

کارایی برخی معیارهای کیفیت تصویر در بهینه‌سازی یک سامانه تبدیل گفتار

داود غروی‌ان

دانشیار

دانشکده مهندسی برق

دانشگاه شهید بهشتی

Gharavian@sbu.ac.ir

مجید بهداد*

دکترای تخصصی مخابرات

دانشکده مهندسی برق

دانشگاه شهید بهشتی

Ma_Behdad@sbu.ac.ir

چکیده

با وجود پیشرفت‌های چشمگیر در سامانه‌های تبدیل گفتار مبتنی بر یادگیری عمیق، هم‌راستا نبودن تابع هدف آموزش با معیارهای ارزیابی ادراکی همچنان یکی از موانع اصلی بهبود کیفیت خروجی است. از طرف دیگر، این حوزه به معیارهای عینی و استانداردشده‌تری برای ارزیابی کامل عملکرد خود نیاز دارد. در سامانه‌های تبدیل گفتار ناموازی مبتنی بر GAN، به‌ویژه در خانواده CycleGAN-VC و از جمله در MaskCycleGAN-VC، تمایزگرها معمولاً برای تشخیص واقعی یا جعلی بودن داده آموزش می‌بینند و به‌طور مستقیم با معیارهایی مانند فاصله مل‌کیسترال، فهم‌پذیری، شباهت گوینده یا طبیعی بودن گفتار همسو نمی‌شوند. این مقاله روشی را با عنوان MetricCycleGAN-VC ارائه می‌کند که با اصلاح تمایزگرهای ثانویه و افزودن خروجی پیوسته، امکان یادگیری رفتار متریک‌های ارزیابی را در قالب یک راهبرد جعبه‌سیاه فراهم می‌کند. در این چارچوب، متریک‌های MoNISQA، SECS، STOI، MoMCD و معیارهای تصویری SSIM، FSIM و GMSD برای هدایت آموزش به کار گرفته می‌شوند. ارزیابی‌ها بر روی دادگان VCC2018 و در چهار حالت تبدیل جنسیتی شامل زن به زن، زن به مرد، مرد به زن و مرد به مرد انجام شده است. نتایج نشان می‌دهد که به کارگیری هر متریک در آموزش، عمدتاً همان متریک را در خروجی بهبود می‌دهد. همچنین، معیار GMSD در میان متریک‌های تصویری توانسته است بهبودی همزمان و هماهنگ در چهار شاخص ارزیابی ایجاد کند. ارزیابی شنیداری نمونه نیز برتری روش پیشنهادی را نسبت به MaskCycleGAN-VC تأیید می‌کند.

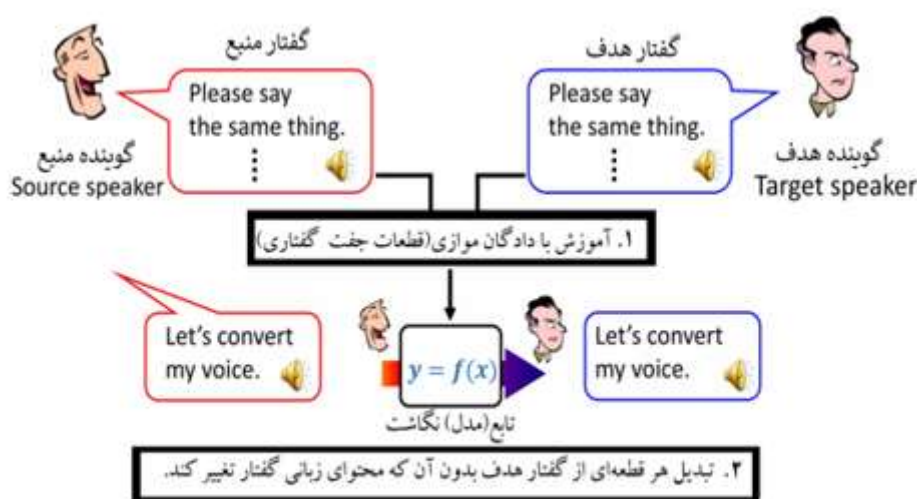
واژگان کلیدی: تبدیل گفتار، CycleGAN-VC، MaskCycleGAN-VC، بهینه‌سازی ادراکی، MetricGAN، NISQA، SECS و GMSD

۱. مقدمه

* نویسنده مسئول

تبدیل گفتار^۱ (VC)، فرایندی است که در آن گفتار یک گوینده‌ی مبدأ به گونه‌ای تغییر داده می‌شود که به گفتار گوینده‌ی هدف شباهت پیدا کند؛ در حالی که محتوای زبانی گفتار حفظ می‌شود. این فناوری در حوزه‌هایی مانند دوبله ویدئویی، شخصی‌سازی صدای سامانه‌های گفتاری، کمک به افراد دارای اختلال گفتاری، بهبود فهم‌پذیری گفتار و کاربردهای مرتبط با شناسایی یا تغییر هویت گوینده اهمیت دارد [۱-۲]. با این حال، کیفیت خروجی سامانه‌های تبدیل گفتار هنوز در همه شرایط با ادراک انسانی همسو نیست و در بسیاری از موارد، خروجی تبدیل‌شده از نظر طبیعی بودن، شباهت گوینده یا فهم‌پذیری^۲ با گفتار طبیعی فاصله دارد و از طرفی در ارزیابی‌های کامل عملکردی در این حوزه، نیاز به معرفی معیارهای عینی استاندارد شده‌تر و مرتبط‌تری با مفهوم کیفیت در VC است [۳-۴].

در نسل‌های اولیه‌ی تبدیل گفتار، اغلب به دادگان موازی یا جفت‌شده^۳ نیاز بود؛ یعنی مطابق شکل ۱، جملات یکسان باید توسط گوینده مبدأ و گوینده هدف ضبط می‌شد تا تابع نگاشت میان دو فضای صوتی آموخته شود.



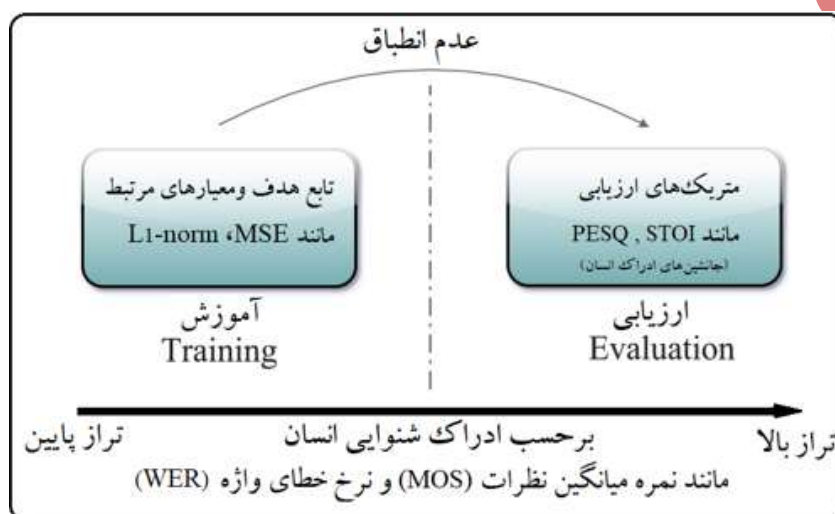
شکل ۱. فرایند یادگیری و تبدیل گفتار موازی با دادگان جفت

پس از یادگیری تابع نگاشت در مرحله‌ی ۱، همان گونه که در این شکل نشان داده شده است، در مرحله‌ی ۲، یعنی فاز تبدیل این تابع نگاشت، برای تبدیل گفتار و تغییر گوینده‌ی گفتار مورد استفاده قرار می‌گرفت. این نیاز، گردآوری داده را پرهزینه و زمان‌بر می‌کرد و استفاده از سامانه‌های تبدیل گفتار را در کاربردهای واقعی محدود می‌ساخت.

با ظهور روش‌های مبتنی بر یادگیری عمیق و به‌ویژه شبکه‌های مولد تخصصی، امکان آموزش با دادگان ناموازی^۴ فراهم شد. خانواده‌ی CycleGAN-VC یکی از رویکردهای مهم در این زمینه است؛ زیرا با استفاده از سازوکار ثبات حلقه^۵ و یادگیری تخصصی^۶، می‌تواند نگاشت میان دو دامنه گفتاری را بدون نیاز به جملات جفت‌شده بیاموزد [۵-۸]. با وجود این مزیت، در سامانه‌های گن‌محور^۷ و بسیاری از دیگر سامانه‌های تبدیل گفتار، یک مسأله‌ی اساسی باقی می‌ماند تابع هدف آموزش الزاماً با معیارهایی که در مرحله ارزیابی کیفیت استفاده می‌شوند؛ همسو نیست. به عبارت دیگر، مثلاً در همان سامانه‌های گن‌محور تمایزگر^۸ در مرحله آموزش معمولاً واقعی یا جعلی بودن نمونه را تشخیص می‌دهد؛ اما این تصمیم الزاماً به معنی بهبود امتیازهایی مانند MCD^۹ [۹]، STOI^{۱۰} [۱۰]، NISQA^{۱۱} [۱۱] یا شباهت گوینده به

عنوان معیارهای ارزیابی ادراکی بلاواسطه برای حوزه‌ی ارزیابی VC نیست. این مسأله در ادبیات پردازش گفتار با عنوان عدم تطابق میان تمایزگر و معیار ارزیابی^{۱۲} یا DEM شناخته می‌شود [۱۲-۱۳].

شکل ۲ شمایی کلی را از مسأله‌ی DEM یا همان مفهوم عدم انطباق بین معیارهای نهفته در تابع هدفی که در آموزش به کار می‌رود و معیارها و متریک‌هایی را که در ارزیابی نهایی خروجی سامانه استفاده می‌شود؛ نشان می‌دهد. طبعاً در نتیجه‌ی این عدم انطباق، ممکن است سامانه در ظاهر از نظر تابع زیان آموزش بهبود یابد؛ اما در معیارهای ادراکی نهایی کیفیت، پیشرفت کافی ایجاد نشود.



شکل ۲. عدم انطباق بین معیارهای به کار رفته در تابع هدف و متریک‌های ارزیابی نهایی عملکرد سامانه

هدف این مقاله ارائه و تحلیل چارچوبی برای کاهش این شکاف در تبدیل گفتار ناموازی در نمونه‌ای از سامانه‌های گن‌محور است و محور اصلی پژوهش، معرفی نحوه وارد کردن متریک‌هایی خاص از حوزه ارزیابی کیفی تصویر در بستری مانند MetricCycleGAN-VC [۱۴] به عنوان توسعه‌ای بر MaskCycleGAN-VC [۸] است. در این روش، تمایزگرهای ثانویه در سامانه‌ی پایه‌ی MaskCycleGAN-VC اصلاح می‌شوند تا علاوه بر نقش تخصصی معمول، بتوانند خروجی پیوسته‌ای متناظر با امتیاز متریک یا همان معیار کیفیت تولید کنند. سپس این خروجی با امتیاز واقعی متریک مقایسه می‌شود و به این ترتیب، تمایزگر به یک ارزیاب متریک‌آگاه^{۱۳} تبدیل می‌شود. نوآوری‌های اصلی مقاله عبارتند از: ارائه چارچوبی جعبه‌سیاه برای ورود متریک‌های ارزیابی گفتار به تابع هدف تبدیل گفتار ناموازی؛ اصلاح تمایزگرهای ثانویه برای تولید خروجی پیوسته متناسب با امتیاز متریک؛ استفاده از متریک‌های گفتاری شامل MoMCD^{۱۴} (همان متریکی که در [۱۵] با نام MMCD معرفی شده است)، STOI، SECS^{۱۵} [۱۴] و MoNISQA^{۱۶} [۱۴] برای پوشش جنبه‌های مختلف کیفیت تبدیل گفتار؛ بررسی امکان استفاده از معیارهای تصویری مانند FSIM^{۱۷} [۱۶]، SSIM^{۱۸} [۱۷] و به طور خاص GMSD^{۱۹} [۱۸] روی شبه‌جفت‌های مل‌اسپکتروگرام؛ و ارزیابی تجربی روش در چهار حالت تبدیل جنسیتی برون‌جنسیتی و درون‌جنسیتی بر روی دادگان VCC2018 با معیارهای ارزیابی عینی ادراکی و نیز روش ذهنی.

۲. مرور پژوهش‌های مرتبط

پژوهش‌های تبدیل گفتار از روش‌های آماری کلاسیک به مدل‌های عمیق و تخصصی تحول یافته‌اند. روش‌های اولیه مانند مدل‌های مبتنی بر GMM، HMM یا VQ نیازمند داده‌های موازی بودند و اغلب در شرایط محدود قابل استفاده بودند [۲] و [۱۹]. در ادامه، روش‌های مبتنی بر خودکدگذار^{۲۰} و جداسازی بازنمایی محتوا و گوینده^{۲۱} مطرح شد. این روش‌ها تلاش می‌کنند محتوای زبانی را از ویژگی‌های گوینده جدا کنند و سپس محتوای منبع را با ویژگی گوینده هدف ترکیب کنند. با این حال، جداسازی کامل محتوا و هویت گوینده دشوار است و نشت اطلاعات گوینده مبدأ می‌تواند کیفیت تبدیل را کاهش دهد [۲۰].

خانواده‌ی CycleGAN-VC راه‌حلی اولیه ولی مؤثر برای مسأله‌ی دادگان ناموازی ارائه کرد. مدل CycleGAN-VC با الهام از ترجمه تصویر به تصویر ناچفت^{۲۲} [۲۱] از نگاشت‌های دوطرفه و زیان ثبات حلقه استفاده می‌کند تا تبدیل میان دو دامنه گفتاری را بدون داده‌های موازی بیاموزد [۵]. مدل CycleGAN-VC2 با بهبود معماری و تابع زیان، کیفیت تبدیل را افزایش داد [۶]. مدل CycleGAN-VC3 با تمرکز بر تبدیل مل‌اسپکتروگرام، گامی دیگر در بهبود کیفیت برداشت [۷]. سپس MaskCycleGAN-VC با افزودن سازوکار پرکردن قاب‌ها (FIF)^{۲۳} و یادگیری خودنظارتی مبتنی بر پوشاندن قاب‌ها، توانست کیفیت تبدیل را در شرایط ناموازی بهبود دهد [۸].

با وجود این پیشرفت‌ها، بیشتر سامانه‌های گن‌محور^{۲۴} برای تبدیل گفتار، هنوز از تابع زیان یا همان تابع هدفی^{۲۵} استفاده می‌کنند که به طور مستقیم برای بهبود متریک‌های ارزیابی ادراکی طراحی نشده است. این مشکل شبیه چالشی است که در به‌سازی گفتار^{۲۶} نیز مطرح شده است و در آن معیارهایی مانند PESQ^{۲۷} [۲۲] و STOI ممکن است در ارزیابی نهایی استفاده شوند؛ اما تابع زیان آموزش بر معیارهای ساده‌تری مانند MSE یا L1 بنا شده باشد [۱۲]. فو و همکاران نشان دادند که بهینه‌سازی مستقیم معیارهای ارزیابی در پردازش سیگنال گفتار در قالب سامانه‌ای با نام MetricGAN [۱۳] می‌تواند برای افزایش کیفیت ادراکی خروجی مؤثر باشد. اما آیا بهره‌گیری از این تکنیک در تبدیل گفتار با دادگانی ناموازی و در شرایطی پیچیده که شامل تغییر گوینده است؛ باز هم ممکن است؟ اگر آری چگونه؟ مدل MetricGAN برای حل این مسأله در به‌سازی گفتار و باهدف وارد کردن متریک‌های مطرح در ارزیابی در بهینه‌سازی، یک رویکرد جعبه‌سیاه را معرفی کرد. در این رویکرد، تمایزگر تلاش می‌کند رفتار یک متریک ارزیابی را یاد بگیرد و مولد را به سمت تولید خروجی‌هایی با امتیاز بالاتر هدایت کند [۱۳]. با این حال، انتقال مستقیم این ایده به تبدیل گفتار ساده نیست؛ زیرا تبدیل گفتار ناموازی مبتنی بر CycleGAN شامل دو مولد^{۲۸}، چند تمایزگر، ساختار حلقوی، داده‌های ناچفت و قیود حفظ محتواست.

در حوزه‌ی تبدیل گفتار، معیار MCD یکی از رایج‌ترین ابزارهای ارزیابی فاصله طیفی است [۹]؛ با این حال، MCD تنها بخشی از کیفیت تبدیل گفتار را نشان می‌دهد و لزوماً به تنهایی بیانگر فهم‌پذیری، طبیعی بودن یا شباهت گوینده نیست. معیار STOI ابزاری است که برای سنجش فهم‌پذیری گفتار در شرایط نویزی طراحی شده است و NISQA [۱۱] نیز به عنوان یک مدل یادگیری عمیق برای ارزیابی غیرتهاجمی کیفیت گفتار مطرح شده است.

ایده‌ی استفاده از معیارهای تصویری در تبدیل گفتار از این مشاهده ناشی می‌شود که بسیاری از سامانه‌های جدید، به‌ویژه MaskCycleGAN-VC، بر روی ویژگی مل اسپکتروگرام^{۲۹} کار می‌کنند. مل اسپکتروگرام را می‌توان به عنوان تصویری برای بازنمایی ساختار زمانی فرکانسی سیگنال گفتار در نظر گرفت. بنابراین، معیارهای سنجش شباهت تصویر مانند FSIM، SSIM و GMSD می‌توانند برای مقایسه‌ی شبه‌جفت‌های مل اسپکتروگرام مورد بررسی قرار گیرند.

تبدیل گفتار (VC) را می‌توان ابتدا با استخراج اطلاعات محتوای منبع و اطلاعات گوینده هدف و سپس بازسازی شکل موج با این اطلاعات به‌دست‌آورد. با این حال، رویکردهای فعلی معمولاً یا اطلاعات ناخالص محتوا را با اطلاعات دارای نشتی از گوینده، استخراج می‌کنند و یا مقدار زیادی داده‌ی همراه با متن را برای آموزش طلب می‌کنند. علاوه بر این، کیفیت شکل موج بازسازی شده می‌تواند به دلیل عدم تطابق بین مدل تبدیل و وکودر، با تنزل همراه باشد. اما به هر روی این رویکرد پنجره‌ای می‌گشاید به سمت و سوی سامانه‌های تبدیل گفتار صفرنما و تک‌نما که در ادامه به بررسی ویژگی‌های شاخص مهمترین و به‌روزترین سامانه‌های ارایه‌شده در سال‌های اخیر برای رفع مشکلات و چالش‌های موجود در تبدیل گفتار صفرنما می‌پردازیم.

پس از نسل اولیه‌ی سامانه‌های یک‌به‌یک^{۳۰} و سپس یک‌به‌چند^{۳۱} برای تبدیل گفتار که عمدتاً بر پایه‌ی شبکه‌های مولد تخصصی یا همان گن، خودکدگذارها^{۳۲} و تفکیک ویژگی‌های گوینده از محتوای گفتار توسعه یافته بودند، تمرکز پژوهش‌ها به سمت یادگیری بازنمایی‌های مستقل از گوینده و مدل‌های بنیادین^{۳۳} معطوف شد. در این رویکردها، هدف اصلی استخراج بازنمایی‌هایی از گفتار است که اطلاعات زبانی را حفظ کرده؛ اما وابستگی به هویت گوینده را تا حد امکان حذف کنند.

نخستین تحول مهم، استفاده از یادگیری خودنظارتی (SSL)^{۳۴} بود. مدل‌هایی مانند HuBERT [۲۳]، WavLM [۲۴]، wav2vec2 [۲۵]، ContentVec [۲۶]، بدون نیاز به برچسب‌گذاری دستی، بازنمایی‌هایی غنی از محتوای گفتار استخراج کردند که در سامانه‌های تبدیل گفتار به عنوان کدگذار محتوا^{۳۵} مورد استفاده قرار گرفتند. این بازنمایی‌ها، نسبت به روش‌های مبتنی بر PPG و ASR، استقلال بیشتری از هویت گوینده داشته و کیفیت تبدیل را به طور محسوسی افزایش دادند [۲۶، ۲۴].

در ادامه، مدل Whisper [۲۷] علاوه بر کاربرد اصلی خود در بازشناسی گفتار به عنوان استخراج‌کننده ویژگی‌های معنایی مقاوم در برابر نویز و لهجه در سامانه‌های VC مورد استفاده قرار گرفت. ویژگی‌های استخراج شده توسط Whisper سبب افزایش پایداری محتوای زبانی در هنگام تبدیل و کاهش خطاهای تلفظ شدند، به‌ویژه در سامانه‌های چندزبانه و تبدیل صفرنمایی^{۳۶} که داده آموزشی محدودی در اختیار دارند. از سال ۲۰۲۴، گرایش جدیدی به سمت مدل‌های انتشار^{۳۷} و تطبیق جریان^{۳۸} شکل گرفت. برخلاف گن‌ها که آموزش ناپایداری داشتند و گاهی زواید صوتی تولید می‌کردند؛ مدل‌های انتشار با فرآیند تدریجی حذف نویز توانستند کیفیت طبیعی‌تر، پیوستگی بهتر و شباهت بیشتر به گوینده‌ی مقصد ایجاد کنند. بسیاری از سامانه‌های جدید، کدگذارهای SSL را با مدل‌های انتشار ترکیب کردند تا هم محتوای گفتار و هم ویژگی‌های آکوستیکی با دقت بیشتری بازسازی شوند. یکی از مهم‌ترین دستاوردهای سال ۲۰۲۴، معرفی WavLLM [۲۸] بود. این مدل در واقع یک مدل بزرگ زبانی گفتاری محسوب می‌شود که دو کدگذار مستقل

را به کار می‌گیرد؛ کدگذار Whisper برای استخراج اطلاعات معنایی و کدگذار WavLM [۲۴] برای استخراج ویژگی‌های هویتی گوینده. سپس این اطلاعات به یک مدل زبانی بزرگ منتقل شده و با استفاده از یادگیری چندمرحله‌ای و لورای پرامپت‌آگاه^{۳۹} امکان انجام همزمان وظایفی مانند تشخیص گفتار، تشخیص گوینده، ترجمه گفتار و درک صوت فراهم می‌شود. در سال ۲۰۲۴، سامانه‌ای مبتنی بر مدل‌های بزرگ زبانی گفتاری^{۴۰} (SLLM) با نام WavLLM، معرفی شد که هرچند مستقیماً یک سامانه‌ی VC نیست؛ اما معماری آن به عنوان زیرساخت نسل جدید سامانه‌های تبدیل گفتار مورد استفاده قرار گرفته است [۲۸].

در سال ۲۰۲۵، سامانه‌ی Vevo [۲۹] یکی از پیشرفته‌ترین سامانه‌های تبدیل گفتار و تقلید صدا معرفی شد. این سامانه با استفاده از HuBERT [۲۳] و یک ساختار VQ-VAE، سه مؤلفه‌ی اصلی گفتار یعنی محتوای زبانی، سبک گفتار و هویت گوینده را به صورت کاملاً خودنظارتی از یکدیگر جدا می‌کند. سپس یک ترانسفورمر^{۴۱} خودبازگشتی برای مدل‌سازی محتوا و سبک و یک ترانسفورمر تطبیق جریان، برای تولید ویژگی‌های آکوستیکی به کار گرفته می‌شود. Vevo سامانه‌ی بدون نیاز به داده‌ی موازی یا آموزش اختصاصی برای هر گوینده، در تبدیل لهجه، احساس، تقلید صدا و تبدیل گفتار صفرنما، عملکردی هم‌تراز یا بهتر از روش‌های پیشین ارائه کرده است [۲۹].

در سال ۲۰۲۶، روند توسعه سیستم‌های تبدیل گفتار بیش از پیش به سمت مدل‌های بنیادین گفتاری و مدل‌های بزرگ زبانی گفتاری (SLLM) حرکت کرد. معماری‌های جدید معمولاً ترکیبی از کدگذارهای SSL مانند HuBERT یا Whisper، SLLM، انتشار یا تطبیق جریان یا وکودرهای عصبی^{۴۲} هستند. این سامانه‌ها علاوه بر تبدیل گوینده، امکان کنترل همزمان سبک، احساس، لهجه، سرعت بیان و حتی تولید گفتار بر اساس دستورهای متنی را فراهم کرده‌اند. همچنین قابلیت تبدیل گفتار هرکس به هرکس^{۴۳} و تقلید صدای صفرنما و یادگیری بدون داده موازی، به ویژگی استاندارد این نسل تبدیل شده است. در مجموع، روند تکامل تبدیل گفتار از مدل‌های مبتنی بر گن و خودکدگذار به سمت استفاده از بازنمایی‌های خودنظارتی، مدل‌های انتشار، مدل‌های تطبیق جریان و مدل‌های بزرگ گفتاری حرکت کرده است. این تحول موجب افزایش طبیعی بودن گفتار تولیدی، حفظ بهتر محتوای زبانی، شباهت بیشتر به گوینده مقصد و امکان انجام تبدیل صفرنمونه‌ای شده و مبنای اغلب سیستم‌های پیشرفته‌ی VC در سال‌های ۲۰۲۵ و ۲۰۲۶ را تشکیل می‌دهد [۲۹]. با این حال و با همه پیشرفت‌ها، هنوز سامانه‌های تبدیل گفتار مبتنی بر گن در برخی کاربردهای پردازش گفتار مطرح هستند و تحقیقات در این حوزه‌ها ادامه دارد؛ مثلاً در [۳۰] از یک سامانه‌ی تبدیل گفتار چنددامنه‌ای گن‌محور برای تبدیل گفتار خشن^{۴۴} به عادی استفاده شده است. این مطالعه امکان‌پذیری روش‌های تبدیل گفتار را در بهبود کیفیت گفتار صداها گرفته نشان می‌دهد. از طرفی دیگر می‌توان گفت یکی از چالش‌های باقیمانده و موجود در حوزه‌ی دانش و فناوری تبدیل گفتار، ارزیابی عینی است که به معیارهای عینی و استانداردشده‌تری برای ارزیابی کامل عملکرد خود نیاز دارد. چرا که روش‌های ارزیابی سنتی، و معیارهای آن مانند فاصله‌ی مل کپسترال (MCD)، خطای جذر میانگین مربعات فرکانس پایه^{۴۵} و فاصله‌ی طیفی مدوله شده^{۴۶} (MSD) عمدتاً بر ویژگی‌های ادراکی تمرکز دارند و اغلب ویژگی‌های تولید گفتار را نادیده می‌گیرند. این موضوع می‌تواند به ناهماهنگی بین شباهت ادراک شده‌ی صدا و جنبه‌های فیزیولوژیکی آن منجر شود و در نتیجه به روش‌های ذهنی مانند نمره نظر میانگین^{۴۷} (MOS) اتکا

شود. اگرچه MOS بینش ارزشمندی ارائه می‌دهد؛ اما پرهزینه و ذاتاً ذهنی است که کاربرد گسترده آن را محدود می‌کند. این پژوهش چارچوبی عینی را نیز برای ارزیابی کیفیت صدا در سامانه‌های تبدیل گفتار طرح و دنبال می‌کند [۳].

۳. روش پیشنهادی

۳-۱. مسأله‌ی پژوهش

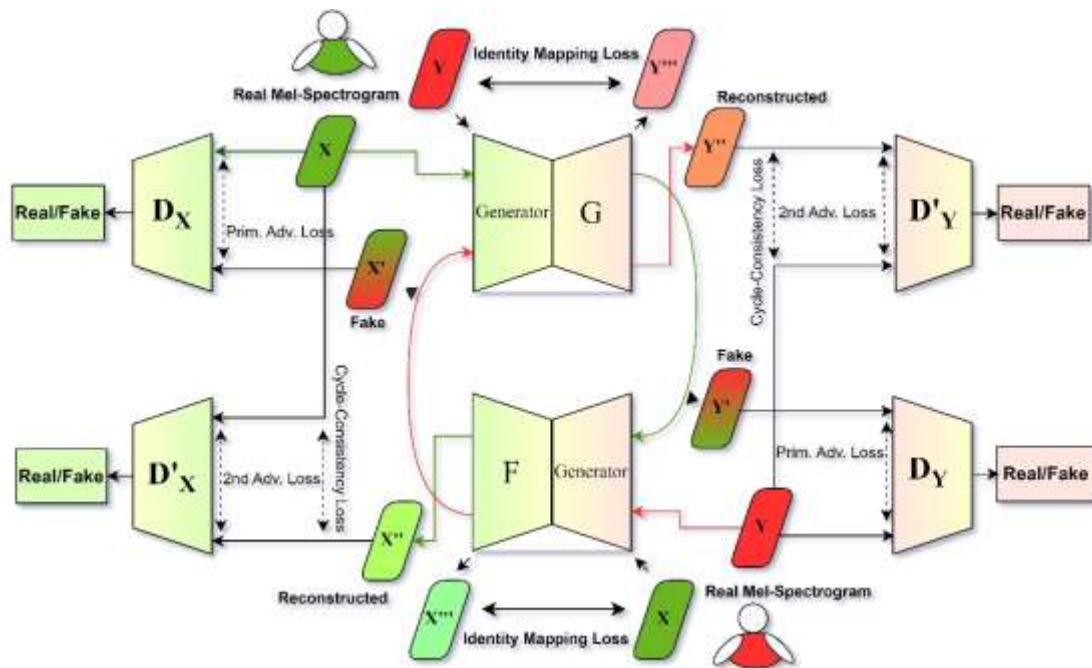
همان گونه که در بالا اشاره شد علی‌رغم پیشرفت‌های چشم‌گیر در VC هنوز توسعه روش‌ها و معیارهای ارزیابی و نیز به‌کارگیری آن معیارهای مطلوب ارزیابی در آموزش سامانه‌ها، جای کار بیشتری دارد. مثلاً MaskCycleGAN-VC، به عنوان به طور نمونه‌ای از سامانه‌های تبدیل گفتار گن‌محور، هدف اصلی یادگیری نگاشت میان دامنه گفتار گوینده مبدأ و گوینده هدف است. در این چارچوب، مولدها تلاش می‌کنند مل‌اسپکتروگرام مبدأ را به دامنه هدف و سپس مجدداً به دامنه مبدأ برسانند و آن را بازسازی کنند. توابع زیان اصلی در این سامانه‌ی VC شامل زیان تخصمی، زیان ثبات حلقه و زیان نگاشت هویت^{۴۸} هستند. این زیان‌ها برای حفظ ساختار کلی تبدیل و پایداری آموزش ضروری‌اند؛ اما به طور مستقیم تضمین نمی‌کنند که خروجی از نظر MCD، STOI، SECS یا NISQA بهتر شود.

مسأله‌ی اصلی این تحقیق آن است که چگونه می‌توان بدون تغییر ساختار درونی و فرمول‌بندی متریک‌های ارزیابی، آن‌ها را وارد آموزش یک سامانه‌ی تبدیل گفتار ناموازی مانند سامانه‌های گن‌محور کرد. پاسخ پیشنهادی، تبدیل تمایزگرهای ثانویه به ارزیاب‌های متریک‌آگاه در این سامانه‌ی گن‌محور است. در این روش، تمایزگر نه تنها واقعی یا جعلی بودن داده را یاد می‌گیرد؛ بلکه می‌آموزد امتیازی پیوسته متناظر با کیفیت نمونه آن هم از منظر عمده‌ترین معیارهای ارزیابی کیفی تبدیل گفتار بازسازی‌شده، تولید کند. این راه‌کار را می‌توان در خانواده‌ی بزرگ سامانه‌های تبدیل گفتار گن‌محور بخصوص سامانه‌های مبتنی بر CycleGAN-VC به کار برد.

۳-۲. ساختار کلی پیشنهادی MetricCycleGAN-VC

هرچند MetricCycleGAN-VC بر پایه‌ی سامانه‌ی MaskCycleGAN-VC [۸] طراحی شده است؛ اما از منظر بهینه‌سازی و تغییر در مدل، تفاوت‌های چشمگیری با آن دارد. در سامانه‌ی پایه، دو مولد برای نگاشت‌های دوطرفه میان دامنه‌های X و Y و چند تمایزگر برای کنترل کیفیت تخصمی وجود دارد. در حالی که در روش پیشنهادی، تمایزگرهای ثانویه اصلاح می‌شوند؛ این تمایزگرها به جای آن که تنها خروجی دودویی^{۴۹} یا PatchGAN تولید کنند؛ خروجی پیوسته‌ای میان صفر و یک نیز ارائه می‌دهند. این خروجی پیوسته باید رفتار متریک انتخاب‌شده را بازنمایی کند.

شکل ۳ تصویری از نحوه عملکرد مدل گن حلقوی را برای تبدیل گفتار ناموازی ارائه می‌دهد. همان طوری که در شکل ۳ قابل مشاهده است؛ شبکه‌ی پایه‌ی ما برای تبدیل گفتار، شبکه‌ای است با دو مولد و چهار تمایزگر. در بیانی ساده، مولد G در این ساختار قرار است مل‌اسپکتروگرام گفتار گوینده مبدأ (X) را به مل‌اسپکتروگرام گفتار گوینده مقصد (Y) تبدیل کند؛ اما در گام‌های نخستین یادگیری فعلاً به چیزی شبیه

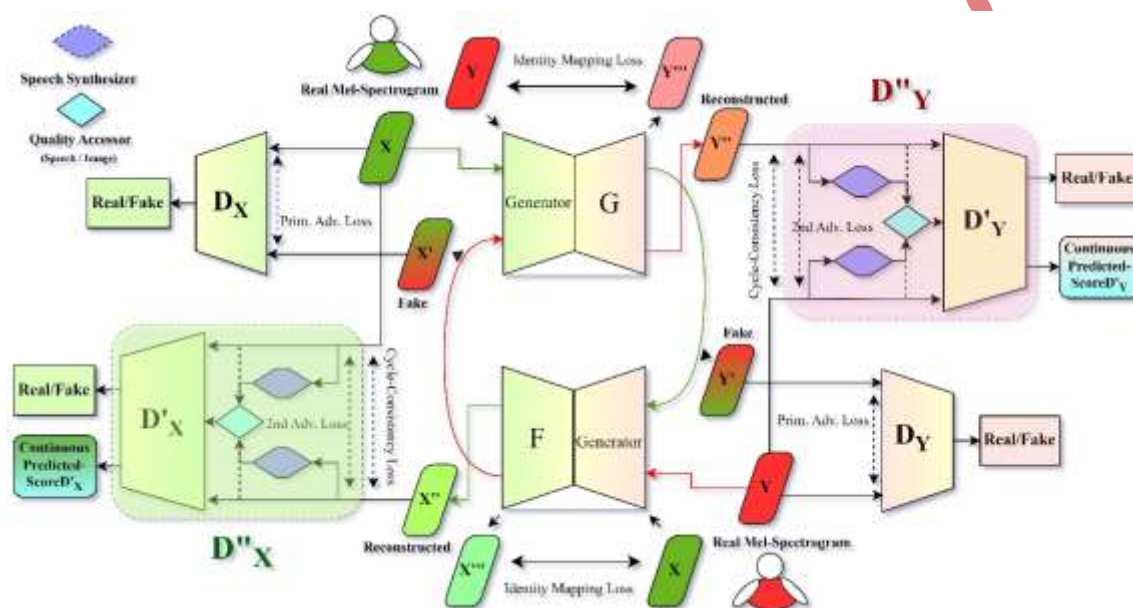


شکل ۳. ساختار و مدل مفهومی آموزش در تبدیل گفتار گن محوری مانند MaskCycleGAN-VC.

آن یعنی (Y') تبدیل می‌کند. برای نزدیک شدن هرچه بیشتر Y' به Y در گام‌های بعدی، تمایزگر اولیه D_Y به این فرآیند کمک می‌کند. برای تضمین انجام درست این فرآیند تبدیل و استفاده بیشتر و کامل‌تر از اطلاعات موجود در زوج در دسترس X و Y در آموزش سامانه، در طرف مقابل و در جهتی معکوس مولد F تلاش می‌کند به سهم خود، این بار مل‌اسپکتروگرام گفتار گوینده مبدأ (Y) را به مل‌اسپکتروگرام گفتار گوینده مقصد (X) تبدیل کند؛ اما در گام‌های نخستین یادگیری فعلا به چیزی شبیه آن یعنی (X') تبدیل می‌کند. برای نزدیک شدن هرچه بیشتر X' به X در گام‌های بعدی، تمایزگر اولیه D_X به این فرآیند کمک می‌کند. در این ساختار، نمونه‌ی بازسازی‌شده X'' با نمونه‌ی اصلی X و نمونه‌ی بازسازی‌شده Y'' با نمونه‌ی اصلی Y یک شبه‌جفت^{۵۰} تشکیل می‌دهند. اما X'' و Y'' چیستند؟ ظرفیت دیگری که در این ساختار بخصوص مسأله وجود دو مولد فراهم می‌کند؛ ایجاد یک حلقه برای کمک بیشتر به حفظ محتوای گفتار در کنار تبدیل گوینده است به ترتیبی که انتظار داشته باشیم در یک حلقه ناشی از عبور X از G و سپس عبور نتیجه از F به همان X اولیه برسیم. این خواسته امکان با تکیه بر "تابع زیان ثابت حلقه" ممکن می‌شود. از طرف دیگر می‌توان در صورت کارکرد صحیح این سامانه‌ی تبدیل گفتار انتظار داشته باشیم هنگامی که Y را به ورودی مولد G می‌دهیم؛ در خروجی دقیقا Y و یا حتی الامکان مشابه‌اش را به ما تحویل دهد و به طور مشابه اگر X را به ورودی مولد F بدهیم؛ در خروجی دقیقا X و یا حتی الامکان مشابه‌اش را به ما تحویل بدهد این خواسته با اعمال تابع زیانی با نام "تابع زیان نگاشت همانی" برای یادگیری مولدهای شبکه ممکن می‌شود. اما اهمیت و کاربرد شبه‌جفت‌هایی که در بالا بصورت زوج مرتب X و X' و Y و Y'' مطرح شد کجاست؟

برای متریک‌های تمام‌مرجع^{۵۱} مانند STOI، MoMCD، FSIM، SSIM و GMSD، این شبه‌جفت‌ها مبنای محاسبه‌ی متریک‌ها هستند. برای متریک نیمه‌تهاجمی^{۵۲} SECS نیز دو ورودی گفتاری مورد نیاز است. برای MoNISQA، که

معیاری غیرتهاجمی^{۵۳} است؛ تنها نمونه‌ی تولیدشده کافی است. بدین ترتیب، روش پیشنهادی می‌تواند متریک‌های تمام‌تهاجمی، نیمه‌تهاجمی و غیرتهاجمی را در یک چارچوب مشترک استفاده کند و این مزیت بزرگی برای MetricCycleGAN-VC است. ساختار شبکه پیشنهادی MetricCycleGAN-VC برای هدف وارد کردن متریک‌ها در تابع زیان، به شکل ۴ است. در تصویر شکل ۴، عمده تفاوت اساسی در دو تمایزگر ثانویه در سامانه‌ی پیشنهادی در مقایسه با سامانه‌ی پایه‌ی شکل ۳ دیده می‌شود. همان طوری که در این شکل قابل مشاهده است دو واحد تبدیل مل‌اسپکتروگرام به شکل‌موج گفتار و یک واحد ارزیاب کیفی، در ورودی دو تمایزگر آن اضافه شده است.



شکل ۴. ساختار آموزش در MetricCycleGAN-VC: تمایزگرهای ثانویه اصلاح‌شده با مازول‌ها و خروجی ثانویه که به ساختار MaskCycleGAN-VC افزوده شده‌است تا دو تمایزگر متریک‌آگاه در سمت راست و بالا و در چپ و پایین این شکل، ایجاد شوند.

۳-۳. تابع زیان متریک‌آگاه

در گن‌های معمول، نمونه‌های بازسازی‌شده یا تولید شده اغلب با برچسب ثابت جعلی (یا از لحاظ عددی صفر)، وارد تابع زیان می‌شوند. در MetricCycleGAN-VC، این برچسب ثابت با امتیاز متریک جایگزین می‌شود. سپس در گام بعدی، خروجی پیوسته تمایزگر با این مقدار مقایسه می‌شود. اگر Q نشان‌دهنده‌ی متریک به‌هنگار شده‌ی کیفیت باشد؛ مقدار $Q(x'', x)$ برای شبه‌جفت بازسازی‌شده و مرجع به عنوان برچسب پیوسته در آموزش تمایزگر استفاده می‌شود. در این تحقیق، Q می‌تواند حاوی یا بیانگر رابطه‌ی ریاضی یک متریک ارزیابی مانند MoMCD یا نماینده‌ی یک بلوک دیاگرام ارزیابی و یا حتی یک شبکه‌ی ارزیاب مانند NISQA باشد. به صورت مفهومی، زیان متریک‌آگاه از دو بخش تشکیل می‌شود: بخش نخست، تمایزگر را برای نمونه واقعی به امتیاز یک نزدیک می‌کند. بخش دوم، تمایزگر را وادار می‌کند امتیاز نمونه‌ی بازسازی‌شده را مطابق مقدار متریک واقعی آن پیش‌بینی کند. در نتیجه، تمایزگر ثانویه از یک دسته‌بند ساده‌ی واقعی جعلی، به یک ارزیاب کیفیت تبدیل می‌شود. این ارزیاب از طریق گرادیان خود مولد را به سمت تولید نمونه‌هایی هدایت می‌کند که امتیاز متریک بالاتری دارند. برای متریک‌های گفتاری، تابع $S(\cdot)$

به معنی تبدیل مل اسپکتروگرام ورودی این تابع، به شکل موج گفتار در خروجی آن است. بنابراین، برای MoMCD یا STOI، امتیاز متریک بر روی $S(x)$ و $S(x'')$ محاسبه می‌شود. در عمل شبکه‌ی MelGAN برای تبدیل مل اسپکتروگرام‌ها به سیگنال گفتار استفاده می‌شود. برای متریک‌های تصویری SSIM، FSIM و GMSD، خود مل اسپکتروگرام‌ها به عنوان ورودی استفاده می‌شوند. این تفاوت مهم است؛ زیرا متریک‌های تصویری می‌توانند هزینه‌ی محاسباتی مربوط به تبدیل مکرر به سیگنال گفتار را کاهش دهند.

۳-۴. متریک‌های مورد استفاده

متریک یا معیار MoMCD، نسخه‌ی اصلاح‌شده‌ی و به‌هنگار شده MCD است که برای استفاده در تابع هدف طراحی شده است. معیار MCD، فاصله‌ی طیفی میان گفتار تبدیل‌شده و گفتار مرجع را با مقداری بین صفر تا بینهایت می‌سنجد؛ در شرایطی که مقدار کمتر آن نشان‌دهنده کیفیت بهتر است. مثلاً برای MoMCD به عنوان یک متریک تهاجمی تابع Q از MCD پس از محاسبه‌ی آن و پس از معکوس کردن و نرمالیزه کردن آن به دست می‌آید.

در این روابط هنگام مقایسه‌ی دو شکل موج یا فایل گفتار (یکی تبدیل‌یافته و دیگری مرجع با مدت زمان برابر)، MCD میانگین فاصله‌ی بین بردارهای MFCC مربوط به آن‌ها را با استفاده از معادله‌ی ۱ بر حسب دسی‌بل محاسبه می‌کند [۹]:

$$MCD(Y, \hat{Y}) = 10 / \ln(10) \sqrt{(Y - \hat{Y})(Y - \hat{Y})} \quad [\text{dB}] \quad (1)$$

که Y و $\hat{Y}$ به ترتیب نشان‌دهنده‌ی بردارهای ویژگی MFCC از گفتار گوینده‌ی هدف و تبدیل‌یافته هستند [۷۰]. کاهش MCD، نشان‌دهنده‌ی افزایش شباهت بین گفتار سنترشده و گفتار طبیعی است. در نتیجه، اگر یک مقداری کاهش یافت، نشان‌دهنده‌ی عملکردی بهبودیافته، برای VC است. MCD از مقدار صفر (کیفیت عالی) تا بی نهایت (کیفیت پایین)، تغییر می‌کند. ما از یک تابع نگاشت معکوس، برای استانداردسازی MCD به عنوان اصلاح‌شده (MoMCD)، در فرآیندهای بهینه‌سازی که از ۰ تا ۱ متغیر است استفاده کردیم که در آن صفر به عنوان بدترین کیفیت و یک به عنوان بالاترین کیفیت تبدیل، است. انتخاب یک تابع نگاشت معکوس و غیرخطی طبق رابطه‌ی ۲ که رفتار MCD را به MoMCD نگاشت می‌کند، منجر به نتایج تجربی برتری، در مقایسه با سایر نگاشت‌های غیرخطی ممکن شد:

$$\text{MoMCD}(Y, \hat{Y}) = 1 - \left[\frac{MCD(Y, \hat{Y})}{MCD_{\max}} \right] \quad (2)$$

که در آن $MCD_{\max}$ بالاترین مقدار MCD در بین تمام نمرات به‌دست‌آمده از همه‌ی قطعه‌های گفتار مرتبط به یک گوینده در داده‌های آموزشی است. بنابراین در این تحقیق MoMCD به عنوان یک نمونه از Q قابل طرح و استفاده است.

متریک STOI معیاری برای فهم‌پذیری گفتار با عددی بین صفر تا یک است که به عنوان مثال دیگری از تابع Q قابل استفاده است. هرچند STOI اساساً برای گفتار نویزی طراحی شده است؛ اما در این چارچوب برای بررسی فهم‌پذیری گفتار تبدیل‌شده به کار می‌رود. این نحوه‌ی کاربرد را می‌توان در قالب نگرشی نو به این معیار مطرح کرده و در پژوهش‌های مستقل و جدید توسعه داد. بهینه‌سازی این معیار برای این کاربرد جدید، می‌تواند زمینه پژوهشی جدیدی در حوزه‌ی ارزیابی‌های کیفی گفتار در کنار معیارهایی چون WER^{۵۴} باشد. معیار SECS شاخص شباهت کسینوسی بردار جاسازی گوینده است. این معیار برای سنجش شباهت گوینده میان گفتار تبدیل‌شده و گفتار هدف استفاده می‌شود. متریک MoNISQA نسخه‌ی به‌هنگار شده‌ی NISQA است که برای ارزیابی کیفیت گفتار از منظر طبیعی بودن است که به صورت غیرتهاجمی در تابع هدف استفاده می‌شود [۱۱].

با الهام گرفتن از [۲۱] در ترجمه‌ی تصویر به تصویر و نیز بهره‌گیری از رویکرد متریک‌آگاه آن در [۳۱]، معیارهای FSIM، SSIM و GMSD از حوزه‌ی ارزیابی کیفیت تصویر وارد چارچوب پیشنهادی ما شده‌اند و بر روی مل‌اسپکتروگرام‌ها اعمال می‌شوند. هدف از استفاده این متریک‌ها در این تحقیق بررسی آن فرض است که شباهت تصویری میان مل‌اسپکتروگرام‌های بازسازی‌شده و مرجع می‌تواند با کیفیت صوتی همبستگی داشته باشد.

۳-۵. معماری شبکه و آموزش

معماری مولد در مدل پیشنهادی که آن را MetricCycleGAN-VC نامیدیم؛ مشابه‌ی مدل MaskCycleGAN-VC است. ورودی مولد، مل‌اسپکتروگرام پوشانده‌شده^{۵۵} همراه با اطلاعات ماسک^{۵۶} است. تعداد باندهای مل‌اسپکتروگرام ۸۰ و تعداد فریم‌های هر نمونه‌ی آموزشی ۶۴ است. سازوکار پرکردن قاب‌ها (FIF) سبب می‌شود مدل از طریق یادگیری خودنظارتی، ساختار زمانی فرکانسی گفتار را بهتر بیاموزد که البته این سازوکار در سامانه‌ی پایه هم وجود دارد. تمایزگرهای معمولی از معماری PatchGAN برای خروجی دودویی یا همان دوحالتی استفاده می‌کنند. در تمایزگرهای ثانویه‌ی اصلاح‌شده، یک خروجی پیوسته نیز افزوده می‌شود. برای تولید این خروجی، در لایه‌ی پایانی از ساختاری تمام‌متصل^{۵۷} همراه با تابع فعال‌سازی سیگموید^{۵۸} استفاده می‌شود تا امتیازی در بازه‌ی صفر تا یک تولید شود. این امتیاز با مقدار متریک محاسبه‌شده مقایسه می‌شود و زیان متریک‌آگاه را تشکیل می‌دهد.

تابع هدف کل MetricCycleGAN-VC شامل زیان‌های تخصصی اصلی، زیان ثبات حلقه، زیان نگاشت هویت، زیان‌های تخصصی ثانویه و زیان متریک‌آگاه است. برای ایجاد تعادل میان حفظ هدف اصلی تبدیل گفتار و افزایش اثر متریک‌ها، وزن‌های مربوط به ثبات حلقه و نگاشت هویت تعدیل شده‌اند و وزن زیان متریک‌آگاه نیز به تابع هدف اضافه شده است.

۴. شرایط آزمایش

۴-۱. داده‌گان

آزمایش‌ها بر روی دادگان VCC2018 انجام شده‌اند. تمرکز بر بخش Spoke از این دادگان است که برای سناریوی تبدیل گفتار ناموازی مناسب است. در این بخش، گویندگان مبدأ و هدف در مرحله آموزش جملات متفاوتی دارند؛ اما در مرحله آزمون، جملات یکسان برای ارزیابی استفاده می‌شود. چهار گوینده انتخاب شده‌اند: SF به عنوان گوینده زن مبدأ، SM به عنوان گوینده مرد مبدأ، TF به عنوان گوینده زن هدف و TM به عنوان گوینده مرد هدف. چهار حالت تبدیل بررسی شده است SF-TF: به عنوان زن به زن، SF-TM به عنوان زن به مرد، SM-TF به عنوان مرد به زن و SM-TM به عنوان مرد به مرد. برای هر گوینده، ۸۱ قطعه گفتار برای آموزش و ۳۵ جمله موازی برای آزمون استفاده شده است. نرخ نمونه‌برداری گفتارها به ۲۲/۰۵ کیلوهرتز کاهش داده شده است. مل اسپکتروگرام‌های لگاریتمی ۸۰ بانده با طول پنجره ۱۰۲۴ و طول پرش ۲۵۶ نمونه استخراج شده‌اند.

۲-۴. تنظیمات آموزش

برای مقایسه‌ی منصفانه، روش تبدیل و سنتز در سامانه‌ی پایه و سامانه‌ی پیشنهادی یکسان نگه داشته شده است. مدل مل‌گن^{۵۹} به عنوان وکودر^{۶۰} برای تولید شکل موج گفتار استفاده شده است. اندازه‌ی دسته^{۶۱} برابر یک و هر نمونه آموزشی شامل ۶۴ فریم است. برای سامانه‌ی پایه، وزن زیان ثابت حلقه برابر ۱۰ و وزن زیان نگاشت هویت برابر ۵ در نظر گرفته شده است. در سامانه‌ی پیشنهادی، این وزن‌ها تعدیل شده‌اند تا اثر زیان متریک‌آگاه در آموزش ظاهر شود. وزن زیان متریک‌آگاه برابر ۱۰ تنظیم شده است. شبکه‌ها با بهینه‌ساز Adam آموزش داده شده‌اند. نرخ یادگیری برای مولد و تمایزگر به ترتیب ۰/۰۰۰۱ و ۰/۰۰۰۲ تنظیم شده است. آموزش عمدتاً تا ۱۰۰۰ دور^{۶۲} ادامه یافته و در برخی موارد برای بررسی پایداری تا ۲۴۰۰ دور نیز دنبال شده است. به دلیل محدودیت امکانات پردازشی، آموزش تا ۶۰۰۰ دور، مشابه برخی شرایط ایدآل سامانه‌ی پایه، انجام نشده است و در آزمایش‌های متعدد نمونه با بالاتر از ۱۰۰۰ دور، این عدد برای نتیجه‌گیری معتبر از آزمایش‌ها مناسب و کافی تشخیص داده شد.

۳-۴. معیارهای ارزیابی

برای ارزیابی، چهار معیار اصلی MCD، STOI، SECS و MoNISQA استفاده شده است. معیار MCD برای سنجش فاصله‌ی طیفی استفاده شده که مقدار کمتر آن بهتر است. معیار STOI برای فهم‌پذیری، SECS برای شباهت گوینده و MoNISQA برای ارزیابی کیفیت از منظر طبیعی بودن گفتار به کار رفته است. در آزمایش‌های متریک‌آگاه، هر بار سامانه با یکی از متریک‌های MoMCD، STOI، SECS، MoNISQA یا یکی از معیارهای تصویری آموزش داده شده و سپس خروجی با چهار معیار اصلی ارزیابی شده است.

۵. نتایج

۵-۱. سامانه‌ی پایه

جدول ۱ نتایج سامانه‌ی پایه‌ی MaskCycleGAN-VC را در چهار حالت تبدیل نشان می‌دهد.

جدول ۱. نتایج سامانه پایه MaskCycleGAN-VC

معیار	SF-TM	SF-TF	SM-TF	SM-TM
MCD	5.8422	5.6913	6.9013	6.3281
STOI	0.5135	0.4385	0.3681	0.4978
SECS	0.8824	0.8488	0.8186	0.8698
MoNISQA	0.7261	0.7547	0.7298	0.6582

نتایج این جدول نشان می‌دهد که رفتار سامانه‌ی پایه در حالت‌های مختلف تبدیل یکسان نیست. برای مثال، مقدار STOI در حالت SM-TF کمتر از سایر حالت‌هاست. در حالی که در حالت SF-TM این متریک مقدار بالاتری دارد. این تفاوت می‌تواند ناشی از تفاوت ذاتی گویندگان، ویژگی‌های صوتی دادگان یا دشواری متفاوت نگاشت در حالت‌های برون‌جنسیتی و درون‌جنسیتی باشد.

۵-۲. نتایج بررسی سامانه‌ی پیشنهادی

الف_ نتایج متریک آگاه تبدیل SF-TM

جدول ۲ نتایج آزمایش‌های متریک آگاه سامانه‌ی پیشنهادی را مرتبط با تبدیل SF-TM نشان می‌دهد:

جدول ۲. نتایج متریک آگاه در حالت SF-TM

معیار	MoMCD	STOI	SECS	MoNISQA	FSIM	پایه
MCD	5.780	5.724	5.676	5.928	5.702	5.842
STOI	0.517	0.527	0.523	0.500	0.516	0.513
SECS	0.885	0.892	0.891	0.864	0.888	0.882
MoNISQA	0.694	0.708	0.680	0.740	0.692	0.726

در حالت SF-TM، آموزش متریک آگاه باعث شد هر متریک در حالت متناظر خود عمدتاً بهبود یابد. به طور نمونه، در حالت MoMCD-aware، مقدار MCD از ۵/۸۴۲ در سامانه‌ی پایه به ۵/۷۸۰ کاهش یافته است. در حالت STOI-aware، مقدار STOI از ۰/۵۱۳ به ۰/۵۲۷ افزایش یافت و در حالت SECS-aware، مقدار SECS از ۰/۸۸۲ به ۰/۸۹۱ افزایش یافت؛ در حالی که برای حالت MoNISQA-aware، مقدار MoNISQA از ۰/۷۲۶ به ۰/۷۴۰ رسید.

ب_ نتایج متریک آگاه تبدیل SF-TF

جدول ۳ نتایج ارزیابی خروجی‌های سامانه‌ی پیشنهادی را در حالت متریک آگاه تبدیل صدای زن به زن (SF-TF)، نشان می‌دهد:

جدول ۳. نتایج متریک آگاه در حالت SF-TF

معیار	MoMCD	STOI	SECS	MoNISQA	FSIM	پایه
-------	-------	------	------	---------	------	------

5.691	5.670	6.359	5.608	6.250	5.642	MCD
0.438	0.432	0.439	0.459	0.476	0.442	STOI
0.849	0.851	0.845	0.872	0.870	0.854	SECS
0.755	0.730	0.806	0.742	0.762	0.750	MoNISQA

در حالت SF-TF نیز الگوی اصلی حفظ شده است. آزمایش MoMCD-aware مقدار MCD را از ۵/۶۹۱ به ۵/۶۴۲ کاهش داده است. در حالی که آزمایش STOI-aware مقدار STOI را از ۰/۴۳۸ به ۰/۴۷۶ رسانده که افزایش قابل توجهی است. همچنین SECS-aware مقدار SECS را از ۰/۸۴۹ به ۰/۸۷۲ افزایش داده و همین‌طور MoNISQA-aware مقدار MoNISQA را از ۰/۷۵۵ به ۰/۸۰۶ رسانده است.

ج. نتایج متریک آگاه تبدیل SM-TF

در جدول ۴ می‌توان نتایج اعمال متریک‌ها در بهینه‌سازی متریک آگاه تبدیل SM-TF را مشاهده و بررسی نمود.

جدول ۴. نتایج متریک آگاه در حالت SM-TF

پایه	FSIM	MoNISQA	SECS	STOI	MoMCD	معیار
6.901	6.442	6.762	6.203	6.921	6.808	MCD
0.368	0.367	0.369	0.383	0.403	0.378	STOI
0.819	0.809	0.821	0.833	0.830	0.811	SECS
0.730	0.726	0.756	0.748	0.741	0.749	MoNISQA

در این حالت تبدیل، مقدار پایه MCD نسبتاً بالا و مقدار STOI نسبتاً پایین است که می‌تواند نشان‌دهنده دشواری بیشتر این تبدیل باشد. نتایج نشان می‌دهد که MoMCD-aware مقدار MCD را از ۶/۹۰۱ به ۶/۸۰۸ کاهش داده است. آزمایش STOI-aware مقدار STOI را از ۰/۳۶۸ به ۰/۴۰۳ رسانده است؛ در حالی که آزمایش SECS-aware، مقدار SECS را از ۰/۸۱۹ به ۰/۸۳۳ افزایش داده و MoNISQA-aware مقدار MoNISQA را از ۰/۷۳۰ به ۰/۷۵۶ رسانده است.

د. نتایج متریک آگاه تبدیل SM-TM

در حالت SM-TM، نتایج متریک آگاه نیز بهبودهای متناظر را نشان می‌دهند. در جدول ۵ نتایج آزمایش‌های مرتبط، قابل مشاهده و بررسی است:

جدول ۵. نتایج متریک آگاه در حالت SM-TM

پایه	FSIM	MoNISQA	SECS	STOI	MoMCD	معیار
6.328	5.918	6.358	5.472	5.587	6.120	MCD
0.498	0.501	0.471	0.521	0.529	0.501	STOI

0.870	0.876	0.864	0.899	0.900	0.875	SECS
0.658	0.695	0.758	0.635	0.685	0.681	MoNISQA

همان طوری که مشاهده می‌شود؛ در حالت MoMCD-aware، MCD از ۶/۳۲۸ به ۶/۱۲۰ کاهش یافته است؛ اما در حالت-STOI aware، STOI از ۰/۴۹۸ به ۰/۵۲۹ افزایش یافته است. در حالت SECS-aware، SECS از ۰/۸۷۰ به ۰/۸۹۹ رسیده است. در حالت MoNISQA-aware، مقدار MoNISQA از ۰/۶۵۸ به ۰/۷۵۸ افزایش یافته است.

۳-۵. جمع‌بندی کمی نتایج بهینه‌سازی با متریک‌های گفتاری

در میانگین چهار حالت تبدیل، بهینه‌سازی با متریک‌های گفتاری باعث کاهش ۰/۱۰۳ دسی‌بل در MCD شده است. مقدار STOI از ۰/۴۵۴ به ۰/۴۸۴ افزایش یافته که بیش از ۶ درصد بهبود را نشان می‌دهد. مقدار SECS از ۰/۸۵۵ به ۰/۸۷۴ رسیده و مقدار MoNISQA از ۰/۷۱۷ به ۰/۷۶۵ افزایش یافته است. اما اثر بهینه‌سازی با FSIM عمدتاً در کاهش MCD قابل توجه است که در ادامه بیشتر به آن خواهیم پرداخت.

با توجه به این که این مطالعه بر خلاف مطالعات معمول بر این چهار معیارهای مهم کیفیت ادراکی تاحدی جدید متمرکز بود و در سایر تحقیقات و مقاله‌ها این موارد پوشش داده نشده بود، امکان مقایسه و بررسی بیشتری وجود نداشت.

۴-۵. توسعه آزمایش‌ها به دو متریک تصویری SSIM و GMSD

در ادامه این تحقیق و با هدف توسعه بررسی‌ها با دو متریک تصویری SSIM و GMSD در کنار FSIM برای ارزیابی و بهینه‌سازی با آن‌ها، آزمایش‌هایی ترتیب داده شد و از این در متریک برای بهینه‌سازی در سامانه پیشنهادی استفاده شد. در واقع هدف از این آزمایش‌ها این بود که بتوان برتری یا ضعف اثر SSIM و GMSD را نسبت به FSIM بهتر مشاهده کرد و با آن بتوان مقایسه کرد. نتایج کامل چهار حالت تبدیل جنسیتی حاصل از سامانه‌ی پیشنهادی با سه متریک تصویری مدنظر، در جدول ۶ قابل مشاهده است. با بررسی نتایج، شرایط خوبی برای SSIM نسبت به بهینه‌سازی با سایر متریک‌ها و نیز سامانه‌ی پایه وجود دارد ولی به طور خاص به علت کم‌تر بودن میانگین MoNISQA (در حالت میانگین کل) از مقدار متناظرش در سامانه‌ی پایه در بهینه‌سازی با این متریک، نمی‌توان این متریک را متریکی دانست که ظرفیت آن را داشته باشد که در همه‌ی معیارهای ارزیابی، به طور همزمان بهبود ایجاد کند. از این منظر می‌توان متریک SSIM را مشابه‌ی FSIM تلقی کرد.

جدول ۶. نتایج بهینه‌سازی بر پایه‌ی سه روش متریک آگاه به کمک متریک‌های تصویری در مقایسه با نتایج چهار روش بهینه‌سازی متریک آگاه مبتنی بر متریک‌های صوتی (میانگین‌گیری شده روی چهار حالت تبدیل جنسیتی و مقادیر معیارهای ارزیابی چهارگانه به تفکیک حالات تبدیل جنسیتی)

پایه				متریک آگاه گفتاری				FSIM-Aware (تصویری)				SSIM-Aware (تصویری)				GMSD-Aware (تصویری)				روش بهینه سازی
																				متریک
SF-TM	SF-TF	SM-TF	SM-TM	SF-TM	SF-TF	SM-TF	SM-TM	SF-TM	SF-TF	SM-TF	SM-TM	SF-TM	SF-TF	SM-TF	SM-TM	SF-TM	SF-TF	SM-TF	SM-TM	ارزیابی
				MoMCD-Aware																
6.191				6.088				5.933				5.771				5.785				MCD میانگین
5.842	5.691	6.901	6.328	5.780	5.642	6.808	6.12	5.702	5.670	6.442	5.918	5.632	5.566	6.061	5.825	5.639	5.617	5.981	5.900	مقادیر ۴ حالت
				STOI-Aware																
0.454				0.484				0.454				0.473				0.473				STOI میانگین
0.513	0.438	0.368	0.498	0.527	0.476	0.403	0.529	0.520	0.444	0.392	0.498	0.528	0.455	0.400	0.509	0.530	0.443	0.403	0.515	مقادیر ۴ حالت
				SECS-Aware																
0.855				0.874				0.856				0.869				0.868				SECS میانگین
0.882	0.849	0.818	0.870	0.891	0.872	0.833	0.899	0.888	0.851	0.809	0.876	0.885	0.860	0.839	0.883	0.891	0.856	0.836	0.887	مقادیر ۴ حالت
				MoNISQA-Aware																
0.717				0.765				0.711				0.715				0.718				MoNISQA میانگین
0.726	0.755	0.730	0.658	0.740	0.806	0.756	0.758	0.705	0.741	0.737	0.660	0.691	0.760	0.742	0.667	0.669	0.749	0.756	0.698	مقادیر ۴ حالت

وقتی در آزمایشی دیگر به همین بررسی‌ها برای GMSD در کنار FSIM و SSIM پرداخته شد نتیجه‌گیری شد که هر چند به‌کارگیری این متریک (GMSD)، در بهینه‌سازی‌ها نسبت به حالت MoMCD-aware مزیت ویژه‌ای از خود نشان می‌دهد؛ اما مهمتر آن که هرچند نسبت به بهینه‌سازی با سایر متریک‌ها بهبود زیادی را ایجاد نمی‌کند لیکن حایز این مزیت مهم و برجسته نسبت به سایر متریک‌ها در بهینه‌سازی‌ها است که همه‌ی معیارهای ارزیابی چهارگانه سامانه‌ی پایه را همزمان و بدون استثنا اما کم و بیش افزایش می‌دهد (در حالت میانگین در بین تبدیل‌های چهارگانه). به بیان دیگر، این متریک، تنها متریکی از بین متریک‌های آزموده‌شده است که شایستگی و جامعیت آن را دارد که در شرایطی که می‌خواهیم همزمان هر چهار معیار ارزیابی بهبودیابند و یا دست‌کم هیچ‌کدام، کاهش در میانگین کلشان نداشته باشند (در برابر افزایش تکی سایرین)، به کارگرفته شود. به عبارت دیگر، با کمک این متریک، می‌توان به هدف بهینه‌سازی چندمعیاره‌ی همزمان دست یافت.

۵-۵. جمع‌بندی کلی بهبود حاصل از بهینه‌سازی با GMSD در برابر متریک‌های گفتاری

در جدول ۷ یک جمع‌بندی کلی از بهبود متریک‌ها (میانگین در حالات مختلف جنسیتی) را می‌توان ناشی از به‌کارگیری آن‌ها در بهینه‌سازی‌ها، ملاحظه نمود.

جدول ۷. جمع‌بندی کلی (میانگین) بهبود متریک‌ها

معیار	سامانه‌ی پایه	سامانه‌ی متریک‌آگاه	بهبود ناشی از متریک‌های گفتاری	سامانه پیشنهادی با GMSD	بهبود ناشی از GMSD
MCD [dB]	6.191	6.088	0.103	5.785	0.406
STOI	0.454	0.484	6.46%	0.473	5.36%
SECS	0.855	0.874	2.24%	0.868	1.99%
MoNISQA	0.717	0.765	6.65%	0.718	0.001%

براساس نتایج جدول معرف کلی (با میانگین‌گیری روی چهار حالت تبدیل جنسیتی)، بهبود متریک‌ها ناشی از بهینه‌سازی متریک‌آگاه و ناشی از بهینه‌سازی با GMSD در جدول ۶، می‌توان دریافت که GMSD در مقایسه با متریک‌های گفتاری (همچنین در مقایسه با FSIM و SSIM که برای اختصار نتایج آزمایش‌های مرتبط در [۳۲] قابل مشاهده است)، ظرفیت بیشتری برای بهبود همزمان معیارهای مختلف دارد. به‌ویژه، آن که GMSD تنها متریک تصویری بررسی شده است که در میانگین کلی، هیچ‌یک از چهار معیار اصلی را نسبت به سامانه پایه کاهش نداده است. این یافته نشان می‌دهد که معیارهای تصویری مبتنی بر مل‌اسپکتروگرام و به طور خاص GMSD می‌توانند در آینده برای طراحی متریک‌های اختصاصی تبدیل گفتار مورد توجه بیشتری قرار گیرند.

در انتهای آزمایش‌ها و ارزیابی‌ها، برای تکمیل نتایج عینی،

یک ارزیابی ذهنی رجحان با فایل‌های صوتی نمونه‌ی خروجی سامانه‌ی پایه و سامانه‌ی پیشنهادی در حالت MoMCD-Aware، با ۱۵ شنونده ارزیاب، انجام شد که نتایج آن، نشان از موفقیت و برتری شنیداری خروجی‌های روش پیشنهادی MetricCycleGAN-VC با کسب امتیاز ۷۴/۲ درصدی در برابر امتیاز ۲۵/۸ درصدی سامانه پایه در یک آزمون آماری منصفانه داشت. به خاطر محدودیت‌های امکانات پردازشی، این فایل‌های صوتی خروجی سامانه، با آموزش طی ۱۵۰۰ دور برای سامانه‌ی پایه و پیشنهادی، به دست آمده بودند؛ با این حال تفاوت‌های کیفی در ارزیابی شنیداری کاملاً محسوس بود. این نتیجه، شواهد عینی حاصل از متریک‌ها را از منظر ادراکی نیز تقویت می‌کند.

۶. بحث

نتایج کمی و کیفی نشان می‌دهند که ورود متریک‌های ارزیابی به تابع هدف یک سامانه‌ی نمونه‌ی تبدیل گفتار، می‌تواند شکاف میان آموزش و ارزیابی را در تبدیل گفتار کاهش دهد. در بیشتر حالت‌ها، هنگامی که سامانه با یک متریک خاص

آموزش داده شده، همان متریک در خروجی بهبود یافته است. این موضوع نشان می‌دهد که تمایزگرهای اصلاح‌شده توانسته‌اند تا حدی رفتار متریک‌ها را یاد بگیرند و مولدها را در جهت مناسب هدایت کنند.

یکی از نکات مهم، ارتباط میان MCD به عنوان معیاری بسیار قدیمی و رایج در تبدیل گفتار و سایر معیارهاست. در بسیاری از حالت‌ها، بهبود STOI یا SECS با کاهش MCD همراه شده است. این موضوع می‌تواند ناشی از آن باشد که MCD شباهت کلی پوش طیفی را می‌سنجد و به نوعی با چند جنبه کیفیت گفتار ارتباط دارد. با این حال، MCD به تنهایی کافی نیست؛ زیرا نمی‌تواند تمام ابعاد ادراکی کیفیت مانند طبیعی بودن یا شباهت هویت گوینده را در حد کفایت پوشش دهد.

رفتار MoNISQA نسبت به سایر معیارها پیچیده‌تر است. در برخی حالت‌ها، بهینه‌سازی MoNISQA بهبود قابل توجهی در امتیاز خود ایجاد کرده، اما در برخی موارد با کاهش سایر معیارها همراه بوده است. دلیل احتمالی، ماهیت غیرتهاجمی و شبکه‌محور بودن NISQA است. چنین متریک‌هایی به داده‌های آموزش خود وابسته‌اند و ممکن است در برابر نمونه‌های کوتاه یا شرایط خاص تبدیل گفتار نوسان داشته باشند.

نتایج متریک‌های تصویری نیز از نظر پژوهشی مهم است. متریک‌های FSIM و SSIM در برخی معیارها بهبود ایجاد کرده‌اند؛ اما در همه ابعاد کیفیت پایدار نبوده‌اند. در مقابل، GMSD توانسته است بهبود همزمان تری ارائه کند. این یافته اهمیت زیادی دارد؛ زیرا در کنار کارایی آن در بهینه‌سازی، استفاده از معیارهای تصویری بر روی مل اسپکتروگرام می‌تواند سیر طراحی متریک‌های جدیدی را باز کند که مستقیماً در فضای ویژگی‌های مورد استفاده سامانه‌ی آموزش عمل می‌کنند و الزاماً نیازی به تبدیل مکرر به موج گفتار ندارند.

از منظر کاربردی، روش پیشنهادی، مزیت جعبه‌سیاه را نیز دارد. یعنی برای استفاده از یک متریک، لازم نیست ساختار درونی یا فرمول مشتق‌پذیر آن شناخته شود. کافی است مقدار متریک برای نمونه‌های آموزشی محاسبه شود و تمایزگر آن را بیاموزد. این ویژگی باعث می‌شود چارچوب پیشنهادی با متریک‌های جدیدی نیز که در آینده معرفی شوند؛ سازگار باشد.

در عین حال، روش پیشنهادی محدودیت‌هایی هم دارد. نخست، هزینه‌ی محاسباتی آموزش افزایش می‌یابد؛ زیرا در هر مرحله باید نمونه‌هایی از مل اسپکتروگرام به شکل موج صوتی تبدیل شده و متریک‌ها محاسبه شوند. دوم، آموزش فقط تا ۱۰۰۰ دور و در مواردی تا ۲۴۰۰ دور انجام شده است؛ در حالی که برای دستیابی به کیفیت شنیداری بالاتر احتمالاً آموزش طولانی‌تری لازم است. سوم، همه متریک‌های موجود برای تبدیل گفتار ایدآل نیستند و برخی از آن‌ها از حوزه‌های دیگری مانند به‌سازی گفتار یا ارزیابی تصویر وام گرفته شده‌اند. بنابراین، طراحی معیارهای اختصاصی‌تر برای تبدیل گفتار همچنان ضروری است.

۷. جمع‌بندی تحلیلی توصیفی و مقایسه‌ای از متریک‌ها

در مجموع متریک‌های موجود و یا پیشنهادی ایدآل نیستند و دارای ضعف‌هایی هم هستند که اگر برطرف شوند؛ بهبودهای چشم‌گیرتری را نتیجه می‌دهند. به همین جهت طرح متریک‌های ارزیابی عینی قدرتمندتر و ادراکی‌تر خاص

تبدیل گفتار ضرورت دارد. با این وجود با آزمایش‌ها و نتایج، نشان داده شد که روش مورد نظر برای افزایش متریک‌های عینی کیفیت، با بهینه‌سازی در متریک‌های مختلف موجود و یا طرح شده در تبدیل گفتار، در حدّ خود عملی و مؤثر است. اما میزان اثرش از لحاظ ذهنی و ادراکی و سنجه‌های نهایی انسانی بسته به ساختار و انسجام متریک انتخابی است؛ به‌خصوص که به نقاط ضعف و قوت متریک‌ها به‌خصوص وقتی بر حالت ویژه‌ی فاز آموزش تبدیل گفتار که شرایط پیچیده‌ای دارد؛ متمرکز شویم:

الف- متریک MCD خیلی کاربردی است؛ و شاید تنها متریک ادراکی موجود در ادبیات VC باشد؛ ولی برای فاز آموزش یک سامانه‌ی تبدیل گفتار معیاری کلی است و از طرف دیگر نزدیکی زیادی به ارزیابی‌ها براساس مل‌اسپکتروگرام دارد. بنابراین بهینه‌سازی با مل‌اسپکتروگرام که در MaskCycleGAN-VC بر مبنای فاصله‌ی بین آنهاست؛ خیلی دور از بهینه‌سازی با MCD نیست. هرچند به هر حال تفاوتی بین فاصله‌ی L1 یا L2 برای مل‌اسپکتروگرام‌ها و متریک MCD وجود دارد. متأسفانه براساس دانش ما تاکنون متریکی مناسب بر مبنای مل‌اسپکتروگرام تعریف نشده‌است که در فاز آموزش سامانه‌های تبدیل گفتار امروزی، به سهولت بتوان به کار برد. به نظر می‌رسد بتوان به کمک معیارهای شباهت تصویری، در کارهای آینده آن را تعریف کرد. البته تاکنون متریک MCD فقط در فاز آزمون تبدیل گفتار مطرح بوده و نه در فاز آموزش. مشکل دیگر این متریک این است که نیاز به تطبیق زمانی^{۶۳} دارد که این دقت و انسجام آن را بر هم می‌زند. ولی به هر حال معیاری نسبی است نه دقیق از لحاظ کیفی و فقط شباهت کلی طیفی را بر مبنای MFCCها^{۶۴} می‌سنجد.

ب- متریک STOI، مختص ارزیابی فهم‌پذیری گفتار صرفاً یک گوینده در برابر نویز است و در تبدیل گفتار، فقط آن را برای فهم‌پذیری وام گرفته‌ایم. این معیار نیاز به قطعات نسبتاً طولی در حد ۳۸۶ میلی ثانیه دارد؛ آن هم بدون سکوت لابلای آن که در دادگان آموزش ما همواره این‌گونه نیست؛ هرچند تطبیق زمانی وجود دارد و این برای چنین معیاری ضرورت دارد. در فاز آزمون بر عکس است و قطعات بلندتری از گفتار در دسترس است؛ ولی در عوض تطبیق زمانی وجود ندارد و این خود دقت ارزیابی‌ها را کاهش می‌دهد؛ چراکه اعمال تطبیق زمانی کیفیت را کاهش داده و این موضوع تاحدی باعث دقیق نبودن ارزیابی‌های کیفی مرتبط می‌شود. تشخیص و حذف نتایج مربوط به قطعات گفتاری نامناسب، چه در فاز آموزش و چه در آزمون، می‌تواند این مشکل را تا حدّ زیادی برطرف کند.

ج- متریک SECS، متریک خوبی برای ارزیابی شباهت گویندگان است؛ ولی وابسته به دادگان و پیش‌آموزش^{۶۵} است که به هر حال وابستگی به دادگان را به همراه دارد و در عمل دقت و اطمینان بالایی را برای دادگان مختلف در همه‌ی شرایط به همراه ندارد؛ به ویژه برای دادگان کوتاه‌مدت در مرحله‌ی آموزش؛ به‌خصوص اگر محتوای گفتار یک گوینده عوض شود که طبعاً به طور کاملی مستقل از آن نیست و زمانی که تفاوت صدای گویندگان هم کم باشد، دچار خطا می‌شود. گاهی برپایه‌ی Resemblyzer به‌کارگرفته شده، مقدار SECS دو گوینده‌ی متفاوت ولی با شیوش یکسان و محتوای گفتار یکسان، بالاتر از مقدار آن برای دو گفتار با یک گوینده در دو محتوا یا شرایط خیلی متفاوت را نتیجه می‌دهد. به‌خصوص در فاز آموزش که دادگان کوتاه هستند؛ گاه ممکن است مشکلاتی از این قبیل، ایجاد شود که کارایی‌اش را کاهش دهد.

د- معیار NISQA هم معیاری است که تابه حال توجه کمی به آن در تبدیل گفتار شده؛ ولی ادعا شده که نسخه‌ی TTS آن دقیقاً برای سنتز و تبدیل گفتار طراحی شده و با دادگان مورد نظر این مقاله یعنی VCC2018، آموزش داده شده است. ولی در عین مزیت سهولت کاربرد از جهت غیرتهاجمی بودن، دارای همین مشکل مبتنی بودن بر شبکه و دادگان است که گاه نوسان‌هایی را در اندازه‌گیری‌ها می‌تواند ایجاد کند؛ به طوری که در مواردی کیفیت گفتار سنتز شده را بالاتر از گفتار طبیعی نشان دهد. به همین دلیل حتی در فاز آزمون و بررسی کیفیت خروجی تبدیل گفتار، گاه نوسان نسبتاً زیادی به همراه دارد که با توجه به قطعات کوتاه گفتار در طی آموزش، این نوسان اصولاً باید بیشتر هم باشد و گاه، باعث اختلال در آموزش شود.

نتیجه‌ی نهایی و مهم این که حوزه‌ی ارزیابی عینی تبدیل گفتار، هنوز جای کار زیادی دارد تا به صحت و دقت بالا در ارزیابی‌های ادراکی برسیم. به هر حال همان گونه که پیش از این هم اشاره شد؛ نقایص موجود در بهترین متریک‌های موجود ارزیابی، نمی‌تواند توجهی بر عدم استفاده از آن‌ها در ارزیابی‌ها و بهینه‌سازی‌ها باشد. همان گونه که امروزه همه‌ی محققان به ناچار برای ارزیابی‌ها در تحقیقات پردازش گفتار، از همین متریک‌ها بهره می‌برند تا زمانی که پیشنهادات بهتری ارائه شود.

۸. نتیجه‌گیری و کارهای آینده

در دنباله تحقیقاتی که در سامانه‌های کاربردی مختلف پردازش گفتار داشتیم [۳۳] و با انگیزه‌ای که از پیشرفت‌های سال‌های اخیر در سامانه‌های VC ناموازی به ویژه ظهور طیف گسترده‌ای از سامانه‌های VC مبتنی بر CycleGAN و کاربردهای گسترده و متعدد آن گرفتیم؛ در این تحقیق پنجره‌ای به روی بحث ارزیابی و بهینه‌سازی این نوع سامانه‌ها با متریک‌های کیفی مختلفی گشودیم. در واقع ما با موفقیت توانستیم معیارهای تمام‌تهاجمی ارزیابی گفتار را در یک سامانه‌ی VC ناموازی بگنجانیم و راه‌حلی برای مسأله‌ی DEM در دسته‌ی وسیعی از سامانه‌های تبدیل گفتار ارائه دهیم. در تکمیل این ایده، توانستیم معیارهای ارزیابی شباهت تصاویر مانند FSIM، SSIM و GMSD را نیز در تابع هدف، وارد کرده و با بهینه‌سازی این متریک‌ها، جنبه‌های کیفی گفتار تبدیل یافته را در سامانه‌ی پیشنهادی خود ارتقا ببخشیم.

به لطف به کارگیری استراتژی جعبه سیاه، این روش، مزیت انطباق معیارهای کیفیت پیچیده و مشتق‌ناپذیر را بدون نیاز به تغییر در ساختار آنها یا ورود به فرمول‌بندی گاه دشوار آن‌ها ارائه می‌دهد. در نتیجه، با توسعه‌ی معیارهای ارزیابی پیشرفته‌تر، روش جعبه سیاه پیشنهادی، به راحتی متریک‌های آینده را نیز در بر گرفته و به سادگی روی آن‌ها قابل پیاده‌سازی است. علاوه بر این، رویکرد ما بستر و پایه‌ای برای ترکیب سنتزکننده‌های گفتار در بهینه‌سازی، آموزش و استخراج تابع نگاشت VC ایجاد می‌کند. این ایده برای داشتن تمایزگرهای بهبودیافته بدون هیچ ماژول یا داده‌ی اضافی، به جز افزایش اندکی در هزینه‌ی محاسبات، اجرا می‌شود. این افزایش هزینه، به دلیل تبدیل برخی از نمونه‌های آموزشی مل اسپکتروگرام به شکل موج سیگنال گفتار (تولید حداکثر ۴ فایل گفتاری در هر تکرار) و افزودن ارزیابی این صداها در میانه‌ی روند آموزش بر حسب یک متریک محبوب و مورد علاقه است (در این مطالعه یکی از متریک‌های

به کارگیری این روش به ما اجازه داد تا یک چارچوب برای پرداختن به مسأله‌ی DEM و حل آن در یک سامانه‌ی VC مبتنی بر CycleGAN استاندارد و با استفاده از مجموعه‌ای از معیارهای ادراکی (سه مورد از تصویر و چهار مورد از گفتار) ایجاد کنیم.

نتایج این تحقیق، می‌تواند افقی جدید به طرح متریک‌های ارزیابی عینی تبدیل گفتار مبتنی بر برخی متریک‌های تصویری، بگشاید که حقیقتاً جای کار بسیاری دارد. طراحی این متریک‌ها مبتنی بر اعمال معیارهای شباهت تصاویر روی ویژگی‌های مورد استفاده در بسیاری از سامانه‌های تبدیل گفتار، یعنی شبه‌جفت‌های تصاویر مل‌اسپکتروگرام، امکان‌پذیر است.

گسترش تحقیق عملی و تجربی در به کارگیری متریک‌های شباهت تصاویر از FSIM به SSIM و GMSD، این ایده‌ی بالا را تقویت کرد. به طوری که مشاهده شد متریک‌های SSIM و GMSD ظرفیت‌های بالایی را در بهبود MCD و STOI و گاه حتی SECS، از خود نشان می‌دهند. این موضوع، خود می‌تواند تأییدی بر ذهنیت و فرضیه‌ی همبستگی قوی‌تر برخی معیارهای تصویری با متریک‌های صوتی باشد. در حدی که گاه به کارگیری آن‌ها، اثرگذارتر از به کارگیری متریک‌های صوتی در بهینه‌سازی‌ها به نظر می‌رسد.

استفاده از MOSNet [۳۴] یا استفاده‌ی مستقیم از نتایج تجربی MOS، با این رویکرد امکان‌پذیر است و کاوش متریک‌ها و معیارهای مناسب برای این منظور، می‌تواند یک حوزه‌ی تحقیقاتی قانع‌کننده و قابل‌پذیرش باشد؛ طبیعتاً پیشرفت علمی مداوم در زمینه‌ی ارزیابی، برای افزایش کیفیت VC مفید خواهد بود. موضوع بالقوه‌ی دیگر برای پژوهش را می‌توان استفاده از MBNet [۳۵]، یا ابزارهای SVSNet [۳۶]، به عنوان توابع هزینه‌ی جایگزین^{۶۶} در بهینه‌سازی تبدیل گفتار، برای افزایش شباهت گویندگان دانست.

بنابراین برای کارهای آینده، چند مسیر پیشنهاد می‌شود: طراحی متریک‌های اختصاصی تبدیل گفتار مبتنی بر مل‌اسپکتروگرام، بررسی آموزش طولانی‌تر با سخت‌افزار پردازشی قوی‌تر برای قطعات بزرگتر گفتار، استفاده از MOSNet یا MBNet به عنوان زیان جایگزین، توسعه‌ی معیارهای دقیق‌تر شباهت گوینده و ارزیابی ذهنی گسترده‌تر با تعداد شنوندگان بیشتر.

۹. منابع

- [1] Childers, D., B. Yegnanarayana, and Ke Wu. "Voice conversion: Factors responsible for quality." In *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, pp. 748-751. IEEE, 1985.
- [2] Sisman, Berrak, Junichi Yamagishi, Simon King, and Haizhou Li. "An overview of voice conversion and its challenges: From statistical modeling to deep learning." *IEEE/ACM transactions on audio, speech, and language processing* 29 (2020): 132-157.
- [3] Dasare, Ashwini, and K. T. Deepak. "Performance assessment of voice conversion models using speech production-based parameters." *Computer Speech & Language* 95 (2026): 101853.

- [4] Wang, Zhichao, Tao Li, Wenshuo Ge, Zihao Cui, Shilei Zhang, and Junlan Feng. "Onevoice: One model, triple scenarios-towards unified zero-shot voice conversion." *arXiv preprint arXiv:2601.18094* (2026).
- [5] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," EUSIPCO, 2018.
- [6] T. Kaneko et al., "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," ICASSP, 2019.
- [7] T. Kaneko et al., "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion," Interspeech, 2020.
- [8] T. Kaneko et al., "MaskCycleGAN-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," ICASSP, 2021.
- [9] Kubichek, Robert. "Mel-cepstral distance measure for objective speech quality assessment." In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1, pp. 125-128. IEEE, 1993.
- [10] Taal, Cees H., Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "A short-time objective intelligibility measure for time-frequency weighted noisy speech." In *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214-4217. IEEE, 2010.
- [11] Mittag, Gabriel, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets." *arXiv preprint arXiv:2104.09494* (2021).
- [12] Fu, Szu-Wei, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai. "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, no. 9 (2018): 1570-1584.
- [13] Fu, Szu-Wei, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement." In *International Conference on Machine Learning*, pp. 2031-2041. PmLR, 2019.
- [14] Behdad, Majid, and Davood Gharavian. "MetricCycleGAN-VC: Forcing CycleGAN-based voice conversion systems to associate with objective quality metrics." *IEEE Access* 12 (2024): 154631-154649.
- [15] Behdad, Majid, and Davood Gharavian. "Improving CycleGAN-VC2 voice conversion by learning MCD-based evaluation and optimization." In *2023 31st International Conference on Electrical Engineering (ICEE)*, pp. 971-976. IEEE, 2023.
- [16] Zhang, Lin, Lei Zhang, Xuanqin Mou, and David Zhang. "FSIM: A feature similarity index for image quality assessment." *IEEE transactions on Image Processing* 20, no. 8 (2011): 2378-2386.
- [17] Wang, Z. "Image quality assessment: Form error visibility to structural similarity." *IEEE Trans. Image Process.* 13, no. 4 (2004): 604-606.
- [18] Xue, Wufeng, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index." *IEEE transactions on image processing* 23, no. 2 (2013): 684-695.
- [19] Walczyna, Tomasz, and Zbigniew Piotrowski. "Overview of voice conversion methods based on deep learning." *Applied sciences* 13, no. 5 (2023): 3100.
- [20] Hsu, Chin-Cheng, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. "Voice conversion from non-parallel corpora using variational auto-encoder." In *2016 Asia-Pacific Signal*

and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1-6. IEEE, 2016.

- [21] Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *2017 IEEE international conference on computer vision (ICCV)*, pp. 2242-2251. Ieee, 2017.
- [22] Rix, Antony W., John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749-752. IEEE, 2001.
- [23] Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021): 3451-3460.
- [24] Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." *IEEE Journal of Selected Topics in Signal Processing* 16, no. 6 (2022): 1505-1518.
- [25] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.
- [26] Qian, Kaizhi, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I. Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. "Contentvec: An improved self-supervised speech representation by disentangling speakers." In *International conference on machine learning*, pp. 18003-18017. PMLR, 2022.
- [27] Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." In *International conference on machine learning*, pp. 28492-28518. PMLR, 2023.
- [28] Hu, Shujie, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan et al. "Wavlm: Towards robust and adaptive speech large language model." In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552-4572. 2024.
- [29] Zhang, Xueyao, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang et al. "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement." In *International Conference on Learning Representations*, vol. 2025, pp. 58436-58459. 2025.
- [30] Chu, Minghang, Jing Wang, Zhiwei Fan, Mengtao Yang, Chao Xu, Yaoyao Ma, Zhi Tao, and Di Wu. "A multidomain generative adversarial network for hoarse-to-normal voice conversion." *Journal of Voice* 40, no. 1 (2026): 107-118.
- [31] Chen, Lei, Le Wu, Zhenzhen Hu, and Meng Wang. "Quality-aware unpaired image-to-image translation." *IEEE Transactions on Multimedia* 21, no. 10 (2019): 2664-2674.
- [۳۲] مجید بهداد، "بهبود سامانه‌های تبدیل گفتار و ارزیابی آن به کمک یادگیری عمیق"، رساله‌ی دکتری مخابرات، دانشگاه شهید بهشتی، ۱۴۰۳.
- [۳۳] معصومه شفیعیان، وحیداحمدیان و مجید بهداد؛ "بازشناسی احساسات از روی گفتار برپایه بهره‌گیری از شبکه‌های عصبی پیچشی و روش افزایش دادگان"، صوت و ارتعاش، ۱۴۰۱، شماره ۲۱، دوره ۱۱، صفحه ۸۵-۹۸.
- [34] Lo, Chen-Chou, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. "Mosnet: Deep learning based objective assessment for voice conversion." *arXiv preprint arXiv:1904.08352* (2019).

- [35] Leng, Yichong, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. "MBNet: MOS prediction for synthesized speech with mean-bias network." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391-395. IEEE, 2021.
- [36] Hu, Cheng-Hung, Yu-Huai Peng, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. "SVSNet: An end-to-end speaker voice similarity assessment model." *IEEE Signal Processing Letters* 29 (2022): 767-771.

پی نوشت

-
- ¹ Voice Conversion
² Intelligibility
³ Parallel-Data(Paired)
⁴ Non-Parallel Data(Unpaired)
⁵ Cycle-Consistency Loss
⁶ Adversarial Learning
⁷ Generative Adversarial Network(GAN)-Based
⁸ Discriminator
⁹ Mel-Cepstral Distance
¹⁰ Short-Time Objective Index
¹¹ Non-Intrusive Speech Quality Assessment (NISQA)
¹² Discriminator Evaluation Mismatch
¹³ Metric-Aware
¹⁴ Modified-MCD
¹⁵ Speaker Embedding Cosine Similarity
¹⁶ Modified-NISQA
¹⁷ Feature Similarity Index Measure
¹⁸ Structural Similarity Index Measure
¹⁹ Gradient Magnitude Similarity Deviation
²⁰ Autoencoder
²¹ Disentanglement
²² Unpaired Image-to-Image Translation
²³ Filling in Frames
²⁴ GAN-Based
²⁵ Loss-Function (Objective Function)
²⁶ Speech Enhancement
²⁷ Perceptual Evaluation of Speech Quality
²⁸ Generator
²⁹ Mel-Spectrogram
³⁰ One-to-One
³¹ One-to-Many
³² Autoencoder
³³ Foundation Models
³⁴ Self-Supervised Learning
³⁵ Content Encoder
³⁶ Zero-Shot
³⁷ Diffusion Models
³⁸ Flow Matching
³⁹ Prompt-Aware LoRA(Low Rank Adaptation)
⁴⁰ Speech Large Language Models (SLLM)
⁴¹ Transformer
⁴² Neural Vocoder

-
- 43 Any-to-Any
44 Hoarse
45 FORMSE
46 Mel-Spectral Distance
47 Mean-Opinion Score
48 Identity Mapping Loss
49 Binary
50 Pseudo-Pair
51 Full-Reference
52 Semi-Intrusive
53 Non-Intrusive Metric
54 Word Error Rate
55 Masked
56 Mask
57 Fully-Connected
58 Sigmoid
59 MelGAN
60 Vocoder(Voice Coder)
61 Batch-Size
62 Epoch
63 Time-Alignment
64 Mel-Frequency Cepstral Coefficient
65 برای جزییات بیشتر به آدرس زیر مراجعه شود:
<https://github.com/resemble-ai/resemblyzer>
66 Surrogate Loss Function