른 특징들, 피치, 길이정보 그리고 세기와 같은 기타 운율정보들과도 관련이 있다는 가정에 근거하여 이들의 상관관계를 분석하였다. 상관관계의 분석을 위해 각 음성의 강도 일반화를 수행하였고, 피치 정보를 얻기 위해 성대의 떨림 신호를 잡아주는 라링고그래프 (laryngograph)를 이용하였으며 각 음소의 길이정보는 HTK (HMM Toolkit)라는 툴에 의해 획득된 길이정보를 수정하여 사용하였다.

2.4. 운율과 영상정보의 관련성 분석

이 장에서는 음운정보 외에도 운율정보 즉 피치나 음성의 크기, 길이 정보와 영상정보간에 관련이 있는가에 대해 살펴본다. 이는 일반적으로 동일한 음소를 발음할 때라도 목소리나, 피치, 그리고 길이정보에 따라 영상이 변화한다는 가정하에 시행된다.

본 논문에서 상관관계를 분석하기 위해 선택한 변수는 각 음소의 운율정보 (강도, 피치, 지속시간)와 입술영역의 높이와 폭이다. 이 영상변수는 얼굴에서 발화에 따라 변화

표 1. 상관관계 분석 결과
Table 1. Correlation analysis result.

음소	발생빈도	넓이			넓이/높이	높이		
		길이	강도	피치		길이	강도	피치
a	2836	-0.0414	0.12729	0.00026	-0.4281	-0.0083	-0.0340	-0.0319
i	1687	0.03672	0.11797	0.08285	0.43753	0.24967	0.10442	0.01314
o	1169	0.03818	0.12850	0.04873	0.53687	0.03916	0.05083	0.05636
u	620	0.04872	0.11739	0.08946	0.47619	0.06904	0.00114	0.03165
eo	1464	0.03020	0.10355	0.07705	0.31623	0.15253	0.05850	0.08719
eu	1878	0.07293	0.13882	0.04454	0.35663	0.14258	0.02508	0.01996
e	614	-0.0801	0.20410	0.11406	0.31438	0.30258	0.08027	0.00907
ya	109	-0.3312	0.24105	0.28933	0.07811	0.27736	0.13834	-0.2283
yeo	408	0.02075	0.07174	-0.1018	0.23813	0.29040	0.04911	-0.0397
yo	52	-0.1014	0.25523	0.34717	0.41753	0.19101	0.10902	0.19080
yu	31	-0.0417	0.18472	0.19663	0.56542	-0.0890	0.05003	0.21970
ye	56	0.20317	0.23010	0.0064	0.44405	0.28430	0.30483	0.05243
wa	185	-0.2111	-0.0232	0.05558	-0.0497	0.45926	0.18954	-0.0767
wi	64	0.24907	0.27403	0.1558	0.43126	0.29636	0.36383	-0.1517
wo	99	-0.0204	0.3208	0.18594	0.40731	0.29575	0.17872	0.11730
waelp	10	0.00993	0.2782	-0.1239	0.39357	0.48570	-0.1154	-0.5179

하지 않는 미간과 코끝간의 거리로 일반화되었다.

$$w = w_0 / ref \quad (1)$$

$$h = h_0 / ref \quad (2)$$

ref : 미간과 코끝간의 거리

그리고 각 음소가 지속되는 동안 운율정보가 변화하기 때문에 지속시간 내에서 각 운율값의 최대값과 평균값을 취해서 상관관계 분석에 사용하였는데 평균값보다는 최대값을 사용했을 때 상관관계가 더 명확히 드러났다. 표 1에서는 시각소 그룹들의 모음에 해당하는 음소의 음성정보와 입술영상에서 폭과 높이의 상관관계를 나타내었다.

III. 입술 파라미터 합성을 위한 음절 단위

본 논문에서는 영상변수의 합성단위로 음소가 아니라 시각소를 기반으로 하는 음절 (CVC: consonant-vowel-consonant)을 사용하였다. 앞과 뒤의 음소의 영향을 [VC+CVC+CV]로 고려하여 최적음절을 결정할 때 사용하도록 하였는데 이는 발음에 따른 입술영상을 관측해볼 때 앞뒤의 모음이 입모양에 영향을 많이 끼치기 때문이다 [8]. 그러므로 자연스러운 연음현상의 구현을 위해서는 CVC 코퍼스뿐만 아니라 그 주변의 음소환경도 고려해야 한다. 시각소의 정의는 표 2에서와 같이 같은 입 모양으로 구분되는 음소들을 하나의 그룹으로 묶었다. 표 3은 임의의 텍스트입력에 대한 합성의 CVC 단위들을 보이고 있다.

표 2. 시각소 테이블
Table 2. Viseme table.

분류	음소
초성	ㄱ(ㄱ,ㅋ,ㅎ), ㄴ(ㄴ,ㄹ), ㄷ(ㄷ,ㅌ,ㅅ,ㅈ,ㅊ), ㅁ(ㅁ,ㅂ,ㅍ)
중성	아, 어, 이, 오, 우, 에, 으
종성	ㄱ(ㄱ,ㅇ), ㄴ(ㄴ,ㄷ,ㄹ), ㅁ(ㅁ,ㅂ)

표 3. 합성단위 그룹핑
Table 3. Unit grouping.

입력 : 그때를 생각하면

sil (g_eu_d) e r eu n d e n g a k a m eo n
sil g eu (d_e_r) eu n d e n g a k a m eo n
sil g eu d e (r_eu_n) d e n g a k a m eo n
sil g eu d e r eu n (d_e_n) g a k a m eo n
sil g eu d e r eu n d e n (g_a_k) a m eo n
sil g eu d e r eu n d e n g a (k_a_m) eo n
sil g eu d e r eu n d e n g a k a (m_eo_n)

IV. 입술의 합성 및 평가

입술모양 합성시 일반적으로 선택하는 변수 단위는 독립적인 음소이다. 이렇게 각 음소들에 해당되는 영상 변수를 얻고 그 변수들을 이용해서 실제로 사용할 영상 변수를 구성하는 방법은 데이터베이스 구축에도 용이하고 합성에도 용이하지만 실제 발음시에는 각 음소가 독립적으로 발음되는 것이 아니라 앞에 오는 음소 (forward coarticulation)와 뒤에 오는 음소 (backward coarticulation)의 영향을 받아 동일한 음소일지라도 다른 모양으로 나타날 수 있기 때문에 단순히 각 음소를 조합하는 방법으로는

표 4. 탐색 테이블의 구조
Table 4. Structure of lookup table.

음절정보 i :
중심음절
시각소1[C] 피치, 강도, 지속시간
시각소2[V] 피치, 강도, 지속시간
시각소3[C] 피치, 강도, 지속시간
영상 프레임수
주성분분석 변수들
.....
선행음소들
시각소1[V] 피치, 강도, 지속시간
시각소2[C] 피치, 강도, 지속시간
영상 프레임수
주성분분석 변수들
.....
후행음소들 수
시각소1[C] 피치, 강도, 지속시간
시각소2[V] 피치, 강도, 지속시간
영상 프레임수
주성분분석 변수들
.....
반복 .....

자연스러움이 떨어지게 된다. 이를 보완하기 위한 방법으로 연음현상의 모델링에 대한 연구가 수행되고 있는데 크게 미리보기 모델 (look-ahead model)[9], 시간잠금모델 (time-locked model)[10], 혼성모델 (hybrid model)[11]과 지수함수모델 (exponential model)[12]의 네 가지가 있다. 그런데 본 논문에서는 코퍼스 (corpus) 기반의 방법을 제시한다.

4.1. 시청각 데이터베이스 구축 및 선택 알고리듬

합성하고자 하는 텍스트에 따른 영상변수를 선택하기 위해 코퍼스와 탐색테이블내의 패턴과의 거리를 계산하여 최적-N개의 합성 단위를 선정한다. 탐색테이블의 구조는 표 4와 같다. 선정된 합성 단위에 Viterbi 탐색을 실시하여 최적의 경로를 선택한다. 선택한 경로에 해당하는 CVC를 추출한다. 과정은 다음과 같다.

1. 입력된 음소정보를 코퍼스 그룹으로 변환한다.
2. 탐색테이블에서 중심의 기준 코퍼스[CVC]에 해당되는 패턴을 찾는다.
3. 이 패턴들에 대해서 운율과 음소상의 거리 (D_{phon}, D_{pros})를 계산하여 최적-N개의 후보를 선정한다.
4. 선정된 후보들에 대하여 비터비 탐색을 실시한다.

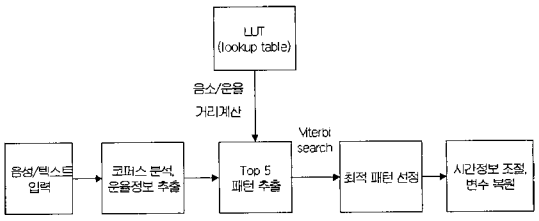


그림 3. 코퍼스의 선택과 합성과정
Fig. 3. Corpus selecting and synthetic processing.

4.2. 운율과 음소거리 계산

중심코퍼스 CVC를 기준으로 운율과 음소정보, 그리고 주성분분석처리된 영상변수들과 이를 중심으로 한 앞쪽과 뒤쪽의 코퍼스에 해당되는 정보들이 배열되어 있다. 이렇게 하나의 CVC에 대한 여러 개의 정보들이 해당 CVC_탐색테이블내에 반복되어 저장되어서 변수의 합성시 각 패턴들에 대한 거리를 계산할 수 있다.

먼저 운율거리 (D_{pros})는 해당 코퍼스그룹 내에서 중심코퍼스를 구성하는 시각소들의 피치, 강도 그리고 지속시간 상의 거리를 계산하고, 각각의 거리에 개별적인 가중치를 주어서 합친 후 모두 더한 값이다.

$$D_{pros} = \sum_i D_{pros_i} \tag{3}$$

$$D_{pros_i} = \lambda_1 D_p + \lambda_2 D_i + \lambda_3 D_d \tag{4}$$

그리고 $D_x (x = p, i, d)$는 기준패턴과 입력 코퍼스의 피치, 강도, 지속시간상의 Mahalanobis 거리이다.

$$D_x = \left| \left(\frac{x_x - x_{ref}}{\sigma_x} \right)^2 \right| \tag{5}$$

이 가중치들은 2장에서 얻은 각 시각소와 운율정보와의 관련성 분석 결과를 전체에서의 각각의 비율로 나타낸 값 ($\lambda_1, \lambda_2 \lambda_3$,)으로 다음과 같은 식에 의해서 구한다.

$$\lambda_1 = \frac{\rho_{p,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \tag{6}$$

$$\lambda_2 = \frac{\rho_{i,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \tag{7}$$

$$\lambda_3 = \frac{\rho_{d,v}}{|\rho_{p,v}| + |\rho_{i,v}| + |\rho_{d,v}|} \tag{8}$$

$\rho_{p,v}$, $\rho_{i,v}$, $\rho_{d,v}$는 각각 음소와 피치, 세기, 길이의 상관관계 값이다.

음소거리 (D_{phon})는 코퍼스그룹의 중심코퍼스를 제외한 앞과 뒤의 코퍼스와 기준패턴과의 음소상의 거리인데,

해당 코퍼스 내의 모든 모음과 자음에 대해 거리 (D_{vowel}, D_{cons})를 구한 후 모두 더한 값이다.

$$D_{phon} = \frac{(D_{cons} + D_{vowel})}{2} \quad (9)$$

4.3. 가중치의 결정

이 장에서는 최적의 패턴을 선택하기 위해 사용되는 음성과 운율거리를 결합하기 위한 가중치의 결정과정에 대해 설명한다. 식 (10)은 가중치를 결정하는 식으로 D_{phon}과 D_{pros}는 운율과 음소 거리값을 나타내고 λ는 가중치를 나타낸다.

$$D = \lambda D_{phon} + (1-\lambda) D_{pros} \quad (10)$$

세 가지의 운율정보를 모두 사용하는 것이 타당함을 보이기 위해 세 가지의 운율정보 (피치, 강도, 지속시간)를 모두 사용할 때와 지속시간만을 사용하는 두 가지 경우에 대해서 실험을 수행하였다. 임의의 파일 50개를 선택하여 0.0부터 1.0까지 가중치를 0.1씩 증가시켜 앞에서 설명한 변수합성을 위한 과정들을 수행한다.

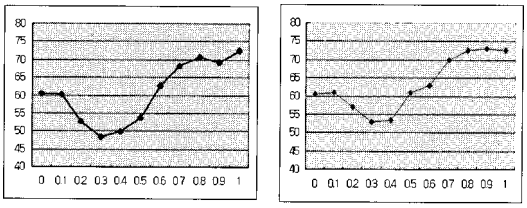
첫 번째의 실험 수행시에는 앞에서 설명한 방법을 그대로 사용하였고, 두 번째의 경우에는 운율거리 계산시 지속시간의 가중치를 1로 하고 나머지는 0으로 한 채 실험을 수행하여 값을 비교하였다. 아래의 식은 에러를 정의한 식이다.

$$e = \frac{1}{N} \sum_{0}^{N} (p_{ref} - p_{syn})^2 \quad (11)$$

여기에서 p_{ref}와 p_{syn}는 실제변수와 합성된 변수값으로 x와 y값이다.

그림 4에서는 각각의 가중치에 대해 세 가지 운율정보를 사용한 경우와 지속시간 정보 한가지만을 사용한 경우를 비교하였다. 여기에서 세 가지 운율정보를 모두 사용한 경우의 합성결과가 더 우수함을 알 수 있다. 다음 장에서 설명하겠지만 여기에서 사용된 변수값들은 비터비 탐색과 복원 과정을 거친 값들이다. 비터비 탐색에 대해서는 다음 절에서 설명한다.

그림 4에서 x축의 값은 음소정보에의 가중치이고 y축은 상대적인 오차값을 나타내고 있다. 여기에서 볼 수 있듯이 음소정보에 가중치 0.3을, 운율정보에 가중치 0.7을 줄 때 오차가 가장 적고 가중치 0.0, 1.0일 오차에 비하여 각각 상대적으로 15%와 27%씩 우수한 성능을 보인다.



(a) 피치, 강도, 지속시간 (b) 지속시간

그림 4. 가중치에 따른 오차율 비교
(x-축 : 가중값, y-축 : 오차)
Fig. 4. Comparison with different weight results.

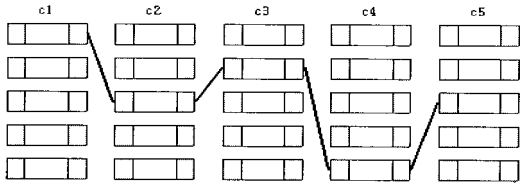


그림 5. 비터비 탐색
Fig. 5. Viterbi search.

4.4. 비터비 탐색

이 논문에서는 연속되는 코퍼스들이 자연스럽게 연결되도록 비터비 탐색 (Viterbi Search)을 사용하여 연음현상을 구현한다.

비터비 탐색은 그림 5와 같이 먼저 위에서 계산한 코퍼스그룹의 거리 D를 기준으로 하여 거리상으로 가장 가까운 패턴 다섯 개를 선택한다. 그리고, 이 후보 패턴들과 다음에 연결될 패턴과의 거리를 계산한다. 여기에서 거리계산의 기준은 각 패턴이 연결되는 부분, 즉 시작부분과 끝부분의 영상변수 (회색으로 표시된 부분)로서 식 (12)에 의해 거리값을 계산하고 선택된 패턴에 대해서 시간정보를 조절한 후 고유치 벡터의 역행렬을 이용해 변수를 복원한다. Viterbi search에 사용되는 패턴의 정보는 표 5와 같다.

표 5. 패턴의 구성
Table 5. Structure of pattern.

앞쪽 코퍼스의 수 시각소1[V] 피치, 강도, 지속시간 시각소2[C] 피치, 강도, 지속시간 영상 프레임 수 주성분분석 변수들
뒤쪽 코퍼스의 수 시각소1[C] 피치, 강도, 지속시간 시각소2[V] 피치, 강도, 지속시간 영상 프레임 수 주성분분석 변수들

$$D_{ij} = \frac{1}{P} \sum_{k=0}^{P-1} (y_{ik} - y_{jk})^2 \quad (12)$$

D_{ij} : 상태 i와 상태e j의 거리

P : 주요소의 수

y_{ik}, y_{jk} : 상태 i와 상태 j의 k번째 요소

4.5. 입술 파라미터의 복원

위의 과정을 거쳐서 선택된 경로의 영상변수는 주성분 분석을 이용하여 추출된 고유치벡터 (eigenvalue vector)로서 실제로 적용하기 위해서는 식 (13)을 이용하여 원래의 데이터 (얼굴에서 각 포인트의 위치정보)로 복원해야 하고 합성대상이 되는 실제의 코퍼스 지속시간과 일치하도록 입력데이터를 기준으로 각 시각소의 시간정보를 수정해야 한다. 다음은 원래의 입술정보로 복원하는 식이다.

$$X' = E' * Y \quad (13)$$

단,

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_{36} \end{bmatrix} = \begin{bmatrix} e_{1,1} .. e_{1,36} \\ e_{2,1} .. e_{2,36} \\ \cdots \\ e_{36,1} .. e_{36,36} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_{36} \end{bmatrix}$$

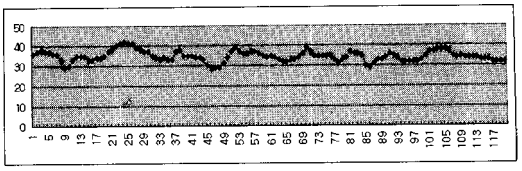
여기서 E'는 36*36의 고유벡터 역행렬이고 Y는 8개의 PCA값과 24개의 0으로 구성된 36*1행렬이다. 이 연산의 결과로 얻어진 X'는 36*1행렬로서 실제 합성에 사용할 얼굴의 스티커 위치를 나타내는 좌표값이다.

복원과정을 거친 후 실제로 변수를 합성하기 위해서는 입력된 음성데이터와 각 음소의 지속시간을 동일하게 조절해야 한다. 실제 데이터의 시간정보를 기준으로 하여 영상변수의 내삽 과정을 거쳐서 시간정보를 조절한다.

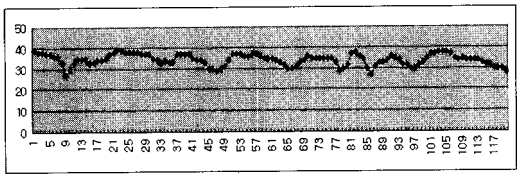
4.6. 합성결과 예제

위의 과정을 거쳐 합성된 결과를 보이기 위해 한 문장을 선택하여 언어정보를 가장 많이 포함하고 있는 입술의 높이에 대해 변화를 비교하였고, 임의의 단어에 대한 얼굴변수를 보였다. 그림 6의 (a)와 (b)에서 x축의 값은 이미지 프레임수이고 y축은 복원된 입술의 높이와 폭이다.

그림 6의 (a)에서 볼 수 있듯이 발화시 입술의 폭보다는 높이의 변화가 뚜렷한 것을 볼 수 있는데 이는 모음의 발화시에 상대적으로 폭보다는 높이의 변화가 크고 자음 시각소 그룹에서 볼 때 ㅁ(ㅁ, ㅂ, ㅍ)그룹을 제외한 다른 시각소들은 그 자체보다는 모음의 영향을 크게 받기 때문이다.



(a) 원래의 변수 (높이)



(b) 합성된 변수 (높이)

그림 6. 합성 결과 비교
Fig. 6. Comparison of synthetic result.

V. 결 론

이 논문에서 영상정보는 음소뿐만 아니라 운율과도 상관관계가 있음을 보였고, 이 결과가 영상합성기에서 효과적으로 쓰일 수 있음을 보였으며 코퍼스를 기반으로 연음현상의 구현을 좀더 용이하게 하였다. 이 결과는 기타의 영상합성기에서 적용할 수 있으며 3차원 영상합성기에서도 유용할 것으로 보인다. 그리고 이 합성기는 음소를 기준으로 하여 구성되기 때문에 보다 나은 합성결과를 얻기 위해서는 정확한 레이블링이 필요하고, 정확한 영상정보 추출방법의 개발, 가장 유사한 패턴을 선택하기 위한 음소, 운율상의 거리 계산법, 그리고 그 가중치의 최적화 등의 개발이 요구된다.

이 논문에서는 화자의 얼굴에 표시를 하는 방법으로 영상정보를 획득하였다. 합성을 위해 코퍼스를 선택할 때 그 주변의 음소를 고려하였으며 합성시 피치, 강도, 지속시간의 세 가지 운율정보를 사용함으로써 지속시간만을 사용하는 경우나 음운정보만을 사용하는 경우보다 더 나은 합성결과를 얻을 수 있었다.

감사의 글

본 논문은 2000년도 학술진흥재단의 선도연구자과제의 연구 결과물 중 하나입니다.

참 고 문 헌

1. W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise" , *Journal of the Acoustical Society of America*, vol. **26**, pp. 212-215, 1954.

2. Q. Summerfield, A. MacLeod, M. McGrath and M. Brooke, *Handbook of Research on Face Processing*, Elsevier Science Publishers, 1989.
3. Tony Ezzat and Tomaso Poggio, "Visual Speech Synthesis by Morphing Visemes," MIT AI Memo No 1658/ CBCL Memo No. 173, May, 1999.
4. F. Parke, "Parameterized models for facial animation," *IEEE Trans. on Computer Graphcis and Applications*, pp. 61-68, November, 1982.
5. Keith Waters, "A Muscle models for Animating 3D Facial Expression," Proceedings of SIGGRAPH 87, pp. 17-24, July, 1987.
6. Yuencheng Lee, Demetri Terzopoulos, and Keith Waters, "Realistic Modeling for Facial Animation," Proc. SIGGRAPH 95, pp. 55-62, 1995.
7. R. Quian, I. Sezan, and K. Matthews, "A robust real-time face tracking algorithm," *Proceedings of Internatinoal Conference on Image Processing*, 1998.
8. A. A. Montgomery, B. E. Walden and R. A. Prosda, "Effects of consonantal context on vowel lipreading," Journal of Speech & Hearing Research, vol. **30**, pp. 50-59, 1987.
9. V. A. Kozhevnikov and L. A. Chistovich, "Rech: artikulyatsiya I Vospriyatiye (TransArticulation andPerception)," Joint Publication Research Service, vol. **30**, pp. 543, 1965.
10. M. M. Cohen and D. W. Massaro, *Models and Techniques in Computer Animation*, Springer-Verlag, pp. 139-156, 1993.
11. S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book*, Campbridge University, 1998.
12. A. Hallgren & B. Lyberg "Visual speech Synthesis with Concatenative Speech," *Proceedings of Auditory-Visual Speech Processing*, 1998 Terrigal (New South Wales, Australia).

저자 약력

● 김 진 영 (Jin Young Kim)

1986년 2월: 서울대학교 전자공학과 졸업 (공학사)
1988년 2월: 서울대학교 전자공학과 석사과정 졸업 (공학박사)
1994년 8월: 서울대학교 전자공학과 박사과정 졸업 (공학박사)
1993년 3월~1994년 12월: 한국통신 소프트웨어연구소 전임연구원
1995년 1월~현재: 전남대학교 전자공학과 재직중
※ 주관심분야: 멀티모달 MMI, 음성신호처리

● 하 영 민 (Yong Min Ha)

2000년 2월: 광주대학교 전자계산학과 졸업 (공학사)
2000년 3월~현재: 전남대학교 정보통신협동과정 석사과정
※ 주관심분야: 입술동기화 및 인터넷 아바타

● 이 화 숙 (Hwa Sook Lee)

1990년 2월: 연세대학교 주거환경학과 졸업 (이학사)
1992년 2월: 연세대학교 주거환경학과 석사과정 졸업 (이학석사)
1998년 2월: 연세대학교 주거환경학과 박사과정 졸업 (이학박사)
1999년 3월~현재: 광주여자대학교 예술디자인 학부 재직중
※ 주관심분야: 실내디자인 및 CAD, 인터넷 아바타