Summary of "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram

Predictions"

This research details the development of a text-to-speech (TTS) synthesis system utilizing WaveNet, conditioned on mel spectrogram predictions. The core methodology involves generating speech from text via a WaveNet architecture, specifically tailored for natural-sounding synthesis.

Methodology & Architecture

The system employs a WaveNet model, a deep generative model known for its ability to capture temporal dependencies in audio signals. Crucially, the WaveNet model is conditioned on mel spectrogram predictions derived from the input text. Mel spectrograms are a representation of sound based on perceptual features, reflecting human auditory processing. The precise details of the WaveNet architecture, including layer configurations and hyperparameter settings, are not explicitly provided in this summary due to a lack of information within the text.

Training Data & Procedure

The specific details concerning the training dataset size, characteristics, and training procedure are not sufficiently detailed within this research paper. The paper does not provide specific quantitative metrics for the training process or the size of the dataset.

Quantitative Results

The paper does not present specific quantitative results or performance metrics regarding the TTS systemâ€™s output quality. It does not provide metrics such as Perceptual Evaluation of Speech Quality (PESQ), Signal-to-Noise Ratio (SNR), or Mean Opinion Score (MOS) derived from subjective listening tests. The paper only states that the system achieves "natural-sounding" synthesis.

Authors & Affiliations

The research was conducted by a team at Google, Inc., with contributions from researchers at the University of California, Berkeley, and an individual identified as â€˜jonathanasdf, rpang, yonghui@google.comâ€™.

Summary of Tacotron 2 Research Paper

This paper details the architecture and performance of Tacotron 2, a neural network system for text-to-speech synthesis.

System Architecture

Tacotron 2 employs a sequence-to-sequence model comprised of two primary components. Initially, a recurrent network maps character embeddings to mean- scale spectrograms. Subsequently, a modified WaveNet model functions as a vocoder, transforming these spectrograms into time-domain waveforms.

Conditioning Input & Spectrogram Generation

The system utilizes mel spectrograms as the conditioning input to the WaveNet vocoder, rather than linguistic, duration, and fundamental frequency (Fâ,€) features. This represents a key design choice investigated within the research.

Performance & Validation

The model achieved a mean opinion score (MOS) of 4.53, which was determined to be comparable to a MOS of 4.58 observed for professionally recorded speech. Ablation studies were conducted to assess the impact of the chosen architecture.

Architectural Reduction

The use of the compact acoustic intermediate representation (mel spectrograms) facilitated a significant reduction in the size of the WaveNet architecture, a crucial aspect of the model design.

1. INTRODUCTION

- **Challenge of TTS:** Generating natural speech from text remains a significant technical challenge despite extensive research [1].
- **Previous Synthesis Techniques:** Historically, concatenative synthesis with unit selection was dominant, followed by statistical parametric speech synthesis. However, these methods often produced muffled and unnatural audio due to issues like boundary artifacts.
- **WaveNet:** The WaveNet model, a generative model producing time-domain waveforms, demonstrates audio quality approaching that of real human speech and is utilized in some existing TTS systems [8, 9, 10, 11]. WaveNetâ€™s

inputs include linguistic features (e.g., predicted log fundamental frequency (Fâ,€) and phoneme durations).

- **Tacotron:** The Tacotron sequence-to-sequence architecture [12] simplifies the speech synthesis pipeline by directly generating magnitude spectrograms from character sequences via a single neural network. It employs the Griffin-Lim algorithm [14] for phase estimation, although the authors note this method produces artifacts and lower audio quality.
- **Proposed Unified Approach:** This paper details a unified, end-to-end neural approach combining a Tacotron-style sequence-to-sequence model for generating mel spectrograms and a modified WaveNet vocoder [10, 15]. The model is trained directly on normalized character sequences and corresponding speech waveforms, aiming to produce speech with natural characteristics indistinguishable from real human speech.
- **Comparison to Existing Systems:** Similar approaches, such as Deep Voice 3 [11] and Char2Wav [16], have been explored, but the proposed systemâ€™s naturalness has not been demonstrated to match human speech. These other approaches differ in intermediate representations and model architectures.

Summary of Model Architecture

This research details a two-component system designed for speech synthesis.

- **Recurrent Sequence-to-Sequence Feature Prediction Network:** The system utilizes a recurrent sequence-to-sequence network with attention. This networkâ€™s primary function is to predict a sequence of mel spectrogram frames, derived from an input character sequence. The underlying assumption is that mel spectrograms represent key acoustic features relevant to speech synthesis. The specific details of the recurrent architecture and attention mechanism are not specified in this section.
- **Modified WaveNet:** A modified version of the WaveNet architecture is employed to generate time-domain waveform samples. This generation process is conditional, reliant on the mel spectrogram frames predicted by the initial recurrent network. The text indicates a modification to the standard WaveNet architecture, but does not outline the specific nature of these modifications.

Summary of Section 2.1: Intermediate Feature Representation

This section details the selection and justification for utilizing mel-frequency spectrograms as an intermediate feature representation within the research.

- **Choice of Representation:** The authors elected for mel-frequency spectrograms to connect the two primary components of their system. This choice was driven by the ease of computation from time-domain waveforms, facilitating independent training of the individual components.
- **Relationship to STFT:** A mel-frequency spectrogram is derived from the short-time Fourier transform (STFT) magnitude. Specifically, it represents a nonlinear transformation applied to the frequency axis of the STFT. This transformation is informed by human auditory system measurements.
- **Frequency Scale Emphasis:** The mel scale emphasizes lower frequencies, which are critical for speech intelligibility, while simultaneously de- emphasizing higher frequencies dominated by noise bursts.
- **Comparison to Linear Spectrograms:** Unlike linear spectrograms, which discard phase information (requiring algorithms like Griffin-Lim for inverse conversion), mel spectrograms discard even more information, presenting a more complex inverse problem. However, the authors contend that, relative to the linguistic and acoustic features used in WaveNet, the mel spectrogram offers a simpler, lower-level representation.
- **WaveNet Applicability:** The authors hypothesize that a modified WaveNet architecture, conditioned on mel spectrograms, will successfully generate audio, essentially functioning as a neural vocoder. The core justification rests on the simpler nature of the mel spectrogram representation compared to other features.

Hereâ€™s a concise, formal summary of the provided text, structured as a bullet point list, suitable for academic review:

- **Spectrogram Generation:** The model generates mel spectrograms using a short-time Fourier transform (STFT) with a 50ms frame size, 12.5ms frame hop, and a Hann window.
- **Mel Filterbank & Dynamic Range Compression:** The STFT magnitude is transformed to the mel scale using an 80-channel filterbank (125 Hz - 7.6 kHz), followed by log dynamic range compression. Filterbank magnitudes are clipped to a minimum of 0.01.
- **Encoder Architecture:**
  - Character input is embedded into a 512-dimensional representation and processed through 3 convolutional layers (5x1 filters, 5 characters each) with batch normalization and ReLU activation.
  - The final convolutional layerâ€™s output is fed into a bi-directional LSTM (512 units, 256 in each direction) to create encoded features.
- **Attention Mechanism:** A location-sensitive attention mechanism (based on cumulative attention weights) is used to improve decoder consistency and mitigate subsequence repetition. Input and location features are projected to

128-dimensional representations.

- **Decoder Architecture:**
  - The decoder is an autoregressive LSTM network.
  - A pre-net (2 fully connected layers, 256 units, ReLU) acts as an information bottleneck and is crucial for learning attention.
  - The pre-net output, attention context vector, and LSTM output are concatenated and projected to predict the spectrogram frame.
- **Post-Net:** A 5-layer convolutional post-net with 512 filters (5x1 shape, batch norm, tanh except final) is used to predict a residual and improve reconstruction.
- **Loss Function:** The summed mean squared error (MSE) from before and after the post-net is minimized. Mixture Density Networks were explored but proved difficult to train.
- **Stop Token Prediction:** A probability is predicted to determine when the output sequence has completed. Generation stops at the first frame where this probability exceeds 0.5.
- **Regularization:** Dropout (0.5) is used in convolutional layers and zoneou (0.1) in LSTM layers. Dropout (0.5) is applied only to layers in the pre-net during inference.
- **Architectural Differences:** The model utilizes vanilla LSTM and convolutional layers instead of CBGH stacks and GRU recurrent layers. There is no "reduction factor"; each decoder step generates a single spectrogram frame.

Summary of WaveNet Vocoder Technical Details

Architecture and Layer Configuration:

The WaveNet vocoder utilizes a modified WaveNet architecture incorporating 30 dilated convolution layers grouped into 3 dilation cycles (dilation rate 2k mod 10, where k = 0-29). Two upsampling layers are employed within the conditioning stack, instead of the typical three, to handle the 12.5 ms frame hop of the mel spectrogram input.

Output Generation:

The model generates 16-bit samples at 24 kHz using a mixture of 10 logistic distributions (Mol). Each mixture component is defined by its mean, log scale, and mixture weight, predicted via a linear projection following a ReLU activation.

Loss Function:

The loss function is based on the negative log-likelihood of the ground truth sample, evaluating the modelâ€™s reconstruction accuracy.

Key Technical Details:

- **Dilation Rate:** The dilation rate of the convolutional layers is dynamically adjusted based on the layer index (2k mod 10).
- **Upsampling:** Two upsampling layers are used instead of three to process the 12.5 ms frame hop.
- **Mixture Distribution:** A mixture of 10 logistic distributions is used to model the audio waveform.

3.1 Training Setup Summary

This section details the training procedure for a multi-network architecture involving a feature prediction network and a modified WaveNet.

- **Feature Prediction Network Training:** The feature prediction network was trained using a maximum-likelihood objective function, employing teacher- forcing. A batch size of 64 was utilized on a single GPU. The Adam optimizer (Î²â, = 0.9, Î²â,, = 0.999, c = 10â»â¶) was implemented with an initial learning rate of 10â»Â³ decaying exponentially to 10â»âµ over 50,000 iterations. Lâ,, regularization with a weight of 10â»â¶ was applied.
- **WaveNet Training:** The modified WaveNet was trained on the ground truth- aligned outputs generated by the feature prediction network. A distributed training approach was used, employing a batch size of 128 across 32 GPUs with synchronous updates. The Adam optimizer (Î²â, = 0.9, Î²â,, = 0.999, c = 10â»â¸) was utilized with a fixed learning rate of 10â»â´. Waveform targets were scaled by a factor of 127.5 to facilitate convergence.
- **Dataset:** All models were trained on an internal US English dataset containing 24.6 hours of speech from a single professional female speaker. The dataset utilized normalized text, with numerical values (e.g., â€œ16â€) represented as their written form.

Summary of Evaluation Results

3.2. Evaluation Methodology

The evaluation of the Tacotron 2 system utilized a human-in-the-loop approach to assess audio quality. During inference mode, which differed from the training phase due to the absence of ground truth targets, the predicted outputs were fed back into the decoding process. A custom test set of 100 fixed examples was randomly selected from the internal dataset, and audio generated from this set was evaluated by human raters via an Amazon Mechanical Turk service. Raters provided subjective mean opinion scores (MOS) on a scale from 1 to 5, with 0.5 increments.

3.2. Evaluation Results

The system achieved a mean opinion score (MOS) of 4.526 Â± 0.066, comparable to the ground truth audio (4.582 Â± 0.053). Side-by-side comparisons revealed a slight, statistically insignificant preference (-0.270 Â± 0.155) for the ground truth audio, primarily attributed to occasional mispronunciations by the system. Further analysis of the 100-sentence test set yielded an MOS of 4.354. Manual error mode analysis identified key issues: pronunciation difficulties, particularly with names, highlighting a challenge for end-to-end training approaches needing diverse usage data. Generalization was tested on 37 news headlines, achieving an MOS of 4.148 Â± 0.124, competitive with WaveNet conditioned on linguistic features (4.137 Â± 0.128).

3.2. Generalization and Additional Findings

The system demonstrated some ability to generalize to out-of-domain text, obtaining a MOS of 4.148 Â± 0.124 when processing news headlines. WaveNet conditioned on linguistic features achieved a similar MOS of 4.137 Â± 0.128. Despite generating more natural and human-like speech according to rater comments, the system still exhibited pronunciation challenges. These results emphasize the need for training data that accurately represents intended usage scenarios.

Abstract

This section details an experimental comparison of WaveNet performance based on training data source â€“ either predicted features or ground truth mel spectrograms. The study evaluated the system's performance, measured by Mean Opinion Score (MOS), across different training and synthesis scenarios.

Methodology

The research compared WaveNet performance when trained using either predicted features or ground truth mel spectrograms. Two synthesis conditions were tested: training WaveNet on predicted features and synthesizing using predicted features, and conversely, training on ground truth features and synthesizing from predicted features.

Results

- **Training on Predicted Features & Synthesis from Predicted Features:** The MOS score achieved was 4.526 Â± 0.066.
- **Training on Ground Truth Features & Synthesis from Predicted Features:** The MOS score achieved was 4.449 Â± 0.060.
- **Training on Ground Truth Features & Synthesis from Ground Truth Features:** The MOS score achieved was 4.522 Â± 0.055.

These results indicate that the highest MOS score (4.526 Â± 0.066) was obtained when WaveNet was trained using predicted features and subsequently used to synthesize from predicted features. Conversely, training on ground truth features and synthesizing using predicted features resulted in a lower MOS score (4.449 Â± 0.060). Training on ground truth features and synthesizing from ground truth features yielded the highest MOS score (4.522 Â± 0.055).

Discussion

The observed differences in performance can be attributed to the characteristics of the predicted spectrograms. The squared error loss utilized in the feature prediction network led to the generation of spectrograms that exhibit oversmoothing and reduced detail. Consequently, when WaveNet was trained on these oversmoothed features and tasked with synthesizing speech, the network lacked the necessary information to generate high-quality speech waveforms. The system's ability to learn effectively was hampered by the inherent discrepancies between the training and inference data.

Summary of Experimental Results on Linear Spectrogram Prediction

This study investigated the performance of Tacotron 2 models trained to predict linear-frequency spectrograms utilizing the Griffin-Lim algorithm for inversion. The experiments involved varying model architectures and training parameters, evaluated through Mean Opinion Scores (MOS).

3.3.2. Linear Spectrograms â€“ Experimental Results:

- **Model Architectures & MOS Scores:** The models were compared based on their ability to predict linear spectrograms. Tacotron 2 models incorporating the Griffin-Lim algorithm yielded an average MOS of 3.944 Â± 0.091. Models incorporating WaveNet achieved higher MOS scores of 4.510 Â± 0.054 and 4.526 Â± 0.066

respectively.

- **Architectural Parameters:** The experiments explored varying numbers of total layers and cycles within the model. Specifically, a model with 30 layers and 3 cycles exhibited an MOS of 3.930 Â± 0.076. A model with 24 layers and 4 cycles achieved an MOS of 4.547 Â± 0.056. Conversely, a model with 12 layers and 2 cycles yielded an MOS of 4.481 Â± 0.059.

- **Dilation Cycle Size & Receptive Field:** The dilation cycle size, measured as samples/ms, varied between 10, 6, and 1. The receptive field, also measured as samples/ms, ranged from 6,139 / 255.8 to 61 / 2.5. These parameters were consistently associated with the MOS scores observed for each model configuration.

Summary of Tacotron 2 Research Paper

This summary details the key aspects of the Tacotron 2 neural Text-to-Speech (TTS) system as presented in the conclusion section of the research paper.

System Architecture

Tacotron 2 employs a sequence-to-sequence recurrent neural network incorporating attention mechanisms for the direct prediction of mel spectrograms. The system is integrated with a modified WaveNet vocoder for audio synthesis.

Synthesis Characteristics

The resulting TTS system produces speech characterized by both Tacotron-level prosody and WaveNet-level audio quality. This signifies a synergistic integration, leveraging the strengths of both components.

Training Methodology

The system is designed for direct training from data, eliminating the requirement for extensive, manually engineered feature extraction.

Sound Quality

The system achieves state-of-the-art sound quality, approximating the characteristics of natural human speech.

This document presents a summary of acknowledgements provided within the research paper.

Authorship and Contributions

The research acknowledges contributions from several key individuals and teams: Jan Chorowski, Samy Bengio, AÃ¨ron van den Oord, and the WaveNet and Machine Hearing teams, for providing discussions and advice.

Technical Support and Evaluation

Heiga Zen and the Google TTS team are credited with feedback and assistance concerning the evaluation of the research.

Review Process

The authors express gratitude to the reviewers for their â€œvery thoroughâ€ feedback.

Okay, here's a bullet point summary of the provided research and technical documents, categorized for clarity:

I. Foundational Techniques & Models:

- **HMM-Based Synthesis (Early Work):**
  - [1] Tanaka, et al. (1998) - Early work demonstrating the use of HMMs for speech synthesis.
  - [30] X. Gonzalo, et al. (2016) â€“ Focuses on HMM-driven unit selection synthesizers for real-time use.
- **Statistical Parametric Speech Synthesis:**
  - A recurring theme across many of the documents, focusing on using statistical models to represent speech.
- **Recurrent Neural Networks (RNNs):**
  - [19] Schuster & Paliwsky (1997) â€“ Introduced Bidirectional RNNs â€“ foundational for sequence modeling.
  - [20] Hochreiter & Schmidhuber (1997) â€“ Introduced Long Short-Term Memory (LSTM) cells â€“ critical for handling long-range dependencies in sequential data.
  - [21] Chorowski, Bahdanau, Serdyuk, Cho (2015) - Attention-based models for speech recognition, highlighting the importance of attention mechanisms.
  - [22] Bahdanau & Cho (2015) â€“ Neural Machine Translation using attention.
  - [26] Zoneout - Regularizing RNNs by randomly preserving hidden activations, offering a method to improve RNN training.
  - [27] PixEInCNN++ - Improvement of the PixelCNN model utilizing discretized logistic mixture likelihood.

II. Deep Learning Approaches for Speech Synthesis:

- **WaveNet:**
  - [28] Van den Oord, et al. (2017) â€“ Introduced Parallel WaveNet: a fast, high-fidelity speech synthesis model.
  - [29] Kingma & Ba (2015) - Introduced the Adam optimization algorithm.
- **LSTM-Based Synthesis:**
  - [31] Zen, et al. (2016) â€“ Developed fast, compact, and high-quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices.

III. Training & Regularization Techniques:

- **Dropout:**
  - [25] Srivastava, Hinton, Krizhevsky, Sutskever, Salakhutdinov (2014) â€“ Introduced Dropout â€“ a simple technique to prevent overfitting in neural networks.

IV. Key Concepts & Research:

- **Attention Mechanisms:** A growing focus across numerous models, capturing relationships between input and output sequences.
- **Mixture Density Networks:** (Bishop, 1994) - Used for modeling probability distributions.
- **Batch Normalization:** (Ioffe & Szegedy, 2015) - Accelerated deep network training.

Do you want me to elaborate on any specific area, provide more detail about a particular model, or perhaps focus on a specific type of research (e.g., attention mechanisms, WaveNet)?