International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 03 | Mar 2023 www.irjet.net p-ISSN: 2395-0072

DIGEST PODCAST

Mitran R[1], Kamleshwar T[2], Vignesh N[3], Sri Harsha Arigapudi[4], Nitin S[5]

[1,2,3,4,5]*UG Student, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore.*

Abstract - *This is the age of instant gratification. Browsing the entire Publication is dawdle , hence we have proposed an application that summarizes all your reading's in a snap of time using AI technologies. The system is composed of Optical Character Recognition(OCR) engine to convert the image to text and transformer to summarise the text, dispensing recurrent networks followed by feature prediction network that maps character embeddings to mel-spectrogram and a GAN based vocoder to convert spectrogram to time based waveforms. Through extensive experiments we demonstrate digest podcast ability to recognize, summarize, speech synthesis for summarized audio generation. Our code can be found at https://github.com/mitran27/Digest_Cassette.*

Index terms: *Optical Character Recognition*, segmentation, summarisation, mel-spectrogram, GAN, speech synthesis.

1.INTRODUCTION

Reading and interpreting a protracted segment of text is a challenging task. The above process in natural scene text is generally divided into three major tasks like Optical Character Recognition, Summarization(interpretation) and synthesizing the audio.

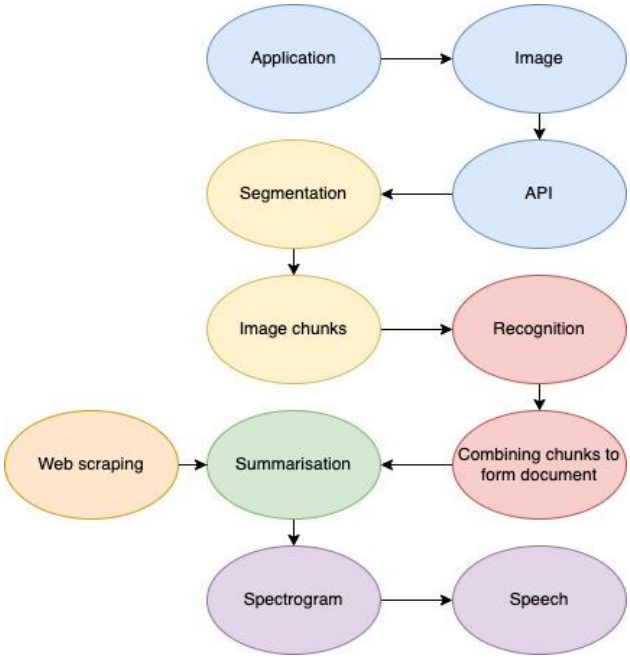


Fig-1: block diagram of digest podcast system architecture

The promise of deep learning architectures to produce probability distributions for applications such as natural images, audio containing speech, natural language corpora. There are very few applications over the decade to combine different techniques to dominate the field. Due to the wide range of vocabulary in the language and speech generated by the model was not promising. We propose a new podcast application which overcome the above drawbacks.

In the proposed application, the architecture is built with the state of the art networks after a bunch of research studies to find better models with their hyperparameters.

The OCR model is built using the scene text detection and text recognition using the computer vision architectures which are proved to give promising results. The summarization model is built using the transformer eschewing recurrent networks and relying entirely on self-attention mechanism, which allows more parallelization to reach the new state of the art model. The choice of abstractive or extractive summarization is given to the user. The podcast model which generates the natural speech from the summarized text is built using two independent networks for feature prediction and speech synthesis. The layers in the model are chosen carefully to avoid artifacts in the sound. Hyperparameters are chosen properly to generate synthetic speech more natural compared to human speech.

2. RELATED WORK

Recognizing tasks/problems have been efficiently addressed by Hidden Markov Model, Recurrent Neural Networks, Long Short-Term Memory (LSTM) network. These methods use artificial neural networks or stochastic models to find the probability distribution of the output. Although these methods are efficient, they have a drawback of limited variation of inputs. So, we use Transformers[1], which uses attention mechanisms to find the output. Transformers is an architecture where each word in the text will pay a weighted attention to the other words in the text, which makes the model work with large sentences. Transformer model exhibits a high generalization capacity of text which makes it state of the art in the NLP domain.

Text Rank algorithm[2] which was inspired by the PageRank algorithm created by Larry Page. In place of web pages, we use sentences. Similarity between any two

sentences is used as an equivalent to the web page transition probability. The similarity scores are stored in a square matrix, similar to the matrix M used for PageRank. Text Rank is an extractive and unsupervised text summarization technique.

Segmentation results can more accurately describe scene text of various shapes such as curve text. However, the post-processing of binarization is essential for segmentation. DBnet[3] suggests trainable dynamic binarization had a greater impact on predicting the coordinates of the text.

Recognition involves extracting the text from the input image using the bounding boxes obtained from the text detection model and the text is predicted using a model containing a stack of VGG layers and the output is aligned and translated using CTC[5] loss function. Baek suggested that using real scene text datasets[4] had significant impact than synthetically generated text dataset.

Connectionist Temporal Classification[5] suggests that training a sequence-to-sequence language model requires computation of loss to fine tune the model, vanilla models utilize cross entropy for finding loss derivatives. Text recognition is sequence to sequence models but suffers from a drawback that they do not have fixed output length and they are not aligned with the input. So, in order to Label Unsegmented Sequence Data, we use CTC which computes loss by finding all the best possible path of the target using conditional probability from the probability distribution of classes available from all time steps.

A sequence to sequence model with attention and location awareness[6] is used to convert the summarized text to spectrograms, which is an intermediate feature for audio using an encoder-decoder GRU[9].

Generative Adversarial Network[7] is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

High-Fidelity Audio Generation and Representation Learning With Guided Adversarial Autoencoder[8] that HiFi-GAN consists of one generator and two discriminators: multi-scale and multi-period discriminators. The generator and discriminators are trained adversarial, along with two additional losses for improving training stability and model performance.

MelGAN[10] has generator which consists of stack of up-sampling layers with dilated convolutions to improve the receptive field and a multiscale discriminator which examines audio at different scales since audio has structure at different levels.

3. MODEL ARCHITECTURE

Our proposed system consists of six components, shown in Figure 1

(1) Semantic segmentation model with differentiable binarization

(2) Temporal recognition network with CTC

(3) Abstractive summarization model using attention mechanism

(4) Extractive summarization algorithm using PageRank algorithm with word2vec embeddings

(5) Feature prediction network with attention and location awareness which predicts mel-spectrogram

(6) Generative model generates time domain waveforms conditioned on mel-spectrogram

3.1. DIFFERENTIABLE BINARIZATION SEGMENTATION NETWORK

DBnet uses an differentiable binarization map to adaptively set threshold for segmentation map to predict the result. Spatial dynamic thresholds found to be more effective than static threshold.

Static binarization

if($P_{i,j}$>threshold)
$B_{i,j}$=1
else
$B_{i,j}$=0

Dynamic binarization

$B_{i,j} = \text{div}(1,1+\exp(-K*(P_{i,j}-T_{i,j})))$

where B is the final binary map,

P is the prediction map,
T is the threshold map,
K is a random value empirically set to 50.

The static binarization is not differentiable. Thus it cannot be trained along with the segmentation network.

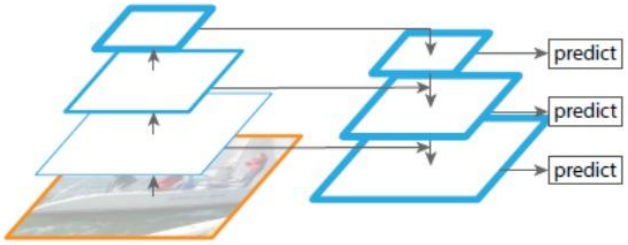


Fig-2: Feature Pyramid Network(FPN)

The segmentation architecture of our method is shown in Figure 2. Initially the image is passed through feature pyramid network and the output features are concatenated and the final prediction is passed through prediction and threshold networks which consists of couple of transposed convolution layers to up sample to its original size.

Our model works well with light weight backbone, with the backbone of RESNET 18. Features were taken at three stages at the scale of (8x,16x,32x) of the original size. The outputs of the FPN produced features of dimension $d_{model} = 256$.

The loss is computed using DICE loss for prediction map and binary map and L1 loss for threshold map. Losses are compared to the ground truth map of the original image.

3.2. RECOGNITION

Temporal image recognition is performed in word images to recognize the words in the text format. This is achieved by constructing a model which consists of VGG network for feature prediction, GRCNN can be used in cases where high accuracy is needed neglecting the speed of the model. The features are transposed to temporal dimension and passed to two layered bi-directional LSTM model with 1024 units to predict the output temporally(Figure 3). The output from the model is not aligned with the input image. So Connectionist Temporal Classification(CTC) model is used because CTC algorithm is alignment free, CTC works by summing the probability of all possible alignments.

Fig-3: Prediction from the LSTM model

CTC is a neural network which uses associated scoring functions to train recurrent networks such as LSTM, GRU to tackle sequence applications where time is variable. CTC is differentiable with output probabilities from the recurrent network, since it sums the score of each alignment(Figure 4), gradient is computed for the loss function with respect to the output and back propagation is done to train the model.

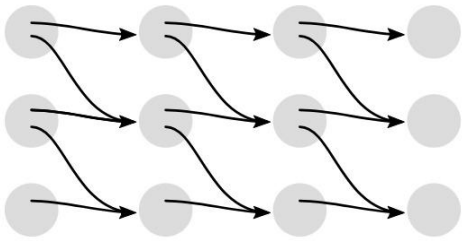


Fig-4: CTC of the output probabilities

3.3. ABSTRACTIVE SUMMARIZATION

The abstractive summarization is achieved by using sequence to sequence models recurrent networks, LSTM, GRU are firmly established in text summarization. The usage of these models restricts their prediction in short range. In contrast, transformer networks allows to capture longer range. We used self-attention network as a building block for transformers with embedding layer for word embeddings and positional encoding for positional embeddings since positional embeddings works for variable sentences.

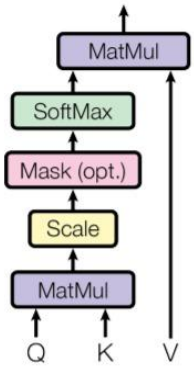


Fig-5: Scaled Dot-Product Attention

The word embedding along with its position are passed to the encoder where attention mechanism is used (Figure 5) to learn by mapping a query and a set of key-value pairs to an output vector. The basic idea is that, if the model is aware of the abstract, then it provides high level interpretation vectors.

A decoder with similar architecture mentioned in encoder is used to obtain the output prediction using the interpreted vectors.

3.4. EXTRACTIVE SUMMARIZATION

The extractive approach involves picking up the significant phrases from the given document, which is

achieved by creating a sentence graph and measuring the importance of each sentence within the graph, based on incoming relationships. The value of the incoming relationships are measured using cosine distance between sentence embeddings. The created graph passed into the Page Rank algorithm to rank the sentences in the document.

The dimensions for the word embeddings are empirically set to 100.

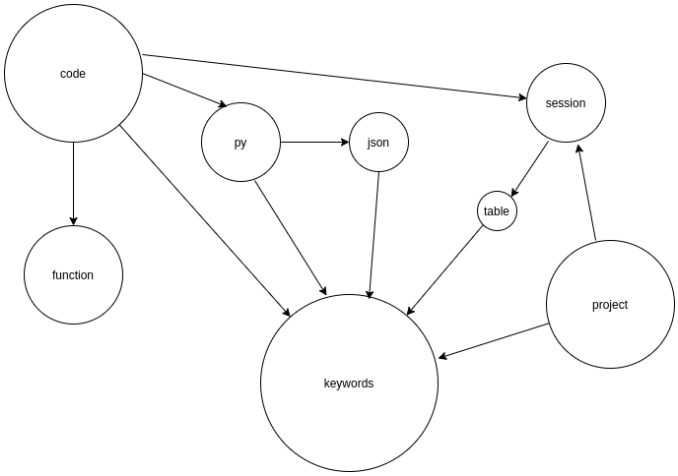


Fig-6: Sentence graph

3.5. ALIGNER – INTERMEDIATE FEATURE REPRESENTATION

Converting text(200 tokens) to audio waveform(200K tokens) is not possible. So we chose low level intermediate representation: mel-spectrogram to act as a bridge between the aligner and the vocoder. Mel-spectrograms scale down the audio to 256x times. Mel-spectrograms are created by applying short-time Fourier transform (STFT) with a frame size 50ms and hop size 12.5ms and STFT is transformed to mel-scale using 80 channel mel-filterbank followed by log dynamic range compression, because humans respond to signals logarithmically.

Since the dimension of spectrograms are more than the dimension of text, sequence to sequence model is used. The model consists of encoder and decoder with attention and locational awareness. Attention mechanism is used to handle long range sequences. Experiments without locational awareness produce made the decoder to repeat or ignore some sequences(Figure 7).

EXPECTED RESULT:
The quick brown fox jumps over a dog

OBTAINED OUTPUT:
The-qui-qui-ck-br-own-own-fox-ps-ov-ov-a-dog

Fig-7: Output without locational awareness

The encoder converts the character sequence to character embedding with dimension 512 which are passed through bi-directional GRU layer containing 512 units(256 in each direction) to generate the encoded features. The encoded features are consumed by the decoder model which contains two layer of auto regressive GRU of 1024 units and an attention network with dimension 128. The attention network is represented by the below equation.

$$a(s_{t-1}, h_i) = v^T \tanh(W_1 h_i + W_2 S_{t-1} + W_3(\text{conv}(a(s_{t-2}, h)))$$

where a represents the attention scores,

s represents the decoder hidden states,

h represents the encoder hidden states,

W_1,W_2,W_3 represents the linear network,

conv represents the convolution network to extract features from the attention scores.

The convolution network for locational awareness is built with kernel size 31 and dimension 32.

A pre-net is built using feedforward network to teacher-force the original previous timestamp to the decoder. Experiments with dimension 128 and 256 made the decoder alter the input rather than attending the encoded features. Dimension 64 with high dropouts(0.5) forced the decoder to attend the encoded features because there is no full access to the teacher-forced inputs. This makes the decoder predict correctly even if there is a slip in the current time stamp input(previous time stamp output).

The output from the decoder is passed through feature enhancer network to predict a residual to enhance the overall construction of spectrogram which consists of four layers of residual networks of dimension 256 with tanh activation and dropouts are added at each layer to avoid overfitting. We minimize the mean squared error (MSE) from before and after the enhancer to aid convergence.

3.6. VOCODER – GANS

Converting spectrograms to time domain waveforms auto-regressively is a tedious and time consuming process.

3.6.1. GENERATOR

A spectrogram is approximately 256x smaller than the audio. A Generative model is built using (8x,8x,2x,2x) up-sampling. Experiments were carried to find the best scaling factors to attain 256x. Firstly, a feature extractor network is used to extract features with a kernel size 7 and dimension 512 to feed the up-sampling layers. Up-sampling layer network consists of transposed convolution with its corresponding scale factor and some convolutional networks to increase the receptive field. Experiments are made for increasing the receptive field, after transposed convolution the receptive field would be low. Receptive field of a stack of dilated convolution increases exponentially with number of layers, while vanilla convolutions increases linearly.

The stack dilation convolution consists of three dilated RESNETs with kernel size 5 and dilations 1,5,25. The receptive field of the network should look like a fully balanced symmetric tree with kernel size as the branch factor. Finally, convolutional layer is used to predict the final audio(dimension of size 1) and an tanh activation function.

3.6.2. DISCRIMINATOR

Discriminators are used to classify whether the given sample is real or fake. We adopt a multiscale discriminator to operate on different scales of audio. Feature Pyramid Networks(Figure 2) are used to extract features at different stages on scale 2x,4x,8x with kernel size 15 and dimension 64. The features from each stage are passed to the discriminator block.

Each discriminator block learns features of different frequency range of the audio by stack of convolutional layers with kernel size 41 and stride 4 and a final layer to predict the binary class(real/fake).

4. EXPERIMENTS AND RESULTS

4.1. TRAINING SETUP

I. Segmentation model was trained on Robust Reading Competition challenges datasets like DocVQA 2020-21, SROIE 2019, COCO-Text 2017 which contained document image dataset along with their coordinates of the words.

II. Recognition network was trained using two major synthetic datasets. MJSynth (MJ) which contains 9M word boxes. Each word is generated from a 90K English lexicon and over 1,400 Google Fonts. SynthText is generated for scene text detection containing 7M word boxes. The texts are rendered onto scene are cropped and used for training.

III. Summarization model was trained using two different datasets. The inshorts dataset which contains news articles along with their headings. CNN/DailyMail non-anonymized summarization dataset. It contains two features. One is the article which contains text of news articles and the highlights which contains the joined text of highlights which is the target summary.

IV. Our training process involved training the feature prediction network followed by training the vocoder independently. This was predominantly done using the LJ speech dataset consisting of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books which also includes a transcription for each clip. The length of the clips vary from 1 to 10 seconds and they have an approximate length of 24 hours in total.

4.2. HARDWARE AND SCHEDULE

We trained our models on NVIDIA P100 GPUs using hyper-parameters described throughout the paper. We trained the segmentation, recognition, summarization models for 12-24 hours(approx.) each. Feature prediction and vocoder were trained for 1-2 weeks(approx.) each.

4.3. OPTIMIZERS AND LOSS

Model	Optimizer	Loss
DBnet	Adam	Dice
Recognition	Adam	CTC
Transformer	Adam	Cross Entropy
Aligner	Adam with weight decay	Mean Squared Error
Vocoder	Adam for discriminator and generator	Hinge loss

Table-1: Optimizers and loss used to train different models

4.4. EVALUATION

We evaluated the OCR model(segmentation + recognition) with real time scene text images. We compared the model with tesseract and Easyocr engines and achieved better results than them. Sample results of this model is shown in Figure 8.

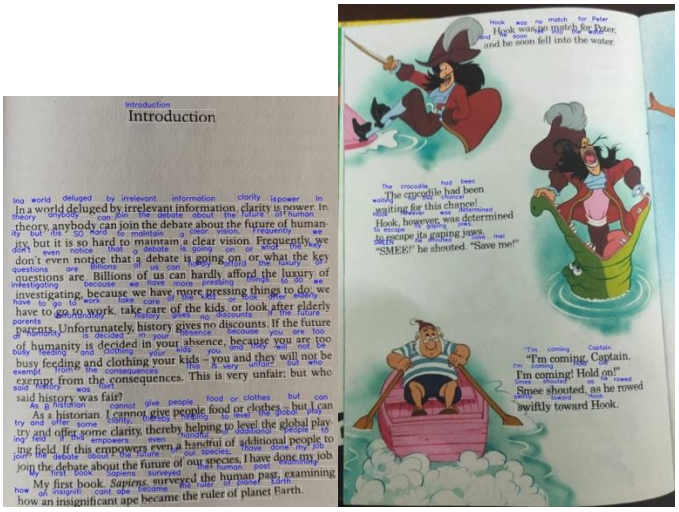


Fig-8: Result of OCR

During the evaluation of transformers, after converting the source text into encoded vectors and modelling the summarized text, during inference mode, the ground truth targets are not known. Therefore, the outputs from the previous step is passed as an input. In contrast to the teacher-forcing method used while training. We randomly selected 50 samples from the test dataset and evaluated the model. The context between the each evaluated result and the ground truth were similar. Sample results of this model is shown below.

Sample Result-1:

Source: summstart ride hailing startup uber s main rival in southeast asia, grab has announced plans to open a research and development centre in bengaluru. the startup is looking to hire around 200 engineers in india to focus on developing its payments service grabpay. however, grab s engineering vp arul kumaravel said the company has no plans of expanding its on demand cab services to india. summend

Model output : summstart uber to open research centre in bengaluru summend

Ground truth : summstart uber rival grab to open research centre in bengaluru summend

Sample Result-2:

Source: summstart wrestlers geeta and babita phogat along with their father mahavir attended aamir khan s 52nd birthday celebrations at his mumbai residence. dangal is based on mahavir and his daughters. the film s director nitesh tiwari and actors aparshakti khurrana, fatima shaikh and sakshi tanwar were also spotted at the party. shah rukh khan and jackie shroff were among the other guests. summend

Model output: summstart geeta babita phogat attend birthday party s laga aamir summend

Ground truth: summstart geeta, babita, phogat family attend aamir s birthday party summend

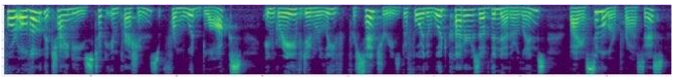
Generating speech during inference to predict the next frame, the output from previous step is taken as input in contrast to teacher-forcing, we generate random text sequences evaluated on the model. The spectrograms predicted from the model is converted to audio using griffin-lim algorithm to check the mapping of character embeddings to the phonons, ignoring the quality of the speech.

Generative vocoder is used to generate high quality speech from the spectrogram. The synthesized audio is rated in respect to their clarity and natural sounding speech. A sample evaluation of this model is shown below.

Sample input:

For although the Chinese took impressions from wood blocks engraved in relief for centuries before the woodcutters of the Netherlands, by a similar process.

Sample output:



5. CONCLUSIONS

These models segmentation, text recognition, extractive and abstractive summarisation, spectrogram and speech synthesis are targeted at summarising the content in the image given and then converting it into audio. With these models, the user will have over 90% accuracy in converting the image with text into audio.

To handle a wide range of words, we can add the entire English vocabulary which requires high performance GPU's. We have trained the models only with English language. Further, the models can be trained with different languages for wider usage. As of now, the OCR model deals with a specific professional style of font. This could be further improved by training it with fonts of different styles. The audio output could be made more human-like by training the model more vigorously.

ACKNOWLEDGEMENT

We extend our gratitude to the Department of Computer Science and Engineering, PSG College of Technology for supporting us throughout the research work.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob, Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need", 2017, In Advances in Neural Information Processing Systems, pages 6000–6010.

[2] K. U. Manjari, "Extractive Summarization of Telugu Documents using TextRank Algorithm," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 678-683, doi: 10.1109/I-SMAC49090.2020.9243568.

[3] Liao, Minghui and Wan, Zhaoyi and Yao, Cong and Chen, Kai and Bai, Xiang, "Real-time Scene Text Detection with Differentiable Binarization", 2020, Proc. AAAI.

[4] Baek, Jeonghun and Matsui, Yusuke and Aizawa, Kiyoharu, "What If We Only Use Real Datasets for Scene Text Recognition? Toward Scene Text Recognition With Fewer Labels", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[5] E. Variani, T. Bagby, K. Lahouel, E. McDermott and M. Bacchiani, "Sampled Connectionist Temporal Classification," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4959-4963, doi: 10.1109/ICASSP.2018.8461929.

[6] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[8] K. N. Haque, R. Rana and B. W. Schuller, "High-Fidelity Audio Generation and Representation Learning With Guided Adversarial Autoencoder," in IEEE Access, vol. 8, pp. 223509-223528, 2020, doi: 10.1109/ACCESS.2020.3040797.

[9] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017, pp. 1597-1600, doi: 10.1109/MWSCAS.2017.8053243.

[10] K. Kumar, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis".