



U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND – ARMY RESEARCH LABORATORY

Shareable TTS Components

Dr. Steve LaRocca
Computer Scientist and Team Lead
Battlefield Information Systems Branch

DISTRIBUTION STATEMENT GOES HERE



The Team



Institutions: U.S. Army Research Laboratory¹, United Tribes Technical College², Cornell University³

Individuals: Zakariya Al Sagheer¹, Katherine Blake³, Vince Iglehart², Stephen LaRocca¹, John Morgan¹, Jerral Murray², Gerardo Cervantes¹, G. Hazrat Jahed¹



TEXT-TO-SPEECH (TTS)



- **A system that converts written text to audible speech**
- **TTS is an important enabling language component**
 - For Speech-to-Speech systems (ASR → MT → TTS)
 - For information delivery tools such as ‘talking books’
 - Ultimately, every language community needs a TTS capability
- **USG has relied on commercial TTS software**
 - Licensed commercial products are cumbersome
 - Recent growth in neural computing for TTS with open tools
 - New prospects for more and better shareable TTS components
- **Publicly available neural implementations of TTS, such as Ito’s implementation of Google’s Tacotron, make creating one’s own shareable components easier**



MAKING A SHAREABLE TTS COMPONENT



- **Recent neural (deep learning) methods simplify data preparation**
 - Google’s 2017 Tacotron project followed by Keith Ito’s implementation
- **Keith Ito’s “LJ English” model built with 24 hours of training data**
 - ARL has developed Android Arabic TTS capability using deep learning methods and only 10 hours of training data
- **Compute time and computer resource requirements are substantial**
 - Aging GPU equipment not up to the task, not compatible with current libraries
- **Shareable data and shareable software is an important aspect**
 - ARL is using single speaker data based on in-house translation materials and VOA-type newswire as prompts
- **Neural TTS computes a spectrogram, then renders that data as synthesized speech using a vocoder**



OUR WORK TO DATE



- **Zak Al Sagheer: created 10 hour Arabic dataset**
 - Trained (K. Ito) Arabic Tacotron model
 - Trained Arabic Tacotron2 model
 - Trained more Arabic models: current success using FastSpeech 2
 - Trained vocoders using Arabic data and neural methods.
- **Hazrat Jahed: created 10 hour Pashto dataset**
- **UTTC (Vince Iglehart and Jerral Murray):**
 - Learned Python programming
 - Trained (K. Ito) Tacotron English model
 - Conducted experiments (formal vs. informal text; full vs. ablated dataset)
 - Surveyed possibilities for Northern Ute and/or Lakota dataset → TTS model
- **Gerry Cervantes provided Android expertise, TensorFlow, tflite**



LET'S SYNTHESIZE SOME ARABIC!



- Demonstration by Zakariya (Zak) Al Sagheer

The screenshot displays a development environment for audio synthesis. On the left, Visual Studio Code shows a Python script named `auto_processor.py` with the following code:

```
import logging
import json
from collections import OrderedDict

from tensorflow_tts.processor import [
    LJSpeechProcessor,
    KSSProcessor,
    BakerProcessor,
    LibriTTSProcessor,
    ZAKSpeeh,
]

CONFIG_MAPPING = OrderedDict(
    [
        ("LJSpeechProcessor", LJSpeechProcessor),
        ("KSSProcessor", KSSProcessor),
        ("BakerProcessor", BakerProcessor),
        ("LibriTTSProcessor", LibriTTSProcessor),
    ]
)

class AutoPro...
def __init__(self, config):
    raise NotImplementedError

class AutoPro...
def from_pretrained(cls, pretrained_path, kwargs):
    with open(pretrained_path, "r") as f:
        config = json.load(f)
    try:
```

In the center, a Notepad++ window displays Arabic text from a news article:

قال شهيد عيان من محافظة دهوك ونيوى العراقيين، الأحد، إنهم سمعوا أصوات انفجارا ورأوا أعضاء في السماء بالتزامن مع إعلان وزارة الدفاع التركية انطلاق عملية "مغلب" ضد مسلحي حزب العمال الكردستاني في مواقعهم الجبلية شمال العراق. ونقلت وكالة "ناس" المحلية عن مصادرها في محافظة نينوى قولهم إن طائرات تركية قصفت في جبال سنجار في المحافظة. وقال مصدر آخر لوكالة إن عمليات قصف طالت مواقع في مدينة زاخو

On the right, GoldWave audio software is open, showing two audio waveforms. The top waveform is for `fs2_015_200k_1m.wav` and the bottom is for `fs2_02_200k_740k.wav`. The waveforms show amplitude over time, with a time scale from 0:00.0 to 0:06.5. The GoldWave interface includes a menu bar (File, Edit, Effect, View, Tool, Options, Window, Help), a toolbar with various editing tools, and a status bar at the bottom showing audio properties like Mono, 13.293, 0.000 to 13.293 (13.293), 0.000, and Wave PCM signed 16 bit, 22050 Hz, 352 kbps, mono.



A UTTC PERSPECTIVE: GOALS/CHALLENGES



- **Tacotron experiment this summer: small data results in a worse model**
- **Implications for building TTS models for under-resourced/under-documented languages**
- **Creating data resources for some of these languages: Northern Ute, Lakota**
- **Challenges for TTS models that are based on Native American language data**
- **TTS models offer new capabilities for communities**



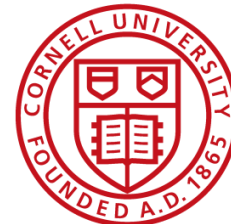
UNITED TRIBES
TECHNICAL COLLEGE



TTS FOR ACADEMICS



- **Teaching materials: introduction to STEM, computational linguistics for linguists and language enthusiasts alike**
- **Second language acquisition: empowering students to practice pronunciation outside the classroom**
- **Experimental materials: a component of the experimental paradigm and a better way to administer instructions to bilingual participants/those with weaker literacy**
- **Language documentation/revitalization: bridging the gap between reading and speaking**
- **Accessibility: free/easy access to screen-readers in many languages for a diverse student body**





GOALS



- **Extend the Ito implementation of Tacotron to build models for additional languages: Pashto, Native American languages, which can be shared. For free.**
- **Make these models transparent and well-documented so that they are easily modified to serve the needs of the military, the academy, and language communities**



CONCLUSION



- **Speech technology including TTS serves military interests because it aids in the communication between Soldiers and local nationals who may not have a language in common or an interpreter available**
- **Speech technology including TTS serves Native American communities because it can help to preserve and revitalize Native American languages**



REFERENCES



- [1] Ito. "Tacotron." (2020). Retrieved June 2020, from <https://github.com/keithito/tacotron>.
- [2] Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., & Junker, M. O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2620-2632).
- [3] Wang, Y., & Skerry-Ryan, R.J., (2020). Expressive Speech Synthesis with Tacotron. Retrieved 21 July 2020, from <https://ai.googleblog.com/2018/03/expressive-speech-synthesis-with.html>.
- [4] Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., ... & Le, Q. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.
- [5] Ren, Y., Hu, C., Qin, T., Zhao, S., Zhao, Z., & Liu, T.Y. (2020) FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. *arXiv preprint arXiv:2006.04558*.
- [6] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (pp. 265-283).