Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6866–6873
Marseille, 20-25 June 2022

SSR7000: A Synchronized Corpus of Ultrasound Tongue Imaging for End-to-End Silent Speech Recognition

Naoki Kimura, Zixiong Su, Takaaki Saeki and Jun Rekimoto
The University of Tokyo, Japan
{kimura-naoki, zxsu}@g.ecc.u-tokyo.ac.jp, takaaki_saeki@ipc.i.u-tokyo.ac.jp, rekimoto@acm.org

Abstract

This article presents SSR7000, a corpus of synchronized ultrasound tongue and lip images designed for end-to-end silent speech recognition (SSR). Although neural end-to-end models are successfully updating state-of-the-art technology in the field of automatic speech recognition, SSR research based on ultrasound tongue imaging has still not evolved past cascaded DNN-HMM models due to the absence of large datasets. In this study, we constructed a large dataset, namely SSR7000, to exploit the performance of end-to-end models. The SSR7000 dataset contains ultrasound tongue and lip images of 7484 utterances by a single speaker. It contains more utterances per person than any other SSR corpus based on ultrasound imaging. We also describe preprocessing techniques to address the data variances that are inevitable when collecting a large dataset and present benchmark results using an end-to-end model. The SSR7000 corpus is publicly available under the CC BY-NC 4.0 license.

Keywords: Silent speech recognition, ultrasound tongue imaging, video corpus, end-to-end speech recognition model

1. Introduction

A silent speech interface (SSI) (Denby et al., 2010) enables us to speak or use voice interfaces without uttering an audible sound. The essential purpose of SSI is to expand the range of applications of voice interfaces in computing. Voice interfaces (Porcheron et al., 2018; Seaborn et al., 2021) based on automatic speech recognition (ASR) are intuitive interfaces that most people can use without training. It is like asking someone else to do things for us. However, the intrinsic nature of vocalization presents various constraints. For example, it is difficult to use in noisy environments. Caution is also warranted when handling information that may jeopardize privacy or confidentiality. The SSI removes this limitation by enabling non-voice interactions. It also allows communication for users in hands-busy settings or for those with a low voice or no voice due to tracheostomy, amyotrophic lateral sclerosis (ALS), or dysarthria.

The key technology for implementing SSI is silent speech recognition (SSR). Silent speech is defined as involving only articulatory movements without vocalization or the use of vocal cords. Traditionally, sensors such as surface electromyography (Maier-Hein et al., 2005; Kapur et al., 2018), electroencephalography (Porbadnigk et al., 2009), a front camera for lip reading (Wand et al., 2016; Assael et al., 2016b; Sun et al., 2018), and ultrasound imaging (Kimura et al., 2019; Cai et al., 2011; Ji et al., 2018a) have been used for SSR. Among these, ultrasound imaging is superior as a non-invasive and safe means of obtaining detailed images of the body, as it is also used during pregnancy (Denby et al., 2010). It can capture tongue movements, which play a vital role in articulation. In addition, the ultrasound probe, the sensor for ultrasound imaging, can be flexible and miniaturized (approx. 1 cm × 2 cm). These are essential factors for future use in wearable applications.

Several studies have focused on SSR based on ultrasound tongue imaging (UTI), and the current state-of-the-art (SOTA) method (Ji et al., 2018b) for this task uses the cascaded DNN-HMM model of speech recognition (Ji et al., 2018a). On the other hand, in the field of ASR, end-to-end models (Kim et al., 2017; Chiu and others, 2018) based on connectionist temporal classification (CTC) (Graves and Jaitly, 2014) or attention-based encoder-decoder (Chan et al., 2016) have become mainstream due to their significantly better performance for large speech corpora. Some of these techniques have also been adopted in the field of lip reading (Assael et al., 2016a; Afouras et al., 2018), which is similar to silent speech recognition tasks, and have achieved SOTA performance with large-scale datasets (Chung and Zisserman, 2016; Alghamdi et al., 2018).The emergence of large datasets has attracted many researchers to the field of lip reading and accelerated research in this area. However, the benchmark dataset of the UTI-based SSR, the Silent Speech Challenge (SSC) dataset (Denby et al., 2013), is relatively small compared with the corpora used for other speech recognition tasks. It is therefore not suitable to exploit performance from end-to-end models.

In this study, we constructed SSR7000, a large-scale corpus of synchronized ultrasound tongue and lip images designed for end-to-end UTI-based SSR. Our dataset comprises approximately 7484 UTI and lip images of silent speech by a single native speaker of English. Table 1 presents a comparison of SSR7000 with other corpora for UTI-based SSR. Our dataset is characterized by a large sample size for a single speaker and a realistic variance among samples, assuming the stories using end-to-end models. The SSR7000 appears to be an extension of SSC (Denby et al., 2013). It shares the same number of speakers and part of the

Table 1: A comparison of UTI-based corpora. Our SSR7000 corpus is characterized by the maximum number of utterances per person. It is approximately three times bigger than SSC (Denby et al., 2013). The TaL (Ribeiro et al., 2021) corpus and SSR7000 used the same hardware and software system.

	SSR7000	SSC	TaL	UltraSuite
Silent speech	Yes	Yes	Almost No	No
Lip camera	Yes	Yes	Yes	No
Number of speakers	1	1	81	113
Max utterances per person (training data)	7384	2342	1582	500
Corpus to read	TIMIT+WSJ0	TIMIT+WSJ0	Mixture	Mixture

corpus but is essentially different. Our dataset has approximately three times the number of sentences and a relatively larger variance among samples compared with SSC (Denby et al., 2013). This is primarily due to the fact that we did not perform a strict calibration to suppress variance, unlike SSC, for each collecting session, considering that calibration will be a barrier when collecting samples even larger than the SSR7000 or when collecting from multiple speakers in a future study. Additionally, for the application used in wearable computing, calibration is not practical for each instance. SSR7000 also provides a preprocessing challenge to reduce data variance, and this paper presents a benchmarking method for this. The UltraSuite repository (Eshky et al., 2018) contains ultrasound and speech data from 58 children with normal development and from 28 children with speech disorders receiving speech therapy. The most recent TAL corpus (Ribeiro et al., 2021) consists of TaL1, a set of six recording sessions of one male native English speaker who is a professional voice talent, and TaL80, a set of 81 recording sessions of a male native English speaker with no professional voice talent experience. The TaL corpus is similar to SSR7000 in that it uses the same fixing device and recording software. However, the TaL corpus is not strictly a silent speech corpus (participants uttered voice), as it is also intended for use in articulatory-to-speech mapping (Hueber et al., 2011; Porras et al., 2019), language learning (Wilson and Gick, 2006; Gick et al., 2008), and phonetics research. Our dataset contains the largest number of utterances per speaker.

Our main contribution is as follows: 1) we have designed and constructed a new dataset "SSR7000" for SSR using end-to-end models, 2) we describe a strategy for recording a large-scale SSR dataset and preprocessing techniques to handle data variances, and 3) we present benchmark results using an end-to-end ASR model. In Section 2, we delineate the data collection method, the characteristics of the data, the preprocessing techniques based on the data properties. In Section 3, we demonstrate how to extract features from the preprocessed data and benchmark the recognition of these features using ESPnet (Watanabe et al., 2018), a speech recognition toolkit. The raw image data, the preprocessed data, and the feature extracted data of the dataset have been packaged and made public. The recognition part of the dataset is available in a form that anyone can reproduce using Google Colab[1].

2. SSR7000

2.1. Dataset Collection

Our SSR7000 corpus is a recording set consisting of 7484 utterances by a single male native English speaker. In this paper, we split the dataset into 7384 training data (100 for validation) and 100 testing data. All utterances were recorded in a silent manner, where the participant did not speak aloud but only moved his articulatory organs. We used an UltraFit system (Spreafico et al., 2018) (Fig. 1) for data acquisition. The system is comprised of a 3D-printed adjustable helmet housing a convex-array ultrasound probe (opening angle: 104°, frequency range: 5–10 MHz, and piezo elements: 128) to the chin of the participant and an NTSC micro-camera for capturing from the front so that the participant's lips were entirely visible in the image.

The Articulate Assistant Advanced (AAA) software (Articulate Instruments Ltd, 2021) was used to record and synchronize the dataset. Ultrasound images were recorded with a field of view of 92 degrees, outputting videos with a resolution of 640 × 445 pixels at 63.51 fps. Lip images were recorded using the micro-camera, outputting videos with a resolution of 640 × 480 pixels at 59.94 fps (greyscale interlaced). The two video streams were synchronized using the SynchBrightUp unit, which is triggered by an audio beep that superimposes a white mark on the video signal and generates a pulse on the audio channel, thereby aligning the first few frames.

Given that the fatigue of the participant could affect the articulation, we limited the collection time to 1 h per session and 3 h per day. We also avoided collecting data for more than 3 days in a row. In addition, we removed the equipment every time we took a break. Between sessions, we did a simple check to ensure that the tongue and lips were visible to the sensors (camera and probe). The SSC (Denby et al., 2013) data collection included a recalibration process to adjust the tongue and lip positions interactively using modules provided by Ultraspeech for each session to make the positioning

[1] https://github.com/supernaiter/ssr7000

consistent. Consistent positioning with strict recalibration is important to achieve high accuracy on the test set, but we omitted it based on the story the SSR7000 supposes, the large dataset for end-to-end models. For this reason, SSR7000 has a larger variance than SSC. This can be observed in the difference in clarity between the SSR7000 "average faces" (Fig. 4-A) and the SSC "average faces" (Fig. 6). This is an important property difference between SSC (Denby et al., 2013) and SSR7000.

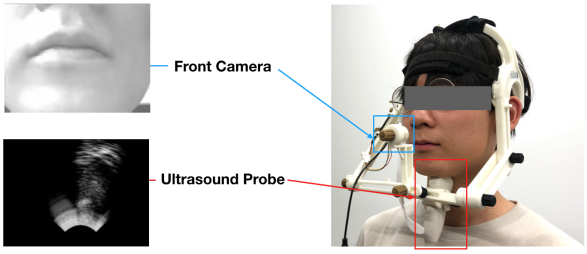

Figure 1: UltraFit stabilizing helmet, which fixes the video camera in front of the participant and fixes an ultrasound probe to the chin of the participant. The same system was used for UltraSuite (Eshky et al., 2018) and TaL (Ribeiro et al., 2021).

For the recording prompts, we chose the TIMIT corpus (Garofolo et al., 1992) as the SSC (Denby et al., 2013) because it includes phonetically balanced 2342 sentences and is suitable for training data. To further extend this, 5042 new sentences were selected from wsj0 (Garofalo et al., 2007) corpus, adding up to 7384 sentences (50 sentences from each corpus are used for validation). For the test set, we selected the same 100 sentences as SSC from the wsj0 (Garofalo et al., 2007). The sentences for the test data are fully independent of the 7384 sentences in the training and validation data. Since the scripts from the WSJ corpus are generally longer than those from TIMIT, we set a maximal duration of 12 s for the recordings compared to the 8 s used in the SSC dataset (Denby et al., 2013). All 7484 sentences were captured in approximately 50 sessions. As mentioned above, the camera and the probe positions differed slightly for each session.

2.2. Ultrasound Tongue Images (UTIs)

Fig. 2 depicts samples of the captured ultrasound tongue image (UTI) sequences. UTIs were captured using a high-gain setting. As with a normal RGB camera, a high-gain setting on the ultrasound imaging probe will make the image brighter, while a low gain setting will make it darker. Since it is difficult to always guarantee the right gain setting in a large dataset, we used a high-gain setting to reliably capture the tongue throughout many sessions over several weeks. Although this high-gain setting resulted in substantial white noise, as the first row of Fig. 2 shows, in all sessions, we were able to avoid the worst-case scenario in which the gain was too low to capture the target tongue. However, the second row of Fig. 2 indicate that when we applied feature extraction using discrete cosine transform (DCT), there was a minimal difference between the reconstructed images, meaning that DCT did not extract important features. Therefore, we designed a filter that removes white noise and emphasizes only the target tongue.

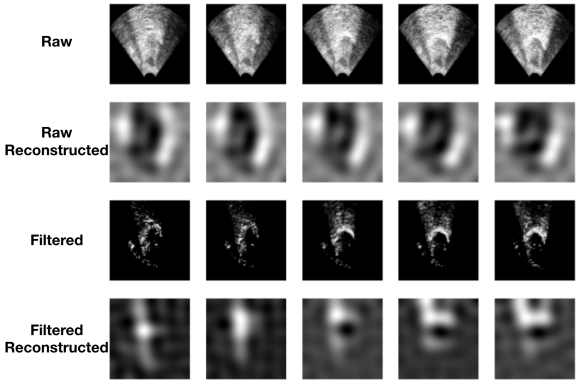

Figure 2: Ultrasound tongue images (UTIs) and reconstructed images using discrete cosine transform (DCT). The first row shows the raw images of the UTIs, which were captured at a high-gain setting and had white noise. The second row shows reconstructed images of those from the first row. The third row shows the UTIs after applying moving average filtering. The reconstructed images in the fourth row are increasingly distinguishable from those without filtering.

2.2.1. Filtering UTI

In still images, white noise is difficult to distinguish from the tongue. However, since white noise is inconsistent over time, it is easy to distinguish the two in video. Therefore, we set a high brightness threshold for each image and performed moving average filtering to emphasize the consistent capture of the tongue over time. Fig. 2 compares filtered (the third row) and raw data (the first row). Considering the frame rate of the ultrasound videos (60 fps), we set the slide window size to 5 and discovered that the noise reduced substantially. The images in the second and fourth rows are the reconstructed images extracted using DCT. In the filtered images, the white noise almost disappears, and the tongue features are emphasized. The fourth row in Fig. 2 shows that filtering succeeded in emphasizing the important DCT features.

2.3. Lip Images

In the top row of Fig. 3, raw lip images taken from different sessions are depicted. The lip positions changed observably in each session. Fig. 4. shows the "average face" from the training data, which is calculated by aggregating the first frame from each of the 7384 utterances and then dividing their sum by the number of utterances (7384). The "average face" in the first frame of the training data is highly blurred. This means that

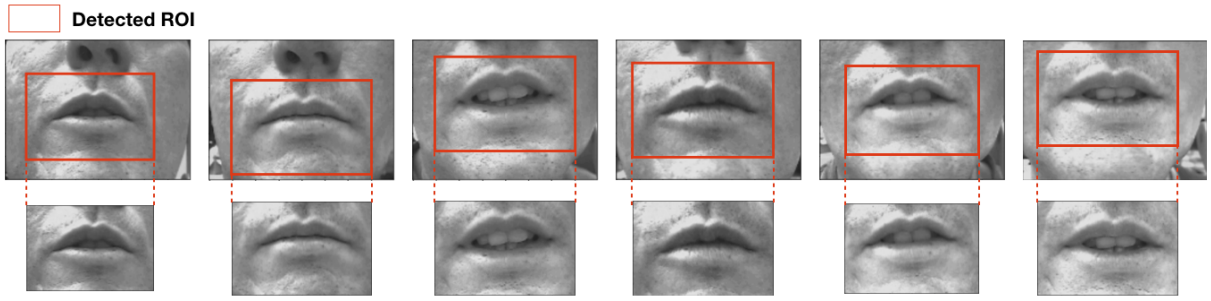

Figure 3: The visualization of cropping for lip videos. We detected ROIs in lip videos using a neural image tracking algorithm and cropped them out for better recognition results.

the lip positions varied and that there was a high variance in the raw data. When comparing average faces from the training data (Fig. 4.A) to those from the test data (Fig. 4.B), it is evident that the test data were not in the distribution of the training data.

2.3.1. Detecting Region of Interest and Cropping

Based on the above observations, we found it necessary to locate the lips and cut out the region of interest (ROI) to improve recognition results. To determine the ROI in the lip videos, we employed a deep-learning image tracking algorithm, GOTURN (Held et al., 2016), to estimate the position of the lips for each frame. For each of the lip videos, we first computed a video-level lip-bounding box by averaging the tracking results at each frame. Subsequently, the bounding box was resized to 80 × 120 and finally fine-tuned manually to ensure that the lips were located approximately at the center. Fig. 3 illustrates how the lip ROIs were detected and cropped out. Fig. 4.C shows the "average face" from the training data with the detection of the ROI and cropping, while Fig. 4.D shows that of the test data. The average training face became much clearer without ROI-cropping (A). The face from the training data (C) and the test data (D) also became more similar.

2.4. Dataset

The SSR7000 is publicly available[2] and it is the first to provide raw data without any preprocessing, which is useful for those interested in improving preprocessing. For those more interested in the recognizer rather than the preprocessing, we have provided the preprocessed data described in this paper. The corpus is publicly available under the CC BY-NC4.0 license.

3. Experiments

3.1. Recognition Pipeline

Our recognition pipeline uses a hybrid CTC/attention-based end-to-end ASR model (Watanabe et al., 2017). We implemented this model based on the VoxForge recipe of ESPnet (Watanabe et al., 2018) with some

[2] https://github.com/supernaiter/ssr7000

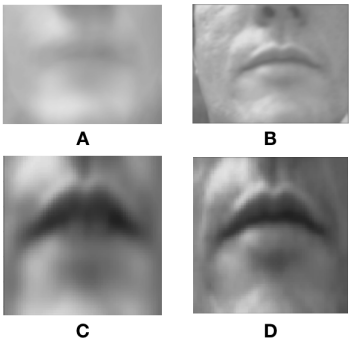

Figure 4: The "average face" of our SSR7000 dataset, which was calculated by adding up the first frame from each of the utterances, then dividing it by the number of utterances. A) shows the "average face" from the training data without cropping or preprocessing. B) shows that from the test data without preprocessing. C) shows that from the training data with preprocessing, and D) shows that from the test data with preprocessing. The more blurred the average face, the greater the variance in the data.

modifications. The detailed model architecture, parameter settings, and configurations are publicly available in the repository alongside the dataset.

As indicated in Fig 5, we utilized the DCT features as inputs to the network. We used SpecAugment (Park et al., 2019) to apply temporal and frequency augmentation, which includes a random time warp (shifting the data sequence along the time axis) for up to five frames, two random time masks (replacing the data in a random time range with zeros) with a length of up to 40, and two random frequency masks (replacing the data in a random frequency range with zeros) up to a width of 5 for DCT 20, 10 for DCT 30, 30 for DCT 60, and DCT 120, respectively.

3.2. Training

The training was run on Ubuntu 18.04 with a GTX1080Ti GPU and converged in approximately 4 h. The model that had the highest accuracy on the validation set was applied to the test data and to calculate the error rates.

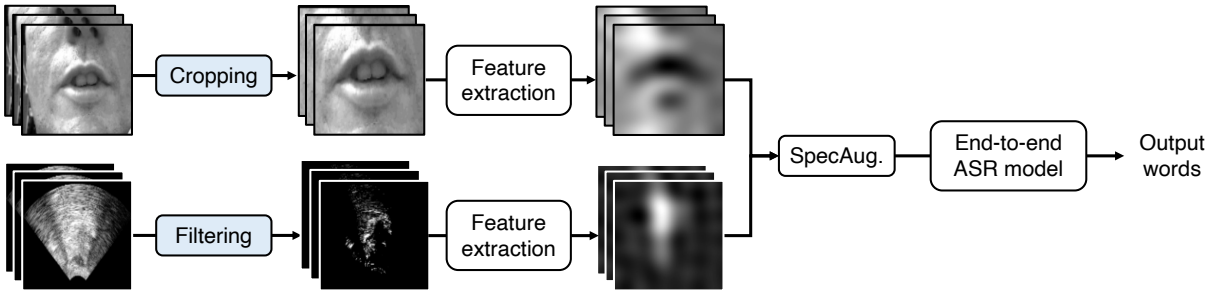

Figure 5: Our recognition pipeline using an end-to-end ASR model. Note that after the feature extraction step, we show images reconstructed from DCT features instead of DCT features themselves for visualization. The lip and tongue images are reconstructed using 30 DCT coefficients.

3.3. Results

We used word error rate (WER) and character error rate (CER) as metrics. While the phoneme level result has been indicated along with WER in past studies on SSR based on UTI, we show CER instead because our recorganizer uses characters as tokens.

3.3.1. Comparison of Preprocessing

Table 2 shows a comparison of the preprocessing performed. Evidently, each preprocessing improved the results as expected. The improvements in preprocessing were substantial (7% with ROI-C, 4% improvement with filtering on both DCT-60 and DCT30 condition), while the change in the error rate was quite minimal when employing various E2E models. This suggests that the main focus of SSR7000 is preprocessing. In particular, ROI-cropping was done semi-automatically by OpenCV, so it can be expected to be greatly improved by aligning the ROI by hand or by inventing superior methods.

Table 2: Comparison of the preprocessing results. ROI-C means ROI-cropping, which is explained in 2.3.1.

		Raw	ROI-C	ROI-C + Filter
DCT30	CER	32.1	25.8	18.1
	WER	59.4	50.4	40.1
DCT60	CER	24.0	18.4	17.6
	WER	48.9	41.2	37.6

3.3.2. Number of Data

Table 3 shows the results of investigating the effect of the number of data on the recognition. For the experiments, 60 DCT features were used. Subsets of 1000, 3000, and 5000 data were randomly selected from all training data. We repeated the random sampling and training process several times to diminish the noise in the results. Overall, we can see a linear improvement in the error rate as the amount of training data increases. This supports our idea of increasing the number of data for the E2E model. The decrease in the error rate has not yet converged, suggesting that adding more data may increase the accuracy. Data augmentation of raw image data as well as SpecAug should be also effective.

Table 3: Comparison of Number of Data

	1000	3000	5000	7284 (all)
CER	51.5	47.4	23.7	17.6
WER	89.5	81.0	50.0	37.6

3.3.3. Number of DCT Dimensions

Table 4 shows the variation in the error rate according to the number of dimensions of the features obtained by DCT. We first attempted 20, 30, and 60 dimensions, and found that the error rate tended to decrease as the number of dimensions increased. When we tried 120 dimensions, however, the error rate increased. Ji et.al (Ji et al., 2018a) reported that 30 dimensions is optimal for the SSC dataset, but when using a more expressive end-to-end model with a large amount of data, as in our current experiment, it is suggested that a larger number of DCT dimensions is appropriate. Based on this result, we set the DCT dimension to 60 in the other comparison experiments.

Table 4: Comparison of the number of DCT dimensions

	20	30	60	120
CER	27.0	18.1	17.6	37.5
WER	62.9	40.1	37.6	67.4

3.3.4. The Lip and Tongue

Table 5 indicates the results of the recognition experiments with lip images and UTIs alone (DCT-60 was used). The lip images and UTIs have been preprocessed respectively. As the "Lip and UTI" column shows, the two modalities were synergistic and had a better error rate than those of UTI or the lip images alone. The lip images alone had a good error rate of 22.7% CER and 46.1% WER. On the other hand, as in Table 5, UTIs alone had a high error rate;72.0% WER. However, considering that the UTIs had greater accuracy than the lip images when tested alone in the TaL

corpus (Ribeiro et al., 2021), the gain might have been set too high.

Table 5: A comparison of lip images, UTIs, and both.

	UTIs	Lip	Lip and UTIs
CER	40.7	22.7	17.6
WER	72.0	46.1	37.6

4. Discussion

4.1. Comparison with SSC

Although the SSC dataset and the SSR7000 were created using different equipment and under different conditions, the test sentences are the same; thus, our results with SSR7000 can be fairly compared to that of previous work with SSC. We previously performed recognition tasks on the SSC dataset using the same pipeline and recorded an error rate of 10.1% CER and 20.5% WER (Kimura et al., 2020) (The best result 6.4% WER is reported by Ji et.al (Ji et al., 2018a)). This is about half the error rate of the SSR7000's best results of 17.6% CER and 37.6% WER, even though the SSR7000 contains roughly three times as much training data. We tried not only the hybrid ctc/attention model (Watanabe et al., 2017), but also the pure attention architecture with the same hyperparameters, and the former model produced the best result above. Therefore, the causes of discrepancy should lies before the recognizer.

Fig. 6 which shows that the lip positions are quite consistent through training data, and it seems quite similar with that of test data. On the other hand, the SSR7000 average face shown in Fig. 4-A is very blurry. This is due to the fact that the lip position is different for each session, as shown in Fig. 3. The preprocessed image of the SSR7000 shown in Fig. 4-C is somewhat clearer than the raw image, but still blurrier than the SSC image (Fig. 6). This strongly depends on the quality of the calibration during the session; how rigorous the calibration is depends on the story the dataset is supposed to tell. The SSR7000 was intended to be the first model on a large dataset to exploit the performance of endto-end ASR models, so only a simple calibration was performed.

There is also room for improvement in the fixation devices for the ultrasound probe and camera. The 3D printed helmet-type fixation device used in this study could not hold the sensors in the same position for a long period of time, and the positions of the sensors moved even during the single session. If we can develop a fixation device that can be easily installed in the same position every time, t will help the calibration between sessions.

Compared to the SSR7000, the SSC (Denby et al., 2013) has a wider angle lens positioned closer to the lips; the SSC is thereby able to successfully capture the frontal tongue movement in addition to the lip movement. The discrepancy between SSR7000 and SSC may propose to adopt SSC's style to capture lip images.

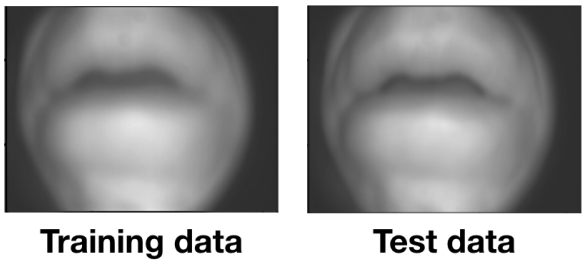

Figure 6: "Average faces (lips)" of the SSC dataset. That of the training data is clearer than that of SSR7000, and the training data and test data are surprisingly similar.

4.2. Preprocessing

As mentioned above, our dataset has a large variance in lip position, which introduces a challenge regarding the method to suppress it (ROI-cropping). We have shown a first benchmark using an existing algorithm provided by OpenCV (GOTURN (Held et al., 2016)). The lip tracing was done automatically, except for the manual adjustment of the few sessions. In order to get the best offline results, it would be useful to use a crowd worker to mark the coordinates of the corners of the mouth for all images. This allows normalization of rotations and size changes, which was difficult to do with GOTURN.

4.3. Feature Extraction

We used the discrete cosine transform used in Ji et al.'s work (Ji et al., 2018a) and TaL (Ribeiro et al., 2021) for feature extraction, but there is still room to experiment with various feature extraction methods; for example, principal component analysis is the first other method to consider. More recent methods, such as using an auto encoder and a variational auto-encoder using neural networks, are good candidates. The use of pre-trained weights (Feng et al., 2020) for the lip images established in the field of lip reading may also be useful.

4.4. Channel Attention

In this pipeline, the lip image features and the UTI features were just stacked and fed into the recognizer (for example, when using DCT-60, the stacked features were 120 dimensions). However, the lip image and the UTI should have different pronunciation strengths. For example, "p", "b", and "m" are not observable from the UTI, but can be inferred from the lip images. On the other hand, pronunciations that mainly use the tongue, such as "r" and "l", cannot be observed from outside the body, so the UTI is important. To reflect these characteristics in the recognizer, it may be effective to train different recognizers for each feature in advance and integrate them afterwards, or to incorporate a mecha-

nism such as a channel attention module (Woo et al., 2018).

5. Conclusion

This paper presented a large dataset for the end-to-end speech recognition model, SSR7000, which comprises 7484 silent speech utterances synchronized with UTIs and Lip Image. Among existing UTI-based corpora, our SSR7000 has the largest number of utterances per person. We also introduced a benchmark preprocessing method and included preprocessed images from the dataset. Silent speech recognition experiments using the E2E model of hybrid CTC/attention were performed and benchmarked. This model will be released with SSR7000. The model will be released together with the SSR7000 so that people who are interested in preprocessing, recognizers, or other techniques can try them without the difficulty of implementing the pipeline.

Afouras, T., Chung, J. S., and Zisserman, A. (2018). Deep lip reading: A comparison of models and an online application. In *Proc. INTERSPEECH*, pages 3514–3518, Hyderabad, India, Sep.

Alghamdi, N., Maddock, S., Marxer, R., Barker, J., and Brown, G. J. (2018). A corpus of audio-visual lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143(EL523).

Articulate Instruments Ltd. (2021). Articulate assistant advanced (aaa). (Accessed on 07/01/2021).

Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2016a). Lipnet: End-to-end sentence-level lipreading.

Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2016b). Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599.

Cai, J., Demby, B., P.Roussel, Dreyfus, G., and L.Crevier-Buchman. (2011). Recognition and real time performance of a lightweight ultrasound based silent speech interface employing a language model. In *Proc. INTERSPEECH*, pages 1005–1008, Florence, Italy, Aug.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. ICASSP*, pages 4960–4964, Shanghai, China, March.

Chiu, C.-C. et al. (2018). State-of-the-art speech recognition with sequence-to-sequence models. In *Proc. ICASSP*, pages 4774–4778, Seoul, South Korea, Apr.

Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *Proc. ACCV*.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., and Brumberg, J. (2010). Silent speech interfaces. *Speech Communication*, 52(4):270–287.

Denby, B., Hueber, T., Cai, J., Roussel, P., Crevier-Buchman, L., Manitsaris, S., Chollet, G., M.Stone, and C.Pillot. (2013). The silent speech challenge archive. `https://ftp.espci.fr/pub/sigma/`.

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J. M., and Wrench, A. (2018). UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Proc. INTERSPEECH*, pages 1888–1892, Hyderabad, India, Sep.

Feng, D., Yang, S., Shan, S., and Chen, X. (2020). Learn an effective lip reading model without pains. *CoRR*, abs/2011.07557.

Garofalo, J., Graff, D., Paul, D., and Pallett, D. (2007). CSR-I (WSJ0) Complete - Linguistic Data Consortium. `https://catalog.ldc.upenn.edu/LDC93S6A`, May.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (1992). Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11.

Gick, B., Bernhardt, B., Bacsfalvi, P., and Wilson, I., (2008). *Ultrasound imaging applications in second language acquisition*, pages 309–322. 01.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *Proc. ICML*, pages 1764–1772, Beijing, China, June.

Held, D., Thrun, S., and Savarese, S. (2016). Learning to track at 100 fps with deep regression networks. In *European Conference Computer Vision (ECCV)*.

Hueber, T., Benaroya, E., Denby, B., and Chollet, G. (2011). Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. pages 593–596, 01.

Ji, Y., Liu, L., Wang, H., Liu, Z., Niu, Z., and Denby, B. (2018a). Updating the silent speech challenge benchmark with deep learning. *Speech Communication*, 98:42–50.

Ji, Y., Liu, L., Wang, H., Liu, Z., Niu, Z., and Denby, B. (2018b). Updating the silent speech challenge benchmark with deep learning. *Speech Communication*, 98:42–50.

Kapur, A., Kapur, S., and Maes, P. (2018). Alterego: A personalized wearable silent speech interface. pages 43–53, 03.

Kim, S., Hori, T., and Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *Proc. ICASSP*, pages 4835–4839, New Orleans, U.S.A., Mar.

Kimura, N., Kono, M., and Rekimoto, J. (2019). Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA. Association for Computing Machinery.

Kimura, N., Su, Z., and Saeki, T. (2020). End-to-End Deep Learning Speech Recognition Model for Silent Speech Challenge. In *Proc. Interspeech 2020*, pages 1025–1026.

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (2005). Session independent non-audible speech recognition using surface electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 331–336, Nov.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Porbadnigk, A., Wester, M., Calliess, J., and Schultz, T. (2009). Eeg-based speech recognition - impact of temporal effects. In *BIOSIGNALS*.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proc. CHI*, pages 1—-12, Montreal, Canada, Apr.

Porras, D., Sepulveda, A., and Csapó, T. (2019). Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging. pages 1–8, 07.

Ribeiro, M. S., Sanger, J., Zhang, J.-X., Eshky, A., Wrench, A., Richmond, K., and Renals, S. (2021). Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. In *Proc. SLT*, pages 1109–1116, Online, Jan.

Seaborn, K., Miyake, N. P., Pennefather, P., and Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Compututing Survey*, 54(4), May.

Spreafico, L., Pucher, M., and Matosova, A. (2018). Ultrafit: A speaker-friendly headset for ultrasound recordings in speech science. In *Proc. Interspeech 2018*, pages 1517–1520.

Sun, K., Yu, C., Shi, W., Liu, L., and Shi, Y. (2018). Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, pages 581–593, New York, NY, USA. ACM.

Wand, M., Koutník, J., and Schmidhuber, J. (2016). Lipreading with long short-term memory. *CoRR*, abs/1601.08188.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., and T, H. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit. *arXiv*, abs/1804.00015.

Wilson, I. and Gick, B. (2006). Ultrasound technology and second language acquisition research. 01.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I.-S. (2018). CBAM: Convolutional block attention module. In *Proc. ECCV*.