Can Phonetics Synthesize Expressive Speech?

Wanying Tian
wta55@sfu.ca
Simon Fraser University

Wenqing Liu
wla161@sfu.ca
Simon Fraser University

Shenyu Gu
shenyug@sfu.ca
Simon Fraser University

1 ABSTRACT

The text-to-speech (TTS) system is designed to convert text input into speech, enhancing accessibility for the visually impaired and aiding language learning. However, the current TTS system such as Google Text-to-Speech and Amazon Polly, produce flat and unengaging audio that audiences may not emotionally relate to. Phonetics in audio processing has the potential to convey emotions in speech. [1] In this study, we aim to develop a system that identifies emotions in texts, applies traditional TTS, and enhances speech expressiveness using phonetic modification. We then evaluate its performance with CoquiTTS, an advanced library for generating emotional speech via voice cloning. The final product not only enhances the overall user experience in chatbot interactions but also aids in speech therapy to help patients practice emotive speech. Additionally, the results have also revealed the significant impact of phonetics on emotional communication in speech.

2 INTRODUCTION

As human-computer interactions become increasingly common, the demand for emotional engagement has become apparent. The evolution of technology bridges the gap between human communication and machine interaction. Introducing the text-to-speech (TTS) system has shaped how we interact with digital devices, making information more accessible, especially for those with visual impairments, and facilitating language learning across various groups. However, despite these advancements, the lack of conveyance of emotion remains a significant challenge that affects the listener's engagement and emotional connection with the content [2].

Traditional TTS technology primarily focuses on converting text into spoken words without much emphasis on the emotional aspects conveyed by human speakers. This often results in speech that sounds flat or monotonic, lacking the emotional richness of natural human communication [2]. Emotional speech, by contrast, is characterized by variations in prosody—such as changes in pitch, volume, and speech rate—that convey feelings and attitudes. However, machines have trouble generating understandable and naturally sounding speech [2].

Inspired by identifying facial emotion through action units [3], our study aims to adapt similar principles to phonetics, exploring how subtle adjustments to phonetic elements can reflect emotional changes in speech [4]. By adjusting phonetic elements, we can mimic the way humans express emotions through pitch, volume, and rate changes.

Through this work, we contribute to the current TTS systems by advancing their capabilities beyond speech output, to communicate, connect, and resonate on a human level, enriching the user experience and making digital interactions more natural and engaging. In addition, our work delves into the relationships between phonetics and emotions communicated through speech, furthering the understanding of how audio processing can be systematically manipulated to encode affective states within synthetic speech.

3 APPROACH

Figure 1 shows the overall workflow of our system which begins with text input, such as "I didn't feel humiliated," which is analyzed by our emotion classification model to detect underlying sentiments, tagging it with an emotion like sadness. This tagged text is then transformed by a text-to-speech model, outputting a neutral voice file, which is subsequently refined by audio enhancement techniques to produce the final speech that reflects the detected emotion, resulting in a sad_voice.wav file.

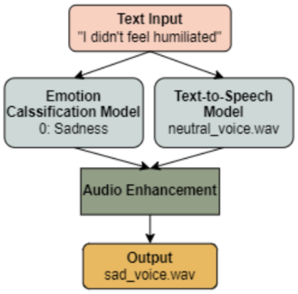


Figure 1: General workflow of system

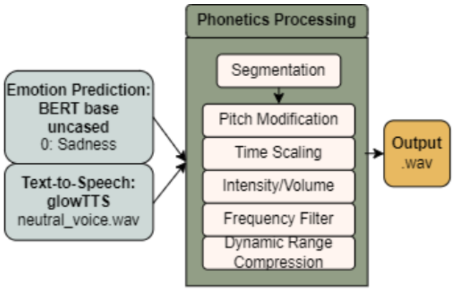


Figure 2: Detailed workflow with approaches

Figure 2 shows a more detailed view of our approaches used for each stage and summarizes the above information. Our system initiates the process with an emotion prediction model, BERT base uncased[5], pre-trained with emotional texts, to determine the underlying emotional context from the input text. This model has been trained to understand and detect emotions accurately, with an accuracy of 95% on the emotion dataset from Hugging Face. Capable of identifying a spectrum of emotions—happiness, sadness, love, anger, fear, and surprise—we have narrowed our focus to texts labeled with happiness, sadness, anger, and surprise to align with the experimental goals.

Each classified emotional text is then saved into a text file named happy_texts.txt, sad_texts.txt, anger_texts.txt, and surprise_texts.txt respectively. Following the emotional classification, these four text files are fed into a text-to-speech engine, specifically the glowTTS model, which generates a neutral baseline voice in a .wav audio format[6]. By choosing this TTS model, we can generate speech that is clear and natural due to its unique architecture that aligns text with speech on a fine-grained level. This feature avoids the robotic cadences commonly associated with earlier text-to-speech systems.

Once we had the neutral speeches for all input texts, we moved to the phonetics processing which is the core of the system. This is where we apply nuanced adjustments to reflect the identified emotion in the speech. The phonetic processing is done through a series of modifications:

- Segmentation[4]: This initial phase breaks down the speech into phonetic units, which can then be individually manipulated. We segmented the original speech into segments of 0.1 seconds.
- Pitch Modification[4]: By altering the pitch of the voice, we can synthesize the speech with the desired emotional tone. For happy speech, we increase the pitch whereas reduce the pitch to convey sadness.
- Time Scaling[4]: Adjusting the duration of phonetic units can significantly affect the emotional delivery, contributing to the perceived speed. For example, we lengthen the speech duration for sadness to mirror the slower pace often associated with this emotion.
- Intensity/Volume[4]: We modify the loudness of the speech to reflect the emotional state. They amplify the emotions of happiness, surprise, and anger, and are toned down to embody sadness.
- Frequency Filter[4]: Implementing a filter on certain frequency ranges can mimic the effect of a voice that is impacted by emotion. Each emotion affects the voice's frequency in unique ways. For instance, by filtering out higher frequencies, we can achieve the flatness that characterizes sad speech.
- Dynamic Range Compression[4]: By compressing the dynamic range, we ensure that the emotional expression is consistent throughout the speech, preventing any sudden shifts and creating a smooth auditory experience.

Audio enhancement is mainly implemented through the *librosa* library, a powerful Python package specifically designed for music and audio analysis. Furthermore, we employ the *matplotlib* library to graphically represent the mel frequency spectrums of both the original and processed audio files. The mel spectrogram is a visual representation of the spectrum of frequencies in a sound or speech signal as they vary with time. By comparing the mel spectrograms of the original and modified audio, we obtain valuable insights into how our adjustments impact the emotional tone of the speech.

As Figure 3 shows, the modified mel spectrogram, which represents a transition from neutral to sadness, shows decreased intensity in the higher frequency bands. This results in a darker visual pattern, aligning with the common perception of sad speech being softer and less varied in pitch. Additionally, the spacing between the bands in the modified spectrum is more stretched, implying a slower pace and longer pauses. These visual cues from the mel spectrograms show how altering the phonetic features can transform neutral speech into one that carries the nuances of other emotions.

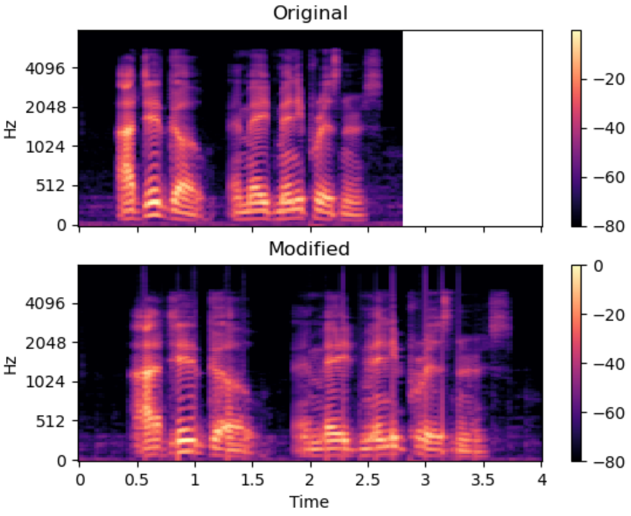


Figure 3: Mel spectrogram: before vs after audio enhancement

From the image, we can see that the time scale has been stretched, and the frequency of overall speeches has been lessened. By actually listening to the generated voice, the sound quality of the voice has been unavoidably lower as the length of the voice has been changed, more details in the piece of sound need to be filled in to maintain the sound quality. Similar to making the time of the speech compact for the happy voice. Due to the rising frequency of the sound due to the shortened sound, the speech voice will be much sharper than the original one, so the post process also needs to omit some of the waves in the speech pieces.

Out of the audio synthesis approach, we also explored deep learning to create emotionally expressive speech. Figure 4 shows the model's workflow.

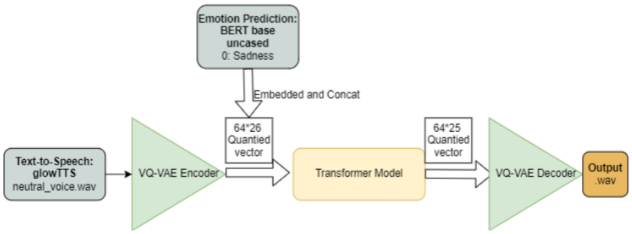


Figure 4: Model architecture

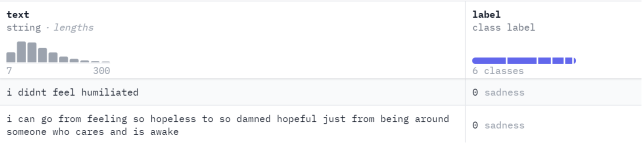
Our model consists of two main components: the VQ-VAE and the transformer. The initial voice input is transformed by the VQ-VAE into a compact form, a matrix token. This token captures the essence of the sound needed for further processing. Then, we inject emotion into the mix by attaching an emotion label that has been converted into a straightforward binary code, which we call a one-hot vector. Next up, the transformer takes over, tweaking the original sound matrix into one that embodies the chosen emotion.

For training, we fed the VQ-VAE with a rich dataset of emotional speech and a complementary text-to-speech dataset to fine-tune the autoencoder's performance. Drawing inspiration from a foundational paper on transformers[7], we tailored our model specifically for audio, simplifying some parts while boosting others, especially those dealing with the details of sound. The final step in our process was aimed at reconstructing the sound from our emotionally infused matrix.

4 DATASET

We applied 2 datasets for this emotional speech synthesis, the Emotional Speech Dataset [8] and the dair-ai/emotion text dataset[9]. The dair-ai/emotion dataset is used as a test dataset for the pre-trained text emotion detection model in the first stage of the project, ESD is used to infer phonetic feature selection for audio enhancement and to train the transformer model for the DL-based approach.

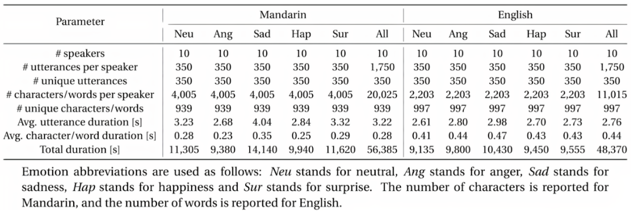
The dair-ai dataset has 416,809 sentences from Twitter and is labeled with corresponding emotions as shown in Figure 5. The length of sentences in the dataset is concentrated into 2 to 168 words. In the label distribution aspects, joy and sadness take the main part of the dataset. In our audio synthesis task, we only picked joy(as happy), sadness (as sad), anger(as angry), and surprise mapped to the next steps.



text string · *lengths* 7 300	label class label 6 classes
i didnt feel humiliated	0 sadness
i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake	0 sadness

Figure 5: Overview of the Dair-ai dataset

The ESD database consists of 350 parallel utterances spoken by 10 native English and 10 native Mandarin speakers and covers 5 emotion classes, neutral, happiness, anger, sadness, and surprise. As shown in Figure 6 there are 2 parts to emotional speech. One is for English and the other one is for Mandarin. Only the English utterances were used as the training dataset.



Parameter	Mandarin						English					
	Neu	Ang	Sad	Hap	Sur	All	Neu	Ang	Sad	Hap	Sur	All
# speakers	10	10	10	10	10	10	10	10	10	10	10	10
# utterances per speaker	350	350	350	350	350	1,750	350	350	350	350	350	1,750
# unique utterances	350	350	350	350	350	350	350	350	350	350	350	350
# characters/words per speaker	4,005	4,005	4,005	4,005	4,005	20,025	2,203	2,203	2,203	2,203	2,203	11,015
# unique characters/words	939	939	939	939	939	939	997	997	997	997	997	997
Avg. utterance duration [s]	3.23	2.68	4.04	2.84	3.32	3.22	2.61	2.80	2.98	2.70	2.73	2.76
Avg. character/word duration [s]	0.28	0.23	0.35	0.25	0.29	0.28	0.41	0.44	0.47	0.43	0.43	0.44
Total duration [s]	11,305	9,380	14,140	9,940	11,620	56,385	9,135	9,800	10,430	9,450	9,555	48,370

Emotion abbreviations are used as follows: *Neu* stands for neutral, *Ang* stands for anger, *Sad* stands for sadness, *Hap* stands for happiness and *Sur* stands for surprise. The number of characters is reported for Mandarin, and the number of words is reported for English.

Figure 6: Overview of ESD database

The limitation of the ESD dataset is that it relies on scripted sentences rather than natural speech transcriptions. This may affect the naturalness of emotion enhancement in generated speech, leading to synthesized outputs that lack the subtlety and variability of human emotions.

5 EXPERIMENTS AND RESULTS

From the two approaches used for audio enhancement: Phonetics Processing and Deep Learning, the DL-based approach did not output a clear human voice, so the evaluation focuses on the phonetics processing method. We compared the phonetics approach against a baseline model, CoquiTTS, which applies voice cloning to generate expressive speech from texts. The experiment is delivered through surveying participants, divided into 3 stages.

The first stage involves preparing voice clips. We randomly choose 5 sentences for each emotion in the research, then we have 20 sentences for evaluation for each model. The selected sentences are used as the input for our method and CoquiTTS to generate the audio of speech. In stage 2, a group of 10 participants were asked to identify the emotion in each clip and record the result of the choice to calculate the accuracy of emotion delivery for each emotion. Finally, another group of 10 participants evaluated the quality of enhancement in terms of naturalness of speech, delivery of emotion, and sound quality.

It is hypothesized that the phonetics adjustment can enhance some emotions with obvious phonetic features, such as sadness, but may potentially miss nuances of subtle emotions delivered in speech. The results of the experiments are shown in Figures 7 and 8.

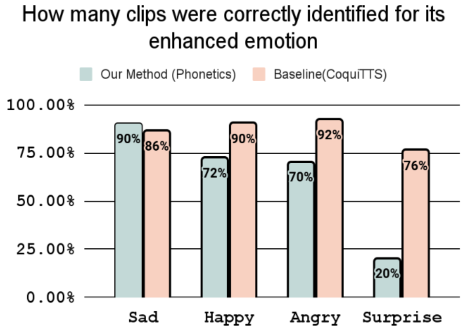


Figure 7: Correctly identify emotion from synthesized audio

Figure 7 presents a comparison between the performance of the phonetic method and a baseline method (CoquiTTS) in terms of correctly identifying enhanced emotions in audio clips. For the emotion of sadness, our system shows a 90% correctness rate, slightly higher than the baseline's 86%. However, the phonetic method only achieves a 20% correctness rate for the emotion of surprise, compared to the baseline's 76%. Also, significant adjustments in pitch and speed lower the sound quality of our method, introducing auditory discrepancies that can affect participants' choices in emotion recognition tasks.

Figure 8 illustrates ratings for synthesized audio clips, with a scoring range from 1 to 5, where 5 indicates the highest level of quality enhancement. For sadness, our method scores a 4, which indicates a high level of enhancement quality, slightly above the baseline (CoquiTTS), which scores a 3.8. As a surprise, our method received a low score of 1.2, which is notably lower than the baseline's score of 3.5. There is a clear indication that for happiness, anger, and especially surprise, the baseline method is preferred for audio quality.

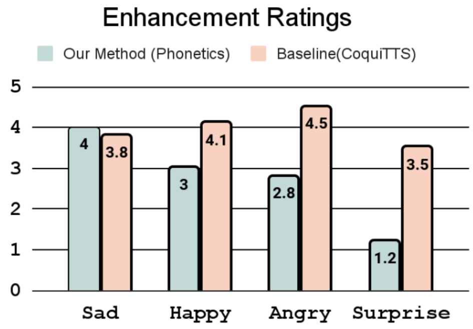


Figure 8: Rate the quality of synthesized audio

As the speech is not audible in the DL-based model output, the evaluation centers on comparing the mel-spectrum of the training outcomes. In Figure 9, the mel-spectrogram reveals a significant presence of background noise, which impedes the human ability to understand the speech content. However, the variation of the spectrogram indicates that the model is capable of implementing varied modifications, suggesting that the model has the potential to alter speech tonality. Enhancements to the model could be achieved by training on additional datasets and enhancing the discriminator component.

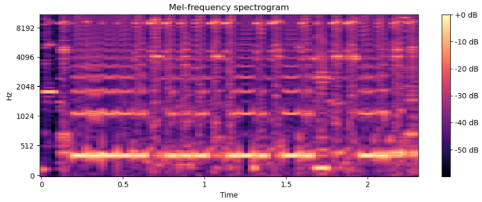


Figure 9: Mel-Spectrogram of Transformers

6 DISCUSSIONS

Our approach successfully enhances the synthesis of sad speech and outperforms CoquiTTS by 4% in performance. We achieve this by carefully adjusting the pitch and extending the duration of speech, emulating the vocal patterns observed in human expressions of sadness. The phonetic features of sadness are easy to identify, setting them apart from the other three emotions we chose.

On the other hand, most surprise clips are mistaken for angry or happy, possibly due to similar phonetic features and a lack of pitch segment selection. Both happiness and anger can exhibit rapid changes in pitch and increased volume, which are also sometimes present in expressions of surprise. This can confuse raters when they identify the emotion of a clip. Pitch segment selection involves identifying and manipulating specific parts of the pitch contour unique to each emotion. For instance, the initial pitch rise might be more pronounced in surprise, anger might show more sustained high-pitch levels, and happiness might feature high pitches that fluctuate gently.

It is also noticed that significant phonetic adjustments can lead to lower sound quality, potentially due to the time stretching and pitch adjustment on the audio. These alterations could introduce noise or degrade the naturalness of the speech, as shown by a decrease in clarity and an increase in auditory distortions. To maintain the sound quality, we can select part of the segments for adjustment to avoid impacting the whole speech or apply post-processing for signal smoothing.

The DL-based model, while capable of reproducing tones, fails to generate intelligible human-like voices. Investigations suggest adding a discriminator during training to improve sound authenticity and incorporating convolutional networks to improve signal processing. Moreover, the Emotional Speech Dynamics (ESD) dataset size is limited for deep-learning-based text-to-speech development. Using a larger speech dataset for pre-training and integrating emotional speech data for fine-tuning could enhance performance in emotional speech generation.

7 CONCLUSION

Incorporating phonetics into text-to-speech technology has shown potential for delivering expressive speech. However, modifying phonetics alone cannot represent the nuances of expressions in human speech. One expression corresponds to many phonetic changes, and different emotions can share the same phonetic. For instance, both happiness and surprise might share a higher pitch despite being distinct emotions. This issue is akin to the challenges encountered with Action Unit (AU) features in facial expression recognition, where certain AUs are common to multiple emotional states, making it difficult to distinguish between them with precision. The nuance of human expression, whether in faces or voices, thus involves more than just altering a single dimension; it requires a rich, multidimensional approach to truly resonate with the nuances of human emotions.

REFERENCES

[1] Y. Xu, "Phonetics of emotion," *Oxford Research Encyclopedia of Linguistics*, 2023. Published online: 23 August 2023.
[2] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: Disadvantages and limitations," *International Journal of Engineering and Technology (UAE)*, vol. 7, pp. 234–239, May 2018.
[3] Y. L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.
[4] S. A. Fulop, "Phonetics and speech processing," in *Encyclopedia of the Sciences of Learning* (N. M. Seel, ed.), p. 769, Boston, MA: Springer, 2012.
[5] N. Rawat, "nateraw/bert-base-uncased-emotion." Hugging Face model hub, Year of access. Accessed on: Insert date accessed.
[6] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *arXiv*, vol. 2005.11129, May 2020. Submitted on 22 May 2020 (v1), last revised 23 Oct 2020 (this version, v2).
[7] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," 2017.
[8] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2021–2021, 2021.

[9] "DAIR.AI emotion dataset." Retrieved from https://huggingface.co/datasets/dair-ai/emotion. Accessed on: Insert date accessed.

Appendix A CONTRIBUTIONS

- Wenqing Liu
  - Stage 1: Emotion classification
  - Stage 2: TTS model
  - Stage 3: Audio Enhancement
  - Poster
  - Report
- Wanying Tian
  - Stage 1: Emotion classification
  - Stage 2: TTS model
  - Stage 3: Audio Enhancement
  - Poster
  - Report
- Shenyu Gu
  - Machine Learning Model
  - Report