ISSN: 2508-7894
KJAI website: http://accesson.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2025.13.1.27

Dynamic Subtitle Allocation: A Practical System with Speaker Separation

Hyun-Seok CHOI[1], Hyuk-Hee HAN[2]

Received: December 11, 2024. Revised: February 24, 2025. Accepted: March 05, 2025.

Abstract

Subtitles help viewers understand the contents of a video by displaying spoken words or their translations. However, when multiple speakers speak simultaneously, it becomes difficult to determine who is speaking through subtitles alone. Additionally, most subtitles are positioned below the scene, which can disrupt visual focus. To address these issues, we develop a dynamic subtitle allocation pipeline that positions subtitles near the speaker's face. A speaker's face is cropped using YOLOv3-based face detection, followed by a lip-reading model associating ground-truth subtitles with the correct speaker. Experimental results show that our method effectively assigns subtitles to the corresponding speakers, even in multi-speaker scenarios, demonstrating its robustness in handling complex audio-visual conditions. This practical approach enhances accessibility, particularly for viewers with hearing impairments, and improves the viewing experience by ensuring seamless subtitle synchronization with speakers. Furthermore, our system operates with standard single-channel audio and requires no additional pre-processing, making it easy to integrate into existing content and deploy in real-world applications.

Keywords: Dynamic Subtitle Allocation, Multi-Speaker Conversations, Lip-Reading Model

Major Classifications: Artificial Intelligence, Vision and Speech Recognition, Accessibility and Assistive Technologies

1. Introduction

Video subtitles enhance accessibility by visually representing spoken dialogue or translations. It can also build up the storytelling and make videos more engaging. However, current subtitles are located statically at the below part or the upper part of the frame for the entire video. The current static subtitle position has several disadvantages. First, it cannot reflect the location of the speaker. This problem can be exacerbated in scenes where multiple people are speaking simultaneously. Secondly, it compels the viewer to focus on a fixed point in the frame that is distant from the primary scene and characters, diverting their attention from the main action. Furthermore, subtitles displayed in fixed positions may obscure critical visual information in the video when they overlap with key elements of the scene.

* This research was supported by the Ministry of Science, ICT(MIST), Korea, under the Project-based AI Talent Fostering Program (RS-2022-00143911) supervised by the Institute for Information Communications Technology Planning Evaluation (IITP).

[1] First Author, PhD Candidate, Graduate School of Convergence Science and Technology, Seoul National University, Korea
Email: guardian1248@snu.ac.kr

[2] Corresponding Author, PhD Candidate, Department of Management Engineering, KAIST (Korea Advanced Institute of Science and Technology, Korea).
Email: hanheukhee@gmail.com

Figure 1: Example of a Static Subtitle in Current Use

Figure 1 illustrates a typical example of subtitles currently in widespread use. In this example, the dialogue of two speakers is merged into a single subtitle, making it difficult to discern which line belongs to which speaker. This ambiguity can distract viewers' attention from important non-verbal cues, such as the speaker's facial expressions, gestures, or the background scenery. Additionally, to enhance the readability of the white text, a black background has been added to the periphery of the subtitle's letters. However, this design inadvertently obscures part of the video frame, impairing the viewer's ability to recognize objects. For instance, in this scene, the subtitles obscure a cup.

To address these issues, we propose a dynamic subtitle allocation pipeline that can be applied to any video featuring subtitles. We position subtitles prominently near the speaker to highlight the scene's more natural and intrinsic aspects. Unlike traditional methods that rely on speech detection or speaker identification (Akahori et al., 2017; Tapu et al., 2019), our model works with standard single-channel audio, making it widely applicable to existing video libraries. Furthermore, by using visual information as the key pivot, our model accurately matches subtitles to their respective speakers, even in situations where multiple individuals are speaking simultaneously.

The rest of the paper is organized as follows. Section 2 discusses related works, highlighting the existing techniques for face detection and visual speech recognition that form the foundation of our approach. Section 3 details the methodology, including the processes of cropping single-speaker videos, applying lip-reading models, mapping subtitles to speakers, and addressing challenges such as subtitle fluctuations and length. Section 4 presents the results, comparing our system's performance with ground-truth subtitles and providing accuracy metrics. Finally, Section 5 concludes and explores potential areas for future work to enhance the value and practical applicability of our research.

2. Related Research

We utilize several deep-learning-based models with pre-trained weights in our system. In this section, we briefly introduce these models.

2.1. Face Detection and Face Recognition

Face detection involves locating human faces in input images, which is crucial for dynamically assigning subtitles near the speakers. Several modern models have emerged to tackle the challenges of real-time and robust face detection.

Zhang et al. (2016) introduced MTCNN, a multi-task framework that integrates face detection with facial alignment. Its pipeline consists of three cascaded stages, each refining the detection results while simultaneously predicting facial landmarks. This method significantly improved detection accuracy in challenging scenarios, such as non-frontal poses and varying lighting conditions.

Liu et al. (2016) marked a significant shift toward single-stage object detection models with SSD (Single Shot MultiBox Detector). Unlike MTCNN, SSD processes an input image in a single forward pass and predicts bounding boxes at multiple scales, improving both speed and accuracy.

Redmon and Farhadi (2018) builds YOLOv3 on the strengths of these earlier models, optimizing for both speed and accuracy through a single-stage object detection framework. YOLOv3 achieves this by using multi-scale predictions, enabling effective detection of faces at various sizes, and by reusing features across the network to reduce computational overhead. YOLOFace, a specialized adaptation of YOLOv3 for face detection, inherits these strengths and further optimizes the architecture for detecting faces specifically. Compared to these earlier models, YOLOFace offers:

1. Speed: YOLOFace operates in a single forward pass, processing images significantly faster than multi-stage approaches like MTCNN, making it ideal for real-time applications.
2. Simplicity: Unlike MTCNN, which requires separate models for each stage, YOLOFace uses a unified network, reducing complexity in both implementation and deployment.
3. Scalability: YOLOFace effectively handles diverse resolutions and configurations, making it more adaptable than SSD in challenging environments.

By leveraging YOLOFace, our system achieves efficient and accurate face detection, making it better suited for

dynamic subtitle allocation in real-time video content compared to both traditional and recent approaches.

Face recognition, on the other hand, is the task of identifying and distinguishing individual faces by learning unique facial features and matching them to previously encountered data. This capability enhances the robustness of the model by ensuring subtitle continuity, even as the speaker moves within the frame or between frames. In our work, face recognition plays a critical role in maintaining subtitle alignment with the correct speaker over time, even in multi-speaker scenarios.

The field of face recognition traces its roots to early template-matching methods. Kanade (1973) pioneered the use of computational approaches for face recognition with his work on an automated system that utilized facial geometry. This approach, while innovative, was limited by its dependency on well-defined and controlled environments. Subsequent research in the 1990s shifted towards eigenfaces, a statistical method developed by Turk and Pentland (1991), which applied principal component analysis (PCA) to represent faces as linear combinations of eigenvectors. This method marked a significant milestone, offering a practical solution for real-world face recognition systems.

Building on these foundations, modern approaches leverage deep learning and convolutional neural networks (CNNs) to achieve remarkable accuracy. Kim et al. (2017) explored the applicability of machine learning models such as TF-SLIM and Inception-V3 for face recognition, demonstrating the effectiveness of CNN algorithms in achieving high accuracy in visual data analysis. Similarly, Son and Lee (2017) highlighted the application of CNN-based face recognition in various scenarios, emphasizing its robustness and adaptability in addressing challenges such as varying lighting conditions, facial expressions, and occlusions. Inspired by these works, we employ advanced face recognition techniques that leverage CNN algorithms to enhance the reliability and accuracy of subtitle assignment in dynamic video content.

2.2. Visual Speech Recognition

Visual Speech Recognition (VSR), also known as lip reading, identifies speech content based on lip movements, even in the absence of audio features. This technology is particularly valuable in noisy environments or scenarios where multiple people are speaking simultaneously, as it enables robust speech detection without relying on clear audio input.

The development of VSR has been significantly advanced by deep learning. Chung et al. (2017) introduced "Watch, Listen, Attend, and Spell" (WLAS), a groundbreaking model that integrated audio and visual modalities to improve transcription accuracy. This approach demonstrated the importance of attention mechanisms in aligning lip movements with spoken words. Following this, Assael et al. (2016) introduced "LipNet," an end-to-end model capable of performing sentence-level lip reading. LipNet bypassed the need for manual feature engineering by directly mapping sequences of lip movements to textual outputs, setting a new benchmark for VSR systems.

Building on these advancements, large-scale datasets like LRS3 (Afouras, Chung, & Zisserman, 2018), containing over 400 hours of labeled audiovisual data, have enabled the training of even more sophisticated models. For our project, we utilize a pre-trained VSR model based on ESPnet and trained on the LRS3 dataset. This model enables precise prediction of spoken content, facilitating accurate subtitle assignment to individual speakers and ensuring high performance even in complex multi-speaker scenarios.

2.3. Dynamic Subtitle Allocation

Dynamic subtitle allocation involves positioning subtitles in proximity to the active speaker, enhancing viewer comprehension and engagement. Several studies have explored this concept using various methodologies:

One of the earliest studies leveraging artificial intelligence for subtitle manipulation was conducted by de Castro et al. (2011). They proposed a system combining automatic speech recognition with synchronization methods tailored for live broadcasting. This study highlighted the importance of aligning subtitles with audio and visual content in real-time applications, offering a significant step forward in improving accessibility and engagement for live television programs.

2.3.1. Speaker-Following Subtitles:

Hu et al. (2014) introduced a method that places subtitles adjacent to the respective speakers by detecting them through audio-visual cues, including lip motion, center contribution, and audio-visual synchrony. And it determines an optimal placement via global optimization. While effective, this approach may require complex computations for real-time applications.

2.3.2. Region of Interest (ROI) Consideration:

Akahori et al. (2017) proposed a dynamic subtitling method that utilizes eye-tracking data to avoid overlapping subtitles with important regions. By calculating the ROI based on multiple viewers' gaze data, they position subtitles immediately under the ROI, preventing obstruction of critical visual content.

2.3.3. Multimodal Systems for Accessibility:

Tapu et al. (2019) developed the DEEP-HEAR framework, a multimodal dynamic subtitle positioning system designed to increase accessibility for deaf and hearing-impaired individuals. Their system places subtitles near the active speaker by fusing text, audio, and visual information to detect and recognize the speaker's identity.

While these approaches have contributed significantly to the field, our method offers distinct advantages:

1. **Adaptability to Existing Content:** Unlike some methods that require multi-channel audio inputs or eye-tracking data, our system operates effectively with standard single-channel audio and does not necessitate additional pre-processing during filming, enhancing its applicability to a wide range of existing video content.
2. **Robustness in Multi-Speaker Scenarios:** By integrating advanced visual speech recognition (lip-reading) models, our approach accurately matches subtitles to speakers even in overlapping speech situations, a challenge for many existing systems.

These innovations position our system as a versatile and efficient solution for dynamic subtitle allocation in diverse multimedia contexts.

3. Methodology

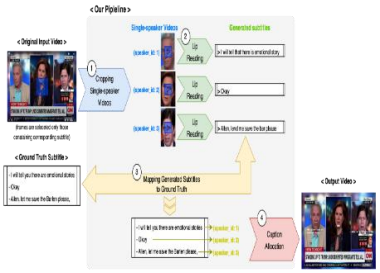

Figure 2: Overall Pipeline of Dynamic Subtitle Allocation

Figure 2 illustrates the overall pipeline of our research. The initial stage of the task involves identifying and separating different speech sources. We extract single-speaker videos from the original input video to distinguish multiple speakers. Subsequently, we apply a lip-reading model to each isolated single-speaker video to generate subtitles. As these generated subtitles may differ from the existing subtitles, which serve as the ground truth, we map them to the most similar ground truth subtitles based on a similarity metric. Finally, we allocate these mapped ground truth subtitles to the space near the independently recognized faces of each speaker.

3.1. Cropping Single-Speaker Video

The baseline model for the initial process is proposed by Afouras et al. (2018). Visual speech recognition requires video input with a single speaker. Therefore, a face recognition model is employed before this step to generate a video featuring only one cropped individual from the original footage. This process comprises two distinct stages: 1) Face detection in each frame and 2) Generation of face-cropped video. This approach ensures that the subsequent visual speech recognition model receives input containing only the relevant speaker, thereby improving the accuracy and efficiency of the recognition process.

For the first step, face detection, we utilize a face detection model with pre-trained weights called YOLOFace. YOLOFace model takes a single frame image as input and returns a set of locations of all faces in that image in a format of bounding boxes. In the second step, generation of face-cropped video, we concatenate the cropped images of detected faces in the first step. However, the problem here is that the YOLO Face only detects multiple faces in the video and does not verify which of the detected bounding boxes belongs to the same person compared to the previous frame. We may solve this problem with a face verification model, but it will be computationally expensive to perform on all frames in the input video. So, we use a relatively simple distance-based approach to determine which detected face belongs to the same person among the previously detected faces. For a distance measure, we sum up an L1 distance between all corners of bounding boxes in the previous frame and the current frame as represented in the equation (1)

$$Speaker\ ID_{t,i} = \underset{j}{argmin} \sum_{k=1}^{4} \left(\left| x_{t,i}^{k} - x_{t-1,j}^{k} \right| + \left| y_{t,i}^{k} - y_{t-1,j}^{k} \right| \right) \quad (1)$$

The equation explains how we define a speaker id of the ith face in current frame t, $Speaker\ ID_{t,i}$, where $x_{t,i}^{k}$ and $y_{t,i}^{k}$ denote the x and y coordinate of kth corner of the ith bounding box in frame t, respectively. In summary, we

calculate the distances between all pairs of faces in subsequent frames and consider those with the closest distance to be the same person. Additionally, we have done another post-processing to the bounding boxes of detected faces. Since the sizes of detected bounding boxes vary through the frames, they cannot be concatenated immediately. Therefore, we calculate the maximum size of bounding boxes in the whole frame for each speaker and slide the maximum-sized bounding box based on the initially detected bounding box's center point in every frame.

3.2. Lip Reading

Now, we have videos of each person cropped during the time shown in each caption. Each video clip is processed with a pre-trained model based on the Lip-Reading Sentences 3 (LRS3) Dataset, which consists of thousands of spoken sentences from TED. The LRS3 dataset is the most extensive publicly open audiovisual English data set that has over 2400 hours of video data. After Lip Reading process, we obtain a CSV file that contains the predicted results for each subtitle time.

3.3. Mapping the Generated Subtitles to Ground-truth

After the lip reading, we obtain N reconstructed sentences and N ground truth subtitles. The task at hand is to establish a correct mapping based on the distance between sentences from these two datasets.
The optimal mapping is determined by minimizing the distance between the paired sentences. We investigate the best method for this mapping process by exploring different algorithms and measurements for calculating the distance between sentences.

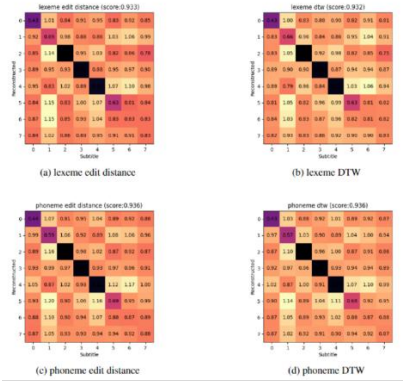

Figure 3: Heatmap Comparison of Distances Between Generated and Ground Truth Subtitles Using Different Algorithms and Measurements

Figure 3 presents a heatmap graph comparing the distances between generated and ground truth subtitles using different algorithms and measurements. The heatmap's axes represent the subtitles' labels, while the values in the squares indicate the distance between each pair of sentences. To determine which method performs best for mapping, we focus on how much lower the distance is for correctly mapped pairs compared to incorrectly mapped pairs. To quantify the overall performance, we define a score as $1 - \frac{tr(A)}{sum(A)}$, where A represents a confusion matrix. The score is higher if the diagonal element (the distance of correctly labeled sentences) is smaller than the others.

3.3.1. Edit Distance versus Dynamic Time Warping

The two algorithms regard a similarity of two sequences as a minimal cost of changing one sentence to another. The cost of adding, deleting, and modifying data in conventional edit distance (Levenstein Distance) equals 1. Dynamic Time Warping differs from the edit distance in that the cost is zero for adding or deleting a duplicate of a time step or adding or deleting zero for both ends. This makes the cost invariant for both globally and locally time-shifted series. However, the DTW algorithm assumes that elements of the sequence are in a metric space, which means the distance between two elements should be derived somehow. Currently, we implement a discrete distance function that returns zero if the two symbols are equal and one otherwise.

3.3.2. Lexemes and Phonemes

Typically, the unit of the sentence under the measurement is the lexeme. We use the raw characters to calculate the distance between the ground truth subtitle and the generated sentence. This would be fine when the reconstruction is always perfect. Nevertheless, in reality, lip reading can make wrong words with similar pronunciations, such as 'syllable' and 'pinnacle'. The problem is emphasized more when the vocabulary for the reconstruction domain is relatively small compared to the ground truth. So, we decided to phonemicize sentences before retrieving distances.

As expected, phoneme distances yielded better results compared with lexeme distances. However, DTW underperformed the conventional edit distance for the unit test dataset. This is due to an incomplete algorithm translation into the sentence domain. Since the element-wise cost is a constant real value in the original algorithm, we must distinguish cost value favoring phonemes with similar pronunciation (such as θ and t). Also, data from our experiments on comparing algorithms are short video clips with single sentences and single speakers, randomly choosing N instances since the sentences are independent. Thus, the DTW algorithm has performed better when applied to real multi-speaker video because it shows more

robustness in several aspects. First, real-life multi-speaker videos have faces that do not speak. Our model has additional code to cope with the mismatching number of subtitles and lip-reading generations, but this leads to performance degradation if unthinkingly applying edit distance. Second, the sentences played simultaneously mostly contain overlapping words since they speak for the same topic. The above concerns require the algorithm to act more robustly, giving DTW more performance on the actual data.

3.4. Caption Allocation and Combining

This research encountered two additional challenges in the caption allocation process. The first issue pertains to the fluctuation of allocated captions, while the second involves managing subtitle length. Each matched subtitle was initially assigned to the lower right portion of the speaker's bounding box. However, due to the temporal variation of the bounding box, the subtitle position exhibited instability. To address this issue, we implement a method where the subtitle positions on the output screen are fixed relative to each speaker's location, ensuring consistent placement throughout the video. The second challenge arises from the excessive length of subtitles when allocated in a single line. We employ a multi-line formatting approach to enhance readability, segmenting subtitles into lines containing six to seven words each. This technique optimizes the visual presentation of the textual information. The final output is produced by merging the processed video, which serves as the visual data, with the corresponding audio track. This integration results in a comprehensive audiovisual product that effectively combines the enhanced subtitle allocation with the original audio content.

4. Results

To evaluate whether the proposed dynamic subtitle allocation method correctly assigns subtitles to their original speakers in multi-speaker videos, we compared the assigned subtitles with the ground truth subtitles across approximately 20 videos and calculated the accuracy.

Figure 4: Comparison of Static and Dynamic Subtitle Allocation for Two-Speaker Video

Figure 4 illustrates an example of a video featuring two speakers. The image presented at the top represents the video with traditional static subtitles. In this scenario, it is not easy to discern which subtitle corresponds to which speaker based solely on the subtitles. This issue becomes particularly pronounced in the third column, where both speakers engage in an argument and speak simultaneously, creating confusion about the attribution of the dialogue. In contrast, the image at the bottom depicts the same video after applying dynamic subtitle allocation. This approach effectively mitigates these challenges. When we measure the subtitle allocation accuracy for this video—spanning 153 seconds and containing 35 subtitle lines—we find that 22 lines of dialogue are correctly assigned to their original speakers using the dynamic subtitle method. This corresponds to an accuracy of approximately 62.86% for this video.

Figure 5: Comparison of Static and Dynamic Subtitle Allocation for Three-Speaker Video

Figure 5 provides an example of a video with three speakers. In the original video, which uses static subtitles, the simultaneous speech of the three speakers results in subtitles appearing in three overlapping rows. Furthermore, the video includes pre-existing news subtitles displayed at the bottom of the screen. The static subtitles overlap with these pre-existing subtitles, significantly reducing readability. In contrast, the version with dynamic subtitle allocation resolves these issues by effectively redistributing the subtitles. In this video, all dialogue lines are accurately attributed to their respective speakers, resulting in a 100% accuracy rate.

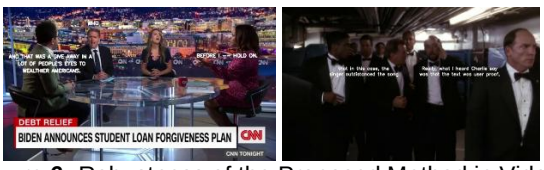

Figure 6: Robustness of the Proposed Method in Videos with Four or More Speakers

Figure 6 presents examples of videos featuring four or more speakers, demonstrating the robustness of the dynamic subtitle allocation method in such cases. The left image depicts a video where four speakers are engaged in a

discussion, with subtitles accurately assigned to each speaker. The right image showcases a scene with eight individuals appearing on screen simultaneously. In this case, the lip-reading model still successfully identifies the active speakers and accurately places the subtitles near the corresponding individuals. These results confirm that the proposed method remains effective even in videos with a larger number of speakers.

5. Conclusion and Future Work

5.1. Conclusion

In this study, we propose a novel dynamic subtitle allocation pipeline to address the limitations of traditional static subtitles in multi-speaker videos. Our method accurately synchronizes subtitles with their corresponding speakers by positioning subtitles near the speaker's face and utilizing advanced techniques such as YOLOv3-based face detection and lip-reading models. This approach effectively mitigates ambiguities in speaker attribution, particularly in scenarios involving overlapping speech, while preserving the visual integrity of the frame.

The experimental results demonstrate the effectiveness of our method, with significant improvements in subtitle allocation accuracy across various multi-speaker scenarios. For instance, in a two-speaker video, our system achieved an allocation accuracy of 62.86%, and in a three-speaker video, it successfully assigned 100% of the subtitles to their respective speakers. These results highlight the practical applicability of our pipeline in addressing the challenges of subtitle placement and speaker attribution in complex audio-visual contexts.

Our approach offers several key benefits, including compatibility with existing video content, reliance on standard single-channel audio, and eliminating the need for extensive preprocessing. This makes our system particularly valuable for enhancing accessibility for viewers with hearing impairments and improving audience engagement in entertainment applications.

In conclusion, the dynamic subtitle allocation pipeline represents a meaningful advancement in subtitle technology, bridging the gap between accessibility and usability in multi-speaker video content. Future work may explore further optimization of lip-reading accuracy and integrating this approach into real-time streaming platforms to extend its practical applicability.

5.2. Future Work

5.2.1. Enhancing Lip Reading with Automatic Speech Recognition (ASR)

The current implementation of our model relies on ground truth subtitles to match textual content with generated subtitles via lip reading. This dependency arises from the limitations of lip-reading accuracy, designed to perform well at the word level, resulting in suboptimal performance at the sentence level. To address these challenges and reconstruct text with high precision, integrating Automatic Speech Recognition (ASR) presents a promising direction. Traditional ASR systems focus on single-speaker transcription, but recent advancements enable reasonable performance in multi-speaker settings, offering the potential to improve subtitle accuracy without relying on pre-existing text.

5.2.2. Dynamic Text Style Adaption

Expanding the visual expressiveness of subtitles by dynamically adjusting their color and size could improve user experience and accessibility. With speaker detection and audio separation completed, audio clips containing only a single voice can be isolated. This enables assigning distinct colors to specific speakers, enhancing clarity, especially in multi-speaker scenarios. Additionally, the subtitle size could be modulated based on audio volume, creating a more engaging and contextually relevant visual representation of speech.

5.2.3. Optimizing Subtitle Placement in Videos

In our initial design, we proposed placing subtitles in non-intrusive positions that preserve critical visual elements of the video. However, the final model positions subtitles centrally beneath each speaker without employing a sophisticated placement algorithm. There remains significant potential for improvement in this aspect. For instance, leveraging existing computer vision techniques, subtitles could be dynamically positioned based on transparency, background contrast, or the salience of video elements. This approach would ensure that subtitles are not only functional but also minimally disruptive to the visual content of the scene.

References

Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*. https://doi.org/10.48550/arXiv.1809.00496

Afouras, T., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04096*. https://doi.org/10.48550/arXiv.1804.04121

Akahori, W., Hirai, T., & Morishima, S. (2017). Dynamic Subtitle Placement Considering the Region of Interest and Speaker Location. *Proceedings of the International Conference on*

Computer Graphics Theory and Applications (GRAPP), 626–635. https://doi.org/10.5220/0006262201020109

Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*. https://doi.org/10.48550/arXiv.1611.01599

Chung, J. S., Senior, A. W., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–3453. https://doi.org/10.1109/CVPR.2017.367

De Castro, M., Carrero, D., Puente, L., & Ruiz, B. (2011). Real-time subtitle synchronization in live television programs. *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE)*, 595–599.

Hu, Y., Kautz, J., Yu, Y., & Wang, W. (2014). Speaker-following video subtitles. *Proceedings of the 22nd ACM International Conference on Multimedia*, 1097–1100. https://doi.org/10.1145/2632111

Kanade, T. (1973). Picture processing by computer complex and recognition of human faces. *Doctoral Dissertation*, Kyoto University.

Kim, M. H., Jo, K. Y., You, H. W., Lee, J. Y., & Baek, U. B. (2017). A study on the evaluation of optimal program applicability for face recognition using machine learning. *Korean Journal of Artificial Intelligence, 5*(1), 10–17. http://dx.doi.org/10.24225/kjai.2017.5.1.10

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 21–37). https://doi.org/10.1007/978-3-319-46448-0_2

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6319–6323). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053841

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. https://doi.org/10.48550/arXiv.1804.02767

Scheibler, R., Zhang, W., Chang, X., Watanabe, S., & Qian, Y. (2022). End-to-end multi-speaker ASR with independent vector analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. https://doi.org/10.48550/arXiv.2204.00218

Son, D. Y., & Lee, K. K. (2017). A study on the recognition of face based on CNN algorithms. *Korean Journal of Artificial Intelligence, 5*(2), 15–25. http://dx.doi.org/10.24225/kjai.2017.5.2.15

Tapu, R., Mocanu, B., & Zaharia, T. (2019). DEEP-HEAR: A Multimodal Subtitle Positioning System Dedicated to Deaf and Hearing Impaired People. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 875–880. https://doi.org/10.1109/ACCESS.2019.2925806

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342