NaijaTTS: A pitch-controllable TTS model for Nigerian Pidgin

Emmett Strickland, Dana Aubakirova, Dorin Doncenco, Diego Torres, Marc Evrard

NaijaTTS: A pitch-controllable TTS model for Nigerian Pidgin

Emmett Strickland[1], *Dana Aubakirova*[2], *Dorin Doncenco*[2], *Diego Torres*[2], *Marc Evrard*[2]

[1]Paris Nanterre University, France
[2]Paris Saclay University, France

emmett.strickland@parisnanterre.fr, dana.aubakirova@université-paris-saclay.fr, dorin.doncenco@université-paris-saclay.fr, diego.torres@université-paris-saclay.fr, marc.evrard@lisn.upsaclay.fr

Abstract

The following report introduces an ongoing project to produce a pitch-controllable speech synthesis model for Nigerian Pidgin, a widely-spoken but poorly-resourced language of West Africa. The first dedicated Nigerian Pidgin TTS model, NaijaTTS, is intended to provide a tool for linguists wishing to study the prosody of this language in an experimental context. In this paper, we present the key objectives of our model, the progress made thus far, and the challenges involved in building a TTS model for this low-resource language.

Index Terms: TTS, prosody, low-resource languages, Nigerian Pidgin

1. Context and overview

Nigerian Pidgin, or Naijá, is an English lexifier creole used as a major lingua franca and vernacular language in Africa's most populous country. With an estimated 75 million speakers, it is by far the world's most widely-spoken creole and one of the most widely-spoken languages in Africa [1]. However, Nigerian Pidgin has been historically underserved and understudied by the field of NLP and the wider linguistic community.

NaijaTTS was conceived to address these shortcomings in two ways. Its first aim is to provide a platform for generating high-quality TTS by leveraging a recently-established corpus of spoken Nigerian Pidgin. Based on the FastSpeech 2 model [2], NaijaTTS is, to our knowledge, the first dedicated TTS model for Nigerian Pidgin and the second model compatible with the language after Meta's Massively Multilingual Speech (MMS) model released several weeks before the submission of this report [3].

Secondly, NaijaTTS will provide a tool for linguists to study the language's prosody through perceptual experiments. Later iterations of NaijaTTS will allow users to directly input pitch information at varying levels of granularity ranging from phoneme-level pitch values to utterance-level contours. This model will therefore provide the opportunity to generate novel utterances and variants of those utterances which differ only by the associated pitch contours. Such a platform will facilitate a wide range of perceptual experiments intended to shed light on the role of tone and melody in semantic and syntactic interpretation.

Thus far, we have succeeded in creating a functional TTS model from a small multi-speaker corpus, which we believe compares favorably against Meta's. In the remainder of this paper, we describe the model's architecture, training data, and the challenges involved in producing a model for this low-resource language.

This Late Breaking Report of the Speech Synthesis Workshop 2023 was not peer-reviewed

2. Architecture and training data

2.1. Dataset

NaijaTTS was trained using the NaijaSynCor corpus, a 30-hour treebank of transcribed Nigerian Pidgin developed between 2017 and 2021 [4]. A subset of this corpus, corresponding to roughly 7.5 hours of speech and 80 speakers, was meticulously transcribed, aligned, and annotated by project members. To ensure the reliability of our dataset, we limited this project to these files. Each audio file was aligned at the utterance level, and the phoneme level using the methodology described in [5] before being adjusted by human annotators. A sound file and corresponding `TextGrid` alignment was then generated for each of the 7469 utterances.

2.2. Core architecture

We based our speech synthesis system on FastSpeech 2 (FS2) [2], a non-autoregressive end-to-end TTS system that is trained directly on ground truth target inputs, from which it extracts duration, pitch, and energy information for use in training. This architecture produces speech from phoneme sequence inputs which we produce by integrating a Nigerian Pidgin pronunciation dictionary adapted from [5]. The FS2 architecture was chosen primarily for its inclusion of a variance adaptor which separately adds predicted duration, pitch, and energy information to the phoneme hidden sequence during the inference stage. This section of the FS2 pipeline can be modified to take in a user-provided pitch vector representing the desired pitch contour. Energy and duration can also be controlled for an even more fine-grained control of prosody.

3. Challenges and adaptations

3.1. Data quality

The NaijaSynCor corpus was primarily recorded for the purposes of transcription and syntactic annotation. Recordings therefore took place in a range of locales, including public spaces and other less-than-ideal environments for speech synthesis. Many files contain noise from animal sounds, background chatter, car honks, and poor acoustic environments. We briefly listened to each recording session and rated their suitability according to personal perceptions. For the training of the text-to-speech model, we tried including only those recorded under acceptable conditions.

Additionally, the corpus primarily consists of spontaneous speech containing a wide variety of dysfluencies. We discarded any utterances containing filled pauses, reparations, abandoned words, or segments deemed incomprehensible to the transcribers.

These constraints reduced the size of our dataset to 3.7 hours of speech and 52 speakers. However, despite the smaller training set, we noticed a marked increase in the quality of the generated speech. Excluding dysfluent speech and files recorded under poor conditions appear to have both contributed to improvements in speech quality.

We also attempted to automatically remove background noise using the speech enhancement model described by [6] and employed by [3]. While this significantly increased the perceived quality of the files used as input, it resulted in somewhat less natural synthesized speech. Nevertheless, the denoiser proved useful when applied directly to the synthesized output, removing many of the perceived artifacts.

3.2. Anonymization

A central goal of this project is to generate realistic speech which cannot be identified with any single speaker represented in the training data. The FS2 architecture allows for the training of separate models for different speakers, provided that each recording is associated with a unique speaker ID. We replaced the speaker IDs in the training set with a binary label corresponding to the speaker's sex, effectively declaring two speakers, respectively comprising all male and all female participants. This initial approach succeeded in yielding anonymous but identifiably male and female voices. While experimenting with our models, we noticed that any changes to the training data, such as the exclusion of certain types of dysfluencies, yielded marked changes to the synthesized voices. We suspect that different approaches to filtering our data significantly altered the representation of the speakers in our training set.

Thus far, we have also noticed changes to the vocal characteristics of our model between generated utterances. At times, two sentences generated from the same model are perceived as different speakers, even if neither can be tied to individuals in our corpus. While our model effectively protects the identities of the different speakers, we suspect that certain sequences of phonemes cause its vocal characteristics to approach those of certain groups of speakers who disproportionately use those sequences in the training data. We consider it essential to resolve this issue going forward, both to ensure a more consistent model and to fully guarantee that it is impossible for a generated utterance to have a voice resembling one used in the corpus.

Inspired by [7], we are currently experimenting with leveraging the speaker representations learned by the model to perform the anonymization. We believe that calculating a single, generalized speaker embedding that is not part of the training set would enable us to generate high-quality and natural-sounding audio, expanding the versatility and applicability of our speech synthesis system. This procedure would provide us with a new voice for our model that is consistent across utterances.

3.3. Comprehensibility

Another central issue going forward is that of comprehensibility. While the naturalness and comprehensibility of our model have improved substantially over the course of our research, the utterances we generate are not always easily understandable due to a mix of artifacts, unusual phonemes, a high speech rate, and unnatural prosody in certain syntactic constructions. However, it is also important to note that many of the utterances used in the training set are themselves difficult to understand when divorced from the broader context of the files in which they appear, as is typical of spontaneous, vernacular speech. In that sense, there may be an upper limit to the comprehensibility of a model trained on such data, and in future tests, we may limit our training set to more prepared oratory, such as radio broadcasts and religious sermons.

Another potential avenue for improvement will be to regenerate our alignments without human intervention. Indeed, the alignments and phonetizations used to train our model are essentially phonetic rather than phonological. A given word form can therefore be transcribed with a variety of speech sounds that do not appear in its canonical form. Furthermore, these transcriptions were also modified by several human annotators who introduced various inconsistencies and errors in our data. We hope that realigning and rephonetizing our corpus will yield a more consistent dataset that is more conducive to training a speech synthesis model.

4. Conclusion and perspectives

In this paper, we have presented the methods employed in an ongoing project to create a natural, pitch-controllable speech for Nigerian Pidgin. Many of the problems and solutions encountered in this project should be of interest to other researchers developing TTS systems from existing corpora of low-resource languages. Even with a small set of training data, excluding utterances that contain dysfluencies or were recorded under poor conditions significantly increased the naturalness and intelligibility of our model. Including recordings of multiple speakers in our training set also appears to be an effective solution to the ethical question of speaker anonymization and one which does not detract from the naturalness of the synthesized speech. Indeed, despite the wide variety of speakers in our training data, we believe that our model already produces speech with naturalness and intelligibility rivaling that of the MMS model, which is likely to have been trained on a single speaker. There remains a great deal of room for progress, particularly when it comes to the comprehensibility of our model and the consistency of its vocal characteristics. However, we are optimistic about the potential solutions outlined in this paper.

5. References

[1] N. Faraclas, *Nigerian Pidgin*. Berlin, Boston: De Gruyter Mouton, 2013, pp. 417–432. [Online]. Available: https://doi.org/10.1515/9783110280128.417

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2022.

[3] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," 2023.

[4] S. Manfredi, B. Caron, K. Gerdes, and M. Courtin, "Naijasyncor: a syntactic treebank, a parser and a wiktionary for naija," in *Summer Conference of the Society of Pidgin and Creole Linguistics*, 2021.

[5] B. Bigi, O. S. Abiola, and B. Caron, "Resources and tools for automated speech segmentation of the african language naija (nigerian pidgin)," in *Human Language Technology. Challenges for Computer Science and Linguistics: 8th Language and Technology Conference, LTC 2017, Poznań, Poland, November 17–19, 2017, Revised Selected Papers 8*. Springer, 2020, pp. 164–173.

[6] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[7] W. Kang, "Speaker anonymization using end-to-end zero-shot voice conversion," Ph.D. dissertation, Massachusetts Institute of Technology, 2022.