بسم الله الرحمان الرحیم

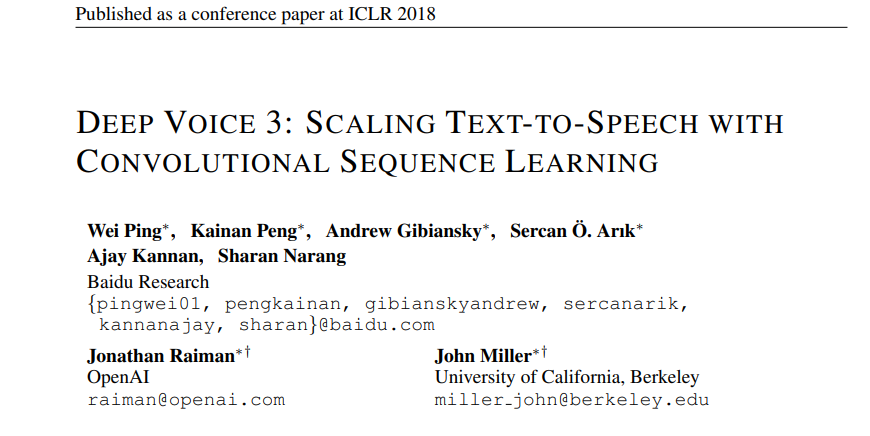
نام و نام خانوادگی : مژگان دهقان آزاد

نام استاد : جناب دکتر مهدی اسلامی

موضوع : صفر شات و مقالات مشابه آن به صورت چکیده

DSP نام درس:

شماره دانشجویی : 40014140111066



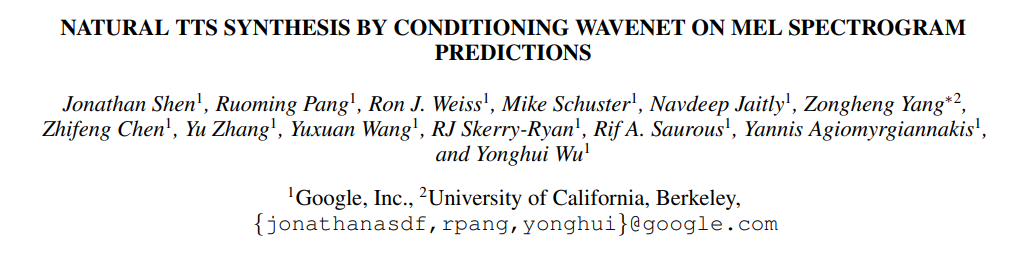
ما Deep Voice 3 را ارائه می‌دهیم، یک سیستم متن به گفتار (TTS) مبتنی بر توجه عصبی کاملاً پیچیده Deep Voice 3 با گفتار عصبی پیشرفته مطابقت دارد

سیستم های سنتز در طبیعی بودن در حالی که تمرین یک مرتبه قدر سریع تر است.

ما Deep Voice 3 را به اندازه‌های مجموعه داده‌ای که برای TTS بی‌سابقه است، تغییر می‌دهیم

که در بیش از هشتصد ساعت صدا از بیش از دو هزار بلندگو است

علاوه بر این، ما حالت‌های خطای رایج شبکه‌های ترکیبی گفتار مبتنی بر توجه را شناسایی می‌کنیم، نحوه کاهش آنها را نشان می‌دهیم و چندین روش مختلف سنتز شکل موج را با هم مقایسه می‌کنیم. ما همچنین نحوه مقیاس بندی استنتاج را توضیح می دهیم ده میلیون درخواست در روز بر روی یک سرور GPU واحد است.



این مقاله تاکوترون 2 را، یک معماری شبکه عصبی برای سنتز گفتار به طور مستقیم از متن این سیستم از یک تشکیل شده است

شبکه پیش‌بینی ویژگی توالی به دنباله تکراری که نقشه‌برداری می‌کند

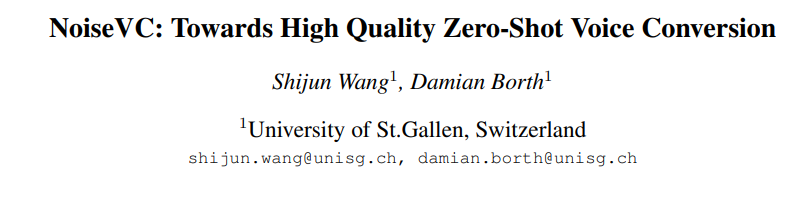
تعبیه کاراکترها در طیف‌نگاری‌های مقیاس mel، به دنبال آن یک مدل WaveNet اصلاح‌شده که به‌عنوان یک کد صوتی برای سنتز دامنه زمانی عمل می‌کند. شکل موج از آن طیف نگارها ی مدل ما به یک میانگین دست می یابد

امتیاز نظر (MOS) 4.53 قابل مقایسه با MOS 4.58 برای گفتار ضبط شده حرفه ای. برای تأیید انتخاب های طراحی خود، ما ارائه می دهیم

مطالعات فرسایشی اجزای کلیدی سیستم ما و ارزیابی تأثیر استفاده از طیف‌ نگارهای mel به عنوان ورودی شرطی‌سازی WaveNet

به جای ویژگی های زبانی، مدت زمان و F0 ما بیشتر نشان می دهیم که استفاده از این نمایش میانی آکوستیک فشرده اجازه می دهد تا

کاهش قابل توجه در اندازه معماری WaveNet داشته باشد



تبدیل صدا (VC) وظیفه ای است که صدا را از صدای هدف به منبع تبدیل می کند بدون اینکه محتوای زبانی از بین برود، به ویژه زمانی که گویندگان منبع و هدف دیده نمی شوند چالش برانگیز است.

در حین تمرین (صفر شات VC).

مدل یا داده های زبانی از پیش آموزش دیده برای انجام تبدیل شات صفر. در همین حال، مدل‌های VC با کوانتیزاسیون برداری (VQ)

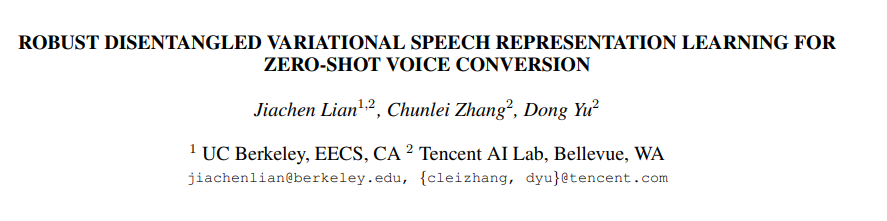
یا Instance Normalization (IN) می‌توانند محتویات را از فایل‌های صوتی جدا کرده و به تبدیل‌های موفقی دست یابند. با این حال،

گره گشایی در این مدل ها به شدت به لایه های تنگنای بسیار محدود متکی است، بنابراین، کیفیت صدا به شدت افزایش می یابد.

قربانی کرد. در این مقاله، ما NoiseVC را پیشنهاد می کنیم،

می تواند محتویات را بر اساس VQ و Contrastive Predictive Coding (CPC) تفکیک کند. علاوه بر این، افزایش نویز برای افزایش بیشتر قابلیت جداسازی انجام می شود. ما چندین آزمایش انجام می دهیم و نشان می دهیم که NoiseVC دارای یک است

توانایی گسستن قوی با یک قربانی کوچک کیفیت



مطالعات سنتی در زمینه تبدیل صدا (VC) پیشرفت کرده است

با داده های آموزشی موازی و بلندگوهای شناخته شده. کیفیت خوب نسخه صوتی با کاوش در ماژول های تراز بهتر به دست می آید

یا توابع نگاشت بیانی. در این مطالعه، ما VC صفر شات را از دیدگاه جدیدی از گسستگی خود نظارتی بررسی می‌کنیم.

یادگیری بازنمایی گفتار به طور خاص، ما با ایجاد تعادل در جریان اطلاعات بین جهانی، به گسستگی دست می‌یابیم

رمزگذار خودکار متغیر متوالی (VAE). یک نسخه صوتی صفر شات با تغذیه یک اسپیکر دلخواه و تعبیه شده انجام می شود

جاسازی محتوا در رمزگشا VAE است

استراتژی آموزشی تقویت داده ها برای ایجاد آموخته ها اعمال می شود

نویز نمایشی ثابت در مجموعه داده های TIMIT و VCTK، ما دستیابی به عملکرد پیشرفته در هر دو ارزیابی عینی،

به عنوان مثال، تأیید بلندگو (SV) در تعبیه و محتوای بلندگو تعبیه و ارزیابی ذهنی، یعنی طبیعی بودن صدا و

شباهت، و حتی با منبع/هدف پر سر و صدا گفته ها همچنان قوی است



تفکیک محتوا و اطلاعات سبک صحبت کردن برای تبدیل صدای غیر موازی شات صفر (VC) ضروری است. مطالعه قبلی ما یک چارچوب جدید با تفکیک شده را بررسی کرد

رمزگذار خودکار متغیر متوالی (DSVAE) به عنوان ستون فقرات

برای تجزیه اطلاعات ما نشان داده‌ایم که جاسازی محتوای جداسازی همزمان و جاسازی بلندگو از یک گفته برای VC صفر شات امکان‌پذیر است. در این مطالعه، ما مسیر را با ایجاد یک نگرانی در مورد ادامه می دهیم

توزیع قبلی شاخه محتوا در خط پایه DSVAE. ما متوجه شدیم که توزیع اولیه تصادفی قبلی باعث می شود

جاسازی محتوا برای کاهش اطلاعات ساختار آوایی در طول فرآیند یادگیری، که یک ویژگی مطلوب نیست. در اینجا، ما به دنبال دستیابی به یک جاسازی محتوای بهتر با آن هستیم اطلاعات آوایی بیشتر حفظ شده است. ما مشروط را پیشنهاد می کنیم

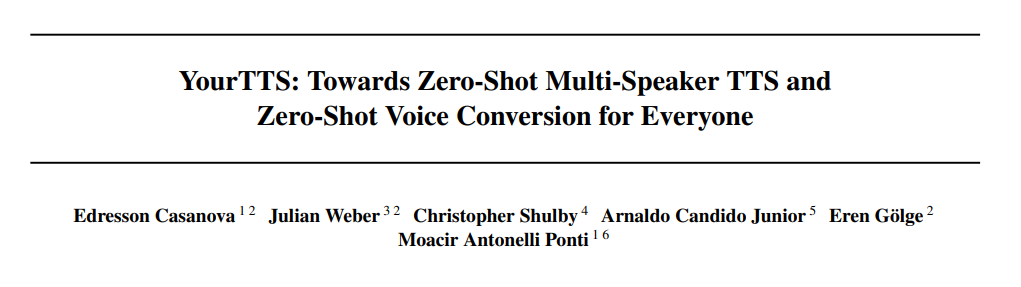
DSVAE، مدل جدیدی که سوگیری محتوا را به عنوان یک شرط فعال می‌کند

به مدل‌سازی قبلی و تغییر شکل محتوای تعبیه‌شده نمونه‌برداری شده از توزیع پسین. در آزمایش ما روی ایمن مجموعه که مجموعه آن

مجموعه داده VCTK، ما نشان می‌دهیم که جاسازی‌های محتوا به دست آمده از DSVAE شرطی بر تصادفی بودن غلبه می‌کند.

و دستیابی به دقت طبقه بندی واج بسیار بهتر، صدای تثبیت شده و عملکرد بهتر VC صفر شات در مقایسه با خط پایه رقابتی DSVAE

است



YourTTS قدرت یک رویکرد چند زبانه را برای کار TTS چند بلندگوی صفر شات به ارمغان می آورد.

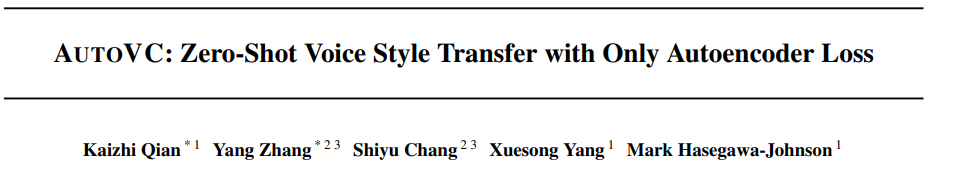
روش ما بر اساس مدل VITS است و اضافه می کند چندین تغییر جدید برای آموزش چند بلندگو و چند زبانه صفر شات. به دست آوردیم

پیشرفته ترین (SOTA) منجر به TTS چند بلندگوی صفر شات و نتایج قابل مقایسه با SOTA در تبدیل صدای صفر شات در مجموعه داده VCTK است.

علاوه بر این، رویکرد ما امیدوارکننده است منجر به یک زبان مقصد با تک سخنران می شود

مجموعه داده، امکان باز کردن امکانات برای TTS چند بلندگوی صفر شات و سیستم های تبدیل صدای صفر شات در زبان های کم منبع. در نهایت، می توان مدل YourTTS را با کمتری تنظیم کرد

بیش از 1 دقیقه سخنرانی و دستیابی به نتایج عالی در شباهت صدا و با منطقی کیفیت این مهم برای اجازه سنتز برای بلندگوهایی با صدای یا ضبط بسیار متفاوت ویژگی هایی که در طول تمرین مشاهده می شود



تبدیل صدای چند به چند غیر موازی، به عنوان:

و همچنین تبدیل صدای صفر شات، مناطق ناشناخته باقی می مانند. الگوریتم های انتقال سبک عمیق، مانند شبکه های متخاصم مولد (GAN) و رمزگذار خودکار متغیر شرطی (CVAE)، به عنوان راه حل های جدید در این زمینه استفاده می شود.

با این حال، آموزش GAN پیچیده و دشوار است و هیچ مدرک قوی دال بر اینکه گفتار تولید شده آن کیفیت ادراکی خوبی دارد وجود ندارد. بر از سوی دیگر، آموزش CVAE ساده است اما انجام می دهد با ویژگی تطبیق توزیع همراه نیست یک GAN. در این مقاله سبک جدیدی را پیشنهاد می کنیم

طرح انتقال که فقط شامل رمزگذار خودکار است با یک گلوگاه به دقت طراحی شده است. ما رسما نشان می دهد که این طرح می تواند انتقال سبک تطبیق توزیع را تنها با آموزش در مورد از دست دادن خود بازسازی به دست آورد. بر اساس این طرح، ما AUTOVC پیشنهادی، که به نتایج پیشرفته ای در تبدیل صدای چند به چند با داده های غیر موازی، و اولین موردی است که انجام می دهد تبدیل صدای صفر شات است