Journal of Korea Robotics Society (2024) 19(4):319-327
https://doi.org/10.7746/jkros.2024.19.4.319 ISSN: 1975-6291 / eISSN: 2287-3961

수화 인식 성능 향상을 위한 입 모양 인식 방법

Approaches of Lip Reading to Assist Sign Language Understanding

사렙 시몬스[1] · 샤미르 프라사드[1] · 임 종 윤[2] · 브루스 맥도널드[3] · 안 호 석†
Caleb Simmons[1], Shameer Prasad[1], Jong Yoon Lim[2], Bruce MacDonald[3], Ho Seok Ahn†

Abstract: With the rapid development of capable machine learning models, comes the possibility for the creation tools and services to benefit those with disabilities. There are currently an estimated 70 million individuals who suffer from hearing impairments. To help overcome this issue, we will be researching the application of using lip-reading machine learning models to help understand and translate those who communicate through sign language (signers). This paper will present the current state of the art lip reading models along with the current limitations. We show that the real-world accuracy of existing models significantly underperforms results reported in controlled testing environments. We propose solutions to these problems as well as our own design approach to create a lipreading model to help understand sign language. This research is preliminary research to use lip reading to enhance the performance of sign understanding, which is a novel idea, and will contribute to sign understanding with lip reading as a future work.

Keywords: Lip Reading, LRW Dataset, Sign Language Recognition

1. Introduction

It is estimated that by 2050, there will be 700 million people affected by hearing impairments by 2050[1]. Unfortunately, existing hearing aids do not provide accurate real time transcriptions in noisy real-world environments, nor can they translate sign interpreters. To help further the development of future hearing aids, we propose researching how current lip-reading models can be applied to signers in understanding and translating mouth movements into words.

In the domain of sign language, hand gestures are often thought of as the primary mode of communication, yet an aspect frequently overlooked is the integration of mouth movements by signers. Mouth movements, or “mouth morphemes,” constitute an integral component of sign languages, helping communicate intonation, emotional emphasis and reinforcing the hand signs meaning. This provides a plausible avenue for lip-reading to be used as a tool to help understand and translate sign-language to a non-signer.

Because of this potential avenue, this paper will explore how the potential of lip-reading can be applied to a sign language recognition scenario. This paper’s main contribution is to research the current lip-reading technology and determine how this may be applied to assist the recognition of sign-language. This research will help lip motion generation of social robots that communicate with humans.

Received : Jul. 31. 2024; Revised : Aug. 16. 2024; Accepted : Sep. 19. 2024
※ This work was supported by the Industrial Fundamental Technology Development Program (20023495, Development of behavior-oriented HRI AI technology for long-term interaction between service robots and users) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), the Te Pūnaha Hihiko: Vision Mātauranga Capability Fund (Te Ara Auaha o Muriwhenua Ngāti Turi: The Journey of Muriwhenua Māori Deaf Innovation, UOAX2124) funded by the Ministry of Business, Innovation & Employment (MBIE, New Zealand), and the Science for Technological Innovation (UOAX2123, Developing a Reo Turi (Māori Deaf Language) Interpreter for Ngāti Turi, the Māori Deaf Community). Ho Seok Ahn* is the corresponding author.
1. Student, CARES, University of Auckland, New Zealand (csim879, spra769@aucklanduni.ac.nz)
2. Senior Research Engineer, CARES, University of Auckland, New Zealand (jy.lim@auckland.ac.nz)
3. Professor, CARES, University of Auckland, New Zealand (b.macdonald @auckland.ac.nz)
† Professor, Corresponding author: CARES, University of Auckland, New Zealand (hs.ahn@auckland.ac.nz)

We describe the current the current datasets available with the benefits and consequences of using each dataset in Section 2. This paper presents the details of current lip-reading machine learning models within Section 3, including the preprocessing of information, network architectures with training and testing procedures. In section 4, we will discuss the results of applying lip-reading models to a general setting i.e., live camera feed, and how these consequences should be addressed to come to a useful solution to assist the hearing impaired. In Section 5, we will discuss possible improvements for developing our lip-reading model to help understand sign language in a real-world use case.

2. Related Works

Signers express gestures with hand motions, but also facial expressions including lip movement for the communication among them. But current sign understanding research focuses on hands and body gesture recognition rather than facial expression. We are developing a sign interpreter robot that interprets signs with nuance of signers to non-signers who cannot understand signs. Here we should generate lip motions with synchronization and lip reading research can give a clue to generate lip motions.

Lip-reading is an active research area with many existing approach's utilizing both traditional computer vison and machine learning based approaches. Before considering the technical aspects of designing a lip-reading model, it is important to understand the nuances in the phonetic expression of the English language. Additionally, human factors also need to be accounted for in the creation of these models.

2.1 Language Factors

Language factors which contribute to these issues are phonemes and homophones as discussed in [2]. Phonemes are letters which are visually identical such as 'p', 'b', and 'm'. Homophones are words which are visually identical such as 'pat', 'bat', 'mat'. Having letters and words which are visually identical cause issues for lip-reading as there is no way to distinguish these letters or words based purely on visual information. This can be a difficult issue to resolve, one potential solution is to take the wider sentence being spoken into context and have a library which can help deduce the most likely word to fill the spot of the homophone compared to using a word-by-word basis for classifying the words being spoken.

An example of this solution is the sentence "I would like to pat the dog", if this sentence were to be classified by a lip-reading model, it would be down to random chance of pat or one of the other homophones to be classified instead. But by taking the context of the sentence into consideration, the model will have a better chance of determining the correct homophone to place in the sentence. Another less impactful issue are plurals, since they are almost identical to the base word except for the letter 's', there is potential of missing the last part of the word being spoken and it classifies the word as the non-plural version, i.e., benefit instead of benefits.

2.2 Human Factors

The other type of factors which can change how a lipreading model would classify words are human factors. The accent in which someone has can change how their mouth moves to say the same word. If the dataset was trained on British English speakers and you were to give it someone who's native language is French, the way they pronounce English words can differ compared to the British, leading to potential miss-classification of the model.

The speed in which someone speaks can affect the classification rate of a model as if you give a model input data of someone speaking significantly faster than the data in which it was trained on, the rate of change in the mouth movements differ, resulting in the same words being miss-classified. Furthermore, the facial structure, skin tone, tattoos or other visual variations outside the training data distribution in the mouth region may lower the accuracy of lip-reading models.

2.3 Existing Datasets

Behind any strongly performing machine learning model, is a high quality, diverse and large dataset in which the models can train and test on, as discussed in [2], there are many datasets for lip-reading training, however, existing datasets have size and diversity limitations them which reduces their generalization capabilities. One such dataset is the GRID dataset[3]. GRID follows a specific sentence structure, with a limited number of words or numbers which can be spoken, in the categories command, color preposition, letter digit and adverb. Each speaker repeats the same predetermined sentence with a vocabulary size consisting of 51 different words. This leads to the input training data to be repetitive and lacking diversity, thus attempting to use

a model trained on the GRID dataset in the wild results in poor accuracy, as real-world people do not speak in that exact structure. Furthermore, this imposes significant constraints on real world users, limiting the range of sentences they can articulate, given the restricted vocabulary comprising only 51 words detectable by the model.

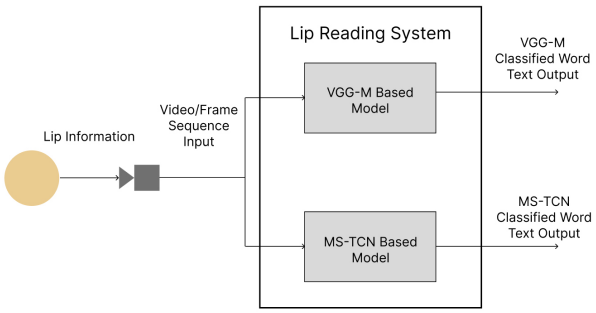
To attempt to bypass the generality problem researchers from Oxford[2] created a new dataset named Lipreading in the wild (LRW) which contains 500 different words with 1000 utterances of each word and each utterance is 1.16 seconds in length containing exactly 29 frames. There are variations on this dataset such as Lipreading Sentences in the wild (LRS) and LRS2 (updated version), however, we will be focusing on the LRW dataset. Each utterance was taken from snippets of BBC TV broadcasts. Compared to other datasets, it has over 10 times the number of classes and utterances. Having a large dataset improves the training and testing accuracy as the more information there is for a model to train on, the greater variations the model will be able to detect. The LRW dataset provides a promising solution to address the issue of poor diversity in lip-reading datasets.

3. Lip Reading System

3.1 Overall System

[Fig. 1] shows the overall System diagram of lip reading system. Our objective is to develop a system capable of capturing lip information from individuals speaking via RGB video recording. This captured lip data will then be fed into a lipreading machine learning model for the purpose of word classification, ultimately predicting the spoken words. The models we test are discussed further down the paper in Section 3.3, however, these models are only word level classification models and not capable of sentence level translation. We will evaluate these existing models' performance as an indicator of their suitability to be adapted into a lip-reading sentence translation tool.

There are many model architectures previously utilized for lip reading. VGG-M models were used as a backbone to the in [2]. Others approaches use ResNet-18 as their backbone[4]. Models can also take advantage of long short-term memory (LSTM)[5] to enhance temporal feature understanding. Networks designed specifically for video understanding, such as Multi-Stage-



[Fig. 1] Overall System diagram for desired system

Temporal Convolution Network (MSTCN)[4], incorporate temporal features directly within the backbone.

Most existing approaches to lip reading utilize models which incorporate some type of temporal mechanism such as LSTM or MS-TCN. The ability to process multi-frame features helps capture important motions of the lips across time, allowing the model to understand both temporal and spatial features coherently.

3.2 Dataset

As discussed in Section 2.3, there are various lip-reading datasets which are publicly available, therefore we need to determine which dataset(s) fit our desired functionality of being diverse, large-scale, and word level rather than sentence level. Out of the eight datasets considered, only the LRW and GRID datasets are word level datasets. Other datasets are labeled and organized by alphabetic letters, digits, or entire sentences.

Since the input data for the GRID dataset is constrained to a certain sentence structure, mentioned in Section 2.3, this does not provide the generality we need, since we want the speaker to be able to speak any word in the vocabulary in any order. This leaves LRW dataset as the most suitable dataset in terms of diversity and suitability to our use case.

The LRW dataset also possesses the desirable quality of being a large-scaled dataset, comprising 500 different words with 1000 utterances each spoken by hundreds of speakers, totaling 500,000 utterances. This is the second largest dataset, following the MV-LRS dataset which has over 5,000,000 utterances. Due to LRW being a large-scale dataset, it shows allow the models trained upon it to be capable of generalizing to real world test cases.

Based on the qualitative evaluation of existing lip-reading datasets, we will be using the LRW dataset for evaluating, comparing and testing different models.

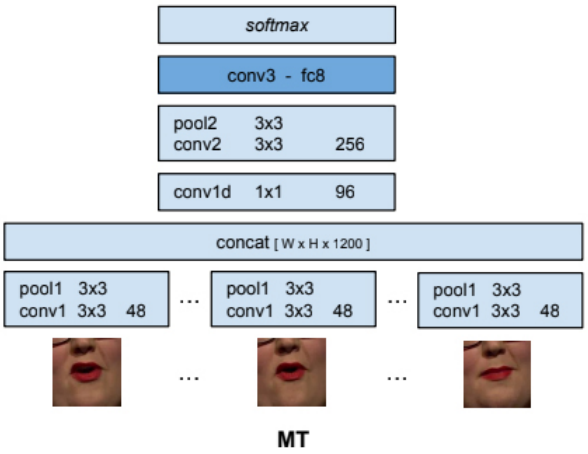
3.3 Models for Lip Reading

We review the reported accuracy of various models in [Table 1] and select two promising models tested on the LRW dataset. One model architecture we consider is [2], refer to [Fig. 2]. This model was built on top of the VGG-M architecture which incorporates a temporal component in the form of Multiple Towers (MT). The first step is to separately pass each input frame into a tower of the MT architecture. The 25 towers each process one frame from the video sequence, with every tower sharing the same weight. Each output feature is then concatenated channel wise with other towers output features. The concatenated features are then passed into a pooling layer and then a convolution layer, creating temporally fused features.

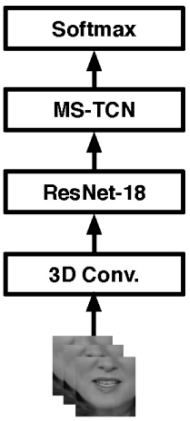
Next, a VGG-M backbone is utilized to perform spatial computation on the temporal features. Due to the architecture being mostly built upon the 2D convolution model VGG-M, both training and inference speed is higher than models designed purely based on the computationally expensive 3D convolution.

[Table 1] The best accuracies grouped by dataset

Dataset	Paper using the dataset	Accuracy
LRW (Oxford DS)[2]	S. Lai[2]	92.30%
	J. S. Chung[12]	76.20%
	P. Ma[13]	88.50%
	B. Martinez[4]	85.30%
AVICAR[6]	Y. Fu[14]	37.90%
AVLetter[7]	G. Zhao[9]	43.5%
	T. Ozcan[15]	52.31%
	A. Krizheysky[16]	54.62%
	C. Szegedy[17]	52.31%
CUAVE[8]	G. Papandreou[18]	83.00%
	M. Wand[5]	79.6%
	Y. Lan[19]	65.0%
	Y. M. Assael[20]	95.20%
	J. S. Chung[12]	97.00%
OuluVS1[9]	Y. Pei[21]	89.7%
	Z. Zhou[22]	85.6%
	S. Petridis[23]	81.80%
OuluVS2[10]	Z. Zhou[22]	73.50%
MV-LRS[11]	J. S. Chung[11]	88.9%
	T. Saitoh[24]	80.30%
	D. Lee[25]	82.20%
	Z. Zhou[22]	70.00%



[Fig. 2] MT based model architecture



[Fig. 3] MS-TCN based model architecture

This can prove beneficial as we want to apply our system to a live setting, and if the model used cannot classify words at a reasonable latency, it would cause delays which could provide a poor user experience. We need to ensure that we achieve a balance of inference latency and prediction accuracy.

Trading off speed for accuracy, we investigate[4] which proposed the MS-TCN as shown in [Fig. 3]. MS-TCN utilizes many 3D convolutional layers organized in a multistage design. Due to the heavy computational demand of 3D convolutional layers, training and more importantly inference latency will be greatly increased compared to [2]. However, MS-TCN also reports higher test accuracy. The improvement in accuracy will need to be evaluated against the increase in inference time, to determine whether the tradeoff is worthwhile.

3.4 Data Augmentation

Data augmentation is another preprocessing step which can

increase the model's accuracy by transforming the existing dataset in various ways to create greater variation in the dataset. This can increase the generalization capabilities of the model through regularization. This is because it prevents the model from overfitting on irrelevant features of the training dataset. Common data augmentations include randomly adjusting brightness, hue, saturation, cropping, resizing, flipping and rotating the image.

In the case of lip-reading, most augmentations such as flipping the image horizontally and cropping the image slightly are viable options, as they both improve the variation of the dataset without producing corrupt training data with a label mismatch. However, an augmentation unsuitable for lipreading would be flipping the image vertically, as this would turn the lips upside down, reversing the direction of the lip movements and thus creating corrupt label mismatched data.

Incorporating this corrupt training data would decrease the accuracy of the model. Data augmentation can greatly benefit the model's accuracy, but only if careful considerations are made to avoid creating corrupt training samples, such as by avoiding unsuitable augmentations. We only adopt horizontal flipping and cropping.

3.5 Preprocessing

Having a well-chosen architecture is important for the performance of a model, but without the properly formatted training data, even the optimal architecture may under perform. This is why thoughtfully preprocessing the input data as well as performing data augmentation on the data is crucial. Preprocessing can take on many forms depending on the model architecture and other considerations. In the case of [4], a series of common operations are performed on the videos in the LRW dataset to enhance the resolution of the mouth region. Firstly, the faces in frame must be detected using a face detection model and then aligned so that the faces are in the center of the frame. Then a 96x96 pixel image is cropped around the mouth as this is the Region of Interest (ROI). The image is then converted to grey scale.

These preprocessing steps are important as it allows the model to focus on the critical features (mouth region) of the data. Due to hardware memory constraints, temporal neural network models usually operate at low spatial resolutions. Without cropping, the mouth region would only occupy a small region of the input, and hence be represented by a small number of pixels. This could lead to shallow receptive field problems. Converting images to grey scale has advantages and disadvantages. Greyscale reduces the computational demands in the first convolution layer which is computationally demanding due to being high in spatial resolution. Greyscale conversion can also be considered a type of regularization, and it removes the variable of lip hue due to things such as lighting, lipstick or skin tone. However, grayscale conversion also removes information which could possibly be useful.

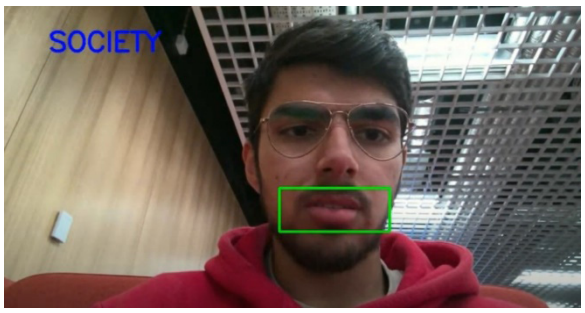
We chose to adopt the preprocessing steps described in [4], including the mouth crop and greyscale conversion. It is notable that the models take in a different number of temporal frames, which is the only key difference in the preprocessing steps.

4. Experimental Results

The primary goal of the experimentation is to determine whether the MT based model architecture performs better than the MS-TCN based model architecture in a fair comparison. We will be evaluating the models based on three main components, top-1 percentage, top-10 percentage, and processing time. By using these quantitative metrics, we can assess the word classification accuracies via the top-1 and top-10 percentages. The top-1 percentage will give us a clear indication if the model has correctly classified the word. The top-10 percentage will determine if the model's prediction was close or completely wrong. The processing time will notify us whether the model is able to perform inference at a responsive rate. This information will help indicate whether the model could feasibly be applied to a real-world application and judge the accuracy verses inference time trade off.

4.1 Test Results with LRW Dataset

Although the results of LRW are reported in ablation studies in various other papers, the testing and preprocessing methodologies could be different, causing inconsistent results. We conduct a fair test of between the MT based model and MS-TCV by utilizing the same dataset, preprocessing and augmentation methodologies. After training and validating the two models, we found that the MT based model architecture performed like the reported accuracy of 65.4% of word accuracy rate with a top 10 accuracy rate of 92.3%, and the MS-TCN based model



[Fig. 4] Example frame taken from live camera testing. Predicted word displayed on top left in blue and green rectangle showing the ROI around the lips

[Table 2] Accuracies of models trained on LRW dataset

Model	Top-1	Top-10
MT Based Architecture	65.4%	92.3%
MS-TCN Based Architecture	85.3%	-

[Table 3] Accuracies of models during the live testing

Model	Process Time	Top-1	Top-10
MT Based Architecture	1.22 s	15%	42%
MS-TCN Based Architecture	1.85 s	26%	62%

architecture also roughly matched the reported top 1 word accuracy rate of 85.3%. The testing script of MS-TCN does not report top 10 accuracy. MS-TCN achieves impressive accuracy on the LRW dataset, however, the test and train dataset although independent, are taken from the same data source modality.

4.2 Test Results from Live Streaming

To evaluate the real-world performance, which may have differences to the data in LRW, we used a webcam to perform live testing. We sampled the required number of frames for the respective model and followed the preprocessing steps before passing it through the model. The result would be shown on the live camera feed as well as printing the top 10 results in the terminal. We kept track of what words we were going to say and recorded if the model classified the words correctly. We chose 50 words randomly times from the LRW dataset to test against the models. Each word was repeated 3 times.

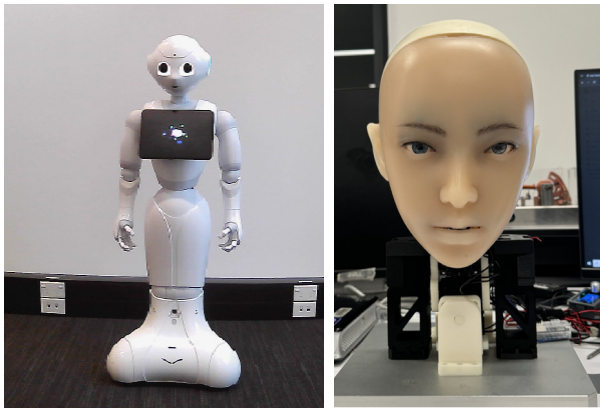
[Fig. 4] shows an example frame for “Society” from the live camera test. [Table 2] shows the test results with LRW dataset, and [Table 3] shows the test results from the live camera test. Testing the MT based model architecture, we found that on average between the three separate tests, the word accuracy rate was 15% and the top 10-word accuracy was 42%. Performing the same testing procedure on the MS-TCN, it produced an average word accuracy rate of 26% and a top 10-word accuracy rate of 62%. These live demonstration results show a significant drop in accuracy compared to the initial training and testing accuracies for both models.

There are many potential reasons for this drop in accuracy. One such issue could be the mismatch of clip length and sample rate in the live video compared to the LRW dataset. Since the LRW dataset consists of pre-recorded clips from BBC TV, they all have a consistent start and end point of frame 1 to frame 29, it is difficult to know exactly when the start of the new 29 frame cycle is for the live testing. When testing with a sliding window approach, we could have started saying a word on the 4th frame in the cycle instead of the 1st frame leading to the first three frames to be the end of another word being spoken. Another possible reason for the lower accuracy is different accents and speaking patterns, as LRW was collected from the BBC with mostly British speakers, whilst the test was performed on a New Zealand English speaker.

Another issue mentioned earlier is the variability of speaking speed. Since the speakers are TV presenters, they all follow a similar speaking speed and pattern, which means if there is a difference in how we spoke compared to the presenters, it would affect the accuracy during the live test. While doing the live test, we tried to match their speech speed and pattern as close as we could, but we cannot guarantee that it was perfect, which potentially led to some misinformation being given to the models. Furthermore, real world users cannot be expected to try match British television presenters.

5. Discussions

Initially, our objective was to replicate the state-of-the art lip-reading model and integrate it into a medium, such as a robot. We are developing different types of service robots as shown in [Fig. 5], for different purposes, i.e. interpreters, receptionists, guides, working at different places, i.e. restaurants, government offices, museums, and these robots should be able to communicate with humans[26,27]. Here, we need lip-synchronization and this lip-reading model and method can be adopted for our next step. Robots capture live camera feed of an individual mouthing



[Fig. 5] The Softbank Pepper robot (Left) and EveR4 H25 robot (Right) that can be used as robot interpreter, which should interact with signers as well as non-signers. Lip-reading research output is applied for lip synchronization of these robots[28-30]

words, then provide the translated output to the user by audibly repeating the words, which can be trained for each robot's lip movement.

To better achieve our initial objective considering the limitations found in testing, we propose refining the scope to a narrower scenario. Instead of a general setting, we aim to develop a limited vocabulary lip-reading model tailored for administrative assistance within medical environments. This entails curating a custom dataset comprising 100 words: 50 common conversational terms ('the', 'and', 'that', 'to', 'I', 'was', 'by', 'not', etc.) and 50 prevalent medical administrative terms ('patient', 'appointment', 'nurse', 'doctor', 'waiting', 'area', 'emergency', etc.). Constraining the dataset to this specific domain allows for a greater number of training samples per class, and a lower chance of errors by reducing the n umber of possible words. This should allow for real-world accuracy, which is acceptably high for users. We chose MS-TCN model as the preferred model for our use case, despite its longer processing time compared to the MT-based architecture, as it significantly outperforms the latter in terms of real-world accuracy. One method to improve real world accuracy is by prompting the user to start speaking after a beep or visual signal, ensuring the speech starts closer to the frame sampling window.

These refinements aim to enhance model accuracy during live testing by mitigating previous issues encountered. These suggestions should allow the solution to be a more viable tool for sign language understanding, addressing the needs of the hearing impaired more comprehensively. This research output will be used for our next research to enhance the accuracy of sign understanding, which is a novel approach. We will develop a receptionist robot that understands human signs, so robots can service signers as well. We target public offices such as government offices and hospital reception.

6. Conclusions

Lip-reading for sign language understanding still faces several challenges to overcome, which are products of the randomness and variations in real speakers. There are phonemes and homophones which are visually identical letters and words respectively. Human factors such as accent, and speed of speech all contribute to making lip-reading difficult. The current lip-reading datasets that are available are not as broad or diverse compared to other large datasets such as datasets MNSIT or ImageNet. This makes it difficult to create a general solution as large-scale datasets are required to train models which generalize to real world testing.

Despite our ambition to create an all-purpose general lip-reading model, we believe this is not feasible with the current datasets available. Future datasets will need a much larger vocabulary from a far greater diversity of sources. However, from the experimentation, we share insights on the performance of lip-reading model architectures and select MS-TCN as the best performing model. Because of limitations, we plan on constraining the problem domain to reduce the scope and difficulty by reducing the number of words which the model can classify. We suggest deploying the reduced scope model within a medical context to allow for satisfactory performance in understanding sign language, thus allowing hearing and speech impaired persons to benefit from being able to communicate in sign language.

References

[1] D. Mamo, *The Indigenous World 2020*, IWGIA, 2020. [Online], iwgia.org/images/yearbook/2020/IWGIA_The_Indigenous_World_2020.pdf.

[2] S. Lai, V. Lepetit, K. Nishino, and Y. Sato, "Lip reading in the wild," *Computer Vision – ACCV 2016*, pp. 87-103, 2017, DOI: 10.1007/978-3-319-54184-6_6.

[3] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recog-

nition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421-2424, Nov., 2006, DOI: 10.1121/1.2229005.

[4] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6319-6323, 2020, DOI: 10.1109/ICASSP40776.2020.9053841.

[5] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long shortterm memory," *arXiv:1601.08188*, 2016, DOI: 10.48550/arXiv.1601.08188.

[6] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang, "Avicar: audio-visual speech corpus in a car environment," *Interspeech*, 2004, [Online], https://www.semanticscholar.org/paper/AVICAR%3A-audio-visual-speech-corpus-in-a-car-Lee-Hasegawa-Johnson/76541dc3140196e34202792a728c29ab7fb334da.

[7] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198-213, 2002., DOI: 10.1109/34.982900.

[8] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, pp. II-2017-II-2020, 2002, DOI: 10.1109/ICASSP.2002.5745028.

[9] G. Zhoa, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254-1265, Nov., 2009, DOI: 10.1109/TMM.2009.2030637.

[10] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "Ouluvs2: A Multiview audiovisual database for non-rigid mouth motion analysis," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, Slovenia, pp. 1-5, 2015, DOI: 10.1109/FG.2015.7163155.

[11] J. S. Chung and A. Zisserman, "Lip reading in profile," *British Machine Vision Conference*, 2017, DOI: 10.5244/C.31.155.

[12] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3444-3453, 2017, [Online], https://openaccess.thecvf.com/content_cvpr_2017/papers/Chung_Lip_Reading_Sentences_CVPR_2017_paper.pdf.

[13] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Toronto, ON, Canada, pp. 7608-7612, 2021, DOI: 10.1109/ICASSP39728.2021.9415063.

[14] Y. Fu, S. Yan, and T. S. Huang. "Classification and feature extraction by simplexization," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 91-100, Mar., 2008, DOI: 10.1109/TIFS.2007.916280.

[15] T. Ozcan and A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models," *Balkan Journal of Electrical & Computer Engineering*, vol. 7, no. 2, pp. 195-201, 2019, DOI: 10.17694/bajece.479891.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012),* vol. 25, pp. 1106-1114, 2012, [Online], https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA, pp. 1-9, 2014, DOI: 10.1109/CVPR.2015.7298594.

[18] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423-435, Mar., 2009, DOI: 10.1109/TASL.2008.2011515.

[19] Y. Lan, R. Harvey, B.-J. Theobald, and R. Bowden, "Improving visual features for lip-reading," *International Conference on Auditory-Visual Speech Processing,* 2010, [Online], https://www.researchgate.net/publication/228084881_Improving_visual_features_for_lip-reading.

[20] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv:1611.01599*, 2016, DOI: 10.48550/arXiv.1611.01599.

[21] Y. Pei, T. K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, pp. 129-136, 2013, DOI: 10.1109/ICCV.2013.23.

[22] Z. Zhou, X. Hong, G. Zhao, and M. Pietikainen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590-605, Sept., 2014, DOI: 10.1016/j.imavis.2014.06.004.

[23] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 2304-2308, 2016, DOI: 10.1109/ICASSP.2016.7472088.

[24] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen, "Concatenated frame image-based CNN for visual speech recognition," *Computer Vision – ACCV 2016 Workshop*, pp. 277-289, Mar., 2016, DOI: 10.1007/978-3-319-54427-4_21.

[25] D. Lee, J. Lee, and K. E. Kim, "Multi-view automatic lip-reading using neural network," *Computer Vision – ACCV 2016 Workshops*, pp. 290-302, 2016, DOI: 10.1007/978-3-319-54427-4_22.

[26] J. Y. Lim, I. Sa, B. MacDonald, and H. Seok Ahn, "A Sign Language Recognition System with Pepper, Lightweight-Transformer, and LLM," *arXiv:2309.16898*, 2023, DOI: 10.48550/arXiv.2309.16898.

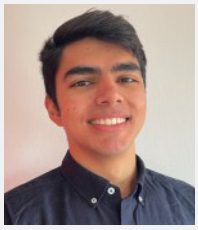
[27] F. Tracey, C. Ying, J. Y. Lim, and H. S. Ahn, "Human Motion Mimicking on Pepper Social Robot," *The 2023 Australasian Conference on Robotics and Automation (ACRA 2023),* Sydney,

Australia, 2023, [Online], https://www.araa.asn.au/conference/acra-2023.

[28] H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, and H. Lee, "Appropriate Emotions for Facial Expressions of 33-DOFs Android Head EveR-4 H33," *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication,* Paris, France, pp. 1115-1120, 2012, DOI: 10.1109/ROMAN.2012.6343898.

[29] H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, and H. Lee, "Designing of Android Head System by Applying Facial Muscle Mechanism of Humans," *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012),* Osaka, Japan, pp. 799-804, 2012, DOI: 10.1109/HUMANOIDS.2012.6651611.

[30] H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, and H. Lee, "Uses of Facial Expressions of Android Head System according to Gender and Age," *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* Seoul, Korea, pp. 2300-2305, 2012, DOI: 10.1109/ICSMC.2012.6378084.



Caleb Simmons

2020~현재 University of Auckland Electrical, Computer and Software Engineering(학사)

관심분야: Computer Vision, Robotics

Shameer Prasad

2020~현재 University of Auckland Electrical, Computer and Software Engineering(학사)

관심분야: Computer Vision, Robotics

임 종 윤

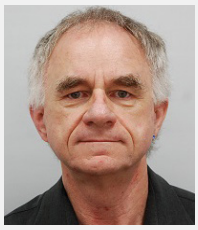
2007 연세대학교 컴퓨터공학과(학사)

2021 University of Auckland Software Engineering(석사)

2007~2021 삼성전자 연구원

2015~현재 The University of Auckland 연구원

관심분야: Computer Vision, NLP, Robotics



Bruce MacDonald

1978 University of Canterbury Electrical Engineering(학사)

1983 University of Canterbury Electrical Engineering(박사)

1995~현재 The University of Auckland, 교수

관심분야: 농업로봇, 로보틱스

안 호 석

2005 성균관대학교 정보통신공학부(학사)

2010 서울대학교 전기컴퓨터공학부(박사)

2010~2012 한국생산기술연구원 선임연구원

2012~2013 일본 전자통신기초기술연구원 선임연구원

2013~현재 The University of Auckland, 부교수

관심분야: 인공지능, HRI, 소셜로봇