تبدیل صدای بهبود یافته غیر موازی بر مبنای روش Cycle-GAN

چکیده

تبدیل صدای غیر موازی(VC) تکنیکی برای یادگیری نگاشت از منبع به گفتار هدف بدون تکیه بر داده های موازی است. این یک کار مهم است، اما به دلیل معایب شرایط آموزش دادن شبکه عصبی همواره چالش برانگیز بوده است. اخیراً، CycleGAN-VC پیشرفتی را ارائه کرده است و بدون اتکا به داده‌ها، ماژول‌ها یا روش‌های تراز زمانی اضافی، عملکردی قابل مقایسه با روش VC موازی داشته است. با این حال، هنوز فاصله زیادی بین هدف واقعی و گفتار تبدیل شده وجود دارد و پر کردن این شکاف همچنان یک چالش است. برای کاهش این شکاف، CycleGAN VC2 پیشنهاد می‌شود که نسخه بهبودیافته CycleGAN-VC است که سه تکنیک جدید را در خود جای داده است: یک هدف بهبودیافته (تلفات خصمانه دو مرحله‌ای)، ژنراتور بهبود یافته (استفاده از شبکه عصبی دولایه بجای تک لایه) و تشخیص دهنده بهبود یافته (PatchGAN). ما روش خود را در یک کار VC غیر موازی ارزیابی کردیم و تأثیر هر تکنیک را با جزئیات تجزیه و تحلیل کردیم. یک ارزیابی عینی نشان داد که این تکنیک‌ها به نزدیک‌تر کردن توالی ویژگی تبدیل‌شده به هدف از نظر ساختارهای کلی و محلی کمک می‌کنند، که ما به ترتیب با استفاده از اعوجاج Mel-cepstral و فاصله طیف مدولاسیون ارزیابی می‌کنیم. یک ارزیابی ذهنی نشان داد که CycleGAN-VC2 از CycleGAN-VC از نظر طبیعی بودن و شباهت برای هر جفت گوینده، از جمله جفت های درون جنسیتی و بین جنسیتی بهتر عمل می کند.

کلمات کلیدی: تبدیل گفتار، تبدیل گفتار غیر موازی، شبکه‌های مولد متخاصم، CycleGAN-VC و CycleGAN

۱- مقدمه

تبدیل صدا (VC) تکنیکی است برای تبدیل اطلاعات غیر زبانی گفتار داده شده در حالی که اطلاعات زبانی را حفظ می کند. VC پتانسیل زیادی برای کاربرد در کارهای مختلف دارد، مانند ابزار کمکی گفتاری [۱، ۲] و تبدیل سبک [۳، ۴] و تلفظ [۵].

یک رویکرد موفق برای VC شامل روش‌های آماری مبتنی بر مدل مخلوط گاوسی (GMM) [۶، ۷ و ۸] روش‌های مبتنی بر شبکه‌های عصبی(NN) با استفاده از ماشین‌های محدود بولتزمن (RBM) [۹، ۱۰]، شبکه‌های عصبی پیش رونده[۱۱، ۱۲، ۱۳]، شبکه‌های کانولوشنی (CNN) [۵]، شبکه های توجه [۱۶، ۱۷]، و شبکه های مولد متخاصم (GAN) [۵] و روش‌های مبتنی بر مثال با استفاده از فاکتورسازی ماتریس غیرمنفی (NMF) [۱۸، ۱۹].

بسیاری از روش‌های VC (از جمله موارد ذکر شده در بالا) به عنوان VC موازی طبقه‌بندی می‌شوند که به در دسترس بودن جفت‌های گفتاری موازی از گوینده‌های منبع و هدف متکی است. با این حال، جمع آوری چنین داده هایی اغلب پر زحمت یا غیر عملی است. حتی اگر دستیابی به چنین داده‌هایی امکان‌پذیر باشد، بسیاری از روش‌های VC به یک روش هم‌ترازی زمانی به عنوان یک پیش‌فرایند نیاز دارند، که ممکن است گاهی اوقات با شکست مواجه شود و نیاز به پیش‌غربال‌گری دقیق یا تصحیح دستی دارد. برای غلبه بر این محدودیت‌ها، این مقاله بر VC غیر موازی تمرکز می‌کند، که بر گفته‌های موازی، رونویسی‌ها یا رویه‌های تراز زمانی متکی نیست.

به طور کلی VC غیر موازی کاملاً چالش برانگیز است و از نظر کیفیت به دلیل معایب شرایط آموزش دادن نسبت به VC موازی پایین تر است. برای کاهش این شرایط شدید، چندین مطالعه یک ماژول اضافی (به عنوان مثال، یک مدول تشخیص خودکار گفتار (ASR) [20، ۲۱]) یا داده‌های اضافی (مانند جفت‌های گفتار موازی در میان سخنرانان مرجع[۲۲، ۲۳، ۲۴، ۲۵]) ترکیب کرده‌اند. اگرچه این ماژول ها یا داده های اضافی برای آموزش مفید هستند، اما آماده سازی آنها هزینه های دیگری را تحمیل می کند و در

نتیجه کاربرد را محدود می کند. برای جلوگیری از چنین هزینه‌های اضافی، مطالعات اخیر استفاده از شبکه‌های عصبی احتمالی (به عنوان مثال، RBN [۲۶] و کدگذارهای خودکار متغیر(VAEs) [۲۷، ۲۸]) را بررسی کرده‌اند، که ویژگی‌های صوتی را در فضای کم‌بعدی رایج و با نظارت بر هویت گوینده قرار می‌دهند. قابل توجه است که آنها از داده های اضافی، ماژول ها و رویه های تراز زمانی آزاد هستند. با این حال، یک محدودیت این است که آنها باید توزیع داده ها را به طور صریحی تخمین بزنند (به عنوان مثال، معمولاً از گاوسی استفاده می شود)، که تمایل دارد از طریق میانگین گیری آماری باعث هموارسازی بیش از حد شود.

برای غلبه بر این محدودیت ها، مطالعات اخیر [۲۷، ۲۹، ۳۰] GAN ها [۳۱] را وارد کرده اند، که می توانند یک توزیع مولد نزدیک به هدف را بدون تقریب صریح بیاموزند، بنابراین از هموارسازی بیش از حد ناشی از میانگین های آماری جلوگیری می‌کنند. در این میان، برخلاف برخی از روش‌های فریم به فریم [۲۷، ۳۰] که در یادگیری وابستگی‌های زمانی مشکل دارند، CycleGAN-VC [۲۹] (منتشر شده در [۳۲]) یادگیری تابع نگاشت مبتنی بر توالی با استفاده از CycleGan [۳۳، ۳۴، ۳۵] با CNN درگاه‌دار [۳۶] و از دست دادن نقشه هویت [۳۷]. این اجازه می دهد تا ساختارهای متوالی و سلسله مراتبی با حفظ اطلاعات زبانی ضبط شوند. با این بهبود، CycleGAN-VC عملکرد قابل مقایسه‌ای با روش VC موازی دارد [۷].

با این حال، حتی با استفاده از CycleGAN-VC، هنوز یک شکاف چالش برانگیز برای پل زدن بین هدف واقعی و گفتار تبدیل شده وجود دارد. برای کاهش این شکاف، CycleGAN-VC2 را پیشنهاد می کنیم که نسخه بهبود یافته CycleGAN-VC است که سه تکنیک جدید را در خود جای داده است: یک هدف بهبود یافته (تلفات خصمانه دو مرحله ای)، یک ژنراتور بهبود یافته (شبکه عصبی کانولوشنی۲-۱-۲ بعدی)، و یک تشخیص دهنده بهبود یافته (PatchGAN) ما تأثیر هر تکنیک را بر روی تسک Spoke (یعنی VC غیر موازی) چالش تبدیل صوتی ۲۰۱۸ (VCC 2018) تجزیه و تحلیل کردیم [۳۸]. یک ارزیابی عینی نشان داد که تکنیک‌های پیشنهادی به نزدیک‌تر کردن توالی ویژگی صوتی تبدیل‌شده به هدف از نظر ساختارهای کلی و محلی کمک می‌کنند، که ما به ترتیب با استفاده از اعوجاج Mel-cepstral و فاصله مشخصات مدولاسیون ارزیابی می‌کنیم. یک ارزیابی ذهنی نشان داد که CycleGAN-VC2 از CycleGAN-VC از نظر طبیعی بودن و شباهت برای هر جفت گوینده، از جمله جفت‌های درون جنسیتی و بین جنسیتی، بهتر عمل می کند.

در بخش ۲ این مقاله، CycleGAN-VC معمولی را بررسی می کنیم. در بخش ۳، CycleGAN-VC2 را توضیح می دهیم که نسخه بهبود یافته CycleGAN-VC است که سه تکنیک جدید را در خود جای داده است. در بخش ۴، نتایج تجربی را گزارش می کنیم. در بخش ۵ با یک خلاصه مختصر نتیجه می گیریم و کارهای آینده را ذکر می کنیم.

۲- CycleGAN-VC معمولی

۲-۱ هدف: تابع ضرر متخاصم تک مرحله‌ای

فرض کنیم که $x \in \mathbb{R}^{Q \times T_x}$ و $y \in \mathbb{R}^{Q \times T_y}$ به ترتیب ویژگی‌های صوتی مربوط به منبع X و هدف Y باشند که Q در اینجا بعد ویژگی و T_x و T_y طول سکانس‌ها هستند. هدف CycleGAN-VC این است که نگاشت $G_{X \to Y}$ را بدون نیاز به داده‌های موازی آموزش ببیند که در آن $x \in X$ را به $y \in Y$ تبدیل می‌کند. CycleGAN VC با الهام از CycleGAN [۳۳]، که در اصل در بینایی کامپیوتر برای ترجمه بدون جفت تصویر به تصویر پیشنهاد شده بود، از تلفات تخاصمی مخالف [۳۱] و از دست دادن ثبات چرخه [۳۹] استفاده می کند. علاوه بر این، CycleGAN-VC برای تشویق حفظ اطلاعات زبانی، از تابع ضرر هویت نگاشت نیز استفاده می‌کند [۳۷].

ضرر خصمانه: برای اینکه یک ویژگی تبدیل شده $G_{X\to Y}(x)$ از هدف y قابل تشخیص نباشد، از ضرر خصمانه استفاده می شود:

$$\mathscr{L}_{adv}(G_{X\to Y}, D_Y) = \mathbb{E}_{y\sim P_Y(y)}[logD_Y(y)] + \mathbb{E}_{x\sim P_X(x)}[log(1 - D_Y(G_{X\to Y}(x)))] \quad (1)$$

جایی که متمایز کننده D_Y سعی می‌کند بهترین مرز تصمیم را بین ویژگی های واقعی و تبدیل شده با به حداکثر رساندن این ضرر پیدا کند، و $G_{X\to Y}$ تلاش می کند تا ویژگی ای را ایجاد کند که بتواند D_Y را با به حداقل رساندن این ضرر فریب دهد.

از دست دادن ثبات چرخه: ضرر خصمانه فقط $G_{X\to Y}(x)$ را برای پیروی از توزیع هدف محدود می کند و سازگاری زبانی بین ویژگی های ورودی و خروجی را تضمین نمی کند. برای منظم کردن بیشتر نقشه برداری، از افت قوام چرخه استفاده می شود:

$$\begin{aligned}\mathscr{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) &\\ = \mathbb{E}_{x\sim P_X(x)}&\left[\left\|G_{Y\to X}(G_{X\to Y}(x)) - x\right\|_1\right]\\ + \mathbb{E}_{y\sim P_Y(y)}&\left[\left\|G_{X\to Y}(G_{Y\to X}(y)) - y\right\|_1\right] \quad (2)\end{aligned}$$

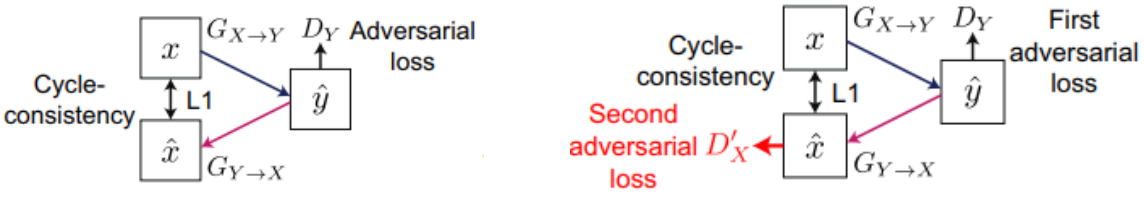
که در آن نگاشت های رو به جلو و معکوس رو به جلو به طور همزمان برای تشبیت آموزش یاد می‌گیرند. این از دست دادن $G_{X\to Y}$ و $G_{Y\to X}$ را تشویق می کند تا یک جفت شبه بهینه از (x, y) را از طریق تبدیل دایره‌ای پیدا کنند، همانطور که در شکل ۱ (الف) نشان داده شده است.

تابع ضرر هویت نگاشت: برای تشویق بیشتر به حفظ ورودی، از تابع ضرر هویت نگاشت استفاده می شود:

$$\mathscr{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{x\sim P_X(x)}[\|G_{Y\to X}(x) - x\|_1] + \mathbb{E}_{y\sim P_Y(y)}[\|G_{X\to Y}(y) - y\|_1] \quad (3)$$

$$\begin{aligned}\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})\\ + \lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) \quad (4)\end{aligned}$$

که در آن λ_{cyc} و λ_{id} پارامترهای مبادله‌ای هستند. در این فرمول، همانطور که در شکل ۱(الف) نشان داده شده است، یک بار برای هر چرخه از ضرر خصمانه استفاده می شود. از این رو، ما آن را «زیان خصمانه یک مرحله‌ای» می نامیم.



الف - تابع ضرر متخاصم تک مرحله‌ای

ب - تابع ضرر متخاصم دو مرحله‌ای

شکل ۱. مقایسه اهداف

۲-۲ ژنراتور: شبکه عصبی پیچشی یک بعدی

CycleGAN-VC از یک CNN یک بعدی (۱ بعدی) [۵] برای ژنراتور استفاده می کند تا رابطه کلی همراه با جهت ویژگی را در حالی که ساختار زمانی را حفظ می کند، ثبت کند. این را می توان به عنوان گسترش زمانی مستقیم یک مدل فریم به فریم در نظر گرفت که ارتباط چنین ویژگی هایی را فقط در هر فریم ثبت می کند. برای گرفتن کارآمد ساختار زمانی با برد وسیع و در عین حال

حفظ ساختار ورودی، ژنراتور از لایه‌های نمونه‌برداری پایین، باقی‌مانده [۴۰] و نمونه‌برداری بالا تشکیل شده است، همانطور که در شکل۲ (الف) نشان داده شده است. نکته قابل توجه دیگر این است که CycleGAN-VC از یک CNN دردار [۳۶] برای ثبت ساختارهای متوالی و سلسله مراتبی ویژگی های صوتی استفاده می‌کند.

۲-۳ تشخیص دهنده: FullGAN

CycleGAN-VC از یک CNN دو بعدی [۵] برای تشخیص دهنده استفاده می کند تا روی ساختار دوبعدی (یعنی بافت طیفی ۲ بعدی [۴۱]) تمرکز کند. به طور دقیق تر، همانطور که در شکل ۳ (الف) نشان داده شده است، یک لایه تماماً متصل در آخرین لایه برای تعیین واقعیت با توجه به ساختار کلی ورودی استفاده می شود. چنین مدلی FullGAN نام دارد

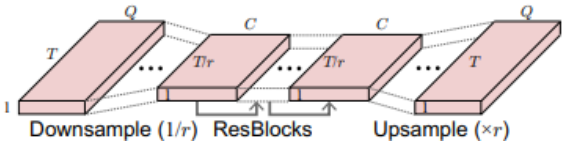
۳- CycleGAN-VC2

۳-۱ تابع هدف بهبود یافته: تابع ضرر متخاصم دو مرحله‌ای

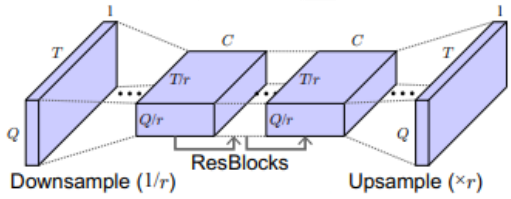
یکی از مشکلات شناخته شده برای مدل‌های آماری، هموارسازی بیش از حد ناشی از میانگین‌گیری آماری است. زیان خصمانه مورد استفاده در معادله. ۴ به کاهش این تخریب کمک می کند، اما از دست دادن ثبات چرخه که به عنوان L1 فرموله شده است، همچنان باعث صاف شدن بیش از حد می شود. برای کاهش این اثر منفی، ما یک تمایز اضافی D'_X را معرفی می کنیم و یک ضرر خصمانه را بر ویژگی تبدیل شده به صورت دایره‌ای تحمیل می کنیم.

$$\mathscr{L}_{adv2}(G_{X\to Y}, G_{Y\to X}, D'_X) = \mathbb{E}_{x\sim P_X(x)}\left[\log\left(1 - D'_X\left(G_{Y\to X}\left(G_{X\to Y}(x)\right)\right)\right)\right] + \mathbb{E}_{x\sim P_X(x)}[log D'_X(x)] \quad (5)$$

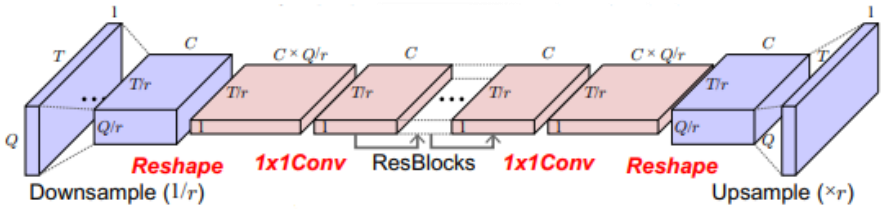
به طور مشابه، ما D'_Y را معرفی می کنیم و یک ضرر خصمانه $\mathscr{L}_{adv2}(G_{Y\to X}, G_{X\to Y}, D'_Y)$ برای نگاشت رو به جلوی معکوس اعمال می کنیم. ما این دو ضرر خصمانه را به معادله ۴ اضافه می کنیم. در این هدف بهبودیافته، همانطور که در شکل ۱(ب) نشان داده شده است، ما دو بار برای هر چرخه از ضررهای خصمانه استفاده می کنیم. از این رو، ما آنها را خسارات خصمانه دو مرحله‌ای می‌نامیم.



الف – شبکه CNN یک بعدی



ب – شبکه CNN دو بعدی



پ – شبکه CNN ترکیبی

شکل۲. مقایسه معماری شبکه مولد. بلوک های قرمز و آبی به ترتیب لایه های پیچشی ۱ بعدی و ۲ بعدی را نشان می دهند. r نشان دهنده نرخ پایین نمونه گیری یا افزایش نمونه است.

FullGAN

PatchGAN پیشنهادی

شکل۳. مقایسه معماری شبکه های تشخیص‌دهنده.

۳-۲ ژنراتور بهبود یافته: شبکه عصبی کانولوشنی ۲-۱-۲ بعدی

در یک چارچوب VC [۵، ۲۹] (از جمله CycleGAN-VC)، یک CNN یک بعدی (شکل ۲(الف)) معمولاً به عنوان یک مولد استفاده می شود، در حالی که در یک چارچوب پس فیلتر [۴۱، ۴۲]، یک CNN دو بعدی (شکل ۲(ب)) ارجح است. این انتخاب‌ها به مزایا و معایب هر شبکه مربوط می شود. یک CNN یک بعدی برای ثبت تغییرات دینامیکی امکان پذیرتر است، زیرا می تواند رابطه کلی را همراه با بعد ویژگی ثبت کند. در مقابل، یک CNN دو بعدی برای تبدیل ویژگی‌ها با حفظ ساختارهای اصلی مناسب تر است، زیرا منطقه تبدیل شده را به محلی محدود می کند. حتی با استفاده از یک CNN یک بعدی، بلوک‌های باقی‌مانده [۴۰] می‌توانند از دست دادن ساختار اصلی را کاهش دهند، اما متوجه می‌شویم که نمونه‌کاهی و نمونه‌افزایی (که برای گرفتن مؤثر سازه‌های برد وسیع ضروری هستند) به علت شدید این تخریب تبدیل می‌شوند. برای کاهش آن، ما یک معماری شبکه ای به نام CNN 2-1-2D ایجاد کرده ایم که در شکل ۲(ج) نشان داده شده است. در این شبکه، کانولوشن دوبعدی برای نمونه برداری پایین و بالا و از کانولوشن یک بعدی برای فرآیند تبدیل اصلی (یعنی بلوک های باقیمانده) استفاده می شود. برای تنظیم ابعاد کانال، پیچیدگی ۱×۱ را قبل یا بعد از تغییر شکل نقشه ویژگی اعمال می کنیم.

۳-۳ تشخیص‌دهنده‌ی بهبود یافته: PatchGAN

در مدل های قبلی پردازش گفتار مبتنی بر GAN [۴۱، ۴۲، ۵، ۲۹]، FullGAN (شکل ۳ (الف)) به طور گسترده مورد استفاده قرار گرفته است. با این حال، مطالعات اخیر در بینایی کامپیوتر [۴۳، ۴۴] نشان می‌دهد که میدان‌های دریافتی با دامنه وسیع تشخیص‌دهنده به پارامترهای بیشتری نیاز دارند، که باعث مشکل در آموزش می‌شود. با الهام از این، FullGAN را با PatchGAN [۴۵، ۴۳، ۴۴] جایگزین کردیم (شکل ۳ (ب))، که از کانولوشن در آخرین لایه استفاده می کند و واقعی بودن را بر اساس پچ تعیین می کند. ما به طور تجربی اثر آن را برای VC غیر موازی در بخش ۴-۲ بررسی می‌کنیم.

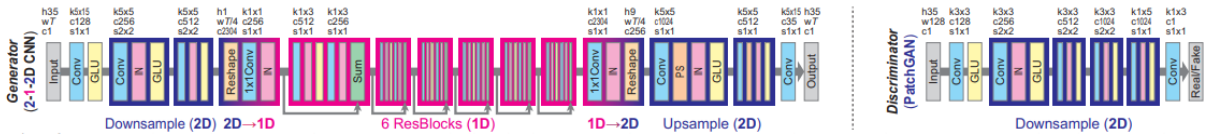
۴- آزمایش‌ها

۴-۱ شرایط آزمایشگاهی

مجموعه داده: ما روش خود را روی تسک Spoke (یعنی VC غیر موازی) [38] VCC 2018 ارزیابی کردیم، که شامل ضبط‌هایی از انگلیسی زبانان حرفه‌ای ایالات متحده است. ما زیرمجموعه‌ای از سخنرانان را انتخاب کردیم تا همه تبدیل‌های بین جنسیتی و درون‌جنسی را پوشش دهیم: VCC2SF3 (SF)، VCC2SM3 (SM)، VCC2TF1 (TF) و VCC2TM1 (TM)، که در آن S، T، F، و M به ترتیب منبع، هدف، ماده و مذکر را نشان می دهد. در ادامه از اختصارات داخل پرانتز (به عنوان مثال SF) استفاده می‌کنیم. ترکیبی از ۲ منبع (SF یا SM)2 × هدف (TF یا TM) برای ارزیابی استفاده شد. هر سخنران به ترتیب مجموعه های ۸۱ (حدود ۵ دقیقه؛ نسبتاً کمی برای VC) و ۳۵ جمله برای آموزش و ارزیابی دارد. در کار Spoke، گوینده منبع و هدف دارای مجموعه‌ای متفاوت از جملات (بدون همپوشانی) هستند تا در یک محیط غیر موازی ارزیابی شوند. ضبط‌ها برای این چالش به ۲۲،۰۵ کیلوهرتز کاهش یافتند. ما ۳۴ ضریب Mel-cepstral (MCEPs)، فرکانس پایه لگاریتمی (log F0) و پیوندهای دوره ای (Aps) را هر ۵ میلی‌ثانیه با استفاده از تحلیلگر WORLD استخراج کردیم [۴۶].

فرآیند تبدیل: روش پیشنهادی برای تبدیل MCEP ها (بعد Q = 34 + 1 شامل ضریب صفرم) استفاده شد. هدف از این آزمایش‌ها تجزیه و تحلیل کیفیت MCEPs تبدیل شده بود. بنابراین، برای بخش‌های دیگر، از روش‌های معمولی مشابه خط پایه VCC 2018 استفاده کردیم [۳۸]. به طور خاص، در تبدیل بین جنسیتی، از روش VC مبتنی بر Vocoder استفاده شد. F_0 بوسیله تبدیل نرمال شده لگاریتمی گاوسی [۴۷] تبدیل شد، APها مستقیماً بدون تغییر استفاده شدند، و کد صوتی WORLD [۴۶] برای سنتز گفتار استفاده شد. در تبدیل درون جنسیتی، از روش VC بدون Vocoder استفاده کردیم [۴۸]. به طور دقیق تر، ما MCEP های دیفرانسیل را با در نظر گرفتن تفاوت بین منبع و MCEP های تبدیل شده محاسبه کردیم. به یک دلیل مشابه، ما از هیچ پس فیلتر [۴۱، ۴۲، ۴۹] یا کد صوتی قدرتمندی مانند Vocoder WaveNet، [۵۰، ۵۱] استفاده نکردیم. گنجاندن آنها یکی از اهداف احتمالی کار آینده است.

جزئیات آموزش: پیاده سازی تقریباً مشابه CycleGAN-VC بود با این تفاوت که تکنیک های بهبود یافته در آن گنجانده شده بودند. جزئیات معماری شبکه در شکل ۴ آورده شده است. برای یک پیش فرآیند، ما MCEPهای منبع و هدف را با استفاده از آمار دیتاست آموزشی، به واریانس صفر میانگین و واحد تبدیل کردیم. برای پایدار کردن آموزش، از حداقل مربعات GAN (LSGAN) استفاده کردیم [۵۲]. برای افزایش تصادفی بودن داده‌های آموزشی، به‌جای استفاده مستقیم از یک جمله کلی، یک بخش (۱۲۸ فریم) را به‌طور تصادفی از یک جمله به‌طور تصادفی انتخاب کردیم. ما از بهینه‌ساز Adam [۵۳] با اندازه دسته‌ی ۱ استفاده کردیم. شبکه‌ها را برای تکرارهای ۲×۱۰۵ با نرخ یادگیری ۰،۰۰۰۲ برای مولد و ۰،۰۰۰۱ برای تشخیص‌دهنده و با ترم مومنتوم β1 برابر با ۰،۵ آموزش دادیم. ما $\lambda_{cyc} = 10$ و $\lambda_{id} = 5$ را تنظیم کردیم. از L_{id} فقط برای ۱۰۴ تکرار اول برای هدایت جهت یادگیری استفاده کردیم. توجه داشته باشید که ما از هیچ داده اضافی، ماژول یا روش تراز زمانی برای آموزش استفاده نکردیم.



شکل۴. معماری شبکه مولد و تفکیک کننده. در لایه های ورودی، خروجی و تغییر شکل، h، w و c به ترتیب نشان دهنده ارتفاع، عرض و تعداد کانال ها هستند. در هر لایه پیچیدگی، k، c و s به ترتیب اندازه هسته، تعداد کانال ها و اندازه گام را نشان می دهند. IN، GLU، و PS به ترتیب نشان دهنده عادی سازی نمونه [۵۴]، واحد خطی دروازه دار [۳۶] و پخش کننده پیکسل [۴۴] هستند. از آنجایی که ژنراتور کاملاً کانولوشنال است [۵۵]، می تواند ورودی با طول دلخواه T را دریافت کند.

جدول ۱. مقایسه MCD [dB]

No.	Method			Intra-gender		Inter-gender	
	CycleGAN-VC2			SF-TF	SM-TM	SM-TF	SF-TM
	Adv.	*G*	*D*				
1	1Step	2-1-2D	Patch	6.86±.04	6.32±.06	7.36±.04	6.28±.04
2	2Step	1D	Patch	6.86±.04	6.73±.08	7.77±.07	6.41±.01
3	2Step	2D	Patch	7.01±.07	6.63±.03	7.63±.03	6.73±.04
4	2Step	2-1-2D	Full	7.01±.07	6.45±.05	7.41±.04	6.51±.02
5	**2Step**	**2-1-2D**	**Patch**	**6.83±.01**	**6.31±.03**	**7.22±.05**	**6.26±.03**
6	CycleGAN-VC [29]			7.37±.03	6.68±.07	7.68±.05	6.51±.05
7	Frame-based CycleGAN [30]			8.85±.07	7.27±.11	8.86±.27	8.51±.36

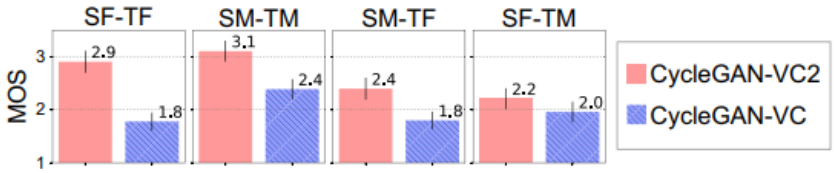
جدول ۲. مقایسه MSD [dB]

No.	Method			Intra-gender		Inter-gender	
	CycleGAN-VC2			SF-TF	SM-TM	SM-TF	SF-TM
	Adv.	*G*	*D*				
1	1Step	2-1-2D	Patch	1.60±.02	1.63±.05	1.54±.03	1.56±.04
2	2Step	1D	Patch	3.31±.36	4.26±.37	2.04±.21	5.03±.32
3	2Step	2D	Patch	1.57±.07	1.54±.01	1.46±.03	1.66±.07
4	2Step	2-1-2D	Full	1.52±.02	1.56±.04	1.47±.01	1.67±.06
5	**2Step**	**2-1-2D**	**Patch**	**1.49±.01**	**1.53±.02**	**1.45±.00**	**1.52±.01**
6	CycleGAN-VC [29]			2.42±.08	2.66±.08	2.21±.13	2.65±.15
7	Frame-based CycleGAN [30]			3.78±.26	2.77±.10	3.32±.06	3.61±.15

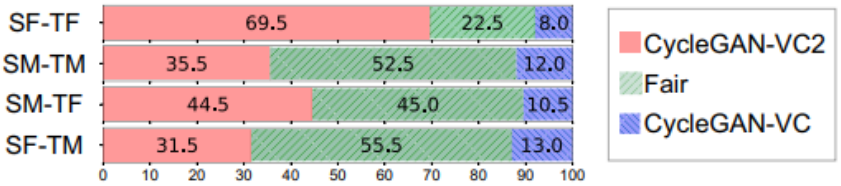
۴-۲ ارزیابی عینی

همانطور که در مطالعات قبلی [۷، ۴۱] بحث شد، طراحی یک معیار واحد که بتواند کیفیت MCEP های تبدیل شده را به طور جامع ارزیابی کند، نسبتاً پیچیده است. متناوبا، ما از دو معیار برای ارزیابی ساختارهای محلی و کلی استفاده کردیم. برای اندازه‌گیری تفاوت‌های ساختاری کلی، از پیچش Mel-cepstral (MCD) استفاده کردیم که فاصله بین تار دریافت و توالی‌های تبدیل‌شده MCEP را اندازه‌گیری می‌کند. برای اندازه‌گیری تفاوت‌های ساختاری محلی، ما از فاصله طیف مدولاسیون (MSD) استفاده کردیم، که به عنوان ریشه میانگین مربع خطا بین هدف و طیف مدولاسیون لگاریتمی تبدیل‌شده MCEPs به‌طور میانگین در تمام ابعاد MCEP و فرکانس‌های مدولاسیون تعریف می‌شود. برای هر دو معیار، مقادیر کوچکتر در دیکته که MCEPهای هدف و تبدیل شده مشابه هستند.

ما MCD و MSD را به ترتیب در جداول ۱ و ۲ فهرست می‌کنیم. برای حذف اثر مقداردهی اولیه، مقادیر میانگین و انحراف استاندارد را در سه مرحله اولیه تصادفی گزارش می‌کنیم. برای تجزیه و تحلیل اثر هر تکنیک، ما مطالعات فرسایشی را روی CycleGAN-VC2 انجام دادیم (شماره ۵ مدل کامل است). ما همچنین CycleGAN-VC2 را با دو روش پیشرفته مقایسه کردیم: CycleGAN-VC [29] و CycleGAN مبتنی بر فریم [۳۰] (پیاده سازی مجدد ما؛ ما علاوه بر این از L_{id} برای تثبیت تمرین استفاده کردیم). مقایسه تلفات متخاصم یک مرحله ای و دو مرحله ای (شماره ۱ و ۵) نشان می دهد که این تکنیک به ویژه برای بهبود MSD موثر است. مقایسه‌های معماری‌های شبکه مولد (شماره‌های ۲، ۳، ۵) و متمایزکننده (شماره‌های ۴، ۵) نشان می‌دهد که آنها به بهبود MCD و MSD کمک می‌کنند. در نهایت، مقایسه با خطوط پایه (شماره‌های ۵، ۶، ۷) تأیید می‌کند که با ترکیب سه تکنیک پیشنهادی، ما به عملکرد پیشرفته‌ای از نظر MCD و MSD برای هر جفت بلندگو دست می‌یابیم.



شکل ۵. MOS برای طبیعی بودن با فاصله اطمینان ۹۵٪.



شکل ۶. میانگین امتیاز ترجیحی (٪) در شباهت گوینده.

۴-۳ ارزیابی ذهنی

ما تست‌های شنیداری را برای ارزیابی کیفیت گفتار تبدیل شده انجام دادیم. CycleGAN-VC [۲۹] به عنوان خط پایه استفاده شد. برای اندازه گیری طبیعی بودن، ما یک آزمون میانگین امتیاز نظر (MOS) (۵: عالی تا ۱: بد) انجام دادیم، که در آن گفتار هدف را به عنوان مرجع در نظر گرفتیم (MOS برای TF و TM ۴،۸ است). ده جمله به صورت تصادفی از مجموعه های ارزیابی انتخاب شدند. برای اندازه گیری شباهت بلندگو، ما انجام دادیم.

آزمون XAB، که در آن «الف» و «ب» گفتار با خط مبنا و روش‌های پیشنهادی تبدیل شدند و «X» گفتار هدف بود. ما ده جفت جمله را به‌طور تصادفی از مجموعه‌های ارزیابی انتخاب کردیم و همه جفت‌ها را در هر دو ترتیب (AB و BA) برای حذف سوگیری در ترتیب محرک‌ها ارائه کردیم. برای هر جفت جمله، از شنوندگان خواسته شد که مورد دلخواه خود ("A"یا "B") یا "منصفانه" را انتخاب کنند. ده شنونده در این آزمون های شنیداری شرکت کردند. شکل های ۵ و ۶ به ترتیب MOS را برای طبیعی بودن و نمرات ترجیحی برای شباهت بلندگو نشان می دهند. این نتایج تأیید می کند که CycleGAN-VC2 از CycleGAN-VC از نظر طبیعی بودن و شباهت برای هر جفت بلندگو بهتر عمل می کند. به ویژه، CycleGAN-VC برای اعمال چارچوب VC بدون Vocoder [۴۸] (مورد استفاده در SF-TF و SM-TM) دشوار است، زیرا این چارچوب به دلیل استفاده از MCEP های دیفرانسیل به خطای تبدیل حساس است. با این حال، MOS نشان می دهد که CycleGAN-VC2 در چنین شرایط دشواری نسبتاً خوب کار می کند.

۵- نتایج

برای پیشبرد تحقیقات در مورد VC غیر موازی، CycleGAN-VC2 را پیشنهاد کرده‌ایم که نسخه بهبودیافته CycleGAN-VC است که سه تکنیک جدید را در خود جای داده است: یک هدف بهبود یافته (تلفات خصمانه دو مرحله‌ای)، ژنراتور بهبودیافته (۲-۱-۲ بعدی)، و تشخیص دهنده بهبود یافته (Patch GAN). نتایج تجربی نشان می‌دهد که CycleGAN VC2 از CycleGAN-VC در هر دو معیار عینی و ذهنی برای هر جفت گوینده بهتر عمل می‌کند. تکنیک‌های پیشنهادی ما به VC یک به یک محدود نمی‌شود، و تطبیق آنها با تنظیمات دیگر (به عنوان مثال، VC چند دامنه‌ای [۵۶]) و سایر برنامه‌ها [۱، ۲، ۴، ۳، ۵] همچنان بعنوان یکی از برنامه‌های آینده باقی می‌ماند.

قدردانی: این کار توسط JSPS KAKENHI 17H01763 پشتیبانی شد.