KIPS Trans. Softw. and Data Eng.
Vol.12, No.10 pp.461~470
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.10.461

A Study on Korean Speech Animation Generation Employing Deep Learning

Suk Chan Kang† · Dong Ju Kim††

ABSTRACT

While speech animation generation employing deep learning has been actively researched for English, there has been no prior work for Korean. Given the fact, this paper for the very first time employs supervised deep learning to generate Korean speech animation. By doing so, we find out the significant effect of deep learning being able to make speech animation research come down to speech recognition research which is the predominating technique. Also, we study the way to make best use of the effect for Korean speech animation generation. The effect can contribute to efficiently and efficaciously revitalizing the recently inactive Korean speech animation research, by clarifying the top priority research target. This paper performs this process: (i) it chooses blendshape animation technique, (ii) implements the deep-learning model in the master-servant pipeline of the automatic speech recognition (ASR) module and the facial action coding (FAC) module, (iii) makes Korean speech facial motion capture dataset, (iv) prepares two comparison deep learning models (one model adopts the English ASR module, the other model adopts the Korean ASR module, however both models adopt the same basic structure for their FAC modules), and (v) train the FAC modules of both models dependently on their ASR modules. The user study demonstrates that the model which adopts the Korean ASR module and dependently trains its FAC module (getting 4.2/5.0 points) generates decisively much more natural Korean speech animations than the model which adopts the English ASR module and dependently trains its FAC module (getting 2.7/5.0 points). The result confirms the aforementioned effect showing that the quality of the Korean speech animation comes down to the accuracy of Korean ASR.

Keywords : Speech Animation, Deep Learning, Viseme, Co-articulation, Blendshape

딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 고찰

강 석 찬† · 김 동 주††

요 약

딥러닝을 활용한 스피치 애니메이션 생성은 영어를 중심으로 활발하게 연구되어왔지만, 한국어에 관해서는 사례가 없었다. 이에, 본 논문은 최초로 지도 학습 딥러닝을 한국어 스피치 애니메이션 생성에 활용해 본다. 이 과정에서, 딥러닝이 스피치 애니메이션 연구를 그 지배적 기술인 음성 인식 연구로 귀결시킬 수 있는 중요한 효과를 발견하게 되어, 이 효과를 한국어 스피치 애니메이션 생성에 최대한 활용하는 방법을 고찰한다. 이 효과는 연구의 최우선 목표를 명확하게 하여, 근래에 들어 활발하지 않은 한국어 스피치 애니메이션 연구를 효과적이고 효율적으로 재활성화하는 데 기여할 수 있다. 본 논문은 다음 과정들을 수행한다: (i) 블렌드쉐입 애니메이션 기술을 선택하며, (ii) 딥러닝 모델을 음성 인식 모듈과 표정 코딩 모듈의 주종 관계 파이프라인으로 구현하고, (iii) 한국어 스피치 모션 캡처 dataset을 제작하며, (iv) 두 대조용 딥러닝 모델들을 준비하고 (한 모델은 영어 음성 인식 모듈을 채택하고, 다른 모델은 한국어 음성 인식 모듈을 채택하며, 두 모델이 동일한 기본 구조의 표정 코딩 모듈을 채택한다), (v) 두 모델의 표정 코딩 모듈을 음성 인식 모듈에 종속되게 학습시킨다. 유저 스터디 결과는, 한국어 음성 인식 모듈을 채택하여 표정 코딩 모듈을 종속적으로 학습시킨 모델 (4.2/5.0 점 획득)이, 영어 음성 인식 모듈을 채택하여 표정 코딩 모듈을 종속적으로 학습시킨 모델 (2.7/5.0 점 획득)에 비해 결정적으로 더 자연스러운 한국어 스피치 애니메이션을 생성함을 보여 주었다. 이 결과는 한국어 스피치 애니메이션의 품질이 한국어 음성 인식의 정확성으로 귀결됨을 보여 줌으로써 상기의 효과를 확인해준다.

키워드 : 스피치 애니메이션, 딥러닝, Viseme, 동시조음, 블렌드쉐입

※ 본 논문은 2021년~2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 ICT R&D 혁신 바우처 지원사업 기금으로 미디어젠 주식회사 주관하에 수행한 '대화형 아바타 개발을 위한 영어/한국어 음성과 동조된 얼굴 모션 합성 솔루션 개발(2021-0-01096)' 과제의 연구 결과임.
※ 본 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2022R1A6A1A03052954)이며, 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2023-00231158, 비전기술을 활용한 편물 검단 및 환편기 예지보전 원격제어 통합모니터링 플랫폼).
† 정 회 원 : 포항공과대학교 인공지능연구원 연구부 선임연구원
†† 비 회 원 : 포항공과대학교 인공지능연구원 연구부 연구부장
Manuscript Received : April 10, 2023
First Revision : August 7, 2023
Accepted : August 24, 2023
* Corresponding Author : Dong Ju Kim(kkb0320@postech.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 서 론

음성을 실제 사람처럼 립싱크하는 스피치 애니메이션 연구는 멀티미디어 및 가상 세계 기술이 고도로 발전하고 있는 오늘날 전 세계적으로 그 중요성이 지속적으로 증가하고 있다. 구체적인 활용 분야로는 메타버스 환경에서의 가상 인간 아바타, 또는 애니메이션 캐릭터, 게임, 영화 등의 제작으로 그 종류와 영역이 다양하다[1-14]. 그런데 스피치 애니메이션 연구는 대상 언어에 관련된 음성 인식, 표정 과학, 그리고 컴퓨터 그래픽스 등의 분야들이 유기적으로 결합 되어야 하는 까다로운 분야이기 때문에 영어를 중심으로는 활발한 학제적 연구가 수행되어왔다[2-5, 7, 12, 14-17].

영어 중심의 활발한 글로벌 연구 상황과 달리, 현재 한국어 스피치 애니메이션 연구는 2000년대 후반에 잠정 중단된 후[18, 19], 2020년에 규칙 기반 동시조음 모델을 이용한 연구[1]가 유일하게 후속 연구로서 발표된 상태이다. 그런데 기존 연구 방식들은 깊은 한국어 음운론 지식이 필요하다는 공통점이 있다[1, 18-24]. 따라서, 한국어 스피치 애니메이션 연구가 잠정 중단된 가장 큰 이유는 기존 연구 방식들의 한국어 음운론 지식 의존성이 연구자들에게 높은 진입장벽으로 작용하기 때문으로 보인다. 또한, 이 이유로 기존 연구 방식은 공용어인 영어와 달리 풍부한 글로벌 연구 인력들의 참여도 쉽지 않다. 이 문제를 해결할 수 있는 대안으로는 언어 비의존적 연구가 고려될 수 있지만, 아직 그 애니메이션 결과물의 품질이 만족스럽지 않은 실정이다[1]. 기존 연구 방식의 이러한 한계를 극복하는 또 다른 방법으로 딥러닝을 활용할 수 있지만, 영어를 중심으로 하는 글로벌 연구 추세와 달리 한국어 스피치 애니메이션 연구는 아직 그 사례가 없다.

따라서, 본 논문은 최초로 지도 학습 딥러닝을 활용해 한국어 스피치 애니메이션을 생성해 보면서, 근래에 들어 활발하지 않은 한국어 스피치 애니메이션 연구를 재활성화하는 효과적이고 효율적인 방향을 파악한다. 구체적으로 본 논문의 기여도는 다음과 같다. 우선, 딥러닝 기술의 활용이 스피치 애니메이션 연구를 그 지배적 기술인 음성 인식 연구로 귀결시키는 효과가 있음을 보임으로써 연구의 최우선 목표를 명확하게 한다. 이 효과는 딥러닝의 귀납적 추론 능력을 활용하여 자소(grapheme)와 표정 사이 함수 관계의 언어 의존성을 실험하여 확인한다. 그리고, 이 효과를 한국어 스피치 애니메이션 생성에 최대한 활용하는 방법도 고찰한다. 본 논문의 나머지 부분에서는 이 효과를 '음성 인식으로의 귀결 효과'로 지칭한다.

2. 배경 지식

2.1 스피치 애니메이션 생성

전통적으로 스피치 애니메이션 연구는 특정 대상 언어에서 음소에 대응하는 표준화된 입 모양 모델인 viseme을 정의하고, 이들이 시간적인 동적 조합에 따라 자연스럽게 연결될 수 있도록 하는 동시조음 모델을 구현한다. 즉, viseme들을 키프레이밍 하는 과정에서 동시조음 모델을 적용하여 연속적인 부드러운 애니메이션을 생성하게 된다. 따라서 일반적으로 스피치 애니메이션 연구에서 동시조음 모델은 viseme 정의를 포함한 동시조음 모델을 의미하는 상위개념 용어로도 사용된다.

2.2 중립 표정 스피치 애니메이션 타입

표준화된 중립 표정 스피치 애니메이션은 입력의 형태에 따라 특정 1인 화자의 음성만 입력받는 타입1, 특정 1인 화자의 음성과 보조 대본을 동시에 입력받는 타입2, 불특정 화자의 음성만 입력받는 타입3, 그리고 불특정 화자의 음성과 보조 대본을 동시에 입력받는 타입4로 구분할 수 있다.

음성만을 입력받는 타입1과 타입3은 시계열 음성 인식 정보를 추출하기 위해서 자체적으로 강제 정열 기능을 구현해야 한다. 한편, 음성과 보조 대본을 동시에 입력받는 타입2와 타입4는 시계열 음소를 추출하기 위해서 Povey et al.[25]와 같은 음성 인식 툴을 바탕으로 하는 외부 강제 정열 도구를 사용할 수 있다[26, 27]. 본 논문은 난이도가 높은 타입3 스피치 애니메이션을 생성하는 방식을 선택한다.

2.3 애니메이션 데이터 인코딩 방식

사람의 얼굴을 3D 애니메이션으로 렌더링하기 위한 인코딩 방식은 facial action coding system (FACS)으로 명명되며 크게 다음의 두 가지 방식으로 나눠볼 수 있다.

랜드마크 방식은 밀집된 많은 독립된 3D 점들의 집합으로 사람의 얼굴을 표현한다. 점들의 수만큼 표현의 자유도를 높일 수 있지만 그만큼 FACS 데이터가 방대해지는 단점이 있다. 또한 딥러닝으로 애니메이션을 추론할 때 나타나는 오류가 각 점들에 대해 독립적으로 발현될 수 있어서 기본적인 얼굴 형상조차 왜곡되어 추론될 수 있다. 가장 큰 단점은 서로 다르게 생긴 얼굴 모형들, 즉 아바타들 간의 동일 표정 전이가 용이하지 않다는 점이다.

블렌드쉐입 방식은 각각의 표정 구성 요소를 나타내는 비교적 소수의 키들을 기저 벡터로 미리 정의하고, 0과 1사이 값으로 정규화된 그 키 값들을 선형 조합하여 사람의 얼굴 표정을 벡터 공간으로 표현한다. 따라서 FACS 데이터를 최소화할 수 있는 장점이 있지만 그만큼 표현의 자유도가 높지 않을 수 있다. 미리 정의된 표정 구성 요소 키들을 사용하므로 딥러닝으로 애니메이션을 추론할 때 오류로 인해 발생할 수 있는 기본적인 얼굴 형상 왜곡에 면역성이 있다. 가장 큰 장점은 서로 다르게 생긴 얼굴 모형 아바타들 간의 동일 표정 전이가 비교적 용이하다는 점이다. 대표적인 애플의 ARkit 호환 블렌드쉐입 방식은 52개의 키들 만을 사용하여 얼굴 표정을 표현하는 FACS 데이터를 구성한다[28].

3D 랜더링 소프트웨어들은 일반적으로 블렌드쉐입을 FACS 데이터로 입력받아 스피치 애니메이션을 생성한다.

3. 관련 연구

3.1 Viseme 과 동시조음 모델

전통적으로 viseme은 고정된 음소와 정적으로 그에 해당하

는 입 모양을 대응시켜 정의한다. 그리고 스피치 애니메이션은 viseme들을 키프레이밍 하는 과정에서 동시조음 모델을 적용하여 연속적인 부드러운 얼굴 표정을 생성하여 제작된다. 이 과정에서 동시조음이 바로 인접한 음소뿐만 아니라 최대 5칸 떨어진 음소의 입 모양까지도 영향을 미친다는 연구 결과가 있었다[29]. 그리고 전통적인 정적인 viseme 모델을, 동시조음 현상을 고려하여 짧은 애니메이션 형태의 동적인 모델로 재정의하는 연구도 있었다[2].

3.2 딥러닝 스피치 애니메이션

귀납적인 딥러닝의 추론 능력을 활용하는 모든 딥러닝 스피치 애니메이션 생성 모델들은 큰 의미에서 Taylor et al.[3]에서 제안된 것과 같은 음성 인식 모듈과 표정 코딩 모듈이 결합된 구조를 가지게 된다. 또한, 표정 코딩 모듈의 구현 방식에 따라 다음과 같이 이분할 수 있다. 첫째는, 전통적인 스피치 애니메이션 제작 방식과 같이 연구자가 개입하여 정적인 viseme을 정의하고 그에 알맞은 동적인 동시조음 모델을 구현하는 과정을 그대로 표정 코딩 모듈에 적용하는 방식이다[4, 30]. 둘째는, 딥러닝의 귀납적 학습 능력으로 음성 인식과 표정 생성이 end-to-end 종단간 학습으로 자연스럽게 융합되도록 하여 viseme 과 동시조음 개념이 없는 표정 코딩 모듈을 구성하는 방식이다[3, 5, 6]. 즉, 후자의 경우 인간의 개입을 최소화할 수 있는 장점이 있다. 본 논문도 후자의 방식을 한국어 스피치 애니메이션 연구에 적용한다.

3.3 한국어 스피치 애니메이션

한국어 viseme 모델 연구는 영어 viseme 연구에 기반하여 한국어에 특화 시키거나[18, 20], 처음부터 한국어 음소에 맞게 자체 정의하는 시도로 발전해 왔다[18, 19, 21-23]. 한편, 한국어 동시조음 모델 연구는 음운 현상을 규칙으로 정의하고 예외에 대해서만 발음 사전을 만드는 규칙 기반 방법이 적합하다는 연구가 있었다[24]. 그러나 한국어 스피치 애니메이션 관련 연구는 2000년대 후반에 중단된 후[18, 19], 2020년에 규칙 기반 동시조음을 활용한 연구[1] 이후로 현재 다시 잠정 중단된 상황이다. 그리고 영어 중심의 글로벌 연구 추세와 달리, 딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 연구는 아직 사례가 없다.

4. 한국어 딥러닝 스피치 애니메이션

본 논문은 불특정 화자의 목소리를 보조 대본 입력 없이 입력받아 표준화된 중립 표정 스피치 애니메이션을 출력하는 기술을 사용한다. 이 절에서는 Fig. 1에 묘사된 스피치 애니메이션 생성의 각 과정들을 자세히 설명한다.

4.1 3D 렌더링 소프트웨어

딥러닝 네트워크가 추론하는 블렌드쉐입 값을 입력받아 한국어 스피치 애니메이션을 3D 렌더링하는 소프트웨어는 공개

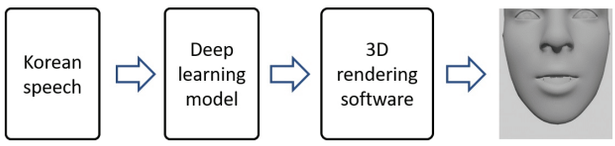

Fig. 1. Steps of Korean Speech Animation Generation in this Paper

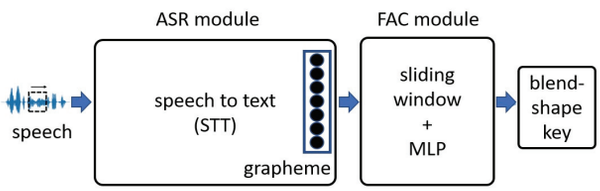

Fig. 2. Deep Learning Model Used in This Paper. It Comprises the Automatic Speech Recognition (ASR) Module and the Facial Action Coding (FAC) Module

소프트웨어인 블렌더[31]를 사용하였다. 블렌더는 렌더링과 더불어 스무딩 기능 등의 다양한 고급 기능을 제공하며, 이러한 기능들을 사용자의 커스텀 파이썬 plug-in 코드로 저장하여 실행시킬 수 있다. 본 논문도 연구 내용에 맞는 객체지향 파이썬 plug-in 코드를 작성하여 애니메이션을 생성하였다.

4.2 딥러닝 모델 개요

음성 인식으로의 귀결 효과를 확인하기 위해, 딥러닝 모델을 Fig. 2와 같이 핵심적 지배 기술인 '음성 인식 모듈 (ASR module)'과 이에 종속되는 '표정 코딩 모듈 (FAC module)'의 파이프라인 방식으로 구성한다. 이 방식은 Taylor et al.[3]이 제안하는 구조이며, 추론 결과에 대한 음성 인식 성능 요인 이외의 다른 요인의 영향을 원천적으로 배제한다. 즉, 음성 인식 모듈이 음성을 입력받아 시계열 자소를 추론하여 출력하면 표정 코딩 모듈은 다시 이 시계열 자소만을 입력받아 시계열 표정을 학습하고 추론한다. 따라서 이 구조의 궁극적 목적은 표정 코딩 모듈을 독립변수인 '표준화된 자소'와 종속변수인 '표준화된 표정' 간의 언어 의존적 딥러닝 회귀모델로 만드는 것이다.

4.3 딥러닝 파이프라인 1단계: 음성 인식 (ASR) 모듈

불특정 한국어 화자의 음성을 입력받아 강제 정렬된 다차원 시계열 자소 시퀀스를 추론하여 출력한다. 즉, 같은 내용을 말하는 서로 다른 음색을 가진 화자들의 말도 거의 동일한 시계열 자소 시퀀스로 대응되게 하는 '음성 표준화' 역할을 담당한다. 음성 인식 모듈을 구현하는 것은 많은 시간과 시행착오가 요구되므로 본 논문의 연구 범위를 벗어난다. 따라서 본 논문에서는 미리 학습시켜 speeech-to-text (STT) 기능을 구현한 공개 패키지 모델들을 사용하였다.

이처럼 음성 인식 모듈을 미리 학습된 STT 모델을 채택하여 구현하는 주요 장점으로는 (i) 노이즈 및 배경음악과 같은 음성이 아닌 신호에 영향받지 않는 면역성이 강하고, (ii) 따로

텍스트 대본을 음성과 같이 입력하지 않아도 시간에 대한 자소의 강제 정렬 효과를 자동으로 얻을 수 있으며, (iii) 특히 생성된 스피치 애니메이션이 무음성 기간에 스피치를 자연스럽게 멈추는 성능이 뛰어나다는 것이다.

본 논문의 초기 개발과정에서는 음성 표준화 모듈을 제작하기 위해 불특정 화자들의 음성이 특정 1인 화자의 음성으로 변환되는 음성변환 (voice conversion) 기술[5]을 채택하였으나, 고품질의 음성 표준화 목적으로 사용하기에는 부적절함을 발견하였다.

4.4 딥러닝 파이프라인 2단계: 표정 코딩 (FAC) 모듈

음성 인식 모듈이 추론하여 출력한 시계열 자소 시퀀스를 입력받아 Table 1이 나타내는 얼굴 하관에 해당하는 32차원의 시계열 블렌드쉐입 시퀀스를 추론하여 출력한다. 즉, 딥러닝이 실제 화자의 모션 캡처 스피치 표정을 학습하여 귀납적 추론을 통해 까다로운 viseme 모델과 동시 조음 모델을 대체하는 '표정 표준화' 역할을 담당한다. 따라서 이 모듈은 viseme을 정의하고 동시조음을 모델링 하기 위해 요구되는 음운론 지식의 높은 진입장벽을 연구자들이 우회할 수 있도록 해준다.

이 표정 코딩 모듈이 viseme 모델과 동시조음 모델을 대체하기 위해서는 매 순간만의 자소 입력을 고려한 추론이 아닌, 과거 자소 입력들까지 고려한 기억 기반 추론이 필요하다. 따라서 이 모듈은 기억 기능이 있는 순환 신경망(recurrent neural network, 즉 RNN)의 발전된 형태인 장단기 메모리 (long short-term memory, 즉 LSTM) 와 같은 정교한 신경망을 사용하여 구현하는 것이 일반적이다[3-5, 32]. 그런데 이들 순환 신경망들은 기본적으로 출력을 다시 입력에 포함 시키는 복잡한 피드백 방식의 구조를 사용하여 기억 기반 추론 기능을 구현한다.

하지만 본 논문은 이 모듈이 파이프라인 1단계 음성 인식

Table 1. Thirty Two Apple ARkit[28] Blendshape Keys Used in This Paper

Blendshape keys	Description
1. jawForward	forward movement of the lower jaw
2. jawLeft	leftward movement of the lower jaw
3. jawRight	rightward movement of the lower jaw
4. jawOpen	an opening of the lower jaw
5. mouthClose	closure of the lips independent of jaw position
6. mouthFunnel	contraction of both lips into an open shape
7. mouthPucker	contraction and compression of both closed lips
8. mouthLeft	leftward movement of both lips together
9. mouthRight	rightward movement of both lips together
10. mouthSmileLeft	upward movement of the left corner of the mouth
11. mouthSmileRight	upward movement of the right corner of the mouth
12. mouthFrownLeft	downward movement of the left corner of the mouth
13. mouthFrownRight	downward movement of the right corner of the mouth
14. mouthDimpleLeft	backward movement of the left corner of the mouth
15. mouthDimpleRight	backward movement of the right corner of the mouth
16. mouthStretchLeft	leftward movement of the left corner of the mouth
17. mouthStretchRight	rightward movement of the left corner of the mouth
18. mouthRollLower	movement of the lower lip toward the inside of the mouth
19. mouthRollUpper	movement of the upper lip toward the inside of the mouth
20. mouthShrugLower	outward movement of the lower lip
21. mouthShrugUpper	outward movement of the upper lip
22. mouthPressLeft	upward compression of the lower lip on the left side
23. mouthPressRight	upward compression of the lower lip on the right side
24. mouthLowerDownLeft	downward movement of the lower lip on the left side
25. mouthLowerDownRight	downward movement of the lower lip on the right side
26. mouthUpperUpLeft	upward movement of the upper lip on the left side
27. mouthUpperUpRight	upward movement of the upper lip on the right side
28. cheekPuff	outward movement of both cheeks
29. cheekSquintLeft	upward movement of the cheek around and below the left eye
30. cheekSquintRight	upward movement of the cheek around and below the right eye
31. noseSneerLeft	a raising of the left side of the nose around the nostril
32. noseSneerRight	a raising of the right side of the nose around the nostril

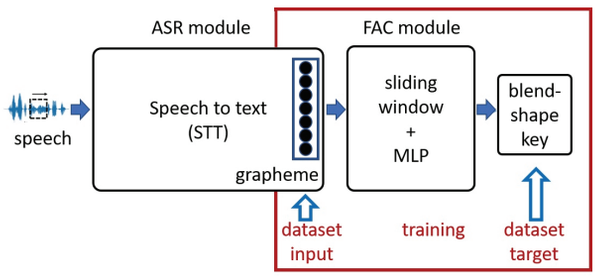

Fig. 3. Deep Learning Training in This Paper. Because the ASR Module is an Off-the-shelf Pre-trained One, Only the FAC Module is Trained (Dependently on the ASR Module).

모듈에 종속되는 모듈로서, 가장 기본적인 구조의 딥러닝 신경망의 보편적 근사화 (universal approximation)를 통해서도 한국어 동시조음 모델을 귀납적으로 구현할 수 있음을 증명하려 한다. 따라서 피드백 방식의 구조를 사용하는 순환 신경망 딥러닝 네트워크의 사용을 지양하였다. 대신, 시계열 자소 시퀀스를 입력으로 받아들이는 슬라이딩 윈도우 (sliding window) 순방향 다층 퍼셉트론 신경망 (multilayer perceptron, 즉 MLP) 을 사용하여 이 모듈을 구성하였다. 이 구성의 장점은 슬라이딩 윈도우의 길이 조절을 통해서 다층 퍼셉트론이 한정된 길이의 자소 시퀀스 입력에만 집중하여 시계열 블렌드쉐입 시퀀스를 학습하고 추론할 수 있도록 한다. 그리고, 총체적으로 순환 신경망에 비해 간단한 구조 덕택으로 짧은 학습 시간이 가능하여 빠른 딥러닝 최적화에도 도움이 된다.

4.5 딥러닝 스피치 애니메이션 Dataset

본 논문에서 사용하는 딥러닝 지도 학습용 dataset은 Fig. 3에 나타낸 것처럼 표정 코딩 모듈만을 학습시키는 dataset input과 dataset target의 형태로 두 단계를 통해서 제작된다. 1단계에서는 실제 화자의 스피치를 모션 캡처하여 가공되지 않은 원시 음성 wav 파일과 이와 동기된 시계열 블렌드쉐입 시퀀스를 수집한다. 원시 음성 wav 파일은 딥러닝 네트워크의 학습 및 추론의 성능 향상을 위해서 노이즈 제거와 음량의 정규화 과정을 거쳐 전처리하였다. 2단계에서는 전처리한 음성 wav 파일을 파이프라인 1단계의 음성 인식 모듈에 입력시켜 정열된 시계열 자소 시퀀스를 추론한다. 2단계에서 추론한 시계열 자소 시퀀스는 dataset input이 되며 1단계에서 수집한 시계열 블렌드쉐입 시퀀스는 dataset target이 된다.

4.6 딥러닝 지도 학습

본 논문에서의 딥러닝 학습은 4.5절에서 설명한 방식으로 파이프라인 2단계 표정 코딩 모듈에 대해서만 수행한다. 상술한 바와 같이 표정 코딩 모듈이 언어 의존적인 '자소-표정' 간 함수 관계를 학습하기 때문에 음성 인식 모듈에 종속되는 모델이 된다.

학습에 사용한 손실 (loss) 함수는 매 batch frame들에서의 참값 (ground truth) 블렌드쉐입 값들과 추론된 블렌드쉐입 값들의 mean absolute error (MAE)로서 다음 Equation (1) 로 표현된다.

$$Loss = \frac{1}{F \times B} \sum_{i=1}^{F} \| BSgt(i) - BSpred(i) \|_1 \quad (1)$$

BSgt(i) 는 i 번째 frame 의 참값 블렌드쉐입 키 벡터를 나타내며, BSpred(i) 는 추론된 i 번째 frame의 블렌드쉐입 키 벡터이다. 그리고 F는 하나의 학습 batch가 포함하는 frame 수를 의미하며, B는 블렌드쉐입 키 벡터의 크기이다.

5. 실험 및 고찰

이 절에서는 '자소-표정' 회귀모델의 정확성에 대한 언어 의존 실험을 통해서 딥러닝 활용이 스피치 애니메이션 연구를 그 지배적 기술인 음성 인식 연구로 귀결시키는 음성 인식으로의 귀결 효과를 확인한다.

실험은 대조용 딥러닝 모델 제작, 교차 검증, 유저 스터디의 3 단계로 진행한다. 이 절의 마지막에서는 딥러닝을 활용한 한국어 스피치 애니메이션 연구에서 음성 인식으로의 귀결 효과를 최대로 활용할 수 있는 방법도 고찰한다.

5.1 대조용 딥러닝 모델 제작

Fig. 2의 파이프라인 구조로, 한국어 음성 인식 성능에서 현저한 차이가 나는 두 한국어 스피치 애니메이션 생성용 딥러닝 모델들을 제작한다. 우선 영어 특화 음성 인식 모듈을 채택한 '비고도화된' 한국어 스피치 애니메이션 생성 모델인 Model-En-ASR을 제작한다. 그리고 한국어 특화 음성 인식 모듈을 채택한 '고도화된' 한국어 스피치 애니메이션 생성 모델인 Model-Ko-ASR을 제작한다.

두 모델의 파이프라인 모듈들은 Table 2에 정리하였다. 지배적 기술인 한국어 음성 인식 모듈을 고도화하면 표정 코딩 모듈도 그에 맞춰 종속적으로 고도화됨을 보이기 위한 목적으로, 두 대조용 모델이 동일한 기본적인 딥러닝 신경망 구조의 표정 코딩 모듈을 채택하게 하였다. 따라서, Model-Ko-ASR은 영어 스피치 애니메이션 생성에 적합한 딥러닝 모델인

Table 2. Pipeline Modules of the Two Comparison Deep Learning Models for Korean Speech Animation Generation. Their ASR Modules are Different, While Their FAC Modules have the Same Basic Structure

	ASR module	FAC module
Model-En-ASR	fine-tuned for **English** 32-dimensional grapheme	18-frame sliding window MLP
Model-Ko-ASR	fine-tuned for **Korean** 50-dimensional grapheme	18-frame sliding window MLP

Model-En-ASR의 영어 특화 음성 인식 모듈만을 한국어 특화 음성 인식 모듈로 업그레이드한 구조라고 생각할 수 있다. 두 모델의 상세한 설명은 다음과 같다.

1) 음성 인식 모듈 (ASR modules)

Meta의 wav2vec2 기술을 사용해 미리 학습되어 hugging face[33]에 공개된 다음 패키지들을 사용한다.

Model-En-ASR에 채택된 영어 특화 음성 인식 모듈은 facebook/wav2vec2-large-960h-lv60-self 패키지이다. 규모가 큰 모델로서 레이블링 없는 dataset을 사용하여 자기 지도 학습으로 사전 학습한 후 다시 960시간 동안 레이블링 된 영어 데이터에 대해 심화 지도 학습하였다. 입력은 16kHz로 샘플링 된 wav 파일을 받아들이며 출력은 50 frames/s 속도로 강제 정렬된 32차원의 영어 자소 소프트맥스 추론 값을 내보낸다.

Model-Ko-ASR에 채택된 한국어 특화 음성 인식 모듈은 fleek/wav2vec-large-xlsr-korean 패키지이다. 마찬가지로 입력은 16kHz로 샘플링된 wav 파일을 받아들이며 출력은 50 frames/s 속도로 강제 정렬된 50차원의 한글 자소 소프트맥스 추론 값을 내보낸다.

2) 표정 코딩 모듈 (FAC modules)

두 대조용 모델이 동일한 기본적인 딥러닝 신경망 구조를 사용하여 이 모듈을 만든다. 즉, 4.4절에서 설명한 바와 같이, 음성 인식 모듈에서 출력되는 시계열 자소 시퀀스를 입력으로 받아 시계열 블렌드쉐입 시퀀스를 추론하여 출력하는 슬라이딩 윈도우 순방향 다층 퍼셉트론 신경망 구조이다. 두 대조용 모델이 동일하게 18 frame의 시계열 자소 시퀀스 슬라이딩 윈도우 크기를 사용한다.

3) 딥러닝 학습

표정 코딩 모듈 학습용 dataset 제작은 5.2절의 교차 검증 결과에서 얻은 정확성과 일관성이 확보될 때까지 중복되지 않는 다양한 문장들의 발화 모션 캡처 데이터를 추가해가며 분량을 증가시켰다. 결과적으로 이 조건을 만족시키는 3시간 분량의 '한국어 자소-한국어 표정' dataset과 3시간 분량의 '영어 자소-한국어 표정' dataset 두 가지를 제작하여 학습에 사용하였다. 비슷한 목적을 위해서 Zhou et al.[4]은 모션 캡처 대신 애니메이터가 정교한 수작업으로 제작한 애니메이션으로 1시간 분량의 dataset을 제작하여 표정 코딩 모듈에 해당하는 딥러닝 네트워크를 학습시켰다.

스피치 모션 캡처 프로그램으로는 애플 iOS 앱인 Face Cap[34]을 사용하여 Table 1의 애플 ARkit[28]와 호환되는 시계열 블렌드쉐입 시퀀스를 추출했다. Dataset 제작에 사용한 다양한 스피치 문장들은 한국어 음성 인식 학습용 dataset인 zeroth-korean corpus[35]에서 사용한 문장들을 활용하였다.

5.2 교차 검증

Equation (1)을 사용한 5-겹 교차 검증으로 두 대조용 모델의 한국어 스피치 애니메이션 블렌드쉐입 추론 손실 값을 정량적으로 측정하고 비교하였다. 즉, 제작된 dataset을 5개의 split들로 분할 하여 테스트용 split 1개, 지도 학습용 split 4개로 조합되는 도합 5 경우에 대해 반복적으로 학습과 테스트를 수행하였다. 따라서, 항상 전체 dataset을 학습 및 테스트에 활용하면서도 테스트 split들이 지도 학습에 포함되지 않은 상태로 두 대조용 모델들의 블렌드쉐입 추론 성능을 검사하였다.

Table 3. Blendshape Loss Values of Model-Ko-ASR and Model-En-ASR, Obtained with the 5-fold Cross Validation. A Lower Loss Value can be a Measure for the Better Model

test set	Model-Ko-ASR (%)	Model-En-ASR (%)
split-1	2.45	2.42
split-2	2.35	2.38
split-3	2.32	2.36
split-4	2.44	2.48
split-5	2.17	2.32

교차 검증을 통한 딥러닝 스피치 애니메이션 모델의 테스트 split별 추론 손실 값은 Table 3에 나타내었다. Model-Ko-ASR과 Model-En-ASR가 모두 2.17 % ~ 2.48 % 내의 우수한 낮은 테스트 추론 손실 값을 보여 주었다. 따라서, 정량적인 블렌드쉐입 추론 손실 값 측정만으로는 두 대조용 딥러닝 모델들의 추론 성능 우열을 변별력 있게 파악하기는 힘들었다. 하지만 교차 검증을 통하여 (i) 한국어 스피치 애니메이션 모션 캡처 dataset이 정확하고 일관성 있게 알맞은 분량으로 제작되었는지, (ii) 딥러닝 네트워크 구성이 dataset을 과적합과 과소적합 없이 학습하여 알맞게 스피치 애니메이션을 추론할 수 있는지, (iii) Model-Ko-ASR과 Model-En-ASR이 한국어 스피치 애니메이션을 생성하는 대조용 모델들로 사용되기에 적합한지 등을 종합적으로 확인할 수 있었다.

5.3 유저 스터디

추론된 블렌드쉐입 값의 정확성을 정량적 비교 실험하는 5.2절의 교차 검증은 두 대조용 딥러닝 모델이 생성하는 최종 한국어 스피치 애니메이션의 자연스러움을 평가하기에 변별력이 부족했다. 따라서, 특징적인 한국어 테스트 스피치들을 준비하고, 딥러닝 모델을 활용해 블렌드쉐입을 추론할 뿐만 아니라, 이를 렌더링하여 스피치 애니메이션을 최종 생성해서 그 자연스러움을 정성적으로 평가하였다.

평가는 Jang et al.[1]의 방식을 좀 더 강화하여 진행하였다. 즉 3명을 추가하여 총 11명의 유저 그룹을 구성하였고 테스트 스피치도 6개를 추가한 총 12개를 사용하였다. 유저 그룹은 연구실 내의 전문 인공지능 연구원들로 구성하였다. Table 4는 이들의 성별, 나이, 그리고 고향에 관한 정보를 나타낸다. Table 5는 사용한 테스트 스피치들을 열거한다. 이들 중 6개의 s2-s7 테스트 스피치들은 Jang et al.[1]에서 사용한 테스트 스피치들이며, 나머지 6개는 본 논문을 위해 새로 추가하였다.

Table 4. Participants Information of the User Study

user #	gender	age	home town region
1	male	26	seoul
2	male	28	daegu
3	male	28	daegu
4	female	24	incheon
5	male	26	seoul
6	male	29	jeollanam-do
7	female	25	incheon
8	female	26	jeju
9	male	28	daegu
10	female	30	gyeonggi-do
11	female	28	gyeonggi-do

Table 5. Test Korean speeches used for the user study.

speech	content
s1	아름다운 한국어.
s2	여기에서 오늘 꼭 알아야 할 내용이 있습니다.
s3	사람들을 돕고 유유히 사라지는 슈퍼맨처럼 한마디 말도 없이 공사 현장으로 돌아갔다.
s4	그러나 지식은 그 종류와 양이 무한하다.
s5	새들도 짐승들도 착한 나무꾼의 친구들이라 나무꾼은 조금도 외롭지 않았어요.
s6	소는 어질고 순해서 어린아이들에게도 순순히 따르고 말도 잘 들었다.
s7	슬기로움은 우연하게 얻어지는게 아니거든.
s8	아름다운 한국어와 아름다운 한글이 우리는 자랑스럽습니다.
s9	고통을 참아야 하며 나아가 고통을 즐길 줄 알아야해.
s10	마음만을 가지고 있어서는 안된다. 반드시 실천하여야 한다.
s11	할 수 있다고 생각하든, 그렇지 않다고 생각하든 간에, 당신 생각이 옳다.
s12	밤사이 어려웠던 문제가 한잠 푹 자고 나면 아침에 해결되어있는 일은 흔한 경험이다.

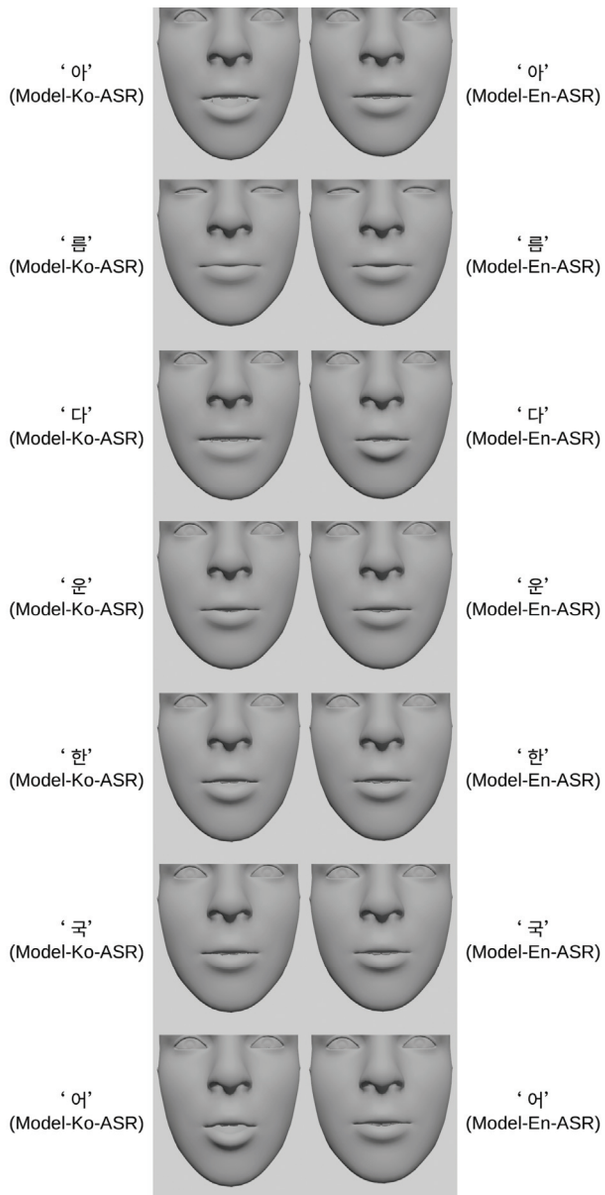

Fig. 4. User Study Speech Animation Snapshot of Korean Speech s1 in Table 5 (generated by Model-Ko-ASR and Model-En-ASR).

유저들은 평가시 Model-Ko-ASR과 Model-En-ASR이 생성한 대조용 한국어 스피치 애니메이션들을 Jang et al.[1]과 동일한 방식인 3회 시청후 5.0 만점으로 익명을 사용하여 채점하였다. Fig. 4는 유저 스터디에 사용된 첫번째 애니메이션 동영상 s1 ‘아름다운 한국어’의 스냅샷을 보여준다.

한국어 스피치 애니메이션의 자연스러움에 대한 유저 스터디 결과는 Table 6에 나타내었다. Model-Ko-ASR이 Model-En-ASR에 비해 압도적으로 높은 점수를 얻어, 훨씬 더 자연스러운 애니메이션을 생성한다는 평가를 받았다. Table 7은 Model-Ko-ASR과 Model-En-ASR 두 모델이 생성한 대조용 한국어 스피치 애니메이션들에 대한 유저 스터디 평가 결과를 이분법적 선호도별로 다시 정리한 결과이다. 실험에 사용된 12개의 테스트 스피치들에 대한 11명의 선택은 평균 86%의 높은 선호도로 Model-Ko-ASR이 Model-En-ASR에 비해 더 자연스러운 한국어 스피치 애니메이션을 생성한다는 결과를 보여 주었다. 여기서 얻게 된 사실은 5.2절의 교차 검증에서 두 대조용 모델이 비슷한 수준의 우수한 손실 함수 값을 정량적으로 얻었다 하더라도, 자연스러움에 대한 평가는 사람의 육안으로 정성 평가하는 결과에서 많은 차이를 나타낸다는 점이다.

유저 스터디 결과에서 주목할 사항은 다음과 같다. 두 대조용 모델들 모두 추론 결과에 대한 음성 인식 성능 요인 이외의 다른 요인의 영향을 원천적으로 배제한다. 그리고 두 대조용 모델들이 한국어 음성 인식 모듈의 성능이 현저하게 차이나지만, 표정 코딩 모듈은 동일한 가장 간단한 딥러닝 신경망 구조를 가진다. 또한 두 대조용 모델들의 표정 코딩 모듈을 단순히 음성 인식 모듈에 종속되게 학습시켰다. 그러나 결과적으로 이 모든 과정을 통하여 Model-Ko-ASR이 Model-En-

Table 6. Naturalness Scores (out of 5.0) of Korean Test Speech Animations of Speeches s1-s12 in Table 5.

test speech	Model-Ko-ASR (out of 5.00)	Model-En-ASR (out of 5.00)
s1	4.00	2.91
s2	3.91	2.64
s3	3.91	2.55
s4	4.36	2.73
s5	3.82	2.64
s6	3.82	2.73
s7	4.09	3.36
s8	4.55	2.18
s9	3.91	2.36
s10	4.64	2.45
s11	4.55	3.00
s12	4.91	2.91
mean	4.20	2.70

Table 7. Naturalness Preference Study (Model-Ko-ASR over Model-En-ASR).

Test speech	Model-Ko-ASR 〉 Model-En-ASR (%)
s1	82
s2	91
s3	91
s4	91
s5	73
s6	73
s7	64
s8	100
s9	91
s10	100
s11	91
s12	91
mean	86

ASR에 비해 훨씬 더 자연스러운 한국어 스피치 애니메이션을 생성한다. 즉, 표준화된 '한국어 자소'와 표준화된 '한국어 표정' 간의 표정 코딩 딥러닝 회귀모델이, 표준화된 '영어 자소'와 표준화된 '한국어 표정' 간의 표정 코딩 딥러닝 회귀모델보다 정확도가 높게 평가되었다. 따라서, 음성 인식 성능이 딥러닝을 활용하여 생성한 스피치 애니메이션의 자연스러움을 지배적으로 결정하는 음성 인식으로의 귀결 효과를 확인할 수 있었다.

5.4 고찰

음성 인식으로의 귀결 효과는 표준화된 자소와 그에 해당하는 표준화된 표정이 각각 독립변수 종속변수로서 언어 의존적인 시계열 딥러닝 회귀모델을 형성한다는 것을 의미한다. 한국어 스피치 애니메이션에 대한 이 회귀모델의 정확도를 높이기 위해서는 다음과 같이 음성 인식 모듈, 모션 캡처 data, 그리고 표정 코딩 모듈을 최적화하는 것이 필요하다.

우선 제작하는 한국어 스피치 애니메이션의 목표 스피치 도메인 범위를 정확하게 결정한다. 그리고 핵심적 지배 기술인 한국어 음성 인식 모듈의 성능을 목표 스피치 도메인에 대해 집중적으로 고도화하여 음성의 표준화 품질을 높인다. 이는 음성 인식 기술이 하나의 언어에 대해서도 대단히 광범위한 연구 분야이기 때문이다. 한편, 학습되지 않은 볼특정 화자의 음성을 인식하여 표준화된 음소로 추론하기 위해서는 최대한 많은 화자에 대한 학습이 필요하다.

그리고 스피치 모션 캡처 data의 표정 coverage와 정확성 및 일관성도 한국어 음운론을 바탕으로 목표 스피치 도메인에 대해 집중적으로 고도화하여 표정의 표준화 품질을 높인다. 본 논문은 딥러닝을 활용하여 까다로운 한국어 음운론 지식을 대체하는 것을 목표로 시작하였지만, 효율적이고 효과적인 모션 캡처 data를 제작하기 위해서는 결국 깊은 한국어 음운론 지식이 상보적으로 필수적임을 알게 되었다. 그리고 귀납적 추론의 특성상 모든 경우의 coverage를 갖는 data를 수집하는 것이 불가능함도 경험하였다. 유저 테스트 준비 과정에서도 한정된 모션 캡처 data의 스피치 coverage 때문에 모든 스피치에 대해 만족스러운 애니메이션 추론을 얻을 수는 없었다.

아울러 모션 캡처 data를 학습하여 표준화된 자소로부터 표준화된 중립 표정을 추론하기 위해서는 정확하고 일관된 표정을 유지할 수 있는 숙련된 1인 화자에 대한 학습이 바람직하다. 음성 인식을 통해 이미 음성의 표준화가 이루어졌기 때문에 다수 화자에 대한 학습은 오히려 일관되지 않은 화자들의 표정 간에 간섭을 발생시켜 의도하지 않은 표정이 추론될 수 있기 때문이다. 현실적으로도, 학습에 필요한 화자의 수와 그들의 숙련도는 정확한 모션 캡처 data를 제작하는 비용과 시간을 결정하는 가장 중요한 요소가 된다.

덧붙혀, 모션 캡처 data 제작은 다음 사항들도 고려하여 세심하게 제작하여야 한다. (i) 화자는 무음성 구간에서 입을 다문 기본 표정을 유지한다. (ii) 음성과 블렌드쉐입 간의 최상의 시계열 동기화 구현을 위해 짧은 모션 캡처 길이를 유지한다. (iii) 모션 캡처 data 자체가, 딥러닝이 귀납적으로 viseme의 정의와 동시조음 모델을 학습해 가는 교재 역할을 하기 때문에 일상적인 발화 표정보다 조금 더 과장된 특징적인 표정을 유지한다. (iv) 무음성 기간에 스피치 애니메이션이 확실하게 입을 다물게 만들기 위해서 무음성 기간 동안 입을 다문 모션 캡처 data도 추가한다. 고도화를 위한 딥러닝 학습과 추론의 반복과정에서 위의 고려 사항들이 추론된 애니메이션의 품질 향상에 큰 영향을 끼치는 것을 관찰할 수 있었다.

마지막으로 표정 코딩 모듈은 간단한 딥러닝 네트워크 구조로 구현한다. 그리고 특별한 고도화 과정 없이, 최적화된 스피치 모션 캡처 data를 이용하여 최적화된 한국어 음성 인식 모듈에 종속적으로 학습시킨다. 본 논문의 실험방식과 같이

음성 인식 모듈들만을 다양하게 교체하며 표정 코딩 모듈을 그에 따라 종속적으로 학습시키는 것은 음성 인식으로의 귀결 효과를 활용하여 스피치 애니메이션 생성을 효율적으로 최적화하는 좋은 방법이다.

6. 결 론

한국어 스피치 애니메이션 연구가 근래에 잠정 중단된 이유는 기존 연구 방식들의 한국어 음운론 지식 의존성이 연구자들에게 높은 진입장벽으로 작용하며, 또한 국제어가 아니기 때문에 풍부한 글로벌 연구가 힘들다는 사실과 관련 있을 가능성이 크다. 본 논문은 이러한 현실을 극복하기 위해 최신 딥러닝 기술의 귀납적 추론 능력을 한국어 스피치 애니메이션 생성에 적용해 보았다. 이를 통해 딥러닝의 도입이 스피치 애니메이션 연구를 음성 인식 연구로 귀결시키는 효과가 있음을 보임으로써 연구의 최우선 목표를 명확하게 하였고, 이 효과를 한국어 스피치 애니메이션 생성에 최대한 활용할 수 있는 방법도 고찰하였다. 본 논문에서 발견한 이 효과가 근래에 들어 활발하지 않은 한국어 스피치 애니메이션 연구를 재활성화하는데 기여할 수 있기를 기대한다.

References

[1] M. Jang, S. Jung, and J. Noh, “Speech animation synthesis based on a Korean co-articulation model,” *Journal of the Korea Computer Graphics Society*, Vol.26, No.3, pp.49-59, 2020.

[2] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, “Dynamic units of visual speech,” in *Proceedings of the ACM SIGGRAPH/Eurographics Conference on Computer Animation*, Lausanne, Switzerland, pp.275-284, 2012.

[3] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, “A deep learning approach for generalized speech animation,” *ACM Transactions on Graphics*, Vol.36, No.4, pp.1-11, 2017.

[4] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, “Visemenet: Audio-driven animator-centric speech animation,” *ACM Transactions on Graphics*, Vol.37, No.161, pp.1-10, 2018.

[5] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “MakeltTalk: Speaker-aware talking-head animation,” *ACM Transactions on Graphics*, Vol.39, No.6, pp.1-15, 2020.

[6] H. X. Pham, Y. Wang, and V. Pavlovic, “End-to-end learning for 3D facial animation from speech,” In *Proceedings of the ACM International Conference on Multimodal Interaction*, New York, pp.361-365, 2018.

[7] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, “Capture, learning, and synthesis of 3D speaking styles,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp.10093-10103, 2019.

[8] A. Nagendran, S. Compton, W. C. Follette, A. Golenchenko, A. Compton, and J. Grizou, “Avatar led interventions in the metaverse reveal that interpersonal effectiveness can be measured, predicted, and improved,” *Scientific Reports*, Vol.12, Iss.1, Article No.21892, 2022.

[9] Speech Graphics, Clients [Internet], https://www.speech-graphics.com/

[10] NVIDIA, Omniverse Audio2Face [Internet], https://www.nvidia.com/en-us/omniverse/apps/audio2face/

[11] NEURAL SYNC, Wav2Lip [Internet], https://www.neuralsyncai.com

[12] T. Ezzat, G. Geiger, and T. Poggio, “Trainable videorealistic speech animation,” *ACM Transactions on Graphics*, Vol.21, Iss.3, pp.388-398, 2002.

[13] F. Shaw and B. Theobald, “Expressive modulation of neutral visual speech,” in *IEEE MultiMedia*, Vol.23, Iss.4, pp.68-78, 2016.

[14] A. Richard, M. Zollhofer, Y. Wen, F. Torre, and Y. Sheikh, “MeshTalk: 3D face animation from speech using cross-modality disentanglement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp.1153-1162, 2021.

[15] L. Xie, L. Wang, and S. Yang, “Visual speech animation,” in Handbook of Human Motion, Springer, Cham, pp. 2115-2144. 2018.

[16] F. I. Parke, “A parametric model of human faces,” PhD thesis, University of Utah, 1974.

[17] Face the FACS, Facial Expressions in Art, Science, and Technology [Internet], https://melindaozel.com

[18] S. W. Kim, H. Lee, K. H. Choi, and S. Y. Park, “A talking head system for Korean text,” *International Journal of Electrical and Computer Engineering*, Vol.3, No.2, pp. 167-170, 2009.

[19] T. E. Kim and Y. S. Park, “Facial animation generation by Korean text input,” *The Journal of The Korea Institute of Electronic Communication Sciences*, Vol.4, No.2, pp. 116-122, 2009.

[20] T. Kim, “A study on Korean lip-sync for animation characters - based on lip-sync technique in English-speaking animations,” *Cartoon and Animation Studies*, No.13, pp. 97-114, 2008.

[21] H. H. Oh, I. C. Kim, D. S. Kim, and S. I. Chien, "A study on spatio-temporal features for Korean vowel lipreading," *The Journal of the Acoustical Society of Korea*, Vol.21, No.1, pp.19-26, 2002.
[22] 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," In *Proceedings of the IEEE International Conference on Ubiquitous Robots and Ambient Intelligence*, Xian, China, pp.78-82, 2016.
[23] I. H. Jung and E. Kim, "Natural 3D lip-synch animation based on Korean phonemic data," *Journal of Digital Contents Society*, Vol.9, No.2, pp.331-339, 2008.
[24] Y.-C. Wang and R. T.-H. Tsai, "Rule-based Korean grapheme to phoneme conversion using sound patterns," in *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, Vol.2, pp.843-850, 2009.
[25] D. Povey et al., "The kaldi speech recognition toolkit," in *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii, US, pp.1-4, 2011.
[26] S. Lim, J. Goo, and H. Kim, "Visual analysis of attention-based end-to-end speech recognition," *Phonetics and Speech Sciences*, Vol.11, No.1, pp.41-49, 2019.
[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Proceedings of the International Speech Communication Association*, Stockholm, Sweden, pp.498-502, 2017.
[28] Apple Inc., ARFaceAnchor.BlendShapeLocation [Internet], https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation
[29] R. D. Kent and F. D. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics*, Vol.5, No.2, pp.115-133, 1977.
[30] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics*, Vol.35, No.4, pp.1-11, 2016.
[31] Blender Online Community, Blender - a 3D modeling and rendering package [Internet], http://www.blender.org
[32] B. Fan, L. Wang, F. K. Soong, and L. Xie. "Photo-real talking head with deep bidirectional LSTM," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Australia, pp.4884-4888, 2015.
[33] Hugging Face, Facebook Models [Internet], https://huggingface.co/facebook
[34] Apple App Store, Face Cap - Motion Capture [Internet] https://apps.apple.com/us/app/face-cap-motion-capture/id1373155478
[35] OpenSLR, Zeroth-Korean [Internet], http://www.openslr.org/40/

강 석 찬

https://orcid.org/0000-0001-9118-2580
e-mail : sukchan.kang@postech.ac.kr
2000년 포항공과대학교 전자전기공학과 (학사)
2000년 ~ 2003년 삼성전자 반도체 System LSI사업부 연구원
2006년 Carnegie Mellon University Electrical & Computer Engineering(석사)
2019년 Georgia Institute of Technology Electrical & Computer Engineering(박사)
2019년 ~ 현 재 포항공과대학교 인공지능연구원 연구부 선임연구원
관심분야 : Computer Systems, Deep Learning, Machine Learning

김 동 주

https://orcid.org/0009-0009-6950-4200
e-mail : kkb0320@postech.ac.kr
2010년 성균관대학교 전기전자컴퓨터공학과 (박사)
2011년 ~ 2015년 대구경북과학기술원 IT 융합연구부 선임연구원
2015년 ~ 2016년 동서대학교 컴퓨터공학부 조교수
2016년 ~ 현 재 포항공과대학교 인공지능연구원 연구부 연구부장
관심분야 : Computer Vision, Face Recongnition, Deep Learning