(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0077821
(43) 공개일자 2023년06월02일

(51) 국제특허분류(Int. Cl.)
G06F 16/70 (2019.01) *G06F 40/232* (2020.01)
G10L 15/02 (2006.01) *G10L 15/08* (2006.01)
G10L 15/16 (2006.01) *G10L 15/26* (2006.01)
G10L 21/0208 (2013.01) *G10L 21/0232* (2013.01)
H04N 21/43 (2011.01) *H04N 21/488* (2011.01)
H04N 5/278 (2006.01)
(52) CPC특허분류
G06F 16/70 (2019.01)
G06F 40/232 (2020.01)
(21) 출원번호 10-2021-0164825
(22) 출원일자 2021년11월25일
심사청구일자 2021년11월25일

(71) 출원인
동서대학교 산학협력단
부산광역시 사상구 주례로 47(주례동, 동서대학교)
(72) 발명자
박승민
부산 사상구 주례로 93
이승용
부산 동래구 여고북로33번길 9
(뒷면에 계속)
(74) 대리인
김종석

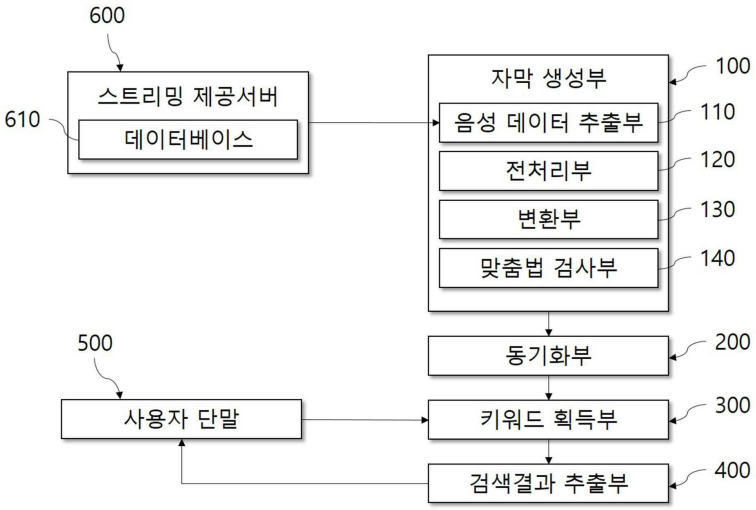
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **인공지능 기술기반 스트리밍 영상 검색 시스템 및 방법**

(57) 요 약

본 발명은 인공지능 기술기반 스트리밍 영상 검색 시스템 및 방법에 관한 것으로, 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막을 생성하는 자막 생성부, 상기 다수 개의 영상 각각의 타임라인을 기준으로 장면(Scene)과 상기 자막을 동기화 하는 동기화부, 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상을 제공할 수 있도록 사용자 단말로부터 키워드를 입력받는 키워드 획득부 및 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)을 검색결과로 추출하는 검색결과 추출부를 포함하고, 상기 검색결과 추출부는, 상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시되는 것을 특징으로 한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 15/02 (2013.01)
G10L 15/16 (2013.01)
G10L 15/26 (2013.01)
G10L 21/0208 (2013.01)
G10L 21/0232 (2013.01)
H04N 21/4307 (2022.08)
H04N 21/4884 (2013.01)
H04N 5/278 (2013.01)
G10L 2021/02085 (2013.01)
(72) 발명자
서부원
부산 수영구 망미번영로55번길 111-1
안지현
서울 도봉구 도봉로110라길 70-14
박상욱
부산 해운대구 해운대해변로 163

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126326
과제번호	2019-0-01817-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신창의인재양성
연구과제명	SW중심대학(동서대)
기 여 율	1/1
과제수행기관명	동서대학교 산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막을 생성하는 자막 생성부;

상기 다수 개의 영상 각각의 타임라인을 기준으로 장면(Scene)과 상기 자막을 동기화 하는 동기화부;

상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상을 제공할 수 있도록 사용자 단말로부터 키워드를 입력받는 키워드 획득부; 및

상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)을 검색결과로 추출하는 검색결과 추출부;를 포함하고,

상기 검색결과 추출부는,

상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시되는 것을 특징으로 하는 인공지능 기술기반 스트리밍 영상 검색 시스템.

청구항 2

제 1항에 있어서,

상기 자막 생성부는,

상기 다수 개의 영상으로부터 음성 데이터를 추출하는 음성 데이터 추출부;

상기 음성 데이터에서 잡음을 분리하고 특정 주파수에 맞게 전처리하는 전처리부;

전처리된 상기 음성 데이터를 음성 텍스트 변환(Speech-to-Text; STT) 기술을 사용하여 텍스트로 변환하는 변환부; 및

상기 텍스트에 대한 맞춤법을 검사한 후 띄어쓰기 또는 오탈자를 정정하고, 상기 텍스트에서 마침표를 기준으로 문장을 구분하는 맞춤법 검사부;를 포함하는 것을 특징으로 하는 인공지능 기술기반 스트리밍 영상 검색 시스템.

청구항 3

제 1항에 있어서,

상기 검색결과 추출부는,

상기 자막 생성부로부터 자막이 생성되지 않을 경우 비지도 학습이 가능하도록 k-평균 알고리즘을 이용하여 상기 키워드와 유사한 하나 이상의 장면(Scene)을 검색결과로 추출하는 것을 특징으로 하는 인공지능 기술기반 스트리밍 영상 검색 시스템.

청구항 4

제 1항에 있어서,

상기 변환부는,

하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하고, 기 저장된 단어에 따라 문장을 단어 단위로 분리하는 인코더(Encoder); 및

하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하고, 콘텍스트(Context) 벡터를 은닉 상태로 하고 상기 전처리부로부터 전처리된 특정 주파수를 이용하여 상기 인코더(Encoder)로부터 분리된 단어를 인식하는 디코더(Decoder);를 포함하는 것을 특징으로 하는 인공지능 기술기반 스트리밍 영상 검색

시스템.

청구항 5

자막 생성부에 의하여, 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막이 생성되는 자막 생성단계;

동기화부에 의하여, 상기 다수 개의 영상 각각의 타임라인이 기준이 되어 장면(Scene)과 상기 자막이 동기화 되는 동기화단계;

키워드 획득부에 의하여, 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상이 제공될 수 있도록 사용자 단말로부터 키워드가 획득되는 키워드 획득단계; 및

검색결과 추출부에 의하여, 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)이 검색결과로 추출되는 검색결과 추출단계;를 포함하고,

상기 검색결과 추출단계는,

상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시되는 것을 특징으로 하는 인공지능 기술기반 스트리밍 영상 검색 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공지능 기술기반 스트리밍 영상 검색 시스템 및 방법에 관한 것으로, 보다 구체적으로 장단기 메모리(Long Short-Term Memory; LSTM) 기반의 음성 텍스트 변환(Speech-to-Text; STT) 기술을 이용하여 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 저장된 다수 개의 영상에서 사용자가 검색하고자 하는 키워드를 포함하는 영상의 장면(Scene)을 검색결과로 추출하는 인공지능 기술기반 스트리밍 영상 검색 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 오늘날 유무선 인터넷으로 행해지고 있는 실시간 미디어 스트리밍 기술은 폭발적으로 증가하고 있다. 일예로 방송국에서 송출하는 영상뿐만 아니라, 개방된 인터넷을 통하여 방송 프로그램, 영화 등 미디어 콘텐츠를 제공하는 서비스인 OTT 서비스와 개인이 직접 콘텐츠를 제작하고 실시간 업로드 및 스트리밍 할 수 있도록 하는 유튜브(Youtube) 서비스 등이 있다.

[0003] 여기서, 스트리밍(Streaming)이란 마치 끊임없고 지속적인 물 흐름처럼 디지털 멀티미디어 데이터를 제공하는 제공자가 실시간으로 전송하면 사용자는 인터넷 네트워크를 통해서 실시간으로 받아볼 수 있는 기술을 일컫는다.

[0004] 이러한 기술의 발달로 수많은 영상기반 콘텐츠가 쏟아져 나오고 있어 특정 콘텐츠에 접근하고자 하는 사용자를 위한 검색 기술이 필요한 실정이다. 이와 관련하여, 관련문헌 1은 동영상 검색 시스템에 관한 것으로, 동영상에서 특정인물이 등장하는 구간을 자동으로 검색하여 디스플레이할 수 있으나, 특정인물이 유명인이라면 해당 유명인의 정보를 기반으로 검색할 수 있지만 기존의 정보가 전무한 경우에는 검색하기 어려운 기술적 한계가 존재한다.

[0005] 또한, 관련문헌 2는 컬러 스펙트럼을 이용한 동영상 검색시스템에 관한 것으로, HSV, HSL 기반으로 핑거프린트 값을 분석하여 사용자가 업로드 한 동영상과 완전일치 또는 부분 일치되는 웹사이트의 동영상을 검색하여 제공할 수 있으나, 사용자가 검색하고자 하는 영상의 색감이 완전히 상이한 경우에는 관련영상이더라도 영상 검색에서 배제되고 이에 따라 한정적인 검색결과를 제공할 수밖에 없는 기술적 한계가 존재한다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) KR 10-0636910

(특허문헌 0002) KR 10-0983629

발명의 내용

해결하려는 과제

[0007] 본 발명은 상기와 같은 문제점을 해결하기 위한 것으로 다수 개의 영상 중에서 사용자가 검색하고자 하는 영상을 텍스트 기반으로 찾는데 용이하도록 장단기 메모리(Long Short-Term Memory; LSTM) 기반의 음성 텍스트 변환(Speech-to-Text; STT) 기술을 이용하여 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 저장된 다수 개의 영상에서 사용자가 검색하고자 하는 키워드를 포함하는 영상의 장면(Scene)을 검색결과로 추출하는 인공지능 기술기반 스트리밍 영상 검색 시스템 및 방법을 얻고자 하는 것을 목적으로 한다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위하여, 본 발명의 인공지능 기술기반 스트리밍 영상 검색 시스템은 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막을 생성하는 자막 생성부; 상기 다수 개의 영상 각각의 타임라인을 기준으로 장면(Scene)과 상기 자막을 동기화 하는 동기화부; 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상을 제공할 수 있도록 사용자 단말로부터 키워드를 입력받는 키워드 획득부; 및 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)을 검색결과로 추출하는 검색결과 추출부;를 제공하고, 상기 검색결과 추출부는, 상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시되는 것을 특징으로 한다.

[0009] 상기 목적을 달성하기 위하여, 본 발명의 인공지능 기술기반 스트리밍 영상 검색 방법은 자막 생성부에 의하여, 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막이 생성되는 자막 생성단계; 동기화부에 의하여, 상기 다수 개의 영상 각각의 타임라인이 기준이 되어 장면(Scene)과 상기 자막이 동기화 되는 동기화단계; 키워드 획득부에 의하여, 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상이 제공될 수 있도록 사용자 단말로부터 키워드가 획득되는 키워드 획득단계; 및 검색결과 추출부에 의하여, 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)이 검색결과로 추출되는 검색결과 추출단계;를 제공하고, 상기 검색결과 추출단계는, 상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시되는 것을 특징으로 한다.

발명의 효과

[0010] 이상과 같이 본 발명에 의하면 장단기 메모리(Long Short-Term Memory; LSTM) 기반의 음성 텍스트 변환(Speech-to-Text; STT) 기술을 이용하여 스트리밍 서비스를 제공하는 스트리밍 제공서버 내 데이터베이스에 저장된 다수 개의 영상에서 사용자가 검색하고자 하는 키워드를 포함하는 영상의 장면(Scene)을 검색결과로 추출함으로써, 다수 개의 영상 중에서 사용자가 검색하고자 하는 영상을 텍스트 기반으로 찾는데 용이한 효과가 있다. 그리고 단순히 영상만을 검색결과로 추출하는 것이 아닌, 영상 내에서도 해당 텍스트가 언급되는 장면을 검색결과로 추출함으로써, 사용자가 자신이 검색하고자 하는 부분을 찾기 위해서 영상 전체를 시청해야하는 번거로움을 해소할 수 있는 현저한 효과가 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 인공지능 기술기반 스트리밍 영상 검색 시스템 구성도이다.

도 2는 본 발명의 일실시예에 따른 자막 생성부의 세부구성을 표시한 도면이다.

도 3은 본 발명의 일실시예에 따라 동기화부의 동작을 표시한 도면이다.

도 4는 본 발명의 일실시예에 따른 k-평균 알고리즘을 표시한 도면이다.

도 5는 본 발명의 인공지능 기술기반 스트리밍 영상 검색 방법 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 본 명세서에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0013] 다르게 정의되지 않는 한 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0015] **인공지능 기술기반 스트리밍 영상 검색 시스템**

[0016] 이하, 본 발명에 따른 실시예를 첨부한 도면을 참조하여 상세히 설명하기로 한다. 도 1은 본 발명의 인공지능 기술기반 스트리밍 영상 검색 시스템 구성도이다. 도 2는 본 발명의 일실시예에 따른 자막 생성부(100)의 세부 구성을 표시한 도면이다. 도 3은 본 발명의 일실시예에 따라 동기화부(200)의 동작을 표시한 도면이다. 도 4는 본 발명의 일실시예에 따른 k-평균 알고리즘을 표시한 도면이다.

[0017] 우선 도 1을 보면, 본 발명의 인공지능 기술기반 스트리밍 영상 검색 시스템은 자막 생성부(100), 동기화부(200), 키워드 획득부(300) 및 검색결과 추출부(400)를 포함한다.

[0018] 보다 구체적으로, 상기 자막 생성부(100)는 스트리밍 서비스를 제공하는 스트리밍 제공서버(600) 내 데이터베이스(610)에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막을 생성한다.

[0019] 상기 자막 생성부(100)는 상기 스트리밍 제공서버(600)에 접근할 수 있는데, 이는 인터넷 네트워크에 연결된 상태에서 가능하다. 여기서, 상기 스트리밍 제공서버(600)는 예컨대 Google, Youtube 등 검색 매체 및 영상 플랫폼을 구축하는 서버일 수 있고, 다수의 단말을 통해서 업로드 된 다수 개의 영상을 저장하거나 또 다른 다수의 단말에 다수 개의 영상을 제공 및 추천할 수 있다.

[0020] 또한, 상기 자막 생성부(100)는 상기 다수 개의 영상 내 음성으로부터 상기 자막을 추출하기 위해서 음성 데이터 추출부(110), 전처리부(120), 변환부(130) 및 맞춤법 검사부(140)를 포함할 수 있다. 우선, 상기 음성 데이터 추출부(110)는 상기 다수 개의 영상으로부터 음성 데이터를 추출할 수 있다. 다시 말하면, 상기 음성 데이터 추출부(110)는 상기 다수 개의 영상 중에서 임의의 하나의 영상을 대상으로 발화자의 음성을 전체적으로 추출할 수 있다. 추출된 상기 음성 데이터는 시간에 대한 진폭으로 형성된 파형을 포함할 수 있다.

[0021] 다음으로, 상기 전처리부(120)는 상기 음성 데이터에서 잡음을 분리하고 특정 주파수에 맞게 전처리할 수 있다. 상기 음성 데이터는 발화자의 음성 외 주변음과 같은 잡음이 포함되어 있는데, 이는 정확도 높은 텍스트 변환을 방해하는 요소임으로 상기 전처리부(120)는 도 2와 같이 음성과 잡음을 분리할 수 있다. 그리고 잡음이 분리된 상기 음성 데이터를 특정 주파수에 맞게 모델링할 수 있다.

[0022] 다음으로, 상기 변환부(130)는 전처리된 상기 음성 데이터를 음성 텍스트 변환(Speech-to-Text; STT) 기술을 사용하여 텍스트로 변환한다. 도 2를 보면, 상기 변환부(130)는 인코더(Encoder)(131) 및 디코더(Decoder)(132)를 포함할 수 있다.

[0023] 일반적으로, 음성 텍스트 변환(Speech-to-Text; STT) 기술은 히든 마르코프 모델(Hidden Markov Model; HMM)이 사용될 수 있는데, 음성을 상태 천이확률 및 각 상태에서의 노드(출력 심볼)의 관찰 확률을 갖는 마르코프 프로세스로 가정한 후에 학습 데이터를 통해 상태 천이 확률 및 노드의 관찰 확률을 추적하고, 추정된 모델에서 입력된 음성이 발생할 확률을 계산하는 인식 기술이다. 종래 히든 마르코프 모델(HMM)의 각 사건은 바로 전 단계에서만 영향을 받고, 구조적으로 지역적인 최적값에 수렴하게 된다.

[0024] 보다 구체적으로, 종래 히든 마르코프 모델(HMM)은 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태, 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 은닉(Hidden)이라는 표현한다. 그리고 마르코프 성질이란 과거와 현재 상태가 주어졌을 때의 미래상태의 조건부 확률 분포가 과

거 상태와는 독립적으로 현재 상태에 의해서만 결정된다.

[0025] 그러나 본원발명은 인코더(Encoder)(131) 및 디코더(Decoder)(132) 내 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함함으로, 과거 상태부터 현재까지 연속된 상태에 의해서 결과가 결정될 수 있고, 이에 따라 보다 정확한 결과값을 도출할 수 있는 현저한 효과가 있다.

[0026] 상기 인코더(Encoder)(131)는 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하고, 기 저장된 단어에 따라 문장을 단어 단위로 분리할 수 있다.

[0027] 우선, 상기 인코더(Encoder)(131) 내 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델은 문장에서 분리된 다수 개의 단어를 모두 입력받은 뒤 마지막 시점의 은닉상태를 상기 디코더(Decoder)(132) 내 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델에 전달할 수 있다. 즉, 상기 인코더(Encoder)(131)는 문장에서 분리된 다수 개의 단어를 순차적으로 입력받은 뒤 마지막 단어까지 입력받으면 모든 단어를 압축해서 콘텍스트(Context) 벡터를 생성한다.

[0028] 상기 디코더(Decoder)(132)는 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하고, 상기 콘텍스트(Context) 벡터를 최초의 은닉 상태로 하고 상기 전처리부(120)로부터 전처리된 특정 주파수를 이용하여 상기 인코더(Encoder)(131)로부터 분리된 단어를 인식할 수 있다. 그리고 상기 디코더(Decoder)(132)는 상기 콘텍스트(Context) 벡터를 입력받으면 번역된 단어를 하나씩 순차적으로 출력할 수 있다.

[0029] 다음으로, 상기 맞춤법 검사부(140)는 상기 텍스트에 대한 맞춤법을 검사한 후 띄어쓰기 또는 오탈자를 정정하고, 상기 텍스트에서 마침표를 기준으로 문장을 구분할 수 있다. 이는, 상기 변환부(130)로부터 변환된 텍스트가 문맥적으로 올바를 수 있도록 추가적인 확인과정을 수행하기 위함이다.

[0030] 다음으로, 상기 동기화부(200)는 상기 다수 개의 영상 각각의 타임라인을 기준으로 장면(Scene)과 상기 자막을 동기화 한다. 도 3을 보면, 상기 자막 생성부(100)로부터 생성된 자막이 문장별로 구분되어 있고, 하기에는 상기 자막을 생성한 음성을 포함한 영상의 타임라인이 표시되어 있다. 그리고 00:08 시간대에 해당하는 장면(Scene), 00:12 시간대에 해당하는 장면(Scene), 00:24 시간대에 해당하는 장면(Scene)이 각각 표시되어 있다. 즉, 상기 동기화부(200)는 상기 다수 개의 영상에서 각각의 타임라인을 기준으로 장면(Scene)에 부합하는 자막을 동기화할 수 있다.

[0031] 한편, 상기 동기화부(200)는 상기 다수 개의 영상 각각에 대한 자막을 단어별로 해시태그화 하여 저장할 수 있다. 자막은 상기 맞춤법 검사부(140)를 통해서 문장 단위로 구분되어 있을 수 있고, 상기 동기화부(200)는 문장을 문법상의 일정한 뜻을 가지는 말의 최소 단위인 단어로 분리할 수 있고, 각 단어의 앞쪽에 해시태그(#)를 덧붙이고 이를 검색할 수 있도록 하는 해시태그화할 수 있다.

[0032] 이때, 상기 동기화부(200)는 상기 해시태그화된 단어들을 단순히 나열하거나 데이터베이스 형식으로 저장하지 않는다. 이렇게 저장되게 되면, 상기 검색결과 추출부(400)로부터 상기 사용자 단말(500)로부터 획득한 키워드와 동일한 해시태그화된 단어를 포함하는 임의의 영상을 검색결과로 추출할 수 있으나, 사용자는 자신이 검색하고자 하는 장면을 찾기 위해서 소정의 시간동안 상기 검색결과로 추출된 영상을 시청한 후 직접 해당 장면을 찾아야하는 번거로움이 존재한다. 따라서 이러한 문제를 해결하기 위하여, 본원발명의 상기 동기화부(200)는 임의의 영상을 대상으로 상기 자막이 해당하는 장면(Scene)에 상기 해시태그화된 단어들을 동기화하여 저장하는 것이 가장 바람직하다.

[0033] 즉, 상기 검색결과 추출부(400)는 해시태그화된 단어들 중에서 상기 키워드와 동일한 해시태그화된 단어들과 함께 저장된 장면(Scene)들을 상기 검색결과로 추출할 수 있도록 함으로써, 사용자는 전체 영상을 확인하지 않고 영상 내에서도 자신이 검색하고자 하는 특정 장면(Scene)을 바로 확인할 수 있는 현저한 효과가 있다.

[0034] 다음으로, 상기 키워드 획득부(300)는 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상을 제공할 수 있도록 사용자 단말(500)로부터 키워드를 입력받는다. 가장 바람직하게 상기 사용자 단말(500)과 유무선 통신 가능하도록 연결될 수 있고, 상기 사용자 단말(500)의 터치스크린 또는 키보드를 통해서 사용자가 검색하고자 하는 키워드가 입력되면 상기 키워드 획득부(300)는 상기 키워드를 획득할 수 있다.

[0035] 다음으로, 검색결과 추출부(400)는 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)을 검색결과로 추출한다. 그리고 상기 검색결과 추출부(400)는 상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말(500)로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시될 수 있다.

[0036] 즉, 상기 검색결과 추출부(400)는 상기 검색결과로 단순히 상기 키워드와 동일한 해시태그화된 단어들이 저장되어 있는 장면(Scene)만을 제공하거나 상기 사용자 단말(500)로부터 상기 검색결과가 선택되면 상기 장면(Scene)에 기 저장된 자막을 상기 장면(Scene)과 동시에 제공할 수 있는 현저한 효과가 있다.

[0037] 또한, 상기 검색결과 추출부(400)는 상기 자막 생성부(100)로부터 자막이 생성되지 않을 경우 비지도 학습이 가능하도록 k-평균 알고리즘을 이용하여 상기 키워드와 유사한 하나 이상의 장면(Scene)을 검색결과로 추출할 수 있다. 여기서, 도 4를 보면 k-평균 알고리즘은 일반적으로 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분선을 최소화하는 방식으로 동작한다. 가장 바람직하게, 상기 검색결과 추출부(400)는 자막과 해시태그화된 단어가 없는 영상을 대상으로 상기 변환부(130)를 통해서 다수 개의 단어가 추출될 수 있고, 각각의 단어를 학습함으로써, 임의의 단어를 포함하는 영상별로 묶어서 군집화시킬 수 있다. 그리고 상기 검색결과 추출부(400)는 임의의 단어가 키워드로 입력되면 상기 키워드와 유사한 하나 이상의 영상 또는 장면(Scene)을 검색결과로 추출할 수 있다.

[0039] **인공지능 기술기반 스트리밍 영상 검색 방법**

[0040] 이하, 본 발명에 따른 실시예를 첨부한 도면을 참조하여 상세히 설명하기로 한다. 도 5는 본 발명의 인공지능 기술기반 스트리밍 영상 검색 방법 흐름도이다.

[0041] 도 5를 보면, 본 발명의 인공지능 기술기반 스트리밍 영상 검색 방법은 자막 생성단계(S100), 동기화단계(S200), 키워드 획득단계(S300) 및 검색결과 추출단계(S400)를 포함한다.

[0042] 보다 구체적으로, 상기 자막 생성단계(S100)는 자막 생성부(100)에 의하여, 스트리밍 서비스를 제공하는 스트리밍 제공서버(600) 내 데이터베이스(610)에 접근하고, 인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 상기 자막이 생성된다.

[0043] 한편, 상기 자막 생성단계(S100)는 상기 다수 개의 영상 내 음성으로부터 상기 자막을 추출하기 위해서 음성 데이터 추출단계(S110), 전처리단계(S120), 변환단계(S130) 및 맞춤법 검사단계(S140)를 포함할 수 있다.

[0044] 우선, 상기 음성 데이터 추출단계(S110)는, 음성 데이터 추출부(110)에 의하여, 상기 다수 개의 영상으로부터 음성 데이터가 추출될 수 있다. 다시 말하면, 상기 음성 데이터 추출단계(S110)는 상기 다수 개의 영상 중에서 임의의 하나의 영상이 대상이 되어 발화자의 음성이 전체적으로 추출될 수 있다. 추출된 상기 음성 데이터는 시간에 대한 진폭으로 형성된 파형이 포함될 수 있다.

[0045] 다음으로, 상기 전처리단계(S120)는 전처리부(120)에 의하여, 상기 음성 데이터에서 잡음이 분리되고 특정 주파수에 맞게 전처리될 수 있다. 상기 음성 데이터는 발화자의 음성 외 주변음과 같은 잡음이 포함되어 있는데, 이는 정확도 높은 텍스트 변환을 방해하는 요소임으로 음성과 잡음이 분리되는 과정을 거칠 수 있다. 그리고 잡음이 분리된 상기 음성 데이터가 특정 주파수에 맞게 모델링될 수 있다.

[0046] 다음으로, 상기 변환단계(S130)는 변환부(130)에 의하여, 전처리된 상기 음성 데이터가 음성 텍스트 변환(Speech-to-Text; STT) 기술이 이용되어 텍스트로 변환될 수 있다. 음성 텍스트 변환(Speech-to-Text; STT)을 위해서, 상기 변환단계(S130)는 인코딩단계(S131) 및 디코딩단계(S132)를 포함할 수 있다.

[0047] 일반적으로, 음성 텍스트 변환(Speech-to-Text; STT) 기술은 히든 마르코프 모델(Hidden Markov Model; HMM)이 사용될 수 있는데, 음성을 상태 천이확률 및 각 상태에서의 노드(출력 심볼)의 관찰 확률을 갖는 마르코프 프로세스로 가정한 후에 학습 데이터를 통해 상태 천이 확률 및 노드의 관찰 확률을 추적하고, 추정된 모델에서 입력된 음성이 발생할 확률을 계산하는 인식 기술이다. 종래 히든 마르코프 모델(HMM)의 각 사건은 바로 전 단계에서만 영향을 받고, 구조적으로 지역적인 최적값에 수렴하게 된다.

[0048] 보다 구체적으로, 종래 히든 마르코프 모델(HMM)은 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태, 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 은닉(Hidden)이라는 표현한다. 그리고 마르코프 성질이란 과거와 현재 상태가 주어졌을 때의 미래상태의 조건부 확률 분포가 과거 상태와는 독립적으로 현재 상태에 의해서만 결정된다.

[0049] 그러나 본원발명은 인코딩단계(S131) 및 디코딩단계(S132)는 과거 상태부터 현재까지 연속된 상태에 의해서 결과가 결정될 수 있고, 이에 따라 보다 정확한 결과값이 도출될 수 있는 현저한 효과가 있다.

[0050] 상기 인코딩단계(S131)는 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하는

인코더(Encoder)(131)에 의하여, 기 저장된 단어에 따라 문장이 단어 단위로 분리될 수 있다.

[0051] 우선, 상기 인코딩단계(S131)는 상기 인코더(Encoder)(131) 내 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 통해서 문장에서 분리된 다수 개의 단어를 모두 입력받은 뒤 마지막 시점의 은닉 상태가 상기 디코더(Decoder)(132) 내 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델에 전달될 수 있다. 즉, 상기 인코딩단계(S131)는 문장에서 분리된 다수 개의 단어가 순차적으로 입력된 뒤 마지막 단어까지 입력받으면 모든 단어가 압축되어 콘텍스트(Context) 벡터가 생성된다.

[0052] 상기 디코딩단계(S132)는 하나 이상의 장단기 메모리(Long Short-Term Memory; LSTM) 신경망 모델을 포함하는 디코더(Decoder)(132)에 의하여, 콘텍스트(Context) 벡터가 은닉 상태로 되고 상기 전처리단계(S120)로부터 전처리된 특정 주파수가 이용되어 상기 인코딩단계(S131)로부터 분리된 단어가 인식될 수 있다. 그리고 상기 디코딩단계(S132)는 상기 콘텍스트(Context) 벡터가 입력되면 번역된 단어가 하나씩 순차적으로 출력될 수 있다.

[0053] 다음으로, 상기 맞춤법 검사단계(S140)는 상기 맞춤법 검사부(140)에 의하여, 상기 텍스트에 대한 맞춤법이 검사된 후 띄어쓰기 또는 오탈자가 정정되고, 상기 텍스트에서 마침표가 기준이 되어 문장이 구분될 수 있다. 즉, 상기 맞춤법 검사단계(S140)는 상기 변환단계(S130)로부터 변환된 텍스트가 문맥적으로 올바를 수 있도록 추가적인 확인과정을 수행하기 위함이다.

[0054] 다음으로, 상기 동기화단계(S200)는 동기화부(200)에 의하여, 상기 다수 개의 영상 각각의 타임라인이 기준이 되어 장면(Scene)과 상기 자막이 동기화 된다.

[0055] 도 3을 보면, 상기 자막 생성단계(S100)로부터 생성된 자막이 문장별로 구분되어 있고, 하기에는 상기 자막을 생성한 음성을 포함한 영상의 타임라인이 표시되어 있다. 그리고 00:08 시간대에 해당하는 장면(Scene), 00:12 시간대에 해당하는 장면(Scene), 00:24 시간대에 해당하는 장면(Scene)이 각각 표시되어 있다. 즉, 상기 동기화단계(S200)는 상기 다수 개의 영상에서 각각의 타임라인이 기준이 되어 장면(Scene)에 부합하는 자막이 동기화 될 수 있다.

[0056] 한편, 상기 동기화단계(S200)는 상기 다수 개의 영상 각각에 대한 자막이 단어별로 해시태그화 되어 저장될 수 있다. 자막은 상기 맞춤법 검사단계(S140)를 통해서 문장 단위로 구분되어 있을 수 있고, 상기 동기화단계(S200)는 문장이 문법상의 일정한 뜻을 가지는 말의 최소 단위인 단어로 분리될 수 있고, 각 단어의 앞쪽에 해시태그(#)가 덧붙여져 이를 검색할 수 있도록 하는 해시태그화가 될 수 있다.

[0057] 이때, 상기 동기화단계(S200)는 상기 해시태그화된 단어들이 단순히 나열되거나 데이터베이스 형식으로 저장되지 않는 것이 가장 바람직하다. 이렇게 저장되게 되면, 다음 단계인 상기 검색결과 추출단계(S300)로부터 상기 사용자 단말(500)로부터 획득한 키워드와 동일한 해시태그화된 단어를 포함하는 임의의 영상이 검색결과로 추출될 수 있으나, 사용자는 자신이 검색하고자 하는 장면을 찾기 위해서 소정의 시간동안 상기 검색결과로 추출된 영상을 시청한 후 직접 해당 장면을 찾아야하는 번거로움이 존재한다. 따라서 이러한 문제를 해결하기 위하여, 본원발명의 상기 동기화단계(S200)는 임의의 영상을 대상으로 상기 자막이 해당하는 장면(Scene)에 상기 해시태그화된 단어들이 동기화되어 저장되는 것이 가장 바람직하다.

[0058] 즉, 상기 동기화단계(S200)를 통해서 상기 검색결과 추출단계(S300)는 해시태그화된 단어들 중에서 상기 키워드와 동일한 해시태그화된 단어들과 함께 저장된 장면(Scene)들이 상기 검색결과로 추출될 수 있다. 이에 따라, 사용자는 전체 영상을 확인하지 않고 영상 내에서도 자신이 검색하고자 하는 특정 장면(Scene)을 바로 확인할 수 있는 현저한 효과가 있다.

[0059] 다음으로, 상기 키워드 획득단계(S300)는, 키워드 획득부(300)에 의하여, 상기 다수 개의 영상에서 사용자가 검색하고자 하는 영상이 제공될 수 있도록 사용자 단말(500)로부터 키워드가 획득된다. 가장 바람직하게 상기 키워드 획득부(300)와 유무선 통신 가능하도록 연결된 상기 사용자 단말(500)의 터치스크린 또는 키보드를 통해서 사용자가 검색하고자 하는 키워드가 입력되면 상기 키워드 획득단계(S300)로부터 상기 키워드가 획득될 수 있다.

[0060] 다음으로, 상기 검색결과 추출단계(S400)는 검색결과 추출부(400)에 의하여, 상기 다수 개의 영상에서 상기 키워드와 동일한 자막이 동기화된 장면(Scene)이 검색결과로 추출된다. 그리고 상기 검색결과 추출단계(S400)는, 상기 검색결과로 추출된 하나 이상의 장면이 상기 사용자 단말(500)로부터 선택되면, 동기화된 상기 자막이 상기 장면(Scene) 일측에 표시될 수 있다.

[0061] 즉, 상기 검색결과 추출단계(S400)는 상기 검색결과로 단순히 상기 키워드와 동일한 해시태그화된 단어들이 저

장되어 있는 장면(Scene)만이 제공되거나 상기 사용자 단말(500)로부터 상기 검색결과가 선택되면 상기 장면(Scene)에 기 저장된 자막이 상기 장면(Scene)과 동시에 제공될 수 있는 현저한 효과가 있다.

[0062] 한편, 상기 검색결과 추출단계(S400)는 상기 자막 생성단계(S100)로부터 자막이 생성되지 않을 경우 비지도 학습이 가능하도록 k-평균 알고리즘이 이용되어 상기 키워드와 유사한 하나 이상의 장면(Scene)이 검색결과로 추출될 수 있다. 여기서, 도 4를 보면 k-평균 알고리즘은 일반적으로 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분선을 최소화하는 방식으로 동작한다. 가장 바람직하게, 상기 검색결과 추출단계(S400)는 자막과 해시태그화된 단어가 없는 영상이 대상으로 STT 변환된 다수 개의 단어가 학습되고, 임의의 단어를 포함하는 영상별로 묶어서 군집화시킬 수 있다. 그리고 상기 검색결과 추출단계(S400)는 임의의 단어가 키워드로 입력되면 상기 키워드와 유사한 하나 이상의 영상 또는 장면(Scene)이 검색결과로 추출될 수 있다.

[0064] 실시예들은 하드웨어, 소프트웨어, 펌웨어, 미들웨어, 마이크로코드, 하드웨어 기술 언어, 또는 이들의 임의의 조합에 의해 구현될 수 있다. 소프트웨어, 펌웨어, 미들웨어 또는 마이크로코드로 구현되는 경우, 필요한 작업을 수행하는 프로그램 코드 또는 코드 세그먼트들은 컴퓨터 판독 가능 저장 매체에 저장되고 하나 이상의 프로세서에 의해 실행될 수 있다.

[0065] 그리고 본 명세서에 설명된 주제의 양태들은 컴퓨터에 의해 실행되는 프로그램 모듈 또는 컴포넌트와 같은 컴퓨터 실행 가능 명령어들의 일반적인 맥락에서 설명될 수 있다. 일반적으로, 프로그램 모듈 또는 컴포넌트들은 특정 작업을 수행하거나 특정 데이터 형식을 구현하는 루틴, 프로그램, 객체, 데이터 구조를 포함한다. 본 명세서에 설명된 주제의 양태들은 통신 네트워크를 통해 링크되는 원격 처리 디바이스들에 의해 작업들이 수행되는 분산 컴퓨팅 환경들에서 실시될 수도 있다. 분산 컴퓨팅 환경에서, 프로그램 모듈들은 메모리 저장 디바이스들을 포함하는 로컬 및 원격 컴퓨터 저장 매체에 둘 다에 위치할 수 있다.

[0066] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 으로 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0067] 그러므로 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

[0068] 100.. 자막 생성부

110.. 음성 데이터 추출부

120.. 전처리부

130,, 변환부

131.. 인코더(Encoder)

132.. 디코더(Decoder)

140.. 맞춤법 검사부

200.. 동기화부

300.. 키워드 획득부

400.. 검색결과 추출부

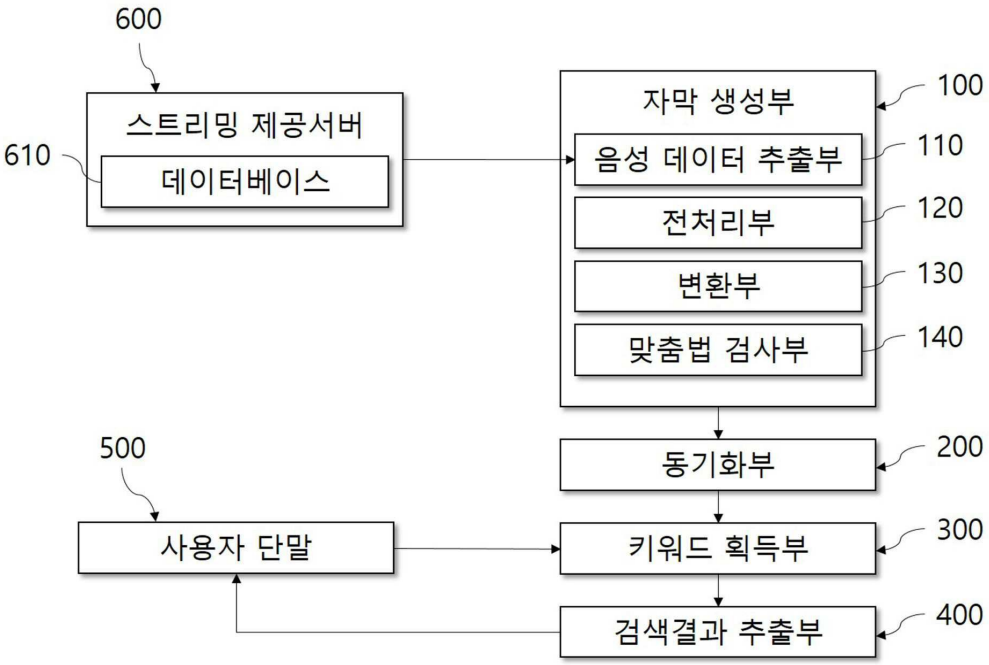
500.. 사용자 단말

600.. 스트리밍 제공서버

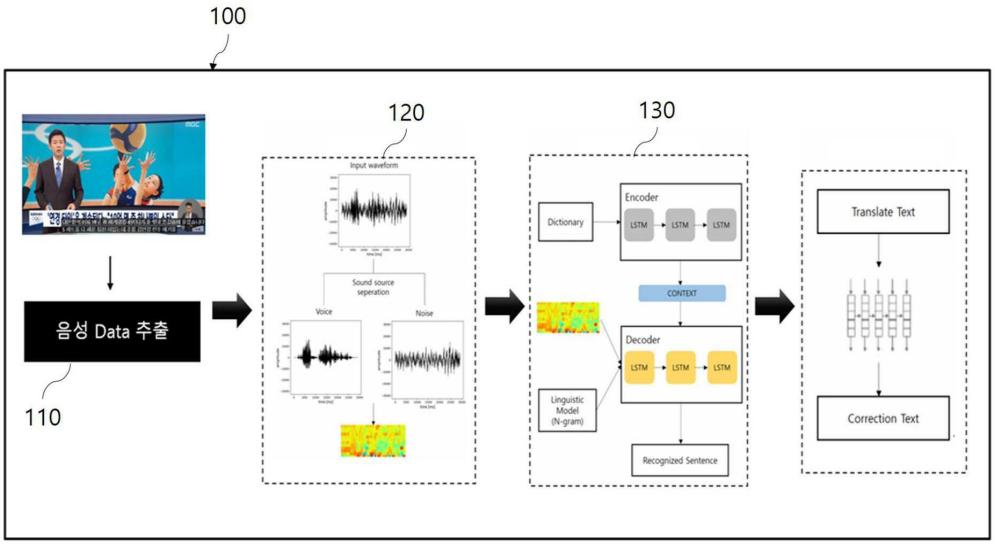
610.. 데이터베이스

도면

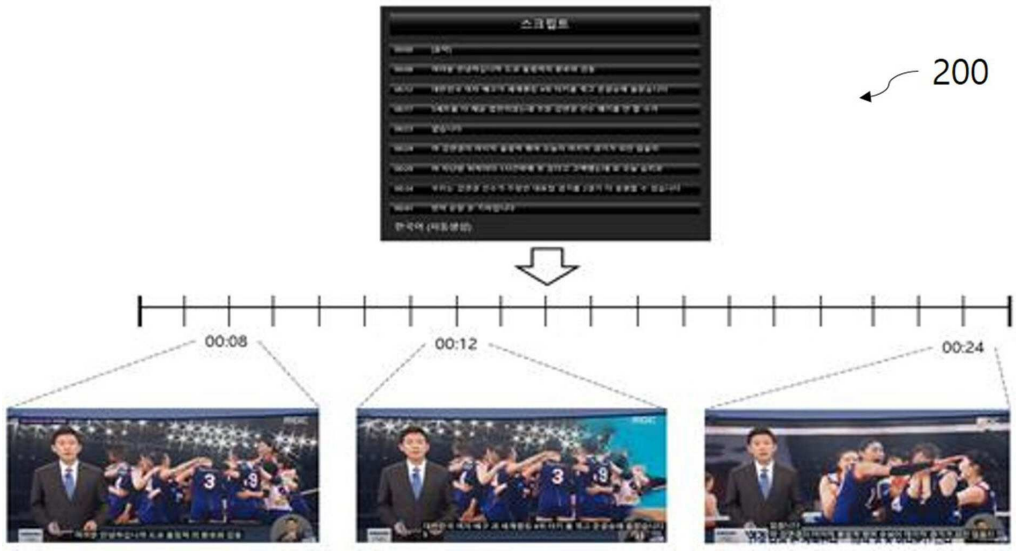
도면1


600
스트리밍 제공서버
610
데이터베이스
자막 생성부
100
음성 데이터 추출부
110
전처리부
120
변환부
130
맞춤법 검사부
140
500
사용자 단말
동기화부
200
키워드 획득부
300
검색결과 추출부
400

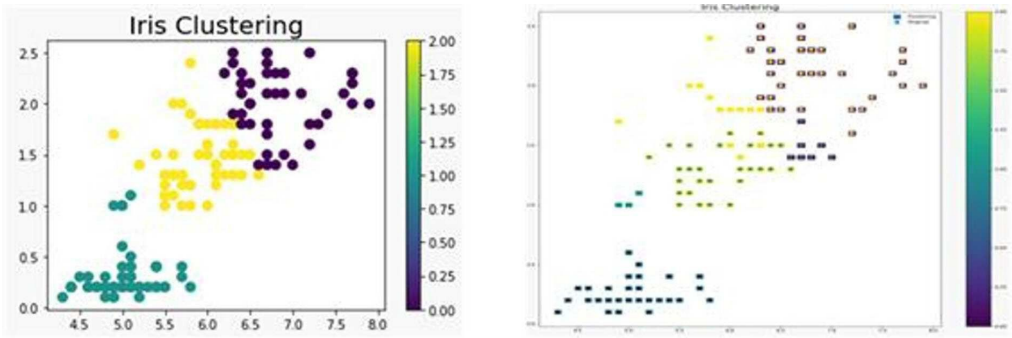
도면2


100
음성 Data 추출
110
120
Input waveform
Sound source seperation
Voice
Noise
130
Dictionary
Encoder
LSTM
CONTEXT
Decoder
Linguistic Model (N-gram)
Recognized Sentence
Translate Text
Correction Text

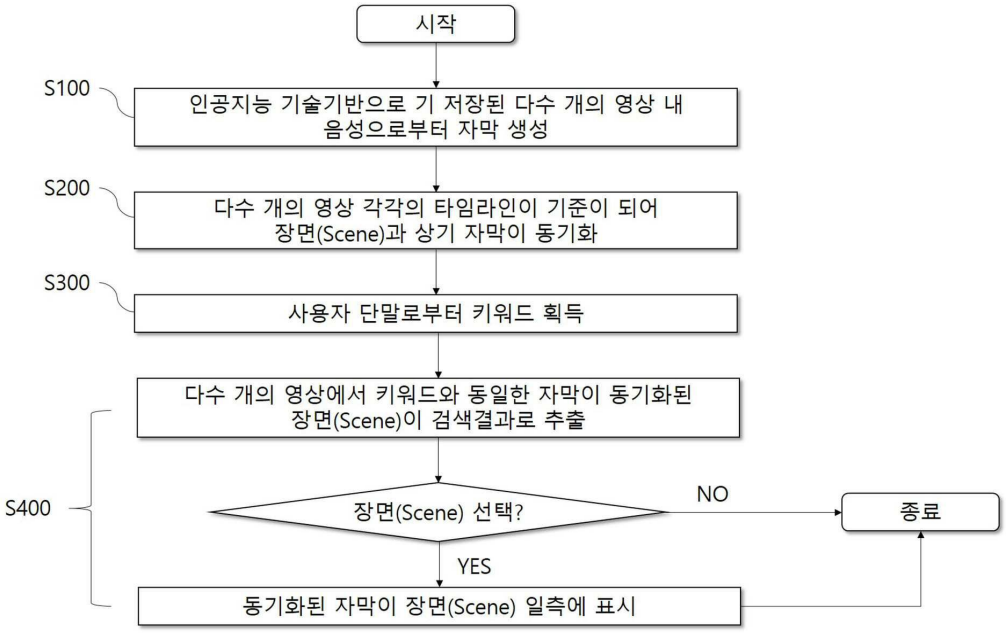
도면3


200
00:08
00:12
00:24

도면4


Iris Clustering

도면5


시작
S100
인공지능 기술기반으로 기 저장된 다수 개의 영상 내 음성으로부터 자막 생성
S200
다수 개의 영상 각각의 타임라인이 기준이 되어 장면(Scene)과 상기 자막이 동기화
S300
사용자 단말로부터 키워드 획득
다수 개의 영상에서 키워드와 동일한 자막이 동기화된 장면(Scene)이 검색결과로 추출
S400
장면(Scene) 선택?
NO
종료
YES
동기화된 자막이 장면(Scene) 일측에 표시