خلاصه مقاله

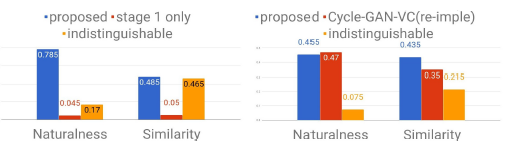
* این مقاله از طریق سایت arXiv preprint arXiv:1804.02812 منتشر شده است
* **نویسنده اول:** اقای Ju-chieh Chou هستند که در اکانت گوگل اسکالر ایشان ذکر شده است که در شرکت تویوتا در شیکاکو فعالیت دارند و در مبحث یادگیری عمیق فعالیت دارند در پروفایل ایشان عدد 7 بعنوان H-index درج شده است و همچنین دو مقاله دیگر بعد از انتشار این مقاله منتشر کردند.

**نویسنده مسئول:** اقای Cheng-chieh Yeh دانشجوی دانشگاه [National Taiwan University](https://scholar.google.co.uk/citations?view_op=view_org&hl=en&org=14962195678536067888) هستند در زمینه های [Speech Processing](https://scholar.google.co.uk/citations?view_op=search_authors&hl=en&mauthors=label:speech_processing) [Text-to-Speech](https://scholar.google.co.uk/citations?view_op=search_authors&hl=en&mauthors=label:text_to_speech) [Voice Conversion](https://scholar.google.co.uk/citations?view_op=search_authors&hl=en&mauthors=label:voice_conversion) [Deep Learning](https://scholar.google.co.uk/citations?view_op=search_authors&hl=en&mauthors=label:deep_learning) فعالیت دارند که H-index ایشان عدد 4 درج شده است و تا کنون 4 مقاله را در رزومه خود دارند.

* این مقاله در سال 2008 منتشر شده است.
* تعداد 118 ارجاع برای این مقاله ثبت شده است.
* عنوان این مقاله به شرح ذیل است

[Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations](https://arxiv.org/abs/1804.02812)

* **چالش ها و کار های پیشین:** به طور کلی، در میان مشکلات دشوار برای رویکردهای VC، نیاز به داده‌های هم‌تراز و هموارسازی بیش از حد سیگنال‌ها دو نمونه هستند که به دقت مورد مطالعه قرار گرفته‌اند. با توجه به مشکلات موجود در به دست آوردن اجسام هم تراز، رویکردهایی با استفاده از مدل‌های مولد مانند رمزگذارهای خودکار متغیر (VAEs) و شبکه‌های متخاصم مولد (GANs) مورد مطالعه قرار گرفتند، زیرا می‌توان آنها را با داده‌های غیر موازی آموزش داد. برخی از کارهای قبلی با موفقیت از VAE برای VC استفاده کردند، اما فریم به فریم صدا را تولید کردند. برخی دیگر از کارهای قبلی توانستند محتوای زبانی را از ویژگی های گوینده در هنگام یادگیری بازنمایی ها جدا کنند، اما بر اساس برخی فرضیات اکتشافی که مورد بررسی قرار داده اند.
* **نوآوری:** ما یک معماری رمزگذار خودکار را پیشنهاد می‌کنیم که قادر به مقابله با چندین فریم در یک زمان است که منجر به نتایج بهتر می‌شود، زیرا در این روش اطلاعات حمل شده توسط فریم‌های همسایه می‌تواند در نظر گرفته شود. این رویکرد همچنین از طبقه‌بندی کننده سخنران آموزش دیده مشترک برای حذف نیاز به مفروضات اکتشافی قبلی استفاده می‌کند. این رویکرد قادر است با جدا کردن ویژگی‌های گوینده از محتوای زبانی، یک مدل واحد را برای تبدیل صدا به بسیاری از سخنرانان مختلف، همه بدون داده‌های موازی، آموزش دهد.
* **پیاده سازی:** ما مدل VC خود را بر روی CSTR VCTK Corpus [36] ارزیابی کردیم. داده های صوتی توسط 109 سخنران به زبان انگلیسی با لهجه های مختلف مانند انگلیسی، آمریکایی و هندی تولید شده است. هر سخنران مجموعه های مختلفی از جملات را بیان می کرد. همانطور که Y در بالا ذکر شد، زیر مجموعه ای از 20 سخنران، 10 زن و 10 مرد انتخاب کردیم. مجموعه داده به طور تصادفی به مجموعه های آموزشی و آزمایشی با درصد 90 و 10 درصد تقسیم شد. ما از طیف‌نگار قدر لگ به عنوان ویژگی‌های صوتی استفاده کردیم. تجزیه و تحلیل طیفی دقیق و تنظیم سنتز مانند کار قبلی بود.
* دستاورد ها : می بینیم که مرحله 2 به طور قابل توجهی کیفیت صدا را از نظر طبیعی بودن و شباهت در ویژگی های بلندگو بهبود بخشیده است.



* در این مقاله در مورد کار های اینده مطلبی ارائه نشده است.