Journal of KIIT. Vol. 24, No. 2, pp. 161-168, Feb. 28, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.2.161

인공지능을 활용한 프레젠테이션 비디오의 화자 음성 및 얼굴모사 영상과 동기화 방법

이재만*, 김선종**

Video Synchronization Method for Presentations in AI-Generated Speaker Voice and Facial Reconstruction

Jae-Man Lee*, Seon-Jong Kim**

이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음

요 약

인공지능(AI) 기술은 화자의 음성과 얼굴을 모사하는 기술뿐만 아니라, 자동화된 프레젠테이션 생성에도 적용되고 있다. 기존 연구들은 텍스트와 이미지를 기반으로 슬라이드를 자동 생성하는 데 초점을 맞추었으나, 발표자의 음성과 얼굴모사를 프레젠테이션 슬라이드와 실시간으로 동기화하는 연구는 부족하였다. 본 논문에서는 사전에 학습된 화자의 음성과 얼굴을 갖는 합성영상을 생성하고, 이를 프레젠테이션 슬라이드와 동기화하는 프레젠테이션 시스템을 제안한다. 본 시스템은 입력 텍스트 내 특수문자를 통해 특정 단어가 발화되는 시점에 슬라이드 텍스트에 효과를 적용하는 방식으로, 실험 1에서 평균 282.17ms, 실험 2에서 평균 316.39ms의 지연시간으로 동기화를 구현하였다.

Abstract

Artificial Intelligence (AI) technology is being applied not only to voice and face synthesis but also to automated presentation generation. While existing studies have focused on automatically generating slides based on text and images, research on the real-time synchronization of a presenter's synthesized voice and face with presentation slides has been limited. This paper proposes a presentation system that generates synthetic videos with a pre-trained speaker's voice and face, and synchronizes them with presentation slides. The system applies effects to slide text at the moment when specific words are spoken, using special characters in the input text as triggers, achieving voice-visual synchronization with an average latency of 282.17ms in Experiment 1 and 316.39ms in Experiment 2.

Keywords

presentation automation, multimodal interaction, real-time synchronization, speech synthesis, facial synthesis

* 부산대학교 IT응용공학과
- ORCID: https://orcid.org/0000-0002-9685-3870
** 부산대학교 IT응용공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-2070-290X

· Received: Nov. 03, 2025, Revised: Jan. 29, 2026, Accepted: Feb. 01, 2026

· Corresponding Author: Seon-Jong Kim
Dept. of IT Engineering, Pusan National University, Korea
Tel.: +82-55-350-5413, Email: ksj329@pusan.ac.kr

Ⅰ. 서 론

최근 인공지능(AI) 기술의 발전은 음성 및 얼굴 모사뿐만 아니라, 자동화된 프레젠테이션 생성에도 혁신적인 변화를 가져오고 있다. 특히, 화자의 음성과 얼굴 특징을 학습한 모델을 활용하면, 실제 화자의 특성을 반영한 음성과 자연스러운 입술 동기화를 생성할 수 있다. 이러한 AI 기반 음성 및 얼굴 모사 기술은 프레젠테이션과 같은 비주얼 콘텐츠에서 몰입감 있는 사용자 경험을 제공하는 핵심 기술로 자리 잡고 있다[1].

기존의 프레젠테이션 자동화 연구들은 주로 텍스트 및 이미지 데이터를 분석하여 슬라이드를 자동 생성하는 데 초점을 맞추고 있었다. 그러나 이러한 방식은 발표자의 개성을 반영하기 어렵고, 음성과 시각적 요소 간의 유기적인 연결이 부족하다는 한계가 있다. 또한, 기존 시스템들은 입력 데이터를 기반으로 결과를 자동 생성하는 방식으로, 사용자가 특정 시점에 효과를 적용하거나 발표 흐름을 직접 제어하기 어렵다는 문제가 있다.

본 연구에서는 이러한 문제를 해결하기 위해, 화자의 음성과 얼굴을 모사한 AI 생성 영상을 프레젠테이션 슬라이드와 실시간으로 동기화하는 시스템을 제안한다. 이 시스템은 텍스트 입력을 통해 화자의 음성과 얼굴 영상을 생성하고, 이를 프레젠테이션 슬라이드와 동기화하며, 중요 텍스트에 실시간 강조 효과를 적용한다. 비교적 단순한 편집이 요구되는 프레젠테이션 영상 제작에 적합하며, 기존 수작업 편집 방식보다 제작 속도를 향상시킬 수 있다.

Ⅱ. 관련 연구

AI 기반의 음성 및 얼굴모사 기술을 프레젠테이션에 접목시키는 연구들이 활발히 진행되고 있다. Glow-TTS와 HiFi-GAN과 같은 생성적 음성 합성 모델들은 고품질 음성을 합성할 뿐만 아니라, 화자의 입술 동기화와 음성 동기화를 자연스럽게 처리하여 비주얼 콘텐츠에서 더욱 몰입감 있는 사용자 경험을 제공하고 있다[2][3]. 또한, 대화형 AI와 음성 인식 기술을 결합하여 가상 면접 시스템을 구축하고, 사용자가 실시간으로 상호작용할 수 있는 몰입감 있는 인터페이스를 제시한 연구들이 있다[4]. GeneFace++는 오디오 데이터를 기반으로 3D 화자의 입술 동기화 및 음성 합성을 실시간으로 처리하는 시스템을 통해 더욱 자연스러운 얼굴모사 및 음성을 제공하여 비주얼 콘텐츠에서 몰입감을 높이고 있다[5].

프레젠테이션 슬라이드 자동 생성 분야에서는 텍스트와 이미지를 결합한 연구들이 주를 이루고 있다. 대규모 언어 모델(LLM, Large Language Model)을 활용하여 텍스트 및 이미지 데이터를 결합한 다단계 접근 방식은 슬라이드 생성 과정에서 더욱 자연스럽고 효율적인 결과를 만들어낸다[6]. 이와 더불어, 컴퓨테이셔널 노트북(Computational notebooks)의 개요를 기반으로 슬라이드를 자동으로 생성하는 연구도 진행되었다[7]. AI 기반 시스템을 통해 맞춤형 오디오 콘텐츠를 제공하여, 슬라이드에 동기화된 오디오를 통해 더욱 개인화된 발표 환경을 구현하는 연구도 제시되었다[8]. 또한, 과학 문서에서 핵심 내용을 추출하여 슬라이드를 자동 생성하는 DOC2PPT와 같은 연구는 발표 준비 시간을 단축하는 데 기여하였다[9].

그러나 이러한 기존 연구들은 주로 텍스트와 이미지를 기반으로 한 프레젠테이션 자동 생성에 중점을 두고 있으며, 발표자의 음성과 얼굴모사 기술을 동기화한 시스템은 상대적으로 부족하다. 또한, 기존 시스템들은 입력 데이터를 기반으로 AI가 자동으로 결과를 생성하는 방식인 반면, 사용자가 특수문자를 통해 효과 시점과 대상을 직접 지정하는 방식의 연구는 거의 없다. 본 연구는 사용자 의도 기반의 실시간 제어 가능성과 음성-얼굴-슬라이드 간 동기화에 초점을 둔다는 점에서 기존 연구와 차별성을 갖는다.

Ⅲ. 합성영상을 이용한 프레젠테이션 시스템

3.1 제안하는 프레젠테이션 시스템

그림 1은 본 논문에서 제안하는 인공지능 기반 프레젠테이션 시스템 응용을 나타낸 것이다. 먼저, 화자의 개인적 특성이 반영된 데이터로 학습된 인공지능 모델(TTS)은 주어진 텍스트에 따라 합성 음성을 생성한다. 이후, 얼굴을 모사하는 인공지능 모

델(Face synthesis)은 생성된 음성과 동기화된 합성영상을 생성하며, 이를 통해 프레젠테이션 영상을 제작할 수 있다.

그림 1. 제안하는 프레젠테이션 시스템
Fig. 1. Proposed presentation system

생성된 합성영상은 프레젠테이션 생성 시스템을 통해 재생되며, 프레젠테이션 슬라이드의 진행 흐름에 맞춰 자동으로 동기화되어 수행되도록 하였다. 특히, 발화 내용과 시각 요소 간의 세부적인 동기화를 구현하기 위해 본 논문에서는 입력 텍스트 내 특정 단어에 특수문자를 부여하고 해당 단어의 발화 시점과 슬라이드 내 시각 효과 시점을 일치시키는 트리거로 사용한다. 따라서, 기존 AI 기반 프레젠테이션 생성 시스템은 주로 거시적인 슬라이드 전환 동기화에 머무르는 반면, 본 시스템은 사용자가 특수문자를 통해 동기화 시점과 대상을 직접 지정하는 방식이다. 따라서 기존 방식과는 목적과 메커니즘이 근본적으로 다르며 본 연구는 사용자 의도를 반영하여 음성-얼굴-슬라이드 요소 간의 유기적인 동기화를 구현하는 데 초점을 둔다.

3.2 사용된 인공지능 모델 및 학습

그림 2에서 확인할 수 있듯이, Glow-TTS 내부의 Text encoder와 Duration predictor는 실시간 추론에 필수적인 병렬 처리를 지원하고, HiFi-GAN의 MRF 기반 Generator는 생성된 Mel-spectrogram을 고품질의 오디오 파형으로 복원하는 역할을 수행한다.

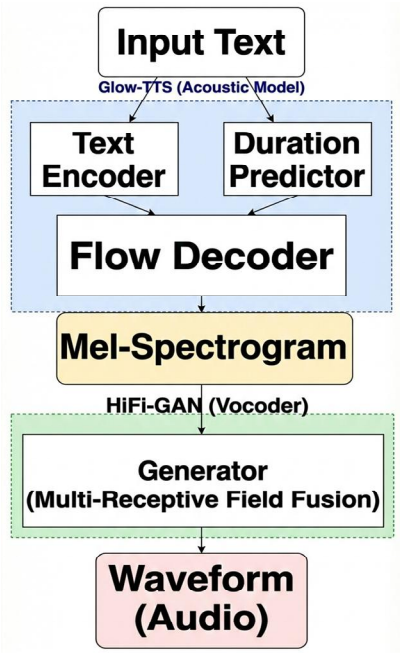


그림 2. 음성합성(TTS) 처리 과정
Fig. 2. Speech synthesis process

음성 생성을 위한 인공지능은 Glow-TTS와 HiFi-GAN 모델을 사용하여 고품질의 합성 음성을 생성하였다. 이 과정에서 약 2,500개의 문장으로 구성된 음성 데이터셋을 구축하여 해당 데이터셋을 기반으로 모델을 학습시켰다. 학습 과정에서는 화자의 음색과 억양, 발음 등의 특성을 최대한 반영할 수 있도록 세심하게 데이터 전처리 작업을 진행하였다. Glow-TTS는 입력 텍스트를 Mel-spectrogram의 변환 과정에서 더 자연스럽고 다채로운 음성을 생성할 수 있도록 설계되었으며, HiFi-GAN은 Glow-TTS에서 변환된 Mel-spectrogram을 재생이 가능한 웨이브폼 오디오 생성에 특화된 모델로, 합성된 음성의 품질을 더욱 향상하였다.

얼굴 합성 인공지능 모델은 GeneFace++를 선택하여, 화자의 얼굴 표정 및 자연스러운 움직임을 생성하였다. GeneFace++는 딥신경망을 사용하여 얼굴의 근육 움직임과 표정을 세밀하게 예측하고, 이를 실시간으로 동기화할 수 있도록 설계되었다. 특히, 이 모델은 얼굴 합성에 있어서 매우 정교한 디테일을 구현할 수 있어, 프레젠테이션에서 화자의 감정선과 표정 변화를 자연스럽게 표현할 수 있다.

그림 3은 음성과 얼굴 합성을 결합하는 과정을 표현하였고 앞서 입력된 텍스트로 합성된 음성에서 음성 특징들을 추출하고 정보를 분석한다. 그리고 음성 특징을 기반으로 얼굴의 랜드마크 68개를 예측하게 되는데 3DMM(3D morphable model)을 활용하여 얼굴 표현을 매핑하여 예측된 얼굴 특징을 바탕으로 비디오 프레임을 생성하고 NeRF(Neural Radiance Fields)와 GAN(Generative Adversarial Network)를 사용하여 원본 스타일을 유지하면서 자연스럽고 사실적인 얼굴을 합성하게 된다. 마지막으로 생성된 얼굴 프레임들을 원본 배경에 맞추어 자연스럽게 렌더링하고 입력한 음성과 결합한 비디오를 생성한다. 또한, 별도의 명시가 없는 한 모델들의 하이퍼파라미터는 기본 설정을 사용하였다.

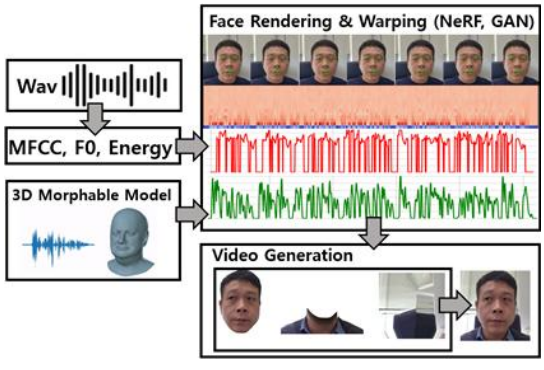


그림 3. 얼굴합성 처리 과정
Fig. 3. Face synthesis process

3.3 슬라이드와 합성영상의 생성

본 시스템은 사전에 준비된 프레젠테이션 파일(PowerPoint)을 입력으로 받아 해당 파일 내의 슬라이드를 분석한다. 슬라이드 내의 텍스트들에 대한 주요 정보를 추출하는데, 이때 텍스트의 위치, 크기, 글씨체와 같은 스타일적인 요소들을 라이브러리를 통해 식별하여 읽어내며 이 텍스트 정보는 시스템 내에서 동기화를 위한 후속 작업에 사용된다.

특히, 텍스트 내에서 특수문자가 쌍으로 표시된 부분이 발견될 경우, 해당 텍스트는 시스템이 사전에 정의해 놓은 텍스트 효과를 적용할 수 있는 대상으로 기억해 둔다. 예를 들어, 텍스트에 포함된 특수문자 쌍을 통해 특정 스타일이나 색상, 크기를 자동으로 인식하고, 해당 효과를 슬라이드 내의 동일한 텍스트에 적용한다. 이러한 과정은 슬라이드마다 일관되게 처리된다.

이 모든 과정이 완료된 후, 사용자는 프레젠테이션 시스템을 통해 합성된 영상이 슬라이드의 특정 영역에 포함되어 재생되는 동안 슬라이드에 나타난 텍스트 효과가 음성 재생 시점에 맞춰 적용되는 것을 확인할 수 있다.

3.4 텍스트와 합성영상의 동기화

그림 4는 텍스트와 합성한 영상 간의 동기화 과정을 시각적으로 표현한 것이다. 먼저, 합성영상에서 음성 영역을 추출한 후, 음성 내에서 텍스트 간 공백을 기준으로 시간을 구간별로 나눈다. 구간별로 나눌 때 글자 수를 가중치로 사용하며 이렇게 분할된 각 텍스트 구간의 시작 시점과 종료 시점을 계산해 텍스트별 표시 시간이 정해진다. 텍스트 내의 단어들은 공백을 기준으로 구분되며, 이때 분할된 단어의 총 개수를 N, i번째 단어의 글자 수를 $L_i\,(i=1,...,N)$으로 정의한다. 이에 따라 전체 텍스트 글자 수 합L_{total}은 식 (1)과 같이 개별 단어 글자 수의 총합으로 산출된다.

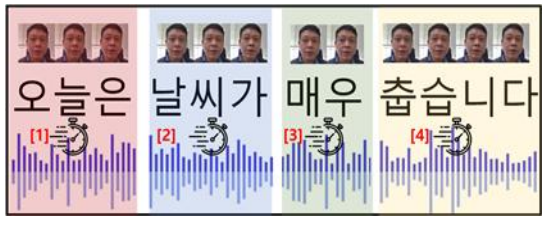


그림 4. 텍스트와 영상의 동기화
Fig. 4. Synchronization of text and video

$$L_{total}=\sum_{i=1}^{N}L_i \tag{1}$$

또한, 합성된 전체 음성의 재생 시간을 T라고 할 때, 각 단어의 표시 시간 D_i는 전체 시간 T를 글자 수 비율에 따라 배분하여 식 (2)와 같이 계산한다. 이를 바탕으로 i번째 단어의 시작 시점 S_i와 종료 시점 E_i는 식 (3), (4)와 같이 순차적으로 결정된다.

$$D_i = T \cdot \frac{L_i}{\Sigma_{k=1}^{N} L_k} \tag{2}$$

$$S_i = \begin{cases} 0 & (i=1) \\ \sum_{k=1}^{i-1} D_k & (i \geq 2) \end{cases} \tag{3}$$

$$E_i = S_i + D_i \tag{4}$$

이후, 시간 정보를 활용하여 슬라이드에서 특정 텍스트가 정해진 시간에 맞춰 효과를 적용하도록 설정한다. 특히, 입력된 텍스트 내의 마크업을 감지하면 해당 시점에 맞춰 텍스트 효과가 자동으로 활성화된다. 이러한 방식으로 합성영상과 슬라이드가 자연스럽게 동기화되며 이는 합성영상이 입력된 텍스트를 기반으로 동일한 음성을 생성하기 때문에 가능하다.

3.5 슬라이드 텍스트와 합성영상의 동기화

입력 문장에 포함된 단어 순서를 슬라이드 전체 텍스트에서 좌->우, 상->하의 순서로 탐색하여 순차적으로 정합시키는 알고리즘을 사용하였다. 먼저, 그림 5(a) 문장의 단어들을 식 (5)와 같이 정의한다.

그림 5. 슬라이드 텍스트와 합성영상의 동기화
Fig. 5. Synchronization of slide text and synthesized video

$$W = (w_1, w, \cdots, w_N) \tag{5}$$

또한, 슬라이드 그림 5(b) 내의 모든 텍스트 객체를 화면상 위치 기준(좌측 상단 -> 우측 하단)으로 정렬한 후, 이를 단일 단어 순서로 변환하여 식 (6)으로 정의한다.

$$S = (s_1, s_2, \ldots, s_M) \tag{6}$$

단어 매칭은 W의 순서를 보존하면서 각 단어 w_1의 매칭 위치는 식 (7)로 결정한다.

$$p_1 = \min jverts_j = w_1 \tag{7}$$

이후 두 번째 단어부터는 이전 단어의 매칭 위치 $p_i - 1$ 이후에서만 탐색하도록 제안하여 식 (8)의 재귀적 형태로 매칭 위치를 정의하였다.

$$p_i = \min jvertj > p_{i-1}, s_j = w_i, i = 2,3,\ldots,N \tag{8}$$

최종적으로 단어 순서 W와 슬라이드 내 등장 위치 p_i 간의 대응 관계는 식 (9)의 정합 형태로 표현된다.

$$M = (w_1, p_1), (w_2, p_2), \ldots, (w_N, p_N) \tag{9}$$

이 과정은 그림에서와 같이 그림 5(a)의 문장을 구성하는 단어들이 슬라이드 그림 5(b) 상의 텍스트 요소 중 동일 단어가 등장하는 위치와 순차적으로 대응하며, 결과적으로 문장과 슬라이드 간의 단어 수준 정렬을 제공한다.

IV. 실험 및 결과

4.1 텍스트 효과

표 1. 특수문자를 쌍으로 감싼 텍스트 효과 목록
Table 1. List of text effects enclosed in special character pairs

Special character	Color	Font specification	Type style
!	Red	Arial (24pt)	Underline
?	Blue	Calibri (20pt)	Italic
#	Green	Times New Roman (18pt)	Italic
*	Purple	Verdana (22pt)	Bold
$	Orange	Georgia (26pt)	Underline, Bold
&	Brown	Courier New (28pt)	Italic, Bold
%	Gray	Tahoma (30pt)	Underline, Italic, Bold

표 1은 특수문자의 쌍에 따라 적용되는 텍스트의 색상, 크기 등의 속성 효과를 나타낸 것으로써 이러한 효과들은 다양한 조합으로 설정할 수 있으며, 합성영상에서 슬라이드를 설명할 때 특정 단어가 언급되면 해당 효과가 슬라이드에 동적으로 적용된다. 합성된 음성이 진행됨에 따라, 관련된 텍스트 효과들이 실시간으로 나타나 슬라이드의 시각적 요소를 강조하며 이를 통해 음성과 슬라이드 내용이 동기화된다.

4.2 동기화 실험

그림 6은 합성영상과 텍스트의 동기화를 위한 실험에 사용한 텍스트이다. 텍스트에는 효과를 프레젠테이션 시스템에서 변환할 수 있는 표 1의 특수문자들이 포함되어 있다. 합성영상이 만들어질 때는 이 특수문자는 분리되어 텍스트만 입력되고 음성 및 합성영상을 생성하게 된다. 특수문자의 분리 과정을 통해 최종적으로 프레젠테이션 시스템에서 슬라이드를 재생할 때, 특수문자는 텍스트에 적용된 시각적 효과를 처리하는 데 중요한 역할을 하게 된다. 합성영상과 텍스트의 동기화가 완료된 후, 각 슬라이드에 포함된 텍스트는 이미 정의된 특수문자에 따라 효과를 순차적으로 적용하여 표시된다.

!기술!들을 하나의 #시스템# $통합$이 목표이며 개발된 기술들을 통합하기 위해 #모듈화# 및 &입력& !텍스트!에 따른 #음성합성#과 쓰리디 캐릭터의 &자동합성& 스크립트의 개발 그리고 최종적으로 입력 텍스트에 따른 나만의 합성 음성과 쓰리디 얼굴 합성 !영상!의 %생성%을 통해 *가상* !콘텐츠!에 응용합니다.

그림 6. 실험 텍스트 1
Fig. 6. Experiment text 1

표 2은 텍스트와 슬라이드 결과의 시간 차이를 보여주고 있다. 표에서 시간은 분리된 텍스트에서 특수문자를 포함하고 있는 텍스트가 인지하고 효과의 적용이 완료되는 시점까지의 지연시간이다. 이것은 프레젠테이션 시스템이 구동하는 PC의 성능에 따라 달라질 수 있는 부분이며 특수문자로 감싼 텍스트의 음성 발화 시점과 슬라이드 효과 적용 시점 간 지연시간은 평균 282.17ms로 측정되었다.

표 2. 텍스트와 슬라이드 효과 적용에 시간 차이
Table 2. Delay between speech and slide effects

Seq	Text	Delay(ms)
1	!기술!들을	380.43
2	#시스템#	285.26
3	$통합$이	262.06
4	#모듈화#	268.31
5	&입력&	261.57
6	!텍스트!에	308.99
7	#음성합성#과	248.60
8	&자동합성&	246.83
9	!영상!의	303.79
10	%생성%을	309.20
11	*가상*	250.56
12	!콘텐츠!에	260.45
Avg	-	282.17

1차!년도!에서는 ?음성? $합성$에 대한 &기술&로써 %인공지능% ?모델?의 *학습*이 필요하고 추가적으로 $데이터$ 세트에 따라서 &텍스트& 기반의 *나만에* 쓰리디#애니메이션#의 생성을 목표로 한다.

그림 7. 실험 텍스트 2
Fig. 7. Experiment text 2

표 3. 실험 2의 음성 텍스트와 슬라이드 효과 적용에 시간 차이
Table 3. Delay between speech and slide effects in experiment 2

Seq	Text	Delay(ms)
1	일차!년도!에서는	420.52
2	?음성?	299.11
3	$합성$에	305.89
4	&기술&로써	303.26
5	%인공지능%	304.19
6	?모델?의	301.24
7	*학습*이	315.23
8	$데이터$	303.06
9	&텍스트&	308.65
10	*나만에*	316.18
11	#애니메이션#의	313.67
12	?생성?을	305.73
Avg	-	316.39

그림 7는 합성영상의 동기화를 위해 추가로 실험을 한 텍스트이며 결과를 표 3에 나타내었다. 표 3의 지연시간은 평균 316.39ms로 두 실험의 평균 지연시간은 비슷하게 나타난다. 여기에 추가로 지연시

간을 증가 또는 감소시간을 적용하게 되면 특수문자로 감싼 텍스트를 읽는 음성이 발현될 때 텍스트에 적용되는 시간을 일률적으로 적용할 수 있다.

4.3 실험 결과

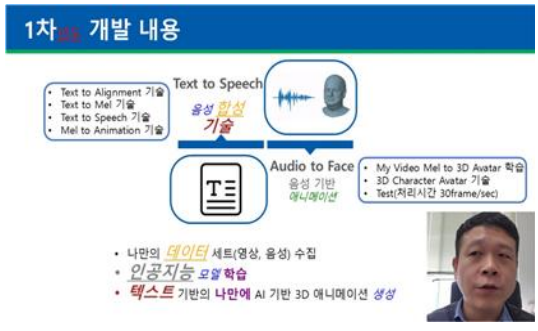
그림 8과 9은 실험 1, 2에서 생성된 텍스트 효과 적용 결과를 보여준다. 그림 8(a)는 주어진 PowerPoint 슬라이드에서 각 텍스트에 효과가 적용된 결과를 확대하여 표시한 것이다. 그림 8(b)는 프레젠테이션 시스템에서 슬라이드와 합성영상을 재생하는 화면으로, 모든 텍스트 효과가 적용된 예시를 나타낸다.

(a) 문장에 텍스트 효과가 적용된 모습
(a) Text with applied effects

(b) 효과가 적용된 슬라이드
(b) Slide with applied effects

그림 8. 실험 1의 모든 텍스트 효과를 적용한 슬라이드 예시 (a), (b)
Fig. 8. Example slide with all text effects applied in experiment 1 (a), (b)

(a) 문장에 텍스트 효과가 적용된 모습
(a) Text with applied effects



(b) 효과가 적용된 슬라이드
(b) Slide with applied effects

그림 9. 실험 2의 모든 텍스트 효과를 적용한 슬라이드 예시 (a), (b)
Fig. 9. Example slide with all text effects applied in experiment 2 (a), (b)

이미 적용된 텍스트 효과는 유지할 수도 있고, 다음 텍스트 효과가 적용될 때 이전 효과를 제거할 수도 있다. 이는 사용자가 선택할 수 있도록 설정이 가능하다. 본 시스템은 슬라이드 텍스트 중 강조가 필요한 부분을 특수문자로 표시하고, 합성영상 재생 중 해당 텍스트가 음성으로 발화되는 시점에 효과를 적용하는 방식으로 동작한다. 실험의 재생 영상은 [10]에서 확인할 수 있다.

합성영상은 일반적으로 슬라이드의 우측하단에 배치되며, 텍스트 효과는 합성영상의 재생 시간에 따라 왼쪽상단에서 우측하단으로 순차적으로 적용된다.

V. 결 론

본 논문은 인공지능 기반의 화자 음성과 얼굴모사 기술을 활용하여 프레젠테이션 슬라이드와 합성영상을 실시간으로 동기화하는 프레젠테이션 시스템을 제안하였다. 기존의 연구들은 음성 합성 및 슬라이드 자동 생성 기술에 초점을 맞추었으나, 본 논문에서는 사용자가 사전에 제작된 프레젠테이션에 음성과 동기화된 합성영상을 추가하고, 특정 텍스트에 특수효과를 적용할 수 있는 프레젠테이션 시스템을 개발하였다.

실험 결과, 본 시스템은 텍스트와 영상 간 동기화 지연시간을 250~400ms 범위 내에서 조정 가능하며, PC별 예측 지연시간을 측정하여 Offset 값으로

활용함으로써 동기화 정확도를 향상할 수 있다. 특수문자를 통한 효과가 음성 재생 시점에 맞춰 적용되어 프레젠테이션 영상 제작 효율성을 기대할 수 있고 사용자 입력 텍스트에 맞춰 음성과 얼굴 영상을 동기화함으로써 프레젠테이션 시스템을 통한 동적인 콘텐츠 생성을 가능하게 하였다.

References

[1] J. Lee and S. Kim, "Development of a Text-based Personalized Content Creation System Using Deep Learning", Journal of the Korean Institute of Information Technology, Vol. 21, No. 12, pp. 67-75, Nov. 2023. https://doi.org/10.14801/jkiit.2023.21.12.67.

[2] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search", arXiv:2005.11129, May 2020. https://doi.org/10.48550/arXiv.2005.11129.

[3] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", arXiv:2010.05646, Oct. 2020. https://doi.org/10.48550/arXiv.2010.05646.

[4] C. Yoon, S. Yang, J. Park, J. Si, Y. Jung, and S. Kim, "Metaverse Virtual Interview Platform Leveraging Generative AI and Speech Recognition", Journal of the Korean Institute of Information Technology, Vol. 22, No. 6, pp. 163-173, Jun. 2024. https://doi.org/10.14801/jkiit.2024.22.6.163.

[5] Z. Ye, J. He, Z. Jiang, R. Huang, J. Huang, J. Liu, Y. Ren, X. Yin, Z. Ma, and Z. Zhao, "GeneFace++: Real-time High-fidelity Lip Sync for 3D Talking Faces from Audio", arXiv:2305.00787, May 2023. https://doi.org/10.48550/arXiv.2305.00787.

[6] X. Zeng, Y. Wang, J. Zhang, and J. Yang, "Enhancing Presentation Slide Generation by LLMs with a Multi-Staged End-to-End Approach", arXiv:2406.06556, Jun. 2024. https://doi.org/10.48550/arXiv.2406.06556.

[7] L. Zhang, Z. Liu, and H. Zhao, "OutlineSpark: Igniting AI-powered Presentation Slides Creation from Computational Notebooks through Outlines", arXiv:2403.09121, Mar. 2024. https://doi.org/10.48550/arXiv.2403.09121.

[8] M. Mansoor, S. Chandar, and R. Srinath, "AI based Presentation Creator With Customized Audio Content Delivery", arXiv:2106.14213, Jun. 2021. https://doi.org/10.48550/arXiv.2106.14213.

[9] T.-J. Fu, W. Y. Wang, D. McDuff, and Y. Song, "DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents", arXiv:2101.11796, Jan. 2021. https://doi.org/10.48550/arXiv.2101.11796.

[10] https://youtube.com/shorts/ceKGRgB1IoM?feature=share. [accessed: Dec. 25, 2025]

저자소개

이 재 만 (Jae-Man Lee)

2011년 8월 : 부산대학교 바이오정보전자전공(공학사)
2014년 2월 : 부산대학교 IT응용공학과(공학석사)
2026년 2월 : 부산대학교 IT응용공학과(공학박사)
관심분야 : 신호 및 영상처리 머신/딥러닝

김 선 종 (Seon-Jong Kim)

1996년 8월 : 경북대학교 전자공학과(공학박사)
1995년 2월 ~ 1997년 2월 : 순천제일대학 제어계측과 전임강사
1997년 3월 ~ 현재 : 부산대학교 IT응용공학과 교수
관심분야 : 신호 및 영상처리, 머신/딥러닝, VR/AR, 스마트 카메라