한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 26, No. 3, P. 49~59

한국어 동시조음 모델에 기반한 스피치 애니메이션 생성

장민정 정선진 노준용*

카이스트 비주얼미디어연구실

{joyful8296, sunjin225, junyongnoh}@kaist.ac.kr

Speech Animation Synthesis based on a Korean Co-articulation Model

Minjung Jang Sunjin Jung Junyong Noh*

KAIST, Visual Media Lab.

요 약

본 논문에서는 규칙 기반의 동시조음 모델을 통해 한국어에 특화된 스피치 애니메이션을 생성하는 모델을 제안한다. 음성에 대응되는 입 모양 애니메이션을 생성하는 기술은 영어를 중심으로 많은 연구가 진행되어 왔으며, 자연스럽고 사실적인 모션이 필요한 영화, 애니메이션, 게임 등의 문화산업 전반에 널리 활용된다. 그러나 많은 국내 콘텐츠의 경우, 스피치 애니메이션을 생략하거나 음성과 상관없이 단순 반복 재생한 뒤 성우가 더빙하는 형태로 시각적으로 매우 부자연스러운 결과를 보여준다. 또한, 한국어에 특화된 모델이 아닌 언어 비의존적 연구는 아직 국내 콘텐츠 제작에 활용될 정도의 퀄리티를 보장하지 못한다. 따라서 본 논문은 음성과 텍스트를 입력받아 한국어의 언어학적 특성을 반영한 자연스러운 스피치 애니메이션 생성 기술을 제안하고자 한다. 한국어에서 입 모양은 대부분 모음에 의해 결정된다는 특성을 반영하여 입술과 혀를 분리한 동시조음 모델을 정의해 기존의 입술 모양에 왜곡이 발생하거나 일부 음소의 특성이 누락되는 문제를 해결하였으며, 더 나아가 운율적 요소에 따른 차이를 반영하여 보다 역동적인 스피치 애니메이션 생성이 가능하다. 제안된 모델은 유저 스터디를 통해 자연스러운 스피치 애니메이션을 생성함을 검증하였으며, 향후 국내 문화산업 발전에 크게 기여할 것으로 기대된다.

Abstract

In this paper, we propose a speech animation synthesis specialized in Korean through a rule-based co-articulation model. Speech animation has been widely used in the cultural industry, such as movies, animations, and games that require natural and realistic motion. Because the technique for audio driven speech animation has been mainly developed for English, however, the animation results for domestic content are often visually very unnatural. For example, dubbing of a voice actor is played with no mouth motion at all or with an unsynchronized looping of simple mouth shapes at best. Although there are language-independent speech animation models, which are not specialized in Korean, they are yet to ensure the quality to be utilized in a domestic content production. Therefore, we propose a natural speech animation synthesis method that reflects the linguistic characteristics of Korean driven by an input audio and text. Reflecting the features that vowels mostly determine the mouth shape in Korean, a co-articulation model separating lips and the tongue has been defined to solve the previous problem of lip distortion and occasional missing of some phoneme characteristics. Our model also reflects the differences in prosodic features for improved dynamics in speech animation. Through user studies, we verify that the proposed model can synthesize natural speech animation.

키워드: 스피치 애니메이션, 동시조음, 강제 음성 정렬, 운율적 요소
Keywords: Speech animation, Co-articulation, Forced-alignment, Prosodic features

*corresponding author: Junyong Noh/KAIST(junyongnoh@kaist.ac.kr)

Received : 2020.06.19./ Review completed : 1st 2020.06.22. / Accepted : 2020.06.25.
DOI : 10.15701/kcgs.2020.26.3.49
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

전 세계에 6,700여 개의 언어가 존재하고 그중 300여 개의 언어만이 문자를 가지고 있지만, 문자를 만든 사람, 목적, 때가 분명하게 밝혀진 문자는 한글이 유일하다. 1443년 세종 대왕은 발음 기관의 모양과 소리의 특성을 살려 한글을 창제하였다. 과학적이고 체계적인 문자에 바탕을 둔 한국어는 언어학적으로 다른 언어와 차별화된 특성을 가진다. 이는 컴퓨터 그래픽스에서 음성에 대응되는 입 모양 애니메이션을 생성하는 데 중요하게 작용한다.

오늘날 스피치 애니메이션 기술은 영화, 애니메이션, 게임, 교육 등 문화산업 전반에 활용되며 현 수준의 콘텐츠를 제작하기 위해 필수적으로 요구된다. 해당 기술은 캐릭터의 감정과 상황을 효과적으로 표현하기 위한 연출 수단으로 자연스러운 입 모양 움직임을 재현하는 것이 보편적이며, 환경에 따라 움직임을 극대화하거나 생략하기도 한다. 전자의 경우, 공용어인 영어를 중심으로 활발하게 연구가 진행되고 있으며, 대량의 모션 캡처 데이터를 통해 실사에 가까운 움직임을 재현하기도 한다.

그러나 국내 콘텐츠는 특히 게임의 경우, 스피치 애니메이션을 생략하거나 음성과 상관없이 단순 반복 재생한 뒤에 성우가 더빙하는 형태로 그 결과가 시각적으로 부자연스럽다. 한국어 스피치 애니메이션 기술에 관한 연구는 2000년대 후반을 끝으로 잠정 중단된 상황이며, 대안인 언어 비의존적 연구는 아직 국내 콘텐츠 제작에 활용할 정도의 퀄리티를 보장하지 못한다. 한국어의 언어학적 특성을 고려하지 않고는 자연스러운 한국어 스피치 애니메이션 생성을 기대하기 어렵기 때문에 본 논문에서는 한국어에 특화된 스피치 애니메이션 생성 모델을 제안한다.

우리는 실제 발화과정에서 텍스트를 문자 그대로 발음하지 않고 여러 음운 현상을 거쳐 발음하며, 심리적으로 같은 소리로 인식하는 음소도 환경에 따라 조금씩 입 모양의 차이를 보인다. 스피치 애니메이션 생성이 까다로운 이유는 이처럼 언어마다 동시조음 (co-articulation) 현상이 발생하여 같은 문자도 환경에 따라 발음할 때의 입 모양이 달라지기 때문이다. 본 논문에서는 한국어에 특화된 동시조음 모델을 정의할 수 있도록 영어와 구분되는 한국어의 특성에 주목하여 방법론을 설정하였다.

본 논문은 크게 3가지 부분에서 기여도를 가진다. 첫 번째로 음소에 대응되는 입 모양(입술과 혀)을 상세하게 정의하였다. 두 번째로 한국어의 언어학적 특성을 반영한 동시조음 모델을 구축하여 실제 입 모양에 가까운 자연스러운 입 모양 애니메이션을 생성할 수 있다. 마지막으로 운율적 요소에 따른 차이를 결과에 반영하여 사실감과 역동성을 더하였다. 제안한 기술을 현업에서 사용한다면 한정적인 블렌드쉐입만으로 환경에 따라 달라지는 다양하고 자연스러운 입 모양의 움직임을 연출할 수 있을 것이다.

2. 관련 연구

1. 한국어 viseme 모델 스피치 애니메이션에서 캐릭터의 립싱크 정확도를 높이기 위해서는 음성과 시각적으로 일치하는 입 모양을 생성하는 것이 중요하다. 이를 위해 각각의 음소에 대응되는 입 모양 (viseme) 모델을 정의하는 작업이 선행된다. 한국어 viseme 정의에 관한 기존 연구는 영어에 다소 의존적인 경향을 보인다. 영어권 애니메이션 산업에서 사용하는 기본적인 입 모양을 한국어에 맞게 수정하여 모든 음절에 대한 입 모양 조합을 제시하거나 [1], 영어에만 존재하는 음소인 /f, v, th, sh/를 제외한 영어 viseme을 한국어 viseme 정의에 그대로 대응시킨다 [2].

이와 반대로 한국어 음소에 맞는 viseme을 자체적으로 정의하려는 시도도 존재한다. 그러나 사실적인 재현을 위해 혀와 치아를 비롯한 입안의 모습도 반영하는 최근 연구와 달리, 겉으로 두드러지게 관찰되는 입술만을 고려하기 때문에 단모음에 대해서만 viseme을 정의하거나 [3, 4] 여기에 자음 중 예외적으로 양순음이나 [5] 치음을 추가하기도 한다 [6]. 이중모음에 대해서는 반모음에 대한 viseme을 따로 정의하는 대신 그와 유사한 단모음 viseme들의 조합으로 대체한다 [2, 5]. 이처럼 입술 모양의 특징이 두드러지게 관찰되는 일부 음소에 한해 viseme을 정의할 경우, 나머지 음소들의 특성은 반영되지 않기 때문에 자연스러움이 떨어지게 된다. 또한, 어떤 음소를 viseme으로 정의하느냐에 따라 스피치 애니메이션이 왜곡될 위험이 있다.

2. 동시조음 모델 음소에 대응되는 블렌드쉐입을 키프레이밍하여 스피치 애니메이션을 생성할 때, 자연스러움을 향상시키기 위해 동시조음 모델을 적용한다 [7]. 이를 위해서는 크게 규칙 기반, 사전 기반, 데이터 기반 방법이 있다. 그중 한국어는 음운 현상을 규칙으로 정의하고 예외에 대해서만 발음 사전을 구축하는 규칙 기반 방법론이 적합하다 [8]. 한자처럼 글자 하나하나가 의미를 나타내는 표의문자와 달리, 한글은 뜻을 파악하기 쉽도록 본래 형태를 유지하면서 소리나는 대로 적는 표음문자로 음운 현상을 규칙으로 정의하기가 용이하기 때문이다.

기존의 연구는 직관적인 관찰에 근거하여 동시조음 모델을 정의한다. 예를 들어 일부 자음은 viseme 매핑을 생략하지만, 일부 자음은 같은 음절 내 모음의 입술이 벌어진 정도의 크고 작음에 따라 viseme을 매핑하거나 생략한다 [2]. 또한, 해당 연구는 텍스트를 음소 단위로 바꿔준 이후에 음성을 합성하지만, 스피치 애니메이션은 음성에 대응되는 입 모양 애니메이션을 생성하는 것이 주목적이다. 따라서 텍스트와 음성 정보를 모두 고려해서 동시조음 현상을 처리할 필요가 있다.

3. 운율적 요소 얼굴 애니메이션 연구에서 운율적 요소는 눈썹의 움직임, 눈의 깜빡임, 고개 끄덕임 등 부수 모션을 생성하는 데 주로 활용된다 [9, 10]. 스피치 애니메이션에서도 운율적 요소에 따른 차이를 반영하려는 시도가 존재하는데, 논문마다 활용하는 요소가 제각각이다. 예를 들어 음의 높이와 세기를 활용하거나 [7] 발화속도만 고려하기도 한다 [11]. 이와 달리 언어병리학에서는 발화속도의 감소와 음의 세기의 증가 모두 입을 크게 벌리고 혀를 많이 움직이는 등의 말 명료도 향상과 관련 있다고 밝히고 있다 [12]. 따라서 본 논문에서는 발화속도와 음의 세기에 따라 입 모양의 크기를 결정하는 모델을 제안하고자 한다.

3. 한국어 viseme 모델 정의

일반적으로 사용되는 viseme 블렌드쉐입은 고정된 형태다. 이에 대해 동시조음 현상으로 인해 viseme 쉐입을 하나로 특정하기 어렵다는 문제가 제기되며 짧은 애니메이션 형태의 역동적인 viseme 모델이 제안되었다 [13]. 그러나 이는 영어의 음소와 문자가 불일치하면서 발생하는 문제로 한국어와는 차이를 보인다. 영어는 일부 자음을 제외한 모든 음소가 입술 모양의 영향을 받는다 [7]. 음절의 구조를 보면 모음뿐만 아니라 자음도 음절의 핵을 이룰 수 있으며, 초성의 경우 최대 3개, 종성의 경우 최대 4개까지 자음이 올 수 있다 [14]. 동시조음 현상이 발생하면 바로 인접한 음소뿐만 아니라 최대 5칸 떨어진 음소의 입 모양에도 영향을 미치면서 [15] 하나의 음소가 여러 입 모양으로 관찰된다.

반면에 한국어에서 자음은 양순음(/ㅁ, ㅂ, ㅃ, ㅍ/)을 제외하고 혀의 위치 및 모양의 영향을, 모음은 주로 입술 모양의 영향을 받는다 [3, 16]. 한국어는 모음만이 음절의 핵을 이루기 때문에 입술 모양은 대부분 모음에 의해 결정된다. 또한, 자음은 초성, 종성에 최대 1개까지 올 수 있다. 동시조음 현상이 발생하더라도 서로 영향을 미치는 범위가 작다. 더 나아가 한국어는 창제 당시 음소와 문자가 서로 일대일 대응되도록 고안되었다. 따라서 고정된 viseme 쉐입이 스피치 애니메이션 생성에 문제가 되지 않는다.

본 논문에서는 기존 연구에서 정의한 viseme을 보완하기 위해 한국어 음운론에 기반하여 입술 모양과 입안의 혀의 위치 및 모양을 고려해 자음과 모음의 viseme 모델을 정의하였다. 또한, 영어와 다른 한국어의 언어학적 특성에 근거하여 입술과 혀 쉐입을 분리한 viseme 모델을 정의하였다.

3.1 한국어 자음 혀 모델

가시적인 영역에 있는 혀끝의 위치와 혀의 모양을 중심으로 시각적으로 유의미한 차이가 없다고 판단되는 음소들은 같은 혀 모델로 분류하였다. 한국어의 자음은 조음 위치, 조음 방법, 발성 유형에 따라 분류할 수 있으며 [14], 그중 발성 유형은 혀 모양에 별다른 차이를 가져오지 않으므로 고려하지 않았다.

더 나아가 같은 음소에 대해서도 음절 혹은 단어 내 음소 위치 등의 환경에 따라 변이음이 발생하는데 [14], 시각적으로 유의미한 차이를 가져오는 경우에만 모델에 반영하였다. /ㅅ/의 변이음이 그에 해당하는데, 일반적으로 /ㅅ/은 혓날이 윗니의 뒤쪽과 잇몸 사이에 가까이 가서 조음될 뿐 닿지는 않는다. 그런데 예외적으로 후행하는 모음이 / ㅑ, ㅒ, ㅕ, ㅖ, ㅛ, ㅟ, ㅠ, ㅣ/인 경우 혀끝이 아랫니 혹은 아랫니와 아랫잇몸 사이의 경계에 닿은 상태로 조음된다는 점에서 차이를 보인다. 이를 바탕으로 한국어 자음은 **Table 1**과 같이 총 7개의 혀 모델로 정의할 수 있으며, 예외적으로 양순음만 입술 모델을 함께 정의하였다.

3.2 한국어 모음 입술 및 혀 모델

한국어의 모음은 단모음과 이중모음으로 나뉜다. 단모음은 처음부터 끝까지 하나의 조음 동작으로만 만들어지는 모음이며, 이중모음은 반모음과 단모음이 결합하여 두 개의 조음 동작으로 만들어지는 모음을 말한다. 모음은 입술의 돌출 여부, 입천장 기준 혀의 전후 위치, 혀의 높이에 따라 세부적으로 분류될 수 있다. 혀 모델은 이들 중 혀의 위치 및 모양에 영향을 주는 기준에 따라 **Table 2**에서 볼 수 있는 것처럼 5가지 모델로 정의할 수 있다.

입술 모델의 경우, 표준 발음법에서는 단모음을 10모음 체계로, 이중모음을 11모음 체계로 규정하고 있지만, 이는 표준어 화자들이 보이는 현실적인 모음 체계로 보기 어렵다. 실제 발화에서 단모음 /ㅚ/와 /ㅟ/는 이중모음으로 조음되며, 단모음 / ㅐ/와 / ㅔ/는 발음상으로는 구분되지 않는다. 따라서 표준어의 현실적인 단모음 체계는 7개의 단모음으로 구성된다고 볼 수 있다 [14].

한편 이중모음은 별도의 입술 모델을 정의하지 않아도 반모음과 단모음 각각의 입술 쉐입을 순차적으로 합성하는 과정으로 이해할 수 있다 [16]. 한국어에는 2개의 반모음 /j/(ㅣ)와 /w/(ㅗ/ㅜ)가 존재하는데, 이들과 결합된 이중모음을 각각 /j/계 이중모음, /w/계 이중모음이라고 부른다. 본 논문에서는 /ㅢ/를 하향 이중모음으로 보고 /j/계 이중모음으로 분류하였다. 지금까지의 논의를 바탕으로 한국어 모음은 단모음과 반모음에 대해 **Table 2**와 같이 총 9개의 입술 모델로 정의할 수 있다.

Table 1: The proposed tongue model for Korean consonants

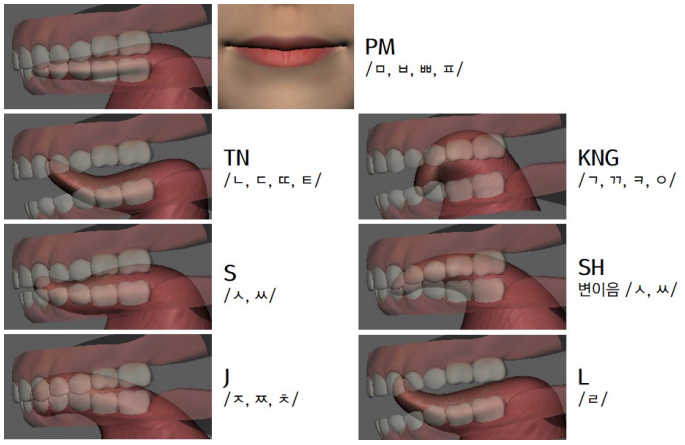
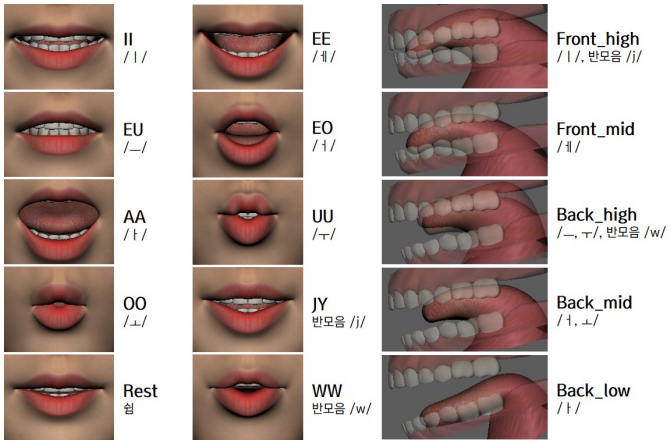


Table 2: The proposed lip and tongue models for Korean vowels



4. 스피치 애니메이션 생성

본 논문에서 제안하는 한국어 스피치 애니메이션 생성 모델은 인풋 단계, 애니메이션 단계, 아웃풋 단계로 구성된다 (**Figure 1**). 인풋 단계에서 스피치 음성과 텍스트를 인풋으로 받으면 강제 음성 정렬 작업을 수행하여 자소 시퀀스를 추출한다. 이후 애니메이션 단계에서 규칙 기반의 동시조음 모델을 적용하여 음소 시퀀스를 생성하며, 이때 입술과 혀를 분리하여 독립적으로 접근한다. 마지막으로 아웃풋 단계에서는 음성에서 운율적 요소를 추출하여 그에 따라 입술 쉐입의 가중치를 결정하고, 입술 및 혀 쉐입을 키프레이밍하여 스피치 애니메이션을 생성한다.

4.1 인풋 단계

인풋 단계는 스피치 애니메이션을 생성하기 위한 전처리 단계에 해당한다. 사용자로부터 스피치 음성과 텍스트를 입력받으면, 음성의 특정 프레임이 어떤 문자에 해당하는지 정렬 정보를 추출함으로써 음성과 텍스트의 타이밍을 맞추는 강제 음성 정렬(forced alignment) 작업이 필요하다 [17]. 본 논문에서는 Montreal Forced Aligner[18]를 사용하여 강제 음성 정렬 작업을 자동화하였다. 해당 툴은 음성 인식 툴 Kaldi[19]를 바탕으로 음운론적 맥락, 음소별 특성, 화자의 특성을 모두 고려하여 언어에 대한 음향 모델(acoustic model)을 학습시킨다. 이를 이용하기 위해서는 대량의 코퍼스(음성 및 텍스트)와 사전이 요구된다. 사전에 등재된 단어로 텍스트가 구성된 경우에만 강제 음성 정렬 작업을 수행할 수 있기 때문에 가능한 모든 단어를 수용하도록 본 논문에서는 한국어의 언어학적 특성을 고려해 맞춤형 사전(A)를 구축하였다.

4.1.1 강제 음성 정렬을 위한 한국어 맞춤형 사전

텍스트가 인풋으로 들어오면 사전에 등재된 단어를 기준으로 토큰화를 수행한다. 그러나 띄어쓰기를 기준으로 단어 토큰화를 수행하는 영어와 달리 한국어는 자연어 처리를 수행하기가 특히 까다롭다. 한국어는 교착어로 동사 및 형용사와 결합하는 어미가 발달하여 활용형이 많은 것이 특징이며, 조사 또한 명사 등과 결합한다. 이를 모두 사전에 등재하는 것은 비효율적이기 때문에 앞 문장성분과 어미 및 조사를 분리하기 위해서는 형태소 토큰화가 적합하다. 따라서 한국어는 형태소 토큰화를 수행하기 위한 맞춤형 사전(B)가 추가로 요구된다.

본 논문에서는 표준국어대사전에서 단어 목록을 확보하였고, 그중 동사와 형용사를 사전에 필요한 형태로 가공하였다. 동사와 형용사는 활용할 때 형태가 변하지 않는 어간과 형태가 변하는 어미의 결합으로 이루어져 있으므로, 원형뿐만 아니라 활용형도 처리할 수 있도록 어간의 형태로 사전에 등재하였다. 다만 환경에 따라 어간도 예외적으로 형태가 변하는 경우가 있는데, 모음 축약 현상과 최현배(1937) 학설의 불규칙 활용 중 어간에 변형이 일어나는 'ㄷ', 'ㄹ', 'ㅂ', 'ㅅ', 'ㅎ', '우', '으', '르' 불규칙 활용 결과를 사전에 추가하였다 [20, 21, 22].

4.1.2 코퍼스

본 논문은 오픈 소스인 Zeroth-Korean 코퍼스를 사용하였다. 115명 화자의 52.8 시간 음성과 텍스트로 구성되어 있다.

4.1.3 자음과 모음의 타이밍 정보 추출

스피치 음성과 텍스트를 입력받으면 학습시킨 한국어 음향 모델을 바탕으로 강제 음성 정렬 작업을 수행한다. 이를 통해 자소 및 쉼의 타이밍 정보가 담긴 자소 시퀀스를 추출하였다.

4.2 애니메이션 단계

애니메이션 단계에서는 규칙 기반의 동시조음 모델을 적용하여 자소 시퀀스를 음소 시퀀스로 변환하며, 각 음소에 대응되는 viseme 블렌드쉐입을 매핑한다. 이때 영어와 한국어의 차이점을 반영하여 입술과 혀 쉐입에 대해 독립적으로 접근한다. 이를 바탕으로 동시조음 현상도 입술과 혀를 분리하여 별도로 처리한다. 동시조음 모델은 크게 Grapheme-to-Phoneme (G2P) 모델에 관한 것과 스피치 애니메이션 생성에 관한 것으로 구분된다.

먼저 G2P 모델은 텍스트를 실제 소리나는 대로 바꿔주는 자연어 처리 영역이다. 본 논문에서는 규칙을 음운 현상의 환경을 고려해 세분화하여 재정의하였으며, 스피치 애니메이션에 맞게 단순화하였다. 이후 자연스러운 한국어 스피치 애니메이션을 생성하는 데 필요한 실질적인 규칙을 마련하였다. 총 12가지 규칙을 바탕으로 한국어에 특화된 동시조음 모델을 정의하였으며, 그중 1-4번이 G2P 모델에, 5-12번이 스피치 애니메이션 생성 규칙에 해당한다. 각 규칙에 대한 설명은 다음과 같다.

1. 자음에 대한 음운 현상에 한정

텍스트를 왼쪽에서 오른쪽으로 읽으면서 규칙을 적용하기 때문에 음운 현상은 대부분 종성과 초성의 만남에서 일어난다. 다만 반모음 첨가와 같이 모음이 변동되는 경우도 있는데, 예를 들어 '피어'는 피여로 발음하는 것도 허용한다. 그러나 이는 필수가 아니기 때문에 동시조음 모델에는 반영하지 않는다. 예외적으로 표준 발음법 제5항의 자음을 첫소리로 가지는 음절의 'ㅢ'는 [ㅣ]로 발음한다는 규칙은 동시조음 모델에서 함께 처리한다.

2. 같은 viseme 클래스 내에서의 변화 생략

선행 과정으로 viseme 모델을 정의할 때, 시각적으로 유의미한 차이가 없는 음소들은 같은 viseme 클래스로 분류하였다. 따라서 음운 현상에 의해 같은 viseme 클래스 내에서 변화가 일어나는 규칙은 생략할 수 있으며, 예시로는 경음화가 있다.

3. 종성을 중심으로 음운 현상 재정의 (예외: 초성 'ㅇ', 'ㅎ')

음운 현상은 대부분 종성과 초성의 만남에서 일어나기 때문에 종성을 중심으로 음운 현상을 재정의할 수 있다. 다만 초성 'ㅇ'과

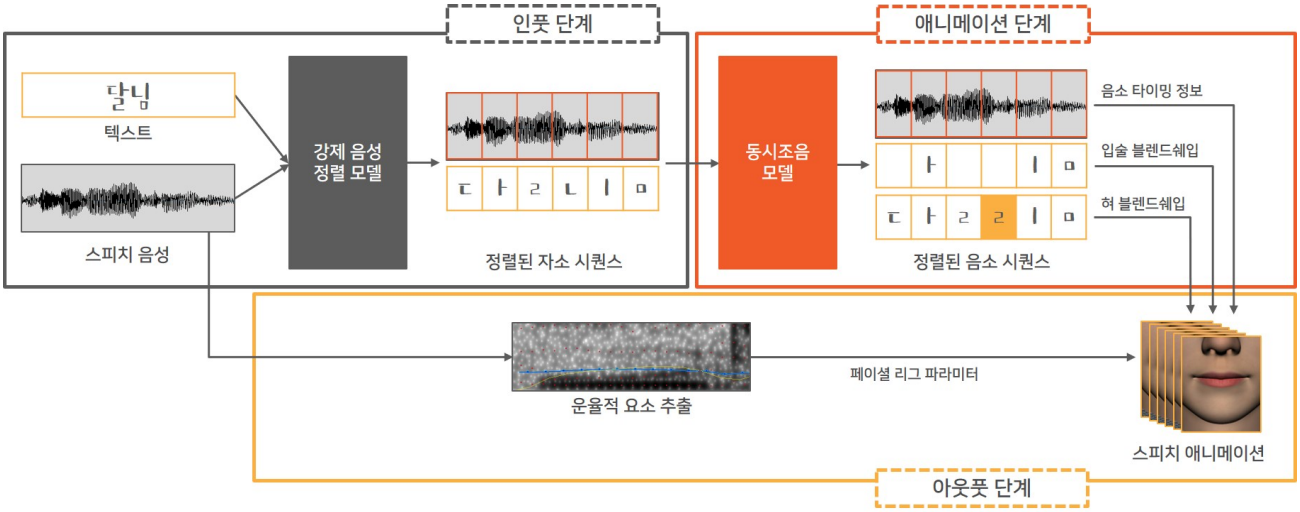


Figure 1: Overview of our Korean speech animation synthesis model

'ㅎ'에 관한 규칙을 별도로 분리한 이유는 두 음소가 초성에 위치할 때 다른 자음과 달리 고유의 입술 모양뿐만 아니라 혀의 위치 및 모양도 없는 탓에 주변 환경의 영향을 많이 받기 때문이다. 따라서 특정 규칙에 해당하지 않는 경우 매핑을 생략한다.

초성 'ㅇ'은 음가가 없어 대부분 연음 현상이 발생하며, 'ㄴ' 첨가에 해당하는지도 고려해야 한다. 또한, 연음 현상도 환경에 따라 음운 규칙과 연음 규칙을 적용하는 순서가 달라진다. 모음으로 시작하는 문법형태소의 경우 곧바로 연음 현상이 발생하는 반면, 실질형태소의 경우 음절의 끝소리 규칙 또는 자음군 단순화를 적용한 이후에 연음 현상이 발생한다. 이는 예시 '닭이'[달기]와 '닭 앞에'[다가페]의 차이에서도 확인할 수 있다. 마찬가지로 초성 'ㅎ'은 발음하기 시작할 때 이미 혀의 위치가 같은 음절 내 중성을 발음할 때와 같은 상태를 취한다 [14].

본 논문에서는 1-3번의 규칙을 바탕으로 표준 발음법에 따른 음운 현상 중 9가지 음운 규칙(제12항 'ㅎ' 받침의 발음, 제8-9항 음절의 끝소리 규칙, 제10-11항 자음군 단순화, 제19항 'ㄹ' 비음화, 제20항 'ㄹ' 비음화 예외, 제20항 유음화, 제17항 구개음화, 제29항 'ㄴ' 첨가, 제12항 격음화)과 3가지 연음 규칙(제13항, 제14항, 제15항)에 대한 G2P 모델을 정의하였다.

4. 음운구 내에서 G2P 모델 적용

음운 현상은 한 단어 내에서만 적용되는 것이 아니라 단어나 문장 간에도 관찰된다. 하지만 두 음소 사이에 음운구 경계가 놓이게 된 경우 음운 현상이 일어나지 않는다 [14]. 본 논문에서는 강제 음성 정렬 작업을 통해 쉼에 대한 타이밍 정보를 확보하였고, 이를 경계로 각 음운구 내에서만 G2P 모델을 적용하였다.

5. 모음의 타이밍에 입술 및 혀 쉐입 매핑

한국어에서 입 모양은 대부분 모음에 의해 결정된다. 따라서 모음의 타이밍에는 그에 대응되는 입술 및 혀 쉐입을 매핑한다.

6. 자음의 타이밍에 혀 쉐입 매핑

자음의 타이밍에는 고유의 혀 쉐입을 매핑하며, 이때 입술 쉐입은 키프레이밍 없이 모음의 타이밍에서 키프레이밍한 입술 쉐입들의 스플라인 보간을 통해 생성된다. 이를 통해 동시조음 현상에 따른 입 모양의 변화를 재현하며, 자음이 가진 고유의 혀 쉐입 특성도 결과에 반영할 수 있다.

7. 양순음의 타이밍에 입술 및 혀 쉐입 매핑

양순음(/ㅁ, ㅂ, ㅃ, ㅍ/)의 경우, 입술 쉐입을 함께 매핑한다. 두 입술이 닿았다가 떨어지면서 나는 소리로, 입술이 닫혀야 한다는 주요한 시각적인 특징을 가지기 때문이다. 이를 위해 본 논문에서는 양순음일 때 입술 쉐입의 가중치를 최대로 설정하였다.

지금까지 다룬 5-7번 규칙은 한국어 동시조음 모델의 기본 원리를 보여준다. 예시 '고막'을 통해 이를 시각적으로 나타내면 **Figure 2**와 같다. 다만 해당 규칙만으로는 자연스러움을 보장하기 어려우며, 환경에 따라 예외적인 입술 쉐입 매핑 혹은 입술 쉐입 간 블렌딩 등의 세부 규칙들이 추가로 요구된다.

8. 치경음, 치경경구개음은 같은 음절 모음의 입술 쉐입 매핑

자음은 양순음을 제외하고 고유의 입술 쉐입 없이 같은 음절 내 모음을 따라간다. 그런데 자음의 분류 기준 중 조음 위치에 따라 아래 조음 기관과 위 조음 기관의 거리를 좁히기 위해 입 모양이 크게 변화하는 경우가 있다. 치경음(/ㄴ, ㄷ, ㄸ, ㅌ, ㄹ, ㅅ, ㅆ/)과 치경경구개음(/ㅈ, ㅉ, ㅊ/)이 이에 해당하는데 대부분 혓날이 아래 조음 기관으로, 앞쪽에 위치한 치경과 전경구개가 위 조음 기관으로 관여한다 [14]. 이는 조음시 두 조음 기관이 서로 밀착되어 있어야 함을 의미한다. 따라서 이웃한 모음이 입을 아래로 크게 벌리는 경우, 조건을 충족하기 위해 예외적으로 임계점 설정이 필요하다. 해당 규칙을 적용하는 자음 혹은 모음의 종류는 다음과 같으며, 총 3가지 환경에서 적용한다.

Figure 2: The process for the co-articulation model application. In each row, the first line represents a phoneme, the second line represents the mapped lip shape, and the third line represents the tongue shape. The first and the second row show the result from applying Rule 5, and Rule 6 and 7, respectively. The lip shapes for the consonants excluding bilabial in the third row are created through interpolation.

자음: /ㄴ, ㄷ, ㄸ, ㅌ, ㅅ, ㅆ, ㅈ, ㅉ, ㅊ/
단모음: / ㅏ, ㅔ, ㅟ/
이중모음: / ㅑ, ㅒ, ㅕ, ㅖ, ㅘ, ㅙ, ㅚ, ㅝ, ㅞ/

(1) 초성에 위치한 자음 + 단모음
(2) 단모음 + 종성에 위치한 자음
(3) 이중모음 + 종성에 위치한 자음

본 논문에서는 모음의 입술 쉐입이 아래로 벌어져 더 이상 치경(경구개)음을 조음할 수 없는 위치에 도달할 때를 기준으로 각 모음에 대한 임계점을 설정하였다. 이에 따라 같은 음절 내 모음 입술 쉐입의 가중치가 임계점을 넘어가면, 임계점에 해당하는 크기의 모음의 입술 쉐입을 자음의 타이밍에도 매핑하였다.

다만 치경음 /ㄹ/은 아래 조음 기관이 혀끝이기 때문에 입이 크게 벌어지더라도 조음이 가능하다. **Figure 3**에서 예시 '아래'와 '아내'의 비교를 통해 치경(경구개)음 예외 규칙 적용 여부에 따른 결과의 차이를 확인할 수 있다. 다만 연구개음 /ㄱ, ㄲ, ㅋ/과 치경음 /ㄹ/ 조음시 일상적인 발화 상황에서 벗어날 경우, 조음상의 편의를 위해 입의 크기를 줄일 수 있다. 그러므로 얼굴 모델의 페이셜 리그에 따라 임계점 설정을 고려할 필요가 있다.

9. 양순음 전후로 원순성 가진 모음 오면 두 입술 쉐입 블렌딩

양순음이 입을 둥글게 오므려서 소리를 내는 원순성을 가진 모음과 이웃하는 경우, 양순음의 타이밍에 양순음과 모음 두 입술 쉐입을 블렌딩한다. 이를 통해 해당 모음의 특성인 원순성의 영향을 반영하고, 양순음과 모음을 보간할 때 이질감을 최소화할 수 있다. 원순성을 가진 모음의 종류는 3가지이며, 총 6가지 환경에서 양순음과 원순성을 가진 모음이 연속될 수 있다.

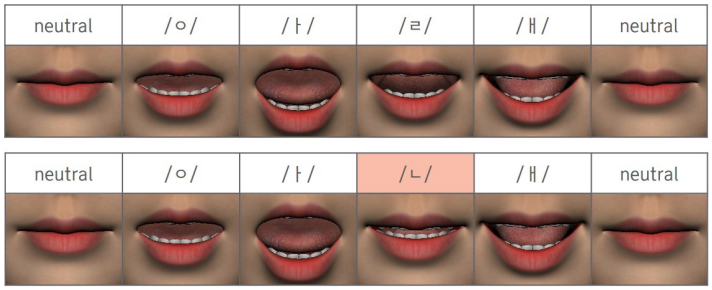


Figure 3: Comparison of the results from the application of the aveolar and the alveo-palatal exception rule

단모음: /ㅗ, ㅜ/
/w/계 이중모음: /ㅘ, ㅙ, ㅚ, ㅝ, ㅞ, ㅟ/
/j/계 이중모음: /ㅛ, ㅠ/

(1) 초성에 위치한 양순음 + 단모음
(2) 초성에 위치한 양순음 + /w/계 이중모음
(3) 단모음 + 종성에 위치한 양순음
(4) /j/계 이중모음 + 종성에 위치한 양순음
(5) 단모음 + 다음 음절 초성에 위치한 양순음
(6) /j/계 이중모음 + 다음 음절 초성에 위치한 양순음

본 논문에서는 이와 같은 경우, 원순성을 가진 모음(/ㅗ, ㅜ/, 반모음 /w/)의 입술 쉐입에 0.4만큼 가중치를 적용하여 양순음의 타이밍에 두 입술 쉐입을 블렌딩하였다. **Figure 4**는 해당 예외 규칙을 적용하기 전과 후의 차이를 보여준다. 양순음의 주요한 시각적인 특징을 유지하면서도 같은 음절 내 모음의 영향으로 입이 동그랗게 모이는 경향성을 보임을 확인할 수 있다.

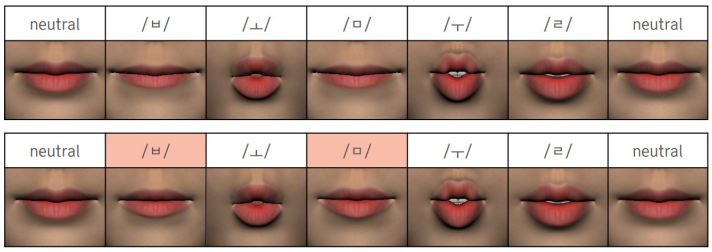


Figure 4: Comparison of the results from before (the first row) and after (the second row) applying the bilabial exception rule

10. 이중모음은 반모음과 단모음의 결합으로 처리

스피치 애니메이션에서 이중모음은 반모음과 단모음을 순차적으로 매핑하는 과정으로 접근할 수 있다. 한국어에서 이중모음의

종류는 총 13가지이며, 이를 본 논문에서 정의한 viseme 모델의 결합으로 나타내면 **Table 3**과 같다.

Table 3: Types of Korean diphthong

/ ㅑ/: 반모음 /j/ + / ㅏ/	/ ㅒ/: 반모음 /j/ + / ㅐ/	/ ㅕ/: 반모음 /j/ + / ㅓ/
/ ㅖ/: 반모음 /j/ + / ㅔ/	/ㅘ/: 반모음 /w/ + / ㅏ/	/ㅙ/: 반모음 /w/ + / ㅐ/
/ㅚ/: 반모음 /w/ + / ㅣ/	/ㅛ/: 반모음 /j/ + /ㅗ/	/ㅝ/: 반모음 /w/ + / ㅓ/
/ㅞ/: 반모음 /w/ + / ㅔ/	/ㅟ/: 반모음 /w/ + / ㅣ/	/ㅠ/: 반모음 /j/ + /ㅜ/
/ㅢ/: /ㅡ/ + 반모음 /j/		

반모음은 모음과 달리 한 음절을 구성하지 못할 정도로 지속시간이 짧기 때문에 블렌딩 과정에서 단모음보다 지속시간을 짧게 설정해야 한다. 그러나 이중모음에서 반모음과 단모음 각각의 비율에 관해서는 밝혀진 바가 없으므로 본 논문에서는 반모음의 지속시간을 단모음의 것의 절반으로 하였다.

11. 이중모음의 반모음과 단모음 입술 쉐입 블렌딩

모음을 반모음과 단모음으로 분할하여 순차적으로 매핑하는 이중모음은 프레임 간 입 모양의 변화가 큰 탓에 부자연스러움을 초래하기 쉽다. 또한, 반모음은 고유의 입술 쉐입을 매핑하지만, 이웃한 단모음의 성질에 따라 입 모양이 크게 달라진다. 따라서 반모음의 타이밍에 두 입술 쉐입을 블렌딩하고, 원순성을 고려해 환경에 따라 블렌딩 규칙을 달리 적용한다. 원순성을 가진 모음은 조음시 원순성을 그대로 유지하는 것이 경제적이므로 입 모양의 변화가 크지 않기 때문이다. 원순성을 가진 이중모음은 2종류이며, 총 3가지 환경에 따라 규칙을 달리 적용한다.

/j/계 이중모음: /ㅛ, ㅠ/
/w/계 이중모음: /ㅘ, ㅙ, ㅚ, ㅝ, ㅞ, ㅟ/

(1) /j/계 이중모음: 단모음의 가중치(α)를 작게 설정하여 블렌딩
(2) /w/계 이중모음: 단모음의 가중치(β)를 작게 설정하여 블렌딩
(3) 그 외: 반모음과 단모음의 가중치(α)를 같게 설정하여 블렌딩

이중모음의 특성을 고려해 반모음이 주변 환경의 영향을 받도록 단모음에 가중치를 두어 반모음과 단모음 두 입술 쉐입을 블렌딩한다. 본 논문에서는 경험적으로 그 값을 α는 0.4, β는 0.1로 설정하였다. **Figure 5**는 예시 '우유'를 통해 해당 예외 규칙의 적용 여부에 따른 결과의 차이를 보여준다.

12. 짧은 쉼은 입술 및 혀 쉐입 생략

스피치 애니메이션을 생성할 때 쉼에서의 입술 모양도 고려할 필요가 있다. 일반적으로 이를 느슨한 상태로 입을 열어두는 것으로 묘사한다 [7]. 그러나 쉼은 문장 간뿐만 아니라 문장 내 단어 간 혹은 단어 내에서도 포착될 수 있는데, 쉼의 길이가 지나치게 짧은 경우까지 모두 매핑할 경우 스피치 애니메이션을 생성했을 때 끊기는 느낌을 줄 위험이 있다. 따라서 임계점을 0.4초로 설정하고, 임계점 미만의 쉼에서는 매핑을 생략하였다.

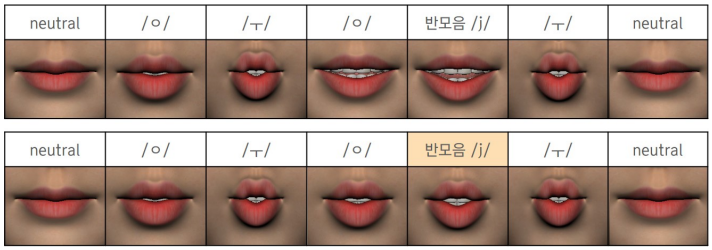


Figure 5: Comparison of the results from before (the first row) and after (the second row) applying the diphthong exception rule

4.3 아웃풋 단계

아웃풋 단계는 스피치 애니메이션을 생성하는 단계로 각 음소의 타이밍에 그에 맞는 입술 또는 혀 쉐입을 키프레이밍한다. 본 논문은 운율적 요소에 따른 차이를 반영하여 같은 텍스트에 대해서도 음성의 특성에 맞는 다양한 스피치 애니메이션을 생성하였다.

4.3.1 운율적 요소에 따른 입술 쉐입 가중치 결정

인풋으로 들어온 음성에서 t번째 모음의 입술 쉐입의 가중치(w_t)는 아래의 식 1에 의해 결정된다.

$$w_t = a(1 - e^{-ci_t})/(\frac{1}{d_t} + a) \qquad (0 \leq w_t \leq 1) \qquad (1)$$

한국어에서 입술 모양은 모음에 의해 결정되기 때문에 운율적 요소 또한 모음의 음의 세기를 활용하며, 이때 $i_t(i_t \geq 0)$는 t번째 모음에서의 음의 세기를 의미한다. 한편, 발화속도는 일반적으로 단위 시간 당 음절 혹은 낱말 수로 표시하지만 [23], 이는 음성의 평균 발화속도를 기준으로 삼기 때문에 음성 내 운율적 요소의 다양한 변화를 반영할 수 없다는 한계가 있다. 따라서 본 논문에서는 음절 길이를 측정 단위로 선정하였으며, 이를 $d_t(d_t \geq 0)$로 표현하였다. 상수 a는 음의 세기와 음절 길이를, c는 음의 세기를 위해 존재하는데, 본 논문에서는 a는 20, c는 0.02로 설정하였다.

5. 실험 및 결과

본 논문은 다음과 같은 환경에서 구현되었다: 인텔 i7-8700 3.20GHz CPU. 또한, 논문에서 제안하는 한국어 스피치 애니메이션 생성 모델은 end-to-end 시스템으로 MAYA에서 구현되었다.

국립국어원에서 제공하는 20대 여성의 서울말 낭독체 발화 데이터 중 5개의 데이터를 입력으로 스피치 애니메이션을 생성했을 때, 각 단계의 소요 시간을 확인해본 결과 **Table 4**와 같다. 평균 1초당 강제 음성 정렬 작업 수행에 1.017초가 걸리며, 운율적 요소를 활용한 애니메이션 생성에 0.338초가 소요된다.

제안하는 운율적 요소를 반영한 모델과 기존의 연구[24, 2]를 비교한 결과는 **Figure 6**과 같다. 인풋 텍스트의 특정 음소에 해

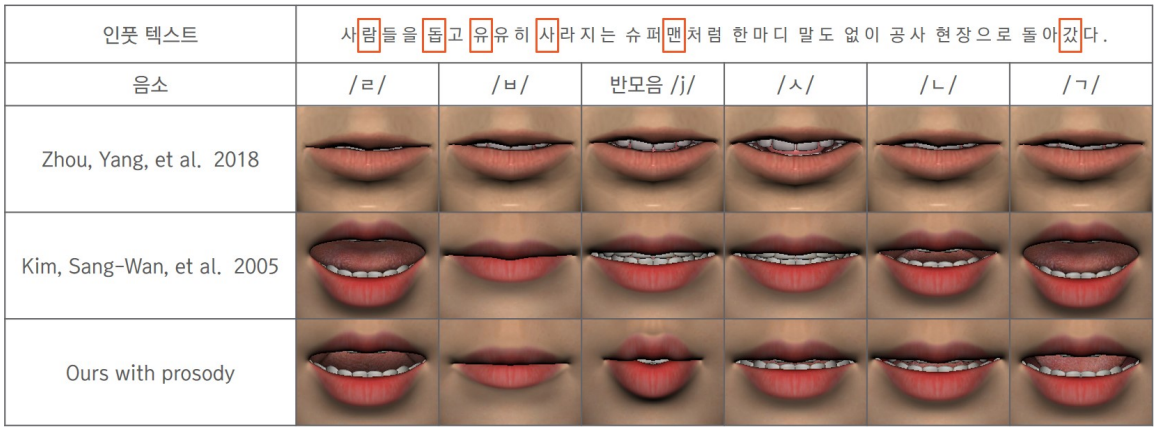


Figure 6: Comparison of results from the proposed model and previous studies. The third and the fourth rows show the results from a language-independent model[24] and a rule-based co-articulation model[2], respectively. The fifth row shows the results from our model reflecting prosodic features.

Table 4: Processing times for each phase

	음성 길이 (s)	글자 수	인풋 단계 (s)	애니메이션 및 아웃풋 단계 (s)
음성 1	5.021	14	6.686	1.509
음성 2	5.002	17	6.678	1.812
음성 3	13.913	54	11.963	4.353
음성 4	13.762	68	12.024	5.009
음성 5	18.608	93	12.724	6.529

당하는 프레임에서의 결과를 보여준다. 언어 비의존적 모델[24]은 음소에 따른 차이를 거의 보이지 않아, 규칙 기반의 동시조음 모델[2]과 중점적으로 비교하였다. 기존의 연구[2]는 음소 /ㄱ, ㄹ/에 따른 차이가 관찰되지 않는 반면, 제안하는 모델은 각각의 혀 쉐입의 특성을 명확하게 반영한다. 또한, /ㄴ, ㅅ/의 경우 기존 연구[2]는 고정된 viseme 쉐입을 정의하지만, 제안하는 모델은 혀 쉐입만으로 음소의 특성을 반영하며 입술 쉐입이 같은 음절 내 모음의 것을 따라간다. 마찬가지로 /ㅂ/에서도 인접한 음소 /ㅗ/의 영향을 반영한다. 마지막으로 기존의 연구[2]는 반모음 /j/를 단모음 / ㅣ/로 대체하여 시각적으로 매우 부자연스러운 반면, 제안하는 모델은 반모음의 입술 쉐입을 정의하고 환경에 따른 차이를 반영하여 원순성을 유지함을 확인할 수 있다. 운율적 요소에 따른 차이를 비교한 결과는 참고 영상에서 확인 가능하다.

6. 정성적 평가

본 논문에서 제안하는 스피치 애니메이션 생성 모델이 실제로 자연스러운 입 모양 애니메이션을 생성하는지 검증하기 위해 유저 테스트를 진행하였다. 중립적인 감정으로 말하는 표준어 낭독체 발화 음성을 활용하여 스피치 애니메이션을 생성하였고, 운율적 요소에 따른 차이를 관찰하기 위해 중립적인 음성과 운율적 요소가 강하게 반영된 음성 2종류를 비교 대상으로 삼았다. 실험에 사용된 데이터는 **Table 5**와 같다. 또한, 본 실험은 기존의 연구에 따른 모델 2개와 본 논문에서 제안하는 모델 2개를 포함한 총 4개의 모델을 비교하였다. 기존 연구는 음성만을 인풋으로 받는 언어 비의존적 모델[24]과 규칙 기반의 동시조음 모델[2]이다. 본 논문에서 제안하는 모델로는 운율적 요소를 반영하지 않은 동시조음 모델과 운율적 요소를 반영한 동시조음 모델을 준비하였다.

Table 5: 6 Audio data used in the user studies. 1-3 correspond to neutral audio, and 4-6 correspond to audio with prosodic features.

	음성 길이	텍스트
음성 1	5.801 (s)	여기에서 오늘 꼭 알아야 할 내용이 있습니다.
음성 2	8.427 (s)	사람들을 돕고 유유히 사라지는 슈퍼맨처럼 한마디 말도 없이 공사 현장으로 돌아갔다.
음성 3	5.049 (s)	그러나 지식은 그 종류와 양이 무한하다.
음성 4	4.885 (s)	새들도 짐승들도 착한 나무꾼의 친구들이라 나무꾼은 조금도 외롭지 않았어요.
음성 5	3.985 (s)	소는 어질고 순해서 어린아이들에게도 순순히 따르고 말도 잘 들었다.
음성 6	7.030 (s)	슬기로움은 우연하게 얻어지는 게 아니거든.

본 실험은 총 8명(여성 3, 남성 5)을 대상으로 진행하였다. 실험 설명 및 휴식 시간 5분, 첫 번째 실험 15분, 두 번째 실험 10분을 포함하여 총 30분간 실험을 진행하였다. 또한, 실험에 사용된 영상의 순서는 랜덤으로 배정하였다.

첫 번째 실험에서는 한 스피치 애니메이션 영상을 3번 반복 재생하여 자연스러움에 대해 5점 척도로 평가하였다. 두 번째 실험은 같은 음성에 대한 2개의 스피치 애니메이션에 대해 pairwised test를 진행하였으며, 참가자는 두 영상 중 더 선호하는 것을 선택하였다. 다만 언어 비의존적 모델[24]의 경우 본 논문과는 다른 얼굴 모델을 사용하기 때문에 모델에서 오는 편향성을 방지하기 위해 첫 번째 실험에서는 3일 뒤에 별도로 평가를 진행하였으며, 두 번째 실험은 모두 하관만 노출하여 비교를 진행하였다.

실험 결과는 다음과 같다. 첫 번째 실험의 경우, oneway ANOVA 테스트를 통해 중립적인 음성(F-value = 11.175, p-value < 0.05)과 운율적 요소가 강하게 반영된 음성(F-value = 29.082, p-value < 0.05) 모두 입 모양 애니메이션 영상의 자연스러움을 판단할 때 모델의 차이가 유의미하게 영향을 미쳤음을 확인하였다. 결과를 보면, 제안하는 운율적 요소를 반영한 모델로 생성한 입 모양 애니메이션이 자연스럽다고 응답한 비율이 73%로 가장 높게 관찰되었다 (**Table 6**). 이를 5점 척도에 따라 점수를 계산해 각각의 음성 데이터에 대한 결과를 확인해보면, 모든 음성에 대해 해당 모델(평균 3.958)의 점수가 가장 높음을 알 수 있다 (**Table 7**). 또한, 동시조음 현상을 규칙 기반으로 처리하는 기존 논문[2](42%)과 비교했을 때도 본 논문(59%)에 따른 결과를 사람들이 더욱 자연스럽다고 느꼈음을 알 수 있다. 언어 비의존적 모델[24]의 경우, 본 실험에 사용한 애니메이션 영상에 대해 자연스럽다고 평가한 경우는 없었다.

Table 6: Evaluation results on naturalness for 6 speech animations. The evaluation on naturalness is expressed on a 5-point scale. −− means strong disagree and ++ means strong agree.

	−−	−	o	+	++	agree
Zhou, Yang, et al. 2018	65	29	6	0	0	0%
Kim, Sang-Wan, et al. 2005	2	27	29	40	2	42%
Ours without prosody	2	4	35	46	13	59%
Ours with prosody	0	8	19	42	31	73%

Table 7: Evaluation results on naturalness for 6 speech animations. 1-6 shows the evaluation results for audio 1-6.

	1	2	3	4	5	6	mean
Zhou, Yang, et al. 2018	1.125	1.625	1.125	1.625	2	1	1.417
Kim, Sang-Wan, et al. 2005	3	3	3.375	2.75	3.125	3.5	3.125
Ours without prosody	3.75	3.875	3.875	3	3.5	3.75	3.625
Ours with prosody	3.875	4.375	4.125	3.75	3.75	3.875	3.958

두 번째 실험에서도 본 논문의 운율적 요소를 반영한 모델이 다른 모델보다 선호도가 높게 나왔으며, 자세한 내용은 **Table 8**에서 확인할 수 있다. 대부분의 음성에서 기존의 연구[24, 2]보다 높은 선호도를 보였으며, 다만 화자가 천천히 크게 말하는 음성 6의 경우 앞선 첫 번째 실험에서 모델 중에 가장 높은 점수를 받았음에도 불구하고 선호도가 38%로 예외적으로 낮게 관찰되었다. 이는 두 입 모양 애니메이션을 동시에 비교하는 과정에서 강한 운율적 요소에 따른 과장된 움직임에 이질감을 느낀 것으로 보인다. 반면에 화자가 작고 빠르게 말하는 경우, 모두 75% 이상의 높은 선호도를 보인 것을 확인할 수 있다.

7. 결론 및 향후 연구

본 논문은 음성과 텍스트를 입력받아 한국어에 특화된 자연스러운 스피치 애니메이션 생성 기술을 제안하였다. 한국어에서 입술

Table 8: Results from the preference evaluation for 6 speech animations. 1-6 shows the evaluation results from audio 1-6.

	1	2	3	4	5	6	mean
Ours > Zhou, Yang, et al. 2018	100	100	100	100	100	100	100%
Ours > Kim, Sang-Wan, et al. 2005	100	100	88	88	75	38	81%
Ours > w/o prosody	50	50	75	75	75	50	63%

모양은 대부분 모음에 의해 결정된다는 특성을 반영하여 입술과 혀 쉐입을 분리한 viseme 모델을 제안하였고, 이를 바탕으로 동시조음 현상을 처리하는 규칙 기반의 동시조음 모델을 정의하였다. 이를 통해 기존의 연구에서 양순음을 제외한 자음에도 viseme 모델을 정의함으로써 입술 모양이 왜곡되거나, 반대로 생략함으로써 해당 음소의 특성이 누락되는 문제를 해결해 자연스러움을 향상시켰다. 또한, 음의 세기, 발화속도와 같은 운율적 요소에 따라 입 모양의 크기에 변화를 주어 같은 텍스트에 대해서도 음성의 특성에 맞는 역동적인 스피치 애니메이션을 생성하였다.

다만 본 연구는 한국어에 특화된 모델로 한국어를 인풋으로 받으며, 다른 언어나 숫자 처리는 불가능하다. 또한, 입 모양 생성에 한정되어 부수 모션과 표정 애니메이션은 생성하지 않는다. 마지막으로 운율적 요소가 강하게 반영된 음성은 입 모양 움직임이 과장되어 부자연스러움을 완전히 해소하기 어렵다는 한계가 있다. 따라서 향후 연구로는 인풋에 한글 외의 다른 문자가 포함되더라도 이를 처리할 수 있도록 전처리 단계에서 한글 표기로 바꾸는 작업을 추가할 계획이다. 더 나아가 입 모양에서 얼굴 전체로 다루는 영역을 확대하고, 실제 사람의 데이터를 접목해 입 모양의 변화가 자연스럽게 이루어지는 방법을 고안할 것이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.R0124-16-0002)

References

[1] 김탁훈, "애니메이션 캐릭터의 한국어 립싱크 연구: 영어권 애니메이션의 립싱크 기법을 기반으로," *만화애니메이션 연구*, pp. 97–114, 2008.

[2] S.-W. Kim, H. Lee, K.-H. Choi, and S.-Y. Park, "A talking head system for korean text," *World Academy of Science, Engineering and Technology*, vol. 50, 2005.

[3] 오현화, 김인철, 김동수, and 진성일, "한국어 모음 입술독해를 위한 시공간적 특징에 관한 연구," *한국음향학회지*, pp. 19–26, 2002.

[4] H.-J. Hyung, B.-K. Ahn, D. Choi, D. Lee, and D.-W. Lee, "Evaluation of a korean lip-sync system for an android robot," *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), IEEE*, pp. 78–82, 2016.

[5] 정일홍 and 김은지, "한국어 음소를 이용한 자연스러운 3d 립싱크 애니메이션," *한국디지털콘텐츠학회 논문지*, vol. 9, no. 2, pp. 331–339, 2008.

[6] 김태은 and 박유신, "한글 문자 입력에 따른 얼굴 에니메이션," *한국전자통신학회 논문지*, vol. 4, pp. 116–122, 2009.

[7] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, p. 127, 2016.

[8] Y.-C. Wang and R. T.-H. Tsai, "Rule-based korean grapheme to phoneme conversion using sound patterns," *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pp. 843–850, 2009.

[9] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, IEEE*, pp. 396–401, 2002.

[10] I. Albrecht, J. Haber, and H.-P. Seidel, "Automatic generation of non-verbal facial expressions from speech," *Advances in Modelling, Animation and Rendering, Springer, London*, pp. 283–293, 2002.

[11] J.-R. Park, C.-W. Choi, and M.-Y. Park, "Human-like fuzzy lip synchronization of 3d facial model based on speech speed," *Proceedings of the Korean Institute of Intelligent Systems Conference, Korean Institute of Intelligent Systems*, pp. 416–419, 2006.

[12] K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria," *Journal of Speech, Language, and Hearing Research*, 2004.

[13] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pp. 275–284, 2012.

[14] 신지영, *한국어의 말소리*. 박이정출판사, 2014.

[15] R. D. Kent and F. D. Minifie, "Coarticulation in recent speech production models," *Journal of phonetics*, vol. 5, no. 2, pp. 115–133, 1977.

[16] 이광희, 고우현, 지상훈, 남경태, and 이상무, "시청각 정보를 활용한 음성 오인식률 개선 알고리즘," *한국정밀공학회 학술발표대회 논문집*, pp. 341–342, 2010.

[17] 임성민, 구자현, and 김회린, "어텐션 기반 엔드투엔드 음성인식 시각화 분석," *말소리와 음성과학*, vol. 11, no. 1, pp. 41–49, 2019.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." *Interspeech*, pp. 498–502, 2017.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society*, 2011.

[20] 김종록, "외국인을 위한 한국어 동사 활용 사전 돌아보기," *한글*, no. 295, pp. 73–134, 2012.

[21] 임홍빈, "한국어의 불규칙 활용에 대하여," *韓國學究論文集*, no. 3, pp. 1–21, 2014.

[22] 양순임, "'ㅎ'불규칙용언의 표기 규정에 대한 고찰," *한민족어문학*, vol. 62, pp. 315–338, 2012.

[23] G. S. Turner and G. Weismer, "Characteristics of speaking rate in the dysarthria associated with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 6, pp. 1134–1144, 1993.

[24] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–10, 2018.

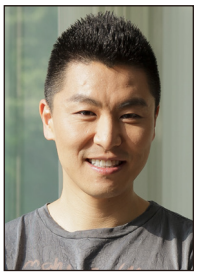
〈저자소개〉

장 민 정

- 2014~2018 고려대학교 국어국문학과/ 미디어학부 학사
- 2018~2020 한국과학기술원 문화기술대학원 석사
- 관심분야: Facial animation, Character animation
- https://orcid.org/0000-0002-7095-8375

정 선 진

- 2011~2015 아주대학교 미디어학부 학사
- 2015~2017 한국과학기술원 문화기술대학원 석사
- 2017~현재 한국과학기술원 문화기술대학원 박사과정
- 관심분야: Facial animation, Character animation
- https://orcid.org/0000-0001-6427-6258



노 준 용

- 2002년 University of Southern California Computer Science 박사
- 2003년~2006년 Rhythm and Hues Studio, Graphics Scientist
- 2006년~현재 카이스트 문화기술 대학원 교수
- 2011년~2014년 카이스트 석좌 교수
- 2016년~현재 카이스트 문화기술대학원 학과장
- 관심분야: 컴퓨터 그래픽스, 컴퓨터 비전, 얼굴 애니메이션, 캐릭터 애니메이션, VR/AR, 몰입형 디스플레이
- https://orcid.org/0000-0003-1925-3326