2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

코퍼스 기반의 립싱크 알고리즘 개발

하영민, 김진영, 정수경
전남대 전자공학과 멀티미디어 DSP 연구실
Tel) 062-530-0472, Fax) 062-530-0472

Development of a lipsync algorithm based on A/V corpus

Young Min Ha , Jin Young Kim, Su Kyung Jung
Multimedia DSP Lab Dept. of Electronic Eng., Chonnam Nat'l Univ.
E-mail : kimjin@dsp.chonnam.ac.kr

요약

이 논문에서는 2차원 얼굴 좌표데이터를 합성하기 위한 음성과 영상 동기화 알고리즘을 제안한다. 영상변수의 획득을 위해 화자의 얼굴에 부착된 표시를 추적함으로써 영상변수를 획득하였고, 음소정보뿐만 아니라 운율정보들과의 영상과의 상관관계를 분석하였으며 합성단위로 시각소에 기반한 코퍼스를 선택하고, 주변의 음운환경도 함께 고려하여 연음현상을 모델링하였다. 입력된 코퍼스에 해당되는 패턴들을 lookup table에서 선택하여 주변음소에 대해 기준패턴과의 음운거리를 계산하고 음성파일에서 운율정보들을 추출해 운율거리를 계산한 후 가중치를 주어 패턴과의 거리를 얻는다. 이중 가장 근접한 다섯개의 패턴들의 연결부분에 대해 Viterbi Search를 수행하여 최적의 경로를 선택하고 주성분분석된 영상정보를 복구하고 시간정보를 조절한다.

1. 서론

인간의 의사소통에 있어서 영상정보[1,2]가 커다란 역할을 한다는 것은 널리 알려진 사실이다. 여러 가지 영상 정보중에서 얼굴은 대부분의 음성정보를 담고 있기 때문에 얼굴의 영상과 음성을 동시에 전달하는 것은 의사소통에 큰 도움이 된다. 이러한 영상의 합성법으로는 해당되는 음성에 따른 비디오프레임을 연결하거나 각각의 이미지를 연결하는 키-프레임(key-frame) 방법[3], 영상 변수를 사용해서 입력된 텍스트나 음성으로부터 영상을 합성하는 변수 기반의 인터폴레이션 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 음성을 이용해 변수화된 얼굴 영상을 합성하는 알고리즘을 제안하고 영상과 음성의 동기화와 자연스러운 발화의 문제에 주안점을 두었다.

2. A/V 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. 음성 데이터는 8kHz로 샘플링되었고 영상 데이터는 초당 30개씩 저장되었다.

2.2 영상처리

그림 1. 영상 데이터

이 논문에서는 발화에 따른 화자의 얼굴 움직임 정보를 얻기 위해 얼굴에 형광 스티커를 부착하여 그 위치정보로 얼굴의 움직임을 제어하는 방법[7]을 사용하였다. 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 PCA(Principal Component Analysis)를 이용하여 36개의 데이터를 8개의 주성분 벡터로 줄이고, 합성시 복원하였다.

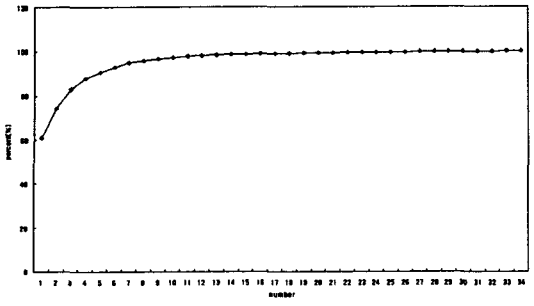


그림 2. 고유치에 따른 데이터 비율

2.3 음성처리

이 논문에서는 영상 데이터가 음소와 관련이 있을뿐만 아니라 음성의 다른 특징들, 피치(pitch)나 길이정보(duration), 그리고 강도(intensity)와 같은 기타 운율정보들과도 관련이 있다는 가정에 근거하여 이들의 상관관계를 분석하였다. 상관관계의 분석을 위해 각 음성의 강도 일반화를 수행하였고, 피치 정보를 얻기 위해 선형예측법(LPC : Linear prediction coding)을 이용하였으며 각 음소의 길이정보는 HTK(HMM Toolkit)이라는 툴에 의해 획득된 길이 정보를 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.
이 논문에서 상관관계를 분석하기 위해 선택한 변수는 각 음소의 운율정보-강도, 피치, 지속시간-와 입술영역의 높이와 폭이다. 이 영상변수는 얼굴에서 발화에 따라 변화하지 않는 미간과 코끝의 거리로 일반화되었다.

$$w = w_o / ref \tag{2.2}$$

$$h = h_0 / ref \tag{2.3}$$

ref : 미간과 코끝의 거리

그리고 각 음소가 지속되는 동안 운율정보가 변화하기 때문에 지속시간내에서 각 값의 최대값과 평균값을 취해서 상관관계 분석에 사용하였고 평균값보다는 최대값을 사용했을때 상관관계가 더 명확히 드러났다.

3. 입술 파라미터 합성을 위한 CVC 단위

본 논문에서는 영상변수의 합성단위로 음소가 아니라 시각소를 기반으로 하는 CVC를 사용하였다. 앞과 뒤의 코퍼스[VC+CVC+CV]도 고려하여 패턴의 결정시에 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.

표 1. 시각소 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 2. 코퍼스 그룹

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)

4. 입술의 합성

4.1 연음현상

영상 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 자연스러움이 떨어지게 된다. 이를 보완하기 위한 방법으로 연음현상의 모델링에 대한 연구가 수행되고 있는데 크게 look-ahead model[9], time-locked model[10], hybrid model[11]과 exponential model[12]의 네 가지가 있다.

4.2 코퍼스의 선택

합성하고자 하는 텍스트에 따른 영상변수를 선택하기 위해 코퍼스와 LUT내의 패턴과의 거리를 계산하는 과정은 다음과 같다.

1. 입력된 음소정보를 코퍼스 그룹으로 변환한다.
2. LUT에서 중심의 기준 코퍼스[CVC]에 해당되는 패턴을 찾는다.
3. 이 패턴들에 대해서 운율과 음소상의 거리 (D_{phon}, D_{pros})를 계산한다.

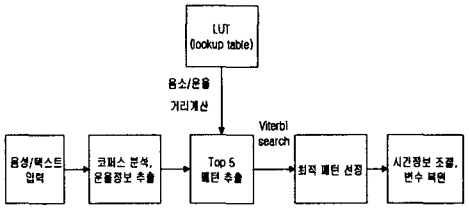


그림 3. 코퍼스의 선택과 변수의 합성과정

4.3 가중치의 결정

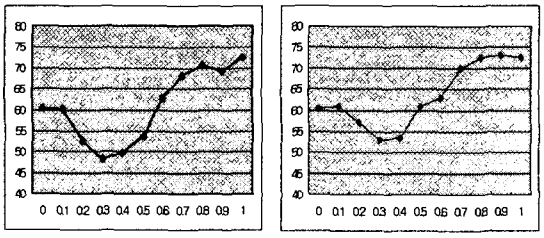
이 장에서는 최적의 패턴을 선택하기 위해 사용되는 음성과 운율거리를 결합하기 위한 가중치의 결정과정에 대해 설명한다.

$$D = \lambda D_{phon} + (1-\lambda) D_{pros} \quad (4.1)$$

세가지의 운율정보를 모두 사용하는 것이 타당함을 보이기 위해 세가지의 운율정보(피치, 강도, 지속시간)를 모두 사용할때와 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.

$$e = \frac{1}{N}\sum_{0}^{N}(p_{ref} - p_{syn})^2 \quad (4.2)$$

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



(a) (b)

((a) 피치, 강도, 지속시간, (b) 지속시간)

그림 4. 가중치에 따른 오차율 비교

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 때와 비교해보면 각각 15%와 27%씩의 차이가 난다.

4.4 운율과 음소거리 계산

중심코퍼스 CVC를 기준으로 운율과 음소정보, 그리고 PCA처리된 영상변수들과 이를 중심으로 한 앞쪽과 뒤쪽의 코퍼스에 해당되는 정보들이 배열되어 있다. 이렇게 하나의 CVC에 대한 여러개의 정보들이 해당 CVC_LUT내에 반복되어 저장되어서 변수의 합성시 각 패턴들에 대한 거리를 계산할 수 있다.
먼저 운율거리(D_{pros})는 해당 코퍼스그룹 내에서 중심코퍼스를 구성하는 시각소들의 피치(pitch), 강도(intensity) 그리고 지속시간(duration)상의 거리를 계산하고, 각각의 거리에 개별적인 가중치를 주어서 합친 후 모두 더한 값이다.

$$D_{pros} = \sum_{i} D_{pros_i} \quad (4.3)$$

$$D_{pros_i} = \lambda_1 D_p + \lambda_2 D_i + \lambda_3 D_d \quad (4.4)$$

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

$$D_x = \left| \left(\frac{x_x - x_{ref}}{\sigma_x} \right)^2 \right| \quad (4.5)$$

이 가중치들은 3장에서 얻은 각 시각소와 운율정보와의 관련성 분석 결과를 전체에서의 각각의 비율로 나타낸 값($\lambda_1, \lambda_2 \lambda_3$,)이다. 그리고 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.

$$D_{phon} = \frac{(D_{cons} + D_{vowel})}{2} \quad (4.6)$$

4.5 Viterbi search

이 논문에서는 Viterbi search를 사용하여 연음현상을 구현하는데 이는 연속되는 코퍼스들이 자연스럽게 연결되도록 하기 위해서이다.

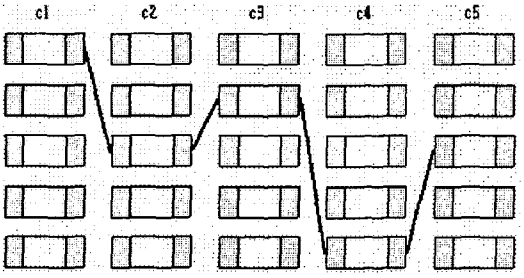


그림 5. Viterbi search

먼저 위에서 계산한 코퍼스그룹의 거리 D를 기준으로 하여 거리상으로 가장 가까운 패턴 다섯개를 선택한다. 그리고, 이 후보 패턴들과 다음에 연결될 패턴과의 거리를 계산한다.
여기에서 거리계산의 기준은 각 패턴이 연결되는 부분, 즉 시작부분과 끝부분의 영상변수(회색으로 표시된 부분)에 대해 계산되고 선택된 패턴에 대해서 시간정보를 조절하고 고유치벡터의 역행렬을 이용해 변수를 복원한다.

$$D_{ij} = \frac{1}{P}\sum_{k=0}^{P-1}(y_{ik} - y_{jk})^2 \quad (4.7)$$

D_{ij} : state i와 state j의 거리
P : 주요소의 수
y_{ik}, y_{jk} : state i와 state j의 k번째 요소

표 3. 패턴의 구성

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

4.6 입술 파라미터의 복원

위의 과정을 거쳐서 선택된 경로의 영상변수는 PCA를 이용하여 추출된 고유치벡터 (eigenvalue vector)로서 실제로 적용하기 위해서는 원래의 데이터로 복원해야 하고 실제의 지속시간과 일치하도록 입력데이터를 기준으로 각 시각소의 시간정보를 수정해야 한다.

$$X' = E' * Y \tag{4.7}$$

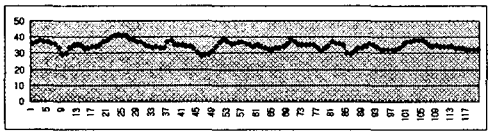
$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_{36} \end{bmatrix} = \begin{bmatrix} e_{1,1} \cdots e_{1,36} \\ e_{2,1} \cdots e_{2,36} \\ \cdots \\ e_{36,1} \cdots e_{36,36} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_{36} \end{bmatrix}$$

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.

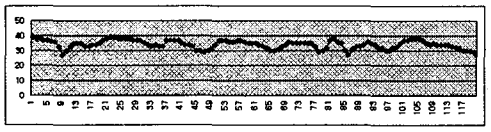
복원과정을 거친 후 실제로 변수를 합성하기 위해서는 입력된 음성데이터와 각 음소의 지속시간을 동일하게 조절해야 한다. 실제 데이터의 시간정보를 기준으로 하여 영상변수의 interpolation/decimation과정을 거쳐서 시간정보를 조절한다.

5. 합성결과

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. x축의 값은 이미지 프레임수이고 y축은 복원된 입술의 높이와 폭이다.



(a) 원래의 변수(높이)



(b) 합성된 변수(높이)

텍스트: 어렸을 때 나는 라틴어 학교에 다녔다.

그림 6. 변수 비교

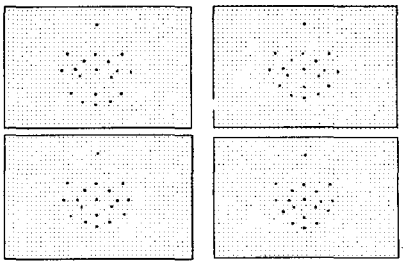


그림 7. 얼굴변수의 합성 (ex) a+m+o)

6. 결론

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 3D 영상합성기에서도 유용할 것으로 보인다. 그리고 이 합성기는 음소를 기준으로 하여 구성되기 때문에 나은 합성결과를 얻기 위해서는 정확한 레이블링이 필요하고, 정확한 영상정보 추출방법의 개발과, 가장 유사한 패턴을 선택하기 위한 음소, 운율상의 거리 계산법과 그 가중치의 최적화가 요구된다.

참고문헌

[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, "Lips. teeth, and the benefits of lipreading", Handbook of Research on Face Processing. A.W.Young and H.D.Ellis Editors, Elsevier Science Publishers. pp.223-233,1989.

[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 Computer Graphcis and Applications, pages 61-68, November 1982

[5] Keith Waters. A Muscle models for Animating 3D Facial Expression, Proc, SIGGRAPH 87, In computer Graphics,Vol.21, No.4, pp.17-24.July 1987

[6] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic Modeling for Facial Animation, Proc. SIGGRAPH 95. In Computer Graphics, 99.55-62, 1995

[7] R. Quian, I. Sezan, and K. Matthews, "A robust real-time face tracking algorithm," in Internatinoal Conference on Image Processing, 1998

[8] Montgomery A.A., Walden B.E. & Prosda R.A. (1987). "Effects of consonantal context on vowel lipreading", Journal of Speech & Hearing Research, 30, 50-59

[9] V.A Kozhevnikov and L.A Chistovich, "Rech:artikulyatsiya I Vospriyatiye (Moscow - Leningrad,1965). TransArticulation and perception". Washington, D.C : Joint Publication Research Service, Vol.30, pp.543,1965.

[10] M.M Cohen and D.W.Massaro, "Modeling Coarticulation in Synthetic Visual Speech", Models and Techniques in Computer Animation, N,M,Thalmann & D.Thalmann (Eds.), Tokyo : Springer-Verlag, pp.139-156.1993.

[11] Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev and Phil Woodland, "The HTK Book"