
بسم الله الرحمن الرحیم

فهرست مطالب

فصل 1: مقدمه

فرآیند تولید انیمیشن چهره از متن فارسی شامل چهار مرحله اصلی است. در مرحله نخست، پردازش تصویر ورودی انجام می‌شود تا چهره موردنظر برای ساخت انیمیشن آماده شود. در این مرحله، تصویر ورودی برش خورده و پیش‌پردازش‌های لازم مانند تنظیم نور، وضوح و مشخصه‌های چهره اعمال می‌شود تا کیفیت نهایی انیمیشن بهبود یابد.

در مرحله دوم، متن فارسی ورودی به صوت تبدیل می‌شود. این فرآیند شامل استفاده از فناوری تبدیل متن به گفتار (TTS) است که گفتار طبیعی و روانی را از روی متن تولید می‌کند. مدل TTS نه تنها متن را به گفتار تبدیل می‌کند، بلکه الگوهای عروضی مانند زیر و بمی صدا، مکث‌ها و تأکیدها را نیز شبیه‌سازی می‌کند تا خروجی صوتی طبیعی‌تر به نظر برسد.

در مرحله سوم، صدای تولیدشده پردازش شده و ویژگی‌های آوایی و عروضی آن استخراج می‌شود. این ویژگی‌ها شامل اطلاعاتی مانند شدت و زمان‌بندی صداها هستند که به یک مدل یادگیری عمیق داده می‌شوند. مدل یادگیری عمیق این اطلاعات را تجزیه‌وتحلیل کرده و ضرایب حرکت صورت را پیش‌بینی می‌کند. این ضرایب تعیین می‌کنند که هر بخش از چهره، از جمله لب‌ها، گونه‌ها و فک، چگونه و با چه شدتی باید در هر لحظه حرکت کند تا هماهنگی دقیقی با گفتار ایجاد شود.

در مرحله نهایی، با استفاده از ضرایب استخراج‌شده، انیمیشن چهره ساخته می‌شود. در این مرحله، یک مدل تولیدی یا رندرینگ، تصویر استاتیک چهره را دریافت کرده و آن را فریم به فریم بر اساس ضرایب حرکت اصلاح می‌کند. در نهایت، فریم‌های تولیدشده با صدای اصلی همگام‌سازی شده و در قالب یک ویدئو خروجی گرفته می‌شوند. این ویدئو شامل حرکات دقیق و هماهنگ چهره است که به‌صورت طبیعی گفتار ورودی را بیان می‌کند.

فصل 2: پیش‌پردازش تصویر

اولین مرحله پردازش تصویر ورودی برای آماده سازی آن برای انیمیشن است. این شامل تشخیص چهره در تصویر، برش دادن آن برای فوکوس روی ناحیه صورت مربوطه و استخراج ضرایب مهم صورت است که بعداً برای ایجاد انیمیشن مورد استفاده قرار خواهند گرفت. هدف از این مرحله این است که قبل از ایجاد حرکات مبتنی بر گفتار واقع گرایانه، تصویر به درستی قالب بندی شده و با داده‌های لازم چهره غنی شده باشد.

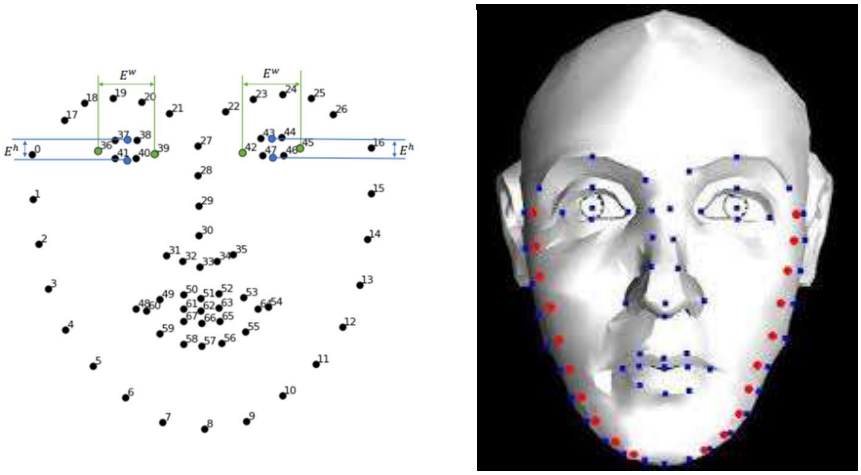
این فرآیند با تشخیص چهره آغاز می‌شود، جایی که سیستم چهره را در تصویر ورودی شناسایی و مکان‌یابی می‌کند. این بسیار مهم است زیرا تشخیص چهره دقیق تضمین می‌کند که انیمیشن روی ناحیه مناسب تصویر تمرکز می‌کند. هنگامی که صورت مشخص شد، تصویر در اطراف ناحیه صورت شناسایی شده برش داده می‌شود و پس‌زمینه غیر ضروری را حذف می‌کند و موقعیت صورت را استاندارد می‌کند. سپس اندازه تصویر برش داده شده به وضوح ثابت تغییر می‌کند تا ثبات در فرآیند انیمیشن حفظ شود.

2-1 مدل‌های مورف‌پذیر سه‌بعدی

در مرحله بعد، سیستم ضرایب مدل مدل‌های مورف‌پذیر سه‌بعدی[1] $(3DMM)$ را استخراج می‌کند که نمایش عددی شکل، بیان و جهت‌گیری صورت است. مدل‌های مورف‌پذیر سه‌بعدی روشی قدرتمند برای نمایش، پردازش و تحلیل چهره‌های سه‌بعدی هستند. این مدل‌ها بر پایه‌ی یک فضای پارامتری ساخته می‌شوند که امکان تغییرات پیوسته در شکل و ظاهر یک چهره را فراهم می‌کند. ایده‌ی اصلی آن‌ها این است که مجموعه‌ای از چهره‌های سه‌بعدی را با استفاده از تحلیل مؤلفه‌های اصلی[2] (PCA) مدل‌سازی کرده و یک فضای برداری ایجاد کنند که در آن هر چهره را می‌توان به‌عنوان ترکیبی خطی از اجزای اصلی تعریف کرد. این روش اجازه می‌دهد که تغییرات چهره در ابعاد مختلف، مانند تغییرات ناشی از حالات چهره، تفاوت‌های فردی و نورپردازی، در یک چارچوب منسجم قابل مدل‌سازی باشند.

[1] 3D Morphable Model (3DMM)

[2] Principal component analysis



شکل 2- 1 مدل مورف‌پذیر سه‌بعدی

مدل‌های مورف‌پذیر سه‌بعدی ابتدا برای بازسازی چهره از تصاویر دوبعدی به کار گرفته شدند و از آن زمان در کاربردهای مختلفی مانند تشخیص چهره، انیمیشن، واقعیت افزوده و ویرایش تصاویر گسترش یافته‌اند. برای ساخت یک DMM3، معمولاً مجموعه‌ای از چهره‌های سه‌بعدی با توپولوژی یکسان مورد استفاده قرار می‌گیرد. با استفاده از تکنیک‌هایی مانند هم‌ترازی[1] و تحلیل آماری، این چهره‌ها در یک فضای مشترک قرار داده می‌شوند تا بتوان یک مدل پارامتری از آن‌ها استخراج کرد. سپس، با تنظیم پارامترهای این مدل، می‌توان چهره‌های جدیدی تولید کرد که ویژگی‌های مختلفی از نمونه‌های اولیه را در خود داشته باشند.

یکی از چالش‌های اصلی در مدل‌های مورف‌پذیر سه‌بعدی، محدودیت آن‌ها در نمایش جزئیات پیچیده‌ی چهره و تغییرات غیرخطی است. برای حل این مشکل، مدل‌های پیشرفته‌تر مانند مدل‌های یادگیری عمیق و شبکه‌های مولد (GAN) توسعه یافته‌اند که می‌توانند دقت و قابلیت تعمیم این مدل‌ها را بهبود بخشند. به‌طور کلی، DMM3ها همچنان به‌عنوان یکی از روش‌های پایه‌ای در تحلیل چهره‌های سه‌بعدی شناخته می‌شوند و با ترکیب آن‌ها با یادگیری عمیق، می‌توان عملکرد آن‌ها را در کاربردهای مختلف بهبود داد.

2-2 بعد از برش تصویر و استخراج ضرایب

پس از برش تصویر و استخراج ضرایب، داده‌های پردازش‌شده برای مراحل بعدی ذخیره می‌شوند. تصویر برش داده‌شده برای تجسم و اشکال‌زدایی ذخیره می‌شود، در حالی که ضرایب استخراج شده به عنوان داده‌های عددی ذخیره می‌شود که انیمیشن چهره را هدایت می‌کند. اگر سیستم نتواند یک چهره را تشخیص دهد یا ضرایب را به درستی استخراج کند، برای جلوگیری از پردازش بیشتر با داده‌های ناقص، خطایی ایجاد می‌شود.

[1] alignment

این مرحله پیش پردازش ضروری است زیرا تضمین می‌کند که تصویر به درستی برای انیمیشن آماده شده است. بدون برش دقیق و استخراج ضریب صورت، انیمیشن به درستی با گفتار هماهنگ نمی‌شود و منجر به حرکات غیرطبیعی یا نادرست می‌شود. با ساختار دادن به داده‌های چهره به این روش، سیستم یک پایه محکم برای مراحل بعدی ایجاد می‌کند، جایی که تصویر همگام با ورودی صدا متحرک می‌شود.

فصل 3: متن به صوت

این جدید شامل تولید یک صوت از متن با استفاده از مدل تبدیل متن به گفتار[1] (TTS) است. به طور خاص، این فرآیند متن نوشته شده را به شکل موج گفتاری با صدای طبیعی تبدیل می‌کند، که بعداً برای متحرک کردن ویژگی‌های صورت در هماهنگی با کلمات گفتاری استفاده می‌شود.

این رویکرد متکی بر مدل MS-TTS گفتار انبوه چند زبانه[2]، facebook/mms-tts-fas، است. این مدل بخشی از ابتکار گسترده‌تر متا[3](فیس‌بوک) برای پشتیبانی از سنتز تبدیل متن به گفتار در صدها زبان از جمله فارسی است. مبتنی بر پیشرفت‌های قبلی در سنتز گفتار است و تحقیقاتی را در مورد مدل‌های VITS[4] انجام می‌دهد که امکان تولید گفتار با کیفیت بالا و رسا را با حداقل داده‌های آموزشی فراهم می‌کند. برخلاف سیستم‌های سنتی TTS پیوسته یا پارامتری، VITS از یادگیری عمیق برای تولید گفتار بسیار طبیعی و روان با مدل‌سازی ویژگی‌های زبانی و آکوستیک گفتار به روشی انتها به انتها[5] استفاده می‌کند.

در این مرحله، متن فارسی ارائه شده در یک نمایش عددی مناسب برای پردازش توسط شبکه عصبی توکن می‌شود. سپس VitsModel یک شکل موج متناظر را ایجاد می‌کند و نه تنها جزئیات آوایی متن ورودی، بلکه آهنگ‌ها و ریتم‌های طبیعی را نیز ثبت می‌کند. این شکل موج متعاقباً به عنوان یک فایل صوتیoutput.wav با نرخ نمونه برداری مناسب ذخیره می‌شود و از سازگاری با مراحل بعدی خط لوله اطمینان حاصل می‌کند.

یکی از مزایای عمده استفاده از mms-tts-fas توانایی آن در تولید گفتار برای زبان‌های کم منبع[6] مانند فارسی است که مدل‌های TTS با کیفیت بالا نسبتا کمیاب هستند. ابتکار MMS متا به طور قابل توجهی تعداد زبان‌های پشتیبانی شده در تشخیص خودکار گفتار (ASR) و وظایف تبدیل متن به گفتار را افزایش داده است، از مجموعه داده های چندزبانه گسترده و تکنیک‌های یادگیری خود نظارت برای دستیابی به عملکرد پیشرفته استفاده می‌کند.

این مرحله، هنگامی که با مرحله پیش پردازش تصویر قبلی ترکیب می‌شود، به ما امکان می‌دهد یک خروجی کاملاً هماهنگ تولید کنیم. هدف این است که یک متن ورودی را بگیرید، آن را به شکل موج گفتاری تبدیل کرده و بعداً این صدا را با تصویر صورت هم‌تراز کنیم تا همگام‌سازی دقیق لب ایجاد شود. این تضمین می‌کند که انیمیشن تولید شده طبیعی به نظر می‌رسد و کلمات گفته‌شده را به درستی دنبال می‌کند.

[1] Text to speech
[2] Massively Multilingual Speech
[3] Meta
[4] Variational Inference Text to Speech
[5] End to end
[6] low-resource languages

فصل 4: تبدیل صدا به ضریب

در این مرحله سیستم شکل موج صوتی تولید شده را از مرحله قبل گرفته و به مجموعه‌ای از نمایش‌های عددی تبدیل می‌کند. این ضرایب حرکت ویژگی‌های صورت، مانند لب‌ها، فک و سایر نکات کلیدی را توصیف می‌کنند و امکان هماهنگی دقیق بین گفتار و انیمیشن صورت را فراهم می‌کنند. این فرآیند برای دستیابی به همگام سازی لب و حالات صورت واقع گرایانه بسیار مهم است.

فرآیند تبدیل به مدل‌های پردازش صوتی مبتنی بر یادگیری عمیق متکی است که شکل موج را تجزیه و تحلیل کرده و الگوهای معنی دار را استخراج می‌کند. سیستم ابتدا فایل گفتار را بارگیری می‌کند و آن را پردازش می‌کند تا ویژگی‌های مربوط به گفتار را استخراج کند. این ویژگی‌ها عبارتند از تراز واج، تغییرات زیر و بمی و ریتم، که تعیین می‌کند دهان و عضلات صورت چگونه باید در پاسخ به صدا حرکت کنند.

هنگامی که ویژگی ها استخراج می‌شوند، آن‌ها به مجموعه‌ای از ضرایب چهره از پیش تعریف شده نگاشت می‌شوند. این ضرایب بر اساس مدل‌های مورف‌پذیر سه‌بعدی (DMM3) هستند که تغییر شکل‌های صورت را به روشی ساختاریافته نشان می‌دهند. هر ضریب جنبه خاصی از حرکت صورت، مانند باز کردن دهان، گرد شدن لب‌ها یا حرکت فک را رمزگذاری می‌کند. سیستم این نقشه‌برداری‌ها را از مجموعه داده‌های بزرگ حاوی داده‌های جفت صدا و حرکت چهره می‌آموزد.

سیستم ابتدا شکل موج گفتار را پردازش کرده و ویژگی‌های آوایی را استخراج می‌کند، مانند اینکه کدام صداها گفته می‌شوند و شدت آن‌ها چقدر است. سپس این ویژگی‌های استخراج‌شده به یک مدل یادگیری عمیق وارد می‌شوند که ضرایب حرکت صورت را پیش‌بینی می‌کند. این ضرایب مشخص می‌کنند که هر قسمت از صورت چگونه باید برای هر بخش از صدا حرکت کند. در نهایت، ضرایب تولیدشده با مدت زمان و زمان‌بندی صدا هماهنگ می‌شوند تا حرکات لب‌ها با گفتار کاملاً منطبق باشند.

این مرحله شکاف بین انیمیشن صوتی و تصویری را پر می‌کند و اطمینان می‌دهد که کلمات گفتاری با حالات صورت و حرکات لب صحیح همراه هستند. بدون این فرآیند تبدیل، انیمیشن فاقد واقع‌گرایی خواهد بود، زیرا لب‌ها و دهان به درستی با گفتار مطابقت ندارند.

با نگاشت موفقیت آمیز صدا به ضرایب حرکت صورت، این مرحله پایه و اساس مرحله بعدی را ایجاد می‌کند، جایی که از این ضرایب برای متحرک سازی چهره استفاده می‌شود و یک انیمیشن گفتاری صاف و طبیعی ایجاد می‌کند.

1-4 شبکه Expnet

ExpNet ماژول است که ورودی صوتی را به ضرایب پویایی بیان چهره تبدیل می‌کند، که بخشی از نمایش مدل سه بعدی Morphable است. چالش اصلی آن این است که نگاشت صدا به بیان یک فرآیند یک به یک نیست. یک نشانه صوتی بسته به ویژگی‌های منحصر به فرد چهره فرد می‌تواند با عبارات مختلفی مطابقت داشته باشد. برای مقابله با این موضوع، ExpNet با ترکیب ضرایب بیان ($\beta 0$) از همان فریم اول به عنوان مرجع، عبارت تولید شده را در هویت سوژه قرار می‌دهد. این مرجع به حفظ ثبات در هویت چهره کمک می‌کند و در عین حال به شبکه اجازه می‌دهد تا بر روی ثبت جنبه‌های پویا حرکات صورت مرتبط با گفتار تمرکز کند.

از نظر فنی، ExpNet از یک رمزگذار صوتی مبتنی بر ResNet استفاده می‌کند که بخش‌های کوتاهی از صدا (که به صورت طیف‌نگاری-مل ۰.۲ ثانیه‌ای نشان داده می‌شود) را به یک فضای ویژگی پنهان تبدیل می‌کند. سپس این ویژگی‌ها از یک لایه نگاشت خطی عبور داده می‌شوند تا ضرایب بیان DMM3 مربوطه را تولید کنند. در اصل، ExpNet شکاف بین نشانه‌های صوتی خام و حالت‌های سه‌بعدی دقیق چهره را پر می‌کند و جزئیات حرکتی واقع‌گرایانه لازم برای لب‌زدن با کیفیت بالا را ارائه می‌کند.

2-4 شبکه PoseVAE

PoseVAE یک ماژول است که برای ایجاد حرکات واقعی و متنوع سر طراحی شده است که مکمل حالات صورت تولید شده توسط ExpNet است. به‌جای پیش‌بینی وضعیت‌های سر مستقیماً از روی صدا، که به دلیل همبستگی نسبتا ضعیف بین نشانه‌های صوتی و حرکات سر سراسری چالش‌برانگیز است. PoseVAE یاد می‌گیرد که حرکت باقیمانده را نسبت به حالت اولیه سر ایجاد کند. در عمل، حالت سر فریم اول به عنوان یک خط مبنا عمل می‌کند و PoseVAE پیش‌بینی می‌کند که چگونه این حالت باید در طول زمان در پاسخ به ورودی صوتی تغییر کند.

این ماژول بر روی یک فریم‌ورک رمزگذار خودکار متغیر شرطی (VAE) ساخته شده است. رمزگذار PoseVAE دنباله‌ای از حالت‌های سر (از یک فیلم آموزشی) را می‌گیرد. نکته مهم این است که این فضای پنهان نه تنها به دنباله ژست‌های سر، بلکه به ویژگی‌های صوتی و سیگنال هویت سبک مشروط است. این شرطی‌سازی به مدل اجازه می‌دهد تا تغییرات سبکی ظریف را در حرکت سر که مخصوص افراد یا زمینه‌های مختلف است، ثبت کند. سپس رمزگشا از این کد پنهان برای بازسازی توالی پوزهای سر استفاده می‌کند، اما به جای پیش‌بینی کامل وضعیت‌ها، تغییر باقی مانده از حالت اولیه را پیش بینی می‌کند.

فرآیند آموزش شامل چندین مولفه از دست دادن است: یک افت بازسازی (معمولاً میانگین مربعات خطا) تضمین می کند که باقیمانده های تولید شده، زمانی که به حالت اولیه اضافه می شوند، دقیقاً با حرکات سر حقیقت زمین مطابقت دارند. از دست دادن واگرایی KL فضای پنهان را تشویق می کند تا از توزیع گاوسی پیروی کند و نمونه برداری صاف و متنوع را در طول استنتاج تسهیل می کند. و یک ضرر خصمانه، واقع گرایی حرکات سر ایجاد شده را بیشتر اصلاح می کند. در نتیجه، PoseVAE می‌تواند طیف گسترده‌ای از حرکات طبیعی و مداوم سر را تولید کند که وقتی با صدا همگام‌سازی می‌شود، به یک ویدیوی کلی سر صحبت واقعی کمک می‌کند.

فصل 5: رندر کردن تصویر متحرک

پس از بدست آوردن ضرایب حرکت صورت از مرحله قبل، سیستم اکنون از آن‌ها برای متحرک‌سازی تصویر صورت استفاده می‌کند. این مرحله مسئول ایجاد یک چهره گفتاری واقعی و هماهنگ است که در آن حرکات دهان و حالات صورت دقیقاً با گفتار ورودی مطابقت دارند. فرآیند رندر، تصویر استاتیک اصلی را با ضرایب حرکت استخراج شده ترکیب می‌کند تا ویدیویی با حرکات و حرکات لب طبیعی به‌نظر برسد.

این مرحله تصویر برش خورده صورت را از مرحله ۱ و ضرایب حرکت صورت را از مرحله ۳ بازیابی می‌کند. سپس یک مدل یادگیری عمیق، تصویر استاتیک صورت را دریافت کرده و آن را فریم به فریم بر اساس ضرایب حرکت اصلاح می‌کند. در ادامه، دنباله‌ای از فریم‌ها تولید می‌شود که در آن، صورت به تدریج مطابق با صدای گفتار حرکت می‌کند. در نهایت، فریم‌های تولیدشده در قالب ویدیویی پردازش شده و با صدای اصلی همگام‌سازی می‌شوند.

این مرحله نهایی است که در آن تمام کارهای قبلی با هم جمع می‌شوند تا انیمیشن لب‌زنی مورد نظر را تولید کنند. موفقیت انیمیشن بستگی به این دارد که حرکات لب چقدر با گفتار مطابقت داشته باشد و حالات صورت به طور طبیعی تغییر کند. کیفیت انیمیشن تحت تاثیر دقت ضرایب تولید شده در مرحله 3 و پیچیدگی مدل رندر استفاده شده در این مرحله است. مدل‌های متحرک چهره مبتنی بر یادگیری عمیق، با حفظ هویت فرد در تصویر و در عین حال افزودن حرکات و حرکات واقعی، نتایج با کیفیت بالا را تضمین می‌کنند.

در خاتمه، این مرحله تصویر ثابت را زنده می کند و به نظر می رسد که فرد به طور طبیعی صحبت می‌کند. خروجی نهایی یک چهره کاملاً متحرک است که به طور واقع گرایانه با سخنرانی ارائه شده حرکت می‌کند و آن را به گامی مهم در ایجاد ویدیوهای لب‌زنی مبتنی بر هوش مصنوعی تبدیل می‌کند.