Using Deepfake Technologies for Word Emphasis Detection

Eran Kaufman
Shenkar College, Israel
erankfmn@gmail.com

Lee-Ad Gottlieb and **Dina Mayzlish** and **Or Tiram** and **Hila Wiesel** and **Nofar Yosef**
Ariel University, Israel
leead@ariel.ac.il

Abstract

In this work, we consider the task of automated emphasis detection for spoken language. This problem is challenging in that emphasis is affected by the particularities of speech of the subject, for example the subject's accent, dialect or pitch. To address this task, we propose to utilize generative speech deep nets to produce an emphasis-devoid speech sample for the speaker. This requires extracting the text of the spoken voice, and then using a pre-recorded voice sample from the speaker to produce an emphasis-devoid speech for this task. By comparing the generated speech with the spoken voice, we are able to isolate patterns of emphasis which are relatively easy to detect.

1 Introduction

We as humans have developed a deep sensitivity to the 'music' of speech, meaning its stress, rhythm and intonation. Intonation in particular may be used to express wonder, cynicism or emphasis, and any one of these may alter (or even completely reverse) the meaning of a sentence.

Let us take for example the simple sentence 'I did not take your bag.' Placing emphasis on different words of the sentence can affect its overall meaning: Emphasizing the subject of the sentence – '*I* did not take your bag' – implies that the bag may still have been taken, but by someone else. Emphasis on the possessive adjective – 'I did not take *your* bag' – implies that I did take a bag, only a different one. And emphasis on the object – 'I did not take your *bag*' – implies that I took a different object of yours, and so on.

Establishing the correct emphasis in a spoken sentence is therefore central to correct interpretation of that sentence. Indeed, written language has long ago adopted tools to convey emphasis or meaning, such as italicization, punctuation marks, and the more recent use of emoji symbols. Hence, understanding and classifying word emphasis is an important task for fields related to human-machine interaction, for example machine translation, spoken information retrieval, automated question response, sentiment analysis and speech synthetics.

Our contribution. The task of automated emphasis detection is complicated by the fact that different languages, dialects or accents already feature inherent differences in emphasis. In addition, different voices resonate at different frequencies. Hence, this makes our task speaker specific. Traditional methods utilize extraction of specific hand picked features such as the fundamental frequency ($F0$), energy and duration of the spoken word. In contrast we propose to address this problem by correlating the sentence of the speaker with the same emphasis-devoid sentence of the same speaker. The cross correlation of the same words spoken by the same subject is high, while the correlation of the same words when one of them is emphasized is significantly lower. In order to obtain an emphasis-devoid sample of the word, we employ the most recent generative speech models: Given a previous sample of the user's voice and any text, the models create a close approximation to the user reciting the text.

An overview of our computational approach is as follows: Our detector is composed of several separate modules. A voice encoder processes the speech sample to produce a representative embedded vector capturing the speaker's voice characteristics. Given the query statement, a speech-to-text (STT) module generates text from the spoken sentence. Then a text-to-speech (TTS) module uses the embedded vector and the text of the spoken sentence to generate an audio waveform of the same text as if produced by the same speaker, but devoid of any special emphasis. This constitutes the 'deepfake' synthesized version of the speech. Finally, an analyzer will compare the query statement and its deepfake. As these differ solely in emphasis, this final step can identify the emphasized words.

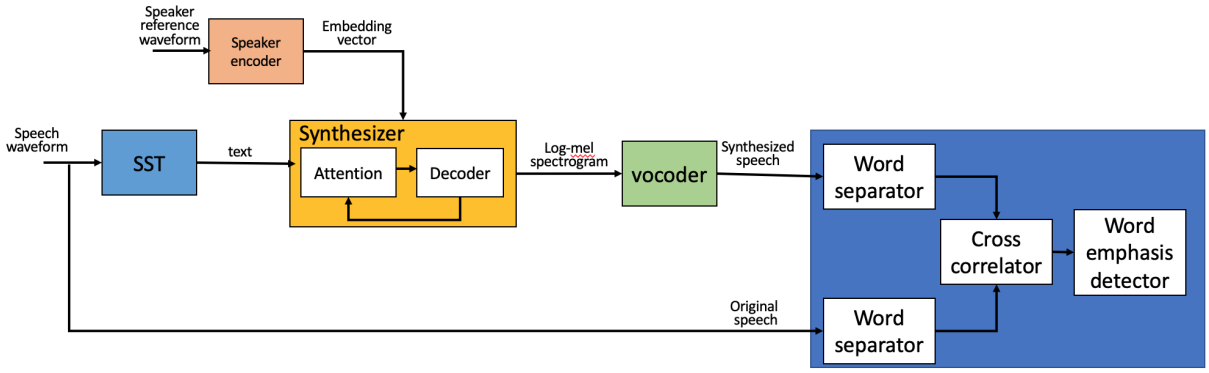

Figure 1: Algorithm workflow.

2 Related Work

Prosody and word emphasis are the subjects of significant research in the field of speech correction, in particular as relates to speech of non-native speakers. They have also attracted much attention in the field of neural TTS synthesis, where attention to emphasis can yield more expressive speech.

Intonation models, such as the Fujisaki (Fujisaki, 1983), Hirst (Hirst, 1992), Rise/Fall/Connection (RFC) (Taylor, 1994) and Tilt models (Taylor, 1998), aim to provide linguistically meaningful interpretations to an utterance. Basic components of intonation events include pitch accents and edge tones (Ladd, 2008). Pitch accents are associated with syllables and signify emphasis, while edge tones occur at the edges of the phrase and give cues such as continuation, question or statement. Kun et al. (Li et al., 2010) used intonation detection in order to detect errors in English speech, and to then provide corrective feedback to speakers of English as a second language. They developed a pitch accent detector based on a Gaussian mixture model, and used features based on energy, pitch contour and vowel duration.

There have been several relevant contributions in the field of generative TTS, with the overarching goal of improving generated prosody. Several variational (Hsu et al., 2019; Zhang et al., 2019) and non-variational (Skerry-Ryan et al., 2018; Wang et al., 2018) models have been suggested for learning latent prosodic representations.

3 Our Work

We present a new approach for emphasis detection based on a comparison between the spoken word and its generative counterpart. Our algorithm uses a sample from a target speaker, and extracts representative features from it. Then given a spoken query statement, the algorithm extracts the text of the query, and produces a 'vanilla' TTS version of this text (that is, TTS with no specific emphasis) using the previously extracted representative features. This emphasis-void speech is then compared to the query statement by cross correlation.

Background. Recent vanilla neural TTS synthesis technologies have achieved realistic synthetic speech generated from a very small sample of a speaker's voice (Ren et al., 2019; Kim et al., 2021; Jia et al., 2018). These TTS models are based on deep neural networks, and are trained using an encoder-decoder architecture. They map input characters or phonemes to acoustic features (for example, mel-spectrograms) or directly to the waveform. The acoustic features can be converted into waveforms via vocoders (van den Oord et al., 2016; Yang et al., 2021).

Our work is based primarily on the SV2TTS TTS architecture (Jia et al., 2018). This specific architecture is composed of three independently trained neural networks:

- A speaker encoder (based on (Wan et al., 2018)), which uses a sample of the speaker's voice to compute a fixed size embedding vector.
- A sequence-to-sequence synthesizer (based on (Shen et al., 2018b)), which constructs a mel-spectrogram from a sequence of grapheme or phoneme inputs, conditioned on the embedding vector.
- An autoregressive WaveNet vocoder (Oord et al., 2016), which converts the mel spectrogram into the time-domain waveform.

Our construction. The workflow of the algorithm is illustrated in Figure 1. Our emphasis detector is composed of five distinct ordered parts:

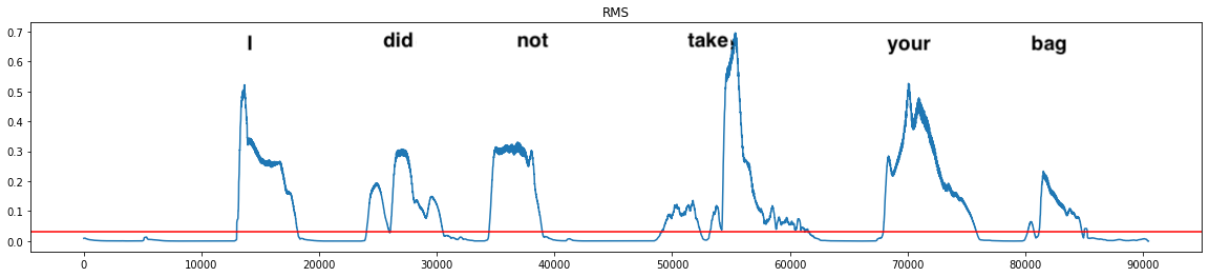

Figure 2: RMS sliding window and word separation.

Step 1: Encoder. The above encoder utilizes a voice sample provided by the speaker to create an embedding vector representing the voice properties of the speaker.

Step 2: Speech to text. The speaker's query statement is inputted into an STT module, which extracts the text of this statement.

Step 3: Text to speech. The TTS module uses the synthesizer described above. Both the text produced from the STT step and the embedding vector produced by the encoding step are fed to the synthesizer, which then produces a waveform.

This waveform is an emphasis-devoid generative model of the speaker reciting the query statement. We recall that vanilla neural TTS systems are not capable of synthesizing emphasis due to the loss of sentiment information (Bai et al., 2022). This computed waveform serves as our baseline for the task of emphasis detection.

Step 4: Waveform comparison. Having computed the synthesized speech, we can compare it to the spoken query statement, to determine which word or words are emphasized. Our comparison technique is detailed in Section 3.1 below.

3.1 Comparison between waveforms

Our premise is that the synthesizer can produce a reasonable imitation of emphasis-devoid speech of the speaker. The emphasis of a word by the speaker may differ from the synthesizer waveform in that the speaker's word is either higher or lower pitched relative to the normal voice produced by the synthesizer. Hence, a cross correlation test between the respective spectrograms of these two waveforms may allow us to identify the special emphasis made by the speaker. Our approach differ from traditional methods in that, we don't assume what features are changed (pitch, duration, energy) but instead we assume that the words themselves are uncorrelated enough with the regular speech in order from a human to detect them.

To effectively compare words, we need to first separate both the synthetic and query speech into their distinct words. This is done using a sliding root mean square (RMS) window, while applying a low threshold to distinguish between spoken and silent parts of the speech (Qi and Hunt, 1993) (see Figure 2). We then compute the fast Fourier transform (FFT) for each individual word, and compare for each word its two spectrograms, which correspond to the synthesized and spoken speech. We discovered two kinds of miscorrelation:

The first is *pitch accents and edge tones*, meaning that the speaker's emphasis of a word is accomplished by modulating regular speech into a higher (or sometimes lower) fundamental frequency $F0$. In this case, the general shape of the spectrogram remains the same, but its central frequency shifts. This is identifiable by the peak of the cross correlation of the two spectrograms.

The second is *continuous modulation*, wherein the speaker modulates the voice up and down several times during the word utterance. Here the spectral distribution is significantly different, and is more evenly spread out compared to the autogenerated waveform. In this case the cross-correlation between the two spectrograms is low for all frequency shifts. Detection of differences due to pitch accent is illustrated in Figure 3: The top line illustrates the above comparison for the word 'bag' in the sentence 'I did not *take* your bag' (i.e., where the word 'take' and not 'bag' is emphasized.) The comparison is between the spoken and generated waveforms' spectrograms. One can see that the spectral analysis of the two waveforms are quite similar, and this is due to the fact that the word 'bag' was not emphasized in this query. The figure showing the cross-correlation between the two spectrograms shows that the peak is close to zero, implying a relatively high correlation between the two waveforms. The bottom line of the figure illus-

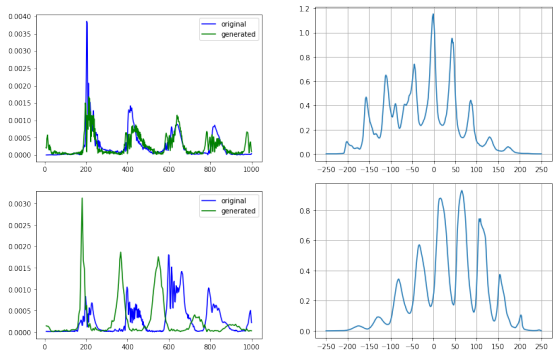

Figure 3: Comparison for the words ‘bag’ (top) and ‘take’ (bottom) in a sentence where ’take’ was emphasized. Left: original and generated waveforms in the frequency domain. Right: their cross correlation.

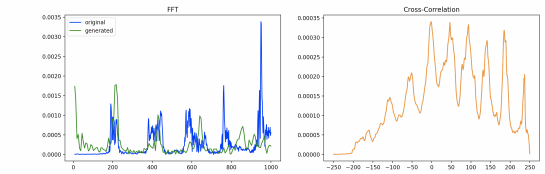

Figure 4: continuous modulation comparison. The x-axis represents the frequency measured in *Hz*, and the y-axis represents energy (left) and the normalized cross-correlation value (right) respectively.

trates the comparison of the word ‘take’ in the same sentence (where ’take’ was indeed emphasised). It is readily seen that the synthesized and spoken spectrum of the waveforms differ significantly. The corresponding cross correlation demonstrates a shift of the peak correlation of the signal by about *80Hz*, which is almost 30% of the fundamental frequency.

The general shape of these two spectrograms does not differ significantly, and so by applying a threshold on the frequency shift (around 10% of $F0$), it is possible to identify if the word was emphasized or not. For continuous modulation the cross correlation peak is below 0.05 of the total energy (see Figure 4), which was regarded as too low.

4 Implementation and experiments

As already mentioned in Section 3 above, our encoder and decoder are adapted from the SV2TTS architecture (Jia et al., 2018), which is itself based on the recurrent sequence-to-sequence Tacotron2 network (Shen et al., 2018a), extended with an attention network to support multiple speakers, similar to the scheme suggested for Deep Voice2 (Gibiansky et al., 2017).

We used the sample-by-sample autoregressive WaveNet (Oord et al., 2016) as a vocoder to invert synthesized mel-spectrograms emitted by the synthesis network into time-domain waveforms. This architecture is composed of 30 dilated convolution layers, similar to what was described in (Shen et al., 2018b). The network is not directly conditioned on the output of the speaker encoder. The mel-spectrogram predicted by the synthesizer network captures the information needed to produce a multi-speaker vocoder. To train the speech synthesis and vocoder neural networks, we used the VCTK dataset (Christophe Veaux, 2017), which contains 44 hours of speech from 109 speakers. We downsampled the audio files to 16 *kHz*, and trimmed leading and trailing silent sequences.

Our word emphasis predictor, described in Section 3 above, computes the word by word cross-correlation between the generated and original words. Since the number of samples for the generated and original words are not of the same in length, a simple linear interpolator is applied in the frequency domain.

For our experiments, we constructed a dataset of over a hundred different voice samples: five different speakers of different gender and accents recited five different sentences, each sentence with word emphasis on one of four different words (in different parts of speech). The five sentences are: (i) “I did not take your bag.” (ii) “Hello, this is our intonation project.” (iii) “There are very few black rhinos left in Africa.” (iv) “I saw her face under the hood.” (v) “Why did you give Sarah the sandwich with mustard.” The above underlined words were the ones given emphasis. These are words which seemed to us plausible as words people would want to emphasis and which may change the original meaning of the sentence.

We obtained an accuracy, precision, recall, and F1 score of 92, 89.14, 89.33, and 89.23, respectively between the ground truth and predicated results.

The project is open source and can be found online.[1] It runs as a python application with three distinct parts: (i) Configuration of a user using live or recorded voice recording. (ii) Recording a sentence from a chosen user to create a synthetic voice. (iii) Word emphasis detector - the recording is converted into text and a separation is applied. Each word is placed in a different box, with emphasized word boxes highlighted. Pressing a box

[1] https://github.com/hila-wiesel/Intonation-Project

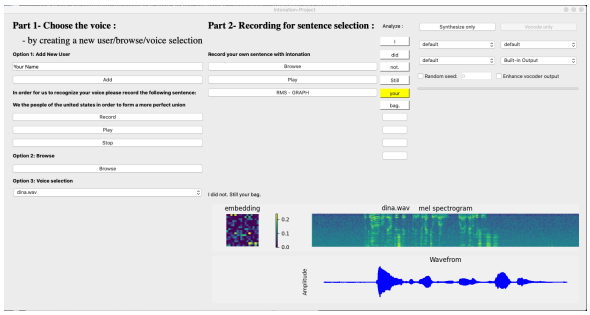

Figure 5: Control panel illustration.

will open the spectrum analysis of the original and synthesized words, alongside the cross correlation between them (See figure 3).

The control panel is shown in figure 5. The original speech in the time domain is represented in blue, the mel-spectrogram of the synthesized speech as outputted from the decoder is found above, and the embedding vector of this specific user used to generate the synthetic words is found alongside. The result highlights the word 'take' which was found to be emphasized by the algorithm in this specific example (which was correct). Videos demonstrating the use of the application are also provided.[2]

5 Conclusions and Future Work

In this paper, we presented the layout and empirical results for our word emphasis detector. This problem is especially challenging in that emphasis is affected by dialect and accent, and also different voices may differ significantly in their resonance. We introduced a novel approach using deep generative speech modules to produce an emphasis-devoid speech sample for any speaker. We used double conversion from speech to text and back to speech again. By comparing the generated and spoken voice, we were able to isolate patterns of emphasis which were relatively easy to detect.

While our focus in this work is on emphasis detection using audio features, emphasis can also be detected using the content of the text itself, for example word order, information structure, syntactic structures, and certain linguistic devices, such as repetition and contrast. Though this approach is not within the scope of our paper, it can be used to enforce the emphasis decision as its purpose.

Our experiments were based on recitation speech as opposed to spontaneous speech or dialogue, but we believe that our approach is general enough to apply to all these modalities.

For future work, we intend to use our technique to further study the effects of emphasis on the fundamental features of utterance, and expand the classification process not only to detect emphasis, but also for other speech related tasks such as, sentiment analysis, machine translation, spoken information retrieval, automated question response, and speech synthetics.

Another important direction would be to investigate the content and purpose behind the emphasised words. For example, do words adjacent to the emphasized word share certain properties? Are less common words less likely to be emphasized? This study could be used to improve emphasis detection and its analysis.

[2] https://www.youtube.com/@intonationdetection-kl7np

References

Qibing Bai, Tom Ko, and Yu Zhang. 2022. A study of modeling rising intonation in cantonese neural speech synthesis. In *Interspeech*, pages 501–505.

Kirsten MacDonald Christophe Veaux, Junichi Yamagishi. 2017. Cstr vctk corpus.

Hiroya Fujisaki. 1983. *Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing*, pages 39–55. Springer New York.

Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In *Advances in Neural Information Processing Systems*, pages 2962–2970.

Daniel Hirst. 1992. Prediction of prosody: An overview.

Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. 2019. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations, ICLR*.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

D Robert Ladd. 2008. *Intonational phonology*. Cambridge University Press.

Kun Li, Shuang Zhang, Mingxing Li, Wai-Kit Lo, and Helen Meng. 2010. Detection of intonation in l2 english speech of native mandarin learners. In *7th International Symposium on Chinese Spoken Language Processing*, pages 69–74.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Yingyong Qi and Bobby R. Hunt. 1993. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Trans. Speech Audio Process.*, 1(2):250–255.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, pages 3165–3174.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018a. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783. IEEE.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018b. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783.

R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *International Conference on Machine Learning ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4700–4709.

Paul Taylor. 1994. The rise/fall/connection model of intonation. *Speech Communication*, 15(1-2):169–186.

Paul A. Taylor. 1998. The tilt intonation model. *5th International Conference on Spoken Language Processing (ICSLP 1998)*.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *ISCA Speech Synthesis Workshop*, page 125.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5167–5176.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *IEEE Spoken Language Technology Workshop, SLT*, pages 492–498.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.