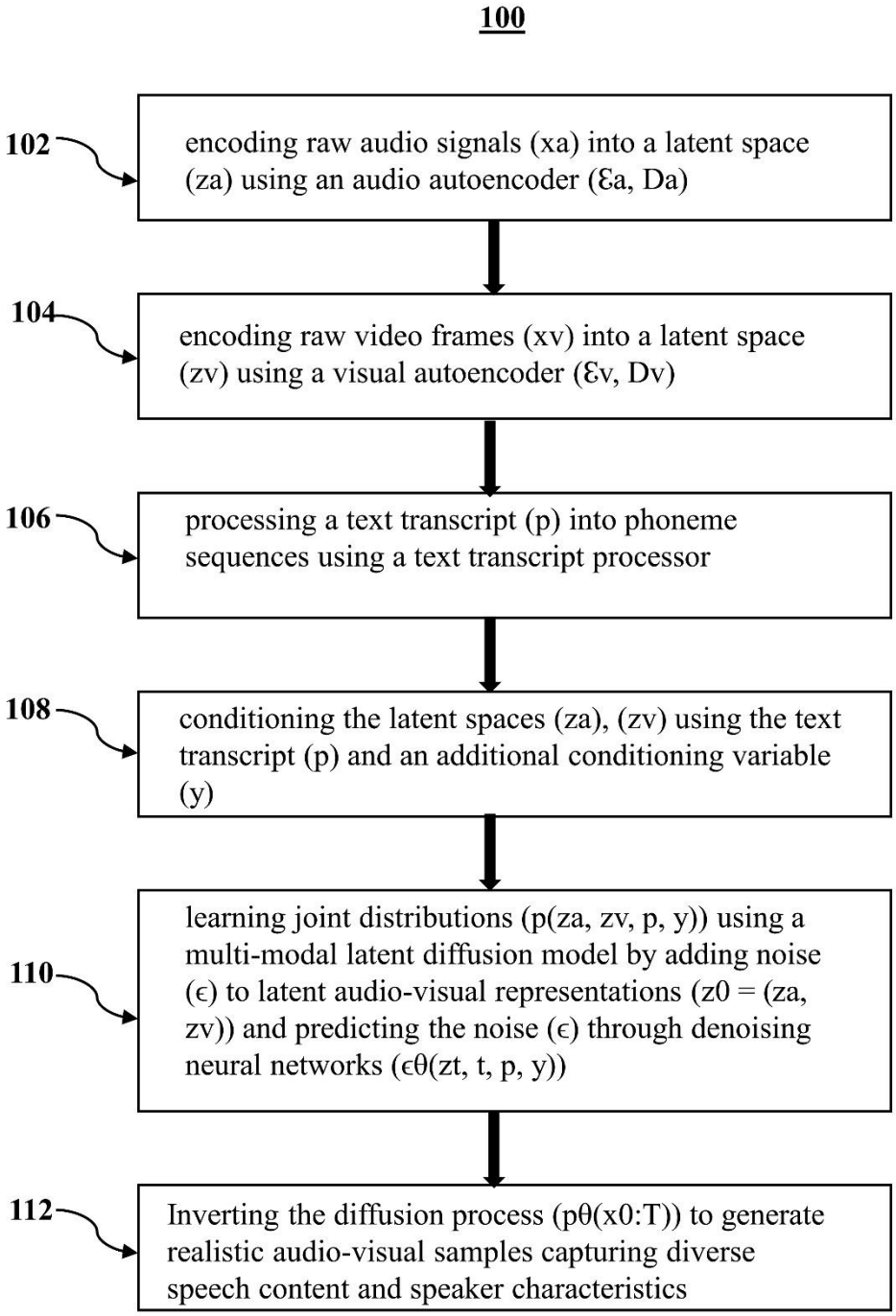


FIG. 1

(SOMA RANI MISHRA)
Reg. No. IN/PA - 1159
L. S. DAVAR & CO.
APPLICANTS' AGENT

200

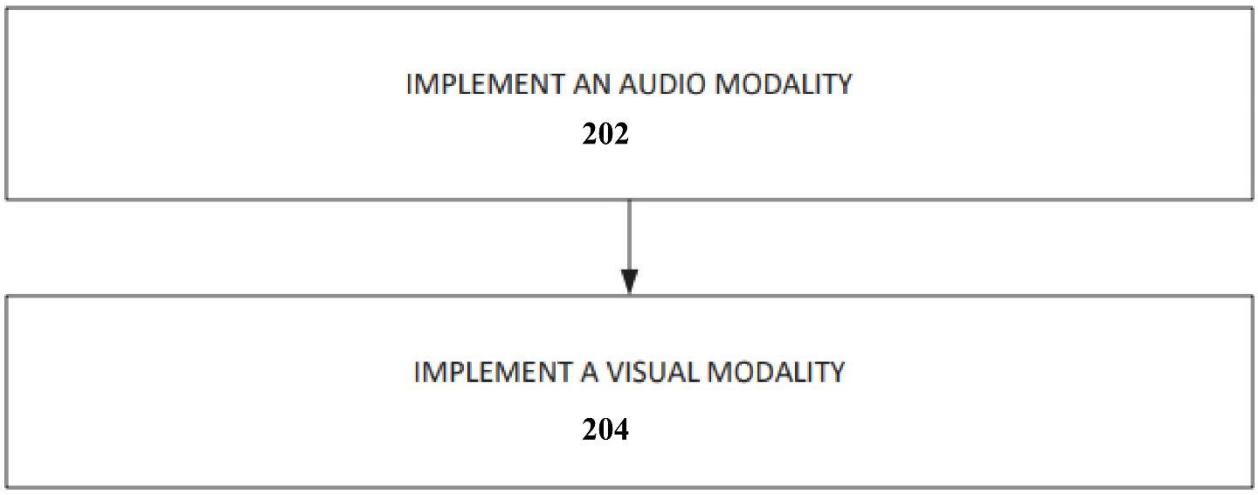


FIG. 2

(SOMA RANI MISHRA)
Reg. No. IN/PA - 1159
L. S. DAVAR & CO.
APPLICANTS' AGENT

300

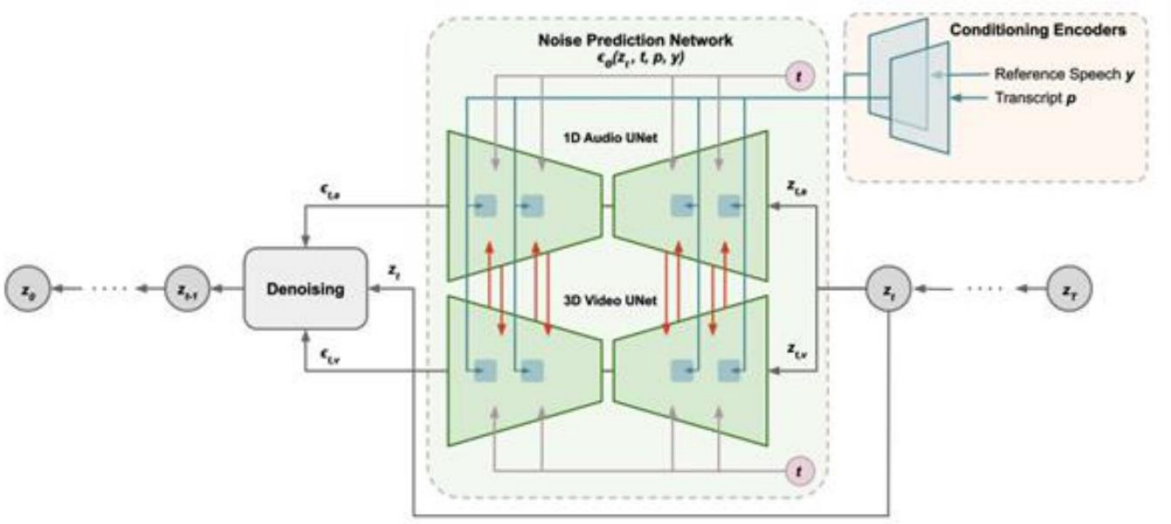


FIG. 3

(SOMA RANI MISHRA)
Reg. No. IN/PA - 1159
L. S. DAVAR & CO.
APPLICANTS' AGENT

202211064831

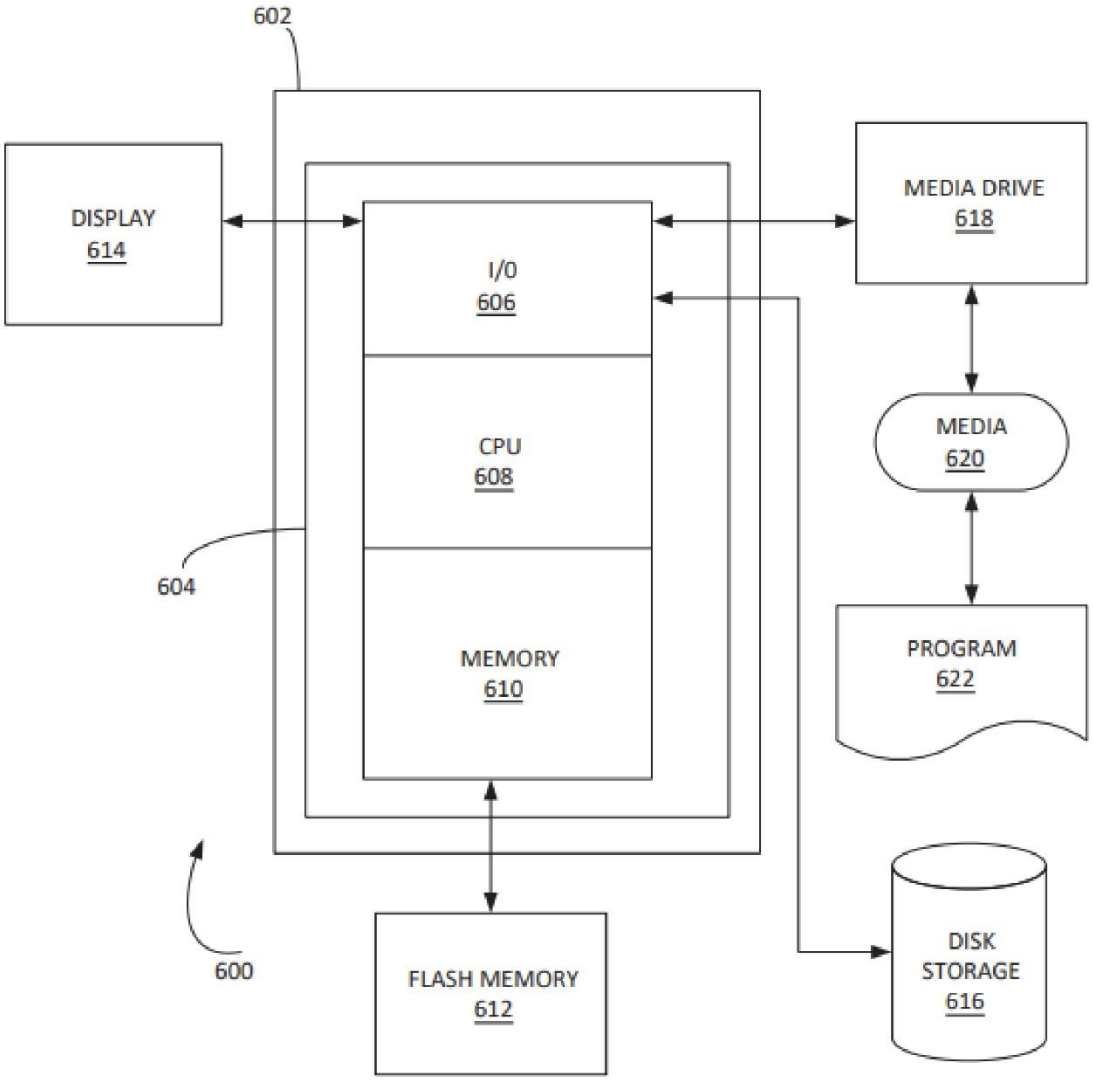


FIG. 4

(SOMA RANI MISHRA)
Reg. No. IN/PA - 1159
L. S. DAVAR & CO.
APPLICANTS' AGENT

FORM 2

THE PATENTS ACT 1970

39 OF 1970

&

THE PATENT RULES 2003

COMPLETE SPECIFICATION

(SEE SECTIONS 10 & RULE 13)

1. TITLE OF THE INVENTION

TEXT-CONDITIONED AUDIO-VISUAL SPEECH GENERATION WITH MULTI-MODAL LATENT DIFFUSION MODELS

2. APPLICANTS (S)

NAME	NATIONALITY	ADDRESS
TENSORTOUR INC	**USA**	**39620, POTRERO DR, NEWARK CA, 94560, USA**

3. PREAMBLE TO THE DESCRIPTION

COMPLETE SPECIFICATION

The following specification particularly describes the invention and the manner in which it is to be performed

5

TECHNICAL FIELD

[0001] The present subject matter described herein relates generally to an audio-visual speech synthesis. The present invention is particularly related to a method and a system for generating text-conditioned audio-visual speech using diffusion models.

BACKGROUND

[0002] Background description includes information that may be useful in understanding the present subject matter. It is not an admission that any of the information provided herein is prior art or relevant to the presently claimed invention, or that any publication specifically or implicitly referenced is prior art.

[0003] In the realm of synthesizing human-like video presentations, the research landscape has seen a multitude of approaches, each striving to animate the human head or face in distinctive ways. Much of the existing research has primarily concentrated on animating facial features such as identity, appearance, head positioning, expressions, and lip movements. Typically, these different facets of video generation are driven by disparate sources, underscoring the intricacy of this domain.

[0004] A notable technique in this field is audio-driven video generation, where facial movements, especially those of the mouth, are synchronized with an audio input. While these methods excel in creating the visual aspects of the video, they often operate within a single modality paradigm, focusing solely on generating the visual elements of the presentation. Yet, real-world applications, especially those necessitating multi-modal synthesis, require the seamless integration of both audio and video components. For instance, applications involving the manipulation of audio-visual speech samples based on text input demand the synthesis of both audio and video content. Traditionally, these tasks have been approached through separate synthesis steps for audio and video, potentially resulting in suboptimal outcomes due to the lack of synchronized and coherent multi-modal output.

[0005] In the current landscape of audio-visual synthesis, several challenges persist, hindering the seamless integration of audio and video components to create realistic human-like interactions. One significant limitation is the lack of precise synchronization between facial movements and speech, leading to unnatural visual representations. Moreover, existing techniques often focus on either audio or video synthesis, lacking the ability to blend both seamlessly for a holistic user experience. Achieving nuanced and realistic facial expressions, a vital aspect of authentic interactions, remains complex. These challenges are further exacerbated by the computational demands of existing methods, limiting accessibility and scalability. Additionally, the adaptability of these techniques across different speakers and scenarios remains a concern, along with the heavy reliance on diverse and high-quality training data.

[0006] Therefore there is a need to overcome the limitations of the existing state of art and enhance synchronization, improve multimodal integration, and optimize computational efficiency.

OBJECTS OF THE DISCLOSURE

[0007] It forms an object of the present disclosure to overcome the aforementioned and other drawbacks/limitations in the existing solutions available in the form of related prior arts.

[0008] It is a primary objective of the present disclosure to seamlessly integrate audio and visual components in real-time, ensuring precise synchronization between speech and facial movements.

[0009] It is another objective of the present disclosure to enhance the generation of facial expressions to mimic human emotions accurately.

[0010] It is another object of the present disclosure to achieve a harmonious blend between audio and video modalities and allow both modalities to influence each other organically, ensuring that speech tone, volume, and content are reflected in facial expressions seamlessly.

[0011] These and other objects and advantages of the present subject matter will be apparent to a person skilled in the art after consideration of the following detailed description taken into consideration with accompanying drawings in which preferred embodiments of the present subject matter are illustrated.

SUMMARY

[0012] A solution to one or more drawbacks of existing technology and additional advantages are provided through the present disclosure. Additional features and advantages are realized through the technicalities of the present disclosure. Other embodiments and aspects of the disclosure are described in detail herein and are considered to be a part of the claimed disclosure.

[0013] The present disclosure provides a method for conditional audio-visual speech generation, comprising encoding raw audio signals (xa) into a latent space (za) using an audio autoencoder ($\mathcal{E}$ a, Da), encoding raw video frames (xv) into a latent space (zv) using a visual autoencoder ($\mathcal{E}$ v, Dv), processing a text transcript ($p = \{p1, p2, ..., pK\} \in P$) into phoneme sequences using a text transcript processor, conditioning the latent spaces (za), (zv) using the text transcript (p) and an additional conditioning variable (y), learning joint distributions (p(za, zv, p, y)) using a multi-modal latent diffusion model by adding noise (ϵ) to the latent audio-visual representations ($z0 = (za, zv)$) and predicting the noise (ϵ) through denoising neural networks ($\epsilon\theta(zt, t, p, y)$) in a sequence of diffusion steps and inverting the diffusion process ($p\theta(x0{:}T)$) to generate realistic audio-visual samples capturing diverse speech content and speaker characteristics.

[0014] In an aspect, the additional conditioning variable (y) includes parameters specifying emotional expression in the generated audio-visual speech.

[0015] In an aspect, the conditioning variable (y) includes parameters related to gender of the speaker, allowing the generated audio-visual speech to exhibit gender-specific characteristics.

[0016] In an aspect, the audio autoencoder (Ɛ a, Da) employs convolutional neural networks (CNNs) to transform raw audio signals into mel-spectrogram representations before encoding into the latent space.

[0017] In an aspect, a style encoder (Se) encodes a reference audio-visual sample to extract specific style characteristics to be applied to the generated audio-visual speech.

[0018] In an aspect, the denoising neural networks (ϵθ) utilize cross-modal connections to align audio and visual modalities.

[0019] In an aspect, the inversion of the diffusion process (pθ(x0:T)) involves generating diverse speech content by manipulating the latent representations (za, zv) based on different text transcripts (p) and additional conditioning variables (y).

[0020] In an aspect, a post-processing step is included wherein generated audio signals are converted into waveforms and video frames are transformed into a video format for seamless integration into audio-visual content.

[0021] In an aspect, the diffusion step include adjusting noise levels based on the complexity of the desired speech content, ensuring realistic generation in varied scenarios.

[0022] In an aspect, the denoising neural networks (ϵθ) employ attention mechanisms to focus on specific phonemes and facial expressions, enhancing the accuracy of noise prediction and speech generation.

[0023] In an aspect, the joint distributions (p(za, zv, p, y)) are learned iteratively, refining the generation process based on user feedback and preferences for specific speaker characteristics and speech styles.

[0024] In an aspect, a system for audio-visual speech synthesis includes a Central Processor Unit (CPU) equipped with modules for encoding real audio-visual speech samples into latent representations, a memory module connected to the CPU for storing machine learning models and an I/O module connected to the CPU, wherein the I/O module is further connected to input devices for user

interaction, including a display, a disk storage unit, and a media drive unit, wherein the modules in the CPU enable the system to generate accurate and synchronized audio-visual speech samples based on real data.

[0025] The foregoing summary is illustrative only and is not intended to be in any way limiting. In addition to the illustrative aspects, embodiments, and features described above, further aspects, embodiments, and features will become apparent by reference to the drawings and the following detailed description.

BRIEF DESCRIPTION OF ACCOMPANYING DRAWINGS

[0026] It is to be noted, however, that the appended drawings illustrate only typical embodiments of the present subject matter and are therefore not to be considered for limiting of its scope, for the present disclosure may admit to other equally effective embodiments. The detailed description is described with reference to the accompanying figures. In the figures, a reference number identifies the figure in which the reference number first appears. The same numbers are used throughout the figures to reference like features and components. Some embodiments of system or methods or structure in accordance with embodiments of the present subject matter are now described, by way of example, and with reference to the accompanying figures, in which:

[0027] Fig. 1 illustrates a flow chart of a method according to the present disclosure;

[0028] Fig. 2 illustrates a flowchart depicting a process of implementing an audio and a visual modality according to the present disclosure;

[0029] Fig. 3 illustrates the multi-modal latent probabilistic diffusion model according to the present disclosure;

[0030] Fig. 4 illustrates a block diagram of a system according to the present disclosure.

[0031] The figures depict embodiments of the present subject matter for illustration only. A person skilled in the art will easily recognize from the following description that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the disclosure described herein.

DETAILED DESCRIPTION OF INVENTION

[0032] The detailed description of various exemplary embodiments of the disclosure is described herein with reference to the accompanying drawings. It should be noted that the embodiments are described herein in such details as to communicate the disclosure. However, the amount of details provided herein is not intended to limit the anticipated variations of embodiments; on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the present disclosure as defined by the appended claims.

[0033] It is also to be understood that various arrangements may be devised that, although not explicitly described or shown herein, embody the principles of the present disclosure. Moreover, all statements herein reciting principles, aspects, and embodiments of the present disclosure, as well as specific examples, are intended to encompass equivalents thereof.

[0034] It should also be noted that in some alternative implementations, the functions/acts noted may occur out of the order noted in the figures. For example, two figures shown in succession may be executed concurrently or may sometimes be executed in the reverse order, depending upon the functionality/acts involved.

[0035] Unless otherwise defined, all terms (including technical and scientific terms) used herein have the same meaning as commonly understood by one of ordinary skill in the art to which example embodiments belong. It will be further understood that terms, e.g., those defined in commonly used dictionaries, should be interpreted as having a meaning that is consistent with their meaning in the

context of the relevant art and will not be interpreted in an idealized or overly formal sense unless expressly so defined herein.

[0036] In the following detailed description of the embodiments of the disclosure, reference is made to the accompanying drawings that form a part hereof, and in which are shown by way of illustration of specific embodiments in which the disclosure may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure, and it is to be understood that other embodiments may be utilized and that changes may be made without departing from the scope of the present disclosure. The following description is, therefore, not to be taken in a limiting sense.

[0037] Hereinafter, a description of an embodiment with several components in communication with each other does not imply that all such components are required. On the contrary, a variety of optional components are described to illustrate the wide variety of possible embodiments of the present disclosure.

[0038] The present invention is directed to a system and a method of text-conditioned audio-visual speech generation with multi-modal latent diffusion models.

[0039] Example definitions for some embodiments are now provided.

[0040] A generative model is a statistical model of a joint probability distribution P (X, Y) P(X, Y) on a given observable variable X and a target variable Y.

[0041] A latent diffusion model may be a diffusion model that, instead of applying a diffusion process directly on a high-dimensional input, projects the input into a smaller latent space to apply the diffusion(s). Here a diffusion model may be a Markov chain(s) trained using variational inference. Diffusion models may learn the latent structure of a dataset by modeling how data points diffuse through the latent space.

[0042] Machine learning (ML) may use statistical techniques to give computers the ability to learn and progressively improve performance on a specific task with data, without being explicitly programmed. Deep learning is a family of machine

learning methods based on learning data representations. Learning may be supervised, semi-supervised, or unsupervised. Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that may teach themselves to grow and change when exposed to new data. Example machine learning techniques that may be used herein include, inter alia: decision tree learning, association rule learning, artificial neural networks, inductive logic programming, support vector machines, clustering, Bayesian networks, reinforcement learning, representation learning, similarity, and metric learning, and/or sparse dictionary learning. Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that may teach themselves to grow and change when exposed to new data. Example machine learning techniques that may be used herein include, inter alia: decision tree learning, association rule learning, artificial neural networks, inductive logic programming, support vector machines, clustering, Bayesian networks, reinforcement learning, representation learning, similarity, and metric learning, and/or sparse dictionary learning. Random forests (RF) (e.g. random decision forests) are an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (e.g. classification) or mean prediction (e.g. regression) of the individual trees. RFs may correct for decision trees' habit of overfitting to their training set. Deep learning is a family of machine learning methods based on learning data representations. Learning may be supervised, semi-supervised, or unsupervised.

[0043] Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement may defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone, 40 dB above the listener's threshold.

[0044] Mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.

[0045] An artificial neural network (ANN) is based on a collection of connected units or nodes called artificial neurons. Each connection may transmit a signal to other neurons. An artificial neuron receives signals then processes them and may signal neurons connected to it. The connections are called edges. Neurons and edges may have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold.

[0046] In one example an audio-visual speech generation framework is provided. The audio-visual speech generation framework may train generative models capable of synthesizing diverse audio-visual speech samples. These may include, inter alia: video clips where the audio track captures a piece of speech, and the video track captures the speaker's face. The audio-visual speech generation framework enables the incorporation of different conditioning mechanisms that enable users to have greater control over the synthesized samples. For example, a text-conditioned version of the audio-visual speech generation model allows users to specify the contents of the speech in the synthesized samples in the form of a text input. Additional control over various aspects of synthesized samples may be gained by adding more conditioning inputs. These may include, for example, an audio-visual speech sample that serves as the reference for identitfication, appearance, and speaking style of the speaker in the synthesized sample.

[0047] The audio-visual speech generation framework is a flexible framework for general audio-visual speech synthesis-based problems. In contrast to existing approaches, our method learns to jointly synthesize both audio and visual modalities that are consistent and synchronized with each other. With different choices of conditioning mechanisms, our method may be used for multiple downstream tasks. For instance, by including text transcript as a conditioning input, the method may be used as an audio-visual generalization of text-to-speech

(TTS) systems. The method may also be used for an audio-driven facial animation by adding an audio source input for driving the audio-driven facial animation and a visual source for identity and appearance. The audio-visual speech generation framework may utilize latent diffusion models. The latent diffusion models may be used for joint modeling of audio-visual speech.

[0048] The audio-visual speech generation framework may be used for a wide variety of audio-visual speech synthesis and manipulation tasks. The audio-visual speech generation framework may be used for recovering corrupted segments (e.g. due to background noise, visual occlusion, etc.) in the talking head videos. Text-conditioned versions of the models may be used for creating talking head videos with dynamic content. The audio-visual speech generation framework may also be used for editing tasks such as correcting grammatical errors and correcting word choices in the videos.

[0049] FIG. 1 illustrates method 100 for conditional audio-visual speech generation with multi-modal latent diffusion models, according to various embodiments. The method 100 may be used in generating audio-visual samples where the audio signal contains the speech following the contents of a given text transcript and the visual signal captures the face of the speaker.

[0050] The objective of the present invention is to model the conditional distribution $p(x_a, x_v\ p, y)$, where $x_a \in R^{Na}$ and $x_v \in R^{Nv\times3\times H\times W}$ are the audio and visual modality of the speech sample to be synthesized, *p* is the text transcript, and *y* is an additional conditioning variable that informs the speaker's voice timbre, speaking style, visual appearance, and other various characteristics in audio-visual speech that are not captured by the text transcript *p*. Here, N_a and N_v represent the number of samples in the audio x_a and the number of frames in the video x_v respectively. The text transcript *p* is represented by a sequence of phonemes $p = \{p1, p2, \ldots, pK\}$, $p_i \in P$, where P is the predefined universal set of phonemes.

[0051] The first step 102 begins with the encoding of raw audio signals (xa) into a lower-dimensional latent space (za) through the use of an audio autoencoder (Ɛ a, Da), and the encoding of raw video frames (xv) into a reduced latent space (zv)

via a visual autoencoder (Ɛ v, Dv). Let ($\mathcal{E}_a$, D_a) and ($\mathcal{E}_v$, D_v) denote the audio and the video autoencoders. Given an audio-video pair (x_a, x_v), encoders $\mathcal{E}_a$ and $\mathcal{E}_v$ learn to map audio $x_a \in R^{Na}$ and video $x_v \in R^{Nv \times H \times W \times 3}$ into the corresponding latent representations $z_a \in R^{na \times ca}$ and $z_v \in R^{nv \times h \times w \times cv}$, while the decoders D_a and D_v learn to reconstruct x_a and x_v back from the latent representations z_a and z_v. Here, c_a and c_v are the number of channels in the latent representation of audio and video modality.

[0052] FIG. 2, illustrates an example process 200 for implementing audio and video autoencoders, according to some embodiments. In step 202, process 200 implements the audio modality. For the audio modality, process 200 trains the autoencoder in the mel-spectrogram domain instead of the raw waveform. Further, a 1D U-Net-like convolutional network is used for the autoencoder with no skip connections between the encoder and the decoder. Process 200 may use a pre-trained vocoder to convert the mel-spectrogram output from the decoder into an audio waveform. Between the raw waveform to the latent representation, process 300 can obtain a total compression factor of $N_a / n_a . c_a$.

[0053] In step 204, process 200 implements a visual modality. For the visual modality, a clip-level autoencoder is trained, where the full video xv is divided into short clips with a length of L frames, and encodes each clip independently. The full latent representation zv is obtained by concatenating the latent codes of individual clips along the temporal dimension. Similar to the audio modality, a U-Net-like architecture is used for the autoencoder but with 3D convolutions. With a spatial downsampling factor of fS = H/h = W/w and a temporal downsampling factor of fT = Nv/nv, an overall compression factor of 3/cv$f$2S fT is obtained in the visual modality. Process 200 trains both the audio and the video autoencoders with a combination of L1 reconstruction loss, perceptual losses, and adversarial losses.

[0054] Also, the audio and video autoencoders may recover the audio and the video from the latent space without loss of any perceptually relevant information.

In this way, these autoencoders obtain a two-way mapping between the raw audio-visual data and their latent representations.

[0055] Also, in the third step 106, a text transcript is processed to transform the textual input (p = {p1, p2, ..., pK} ∈ P) into phoneme sequences. In the fourth 108 step, the latent spaces (za), (zv) are conditioned using the text transcript (p) and an additional conditioning variable (y). Further, the additional conditioning variable (y), which encapsulates emotional expression, speaker characteristics, gender-specific traits, and style elements is added.

[0056] Further, a multi-modal latent probabilistic diffusion model is implemented in the fifth step 110 (Ref. **FIG. 3**). The multi-modal latent probabilistic diffusion model (300) is characterized by its capacity to add controlled noise (ϵ) to the conditioned latent audio-visual representations (z0 = (za, zv)), and operates in a sequence of diffusion steps. The process involves denoising neural networks (ϵθ), which utilize cross-modal connections to align audio and visual modalities accurately. These networks predict noise (ϵ) in the latent representations based on the input text transcript (p) and the additional conditioning variable (y), ensuring the cohesiveness and realism of the generated audio-visual speech samples.

[0057] The multi-modal latent probabilistic diffusion model (300) are latent variable models that model the data distribution q(x0) with $p\theta(x0) = \int p\theta(x0{:}T)dx1{:}T$ where x1, . . . xT is the latent variables. The joint distribution is defined as a Markov chain that starts with the normal distribution p(xT) =N (xT ; 0, I) and denoises the pure noisy sample xT into x0 via T Gaussian transitions,

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\Pi_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \quad (1)$$

$$\text{with } p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)) \quad (2)$$

[0058] In diffusion models, the posterior distribution $q(x_{1:T}\ x_0)$, i.e. the diffusion or the forward process, is set to a fixed Markov chain that gradually adds noise to the data and the reverse process $p_\theta(x_{0:T})$ (Eqs. (1) and (2)) learns to invert that diffusion process.

[0059] The multi-modal latent probabilistic diffusion model (300) are trained by maximizing a variational lower bound on $p\theta(x0)$. It has been shown that a reweighted version of the objective with a certain reparameterization simplifies to prediction of the noise in a noisy version xt of the clean data x_0 using a sequence of denoising neural networks $\epsilon\theta(xt, t)$,

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{\boldsymbol{x}_0,t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right] \qquad (3)$$

[0060] Here, x_0 is sampled from the training dataset, $\epsilon \sim N(0, I)$ is the noise sampled from the normal distribution and the time step t is uniformly sampled from $\{1, \ldots, T\}$.

[0061] Further, consider an audio-video sample (x_a, x_v) from the training dataset. Let *p* be the transcript of the speech in the sample represented as a phoneme sequence. Let *y* be some conditioning variable e.g. another audio-visual speech sample with the same speaker as (x_a, x_v) for style/appearance reference.

[0062] We first get the latent representations $z_a = \mathcal{E}_a(x_a)$ and $z_v = \mathcal{E}_v(x_v)$ using the pre-trained autoencoders (Sec. 2.1). We learn the audio-visual diffusion model by adding noise to the latent audio-visual representation $z_0 = (z_a, z_v)$ as

$$\boldsymbol{z}_t = (\boldsymbol{z}_{t,a}, \boldsymbol{z}_{t,v}) = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \qquad (4)$$

and predicting noise ϵ in the noisy latent representation z_t with a neural network $\epsilon_\theta(z_t, t, \boldsymbol{p}, \boldsymbol{y})$. Here, $\bar{\alpha}_t$ =------β_s, with β_s being the variance of the *s*-th diffusion step in the forward process [7]. The training objective for our diffusion model is given by,

$$\mathcal{L}_{\text{AVDM}} = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{p},\boldsymbol{y},t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{p}, \boldsymbol{y})\|_2^2\right] \qquad (5)$$

[0063] A two-stream architecture for $\epsilon_\theta(z_t, t, \boldsymbol{p}, \boldsymbol{y})$ may be used that has two separate parallel UNets for the audio and visual modality consisting of one and

three-dimensional convolutional layers respectively. The two streams communicate with each other using a number of cross-modal connections. Since $z_{t,a}$ and $z_{t,v}$ typically have different temporal resolutions, we either use a nearest-neighbor based resampling or a learnable cross-attention to align the two modalities along the temporal dimension.

[0064] The text p and all the modalities in y are first encoded using separate encoders, and then the encoded representations are used for conditioning the core noise prediction UNets via cross-attention mechanisms.

[0065] The final step 110 involves the inversion of the diffusion process ($p\theta(x0{:}T)$), which results in the generation of highly realistic and synchronized audio-visual samples. These samples encompass diverse speech content and speaker characteristics, making them ideal for applications requiring immersive and lifelike audio-visual experiences.

[0066] In an embodiment, a style encoder (Se) is integrated that extracts specific style characteristics from a reference audio-visual sample. These characteristics are then seamlessly incorporated into the generated audio-visual speech, adding layers of depth and authenticity to the final output.

[0067] FIG. 4 depicts a system 600 that may be configured to perform any one of the processes provided herein. In this context, the 600 may include, for example, a processor, memory, storage, and I/O devices (e.g., monitor, keyboard, disk drive, Internet connection, etc.). However, the system 600 may include circuitry or other specialized hardware for carrying out some or all aspects of the processes. In some operational settings, the system 600 may be configured as a system that includes one or more units, each of which is configured to carry out some aspects of the processes either in software, hardware or some combination thereof.

[0068] Figure 6 depicts the system 600 with a number of components that may be used to perform any of the processes described herein. The main system 602 includes a motherboard 604 having an I/O section 606, one or more central processing units (CPU) 608, and a memory section 610, which may have a flash

memory card 612 related to it. The I/O section 606 can be connected to a display 614, a keyboard and/or other user input (not shown), a disk storage unit 616, and a media drive unit 618. The media drive unit 618 can read/write a medium 620, which can contain programs 622 and/or data. The system 600 can include a web browser. Moreover, it is noted that computing system 600 can be configured to include additional systems in order to fulfill various functionalities. Computing system 600 can communicate with other computing devices based on various communication protocols such as Wi-Fi, Bluetooth® (and/or other standards for exchanging data over short distances including those using short-wavelength radio transmissions), USB, Ethernet, cellular, ultrasonic local area communication protocol, etc.

TECHNICAL ADVANTAGES

[0069] An essential feature of our method lies in its adaptability to varied scenarios. By adjusting noise levels based on the complexity of the desired speech content, the method ensures the generation of realistic audio-visual samples in diverse situations. Attention mechanisms embedded within the denoising neural networks ($\epsilon\theta$) focus on specific phonemes and facial expressions, enhancing the accuracy of noise prediction and, subsequently, the quality of speech generation.

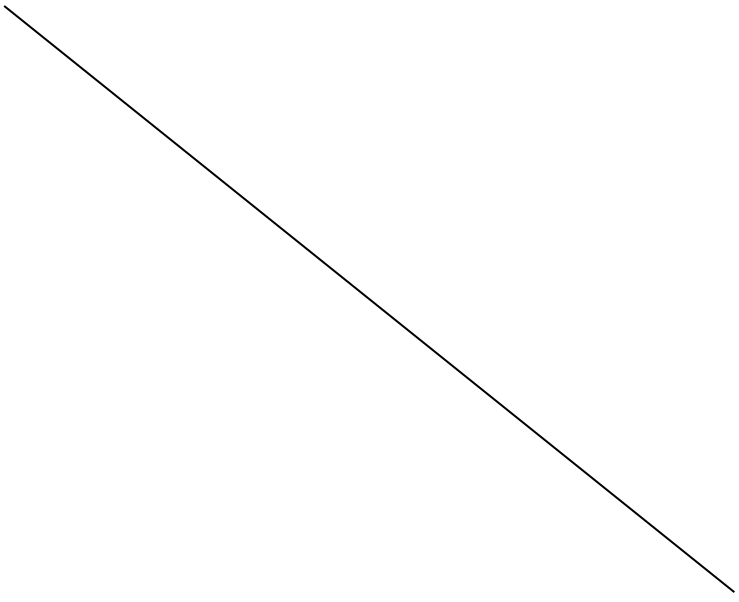
[0070] The joint distributions (p(za, zv, p, y)) are continually refined based on user feedback and preferences. This iterative process, informed by user interactions, ensures that the generated speech content captures nuanced speaker characteristics, diverse speech styles, and specific emotional expressions accurately.

[0071] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (e.g., bodies of the appended claims) are generally intended as "open" terms (e.g., the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such

an intent will be explicitly recited in the claim, and in the absence of such recitation, no such intent is present. For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases “at least one” and “one or more” to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles “a” or “an” limits any particular claim containing such introduced claim recitation to disclosures containing only one such recitation, even when the same claim includes the introductory phrases “one or more” or “at least one” and indefinite articles such as “a” or “an” (e.g., “a” and/or “an” should typically be interpreted to mean “at least one” or “one or more”); the same holds true for the use of definite articles used to introduce claim recitations. Also, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should typically be interpreted to mean at least the recited number (e.g., the bare recitation of “two recitations,” without other modifiers, typically means at least two recitations or two or more recitations). Furthermore, in those instances where a convention analogous to “at least one of A, B, and C, etc.” is used, in general, such construction is intended in the sense one having skill in the art would understand the convention (e.g., “a system having at least one of A, B, and C” would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). In those instances, where a convention analogous to “at least one of A, B, or C, etc.” is used, in general, such construction is intended in the sense one having skill in the art would understand the convention (e.g., “a system having at least one of A, B, or C” would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms or both terms. For

example, the phrase “A or B” will be understood to include the possibilities of “A” or “B” or “A and B.”

[0072] It will be further appreciated that functions or structures of a plurality of components or steps may be combined into a single component or step, or the functions or structures of one-step or component may be split among plural steps or components. The present disclosure contemplates all of these combinations. Unless stated otherwise, dimensions and geometries of the various structures depicted herein are not intended to be restrictive of the disclosure, and other dimensions or geometries are possible. Also, while a feature of the present disclosure may have been described in the context of only one of the illustrated embodiments, such feature may be combined with one or more other features of other embodiments, for any given application. It will also be appreciated from the above that the fabrication of the unique structures herein and the operation thereof also constitute methods in accordance with the present disclosure. The present disclosure also encompasses intermediate and end products resulting from the practice of the methods herein. The use of “comprising” or “including” also contemplates embodiments that “consist essentially of” or “consist of” the recited feature.



We claim:

1. A method (100) for conditional audio-visual speech generation, comprising:

encoding (102) raw audio signals (xa) into a latent space (za) using an audio autoencoder (Ɛ a, Da);

encoding (104) raw video frames (xv) into a latent space (zv) using a visual autoencoder (Ɛ v, Dv);

processing (106) a text transcript (p) into phoneme sequences using a text transcript processor;

conditioning (108) the latent spaces (za), (zv) using the text transcript (p) and an additional conditioning variable (y);

learning (110) joint distributions (p(za, zv, p, y)) using a multi-modal latent diffusion model by adding noise (ϵ) to latent audio-visual representations (z0 = (za, zv)) and predicting said noise (ϵ) through denoising neural networks ($\epsilon\theta$(zt, t, p, y)); and

inverting (112) the diffusion process ($p\theta$(x0:T)) to generate realistic audio-visual samples capturing diverse speech content and speaker characteristics.

2. The method (100) as claimed in claim 1, wherein the additional conditioning variable (y) includes parameters specifying emotional expression in the generated audio-visual speech.
3. The method (100) as claimed in claim 2, wherein the conditioning variable (y) includes parameters related to gender of the speaker, allowing the generated audio-visual speech to exhibit gender-specific characteristics.
4. The method (100) as claimed in claim 1, wherein the audio autoencoder (Ɛ a, Da) employs convolutional neural networks (CNNs) to transform raw audio signals into mel-spectrogram representations before encoding into the latent space.
5. The method (100) as claimed in claim 1, wherein the visual autoencoder (Ɛ v, Dv) utilizes 3D convolutional networks to encode raw video frames into the reduced latent space (zv).

6. The method (100) as claimed in claim 1, wherein a style encoder (Se) encodes a reference audio-visual sample to extract specific style characteristics to be applied to the generated audio-visual speech.
7. The method (100) as claimed in claim 1, wherein the denoising neural networks ($\epsilon\theta$) utilize cross-modal connections to align audio and visual modalities.
8. The method (100) as claimed in claim 1, wherein the inversion of the diffusion process ($p\theta(x0{:}T)$) involves generating diverse speech content by manipulating the latent representations (za, zv) based on different text transcripts (p) and additional conditioning variables (y).
9. The method (100) as claimed in claim 1, further comprises a post-processing step wherein generated audio signals are converted into waveforms and video frames are transformed into a video format for seamless integration into audio-visual content.
10. The method (100) as claimed in claim 1, wherein the diffusion step includes adjusting noise levels based on the complexity of the desired speech content, ensuring realistic generation in varied scenarios.
11. The method (100) as claimed in claim 1, wherein the denoising neural networks ($\epsilon\theta$) employ attention mechanisms to focus on specific phonemes and facial expressions, enhancing the accuracy of noise prediction and speech generation.
12. The method (100) as claimed in claim 1, wherein the joint distributions (p(za, zv, p, y)) are learned iteratively, refining the generation process based on user feedback and preferences for specific speaker characteristics and speech styles.
13. A system (400) for audio-visual speech synthesis, comprising:

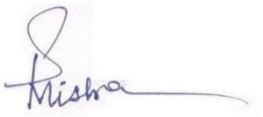
a Central Processor Unit (CPU) (608) equipped with modules for encoding real audio-visual speech samples into latent representations;

a memory module (610) connected to the CPU (608) for storing machine learning models; and

an I/O module (606) connected to the CPU, wherein the I/O module (606) is further connected to input devices for user interaction, including a display (614), a disk storage unit (616), and a media drive unit (618),

wherein the modules in the CPU (608) enable the system (400) to generate accurate and synchronized audio-visual speech samples based on real data.

Dated this 10th day of November, 2023



SOMA RANI MISHRA
IN/PA-1159
OF L. S. DAVAR & CO.
APPLICANTS' AGENT

ABSTRACT

TITLE : TEXT-CONDITIONED AUDIO-VISUAL SPEECH GENERATION WITH MULTI-MODAL LATENT DIFFUSION MODELS

The present subject matter disclosed herein relates to a method (100) for conditional audio-visual speech generation, comprising encoding (102) raw audio signals (xa) into a latent space (za), encoding (104) raw video frames (xv) into a latent space (zv), processing (106) a text transcript (p), conditioning (108) the latent spaces (za), (zv), learning (110) joint distributions (p(za, zv, p, y)) by adding noise (ϵ) to latent audio-visual representations (z0 = (za, zv)) and predicting said noise (ϵ) through denoising neural networks (ϵθ(zt, t, p, y)) and inverting (112) the diffusion process (pθ(x0:T)) to generate realistic audio-visual samples. The system includes a Central Processor Unit (CPU) (608), a memory module (610) connected to the CPU (608), and an I/O module (606) connected to the CPU, wherein the modules in the CPU (608) enable the system (400) to generate accurate and synchronized audio-visual speech samples based on real data.

REFER TO FIG. 1 AND 4