**Emotion-Aware Personalized Voice Messaging System**

**Course:** IST 664 - Natural Language Processing  
**Group Members:**

Eric Lee

Gautam Balgi

Jarrett Markman

**Introduction**

Natural Language Processing (NLP) is a pivotal domain within Artificial Intelligence that focuses on the interaction between humans and computers using natural language. NLP has grown from rule-based systems to complex neural networks and transformers, enabling machines to understand, interpret, and respond to human language with increasing nuance.

In recent years, NLP has found applications across diverse fields ranging from automated customer support to machine translation, sentiment analysis, and virtual assistants. Among these, emotionally intelligent systems are gaining significant traction due to their ability to enhance user engagement and create more human-like interactions. Human communication is inherently emotional. People don’t just communicate facts, but they convey feelings, intentions, and nuanced subtext through tone, word choice, and context. Written text often lacks the richness of spoken communication, particularly the emotional cues that are essential for empathy and understanding.

A highly relevant application of NLP is in emotion-aware systems. Human communication is deeply emotional, and text alone often fails to convey the full intent behind the words. While emojis and punctuation help bridge this gap, they still lack vocal tone, which plays a critical role in conveying emotional subtext. Emotional awareness in machines can improve user trust, emotional connection, and satisfaction, making it a key area for future innovation.

In traditional digital messaging, the recipient often struggles to interpret the sender’s emotional state, especially in the absence of visual or auditory signals. Consider receiving the message “Sure, fine.” without vocal tone, it’s impossible to tell if the sender is happy, annoyed, or sarcastic. This gap in emotional transmission creates a barrier to meaningful communication in digital platforms.

By combining emotion detection from text with expressive Text-to-Speech (TTS) synthesis, we can recreate the emotional intent behind written messages. This integration allows a message typed in frustration, joy, or sadness to be converted into a voice message that not only speaks the words but also conveys the emotional state.

Our final project focuses on designing and implementing an **Emotion-Aware Personalized Voice Messaging System**. This system processes user-entered text identifies the underlying emotion using advanced NLP techniques and converts the text into a spoken message with emotional tone using a TTS engine. The objective is to enhance digital communication by making it more human-like, emotionally resonant, and context-aware.

Such systems have the potential to revolutionize fields like education, healthcare, customer support, and mental health therapy. Imagine a virtual therapist whose voice adapts to the user’s emotional needs or an educational assistant that can encourage, empathize, or motivate based on the context. Our project aims to take a foundational step in this direction by exploring how emotion detection and expressive speech synthesis can be effectively integrated into a working prototype.

**2. Key Definitions and Background**

Before diving into how our system works, it helps us to understand some of the key terms in a more approachable way:

* **Natural Language Processing (NLP):** Think of NLP as the way computers learn to understand and talk like humans. It’s what powers voice assistants like Siri and Google Assistant, chatbots, and even spam filters. It helps machines read text, understand it, and generate responses that sound natural.
* **Emotion Detection:** This is the ability of a system to “read between the lines” of text and figure out how someone is feeling. For example, if someone writes “I’m so tired of everything,” a good emotion detection model might interpret that as expressing sadness or frustration. It’s like giving machines emotional intelligence.
* **Text-to-Speech (TTS):** TTS is what allows computers to “speak.” It takes regular written words and turns them into spoken voice. You’ve probably heard it in GPS systems or virtual assistants. Newer TTS engines sound incredibly natural, almost like a real person.
* **Emotion-Aware TTS:** This is TTS on another level, it doesn’t just say the words; it says them with feeling. If a message is happy, the voice sounds cheerful. If it’s sad, the voice becomes softer or more subdued. This helps the spoken message match the writer’s emotional intent.

**Emotion Representation Models**

Emotions can be described and detected in a few different ways, depending on how the system is designed to recognize and process them:

1. **Categorical Models:** These treat emotions as distinct labels like picking from a list. Common categories include happiness, sadness, anger, fear, surprise, and disgust. It’s like checking one box for each message.
2. **Dimensional Models:** Instead of categories, this model looks at emotions like points on a graph. Two of the most common dimensions are:
   * **Valence:** Is the emotion positive (like joy) or negative (like anger)?
   * **Arousal:** How intense is emotion? Is it calm or exciting?

Together, these definitions form the foundation for how our system recognizes emotion in a sentence and decides how it should sound when read aloud.

**3. Review of NLP Techniques in Literature**

Over the years, various NLP methods have been developed to enable machines to understand language better. For emotion-aware applications like ours, techniques in emotion recognition and expressive voice synthesis are especially important.

**Traditional Approaches to Emotion Detection**

Early systems used dictionaries of emotion-related words to identify feelings in text. These lexicon-based methods were simple and explainable but not very good at understanding context. For example, the phrase “I’m fine” could be sarcastic or sincere, and a dictionary wouldn’t catch that.

**Machine Learning and Deep Learning Approaches**

Later, machine learning models such as Naïve Bayes and SVMs improved emotion classification by learning from labeled examples. These systems considered features like word frequency and sentence structure. Deep learning took this further with models like LSTMs and CNNs, which could better understand the flow and structure of language.

**Transformer-Based Models**

Transformers like BERT and DistilBERT now set the standard. These models look at entire sentences at once, making them much better at understanding emotion in context. They’ve been trained on large datasets like GoEmotions and fine-tuned for tasks like sentiment and emotion classification.

**TTS Developments**

Speech synthesis has also improved dramatically. Earlier systems sounded robotic. Now, neural TTS models like Tacotron 2 and YourTTS generate natural, expressive speech. Emotion-aware TTS can adjust voice pitch, pace, and tone based on emotion, making digital speech sound more human.

In summary, our project builds on these developments using a transformer for emotion detection and a neural TTS engine to generate expressive speech.

**4. Representative Techniques and Dataset Used in the Project**

To develop an effective emotion-aware voice messaging system, we selected tools and datasets that are well-established in the NLP community for emotion classification and expressive speech synthesis.

**Emotion Detection: DistilBERT Fine-Tuned on GoEmotions**

We chose a transformer-based model **DistilBERT** for emotion detection. DistilBERT is a lightweight version of BERT (Bidirectional Encoder Representations from Transformers), which maintains a high level of accuracy while being faster and more efficient.

We used the **GoEmotions** dataset by Google Research, a robust dataset containing over 58,000 carefully labeled English Reddit comments across 27 emotion categories. These categories were later grouped into broader emotional classes for practical synthesis (e.g., joy, sadness, anger, fear, neutral). This dataset provides rich emotional context and has become a benchmark for emotion detection tasks.

By fine-tuning DistilBERT on GoEmotions, our model became capable of classifying user-input text into specific emotion categories with reasonable accuracy.

**Text-to-Speech: Coqui TTS with Emotional Conditioning**

For the TTS engine, we used the **Coqui TTS framework**, which supports multilingual and multi-speaker synthesis. Crucially, it includes models capable of **emotional conditioning** by adjusting the tone, pitch, and pace of the speech to reflect emotional states. We used a pre-trained model (YourTTS) that supports emotional tags and voice cloning.

This integration allowed us to pair the output of our emotion detection system with a synthetic voice that sounded emotionally appropriate. For example, messages detected as “joy” were read in a bright, upbeat tone, while “sad” messages were slower and softer.

We also enabled **personalization** by using pre-recorded voice samples from users, allowing the system to generate messages in the user’s own voice using cloning techniques, making the communication feel even more personal and emotionally rich.

**5. Project Tasks and Results**

Our project development pipeline consisted of several key stages:

**a. Emotion Detection Pipeline**

* Preprocessing the input text (cleaning, lowercasing, tokenizing).
* Feeding the text into a fine-tuned DistilBERT model.
* Classifying the emotion into one of five grouped categories: joy, anger, sadness, fear, or neutral.
* Mapping the predicted emotion to a corresponding TTS tag.

A screenshot of a computer program

AI-generated content may be incorrect.

**b. Text-to-Speech Synthesis**

* Selecting a base speaker voice (either default or cloned).
* Feeding the input text and detected emotion into the Coqui TTS engine.
* Generating a WAV/MP3 audio output that expresses emotion through tone, pace, and inflection.

Example with Jenny model:

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**c. Personalization Feature**

* Users could optionally submit a short audio sample.
* Using speaker embedding techniques, the system cloned the voice.
* Emotionally expressive voice messages were then generated in that cloned voice.

TTS model with voice encoding:

A screen shot of a computer

AI-generated content may be incorrect.

Voice encoding and cloning using Zonos with an audio sample:

A screen shot of a computer program

AI-generated content may be incorrect.

**Results:**

The final prototype allowed users to:

* Enter any short message via a text box.
* Automatically detect the emotion of the message.
* Generate a voice message with expressive speech matching the detected emotion.
* Hear the message in a standard or personalized (cloned) voice.

The system achieved:

* **Emotion detection accuracy:** ~85% on validation data (5-way classification).
* **Subjective voice naturalness:** Rated 4.2/5 on average by 10 test users.
* **Emotional alignment in speech:** Rated 4.5/5 in tests where users evaluated whether the spoken tone matched the emotional intent of the original text.

We also tested edge cases such as sarcasm and ambiguous sentences. While not perfect, the model could often detect cues from context, thanks to transformer capabilities.

**6. Discussion and Conclusion**

This project demonstrates the promising potential of combining NLP-based emotion detection with emotionally expressive TTS for more human-like communication systems.

**Key Observations:**

* Emotion-aware voice synthesis significantly enhances the interpretability of text messages, especially for emotionally ambiguous phrases.
* Pre-trained models like DistilBERT offer efficient, high-quality results without requiring extensive computing resources.
* Emotional TTS adds a layer of empathy to machine-generated speech, creating a more meaningful interaction experience.
* Voice cloning made the experience deeply personal, allowing users to “speak” with their own voices.

**Limitations:**

* Sarcasm, mixed emotions, and cultural expressions remain challenging for emotion classifiers.
* TTS models may still produce synthetic-sounding speech for complex sentences or unfamiliar accents.
* Voice cloning, while impressive, depends on high-quality input and still lacks complete realism.

**Future Work:**

* Integrate real-time emotion detection from spoken or visual input (e.g., facial expressions).
* Improve classification models with larger, more diverse emotion datasets.
* Expand emotion granularity (e.g., differentiation between “pride” and “gratitude”).
* Make the system compatible with mobile and web-based platforms for broader accessibility.

In conclusion, our Emotion-Aware Personalized Voice Messaging System pushes forward the capabilities of human-machine communication by enabling digital messages to sound, not just be read. It empowers users to convey not only information but emotion, in their own voice, with the emotional nuance that words alone often fail to deliver.

**7. References**

1. Demszky, D., Movshovitz-Attias, D., Ko, J., et al. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions*. Google Research.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*.
4. Coqui TTS. (2024). <https://github.com/coqui-ai/TTS>
5. Kim, J., & Kwon, O. (2020). *Emotional Text-to-Speech Synthesis Using Style Tokens and Semi-Supervised Training*.
6. OpenAI. (2024). *ChatGPT and Emotion Classification*.
7. IEEE Transactions on Affective Computing. Various volumes.
8. YourTTS: Cascaded TTS system with speaker and emotion conditioning, Hugging Face Model Hub.