



درس:

پردازش سیگنال‌های دیجیتال

موضوع:

**Introducing Translatotron: An End-to-End
Speech-to-Speech Translation Model**

استاد:

جناب آقای دکتر مهدی اسلامی

دانشجو:

حمیدرضا پورمحمد

شماره دانشجویی:

۴۰۰۱۴۱۴۰۱۱۱۰۳۳

Introducing Translatotron: An End-to-End Speech-to-Speech Translation Model

Wednesday, May 15, 2019

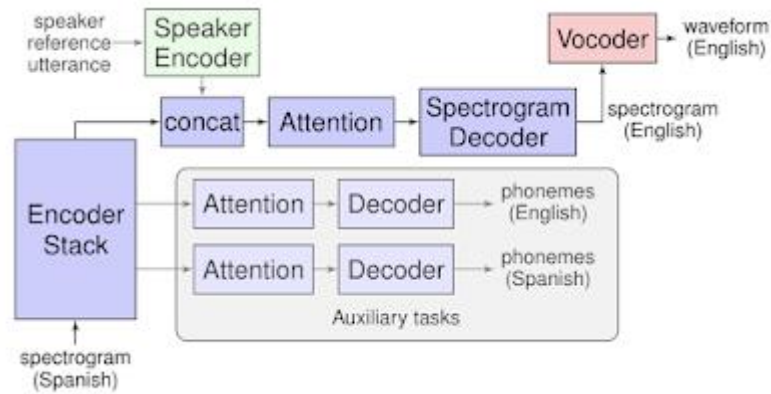
Speech-to-speech translation systems have been developed over the [past several decades](#) with the goal of helping people who speak different languages to communicate with each other. Such systems have usually been broken into three separate components: [automatic speech recognition](#) to transcribe the source speech as text, [machine translation](#) to translate the transcribed text into the target language, and [text-to-speech synthesis](#) (TTS) to generate speech in the target language from the translated text. Dividing the task into such a *cascade* of systems has been very successful, powering many commercial speech-to-speech translation products, including [Google Translate](#).

In “[Direct speech-to-speech translation with a sequence-to-sequence model](#)”, we propose an experimental new system that is based on a single [attentive sequence-to-sequence model](#) for direct speech-to-speech translation without relying on intermediate text representation. Dubbed *Translatotron*, this system avoids dividing the task into separate stages, providing a few advantages over cascaded systems, including faster inference speed, naturally avoiding compounding errors between recognition and translation, making it straightforward to retain the voice of the original speaker after translation, and better handling of words that do not need to be translated (e.g., names and proper nouns).

Translatotron

The emergence of end-to-end models on speech translation started in 2016, when researchers demonstrated [the feasibility](#) of using a single sequence-to-sequence model for speech-to-text translation. In 2017, we demonstrated that such end-to-end models can [outperform cascade models](#). Many approaches to further improve end-to-end speech-to-text translation models have been proposed recently, including our effort on [leveraging weakly supervised data](#). Translatotron goes a step further by demonstrating that a single sequence-to-sequence model can directly translate speech from one language into speech in another language, without relying on an intermediate text representation in either language, as is required in cascaded systems.

Translatotron is based on a sequence-to-sequence network which takes source [spectrograms](#) as input and generates spectrograms of the translated content in the target language. It also makes use of two other separately trained components: a neural [vocoder](#) that converts output spectrograms to time-domain waveforms, and, optionally, a speaker encoder that can be used to maintain the character of the source speaker’s voice in the synthesized translated speech. During training, the sequence-to-sequence model uses a [multitask objective](#) to predict source and target transcripts at the same time as generating target spectrograms. However, no transcripts or other intermediate text representations are used during inference.



Model architecture of Translatotron.

Performance

We validated Translatotron’s translation quality by measuring the [BLEU score](#), computed with text transcribed by a speech recognition system. Though our results lag behind a conventional cascade system, we have demonstrated the feasibility of the end-to-end direct speech-to-speech translation.

Compared in the audio clips below are the direct speech-to-speech translation output from Translatotron to that of the baseline cascade method. In this case, both systems provide a suitable translation and speak naturally using the same canonical voice.

Input (Spanish)	▶ 0:03 / 0:03 ——— 🔊 ⋮
Reference translation (English)	▶ 0:03 / 0:03 ——— 🔊 ⋮
Baseline cascade translation	▶ 0:01 / 0:01 ——— 🔊 ⋮
Translatotron translation	▶ 0:02 / 0:02 ——— 🔊 ⋮


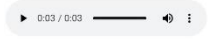

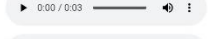

You can listen to more audio samples [here](#).

Preserving Vocal Characteristics

By incorporating a speaker encoder network, Translatotron is also able to retain the original speaker’s vocal characteristics in the translated speech, which makes the translated speech sound more natural and less jarring. This feature leverages previous Google research on [speaker verification](#) and [speaker adaptation for TTS](#). The speaker encoder is pretrained on the speaker verification task, learning to encode speaker characteristics from a short example utterance. Conditioning the spectrogram decoder on this encoding makes it possible to synthesize speech with similar speaker characteristics, even though the content is in a different language.

The audio clips below demonstrate the performance of Translatotron when transferring the original speaker’s voice to the translated speech. In this example, Translatotron gives more accurate translation than the baseline cascade model, while being able to retain the original speaker’s vocal characteristics. The Translatotron output that retains

the original speaker's voice is trained with less data than the one using the canonical voice, so that they yield slightly different translations.

Input (Spanish)	
Reference translation (English)	
Baseline cascade translation	
Translatotron translation (canonical voice)	
Translatotron translation (original speaker's voice)	

More audio samples are available [here](#).

Conclusion

To the best of our knowledge, Translatotron is the first end-to-end model that can directly translate speech from one language into speech in another language. It is also able to retain the source speaker's voice in the translated speech. We hope that this work can serve as a starting point for future research on end-to-end speech-to-speech translation systems.

Acknowledgments

This research was a joint work between the Google Brain, Google Translate, and Google Speech teams. Contributors include Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, Mengmeng Niu, Quan Wang, Jason Pelecanos, Ignacio Lopez Moreno, Tom Walters, Heiga Zen, Patrick Nguyen, Yu Zhang, Jonathan Shen, Orhan Firat, and Yonghui Wu. We also thank Jorge Pereira and Stella Laurenzo for verifying the quality of the translation from Translatotron.

معرفی Translatotron: یک مدل ترجمه از پایان گفتار به گفتار

چهارشنبه ۱۵ می ۲۰۱۹

سیستم های ترجمه گفتار به گفتار هوش مصنوعی گوگل در چندین دهه گذشته با هدف کمک به افرادی که به زبان های مختلف صحبت می کنند برای برقراری ارتباط با یکدیگر توسعه یافته اند. چنین سیستم هایی معمولاً به سه بخش جداگانه تقسیم می شوند: تشخیص خودکار گفتار برای رونویسی گفتار مبدأ به عنوان متن، ترجمه ماشینی برای ترجمه متن رونویسی شده به زبان مقصد، و سنتز متن به گفتار (TTS) برای تولید گفتار در هدف. زبان از متن ترجمه شده تقسیم کار به چنین آبشاری سیستم ها بسیار موفق بوده است و بسیاری از محصولات تجاری ترجمه گفتار به گفتار، از جمله Google Translate را تامین می کند.

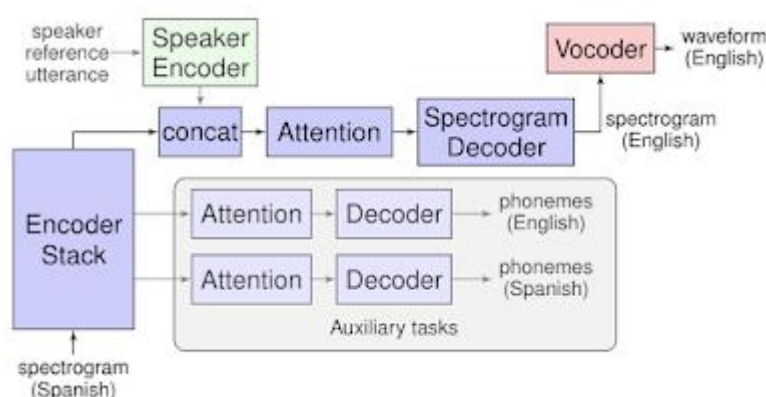
در "ترجمه مستقیم گفتار به گفتار با مدل توالی به دنباله"، ما یک سیستم آزمایشی جدید را پیشنهاد شده است که مبتنی بر یک مدل توالی به دنباله متمرکز برای ترجمه مستقیم گفتار به گفتار بدون تکیه بر حد متوسط است. نمایش متن دوبله *Translatotron*، این سیستم از تقسیم کار به مراحل جداگانه جلوگیری می کند، چندین مزیت را نسبت به سیستم های آبشاری ارائه میدهد، از جمله سرعت استنتاج سریعتر، به طور طبیعی اجتناب از اشتباهات ترکیبی بین تشخیص و ترجمه، حفظ صدای گوینده اصلی پس از ترجمه و مدیریت بهتر. کلماتی که نیازی به ترجمه ندارند (مثلاً نام ها و اسم های خاص).

Translatotron

ظهور مدل های پایان به انتها در ترجمه گفتار در سال ۲۰۱۶ آغاز شد، زمانی که محققان امکان استفاده از یک مدل توالی به دنباله واحد را برای ترجمه گفتار به متن نشان دادند. در سال ۲۰۱۷، که چنین مدل های سرتاسری می توانند از مدل های آبشاری بهتر عمل کنند. رویکردهای بسیاری برای بهبود بیشتر مدل های ترجمه گفتار به نوشتار سرتاسر اخیر پیشنهاد شده اند، از جمله برای استفاده از داده های تحت نظارت ضعیف *Translatotron*. با نشان دادن اینکه یک مدل توالی به دنباله منفرد می تواند مستقیماً گفتار را از یک زبان به گفتار در زبان دیگر ترجمه کند، بدون تکیه بر نمایش متن میانی در هر یک از زبان ها، همانطور که در سیستم های آبشاری لازم است، قدمی فراتر می رود.

Translatotron براساس یک شبکه دنباله به دنباله است که طیف نگارهای منبع را به عنوان ورودی می گیرد و طیف نگاری های از محتوای ترجمه شده را در زبان مقصد تولید می کند. همچنین از دو جزء آموزش دیده دیگر استفاده می کند: یک کد صوتی عصبی که طیف نگارهای خروجی را به

شکل موج‌های حوزه زمان تبدیل می‌کند، و به صورت اختیاری، یک رمزگذار بلندگو که می‌تواند برای حفظ شخصیت صدای گوینده منبع در گفتار ترجمه‌شده سنتز شده استفاده شود. در طول آموزش، مدل توالی به دنباله از یک هدف چندکاره برای پیش‌بینی متن منبع و هدف همزمان با تولید طیف‌نگارهای هدف استفاده می‌کند. با این حال، هیچ رونوشت یا دیگر نمایش‌های متنی میانی در طول استنتاج استفاده نمی‌شود.



معماری مدل Translatotron

عملکرد

کیفیت ترجمه Translatotron را با اندازه‌گیری امتیاز **BLEU**، که با متن رونویسی شده توسط یک سیستم تشخیص گفتار محاسبه شده است. اگرچه نتایج از یک سیستم آبخاری معمولی عقب‌مانده است، اما امکان‌پذیری ترجمه مستقیم گفتار به گفتار انتها به انتها را نشان داده است.

در مقایسه در کلیپ‌های صوتی، خروجی ترجمه مستقیم گفتار به گفتار از Translatotron به روش آبخاری پایه است. در این حالت، سیستم ترجمه مناسبی را ارائه می‌دهند و به طور طبیعی با استفاده از صدای متعارف یکسان صحبت می‌کنند.

حفظ ویژگی‌های صوتی با ترکیب یک شبکه رمزگذار بلندگو

Translatotron همچنین می‌تواند ویژگی‌های صوتی گوینده اصلی را در گفتار ترجمه‌شده حفظ کند، که باعث می‌شود گفتار ترجمه‌شده طبیعی‌تر به نظر برسد و کمتر دل‌انگیز باشد. این ویژگی از تحقیقات قبلی **Google** در مورد تأیید بلندگو و سازگاری بلندگو برای **TTS** استفاده می‌کند. رمزگذار بلندگو در مورد کار راستی‌آزمایی بلندگو از قبل آموزش دیده است، و رمزگذاری ویژگی‌های بلندگو را از یک مثال کوتاه می‌آموزد. شرطی شدن رمزگشای طیف‌نگاری روی این رمزگذاری، ترکیب گفتار با ویژگی‌های گوینده مشابه را ممکن می‌سازد، حتی اگر محتوا به زبان دیگری باشد.

کلیپ‌های صوتی عملکرد **Translatotron** را هنگام انتقال صدای گوینده اصلی به سخنرانی ترجمه شده نشان می‌دهد. **Translatotron** ترجمه دقیق‌تری نسبت به مدل آبشار پایه ارائه می‌دهد، در حالی که می‌تواند ویژگی‌های صوتی گوینده اصلی را حفظ کند. خروجی **Translatotron** که صدای گوینده اصلی را حفظ می‌کند، با داده‌های کمتری نسبت به صدای متعارف آموزش داده می‌شود، به طوری که ترجمه‌های کمی متفاوت را ارائه می‌دهند.

نتیجه‌گیری

تا جایی که ما می‌دانیم، **Translatotron** اولین مدل سرتاسری است که می‌تواند مستقیماً گفتار را از یک زبان به گفتار در زبان دیگر ترجمه کند. همچنین قادر است صدای گوینده منبع را در گفتار ترجمه شده حفظ کند. ما امیدواریم که این کار بتواند به عنوان نقطه شروعی برای تحقیقات آینده در مورد سیستم‌های ترجمه گفتار به گفتار پایان به انتها باشد.

قدردانی

این تحقیق یک کار مشترک بین تیم‌های **Google Brain**، **Google Translate** و **Google Speech** بود. مشارکت کنندگان عبارتند از **Fadi Biadsy**، **Ron J. Weiss**، **Ye Jia**، **Mengmeng Niu**، **Zhifeng Chen**، **Melvin Johnson**، **Wolfgang Macherey**، **Heiga**، **Tom Walters**، **Ignacio Lopez Moreno**، **Jason Pelecanos**، **Quan Wang**، **Patrick Nguyen**، **Yu Zhang**، **Jonathan Shen**، **اورهان فیرات** و **یونگهوی وو** ما همچنین از خورخه پریرا و استلا لورنزو برای تأیید کیفیت ترجمه از **Translatotron** تشکر می‌کنیم.

- تعداد مقالات جدید (بعد از ۲۰۱۸): ۱۵ مقاله جرنال و ۷۸ مقاله پذیرفته شده می‌باشد.

- این مخزن نشان می‌دهد که چگونه می‌توان با استفاده از **Streamlit** یک برنامه ساده انتقال صدا ایجاد کرد. کد این دمو بر اساس مخزن **Real-Time-Voice-Cloning** است.

این برنامه به شما امکان می‌دهد:

صدای خود را ضبط کنید

تعبیه شدن بلندگو را تجسم کنید

گفتار را بر اساس صدای ضبط شده ترکیب کنید