Text to Speech Synthesis For Setswana

Boago Okgetheng [1] Gabofetswe Malema [1]

[1]University Of Botswana

Abstract

This research paper investigates the development of an efficient text-to-speech (TTS) framework for Setswana, a complex Bantu language. We explore techniques including syllable concatenation, positioning, and transformer-based models to address its unique linguistic features. Challenges and limitations are evaluated, with potential directions for future research.

Introduction

- Text-to-speech (TTS) synthesis is a technology that converts written text into spoken language [1]. TTS systems have gained widespread applications in assistive technology, language learning, navigation systems, and entertainment. Developing TTS for lesser-resourced languages, like Setswana, presents numerous challenges due to the language's unique linguistic characteristics.
- TTS for Setswana holds promise in enhancing accessibility, improving language preservation, and promoting literacy. This technology can facilitate communication for individuals with visual impairments, aid in language learning, and enable navigation systems to provide directions in local languages.
- Setswana's linguistic complexities pose challenges for TTS synthesis. Its tonal nature, morphological richness, intricate phonetics, and syllable structure require specialized approaches. The lack of comprehensive linguistic resources further complicates TTS development.

Setswana Speech

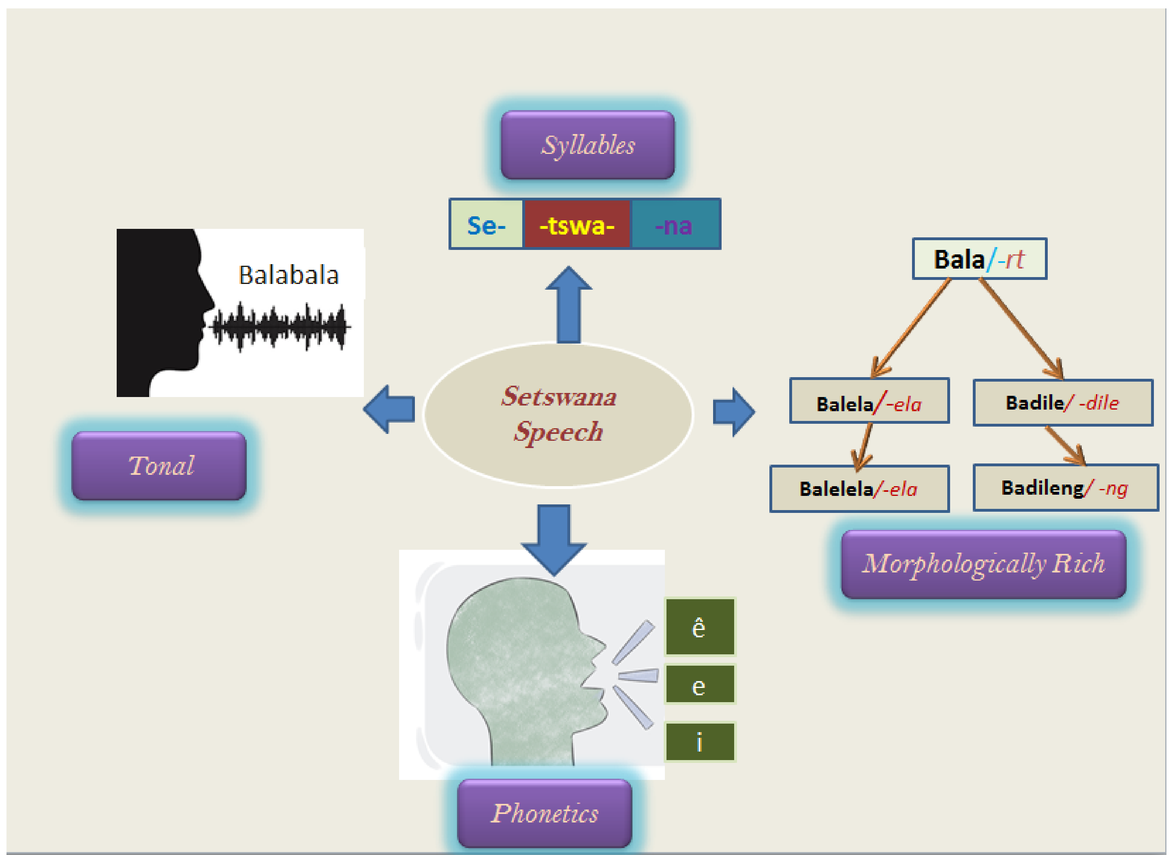

Figure 1. Setswana Speech Issues

Setswana's Linguistic Complexities:

- Tonal nature and semantic nuances.
- Morphological richness and agglutinative features.
- Complex phonetics and intricate syllable structure.

Literature Review

Various TTS techniques have been explored in the literature. Concatenative methods use pre-recorded speech units, while parametric methods employ statistical models to generate speech. Recent advancements in deep learning, particularly transformer-based models, have shown promising results in TTS [2].

Concatenative methods involve recording individual speech units and concatenating them to synthesize speech. Parametric methods employ models trained on linguistic and acoustic features to generate speech. Transformer-based models, like Tacotron and WaveGlow, have shown significant success in capturing linguistic nuances.

Concatenative methods often lack naturalness due to the challenges of selecting appropriate units and transitions. Parametric models provide better control but may struggle with capturing intricate linguistic features accurately. Transformer-based models offer high-quality output but require substantial training data and computational resources.

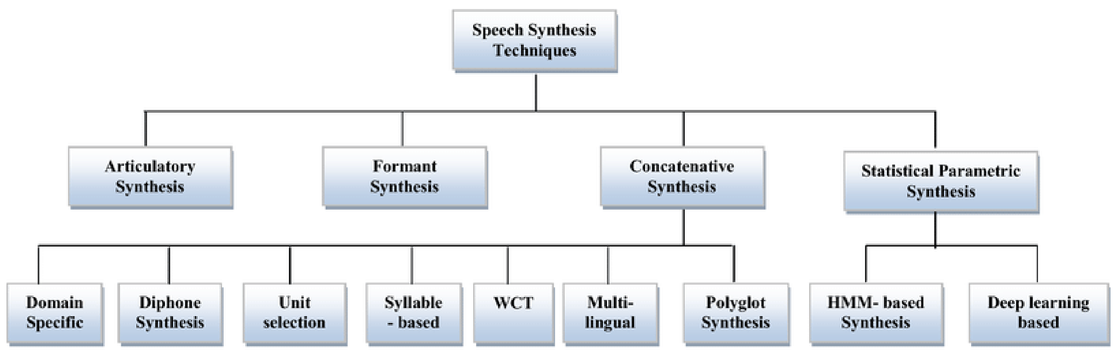

Figure 2. Classification of TTS synthesis techniques [3]

Proposed Solution

1. Syllable Concatenation
   Our method involves these steps:
   - **Recording Syllables:** We recorded individual Setswana syllables and stored them in a database.
   - **Syllable Identification:** Input text is transformed into constituent syllables using a specialized Setswana syllable algorithm.
   - **Matching with Audio:** Identified syllables are matched with pre-recorded audio segments from the Audio Knowledge database.
   - **Concatenation Algorithm:** Using syllable frequency and pitch characteristics, our algorithm seamlessly connects syllables to preserve natural pitch and reduce gaps.
   - **Synthesized Audio:** The process yields synthesized audio with tonal and rhythmic attributes of Setswana speech.

   Illustrating of traditional syllable concatenation method in table 1.

Word	Syllable 1	Syllable 2	Syllable 3	Syllable 4
Balela	Ba	le	la	
Boletsa	Bo	le	tsa	
Babalela	Ba	ba	le	la

Table 1. Traditional Syllable Concatenation.

   Remarks: Requires a comprehensive syllable database; challenges in transition smoothing.

2. Syllable Positioning
   Our approach addresses tone and pitch variations influenced by syllable positions within Setswana words:
   - **Algorithm Development:** Recognizing syllable position importance, we created an algorithm.
   - **Morphological Richness:** Setswana's richness due to affixes and suffixes inspired our method.
   - **Categorizing Affixes:** Algorithm categorized affixes using research findings (Malema et al.).
   - **Position-Dependent Library:** We built a library of segments based on categorized affixes.
   - **Position-Dependent Extraction:** Using syllable positions as input, our algorithm extracted corresponding segments from the Audio Knowledge database.
   - **Coherent Concatenation:** Extracted segments were concatenated to ensure coherence and capture intended tone and pitch variations.
   - **Start, Mid, and End Positions:** By considering syllable positions, including start, mid, and end, we aimed for natural and coherent audio synthesis.

   Illustrating the improvement over the traditional syllable concatenation method, table 2 provides examples:

Word	Syllable 1	Syllable 2	Syllable 3	Syllable 4
Balela	Ba	lela		
Boletsa	Bo	letsa		
Babalela	Ba	ba	lela	

Table 2. Syllable Positioning

   Remarks: Complex rules for position-dependent phonetics; context-sensitive morphemes.

3. Transformers (Tacotron Model)
   Employed the acclaimed Tacotron model for advanced Setswana TTS.
   - **Dataset:** Curated comprehensive Setswana speech dataset for training.
   - **Training Process:** Tacotron mapped text to spectrogram representations.
   - **Speech Generation:** Transformed linguistic content into spectrogram frames.

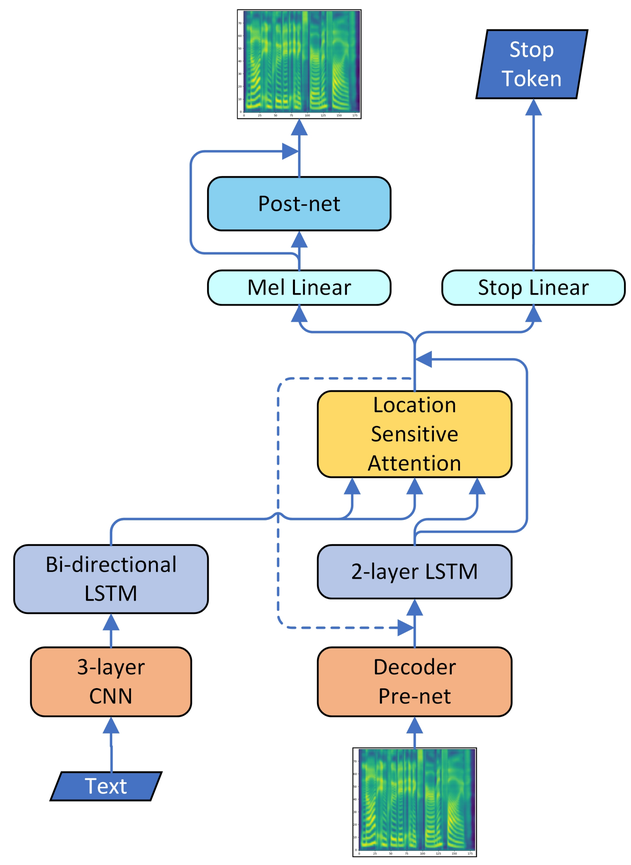

Figure 3. Tacotron Model [4]

Remarks: Success tied to dataset quality and linguistic complexities. It is data hungry however it could be the possible best solution for word text to speech.

Conclusion

Developing a Setswana TTS system is crucial for preserving linguistic heritage and enhancing accessibility. This paper's proposed methods address Setswana's linguistic intricacies, although challenges remain. Future research should focus on collecting linguistic resources, refining models, and exploring data augmentation techniques to enhance TTS performance.Tacotron's potential for high-quality, contextually rich Setswana speech synthesis. Address challenges and dataset diversity.

References

[1] Alex Acero.
An overview of text-to-speech synthesis.
2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421), pages 1–, 2000.

[2] Helal Uddin Mullah.
A comparative study of different text-to- speech synthesis techniques.
2015.

[3] Soumya Priyadarsini Panda, Ajit Kumar Nayak, and Satyananda Champati Rai.
A survey on speech synthesis techniques in indian languages.
Multimedia Systems, 26:453–478, 2020.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu.
Neural speech synthesis with transformer network.
In *AAAI Conference on Artificial Intelligence*, 2018.