Digital Signal Processing Filtering and Deep Learning for Live Captioning and Music Understanding

Andre Massanobu Shibata

***Problem Statement*— The goal of this project is to use a combination of Deep learning speech recognition algorithms, and Digital Signal Processing to be able to automatically caption in real time a conversation in between two people, being able to recognize sentences and words. Additionally, Digital Signal Processing algorithms and filters will be used to identify different musical instruments being played, and subtitles will be generated describing the musical notes being played by the different instruments in the background. In the situation that the song being played contains vocals, the lyrics of the song should also be identified and generated in real time.**

Keywords—Convolutional Neural Networks, Deep Learning, Music Understanding, Live Caption, Artificial Intelligence, Digital Signal Processing.

I. INTRODUCTION

Music is an essential part of the life of many individuals in the modern world. Not only music can be appreciated as a form of entertainment, but it also contains an immense cultural value, with it being very particular and meaningful to various cultures. The concept of music understanding, in a computing sense, is about what processes and methods can be used to make a computer understand music. Music, such as any other sound, can be broken down into a signal that can be processed into a digital environment and then manipulated.

Via the use of Digital Signal Processing, commonly abbreviated as DSP, there are numerous operations that can be performed to modify and manipulate sounds. An example of such procedures, would be to use an algorithmic version of one of a filter. With filters, such as high-pass filters, we are capable of filtering the signal, such that only high frequencies are left on the signal, with filters such as band-pass signals, we are capable of only allowing a certain range of frequencies to pass through, and many others. From the perspective of a musician, or a sound engineer, Data Signal Processing filters are essential, especially when considering common uses, such as removing background noise, making certain musical notes clearer, removing unwanted frequencies, and others.

Besides instrumentals, it is also very common for songs to contain vocals. Vocals are also signals that can be processed and manipulated via the usage of DSP algorithms. In order to correctly identify the words that are being sung, the area of speech recognition is very important. With the usage of DSP algorithms, alongside with the usage of Deep Learning, we are capable of converting spoken words and sentences into text. The idea of converting speech into text in real time is also commonly known as Live Caption, as they mean the creation of subtitles for a conversation in real time.

II. APPLICATIONS

The usage of live captions would greatly facilitate the communication between two individuals, as the presence of real time subtitles at any moment could help avoid misunderstandings due to a person mishearing a word. Additionally, it would assist those in need of hearing aids, as it would allow for spoken and auditory communication to become visual, meaning that conversations could be read instead of heard.

In addition, seeing musical notes as they are being heard, could help beginner music students, people who suffer from amusia, and people with some level of deafness, visualize and understand musical notes more easily, as they would be capable of seeing the note that is currently being played instead of depending solely on hearing it..

From a musician's perspective, with the real-time subtitling, it would become very clear if a note was played incorrectly, and the instrument that produced the irregular musical note would also be easily identified, and an action to correct the pitch played by their instrument could be quickly taken..

The real-time captions of the notes being played could also lead to live creation of sheet-music, which would save time from the musicians having to do it by themselves, and would allow musicians who depend on musical sheets to be able to play any song they want..

III. RECENT FINDINGS

Since the topic of implementing DSP and Deep learning for music understanding with an intention of captioning lyrics is very specific, a decision was made to separate the search of articles in between Music Understanding articles, and Speech recognition articles. The topics of Music Understanding and Speech recognition were chosen as they would be able to provide an overview of how computers understand music, and on how computers are capable of processing speech.

3.1 Music Understanding

When searching for research papers on the topic of music understanding, not many recent articles were found by the group, with the most relevant articles being from the early 2000s. Even though the articles found were fairly old, they helped to further my understanding of the basic fundamentals and common tools used to implement music understanding.

A keynote made by Roger B. Dannenberg[1] in the Brazilian Symposium on Computer Music in 2000, Roger argued on the importance of computers as tools for musicians, and mentioned how he believed that although Deep Learning and Artificial Intelligence, also commonly known as AI, and music as a theme are seemingly unrelated, the use of such technologies could greatly benefit the bettering of Music Understanding. To further his point, Dannenberg presented his project on computer accompaniment. Computer accompaniment systems are systems that as they "hear" musicians perform songs, the computer tries to behave in the same way as a human accompaniment, which means, it should play the same notes and tunes played by the musician, as Dannenberg described, "expressively and musically." One of the experiments Dannenberg performed with his computer accompaniment system model was to perform an ensemble, and see how the computer would react to multiple people playing at the same time. As a result, he learned that initially, the computer would not work well, as it would not be capable of deciding which musical performer to follow. He then concluded that some level of, as he described, "self-awareness" is needed for such a system to be able to successfully accompany an ensemble. In order to achieve self-awareness, machine learning and statistical approaches to musical understanding are essential.

In 2004, Masataka Goto[2] published a paper describing a real time music-scene-description system capable of accurately detect melody and bass frequencies in the signal setting of a recorded song, which should theoretically be unachievable in the real world, since musical producers usually remove unwanted noise from musical records.. Music-scene-description is to be able to describe musical signals. A problem highlighted by Goto on this topic, is the system's difficulty of identifying the fundamental frequencies of individual instruments when multiple instruments are being played simultaneously. The reason given by Goto is that there is a lot of frequency overlap between notes of different instruments. In order to achieve his goal, Goto introduces a new approach for resolving his music-scene-description problem, a fundamental frequency, which he abbreviates as F0, estimation method called PreFEst. PreFEst stands for predominant-F0 estimation method. Even though Goto did not fully explore his method's capabilities of identifying multiple sources of sound, in the discussion section of his paper, Goto contemplates the idea of investigating different source frequencies using his method. Although Goto's work is focused on the analysis of bass and melody lines that are different from musical notes, his work also provides important information on the topic of analyzing music from a Digital Signal Processing approach.

3.2 Speech Recognition

A journal article called Automated audio captioning: an overview of recent problems and challenges[3], summarizes what is the current scientific scenario for the generation of subtitles in real time. It also highlights what are current difficulties and challenges associated with the topic. The article starts by mentioning the importance of Deep Learning algorithms, such as Recurrent Neural Networks, also known as RNNs, and Convolutional Neural Networks, also known as CNNs. The authors propose a structure of speech signal processing where the signal is converted into a Mel Spectrogram, fed into an audio encoder, which extracts the features of the audio, the features then are passed through a text decoder, and captions are generated. With this model, the article discusses the different options for the audio encoders and decoders, with the encoders usually being composed of transformers, RNNs, CNNs, or Convolutional Recurrent Neural Networks, also known as CRNNs, which are a combination of RNNs and CNNs. It is also mentioned how initially, RNNs were very popular to use as acoustic encoders, but with time the trend was to shift towards the use of CNNs. For the text decoding of the audio features, the paper suggests the use of RNNs alongside word embeddings.

The article also comments on the increase in use of word embeddings for text decoding, and gives some examples of already trained word embedding models, such as Word2Vec and GloVe. Transformers have also been shown to perform very efficiently for text decoding. Furthermore, the journal talks about the common methods for training audio models, with the most popular training methods being cross-entropy training, reinforcement learning, and transfer learning. The article also gives an overview on the common evaluation metrics used to evaluate live captioning models, with the conventional evaluation metrics being: BLEU, ROUGE, METEOR, CIDEr, SPICE and SPIDEr. The paper also talks about important datasets for the topics of audio description, with the two datasets being highlighted being AudioCaps, which contains 51k audio clips that are divided into training, test, and validation sets, Clotho, which is used for official ranking for the Detection and Classification of Acoustic Scenes and Events challenges, MACS which has data from places such as airports, public areas, and parks, and AudioCaption, which has Mandarin-annotated data. The last thing that the paper goes over is the many challenges that the automated captioning faces. The current data available is listed as a problem because of how limited they are, not being able to cover all real life scenarios. Another challenge is to explore models different from the acoustic encoder to text decoder model currently being utilized. Additionally, evaluation is also considered a problem, when it is noticed that the metrics utilized can sometimes not be congruent with human judgements. This technology should also be able to replicate conversations and audio in a natural and non-machine-generated way, thus diversity and stylization of captions is also a challenge of note.

Another paper on the topic of live captioning and speech recognition is a journal article published in 2019 called An Overview of End-to-End Automatic Speech Recognition[4]. This paper proposes End-to-End deep learning models for speech recognition. Firstly, the authors talk about the complexity and difficulty of use of Hidden Markov Models, also known as HMM, and Gaussian mixed models, commonly referred to as GMM. The paper also compares different Automatic Speech Recognition, or ASR models.

The website AssemblyAI[5] provided a very useful list of free Application Programming Interfaces, also known as

APIs, and Open Source Engines that can be used for text to speech related projects. The APIs identified by the website were the AssemblyAI, Google Speech-to-Text, and AWS Transcribe APIs. Additionally, the website highlights the following Open Source Engines: Deep Speech, Kaldi, Wav2Letter, SpeechBrain, Coqui and Whisper.

Commercial Products related to the area of live captioning were also researched. Some of the identified products were live caption glasses, such as Xanderglasses[6] and XRAI Glasses[7]. Both companies present glasses capable of subtitling a conversation in real time, and have it as a goal to assist those who suffer from hearing loss. A recent article by David Copithorne[8] speculates that the market for live caption glasses will become more and more popular, with effective software becoming available for AR glasses, such as the XREAL, formerly known as NREAL, glasses, which are popular AR glasses.

IV. Next Steps

After discussing the project with Professor Hamid Nawab, I have been instructed to further my studies on the topic of Time-Dependant Fourier Transforms, as for a Digital Signal processing perspective, Time-Dependant Fourier Transforms will be very useful for understanding speech recognition.

Another important point to further is to do more research on acoustic encoders for audio feature extraction, and text decoders. In the future, more papers on these topics will be studied, and example open source projects will be looked for.

The databases presented in the article Automated audio captioning: an overview of recent progress and new challenges[3] will also be closely studied and evaluated on their possible uses. Word embeddings will also be further researched and studied. The different pre-trained word embeddings will also be analyzed and studied in order to find a fitting embedding for the project.

Further studies on the Deep Learning RNN and CNN algorithms used will be performed. The articles either gave a brief overview or a complex high level description on the implementation of the neural networks. More research will be necessary to properly understand the implementation of CNNs and RNNs as acoustic encoders and text decoders. Furthermore, a research on transformers will be performed in order for a decision to be made on whether to utilize CNNs and RNNs for the encode-decode model, or to utilize transformers for such tasks.

Because the research was divided into a segment on music understanding and a segment on speech recognition, not much intersection in between the two segments were found. Further research is needed to understand how to integrate music understanding with speech recognition in order to generate live captioning for song lyrics and for instrumentals.

It has also been noted that not many recent papers on the topic of music understanding have been found during preliminary research. Further analysis will be performed to find out if this consisted of human error, or if there is a reason on why not many recent papers on music understanding have been identified. Besides academic papers, research will be done on whether any commercial products with relevant features related to this topic have been found.

The project will also be re-evaluated to further understand what are the features of the project that matter the most when creating a Minimum Value Product, or MVP. The more the topic was researched, the true complexity of the project also became clearer, thus re-evaluation is needed to understand which parts of the topic will be focused on throughout the semester.

V. Conclusion

Artificial intelligence has been speculated to be an important aspect for implementing music understanding for many years. It is also a very important aspect for live captions and speech recognition algorithms, as the neural networks are commonly used as the base for encoders and decoders which turn the sounds into text.

Although extensive research has been done both on the topic of music understanding, and also on the topic of speech recognition and live captioning, and also although both topics contain a lot of intersecting concepts, no articles involving both topics were found. This means finding how to integrate the algorithms and knowledge from both topics will be a major challenge for this project.

VI. ChatGPT Segment

When given the prompt: “write a paper on the topic of deep learning and digital processing for music understanding and live captions”, ChatGPT output the following:

Title: Deep Learning and Digital Processing for Music Understanding and Live Captions

Abstract **—Music is a universal language that has the power to convey emotions, stories, and cultural expressions. As technology continues to advance, the intersection of deep learning and digital signal processing has opened up exciting possibilities for enhancing music understanding and accessibility. This paper explores the synergies between deep learning techniques and digital processing in the context of music, with a specific focus on live captions. We discuss the current state of the field, recent advancements, and the potential impact on music enthusiasts and individuals with hearing impairments.**

1. Introduction*:*

Music is a fundamental form of human expression, cherished for its capacity to evoke emotions and communicate ideas. However, not everyone can fully enjoy the beauty of music, especially those with hearing impairments. Live captions, the real-time conversion of spoken or sung words into text, have been a vital tool for making music accessible to a wider audience. Traditional approaches to live captioning often rely on human transcriptionists, but they can be costly, error-prone, and limited in scalability.

Deep learning, a subset of artificial intelligence, has revolutionized various fields, including computer vision and natural language processing. In recent years, researchers and engineers have started to harness the power of deep learning in combination with digital signal processing (DSP) techniques to create innovative solutions for music understanding and live captions.

2. Deep Learning for Music Understanding

2.1 Music Genre Classification:

Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown remarkable results in automatically classifying music genres. These models analyze audio features like spectrograms, mel-frequency cepstral coefficients (MFCCs), and chroma features to categorize music into genres. This capability can enhance music recommendation systems and playlist generation.

2.2. Music Emotion Recognition:

Understanding the emotional content of music is a complex task, but deep learning models have made significant strides in this area. Sentiment analysis of lyrics combined with audio features allows systems to automatically determine the emotional tone of a song, providing valuable insights for music therapy and personalized music recommendations.

2.3. Music Transcription:

Deep learning has also been applied to music transcription, where it converts audio signals into musical notation. This can be beneficial for musicians, as it automates the process of converting recorded music into sheet music, making it easier to analyze and learn.

3. Digital Processing for Live Captions

3.1. Automatic Speech Recognition (ASR):

ASR systems, powered by deep learning, have become essential for live captions in music performances. These systems can accurately transcribe lyrics and spoken commentary in real-time. They utilize recurrent neural networks with attention mechanisms to adapt to various accents, dialects, and music genres.

3.2. Noise Reduction and Signal Enhancement:

Live music events often have challenging acoustic conditions, including background noise and reverberation. DSP techniques, combined with deep learning, can help improve the quality of transcriptions by reducing noise and enhancing the clarity of the audio signal.

3.3. Multimodal Approaches:

To provide a comprehensive live captioning experience, combining audio and video processing with deep learning can generate synchronized captions, highlighting musical nuances, facial expressions, and stage dynamics to offer a more immersive experience to all viewers.

4. Challenges and Future Directions

While deep learning and digital processing hold immense promise for music understanding and live captions, several challenges remain:

1. Privacy concerns related to audio and video data processing.
2. The need for larger and more diverse datasets to improve model accuracy.
3. Addressing bias and cultural nuances in music understanding and transcription.
4. Real-time processing requirements for live events.
5. Future directions in this field include exploring generative models for automatic music composition, creating AI-driven music therapy tools, and developing more accessible and affordable live captioning solutions.

5. Conclusion

The fusion of deep learning and digital signal processing has the potential to revolutionize music understanding and live captions, making music more accessible to all. As technology continues to advance, it is essential to address the challenges and ethical considerations while pushing the boundaries of what is possible in the realm of music accessibility.

This interdisciplinary approach not only benefits individuals with hearing impairments but also enhances the music experience for everyone, fostering inclusivity and preserving the universal language of music.

VII. ChatGPT Analysis

The only alterations to the ChatGPT response were formatting to ensure that the paper written by the chatbot fit with the rest of the paper.

One of the main differences between the paper written by the author and the paper written by ChatGPT is the fact that the paper written has a focus on researches and findings on the topics of music understanding and live captions, including commercial products that use such concepts, and useful resources, such as sound databases, APIs, and free relevant open source engines, while ChatGPT provided a paper listing and explaining relevant topics and concepts that are very important to properly understand what music understanding and live captioning are, and how deep learning and Digital Signal Processing can assist with those topics. This means that ChatGPT was not capable of identifying recent findings on the subjects, and instead, explained them.

VIII. REFERENCES

[1] R, Dannenberg, “Artificial intelligence, machine learning, and music understanding”, VII Brazilian Symposium on Computer Music, Curitiba, Brazil, 2000.

[2] M. Goto, “A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals,” *Speech Communication*, vol. 43, no. 4, pp. 311–329, Mar. 2004. doi:10.1016/j.specom.2004.07.001

[3] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, “Automated audio captioning: An overview of recent progress and new challenges,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, 2022. doi:10.1186/s13636-022-00259-2

[4] D. Wang, X. Wang, and S. Lv, “An Overview of End-to-End Automatic Speech Recognition,” Symmetry, vol. 11, no. 8, p. 1018, Aug. 2019, doi: 10.3390/sym11081018.

[5] K. Foster, “The top free speech-to-text apis, AI models, and Open Source Engines,” News, Tutorials, AI Research, https://www.assemblyai.com/blog/the-top-free-speech-to-text-apis-and-open-source-engines/ (accessed Sep. 17, 2023).

[6] “Xanderglasses,” Xander, https://www.xander.tech/xanderglasses (accessed Sep. 17, 2023).

[7] “About Us,” XRAI Glass, https://xrai.glass/about/ (accessed Sep. 17, 2023).

[8] D. Copithorne, “Captioning glasses you can wear all day,” Hearing Tracker, https://www.hearingtracker.com/news/captioning-eyeglasses-that-can-be-worn-all-day (accessed Sep. 17, 2023).