Vocoderعصبی چند بلندگو

اوریول باربانی 2،1، آنتونیو بونافونته 1 و سانتیاگو پاسکوال

Universitat Politècnica de Catalunya, Barcelona, Spain
2École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

چکیده :
سنتز گفتار پارامتریک آماری (SPSS) نسبت به سنتز گفتار مبتنی بر انتخاب واحد، که فناوری تجاری غالب در دهه 2000 بود، انعطاف بیشتری را ارائه می دهد. اگرچه انعطاف پذیری بیشتری را ارائه می دهد و به کل پایگاه داده peech در استقرار نیاز ندارد، سیستم های کلاسیک طبیعی بودن کمتری نسبت به روش های انتخاب واحد دارند. یادگیری عمیق به لطف معماری های تکرارشونده با ارائه گفتار با کیفیت بالا و در عین حال حفظ انعطاف پذیری مطلوب در انتخاب پارامترهایی مانند بلندگو، لحن و غیره، بهتر از SPSS عمل کرده است. این مقاله دو پیشنهاد را برای بهبود سیستم‌های متن به گفتار مبتنی بر یادگیری عمیق ارائه می‌کند. مدل پایه که با تطبیق SampleRNN به دست آمد، توانست با یک مدل واحد صدا را از بلندگوهای مختلف تولید کند و پس از اجرای ترکیبی از دو رویکردی که در این مقاله مورد بحث قرار خواهد گرفت، نتایج پیشرفته‌ای به دست آمد. پیشنهاد اول با نرمال‌سازی ویژگی‌های معمولی که منشأ چنین ویژگی‌هایی را در نظر نمی‌گیرند متفاوت است، که می‌تواند تفاوت‌های ذاتی مانند مدل‌سازی بلندگوهای مختلف با یک شبکه داشته باشد. هدف از این کار دستیابی به ویژگی های صوتی با مقادیر مشابه در بین بلندگوها است، به عنوان مثال:
مردها صدای کمتری نسبت به زن دارند، اما تنوع آن در یک گفتار معین، یعنی لحن، در هر دو مورد چندان متفاوت نیست .به لطف این، شبکه می تواند به راحتی الگوهای این ویژگی ها را در همه بلندگوها با یک مدل واحد مدل کند. پیشنهاد دوم، به نام نگاه به جلو، شامل تغذیه اطلاعات فریم های آینده به شبکه با هدف مدل سازی بهتر سیگنال گفتار و جلوگیری از ناپیوستگی های احتمالی است.
شنوندگان انسانی سیستمی را ترجیح می دهند که هر دو تکنیک را ترکیب می کند، که در مقیاس میانگین امتیاز نظر (MOS) با مجموعه داده متعادل به نرخ 4 می رسد و از سایر مدل ها بهتر عمل می کند.
اصطلاحات فهرست: یادگیری عمیق، سنتز گفتار، شبکه‌های عصبی مکرر، تبدیل متن به گفتار، SampleRNN، سری زمانی

1-مقدمه :

یادگیری عمیق تقریباً در تمام شاخه های مهندسی در دهه های گذشته متحول شده است و همچنین با موفقیت برای تبدیل متن به گفتار (TTS) به کار گرفته شده است، جایی که عملکرد پیشرفته ای را ارائه می دهد و بر رویکردهای کلاسیک غلبه می کند.
مسئله سری زمانی کاملاً توسط شبکه‌های عصبی بازگشتی (RNN) و انواع آن‌ها مورد استفاده قرار گرفته‌اند که باعث می‌شود آنها به نتایج جالبی در زمینه سنتز گفتار منجر شوند. علاوه بر این، مدل‌های مولد عمیق می‌توانند نمونه گفتار را با نمونه تولید کنند همانطور که برای اولین بار در [1] Wavenet پیشنهاد شد، که دامنه‌های شکل موج بسیار ریز را به دست آورد و از مدل‌های قبلی سنتز پارامتریک گفتار آماری آماری (SPSS) بهتر عمل کرد. این مقاله دو مورد از پیشنهادات ارائه شده در پایان نامه کارشناسی نویسنده اصلی [2] را نشان می دهد که برای مدل سازی بهتر گفتار تولید شده با یک مدل مبتنی بر یادگیری عمیق چند گوینده به کار گرفته شد. یادگیری عمیق با ارائه گفتار با کیفیت بالا و در عین حال حفظ انعطاف پذیری سیستم هایSPSS ، از سیستم های کلاسیک بهتر عمل کرده است. برای دستیابی به یک سیستم پیشرفتهTTS ، [3] Samplernnبرای تولید گفتار منسجم به زبان اسپانیایی قابل انتساب به سخنرانان مختلف اقتباس شد. انگیزه اصلی این کار تعمیم سیستم ارائه شده در [4] برای بسیاری از سخنرانان به عنوان یک ساختار شبکه عصبی عمیق مشترک (DNN) بود، زیرا به نتایج بهتری در تولید گفتار با کیفیت نسبت به یادگیری پارامترهای یک بلندگوی مجزا دست می‌یابد.[5] . در این مورد، یادگیری ساختار مشترک پایه را می توان به یک بلندگوی جدید انتقال داد تا به سازگاری بلندگو مانند [6] با داده های آموزشی محدود دست یابد و نتایج خوبی در طبیعی بودن و شباهت به بلندگوی اصلی به دست آورد. این پیشنهادها در ابتدا برای بهبود گفتار به‌دست‌آمده با یک شبکه مولد عمیق که قادر به مدل‌سازی چندین سخنران با ساختار یکسان بود، طراحی شد. با این وجود، نرمال سازی وابسته به بلندگو می تواند به عنوان یک تکنیک پیش پردازش جدید در مسائل مختلف مورد استفاده قرار گیرد و رویکرد نگاه به جلو را می توان به مدل سازی سری های زمانی تعمیم داد. مدل‌های پیشرفته TTS فعلی مانند[1] WaveNet ، [7] Tacotronیا [8] VQ-VAE قبلاً چندین بلندگو را با یک مدل منحصر به فرد مدل می‌کنند، اما نرمال‌سازی‌های وابسته به بلندگو را اعمال نمی‌کنند، که نشان داده شده است که نتایج را بدتر می کند. ولین پیشنهاد انجام تبدیل صدا بود، که تکنیکی برای تغییر شکل موج گفتار است که آزادانه اطلاعات غیر فرازبانی را در حالی که اطلاعات زبانی را حفظ می کند، تبدیل می کند .[9] این بدان معنی است که لحن، مکث و متن گفتاری دقیقاً یکسان است اما گوینده تغییر می کند. این افزونگی توسط شبکه شناسایی می‌شود که وزن‌های کمی را به هویت گوینده اختصاص می‌دهد و آن را بی‌ربط می‌کند. با این وجود، این ورودی برای تبدیل صدا برای انتخاب بلندگوی مورد نظر مورد نیاز است. هدف نرمال‌سازی وابسته به بلندگو این است که با جدا کردن ویژگی‌ها از بلندگو به هویت گوینده اهمیت دهد و در نتیجه این شبکه مجبور می شود از هویت گوینده برای تولید گفتار طبیعی

برای هر کاربر استفاده کند. رویکرد نگاه به جلو، علیت مدل‌سازی سری‌های زمانی را زیر سوال می‌برد، که مورد نیاز نیست مگر اینکه ویژگی‌های ورودی در زمان واقعی استخراج شوند و بنابراین از قبل شناخته نشده باشند. در مورد سیستم های TTS، متنی که گفته می شود از قبل مشخص است و بنابراین، ویژگی های صوتی تمام سیگنال ها مشخص است. علاوه بر این، در گفتار طبیعی، واج ها بسته به زمینه متفاوت به نظر می رسند و بنابراین بسته به واج های آینده می توانند تغییر کنند (هم مفصلی). با دادن اطلاعات رفتار آینده دنباله پیش بینی شده، هیچ ناپیوستگی وجود ندارد و مصنوعات کاهش می یابد. این به گفتار با کیفیت بهتری که توسط شنوندگان رتبه‌بندی می‌شود ترجمه می‌شود. این دو پیشنهاد در چهار پیکربندی مختلف به دست می‌آیند که در این کار بیشتر مورد بررسی و مقایسه قرار خواهند گرفت.

شکل 1

همانطور که توسط کاربران رتبه بندی شده است، نرمال سازی وابسته به بلندگو به نتایج قابل ملاحظه ای بهتر در طبیعی بودن برای تولید گفتار مدل سازی شده با مجموعه داده های متعادل در صورت ترکیب با رویکرد نگاه به جلو دست می یابد. در مورد پروپوزال دوم، عملکرد بهتری نسبت به نمرات کسب شده قبلی دارد و هنگامی که با نرمال سازی وابسته به بلندگو ترکیب می شود، به نتایج پیشرفته می رسد.

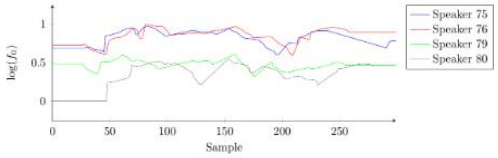
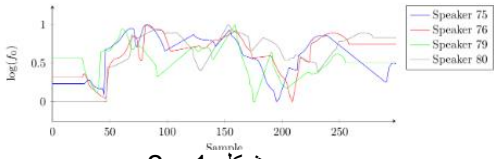


Figure 1: *Classical speaker-independent normalization*



شکل 1 و 2

2-پروپوزال ها :

2.1 عادی سازی ویژگی های وابسته به بلندگو

ویژگی‌هایی که به شبکه عصبی تغذیه می‌شوند، اغلب قبلاً برای کنترل میزان فعال‌سازی‌ها و گرادیان‌ها در تمرین نرمال‌سازی می‌شوند. با فرضیه داشتن ویژگی های وابسته به گوینده، یک نرمال سازی مستقل برای هر یک از گوینده ها برای جداسازی ویژگی های گفتار از منبع ارائه شد. مقادیر حداکثر و حداقل برای هر یک از پارامترها در پارتیشن آموزشی یافت شد، بنابراین ممکن است برخی از ویژگی‌های پارتیشن‌های قطار یا اعتبارسنجی از مرزها فراتر روند.

این رویکرد را می توان با سایر توابع نرمال سازی مانندz-score ، یعنی نرمال سازی آماری نیز به کار برد .این آخرین گزینه قبل از نوشتن این مقاله به دلیل نتایج کم بهبود این اصلاح تنها آزمایش نشده بود (جدول 1 را ببینید).

با این وجود، همانطور که در همان جدول مشاهده می شود، این رویکرد در صورت ترکیب با رویکرد نگاه به جلو (توضیح داده شده در بخش 2.2) از سایر مدل ها بهتر عمل می کند .بنابراین، یک نرمال سازی آماری نیز می تواند در کار آینده آزمایش شود.

هدف این پیشنهاد اهمیت دادن به هویت گوینده است تا امکان تبدیل صدا بدون نیاز به نقشه‌برداری پیچیده از ویژگی‌ها را فراهم کند. الهام از رفتار گام برای هر بلندگو به دست آمد، که برای نرمال سازی مستقل از بلندگو و وابسته به بلندگو به ترتیب در شکل های 1 و 2 نشان داده شده است. این نمودارها تکامل فرکانس اصلی لگاریتمی را برای چهار گوینده مختلف شامل دو مرد و دو زن که متن قبلی را می‌خوانند و بنابراین پس از نرمال شدن به دنبال رویکرد وابسته به گوینده، بسیار شبیه هستند، نشان می‌دهد.

توجه داشته باشید که به دلیل مدت زمان متفاوت واج ها و مکث ها مقداری جابجایی زمانی وجود دارد، اما سیگنال هنوز بسیار شبیه است. پس از عادی سازی مستقل از بلندگوی کلاسیک (شکل 1)، تشخیص زن (75، 76) و مرد (79، 80) بسیار آسان است.

این بدان معنی است که انجام تبدیل صدا غیرممکن است زیرا شبکه به شناسه بلندگو نیاز ندارد تا این اطلاعات ضمنی در ویژگی‌ها باشد. به همین دلیل است که این افزونگی به بیهودگی این ورودی که هنگام تلاش برای تغییر هویت گوینده به میل مشاهده می شود، تبدیل می شود. اگر لحن قابل مقایسه باشد، رفتار گام پس از نرمال شدن توسط گوینده بسیار مشابه است. با این وجود، سایر ویژگی‌هایی که به شبکه داده می‌شوند (به بخش بعدی مراجعه کنید) منجر به عادی‌سازی‌های بسیار مشابه برای هر دو رویکرد مستقل از بلندگو و وابسته به بلندگو شدند.

2.2 نگاه به جلو

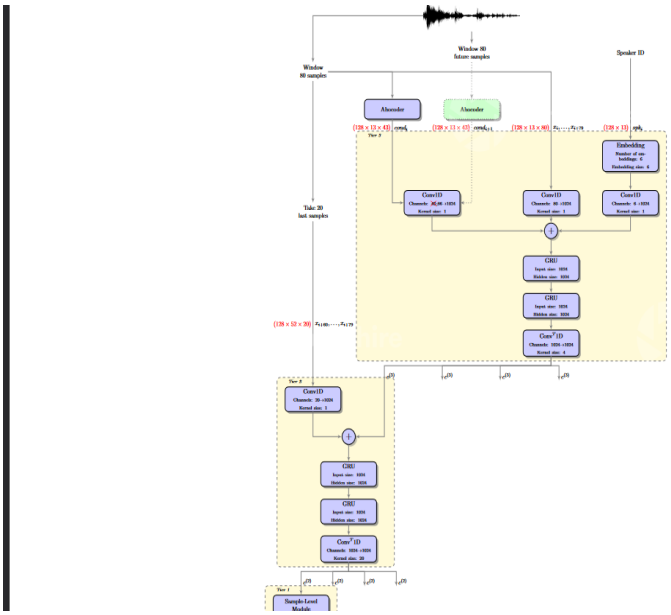
در مدل‌سازی توالی‌های غیرواقعی مانند تولید گفتار در یک سیستمTTS ، ویژگی‌هایی که به شبکه داده می‌شود، از قبل مشخص است. این بدان معناست که برخلاف یک تماس تلفنی احتمالی که در آن هر دو طرف در زمان واقعی صحبت می‌کنند، ویژگی‌هایی که توالی را در گام‌های زمانی آینده مشخص می‌کنند همیشه شناخته شده‌اند و بنابراین می‌توان برای مدل‌سازی بهتر سیگنال تولید شده استفاده کرد. نتایج در یک مدل بزرگ‌تر است زیرا تعداد ویژگی‌ها در هر مرحله تکرار می‌شود، اما کیفیت بهتری را بدون نیاز به ویژگی‌های بیشتر به دست می‌آورد.

3-راه اندازی آزمایشی :

مجموعه داده گفتاری مورد استفاده برای آموزش مدل توسط شش صدای اسپانیایی از پروژه [10] TC-STAR تشکیل شد که نیمی از آنها مرد و نیمی دیگر زن هستند. پایگاه داده نامتعادل بود و یکی از سخنرانان زن به سختی یک چهارم زمان ضبط گفتار را در مقایسه با دیگران داشت. علیرغم اینکه در برخی کارها مانند [11]، توصیه می شود داده ها را به ازای هر کاربر متعادل کنید تا همه آنها تقریباً به همان میزان نمونه برای آموزش داشته باشند. برای جلوگیری از محدود کردن تمام سخنرانان به جای یک ساعت، تنها به 14 دقیقه سخنرانی استفاده شد مدت زمان کل مجموعه داده شامل شش سخنران 5.25 ساعت بود که به 80% برای آموزش، 10% برای اعتبار سنجی و 10% برای تست تقسیم شد.

3.1 نمونه rnn

مدل پایه SampleRNN بود، یک مدل تولید صوت عصبی بدون قید و شرط [3] که شامل دو ماژول تکراری است که با نرخ‌های ساعت مختلف اجرا می‌شوند و هدف آن مدل‌سازی وابستگی‌های کوتاه‌مدت و بلندمدت سیگنال‌های گفتاری، و یک ماژول با اتورگرسیو است .پرسپترون های چند لایه (MLPs) که نمونه به نمونه گفتار را پردازش می کنند.
معماری تکراری مورد استفاده برای این مدل واحد بازگشتی در دار [12] (GRU) است که با پیاده‌سازی ماژول‌های حافظه کوتاه‌مدت بلند مدت (LSTM) پیشنهاد شده توسط نویسندگان SampleRNN متفاوت است. معماری سه لایه انعطاف‌پذیری را در تخصیص مقدار منابع محاسباتی برای مدل‌سازی سطوح مختلف انتزاع فراهم می‌کند و نتایج بسیار کارآمدی در حافظه در طول آموزش دارد. خروجی نهایی مدلSampleRNN ، احتمال همان مقدار نمونه فعلی مشروط به تمام مقادیر قبلی توالی است که می‌تواند طبق قانون زنجیره بیان شده در رابطه (2) بیان شود.



شکل 3

خروجی از یک توزیع Multinoulli پیروی می کند که به دلیل طبیعی بودن سیگنال های گفتاری که دارای ارزش واقعی هستند می تواند غیر شهودی باشد، اما به نتایج بهتری دست می یابد زیرا هیچ توزیعی از داده ها را فرض نمی کند و بنابراین می تواند به راحتی توزیع های دلخواه را مدل کند. در این کار، نمونه‌های گفتار با 8 بیت کوانتیزه می‌شوند، بنابراین دارای 256 مقدار ممکن است. به منظور تولید گفتار منسجم، مدل مانند [4] با ویژگی‌های آکوستیک به‌دست‌آمده با[14] Ahocoder ، یک کد صوتی با کیفیت بالا

هارمونیک و نویزی که مجموعه‌ای از ویژگی‌ها را پیش‌بینی می‌کند که می‌تواند سیگنال‌های گفتاری را مشخص کند، شرطی شد. مدل اقتباس شده در شکل 3 نشان داده شده است. توجه داشته باشید که رویکرد نگاه به جلو، معماری را تغییر می‌دهد، زیرا بلوک D-1 Convolutionسمت راست، اندازه ورودی آن را دو برابر می‌کند (مقدار اصلی 43 در شکل خط زده شده و با 86 جایگزین شده است تا هر دو ویژگی قاب فعلی و آینده را بپذیرد. متفاوت از مدل اصلی[3] SampleRNN است و به غیر از افزودن قبلاً ذکر شده از تهویه‌کننده‌های صوتی که امکان سنتز گفتار منسجم را فراهم می‌کنند، نویسندگان بلوک‌ها را در سمت چپ شکل نیز ترکیب کرده‌اند. هدف از اینها تمایز بین همه بلندگوهای پایگاه داده و در نتیجه محاسبه تعبیه از شناسه است که همچنین برای شرطی کردن مدل در امتداد ویژگی های فوق الذکر استخراج شده با Ahocoder استفاده می شود.

پارامترهای آکوستیک در فریم‌های 15 میلی‌ثانیه با جابه‌جایی هر 5 میلی‌ثانیه استخراج می‌شوند و عبارتند از:

40 ضرایب Mel-cepstral

- حداکثر فرکانس صداFV
- مقدار لگاریتمی F0
- پرچم صدادار/بی صدا uv

سیگنال کانتور گام برای سیگنال های صوتی و بدون صدا بسیار متفاوت عمل می کند.

در حالت اول یک سیگنال پیوسته است و در حالت دوم وجود ندارد و Ahocoder آن را به صورت -1010 نشان می دهد. برای مقابله با این ناپیوستگی در آماری که می‌تواند مدل را توسط این نقاط پرت تنزل دهد، خروجی Ahocoder پس پردازش می‌شود و کانتور گام به‌صورت لاگ خطی برای بخش‌های بدون صدا درون‌یابی می‌شودسپس این ویژگی‌ها به دنبال عادی‌سازی وابسته به بلندگوی پیشنهادی یا عادی‌سازی مستقل از بلندگوی کلاسیک مقیاس‌بندی می‌شوند. سپس ویژگی های نرمال شده برای مطابقت با نمونه های گفتاری مورد استفاده در آموزش مرتب می شوند و یک شناسه بلندگو به عنوان ورودی مستقل (نگاه کنید به شکل 3) به سیستم اضافه می شود..
هدف آن مدل سازی بهتر هر یک از صداهای مختلف و انجام تبدیل صدا است. مدل شرطی توزیعی را خروجی می دهد که نه تنها به نمونه های قبلی بستگی دارد، بلکه به ویژگی های به دست آمده با AHocoder و به هویت بلندگو نیز بستگی دارد. بنابراین، بیان مدل تطبیق‌شده از رابطه (3) پیروی می‌کند، که در آن lt مخفف یک بردار 49 بعدی است که نتیجه یک جاسازی با بعد 6 است که هویت بلندگو و بردار صوتی 43 بعدی مربوط به تحلیل را نشان می‌دهد. توجه داشته باشید که در صورت پیروی از رویکرد نگاه به جلو، به دلیل دو برابر شدن اندازه بردار صوتی، یک بردار 92 بعدی خواهد بود. استراتژی یادگیری آموزش هر یک از مدل‌های مشتق‌شده از پیشنهادات قبلی با نزول گرادیان تصادفی کوچک (SGD) با استفاده از اندازه کوچک دسته‌ای 128 و به حداقل رساندن Log-Likelihoodمنفی بود. بهینه ساز انتخاب شده برآورد لحظه تطبیقی[15] (ADAM) ، یک الگوریتم SGD با نرخ یادگیری تطبیقی، و مقدار اولیه 10-4 بود که با یک زمانبندی شناخته شده کنترل کننده نرخ خارجی افزایش یافت. این دو نقطه عطف در دوره‌های 15 و 35 حضور داشت. در هر یک از آن نقاط عطف، نرخ یادگیری فعلی با ضریب 0.1 کاهش می‌یابد که به تغییرات ناگهانی در منحنی ضرر که در دوره‌های اول نشان داده شده بود، حمله می‌کند. نرمال سازی وزن [16] نیز در لایه های کانولوشن 1 بعدی برای افزایش سرعت تمرین استفاده شد.

3.2 ارزیابی ذهنی

یک آزمون میانگین امتیاز نظر (MOS) برای مقایسه عمیق‌تر بین چهار آزمایش به دست آمده از ترکیب دو پیشنهاد انجام شد . MOS طبیعی بودن را در مقیاسی از اعداد صحیح طبیعی از 1 تا 5 درجه بندی می کند، به این معنی که این مرزها به ترتیب بد و کیفیت عالی هستند. داوطلبانی که در آزمون شرکت کردند، می‌توانستند به تعداد دفعات مورد نیاز به ضبط‌های مختلف گوش دهند تا سیستم‌ها را با هم مقایسه کرده و به آنها امتیاز دهند. برای هر جمله، رونویسی صدا برای سهولت گوش دادن ارائه شد و فایل‌های صوتی هر یک از سیستم‌های مختلف که همان جمله را ترکیب می‌کردند، در کنار هم برای مقایسه قرار گرفتند.

4.نتایج

نتایج نشان داده شده در جدول 1 با میانگین ارزیابی 25 داوطلب برای هر سیستم به دست آمد. برخی از نظرات نوشته شده توسط داوطلبانی که در آزمون شرکت کردند، تفاوت کیفیت بین مردان و زنان را برجسته کرد. به همین دلیل است که جدول زیر جدایی بین

جنسیت ها انجام می دهد که نشان می دهد بهترین نتایج در واقع با صدای مردان به دست می آید. این امر به پایگاه داده گفتار نامتعادل نسبت داده شد که شامل یک سخنران زن با تنها 25 درصد ضبط در مقایسه با سایر سخنرانان است. ظاهراً این موضوع در مدل‌سازی صداهای زنانه که پر سر و صداتر بودند، تأثیر داشت. نمونه های متعلق به بهترین سیستم ترکیبی از هر دو پیشنهاد را می توان در GitHub نویسنده اصلی 1 شنید.

5.نتیجه گیری

عادی سازی وابسته به بلندگو برای اهداف تبدیل صدا کافی نبود، بنابراین معماری های پیچیده تری در [2] پیشنهاد شد. با این وجود، شنوندگان انسانی گفتار مدل‌سازی شده با نرمال‌سازی وابسته به سخنران را ترجیح می‌دهند و با توجه به شباهت ویژگی‌های نرمال‌شده برای هر سخنران، کمیت‌سازی بهتر را می‌توان برای برنامه‌های کدگذاری یا برای استقرار شبکه‌های عصبی با منابع محدود اعمال کرد. در حالی که به نظر نمی‌رسد نرمال‌سازی وابسته به بلندگو نتایج به‌دست‌آمده با مقیاس‌بندی ویژگی کلاسیک مستقل از بلندگو را بهبود بخشد، وقتی با رویکرد نگاه به جلو ترکیب شود، با مجموعه داده‌های متعادل مردانه به امتیاز 4 می‌رسد. به طور خلاصه، با ترکیب این دو پیشنهاد، یک امتیاز پیشرفته MOS برای یک سیستم سنتز گفتار چند گوینده به دست آمده است. هر دوی این رویکردها نوآوری‌هایی بودند که در این پایان‌نامه معرفی شدند و نتایج نشان می‌دهد که می‌توانند برای سایر سیستم‌های TTS و همچنین برای دسته‌ای از برنامه‌های کاربردی دیگر که شامل ویژگی‌هایی از منابع مختلف و مدل‌سازی توالی‌های بی‌درنگ هستند، سودمند باشند.

Normalization: Look ahead:	Spk-D No	Spk-Ind No	Spk-D Yes	Spk-Ind Yes
Female:	3.3	3.3	3.8	3.6
Male:	3.6	3.6	4.0	3.8
Total:	3.5	3.5	3.9	3.8

Table 1: *Table with subjective results comparing proposed methods*

جدول 1

6.مراجع

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 1–15, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[2] O. Barbany Mayor, "Multi-Speaker Neural Vocoder," Bachelor's thesis, Universitat Politècnica de Catalunya, 2018.

[3] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," ICLR, pp. 1–11, 2017. [Online]. Available: http://arxiv.org/abs/1612.07837

[4] A. Bonafonte, S. Pascual, and G. Dorca, "Spanish Statistical Parametric Speech Synthesis using a Neural Vocoder," InterSpeech, 2018. [Online]. Available: https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2417.pdf

[5] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for DNN-based TTS synthesis," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4475–4479, 2015.

[6] A. W. Black, H. Zen, and K. Tokuda, "Statistical Paramet-

ric Speech Synthesis," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 1229–1232, 2007.
[7] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," CoRR, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135
[8] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in NIPS, 2017. [Online]. Available: http://arxiv.org/abs/1711.00937
[9] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 08-12-Sept, pp. 1632–1636, 2016.

[10] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. V. D. Heuvel, H. Hain, X. S. Wang, and M. N. Garcia, "TC-STAR : Specifications of Language Resources and Evaluation for Speech Synthesis," Proceedings of the Language Resources and Evaluation Conference LREC06, pp. 311–314, 2006.
[11] S. Pascual de la Puente, "Deep learning applied to speech synthesis," Master's thesis, Universitat Politècnica de Catalunya, 2016.
[12] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," CoRR, 2014. [Online]. Available: http://arxiv.org/abs/1406.1078
[13] ITU-T. Recommendation G. 711, "Pulse Code Modulation (PCM) of voice frequencies," 1988.
[14] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus Noise Model based Vocoder for Statistical Parametric Speech Synthesis," IEEE Journal on Selected Topics in Signal Processing, vol. 8, no. 2, pp. 184–194, 2014.
[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980
[16] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," CoRR, vol. abs/1602.07868, 2016. [Online]. Available: http://arxiv.org/abs/1602.07868