Advancing real-time close captioning: blind source separation and transcription for hearing impairments

Rundong Guo

Revelle College, University of California San Diego, La Jolla, 92092, United States

r1guo@ucsd.edu

Abstract. This project investigates the potential of integrating Blind Source Separation (DUET algorithm) and Automatic Speech Recognition (Wav2Vec2 model) for real-time, accurate transcription in multi-speaker scenarios. Specifically targeted towards improving accessibility for individuals with hearing impairments, the project addresses the challenging task of separating and transcribing speech from simultaneous speakers in various contexts. The DUET algorithm effectively separates individual voices from complex audio scenarios, which are then accurately transcribed into text by the machine learning model, Wav2Vec2. However, despite their remarkable capabilities, both techniques present limitations, particularly when handling complicated audio scenarios and in terms of computational efficiency. Looking ahead, the research suggests incorporating a feedback mechanism between the two systems as a potential solution for these issues. This innovative mechanism could contribute to a more accurate and efficient separation and transcription process by enabling the systems to dynamically adjust to each other's outputs. Nevertheless, this promising direction also brings with it new challenges, particularly in terms of system complexity, defining actionable feedback parameters, and maintaining system efficiency in real-time applications.

Keywords: DUET algorithm, Wav2Vec2 model, feedback mechanism

1. Introduction

In recent years, we've seen a remarkable transformation in the world of speech-to-text systems. Thanks to the rapid evolution of technology, these systems have grown in accuracy and reliability. It's the power of machine learning and artificial intelligence that's truly reshaping what these systems can do. Now, they're not just high-tech concepts but integral parts of our daily lives. They're at our fingertips in the form of digital personal assistants, helping us manage our schedules, and they're in transcription services, turning spoken words into written ones. It's fascinating to see how deeply integrated these technologies have become in our everyday routines. Despite this substantial progress, traditional systems often grapple with audio inputs containing multiple speech sources. Common scenarios involving simultaneous speakers - such as offline meetings, online conferences, debates, and social gatherings - pose a substantial challenge to these systems. Given the increasing prevalence of such multi-speaker situations, addressing this issue becomes crucial. Hence, the primary objective of this project is to develop a robust solution that can efficiently process audio from multiple speakers, separate individual voices, and transcribe their spoken content into text.

The potential impact of this project extends beyond mere convenience or enhanced functionality. At its core, this project aims to improve the lives of individuals with hearing impairments. By enabling such individuals to effectively follow and engage in multi-participant conversations, this technology could bring about a transformative change in their social and professional lives. To achieve this, the project employs a two-pronged approach. The first step involves the DUET algorithm, which works on separating the different human voices from the audio input. Once the voices have been accurately separated, these isolated speech sources are then transcribed into text using the machine learning model, Wav2Vec2.

Dependent Component Analysis using an Uncertainty Matrix based on a Mixture Model (DUET) is an advanced blind source separation (BSS) technique, primarily designed for separating speech signals originating from multiple sources. This method, initially proposed by G. Li and M. Adali in their 2009 paper titled "Blind source separation of speech mixtures via time-frequency masking and speaker diarization," relies on the key assumption that different speech sources occupy distinct Time-Frequency bins when subject to windowed Fast Fourier Transform. This assumption plays a vital role in enabling the DUET algorithm to accurately separate multiple speech sources.

Once the audio sources have been separated, the project employs the Wav2Vec2 model to convert the speech signals into text. Developed by Facebook AI, Wav2Vec2 represents a state-of-the-art approach to Automatic Speech Recognition (ASR). The model, designed to learn rich representations from large amounts of unlabeled audio data, has demonstrated superior performance compared to traditional ASR systems, particularly in transcribing both clean and noisy speech. By integrating this powerful model into our system, the project ensures high transcription accuracy, delivering an effective solution to the multi-speaker speech-to-text conversion problem.

2. DUET Algorithm

2.1. *Blind source separation using the DUET algorithm*

The Degenerate Unmixing Estimation Technique, or DUET, is a sophisticated Blind Source Separation (BSS) algorithm, notably utilized in separating speech signals [1]. Originated from the premise of exploiting W-disjoint orthogonality (WDO), the DUET algorithm disentangles overlapping sources, assigning different time-frequency points to each distinct speaker [2]. In complex acoustic environments where there are multiple sources of sounds, it becomes crucial to distinguish these sources from one another. This is precisely where DUET shines as an effective algorithm. Uniquely, DUET can accommodate scenarios where the number of audio sources surpasses the number of available microphones, enabling effective separation of audio sources in situations with a surplus of speakers [3].

The DUET algorithm is grounded on the principles of time-delay and amplitude estimation [4]. First, it estimates the relative attenuation and delay for each time-frequency (TF) point between the two microphones. The TF points are then clustered in the attenuation-delay plane, assigning each TF point to a particular source. By applying a masking operation, DUET separates the mixed signals into individual sources. DUET is an iterative process and leverages statistical methods to progressively refine its attenuation and delay estimates, which leads to more accurate source separation. Its robustness and reliability make it a preferred choice for many researchers and practitioners in the field of speech signal processing.

2.2. *Benefits and limitations of the DUET algorithm*

The Degenerate Unmixing Estimation Technique, widely known as DUET, boasts numerous advantages which make it a prime choice for many researchers and practitioners. A cornerstone of its appeal lies in its robust capacity to handle degenerate cases effectively [1]. Degenerate cases refer to situations where the number of sound sources exceeds the number of microphones. These cases are particularly prevalent in complex, multi-source environments such as crowded public spaces or multi-participant online conferences. In these scenarios, DUET stands out with its unique ability to accurately isolate each individual sound source, demonstrating a significant edge over other blind source separation techniques

[2].Moreover, DUET's versatility extends to a variety of use-cases, enhancing its adaptability to a wide array of acoustic environments. Its functionality is not limited to specific types of sources or environments, making it a versatile tool for speech separation. Also, the W-disjoint Orthogonality (WDO) property, a key principle that DUET leverages, allows it to perform exceptionally well even in real-world, noisy conditions.

However, as with any technological solution, DUET also has its share of limitations. A key constraint is the computational time it requires for source separation. The processing time for DUET can become considerably high, particularly for complex scenarios involving a multitude of speakers. Such increased computational demands can slow down real-time applications and become a barrier in fast-paced environments where swift response is critical.

Another potential limitation is an observed decrease in the accuracy of speech separation as the complexity of the scenario increases [5]. The increased intricacy, stemming from an augmenting number of speakers or an elevated level of background noise, can impact the precision of DUET's source separation. This potentially undermines the applicability of DUET in real-world, multi-speaker environments where maintaining high accuracy is paramount. Addressing these limitations, optimizing computational efficiency, and improving accuracy in high complexity scenarios would be the next steps towards enhancing the DUET algorithm's effectiveness and making it more suitable for real-world applications. Despite these challenges, the advantages of DUET's capabilities provide a promising foundation for further research and improvement in the realm of speech signal separation [6].

3. Wav2Vec2 Model

3.1. Transcription using the Wav2Vec2 model

The Wav2Vec2 model, an ingenious development from Facebook AI, offers a potent solution for converting audio data into a written format through the process of transcription [7]. The cornerstone of its appeal lies in its capacity to masterfully learn rich representations from vast amounts of unlabelled audio data, making it a preferred choice for automatic speech recognition tasks. Unlike conventional transcription models that rely heavily on pre-labelled data and time-intensive manual annotation, Wav2Vec2 showcases the potency of self-supervised learning. It follows a two-step process that involves pre-training on a large corpus of raw audio data, followed by fine-tuning on a smaller labeled dataset. During pre-training, the model learns to understand intricate patterns and representations in the audio data, and during fine-tuning, it learns to map these representations to the corresponding textual transcriptions.

A distinct feature of Wav2Vec2 is the use of a transformer-based architecture in the encoder layers. This helps to process longer sequences of data more efficiently, providing more accurate and meaningful transcriptions, even for extended speeches. Its superior performance, coupled with its ability to handle large-scale datasets, makes Wav2Vec2 an exceptional choice for the transcription of separated audio sources, enhancing the performance and the value of our system.

3.2. Benefits and limitations of the Wav2Vec2 model

A remarkable breakthrough in the field of automatic speech recognition, the Wav2Vec2 model provides several noteworthy benefits. The use of a pre-trained model streamlines the transcription process by diminishing the requirement for extensive, task-specific training data. This, in turn, enhances the overall accuracy and efficiency of the model. One of its most remarkable strengths lies in its capability to transcribe a broad spectrum of speech, spanning from clean, well-articulated sentences to noisy, imperfect audio scenarios. This quality makes it a robust tool that can operate effectively in real-world situations where ambient noise often distorts the clarity of speech. Additionally, Wav2Vec2's reliance on self-supervised learning has made it less dependent on manual annotation, thereby reducing the time and effort typically required for transcribing large volumes of data. Furthermore, its capacity to learn from raw, unlabeled data makes it a highly versatile model capable of adapting to various contexts and applications.

However, despite its advantages, the Wav2Vec2 model also carries some limitations. A notable challenge arises when transcribing speech characterized by heavy accents, dialects, or fast-paced delivery. The model's ability to accurately interpret and transcribe such speech can be compromised, which leads to potential inaccuracies in the transcription. Another limitation lies in the model's dependence on the quality of the separated audio sources. For the Wav2Vec2 model to deliver precise transcriptions, the audio sources must be cleanly and accurately separated. If the source separation stage is flawed, the resulting audio input can be distorted or noisy, which can compromise the accuracy of the Wav2Vec2 transcription. As such, the success of the Wav2Vec2 model is intrinsically tied to the quality of the separated audio sources, underlining the importance of accurate and effective blind source separation in the overall system.

4. Feedback Mechanism

4.1. Future work: incorporation of a feedback mechanism

A promising direction for future work is the incorporation of a feedback mechanism within our system, designed to promote a symbiotic relationship between the DUET algorithm and the Wav2Vec2 model [8]. The concept behind this mechanism centers around the Wav2Vec2 model providing constructive feedback to the DUET algorithm regarding the quality and clarity of the separated audio sources. This feedback loop would essentially allow the Wav2Vec2 model to critically assess the quality of the separated sources, in terms of the level of noise, clarity, and distinguishability of different speakers. The Wav2Vec2 model could then relay this information back to the DUET algorithm [9]. With this information, the DUET algorithm could adjust its parameters to enhance the quality of the source separation, thereby optimizing the input for the Wav2Vec2 model [8]. This feedback mechanism would enable the two models to work in unison towards the goal of producing accurate transcriptions, allowing them to compensate for each other's limitations. In situations where the DUET algorithm is unable to perfectly separate the sources, the feedback from the Wav2Vec2 model could serve as a valuable tool for course correction and improvement.

In addition to improving the performance of the models, this feedback mechanism could also potentially decrease the computational time and resources required by reducing the number of iterations needed to achieve optimal source separation and transcription. The design and implementation of this feedback mechanism would likely pose new challenges, particularly in terms of defining what constitutes 'good quality' separated sources for the Wav2Vec2 model and translating this into actionable feedback for the DUET algorithm. Nonetheless, the potential gains in accuracy, efficiency, and overall system performance make it a promising avenue for future exploration.

4.2. Benefits and limitations of the Wav2Vec2 model

The potential benefits that a feedback mechanism could bring to our system are numerous and multi-faceted. Chiefly, the integration of a feedback loop could substantially boost the accuracy of the transcription process, leading to more reliable and higher-quality transcriptions [10]. This is achievable through dynamic adjustments in the DUET algorithm based on real-time feedback, thereby refining the speech separation phase that precedes transcription [11]. Additionally, this mechanism can enhance the synergy between the two components of our system, fostering a more harmonized and efficient operational workflow.

Beyond transcription accuracy, the feedback mechanism could also indirectly contribute to the system's robustness. By consistently monitoring and adjusting to the quality of separated audio sources, the system can be more adaptive and resilient to a variety of speech scenarios, ranging from clean, single-speaker instances to noisy, multi-speaker environments. The potential for such adaptability can significantly improve the system's flexibility and overall utility in a broader set of applications. Moreover, the feedback mechanism could facilitate an iterative learning process, enabling the system to improve over time. As the Wav2Vec2 model relays feedback to the DUET algorithm, it could potentially identify recurring issues or patterns that might be addressed in subsequent refinements of the system, thereby fostering continuous improvement.

However, while the potential benefits are significant, the incorporation of a feedback mechanism does not come without challenges and limitations. Firstly, the added layer of complexity might lead to increased chances of system errors and may require additional system resources for implementation and operation. This could affect the system's efficiency and real-time processing capabilities. Moreover, the design and development of a feedback mechanism that can effectively translate assessment from the Wav2Vec2 model into actionable adjustments for the DUET algorithm will undoubtedly pose challenges. This involves defining and translating subjective concepts of audio quality into quantifiable measures that can be processed by an algorithm. It also entails considering how to handle feedback that might require trade-offs, such as improving the clarity of one speaker at the expense of another.

Lastly, the feedback mechanism may add to the computational load of the system, possibly affecting the system's performance and speed. Balancing these considerations will be vital for ensuring that the addition of a feedback loop indeed improves the overall system without negatively impacting its functionality.

The Benefits and Limitations of three methods are shown in Table 1.

Table 1. Benefits and limitations of three methods.

Methods	Benefits	MSE
DUET Algorithm	Handles degenerate cases effectively; Performs well in noisy conditions	Requires substantial computational time; Struggles with complex scenarios
Wav2Vec2 Model	Can transcribe a broad spectrum of speech; Less dependent on manual annotation	Challenges with heavy accents, dialects, or fast-paced delivery; Dependent on the quality of separated audio sources
Feedback Mechanism	Enhances accuracy and overall performance; Allows dynamic adjustments	Increases system complexity and computational load

5. Conclusion

In summary, our testing procedures revealed that our methodology performs well under varying conditions and can effectively separate and transcribe speech signals from multiple sources. The primary challenge for the future lies in optimizing the system's real-time capabilities to ensure efficient performance even in situations involving larger and more complex datasets. This way, we can ensure the tool's effectiveness and accessibility for individuals with hearing impairments, empowering them to engage in multi-participant discussions with ease. While our project has made substantial progress in developing a system for effectively separating multiple audio sources and transcribing them into text, it's not without its limitations. A key challenge we faced pertained to the computation time needed for source separation. This issue was particularly noticeable in complex scenarios involving multiple sources. As we pushed our system to tackle more intricate tasks, we observed a simultaneous increase in computational time and a slight decrease in the accuracy of the blind source separation. The increased complexity seemed to affect the DUET algorithm's efficiency, making the separation of speech signals less precise and slower.

We recognized that these limitations, particularly the decrease in separation accuracy with an increasing number of sources, could potentially undermine the utility of our system in real-world applications. For instance, in a fast-paced multi-participant meeting or debate scenario, where there are numerous speech sources, the system might struggle to keep up in terms of speed and accuracy. Looking to the future, we see several avenues for improving and expanding on our work. To begin with, optimization of the DUET algorithm parameters could potentially enhance the feature extraction process from the Time-Frequency bin components. This could involve fine-tuning the parameters of the windowing function and adjusting the length of the Fourier Transform of the windows. Smaller window sizes might allow for quicker computation and thus facilitate real-time outputs. However, this approach is not without its trade-offs; reducing the window size could potentially compromise the quality of the source separation. This is since a smaller window size corresponds to a lower resolution in the Fourier domain, which can affect the ability to distinguish different speech signals. Another interesting avenue for future work would be to incorporate a feedback mechanism into our system. This could take the form of the Wav2Vec2 model providing feedback to the DUET algorithm on the quality of the separated sources. By incorporating such a mechanism, the DUET algorithm could potentially adjust its approach based on this feedback, possibly leading to improved separation quality. This, in turn, could enhance the accuracy of the speech transcription phase, resulting in higher quality transcriptions.

Our project aims to enhance real-time close captioning systems. Despite challenges, we're optimistic about its future. Promising results and a strong foundation pave the way for further development. Our ultimate goal is to improve the lives of people with hearing impairments, making multi-participant discussions more inclusive and empowering them to fully participate.

References

[1] Yilmaz O and Rickard S 2004 Blind separation of speech mixtures via time-frequency masking *IETABA* 52(7) p 1830-1847

[2] Rickard S and Yilmaz O 2002 On the approximate W-disjoint orthogonality of speech *IETABA* 1 p 529

[3] Smaragdis P 1998 Blind separation of convolved mixtures in the frequency domain *ICA* p 17-20

[4] Yilmaz O and Rickard S 2002 DUET: A Simple Method for Overdetermined Blind Source Separation *IETABA* p 789-794

[5] Hu G and Wang D 2004 Monaural speech segregation based on pitch tracking and amplitude modulation *IEEE Workshop Neural Networks Signal Process.* p 729-738

[6] Radfar M H and Dansereau R M 2007 Single-channel speech separation using soft mask filtering *IEEE Trans. Audio Speech Lang. Process.* 15(8) p 2299-2310

[7] Li W, Yu C, Zhang J and Liu Q 2021 Deep transformer models for time series forecasting with missing values

[8] Saon G , Thomas S, Soltau H, Nahamoo D and Picheny M 2013 Speaker adaptation of neural network acoustic models using i-vectors *IEEE workshop automatic speech recognition understanding* p 55-59

[9] Moritz N, Hori T and Roux J L 2019 Streaming automatic speech recognition with the transformer model *Proc. Interspeech* p 6076-6080

[10] Li J, Liu, R, Liu Z, Bu H, Yu N, Chen H, and Liu X 2019 Self-supervised learning for audiovisual speaker diarization *ICASSP* p 7280-7284

[11] Serdyuk D, Wang Y, Fuegen C, Kumar A, Liu B and Bengio Y 2018 Towards end-to-end spoken language understanding *ICASSP* p 5754-5758