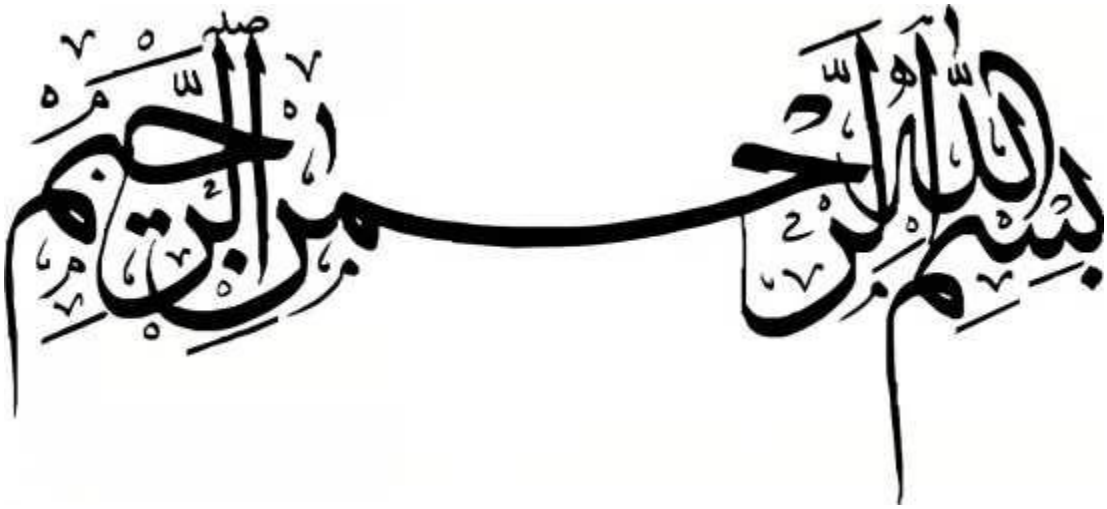


نام و نام خانوادگی : مژگان دهقان آزاد

نام استاد : جناب دکتر مهدی اسلامی

موضوع مقاله : تبدیل صدای آواز شات صفر

نام درس :DSP

شماره دانشجویی : ۴۰۰۱۴۱۴۰۱۱۱۰۶۶

چکیده

در این مقاله، ما استفاده از تعبیه بلندگو را پیشنهاد می کنیم شبکه هایی برای انجام تبدیل صدای آواز بدون شات، و دو معماری را برای تحقق آن پیشنهاد کنید. کاربرد شبکه های تعبیه کننده بلندگو نه تنها قابلیت انطباق با صداهای جدید را در لحظه امکان پذیر می کند، بلکه اجازه می دهد آموزش مدل بر روی داده های بدون برچسب. این نه تنها تسهیل می کند مجموعه ای از داده های صوتی آواز مناسب، بلکه اجازه می دهد شبکه‌هایی که باید قبلاً بر روی پیکره‌های گفتاری بزرگ آموزش داده شوند بهبود در مجموعه داده های صوتی آواز، بهبود شبکه تعمیم است.

ما اثربخشی الگوریتم‌های تبدیل صدای آواز بدون شات پیشنهادی را توسط هر دو به معنای کیفی و کمی میبینیم.

۱. مقدمه

تبدیل صدای آواز (SVC) تغییر شکل است

اجرای آواز از یک خواننده به خواننده دیگر. می توان از آن برای دستکاری های خلاقانه استفاده کرد صدایی که بسیار فراتر از کشش زمانی سنتی و تغییر گام/فرمانت [۱]

روش‌های SVC باید یاد بگیرند که محتوای بلندگو را از ویژگی‌های صوتی [۲] جدا کنند.

حفظ دقیق اطلاعات صوتی و صدای ورودی در خروجی تبدیل شده نسبت به روش‌های مشابهی که برای گفتار به کار می‌رود، صدای آواز گام بزرگ‌تری را نشان می‌دهد.

محدوده و به طور کلی انتقال آهسته تر بین آوایی واحدهایی که شبکه های تبدیل باید بتوانند آنها را تطبیق دهند [۳، ۴].

اکثر رویکردهای SVC به نوعی از Vocoder متکی هستند که شکل موج های صوتی را سنتز می کند. سپس وظیفهSVC تبدیل به یکی از ویژگی‌های رمزگذار صوتی از هر اجرای یک خواننده منبع به صدای هدف می‌شود برخلاف رویکردهای تبدیل صدا که معمولاً از کدهای صوتی عصبی مانند WaveNet [۵] یا WaveRNN [۶]

به عنوان سینت سایزر گفتار پشتیبان آنها،SVC و الگوریتم‌های سنتز آواز تمایل دارند از طراحی دستی استفاده کنند.

کدهای صوتی مانند WORLD [۷] برای مدل سازی آکوستیک و (به استثنای برخی موارد مانند[۸]).

این به این دلیل است که آنها صریحاً گام را از تمبرال جدا می کنند اجزاء [۳، ۴، ۹]. بر این اساس، امکان یادگیری وجود دارد دگرگونی های تمبرال با حفظ گام، که است معمولاً هنگام استفاده از کد صوتی عصبی [۲] تضمین نمی شود.

این ممکن است به قیمت کاهش بیانی نسبت به صداگذارهای عصبی باشد، اما قابل قبول است با توجه به ویژگی های حفظ گام آن[۴] شبکه های متخاصم مولد (GANs) [۳،۹،۱۰] رمزگذارهای خودکار متغیر و (VAEs) [۱۱] به انتخاب های محبوبی برای یادگیری تبدیل ویژگی هایVocoder تبدیل شده و هم برای سنتز آواز و هم برایSVC ، استراتژی های مختلف برای مدل سازی چندین صدای هدف مورد بررسی قرار گرفته‌اند.

به طور خاص، برای انطباق سیستم ها برای صداهای جدید که در طول آموزش مدل دیده نمی شوند. یکی از این راهبردها شامل تخصیص است جاسازی تصادفی به صدای غیبی و از سرگیری آموزش مدل بر روی داده های صدای غیبی برای به روز رسانی

این جاسازی و انجام هر گونه اصلاحات لازم برای مدل [۱۲، ۱۳]. اخیراً الگوریتم های تبدیل در حوزه گفتار، از شبکه‌های جاسازی بلندگوی از پیش آموزش‌دیده‌ای که برای [۱۴] کارهای تأیید بلندگو طراحی شده‌اند استفاده کرده‌اند.

به منظور رمزگذاری هویت گوینده [۱۵] این رویکردها این مزیت را دارند که پس از آموزش گوینده با تعبیه شبکه بر روی بسیاری از بلندگوها، ریتم های الگوی تبدیل را می توان به صورت صفر شات با صداهای جدید تطبیق داد بدون نیاز به آموزش بیشتر مدل و با تعداد کمی به عنوان نمونه ای از صدای غیبی در این مقاله، ما تبدیل صدای صفر شات را تطبیق می دهیم روش شناسی [۱۵] با استفاده از شبکه های تعبیه کننده بلندگو است.

برای کاربرد SVC ما از Vocoder WORLD استفاده می کنیم و دو معماری را برای

برای اجرای صفر شات SVC پیشنهاد کنید.

ما نشان می دهیم که ماهیت صفر شات الگوریتم است امکانSVC روی داده های

بدون برچسب را فراهم می کند علاوه بر این، ما مطرح می کنیم که سیستم های SVC برای آموزش اولیه در بزرگ قابل قبول هستند

مجموعه داده‌های گفتاری که به طور گسترده‌تری در دسترس هستند، به دنبال آن قرار گرفتند با انطباق مدل بر روی مجموعه داده های صدای آوازخوان کوچکتر به بهترین دانش ما، این اولین کاری است که باید به آن پرداخته شود.

SVC صفر شات بر خلاف الگوریتم های سنتز آواز، مانند[۴، ۱۰، ۱۳]

همانطور که نیازی به حاشیه نویسی از پیش تعریف شده ندارد انتقال های آوایی یا گام، زیرا این اطلاعات از ویژگی های صوتی عملکرد منبع استخراج می شود

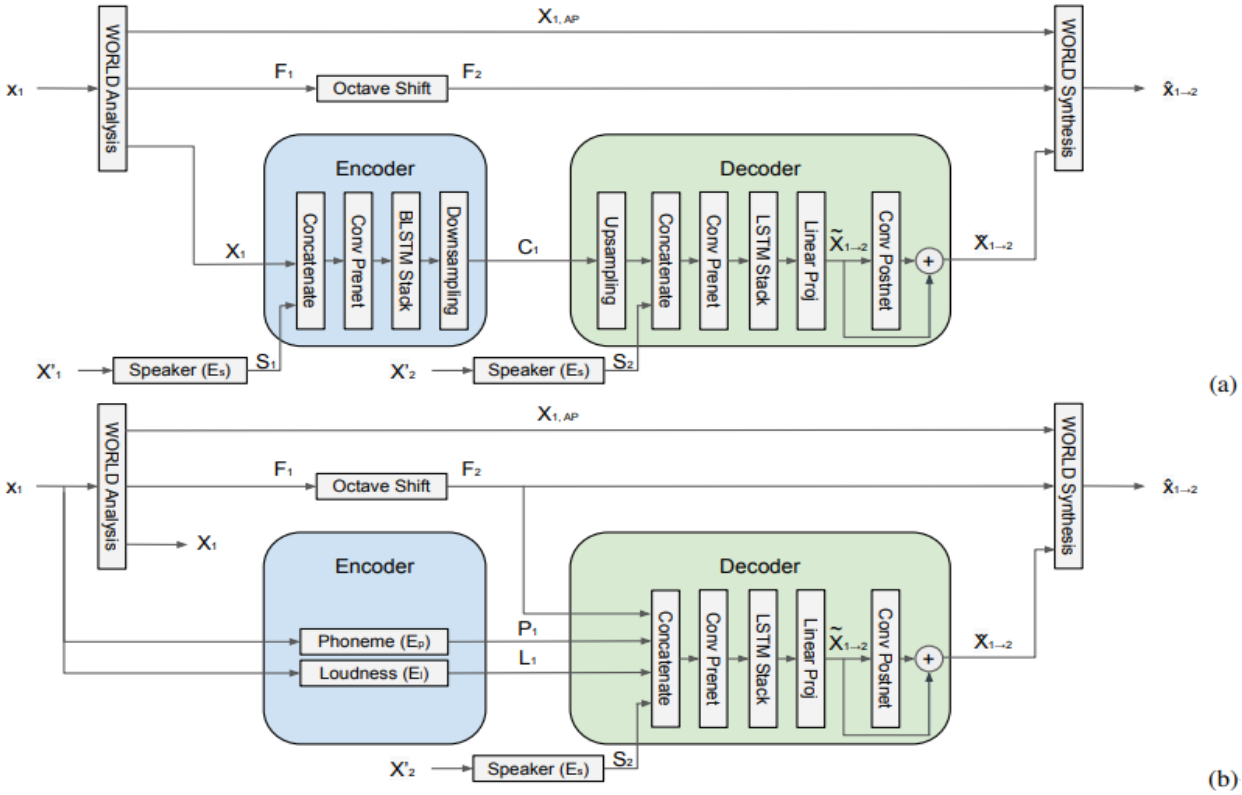
ساختار باقی مانده این مقاله به شرح زیر است:

ما دو معماری را برای SVC صفر شات در بخش پیشنهاد می کنیم.

ارزیابی عملکرد مدل از طریق ابزارهای کمی و کیفی در بخش ۳ در نهایت، ما نتیجه گیری می کنیم و به کارهای آینده در بخش ۴ اشاره کنید.

۲. الگوریتم های SVC

ما از Vocoder WORLD برای تجزیه و تحلیل و سنتز استفاده می کنیم از آواز های آواز به دلیل توانایی آن در جداسازی تن و اجزای زمین به طور خاص، سیستم تجزیه می شود یک سیگنال صوتی به یک پوشش طیفی هارمونیک و یک پاکت تناوب، بر اساس یک برآورد تونالیته



شکل ۱. (الف) معماری شبکه رمزگذار ثابت AutoVC و(ب) برای SVC صفرشات

فرکانس اساسی وظیفه تبدیل در درجه اول شامل تبدیل پوشش طیفی هارمونیک است که پوشش تناوبی را بدون تغییر باقی می گذارد. مانند در [۴]، ابعاد طیف هارمونیک را کاهش می دهیم پاکت تا ۶۰ ضریب در هر مرحله زمانی، با استفاده از تاب خوردگی فرکانس کوتاه در حوزه مغزی با یک ضریب تاب برداشتن تمام قطب [۱۶] α = 0.45 در نظر می گیریم. دو معماری مختلف برای SVC همانطور که در شکل ۱ نشان داده شده است با الهام از [۱۸،۲، ۱۵، ۱۷] .

AutoVC ۱.۲

اولین معماری اقتباسی از معماری AutoVC [۱۵] برای آواز خواندن است که بر روی هارمونیک عمل می کند پاکت های طیفی استخراج شده ازWORLD(بجای Mel طیف‌نگاری‌هایی که در نهایت به WaveNetوارد می‌شوند Vocoder همانطور که در کار اصلی است) شامل می شود از یک شبکه جاسازی بلندگو(۰)Es که به عنوان ورودی a میباشد.

طیف سنجی مل و یک تک بعدی ثابت تولید می کند تعبیه بلندگو، یک رمزگذار محتوا (۰)E که به عنوان پوشش طیفی هارمونیک و جاسازی بلندگو را از یک منبع وارد کرده

ویک رمزگذاری پنهان ایجاد می کند و یک شبکه رمزگشا D(۰) که پوشش طیفی هارمونیک تبدیل شده را از یک رمزگذاری نهفته می سازد و هدف قرار دادن بلندگو

ورودی رمزگذار، x1 است که از یک عبارت منبع X1 محاسبه می شود

این به یک بلندگوی منبع متصل است که S1 =Es (X_1') پوشش طیفی هارمونیک را تعبیه کرده است.

در هر مرحله زمانی، جایی که X_1' یک طیف نگار Mel از عبارت یکسان بالقوه متفاوت X_1' از همان سخنران منبع رمزگذار از یک کانولوشن تشکیل شده است متشکل از prenet سه لایه کانولوشنیک ۱ بعدی با۵۱۲ کانال خروجی و اندازه هسته ۵ ، هر کدام به دنبال آن عادی سازی دسته ای و فعال سازیReLU این نتیجه است.

که از دو لایه LSTM دو طرفه با ابعاد سلول جلو و عقب ۳۲ عبور می کند که کدگذاری بعد ۶۴ را به دست می دهد. این به طور موقت نمونه برداری شده است با ۳۲، محتوای رمزگذاری کننده C1 را به دست می دهد. گنجاندن از S1

در شبکه رمزگذار به رمزگذار کمک می کند تا راحت تر بتواند رمزگذاری مستقل از

بلندگو را یاد بگیرید.

رمزگشا با نمونه برداری از رمزگذاری C1پنهان به وضوح زمانی اولیه آن شروع می شود. با توجه به مل طیف گرا $X_{۲}{}'$ برخی گفته ها $x_{۲}{}'$ از همان بلند گوی هدف x2 جاسازی بلندگو عبارت هدف $S2 = Es(X_2{}')$ با نمونه آپلود شده الحاق می شود رمزگذاری ویژگی‌های به هم پیوسته از یک پرشبکه کانولوشنال مشابه آنچه در رمزگذار وجود دارد، عبور می‌کنند

توسط سه لایه به LSTM با ابعاد سلول ۱۰۲۴ خروجی های لایه LSTMبه صورت خطی به ابعاد۶۰ پیش بینی می شوند که به عنوان تخمین اولیه عمل میکند.

$\tilde{X}_{1\to2}$ از تبدیل شده پوشش طیفی هارمونیک این برآورد اولیه اصلاح شده است

با استفاده از یک پست شبکه کانولوشن متشکل از پنج تا۱ بعدی لایه های کانولوشن با اندازه هسته ۵ . نرمال سازی دسته ای به چهار لایه اول اعمال می شود و هر کدام خروجی ۵۱۲ کانال لایه نهایی هیچ فعال سازی اعمال نمی کند و خروجی ۶۰ کانال. طیف هارمونیک تبدیل شده است.

$\hat{X}_{1\to2}$ با افزودن خروجی postnet به $\tilde{X}_{1\to2}$ تولید میشود.

در حین تمرین، تنظیم کردیم

$x_1 = x_2$،

$S1 = S2$ ،

و بر این اساس $X1 = X2$ ،

$\tilde{X}_{1\to1} = \tilde{X}_{1\to2}$،

$\hat{X}_{1\to1} = \hat{X}_{1\to2}$،

تابع هدف مورد استفاده برای آموزش AutoVC است.

$$\begin{aligned}\mathcal{L} =& \mathrm{E}[|X_1 - \hat{X}_{1\to1}|_2^2] + \\ & \mu\mathrm{E}[|X_1 - \tilde{X}_{1\to1}|_2^2] + \\ & \lambda\mathrm{E}[|E(X_1, S_1) - E(\hat{X}_{1\to1}, S_1)|_1] \end{aligned} \quad (1)$$

اولین عبارت، از دست دادن بازسازی بین پوشش های طیفی هارمونیک اصلی و بازسازی شده است. این عبارت دوم یک از دست دادن بازسازی بین اصلی است

و در ابتدا پوشش های طیفی هارمونیک را تخمین زدند که به طور تجربی به همگرایی مدل کمک می کند. اصطلاح سوم الف است از دست دادن رگرسیور نهفته [۱۹] جریمه کردن تفاوت در کدگذاری بین طیف هارمونیک اصلی و تبدیل شده

پاکت نامه ها در عمل، هایپرپارامترهای λ و μ را می توان روی ۱ تنظیم کرد [۱۵]. این مدل به عنوان رمزگذار خودکار آموزش داده شده است

امیدوارم که گلوگاه آن به اندازه ای کوچک باشد که بتوان از هم جدا شود هویت گوینده اما به اندازه کافی بزرگ است که امکان دقیق را فراهم کند

بازسازی

در طول استنتاج، S2 را می توان روی تعبیه بلندگوی برخی از خواننده های هدف تنظیم کرد تا یک تبدیل را انجام دهد. داده شده

یک کانتور گام منبع F1 استخراج شده در طول تجزیه و تحلیل WORLDکانتور گام ، هدف F2 باید برای تطبیق با رجیستر خواننده هدف تنظیم شود، و بنابراین:

$$F2 = F1 + F\Delta 1 \rightarrow 2$$

تغییر گام $F\Delta 1\rightarrow 2$را می توان به طور خودکار با اندازه گیری گام های میانه تعیین کرد عملکرد منبع و هدف، و در نظر گرفتن تفاوت آنها به نزدیکترین اکتاو گرد شده است.

پوشش طیفی تناوب عملکرد منبع$X_{1,AP}$ همانطور که هست استفاده میشود این شکل موج صوتی تبدیل شده $x\hat{}1\rightarrow 2$ با تغذیه محاسبه می شود

پوشش طیفی هارمونیک تبدیل شده، پوشش طیفی تناوب منبع، و کانتور گام هدف F2 به عنوان ورودی به موتور سنتز WORLD است.

۲. ۲ مدل رمزگذار ثابت

معماری دوم مشابه AutoVC است، اما جایگزین می شود رمزگذار$E(\cdot)$

با تعدادی سیگنال شرطی، مانند مواردی که در [۲] یافت می شود. با طراحی، این تهویه سیگنال‌ها ورودی را به روشی مستقل از بلندگو با استفاده از ویژگی‌های صریح، شبیه به شبکه‌های انتقال صدا رمزگذاری می‌کنند

در [۱۸] ما محتوای زبانی را با استفاده از پسین‌گرام‌های آوایی (PPGs) استخراج‌شده از طبقه‌بندی‌کننده واج می‌گیریم(۰)Epمانند [۱۷] ،.

طبقه بندی کننده فرکانس ۴۰ Mel را عبور می دهد ضرایب مغزی (MFCCs)

در هر فریم از طریق دو LSTM دو طرفه با ۱۲۸ واحد در هر جهت. یک فینال

لایه متراکم با فعال سازی softmax طبقه بندی کننده را ایجاد می کند .

خروجی، که با برچسب های حقیقت زمینی با استفاده از a مقایسه می شود .

از دست دادن متقابل آنتروپی طبقه ای در طول تمرین. ما آموزش دیدیم که

شبکه در مجموعه داده[20] TIMIT، با استفاده از موارد تجویز شده آن ،

مجموعه های آموزشی و تستی مجموعه داده شامل صدا و مهرهای زمانی سطح

نمونه انتقال آوایی از یک از ۶۱ کلاس (از جمله کلاس سکوت). خروجی از بنابراین،

طبقه‌بندی‌کننده واج یک بردار ۶۱ بعدی در هر فریم زمانی است. دقت طبقه بندی در

مجموعه تست ۶۵ درصد است که برای عمل کردن کافی است یک نماینده مستقل از

گوینده از محتوای زبانی است.

ما اطلاعات بلندی صدا (L)را با استفاده از مراحل محاسباتی (۰)El مثل [۲۱]

استخراج میکنیم یکA-weighted را محاسبه می کنیم.

طیف قدرت، که تاکید بیشتری بر بالاتر دارد فرکانس ها نتیجه در تمام فرکانس ها .جمع می شود و به دسی بل تبدیل می شود تا یک مقدار بلندی صدا ایجاد شود

(درdbA) در هر مرحله زمانی. در نهایت، ما هدف را درج می کنیم.

کانتور زمین F2رمزگشا کانتور گام هدف F2 را به هم متصل می کند.

P1 = Ep (x1)، L1 = El(x1)

با بلندگوی هدف کهS2 را تعبیه کرده است. گنجاندن این شرطی سازی های مختلف سیگنال تلاش برای به حساب آوردن تغییرات تیمبرال که ممکن است به عنوان تابعی از زیر و بم و پویایی یک عملکرد خاص تغییر می کند، در حالی که به رمزگشای محتوای زبانی زیرین آن دستور می دهد. شبکه رمزگشا تقریباً است.

مشابه آنچه در AutoVC وجود دارد، با این تفاوت که ما عملیات نمونه برداری را حذف می کنیم زیرا دیگر نیازی به ایجاد یک گلوگاه اطلاعاتی برای از هم گسیختگی بلندگو نداریم. ما از این معماری به عنوان مدل رمزگذار ثابت یاد می کنیم، زیرا

همه سیگنال‌های شرطی یا بدون شبکه عصبی محاسبه می‌شوند، یا با استفاده از یک شبکه عصبی از پیش آموزش‌دیده که وزنه ها در حین تمرین شبکه رمزگشا منجمد می شوند. هدف آموزش مشابه همان است که در(1) .Eqnبا این تفاوت که اصطلاح سوم دیگر قابل اجرا نیست بنابراین حذف میشود . توجه داشته باشید که در این مورد، پوشش طیفی هارمونیک X1منبع هرگز به عنوان ورودی به شبکه ارسال نمی شود بلکه به عنوان یک هدف برای بازسازی در طول آموزش استفاده می شود.

۳. ۲ مقایسه معماری

ما بطور تصوری مزایا و معایب بالقوه مرتبط با معماری های پیشنهادی در اینجا را مورد بحث قرار می دهیم مزیت اصلی معماریAutoVC این است که به مجموعه آموزشی

اختصاصی برای استخراج اطلاعات آوایی متکی نیست. این اطلاعات توسط خود رمزگذار در طول آموزش مدل یاد می شود. این به طور بالقوه می تواند مفاهیم بهتری برای کاربردهای بین زبانی دارند، در موردی که مجموعه ای از برچسب های واجی خود یک مجموعه داده است تعصب زبانی را معرفی می کند [۲۲]. با این حال، متحمل می

شود برخی از خطرات، زیرا رمزگذار صرفاً مسئول یادگیری تمام تغییرات صدا در صدا است. همچنین مستلزم آن است یک نمونه برداری/نمونه برداری موقت از رمزگذاری آن به یک گلوگاه اطلاعاتی برای از هم گسیختگی بلندگو ایجاد کنید که پیامدهای تأخیر اضافی در آن دارد رمزگشا معماری رمزگذار ثابت از نظر محاسباتی فشرده‌تر است، زیرا طبقه‌بندی‌کننده واج به طور قابل‌توجهی کوچک‌تر از شبکه رمزگذار در AutoVC است. آن را نیز از نیاز به نمونه برداری موقتی اجتناب می کند معایب اصلی این معماری تکیه است در مورد داده ها برای آموزش یک طبقه بندی کننده واجی، و همچنین این واقعیت که بیان آن محدود به آن چیزی است که توسط سیگنال های شرطی سازی ارائه می شود.

۴.۲ مدل پس‌زمینه جهانی(UBM)

در حالی که ما می توانیم به سادگی شبکه های SVC را "از ابتدا" آموزش دهیم در مورد آواز خواندن مجموعه داده‌های صوتی، ما از این واقعیت جالب استفاده می‌کنیم که استفاده از تعبیه‌های بلندگو برای رمزگذاری هویت صوتی (به‌جای برچسب‌های تک داغ) به سیستم اجازه می‌دهد تا بر روی داده‌های بدون برچسب آموزش داده شود.

مسلماً هر "تمیز" هم اکنون می‌توان از کلیپ صدای نوازش یا آواز برای آموزش سیستم‌های SVC استفاده کرد. به طور کلی درک می شود که وجود دارد داده های گفتاری به طور قابل توجهی بیشتر از صدای آواز آنهاست داده ها برای اهداف تحقیق نامگذاری وام گرفتن از جامعه تشخیص گفتار، یک پیشآموزش اولیه در بدنه‌های گفتاری بزرگ مانند آموزش یک UBM [۲۳] است.

کدام شبکه های دیگر را می توان برای موارد خاص تر وظیفه SVC تطبیق داد ما امیدواریم که چنین مدلی در خدمت باشد به عنوان یک شرط اولیه بهتر برای آموزش شبکهSVC نسبت به وزن های تصادفی و این که سیستم به دست آمده در حداقل به صداهای بیشتر تعمیم دهید.

۳. نتایج تجربی

۳. ۱ راه اندازی آزمایشی

دو مجموعه داده برای آموزش شبکه های تبدیل استفاده می شود در این کار

ما از پیکره استفاده می کنیم که شامل VCTK بیش از ۴۰ ساعت سخنرانی از ۱۰۹ سخنران [۲۴]. این مجموعه به عنوان یک مجموعه داده گوینده نظارت شده برای مقایسه عمل می کند.

UBM عملکرد بین شبکه های zeroshotتحت نظارت و (بدون نظارت)، و همچنین مجموعه داده به اندازه کافی بزرگ برای آموزش یک برای تنظیم دقیق مدل بیشتر. همانطور که در[۱۵]است.

ما ۹۰٪ از داده های هر سخنران را برای آموزش حفظ می کنیم و باقیمانده را به عنوان یک مجموعه آزمایشی ذخیره کنید. علاوه بر این، ما ازa استفاده می کنیم .

مجموعه داده اختصاصی و بدون برچسب متشکل از ۷ ساعت خواندن داده های صوتی، که ما به سادگی آن را مجموعه دادهSVC می نامیم .

باز هم، ما ۹۰٪ از داده ها را برای آموزش حفظ می کنیم و ذخیره می کنیم

باقی مانده به عنوان یک مجموعه آزمایشی توجه داشته باشید که عدم وجود برچسب در این مجموعه داده هیچ مشکلی برای آموزش شبکه صفر شات ایجاد نمی کند

ما از جاسازی اسپیکر منبع باز استفاده می کنیم شبکه ۱ برای به حداقل رساندن اتلاف انتها به انتها تعمیم یافته از قبل آموزش دیده است [۱۴].

این شبکه تعبیه کننده بلندگو یک بلندگوی ۲۵۶ بعدی را از یک باند ۴۰ تولید می کند طیف نگار Mel با استفاده از معماری LSTM و حفظ تنها خروجی از مرحله زمانی نهایی در طول آموزش، ما از یک گفته کامل برای x_1'استفاده می کنیم

در حالی که x1 برش دوم از همین گفته شبکه تعبیه کننده بلندگو و طبقه بندی واج از پیش آموزش داده شده اند و در طول آموزش شبکه های تبدیل منجمد شد.

همه مدل ها با فرکانس ۱۶ کیلوهرتز با نرخ فریم کار می کنند

۱۲/۵میلی‌ثانیه و با استفاده از اندازه دسته ۲ آموزش دیدند بهینه سازADAM

و نرخ یادگیری ۱۰- ۳ است.

ما چهار پیکربندی را برای هر معماری مدلی که در اینجا توضیح داده شده است

اولین پیکربندی، VCTK (یک داغ)، است که آموزش میدهیم بر روی مجموعه

VCTK با استفاده از برچسب های ارائه شده توسط مجموعه داده که به یک نمایش یک داغ تبدیل می شوند و به عنوان S1 به شبکه تغذیه میشود این پیکر بندی خدمت می کند به عنوان یک پایه برای مقایسه با همتای صفر شات خود

پیکربندی دومVCTK (شات صفر)، آموزش داده شده است مجموعه VCTK با استفاده از تعبیه‌های بلندگو برایS1 را دو پیکربندی اول هر کدام برای ۱۵۰۰۰۰ مرحله آموزش داده شده اند.

در پیکربندی سوم، SVC (شات صفر)، معماری‌های zeroshot را روی مجموعه داده SVC برای ۵۰۰۰۰۰ مرحله آموزش می‌دهیم که در پیکربندی نهایی VCTK→SVC (شات صفر)، پیکربندی دوم به عنوان حالت اولیه استفاده می شود و آموزش برای ۳۵۰۰۰۰ مرحله در مجموعه داده SVC (در مجموع ۵۰۰۰۰۰ مرحله) از سر گرفته شد. برای نمونه های صوتی لطفا به سایت مراجعه کنید.

سایت دمو ، مرتبط با این مقاله

۳. ۲ ارزیابی عملکرد

ما شبکه ها را از نظر کیفی و کمی ارزیابی می کنیم به معنای هدف اصلی این مقاله نشان دادن آن است در واقع می توان از شبکه های تعبیه شده بلندگو استفاده کرد صفر شات آموزش شبکه های SVC از آنجایی که ما بی خبر هستیم از هر روش منتشر شده دیگری برای SVC صفر شات مانند همانطور که در اینجا معرفی شد و به منظور ارائه برخی در قالب تجزیه و تحلیل مقایسه ای، ما توجه خود را به تجزیه و تحلیل تفاوت در نتایج بین پیکربندی های آموزشی که در اینجا ذکر شده است متمرکز می کنیم برای ارزیابی کمی ما، ما گزارش تلفات بازسازی برای هر شبکه (اولین اصطلاح در معادله(۱) که وقتی بر روی هارمونیک محاسبه می شود.

پوشش های طیفی، به طور موثر به عنوان یک متریک اعوجاج مغزی Mel عمل می کند. برای ارزیابی کیفی خود، نظرسنجی هایی را با ۱۵ شرکت کننده در سازمان خود انجام دادیم کسانی که تجربه شنیداری انتقادی دارند و جدول بندی شده اند میانگین نمرات نظر (MOS) ما نظرسنجی های جداگانه انجام می دهیم برای کیفیت

تبدیل کلی و شباهت به هدف صدا در حالی که ما نمونه هایی از هر دو معماری ارائه می دهیم در مطالب تکمیلی این کار، ما خود را به نمونه‌های تولید شده از انواع آموزشی محدود می‌کنیم معماری رمزگذار ثابت برای ارزیابی های ذهنی را

اولین دلیل برای این محدودیت صرفاً به حداقل رساندن آن است تعداد گزینه های گوش دادن به طوری که شرکت کنندگان در نظرسنجی را تحت تاثیر قرار ندهد. دلیل دوم این است زیرا گنجاندن برچسب های یک بلندگوی داغ برایS1در شبکه رمزگذار AutoVC به این ورودی نیاز دارد نمونه های منبع از مجموعه بلندگوهای بسته آن می آیند بنابراین استفازه از پیکربندی آموزشیVCTK (onehot) در AutoVC در صدای آواز عملا امکان پذیر نیست نمونه هایی بدون حذفS1 از شبکه پیشرو به یک مقایسه بالقوه ناعادلانه نتایج تجزیه و تحلیل کمی ما در هر دو مورد ارزیابی قرار گرفت مجموعه داده هایVCTK وSVCدر جدول به ترتیب ۱و۲ نشان داده شده

در هر دو معماری، ما می توانیم تایید کنیم که جایگزینی برچسب‌های تک داغ با تعبیه‌های بلندگو به طور چشمگیری به عملکرد تبدیل لطمه نمی‌زند.

	AutoVC	Fixed Encoder
VCTK (one-hot)	0.1837	**0.1882**
VCTK (zero-shot)	**0.1634**	0.1891
SVC (zero-shot)	0.2930	0.3590
VCTK→SVC (zero-shot)	0.2557	0.3232

	AutoVC	Fixed Encoder
VCTK (one-hot)	N/A	N/A
VCTK (zero-shot)	0.3007	0.4314
SVC (zero-shot)	0.1650	0.1959
VCTK→SVC (zero-shot)	**0.1439**	**0.1850**

جدول ۱. از دست دادن بازسازی در مجموعه آزمایش VCTK

جدول ۲. تلفات بازسازی در مجموعه تست SVC

در واقع می بینیم که برای معماریVCTK، AutoVC(صفر شات)

در واقع بهتر از VCTK (یک داغ)، در حالی که ارائه قابلیت های اضافه شده

از تطبیق صفر شات به صداهای نادیده جدید این نتیجه با یافته های [۱۵] مطابقت دارد. توجه داشته باشیم که هنگام استفاده مستقیم ازVCTK (صفر شات) در نمونه های صوتی آواز، یا هنگام اعمال شبکه‌هایSVC کاهش قابل مستقیماً در VCTK توجهی در عملکرد ارزیابی شده از نظر کمی وجود دارد.

نمونه‌ها، نشان می‌دهد که واقعاً بین حوزه‌های گفتار و صدای آواز ناهماهنگی وجود دارد. وجود دارد بهبود مداوم در هنگام استفاده از استراتژی تطبیق پیشنهادی ما، با VCTK→SVC(شات صفر) بهتر از SVC (شات صفر)، هم در حوزه گفتار و هم در موارد دیگر است.

مهمتر از همه، در حوزه صدای آواز مورد علاقه. در کل، بهترین روش اجرا برای آواز خواندن مبتنی بر صدا در این ارزیابی کمی با استفاده ازAutoVCآموزش دیده است.

با این حال، پیکربندی آموزشی VCTK→SVC(شات صفر)

مدل رمزگذار ثابت محاسباتی سبک‌تر و ثابت، به‌طور قابل‌توجهی به خوبی عمل می‌کند. شایان ذکر است کهVCTKپیکربندی (یک داغ) برای ارزیابی قابل اجرا نیست. مجموعه دادهSVC زیرا توانایی آنی را برای سازگاری با صداهای جدید ندارد.

در واقع می بینیم که برای معماریVCTK، AutoVC (صفر شات) در واقع بهتر از VCTK(یک داغ)، در حالی که ارائه قابلیت های اضافه شده از تطبیق صفر شات به صداهای نادیده جدید این نتیجه با یافته های [۱۵] مطابقت دارد. توجه داشته باشیم که هنگام استفاده مستقیم از VCTK(صفر شات) در نمونه‌های صوتی آواز، یا هنگام اعمال شبکه‌های SVC مستقیماً درVCTKکاهش قابل توجهی در عملکرد ارزیابی شده از نظر کمی وجود دارد.

نمونه‌ها، نشان می‌دهد که واقعاً بین حوزه‌های گفتار و صدای آواز ناهماهنگی وجود دارد. وجود دارد بهبود مداوم در هنگام استفاده از استراتژی تطبیق پیشنهادی ما، با VCTK→SVC(شات صفر)، (شات صفر) بهتر از SVC هم در حوزه گفتار و هم در موارد دیگر است

مهمتر از همه، در حوزه صدای آواز مورد علاقه. در کل، بهترین روش اجرا برای آواز خواندن مبتنی بر صدا در این ارزیابی کمی با استفاده ازAutoVCآموزش دیده

با این حال، پیکربندی آموزشیSVC←VCTK(شات صفر) مدل رمز گذار ثابت

محاسباتی سبک‌تر و ثابت، به‌طور قابل‌توجهی به خوبی عمل می‌کند. شایان ذکر است که VCTK پیکربندی (یک داغ) برای ارزیابی روی قابل اجرا نیست مجموعه داده

SVC زیرا توانایی آنی برای سازگاری با صداهای جدید ندارد.

نتایج تجزیه و تحلیل کیفی ما، تبدیل آواز خواندن اجراهای صوتی با استفاده از صدای هدف از هر دو مجموعه تستSVC و VCTK در جداول ۳ و ۴ نشان داده شده است

به ترتیب. اول از همه، ما آن بلندگو را مشاهده می کنیم به طور کلی می توان از شبکه های تعبیه شده برایSVCشات صفر استفاده کرد. ما توجه داشته باشید که شبکه های

تبدیل آموزش دیده است گفتار را می توان در آواز خواندن استفاده کرد، اما آنها مقداری دارند مشکل حفظ پوشش های طیفی ثابت روی حروف صدادار طولانی. در نهایت، در حالی که به طور رسمی بخشی از ارزیابی موضوعی نیستیم، ما به طور غیررسمی عملکرد قابل مقایسه ای را بین معماری ها مشاهده می کنیم، با ترجیح نسبت به یک معماری بر دیگری بر اساس هر مورد.

با صداهای هدف ازVCTK هیچ چیز قابل توجهی وجود ندارد تفاوت بین شبکه های آموزش دیده با استفاده از یک بلندگوی داغ برچسب‌ها یا استفاده از جاسازی‌های بلندگوی صفر شات، اما دومی به طور طبیعی اجازه می‌دهد تا با صداهای جدید سازگار شود. در حالی کهSVC (شات صفر) برای سازگاری با ویژگی های آواز آموزش داده شده است.

صدا، با داده های کمتری آموزش دیده است و در معرض آن قرار گرفته است صداهای کمتر اگرچه به دلیل ماهیت شات صفرش توانست صداهایی شبیه صداهای هدف VCTK تولید کند و در مقایسه با روش‌های دیگر کار می‌کرد به طور قابل توجهی کمترین MOS را در این مورد دریافت کرد.

شبکه ها آموزش داده شده بر روی مجموعه دادهSVC هنگام استفاده از صداهای هدف از مجموعه تست SVC موفق تر هستند.

(و باز هم هستند سازگاری بهتر با مقیاس زمانی انتقال آوایی در آواز خواندن)

در این مورد، مقداری افت کیفیت وجود دارد برای سیستمی که با استفاده از پیکربندی VCTK (صفر شات) آموزش داده شده است، و پیکربندی VCTK (یک داغ) نیست

حتی قابل اجرا ما دوباره شاهد بهبودی برای شبکه هایی هستیم که با استفاده از VCTK→SVC (صفر شات) آموزش دیده اند در واقع SVC (شات صفر) در این سناریو VCTK→SVC پیکربندی آموزشی (صفر شات) از نظر کیفیت کلی برای SVC و VCTK از سایر روش ها بهتر عمل می کند.

صداهای هدف VCTK (شات صفر) و VCTK→SVCپیکربندی‌های آموزشی (شات صفر) بهترین عملکرد را دارند از نظر شباهت صدا برای صداهای هدف VCTKو SVC .به ترتیب در نهایت، ماهیت صفر شات خود را بیشتر مثال می زنیم روش پیشنهادی با قرار دادن سیستم ما در معرض صداهای هدف خارج از مجموعه داده هایSVC و VCTKاست.

این نمونه ها بدون هیچ گونه آموزش بیشتر مدل ها و با استفاده از فقط ۱-۲ ثانیه صدا از صدای هدف به ترتیب برای محاسبه تعبیه‌های بلندگو در حالی که کیفیت و صدا بدیهی است که شباهت می تواند با مدل بیشتر بهبود یابد مشخص است که داده های بیشتری را از صدای هدف تنظیم کنید که سیستم می تواند تبدیل های معقولی شبیه صداهای مواد مرجع را در به ثورت شات صفر ایجاد کند.

	Quality	Similarity
VCTK (one-hot)	2.377	2.828
VCTK (zero-shot)	2.447	**3.051**
SVC (zero-shot)	2.289	2.549
VCTK→SVC (zero-shot)	**2.476**	2.664

	Quality	Similarity
VCTK (one-hot)	N/A	N/A
VCTK (zero-shot)	2.154	2.610
SVC (zero-shot)	2.477	2.772
VCTK→SVC (zero-shot)	**2.674**	**2.937**

جدول ۳. میانگین نمرات نظر در مورد آواز خواندن با هدف

صداها از مجموعه تست VCTK با مدل رمزگذار ثابت

جدول ۴. میانگین نمرات نظر در مورد آواز خواندن با هدف

صداها از مجموعه تست SVC با مدل رمزگذار ثابت

۴. نتیجه گیری

در این مقاله، ما کاربرد شبکه‌های جاسازی بلندگو برای SVCصفر شات را پیشنهاد می‌کنیم. ما دو معماری را برای اجرای SVC صفر شات با استفاده از WORLD پیشنهاد می کنیم.

Vocoder برای مدل سازی صدای آواز. به طور کلی، ما آن را پیدا می کنیم

تعبیه‌های بلندگو در واقع می‌توانند مستقیماً برای zeroshot SVC استفاده شوند. علاوه بر این، شبکه‌های شات صفر که برچسب‌های بلندگوی تک داغ را با جاسازی‌های بلندگو جایگزین می‌کنند، و همچنین (یا حتی بهتر از) مجموعه بسته تحت نظارت آنها

همتایان، با مزایای بسیار ارزشمندی که آنها دارند می تواند بر روی داده های بدون برچسب آموزش داده شود و به طور بالقوه می تواند سازگار شود به صداهای جدید بدون نیاز به آموزش بیشتر. علاوه بر این، ما نشان می‌دهیم که آموزش شبکه‌های

zeroshot SVC. با تطبیق یک مدل اولیه آموزش‌دیده، مزایایی دارد.

حجم زیادی از داده های گفتاری در کار آینده، ما عوامل نهفته یادگیری را بررسی خواهیم کرد که می‌توانند بیشتر اجازه دهند دستکاری بیانی نتایج تبدیل در حالی که برخی پیشرفت اولیه برای این منظور با استفاده از گاوسی انجام شده است.

مخلوط VAE (GMVAEs) [۱۱] تا حد زیادی بوده است.

محدود به مصوت های خوانده شده ما احتمالاً می توانیم این را تعمیم دهیم

صدای آواز خواندن عملی تر با استفاده از شرطی سازی سیگنال های مورد استفاده در

این کار ما همچنین علاقه مند به جایگزینیVocoder با Vocoder WORLD

های آموخته شده بر اساس هستیم با پردازش سیگنال دیجیتال متمایز همانطور که

در [۱۸، ۲۵]، به منظور فعال کردن تمرینات سبک وزن از پایان به انتها است.

۵. قدردانی

نویسنده مایل است از فرانسوا ژرمن و همه تشکر کند بازبینان ناشناس برای نظرات

ارزشمندشان در حین تهیه این مقاله که به طرز چشمگیری کیفیت این کار بهبود

یافت.

6. REFERENCES منابع

[1] K. Lent, “An efficient method for pitch shifting digitally sampled sounds,” Computer Music Journal,vol. 13, no. 4, pp. 65–71, 1989.

[2] S. Nercessian, “Improved zero-shot voice conversion using explicit conditioning signals,” in Proc. of Interspeech 2020, 2020, accepted.

[3] W. Zhao, W. Wang, Y. Sun, and T. Tang, “Singing voice conversion based on WD-GAN algorithm,” in Proc. Of the 2019 IEEE 4th Advanced Information Tech., Electronic and Automation Control Conference (IAEAC), 2019, pp. 950–954.

[4] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” in Proc. of Interspeech 2017,2017.

[5] A. van den Oord et al., “WaveNet: A generative model for raw audio,” arXiv:1609.03499, 2016.

[6] N. Kalchbrenner et al., “Efficient neural audio synthesis,” arXiv:1802.08435, 2018.

[7] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Transactions on In formation and Systems, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[8] E. Nachmani and L. Wolf, “Unsupervised singing voice conversion,” in Proc. of Interspeech 2019, 2019, pp. 2583–2587.

[9] B. Sisman, K. Vijayan, M. Dong, and H. Li, “SIN GAN: Singing voice conversion with generative adversarial networks,” in Proc. of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2019, pp. 112–118.

[10] P. Chandna, M. Blaauw, J. Bonada, and E. Gomez, “WGANSing: A multi-voice singing voice synthesizer based on the wasserstein-gan,” in Proc. of the 27th European Signal Processing Conference, 2019.

[11] Y. Luo, C. Hsu, K. Agres, and D. Herremans, “Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[12] S. O. Arik et al., “Deep voice 2: Multispeaker neural text-to-speech,” Advances in Neural Information Processing Systems, vol. 30, pp. 2962–2970, 2017.

[13] M. Blaauw, J. Bonada, and R. Daido, “Data efficient voice cloning for neural singing synthesis,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[14] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 4879—4883.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “AutoVC: Zero-shot voice style transfer with only autoencoder loss,” in Proc. of the International Conference on Machine Learning, 2019.

[16] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel-generalized cepstral analysis - a unified approach to speech spectral estimation,” in Proc. of the International Conference on Spoken Language Processing, 1994.

[17] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in Proc. of the IEEE International Conference on Multimedia and Expo, 2016, pp. 1–6.

[18] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, “DDSP: Differentiable digital signal processing,” in Proc. Of the International Conference on Learning Representations, 2020, pp. 26–30.

[19] J. H. Lee, H. S. Choi, and K. Lee, “Audio query-based music source separation,” in Proc. of the International Society for Music Information Retrieval Conference, 2019.

[20] J. S. Garapolo et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Philadelphia: Linguistic Data Consortium, 1993.

[21] L. Hantrakul, J. Engel, A. Roberts, and C. Gu, “Fast and flexible neural audio synthesis,” in Proc. of the International Society for Music Information Retrieval Conference, 2019.

[22] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, “Crosslingual voice conversion with bilingual phonetic posteriorgram and average modeling,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6790—6794.

[23] T. Hasan and J. H. L. Hansen, “A study on universal background model training in speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 1890–1899, 2011.

[24] C. Veaux, J. Yamagishi, and K. MacDonald, CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. Edinburgh: The Centre for Speech Technology Research (CSTR), University of Edinburgh, 2016.

[25] X. Wang, S. Takaki, and J. Yamagishi, “Neural source filter-based waveform model for statistical parametric speech synthesis,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 5916–5920