درس:

پردازش سیگنال‌های دیجیتال

موضوع:

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

استاد :

جناب آقای دکتر مهدی اسلامی

دانشجو:

حمیدرضا پورمحمّد

شماره دانشجویی :

۴۰۰۱۴۱۴۰۱۱۱۰۳۳

پاییز ۱۴۰۱

Building a Real Time Voice Transfer App with Streamlit and Python

dataroots

Real Time Voice Transfer

dataroots

What is Voice Transfer?

- Voice cloning
- Artificial simulation of a person's voice
- Applications
  - For people who lost their voice
  - Transferring a voice across languages
  - Generate speech from text in low resource settings

dataroots

Context: Voice Cloning

- Large amounts of high quality recordings is **impractical for many speakers**
- Deep neural network trained on a corpus of hours of recorded speech from a single speaker
- **Giving a new voice to such a model**
  - highly expensive
  - record a new dataset
  - retrain the model

Goal: Transfer Learning

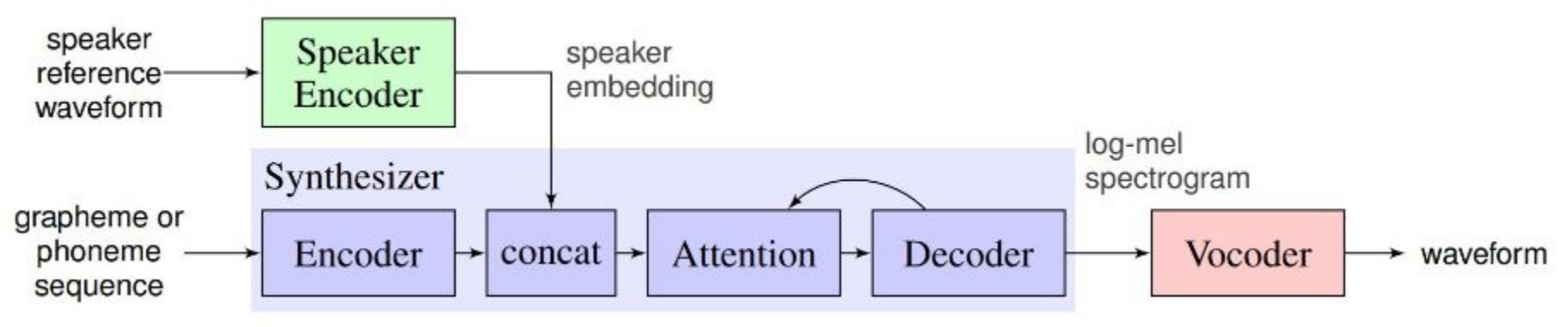
- Key idea of Transfer Learning
  - transfer knowledge from one task with a lot of labelled data
  - to related tasks with very little labelled data
- Text to speech
- Zero-shot setting
  - transfer to voices unseen in the training set

Approach: Voice Cloning

- Decouple speaker modeling from speech synthesis
- Speaker-discriminative embedding network
- Text to speech network
  - conditioned on embedding unique to speaker

Framework: Overview

- Real-time Voice Cloning (Jia et al., 2018)
  - A speaker encoder: GE2E loss (Wan et al., 2017)
  - A synthesizer: Tacotron (Wang et al., 2017)
  - A vocoder: Wavenet (van den Oord et al., 2016)

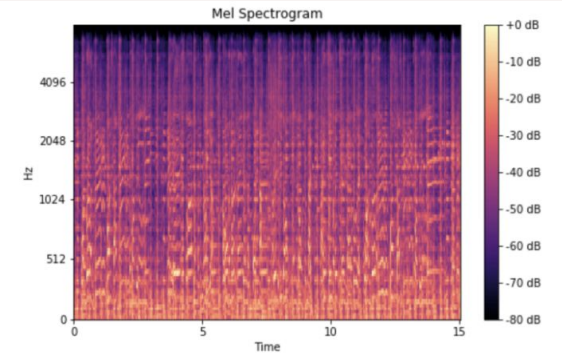


dataroots

Figure from Jemine (2019).

Stage 1: Speaker Encoder

- A speaker encoder: GE2E loss (Wan et al., 2017)
  - The reference speech is a sequence of log-mel spectrogram from a speech utterance
  - Embedding captures the unique characteristics of the speaker
  - Embeddings of utterances from the same speaker have high cosine similarity



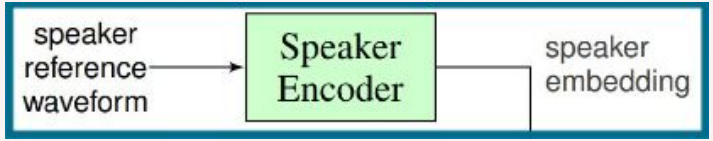
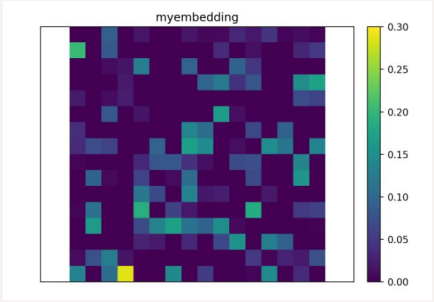
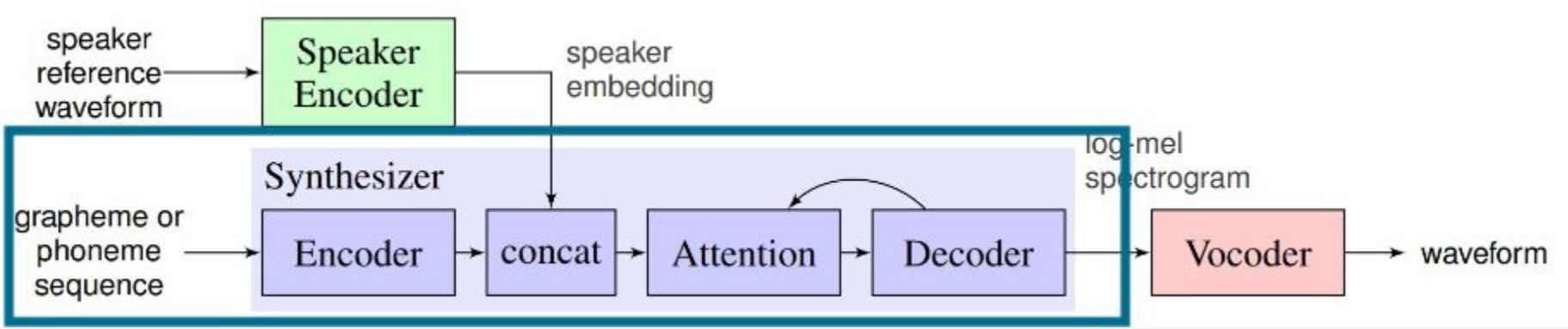


Figure from Jemine (2019).



Stage 2: Synthesizer

- Synthesizer: Tacotron (Wang et al., 2017)
  - Extend attention Tacotron 2 to support multiple speakers (Jia et al., 2018)
  - Embedding vector is concatenated with the synthesizer encoder output at each time step.

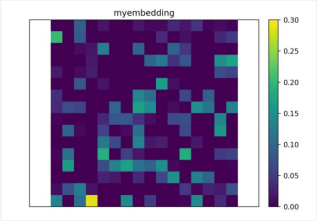


dataroots

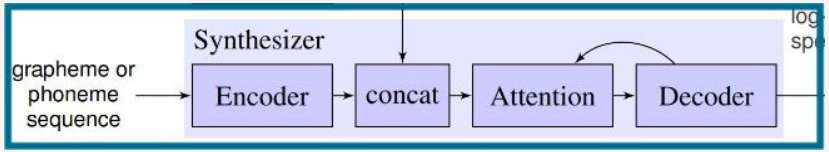
Figure from Jemine (2019).

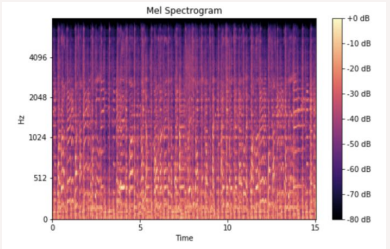
Stage 2: Synthesizer

- Synthesizer: Tacotron (Wang et al., 2017)
  - The synthesizer is trained on pairs of text transcript and target audio.
  - The text is mapped to a sequence of phonemes,
  - Trained in a transfer learning configuration



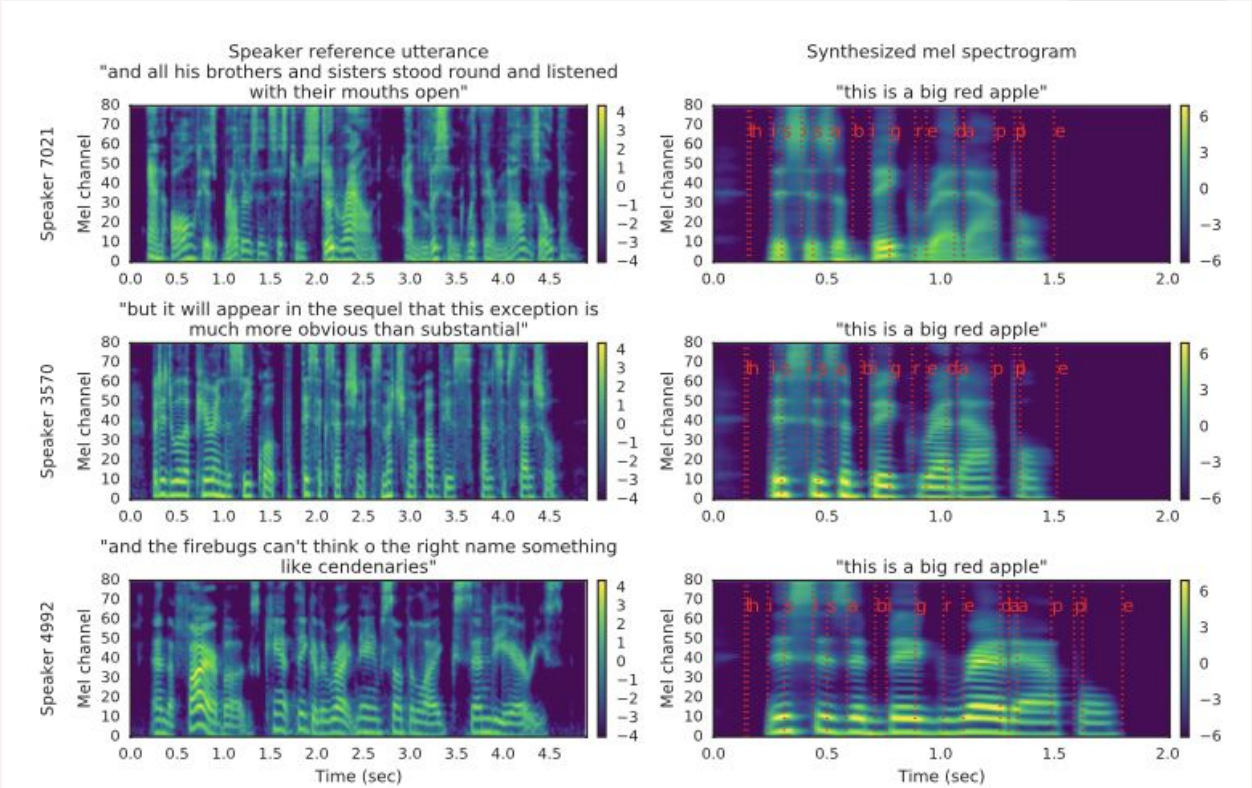
“Hello world”





dataroots

Stage 2: Synthesizer

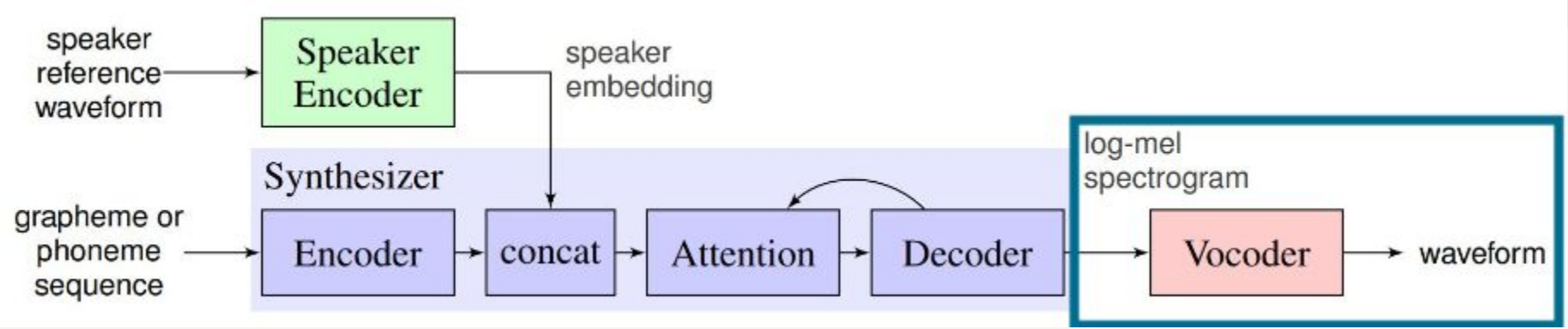


dataroots

Figure from Jia et al. (2018)

Stage 3: Vocoder

- Vocoder WaveNet (Kalchbrenner et al., 2018)
  - autoregressive WaveNet [19] as a vocoder to invert synthesized mel spectrograms into time-domain waveforms

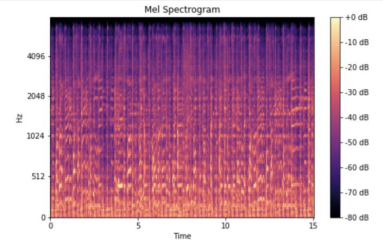


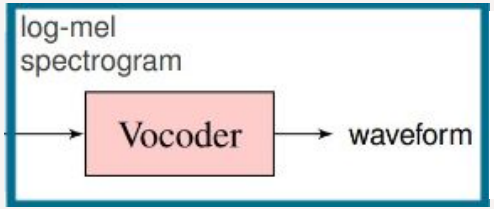
dataroots

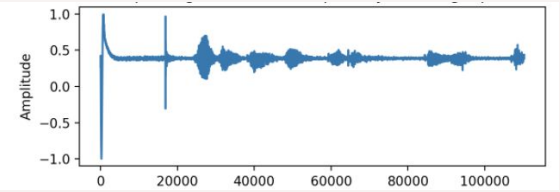
Figure from Jemine (2019).

Stage 3: Vocoder

- Vocoder WaveNet (Kalchbrenner et al., 2018)
  - synthesized mel spectrogram captures all details for high quality synthesis of a variety of voices
  - allowing a multispeaker vocoder by training on data from many speakers







dataroots

Datasets

- VCTK
  - 44 hours of clean speech
  - 109 speakers
  - British accents
- LibriSpeech
  - 2 clean training sets
  - comprising 436 hours of speech
  - from 1,172 speakers

Building Voice Transfer with Streamlit

dataroots

What is Streamlit?

- Data scientists build apps
  - dashboard, data browser, etc.
- Ad hoc building flow
  - jupyter notebook > python script > flask app > need more features...
  - maintainability

What is Streamlit?

- Streamlit is an app framework for data scientists
- Key Idea
  - Make webapps as easy as writing python scripts
  - Use traditional iterative scripting process
  - Instead of layout and event flow
- Workflow
  - Start with python script
  - Slightly annotate to make it an app

What is Streamlit?

- Embrace python scripting
  - everything you can do in a python script
  - you can do in streamlit
- Treat widgets as variables
  - substitute variables with a widget such as st.slider()
  - reuse variables as widgets iteratively
- Reuse data and computation
  - cache computation

Demo

More info

- https://www.streamlit.io/
- https://github.com/datarootsio/rootslab-streamlit-demo

References (1)

- Jia, Ye, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." Advances in neural information processing systems. 2018.
- Wan, Li, et al. "Generalized end-to-end loss for speaker verification." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

References (2)

- Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
- Kalchbrenner, Nal, et al. "Efficient neural audio synthesis." arXiv preprint arXiv:1802.08435 (2018).

References (3)

- Jemine, C. (2019). Master thesis : Real-Time Voice Cloning. (Unpublished master's thesis). Université de Liège, Liège, Belgique. Retrieved from https://matheo.uliege.be/handle/2268.2/6801

The End

dataroots

ساخت صدای زمان واقعی انتقال برنامه با Streamlitو Python

dataroots

صدای زمان واقعی
منتقل کردن

dataroots

انتقال صدا چیست؟

- شبیه سازی صدا
- شبیه سازی مصنوعی صدای یک فرد
- برنامه های کاربردی
  - برای افرادی که صدای خود را از دست داده‌اند
  - انتقال صدا به زبان‌ها
  - تولید گفتار از متن در تنظیمات منابع کم

dataroots

زمینه: شبیه سازی صدا

- مقادیر زیاد ضبط با کیفیت بالا برای بسیاری غیرعملی است بلندگوها
- شبکه عصبی عمیق که بر روی مجموعه ای از ساعت ها گفتار ضبط شده از یک سخنران آموزش داده شده است
- دادن صدای جدید به چنین مدلی
  - بسیار گران است
  - یک مجموعه داده جدید را ضبط کنید
  - مدل را دوباره آموزش دهید

dataroots

هدف: آموزش انتقالی

- ایده کلیدی آموزش انتقالی
- انتقال دانش از یک کار با داده های برچسب گذاری شده زیاد • به کارهای مرتبط با داده های برچسب گذاری شده بسیار کم • متن به گفتار • تنظیم شات صفر • انتقال به صداهایی که در مجموعه آموزشی دیده نمی شوند

dataroots

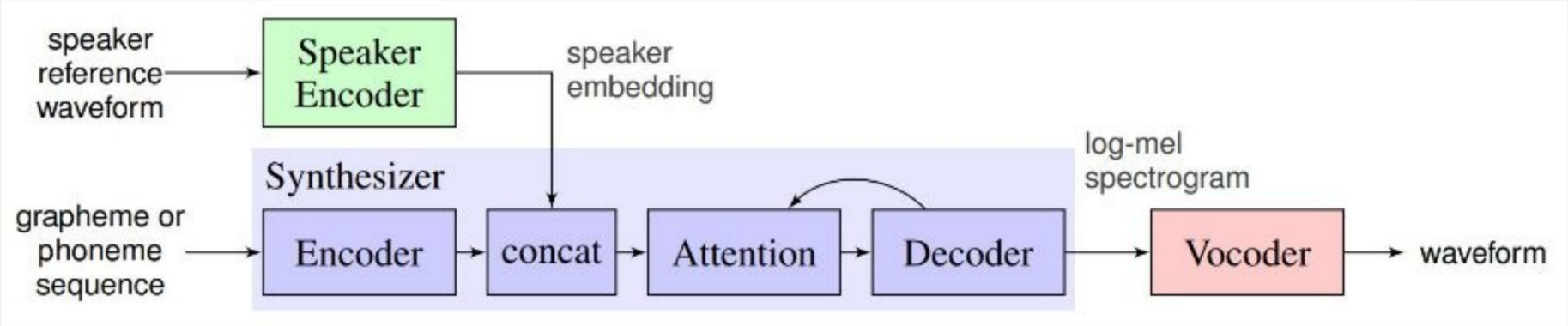
رویکرد: شبیه سازی صدا

- جداسازی مدل‌سازی سخنران از ترکیب گفتار
- شبکه تعبیه‌کننده متمایز سخنران
- شبکه متن به گفتار

- مشروط به تعبیه منحصر به فرد برای بلندگو

چارچوب: بررسی اجمالی

- شبیه سازی صدا در زمان واقعی (جیا و همکاران، 2018)
- رمزگذار بلندگو: از دست دادن GE2E(وان و همکاران، 2017)
- سینت سایزر: تاکوترون (وانگ و همکاران، 2017)رمزگذار صدا: Wavenet (van den Oord و همکاران، 2016)



dataroots

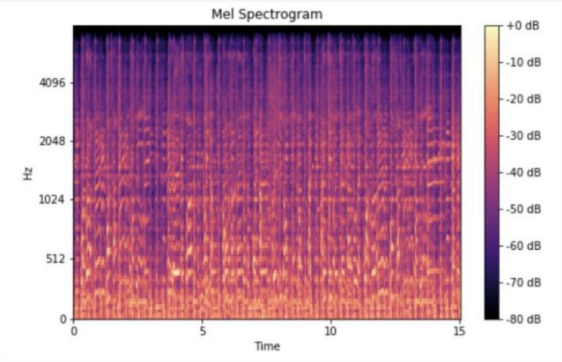
شکلی از Jemine (2019).

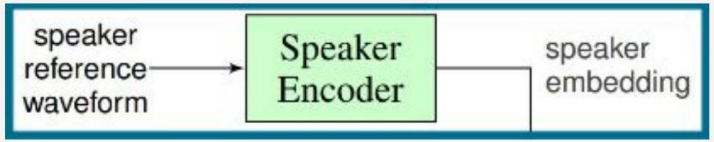
مرحله 1: رمزگذار بلندگو

- رمزگذار بلندگو: از دست دادن GE2E (Wan et al., 2017) • گفتار مرجع دنباله ای از طیف نگاری log-mel از یک گفتار است.

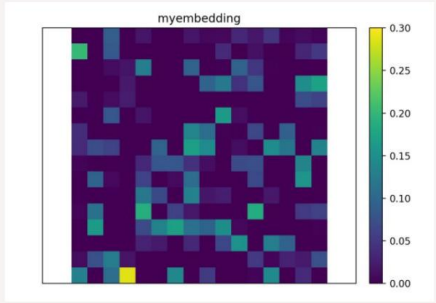
بیان

- جاسازی ویژگی های منحصر به فرد گوینده را نشان می دهد • جاسازی گفته های یک گوینده شباهت کسینوس بالایی دارد





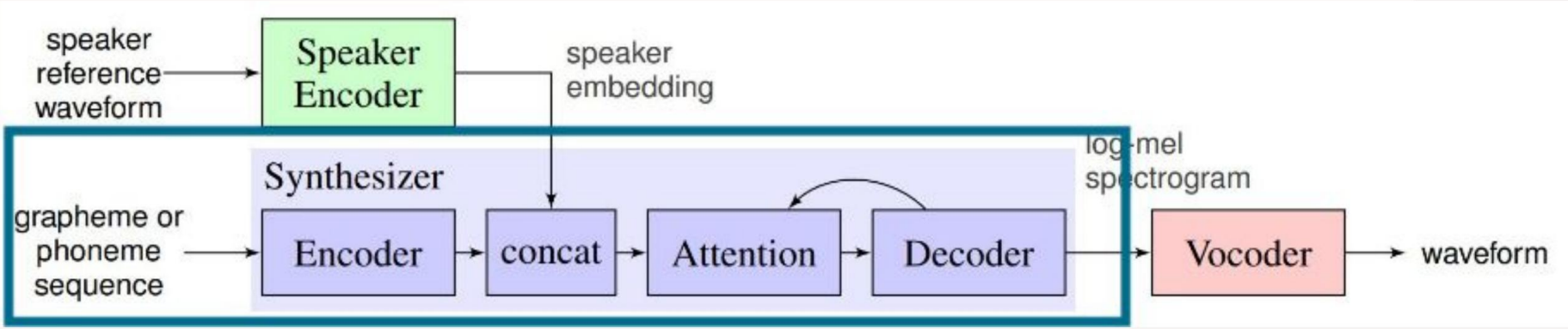
شکلی از Jemine (2019).



مرحله 2: سینت سایزر

- سینت سایزر: تاکوترون (وانگ و همکاران، 2017)
  - توجه تاکوترون 2را برای پشتیبانی از چندین بلندگو گسترش دهید (جیا و همکاران، 2018)
  - بردار جاسازی با خروجی رمزگذار سینتی سایزر در هر کدام الحاق شده است

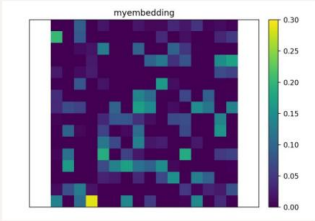
مرحله زمانی



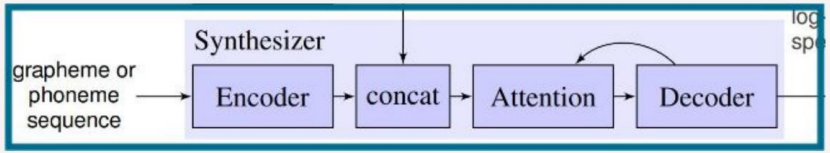
dataroots شکلی از Jemine (2019).

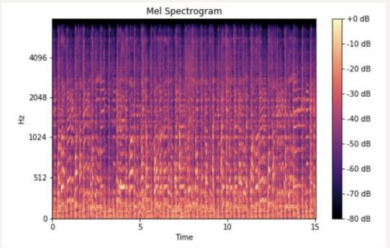
مرحله 2: سینت سایزر

- سینت سایزر: تاکوترون (وانگ و همکاران، 2017)
  - سینت سایزر روی جفت رونوشت متن و صدای هدف آموزش داده شده است. •متن به دنباله ای از واج ها نگاشت شده است، •در پیکربندی یادگیری انتقالی آموزش دیده است



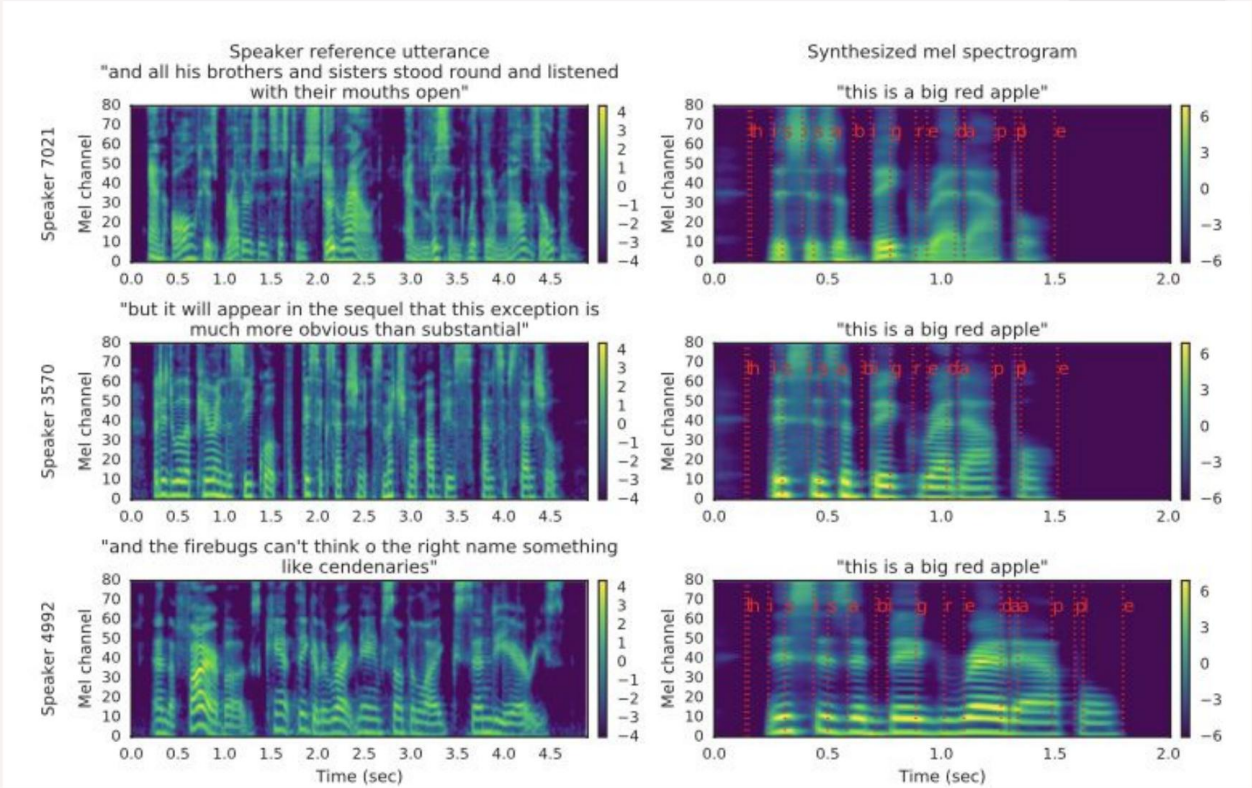
"سلام دنیا"





dataroots

مرحله 2: سینت سایزر

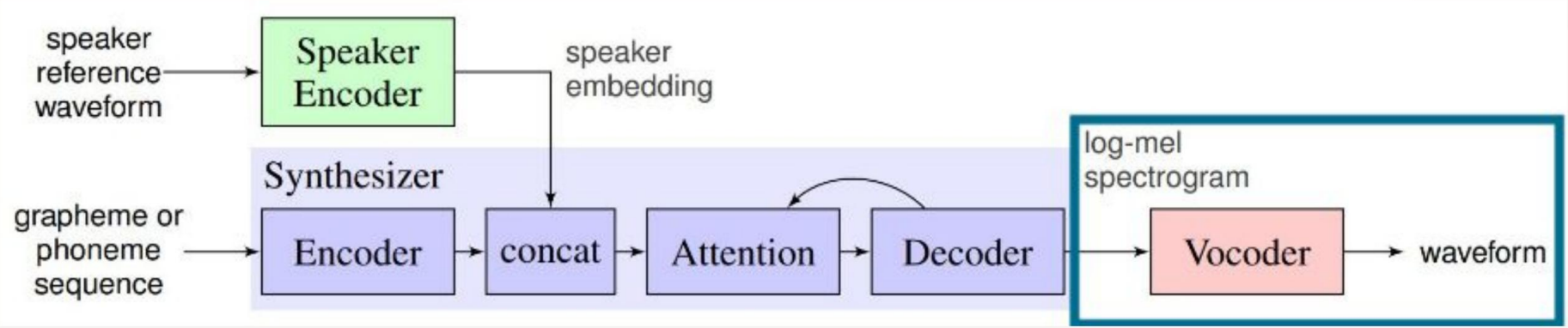


شکل از جیا و همکاران. (2018)

3: Vocoder مرحله

- Vocoder WaveNet (Kalchbrenner et al., 2018)

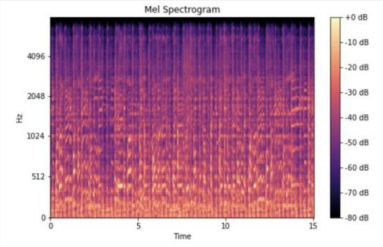
• WaveNet اتورگرسیو [19]به عنوان یک رمزگذار صوتی برای معکوس کردن طیف نگارهای melسنتز شده به شکل موج های حوزه زمان

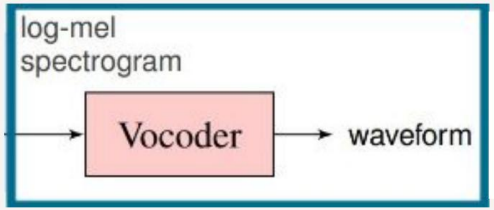


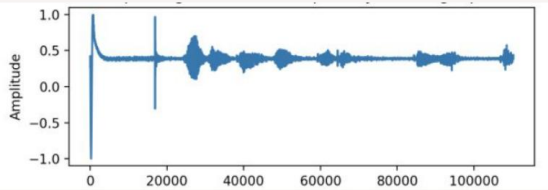
dataroots شکلی از Jemine (2019).

3: Vocoder مرحله

- Vocoder WaveNet (Kalchbrenner et al., 2018)
  - طیف نگار mel سنتز شده تمام جزئیات را برای سنتز با کیفیت بالا انواع صداها ضبط می کند
  - با آموزش داده‌های بسیاری از بلندگوها، به یک کد صوتی چندگوشی اجازه می‌دهد







dataroots

مجموعه داده ها

- VCTK
  - 44 •ساعت سخنرانی تمیز 109 •
    سخنران
    •لهجه های بریتانیایی
- LibriSpeech
  - 2 •ست آموزشی تمیز
    •شامل 436ساعت سخنرانی
    •از 1172بلندگو

صدای ساختمان
انتقال با
Streamlit

dataroots

Streamlitچیست؟

- دانشمندان داده اپلیکیشن می سازند
  - داشبورد، مرورگر داده و غیره
- جریان ساختمان موقت
  - نوت بوک < jupyterاسکریپت پایتون <برنامه < flaskبه ویژگی های بیشتری نیاز دارید... • قابلیت نگهداری

dataroots

Streamlitچیست؟

Streamlit •یک چارچوب برنامه برای دانشمندان داده است •کلید ایده

•برنامه های وب را به آسانی نوشتن اسکریپت های پایتون کنید •از فرآیند اسکریپت نویسی تکراری سنتی استفاده کنید •به جای طرح بندی و جریان رویداد

•گردش کار

•با اسکریپت پایتون شروع کنید •کمی حاشیه نویسی کنید تا به یک برنامه تبدیل شود

Streamlitچیست؟

•اسکریپت نویسی پایتون را در آغوش بگیرید •هر کاری را که می توانید در یک اسکریپت پایتون انجام دهید •می توانید در استریم لایت انجام دهید •ویجت ها را به عنوان متغیر در نظر بگیرید

•جایگزینی متغیرها با ویجتی مانند st.slider() • استفاده مجدد از متغیرها به عنوان ویجت به صورت مکرر

•استفاده مجدد از داده ها و محاسبات

•محاسبات کش

dataroots

Demo

اطلاعات بیشتر

- https://github.com/datarootsio/rootslab-streamlit-demo
- https://www.streamlit.io/

مراجع (1)

- جیا، یه، و همکاران. "انتقال یادگیری از تایید بلندگو به ترکیب متن به گفتار با چند سخنران." پیشرفت در سیستم های پردازش اطلاعات عصبی .2018

- وان، لی و همکاران. «از دست دادن انتها به انتها برای بلندگو تایید." کنفرانس بین المللی IEEE 2018در آکوستیک، پردازش گفتار و سیگنال .2018 ،IEEE (ICASSP).

dataroots

مراجع (2)

- شن، جاناتان، و همکاران. سنتز ttsطبیعی با شرطی سازی wavenetروی پیش‌بینی‌های طیف‌نگاری mel." کنفرانس بین‌المللی IEEEدر آکوستیک، گفتار و پردازش سیگنال (ICASSP) 2018. IEEE، 2018.
- کالچبرنر، نال، و همکاران. "سنتز صوتی عصبی کارآمد." arXiv پیش چاپ arXiv:1802.08435 (2018).

dataroots

مراجع (3)

- جمین، سی .(2019)پایان نامه کارشناسی ارشد: شبیه سازی صدا در زمان واقعی.
(پایان نامه کارشناسی ارشد منتشر نشده). دانشگاه لیژ، لیژ، بلژیک. برگرفته از
https://matheo.uliege.be/handle/2268.2/6801

dataroots

The End

dataroots