Audio-Visual Synchronization for Ultrasound Tongue Imaging

Project report submitted
in partial fulfillment of the requirement for the degree of

Bachelor of Technology

Submitted By:

B. Ieeshasree (19BCS021)

Deepti. P. Ranjolka (19BCS037)

M. Sai Chaitra (19BCS065)

M. Roja (19BCS067)

Under the Guidance of,

Dr. Nataraj. K. S

Assistant Professor

Department of Electronics & Communication Engineering

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DHARWAD

CERTIFICATE

This is to certify that the Project Work entitled— "Audio-Visual Sychronization for Ultrasound Tongue Imaging" is a bonafide work carried out by B. Ieeshasree(19BCS021), Deepti. P. Ranjolkar(19BCS037), M. Sai Chaitra(19BCS065), M. Roja(19BCS067) in fulfillment for the Major Project of Bachelor of Technology in Computer Science & Engineering of the Indian Institute of Information Technology Dharwad during the year 2022-2023. The Project Report has been approved as it satisfies the academics prescribed for the Bachelor of Technology degree.

Signature of Supervisor(s)

Dr. Nataraj K S

Dept. of ECE

(May, 2023)

DECLARATION

We declare that this written submission represents our ideas in our own words and where other's ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

B. Ieeshasree

19BCS021

Deepti. P. Ranjolkar

19BCS037

M. Sai Chaitra

19BCS065

M. Roja

19BCS067

APPROVAL SHEET

This project report entitled "Audio-Visual Synchronization for Ultrasound Tongue Imaging" by Bellamkonda Ieeshasree, Deepti P Ranjolkar, Machireddy Sai Chaitra, and Mankena Roja is approved for the degree of Bachelor ofTechnology in Computer Science and Engineering.

Supervisor(s)

Dr. Nataraj K S

Assistant Professor,

Department of **ECE.**

Head of the Department

Dr. Pavan Kumar C

Assistant Professor,

Department of **CSE.**

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. **Nataraj** K S, Asst. Professor, Department of ECE-IIIT Dharwad for his guidance and constant support throughout the course of this major project. We would also like to thank all the faculty and administration of the institute who ensured the needs were fulfilled for the completion of this project.

Date: May 2023
Place: Dharwad

ABSTRACT

Audio-visual synchronization is the task of determining the time. Offset between speech audio and a video recording of the articulators. In child speech therapy, audio and ultrasound videos of the tongue are captured using instruments that rely on hardware to synchronize the two modalities at recording time. Hardware synchronization can fail in practice, and no mechanism exists to synchronize the signals post hoc. To address this problem, we employ a two-stream neural network that exploits the correlation between the two modalities to find the offset. We have reviewed many research papers, which are mentioned in section 2. In section 3 we have discussed us tools that are used for the visualization of ultrasound data. Data preparation, pre-processing, creation of test, training, and validation sample using self-supervision is discussed in section 4. The main objective is to synchronize between audio and UTI using an ultra-sync model. To increase accuracy and make the model more efficient we increased the number of layers and used Sync Net which employs a two-stream neural network and self-supervision. Section 5 describes the model implementation and the next section is followed by the analysis of results. We conclude the report with a summary and future directions in Sections 6&7.

TABLE OF CONTENT

1. **INTRODUCTION & OBJECTIVE 08**
   - **1.1 Ultrasound Tongue Imaging 09**
   - **1.2 Audio Visual Synchronization 10**
2. **REVIEW OF LITERATURE 11**
3. **ULTRA-SUITE TOOLS 15**
4. **DATASET-UXTD 16**
   - **4.1 Data Preparation 16**
   - **4.2 Self-Supervision Sample Creation 16**
   - **4.3 Create Test, Train & Validation Samples 17**
5. **MODEL & IMPLEMENTAION 17**
6. **RESULTS 20**
7. **CONCLUSION 21**
8. **REFERENCES 21**

1. INTRODUCTION & OBJECTIVE

Complex tongue, lips, and other articulator motions are required for speech production. For analyzing these motions in real time, ultrasound imaging has shown to be a significant tool. On the other hand, the proper synchronization of ultrasound pictures with audio recordings is critical for analyzing speech production data. We describe a method for automated audio-visual synchronization of ultrasound tongue imaging in this study, which can increase the accuracy and efficiency of speech research. In recent years, the field of speech technology has experienced tremendous developments, notably in the creation of tools for analysing speech output data. One such instrument is ultrasound tongue imaging, which offers detailed pictures of the tongue's motions during speaking. However, manually synchronizing ultrasound pictures with audio recordings takes time and is prone to mistakes. In this report, we use a unique approach for automated audio-visual synchronization of ultrasound tongue imaging, which will allow users to analyse speech data more efficiently and precisely. Ultrasound tongue imaging is a great technique for researching speech production because it allows researchers to see real-time motions of the tongue and other articulators. Synchronizing ultrasound pictures with audio recordings, on the other hand, is a difficult process that necessitates perfect timing and alignment. In this report, we provide a new approach for automated audio-visual synchronization of ultrasound tongue imaging that employs deep learning techniques to increase the synchronization process's accuracy and reliability. We feel that our technique will be a helpful addition to the toolset of speech researchers and users, allowing them to analyse speech output data more readily and efficiently.

A non-invasive method of examining the vocal tract during speech production is ultrasound tongue imaging (UTI). In order to diagnose patients, create treatments, and track therapy progress, instrumental speech therapy uses simultaneous ultrasound recordings of the patient's tongue and speech audio. Based on synchronization criteria for broadcast audio-visual signals, the two modalities must be perfectly synced with a minimum shift of +45ms if the audio leads and -125ms if the audio lags. Beyond this range, mistakes can make the data useless; in fact, synchronization errors do happen and, if left unfixed, can result in a large amount of lost effort. There is currently no mechanism in place to automatically rectify these mistakes, and although manual synchronization is technically feasible when certain audio-visual cues, including stop consonants, are present.

The main objective of our project is to sync audio and ultrasound images, for our project we are using UXTD dataset which has images of several ultrasound images forming a video and a corresponding audio. We are aiming to sync these two modalities with more accuracy than the existing one that is Ultra-sync model.

1.1 Ultrasound Tongue Imaging:

Ultrasonography tongue imaging visualizes the tongue's surface using diagnostic ultrasonography. In B-mode (brightness mode), a linear array of transducers scans a physical surface and returns a matrix of reflection intensities (scan lines echo returns) for each scan. Ultrasound data can be saved efficiently as raw reflection data with the metadata needed to change it into real-world proportions for visualization, or it can be modified at the time of recording and stored as movies. Figure 1 depicts an ultrasound frame in both raw and converted versions.

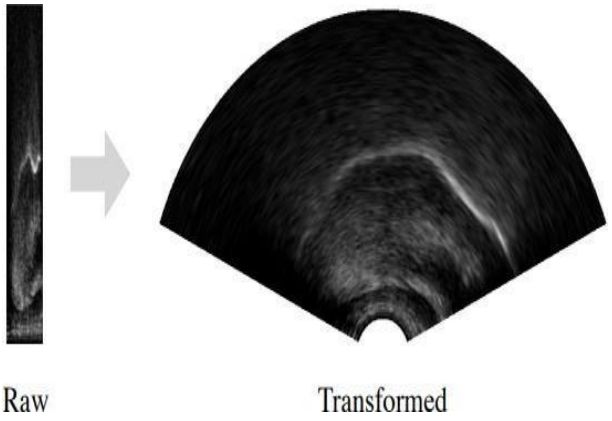

Figure 1: Each ultrasound frame is collected as a raw reflection data matrix (scan lines echo returns) and then translated into real-world proportions for display.

The ultrasound probe is inserted beneath the speaker's chin to image the tongue, recording either a mid-sagittal or a coronal view of the tongue's surface, depending on the probe's direction. Figure 2 depicts ultrasonography tongue images from the mid-sagittal and coronal planes. Ultrasound is clinically safe, non-invasive, portable, and inexpensive.

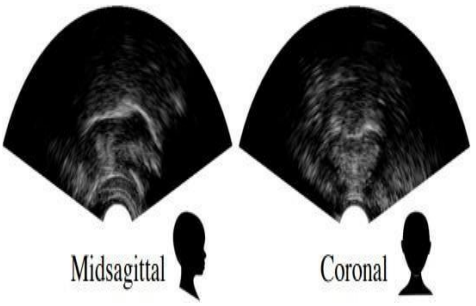

Figure 2: Examples of mid-sagittal and coronal ultrasound tongue image for child

Ultrasound tongue imaging can be used in speech and language therapy to detect a variety of speech difficulties and to offer visual biofeedback in therapy for several types of speech sound abnormalities, including those caused by a cleft lip or palate. During the intervention, ultrasonography can be used to assess the patient's development objectively or to supplement verbal input and contribute to positive reinforcement. When transcribing the speech of children with cleft lip and palate, ultrasound has been proven to boost inter-annotator agreement and aid annotators in finding latent articulation errors.

1.2 Audio-visual Synchronization:

Although ultrasound tongue imaging is frequently paired with audio, proper synchronization is required for successful analysis. When hardware synchronization fails, data usability suffers. Lip synchronization criteria have been explored, but no ultrasonography studies have been undertaken. The authors study lip thresholds to see if they hold for ultrasonography and improve on past work by building a model on data from several domains to generalize to additional data in this research. Speech audio is produced by articulatory movement, which is essentially related to other representations of this movement, including movies of the lips or tongue. Exploiting this correlation to get the offset is an alternative to using hardware to find the offset. Earlier methods looked into how different representations and feature extraction methods affected the discovery of dimensions with strong correlation. Recently, neural networks have been used for the task because they immediately learn features from input. To learn cross modal embeddings, Sync Net employs a two-stream neural network and self-supervision. These embeddings are then utilized to synchronize audio with lip videos. When using manual evaluation, it obtains accuracy that is nearly perfect (>99%), and lip-sync error is undetectable by a person. Since then, it has been expanded to incorporate various sample creation strategies for self-monitoring and training goals. We use the original approach since it is much easier to train and much less expensive than the more current variations.

2. Review of Literature:

In this literature review we have gone through various papers like “Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans (L)”, “Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)”, “One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning”, “One-Shot High-Resolution Editable Talking Face Generation via Pre-trained Style GAN”, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild” and at last we have “Automatic audio-visual synchronization for ultrasound tongue imaging”, we will briefly go through each one of the above mentioned paper below.

- **PAPER-1**

Synchronized and noise-robust audio recordings during real time magnetic resonance imaging scans (L)

This paper describes a data acquisition setup for recording, processing and running speech from a person in an MRI Scanner. The main focus is ensuring synchronicity between image and audio acquisition and getting a good signal-to-noise ratio to facilitate further speech analysis and modelling. The audio setup itself features two fibre optical microphones and a noise-cancelling filter. The real-time data acquisition routine was written in MATLAB and it uses a data acquisition toolbox. The two noise cancellation methods described are Direct NLMS and Model-based NLMS.

In Direct NLMS the gradient noise is filtered by two independent linear systems which represent the acoustic characters of the room. In Model-based NLMS a much-improved noise reduction was achieved using an artificial reference signal generated based on a model for MRI gradient noise rather than a reference signal captured during the scan. The main advantage is that there is no leakage of the desired signal into the reference channel. And the disadvantage is it doesn’t account for other noise sources like cryogen pumps. Overall, Model-based NLMS achieves good noise suppression and is easily implementable.

- **PAPER-2**

Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)

This paper is about a new database that contains recordings of people making different speech sounds while their mouth movements are captured using two different methods- real-time magnetic resonance imaging (rtMRI) and electromagnetic articulography (EMA). This database was created to help researchers to understand how people produce speech sounds, and it can be used to investigate different research questions related to speech production. The rtMRI technique provides high-quality images of the mouth and throat while the person is speaking, this helps researchers to identify different movements of the tongue, lips, and other parts of the mouth during speech production. On the other hand,

EMA captures the movements of the articulators, which are the parts of the mouth that create the different sounds in speech. By combining these two methods, researchers can better understand how speech sounds are produced. The database contains recordings of ten speakers making different speech sounds. They can also investigate the relationship between the movements of the articulators and the acoustics of the speech sounds. The database is a valuable resource for researchers who want to study speech production and improve our understanding of how we communicate.

- **PAPER-3**

A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild

Wav2Lip is the first speaker-independent model to generate videos with lip sync accuracy that matches real synced videos. This paper discusses the problem of lip-syncing a talking face video of an arbitrary identity to match a target speech segment. In this paper, they adopted a powerful lip sync discriminator that can enforce the generator to produce accurate, realistic lip motion consistently. After re-examining the current evaluation protocols, they introduced new rigorous evaluation benchmarks derived from three standard test sets named LRS2, LRS3, and LRW. They used a pre-trained discriminator that is already quite accurate at detecting lip sync errors. The generator is trained to minimize L1 construction loss between generated frames and ground truth to reduce loss in quality, they trained a simple visual quality discriminator in a GAN setup along with the generator. We have two discriminators, one for sync accuracy and the other for better visual quality. The wav2Lip model produces significantly more accurate lip synchronization in dynamic, unconstrained talking face videos. Quantitative metrics indicate that the lip sync in the generated videos is almost as good as real synced videos.

- **PAPER-4**

One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning

According to the paper "One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning" study, a system for creating realistic talking face movies of a particular individual using only a single image and corresponding audio recording is proposed. The technology makes use of a deep neural network architecture that has been trained to understand how the auditory and visual characteristics of a person's face correlate with one another.

A visual decoder and an audio encoder make up the model's two components. The visual decoder creates the correct facial movements and expressions while the audio encoder extracts the pertinent information from the audio signal. In order to produce realistic talking face videos, the model must be trained to link auditory elements to the proper facial movements.

In addition, a brand-new dataset dubbed VoxCeleb1-HQ, a high-quality variation of the VoxCeleb1 dataset, is introduced in the work. The suggested model is tested and trained using this dataset. The outcomes demonstrate that the system is able to produce talking face films of a high calibre that closely match the audio input.

Overall, the research offers a novel solution to the issue of one-shot talking face generation, which may find use in virtual reality, entertainment, and human-computer interaction, among other areas.

Before going into the literature review. Let us first discuss GAN architecture.

Brief introduction of GAN architecture: -

Ian Goodfellow developed generative adversarial neural networks in June 2014. It is an unsupervised learning algorithm where the input is unlabelled. It is majorly used in image processing and image reconstruction.

The basic architecture of GANs: -

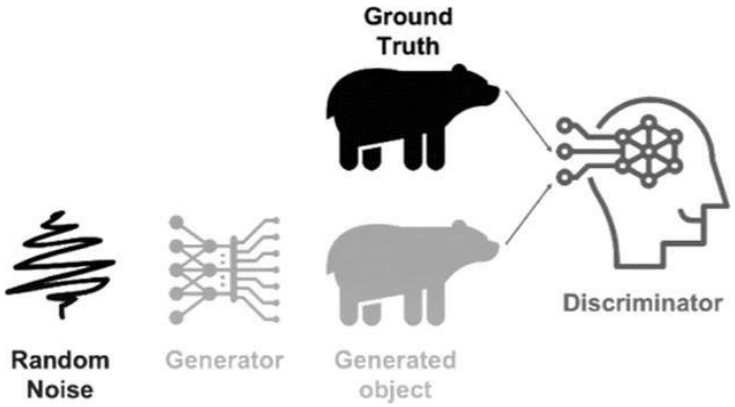

Fig 3: Generative Adversarial Network methodology

There are two neural networks in generative adversarial neural networks

1. Generator
2. Discriminator

The generator takes the features of the given application, in our case, the acoustic features like facial expressions and pose of a given image. It adds random noise to the features we had taken, combines those, and feeds them into the neural network. The generator gets trained and gives the output as a generated static image.

The discriminator takes the generator's output and compares it with the original image given as training data. The discriminator checks whether the generated output is a real or fake image. We take the loss function of both the generator and discriminator and gradient them to their optimal parameters. The generator will train until it fools the discriminator. We are fine-tuning data and refining our model parameters through this process.

Here is another research paper by the name,

- **PAPER-5**

Style Heat: One- Shot high Resolution editable talking Face Generation via pre-trained style Gan

This paper talks about a new method that has been developed to create talking faces that is very realistic and can be controlled and modified using just one picture. This is done using advanced technology that can analyse and replicate the moments and features of the face. This new approach is faster and simpler than the traditional methods which require many images and videos of the subject. It can be used in various fields such as movies, videos, games, virtual reality, and assistive technologies. Here the dataset used in the VOXceleb 2 subset is a collection of data consisting of 38,000 video clips, with a total duration of 138 hours. This data subset is a part of the larger VOXceleb2 dataset, which is a collection of audio-visual recordings of celebrities from various sources such as TV shows, interviews, and speeches.

The approach used here is a two-stage training process to generate high-quality talking faces using just one input image, The first stage uses a pre-trained model called style gan to generate a high-quality base face image. The 2nd stage involves training a new neural network module called style-heat to synchronize facial movements with input audio. Limited to facial animation only, can't be used for generating the entire human body. Requires pre-trained style gan, limited control over facial expressions, may not be able to convey all possible emotions, and not fine-grained i.e. not able to generate specific details of the facial expressions.

StyleHeat Architecture: -

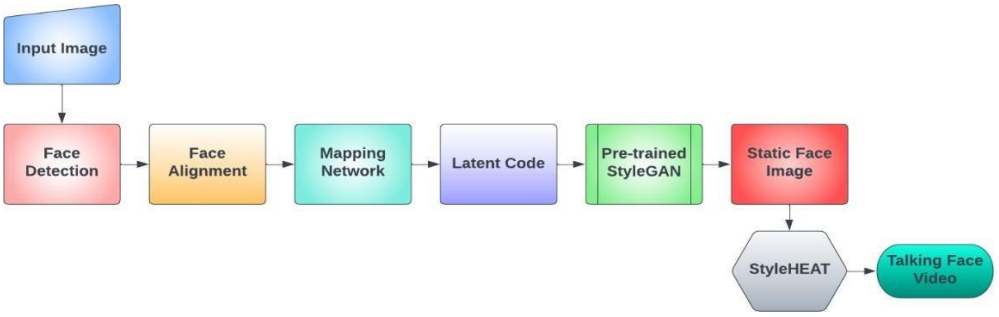

Fig 4: StyleHeat Architecture

First, the input image is fed to the model then it detects the face in the input image and excludes any other background or objects then aligns the detected face to normalize the pose and expression, this aligned face image is fed into a mapping network to encode variations in pose, expression & appearance as a low-dimensional code & uses that output code as an input to a pre-trained style Gan generators, which generates a set of high-quality realistic static face images, the style heat here adds facial expressions & movements to the static face images to create a talking face video, the final output is a talking face video that shows the input face image speaking with a variety of expressions & movements.

In conclusion, the method in this paper is a significant contribution to the field of computer vision & Generative models. The proposed method has potential applications in entertainment, virtual reality & Human-Computer interaction.

3. ULTRA-SUITE TOOLS

Ultra-suite tools are used to interpret, transform and visualize the ultrasound data. These devices might be used for a number of things, including:

1) Helping with development or deployment: The tools may consist of scripts, libraries, or other materials to help with the creation or distribution of Ultra Suite applications.

2) Simplifying workflows: The solutions may include automation or workflow optimization functionalities to make routine Ultra Suite-related chores or procedures easier.

3) Ultrasound Analysis Tools: Inside the ustools directory, you can expect to find Python files that implement various functionalities related to ultrasound tongue imaging. This may include modules for reading and parsing ultrasound data files, visualizing ultrasound frames, extracting speech features from ultrasound data, and performing transformations on ultrasound data.

There are 2 Python files in ustools module: -
1)Ultrasound process
2)Ultrasound animate

Each utterance is represented as a tuple of four files:

- The prompt file: .txt
- The audio file: .wav
- The ultrasound file:. ult
- The parameter file: .param

Ultrasound process: In this Python file, the work is being done on only 2 files namely, the ultrasound and the parameter file. The ultrasound file is a series of ultrasound images showing the child's tongue in midsagittal view. A collection of parameters for interpreting the ultrasound are contained in the parameter file.
A few steps are carried out here:
1). The sample utterance tuple's storage directory must be pointed to in the first step.
2). Then the prompt files are parsed
3) We read the ultrasound file
4)Then we try to reshape the ultrasound
5)Visualize the ultrasound
6) Transform ultrasound from raw to world proportions

Utterance animates: In this Python file we video of an utterance is being created, all four files are used to create a video of the utterance.
The frame rate used to record the ultrasound was roughly 121.5 fps. Here, we lower the frame rate to 60 frames per second, which produces a good-quality video. This can be further decreased to 25 frames per second, which is still acceptable.

4. DATASET-UXTD

For our experiment, we are selecting utterances whose recordings have been synchronized correctly at the time of recording of the dataset. This will allow us to control how our model is trained and check its performance using ground truth synchronization.

We are using the Ultra Suite database: this is a repository that contains ultrasound and acoustic data recorded during the children's speech therapy sessions, there are total 3 datasets, namely UXTD, UXSSD, and UPX but for our experiment, we are using only the **UXTD** dataset recorded with typically developing children, the dataset has 58 speakers (31 females and 27 males) of age between 5 to 12 years.
The utterances from speakers have been divided into a few categories by the type of task given to the children and are labeled as:

1. Words
2. Non-words
3. Sentences
4. Articulatory
5. Non-speech
6. Conversations

And each of the utterances has 3 files: An audio file, an Ultrasound file, and A parameter file
The audio file is a RIFF wave file, sampled at 22.05KHz, which has recordings of the speech of the child and the therapist. **The ultrasound file** contains a sequence of ultrasound frames that has captured a midsagittal view of the child's tongue. Each column in a 2D matrix representing a single scanline's ultrasonic reflection intensities makes up a single ultrasound frame. Each ultrasound frame has 63 scan lines with a total of 412 data points and is captured at a rate of ~121.5 frames per second. Grayscale graphics can be used to represent raw ultrasound frames, which can then be interpreted as videos. And at last, we have **the Parameter file that** has a synchronization offset value (in milliseconds) for this dataset was obtained using hardware synchronization at the moment of recording and verified by the therapists.

4.1 Data Preparation

They have removed "Non-speech" (E) type utterances from the training data. These are coughs that have been recorded to obtain extra tongue shapes or swallowing motions that have been captured to obtain a hard palate trace. Then applying the offset, which ought to be positive if the audio leads and negative if the audio lags. The offset is always positive in this dataset. Cutting the leading audio and applying it to reduce the longer signal's end to make it fit the duration.
In order to analyse the ultrasound more efficiently, they first lower the frame rate from ~121.5 frames per second to 24.3 frames per second while keeping 1 out of every 5 frames. The frame size is then reduced utilizing maximum pixel value from 63x412 to 63x138 by downsampling by a factor of (1, 3). Although there are fewer pixels per vector (138 instead of 412), there are still 63 ultrasonic vectors. Eliminating empty zones is the last pre-processing step. By zeroing in on audio clips that contained individually identifiable information, Ultra Suite was previously made anonymous. The zero regions

from audio and corresponding ultrasound have been eliminated as a preliminary processing step. Additionally, they tried employing vocal activity detection to eliminate areas of quiet but found that keeping them in place produced better results.

4.2. Self-Supervision Sample Creation

Positive and negative training pairs are necessary to train our model. The model reads ~200 ms-long, brief clips from each modality that are generated using the formula t = l/r, where l is the number of ultrasound frames per window (5 in our case), and r is the utterance's ultrasound frame rate (24.3). Then they divided the ultrasound into five-frame non-overlapping windows for each recording. They use a window length of ~20ms, calculated as t/(l*2), and a step size of ~10ms, computed as t/(l*4), to extract MFCC features (13 cepstral coefficients) from the audio. As a result, we get the input sizes in Figure 2(page no 7).

Pairs of ultrasound windows and their associated MFCC frames are considered positive samples. To achieve a balanced dataset, many negative samples are generated as positive samples by randomly pairing ultrasound windows to MFCC frames within the same sentence. Then they obtain a total of 243,764 samples for UXTD (13.5hrs).

4.3 Create Test, Train & Validation Samples

They wanted to find out if the approach applies to data from both new and well-known speakers in new sessions that were recorded. They choose a subset of speakers from each dataset and hold out all of their data either for testing or validation in order to emulate this. Additionally, they only let each of the remaining speakers talk for one full session, using the rest of their data for training. As they choose speakers and sessions, they strive to reserve roughly 80% of the developed samples for training, 10% for validation, and 10% for testing. The UXTD speakers each recorded one session; however, the lengths of the sessions vary. A total of 45 speakers are set out for training, 5 for validation, and 8 for testing.

5. MODEL & IMPLEMENTATION

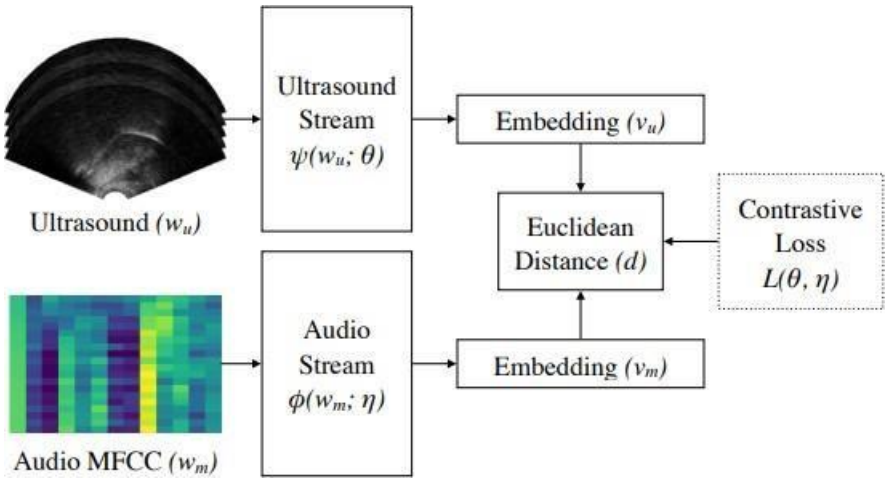

Figure 4: Ultra Sync Model Architecture

The model Ultra Sync, consists of two streams: one stream takes a short segment of ultrasound as input, and the other stream takes the corresponding audio segment as input. These inputs are high-dimensional and of different sizes. The goal is to learn a mapping from the inputs to low-dimensional vectors of the same length. The Euclidean distance between the two vectors should be small when they correlate and large otherwise.

The model architecture is illustrated in Figure 4. The visual data (ultrasound) and audio data (MFCC) are mapped to low-dimensional embeddings of the same size. The mapping functions are denoted as $\psi(u; \theta) \rightarrow v$ for the visual data and $\varphi(m; \eta) \rightarrow a$ for the audio data, where θ and η are the parameters of the mapping functions.

To train the model, a contrastive loss function is used. The loss function, denoted as $L(\theta, \eta)$, minimizes the Euclidean distance, $d = ||v - a||^2$, between v and a for positive pairs ($y = 1$) and maximizes it for negative pairs ($y = 0$). The loss function is defined as follows:

$$L(\theta, \eta) = \frac{1}{N} \sum_{n=1}^{N} y_n d_n^2 + (1 - y_n)\{max(1 - d_n, 0)\}^2$$

Here, N represents the number of training samples, y_n is the label indicating whether the pair is positive (1) or negative (0), and d_n is the Euclidean distance between v and a for the nth sample.

Once the model is trained, it can be used to predict the synchronization offset for an utterance. A discretized set of candidate offsets is considered, and the average distance is calculated for each candidate offset across utterance segments. The candidate offset with the minimum average distance is selected. It is important to note that the candidate set is independent of the model and is chosen based on task knowledge.

Algorithm 1: Synchronisation algorithm

```
Input: ultrasound, audio, and candidate offsets
for each candidate do
    Apply candidate to utterance
    Create windows of ultrasound and audio
    for each window do
        Calculate the distance between ultrasound
          and audio using UltraSync
    end
    Calculate the mean distance
end
Select the offset with the smallest mean distance
```

Figure 5: Optimizing Synchronization between Ultrasound and Audio using ultra–Sync Algorithm

The algorithm described is a synchronization algorithm designed to find the optimal offset between ultrasound and audio data using the ultra-sync model. The algorithm takes as input the ultrasound and audio data, along with a set of candidate offsets to be applied to the utterance.

For each candidate offset, the algorithm applies the offset to the utterance and creates windows of ultrasound and audio data. Then, for each window, the algorithm calculates the distance between the ultrasound and audio using the ultra-sync model, which has been trained to map high-dimensional inputs to low-dimensional embeddings.

After calculating the distance for each window, the algorithm calculates the mean distance by taking the average of all the distances. Finally, the algorithm selects the offset with the smallest mean distance, indicating the offset that provides the best synchronization between the ultrasound and audio data.

PRE-PROCESSING:

Pre-processing data is done by generating positive and negative samples from the original Ultra Suite data. We tallied the number of samples for each data set, speaker, and session, allowing us to pick training, validation, and testing splits depending on the number of samples. Later, establish data frames with the names of each subset's samples (train, val, or test). It also reorders the samples in the training set in preparation for the next phase. For a detailed explanation refer to section 4.

MODEL TRAINING:

By creating batches of samples utilizing the data frames from the previous phase. We tried a few batch sizes before settling on 64 (32 positives and 32 negatives for the same samples).

The model must then be trained. It then generates a model file and a results file. The loss of training, validation, and test data is reported in the results file. It also provides a basic binary classification accuracy by applying a 0.5 threshold to expected distances.

PREDICTION AND EVALUATION

The process begins by retrieving the true offsets for the training, validation, and test sets, where the true offsets in the test set serve as the ground truth values for evaluation. Next, the model retrieves the true offsets from the training data and bins them to generate "offset candidates" for prediction. Using the model, we predict the distance for each candidate and select the candidate with the smallest average distance. Additionally, the model calculates the distance between the ground truth and the prediction to determine the overall accuracy. Furthermore, we conduct analysis across different attributes, calculating the mean discrepancy in addition to the accuracy.

6. RESULTS

Dataset	Subset	N	Accuracy	Discrepancy_Mean	Discrepancy_SD
UXTD	new speakers	455	64.80%	97	357
All		455	64.80%	97	357

Table 1: Performance Evaluation Metrics for Dataset Subsets

We achieved 64.8% accuracy with a Mean & Standard deviation discrepancy of 97± 357ms. In this context, "New speakers" refers to a subset of the UXTD dataset that consists of recordings from speakers who were not part of the training data.

Utterance Type	N	Accuracy	Discrepancy_Mean	Discrepancy_SD
Words (A)	108	88.90%	0	32
Non-words (B)	22	86.40%	1	31
Sentence (C)	0	nan%		
Articulatory (D)	325	55.40%	136	415
Conversation (F)	0	nan%		
All	455	64.80%	97	357
A, B, C and F	130	88.90%	0	32

Table 2: Performance Evaluation Metrics by Utterance Type

Utterance Types:
Words (A): 108-word utterances.
Non-words (B): 22 non-word utterances.
Sentence (C): No utterances.
Articulatory (D): 325 articulatory utterances.
Conversation (F): No utterances.
All: Aggregated results for all utterance types combined.

The "UXTD" dataset subset includes evaluations for new speakers. It consists of 108-word utterances with an accuracy of 88.90% and 22 non-word utterances with an accuracy of 86.40%. No sentence or conversation utterances are available. Additionally, 325 articulatory utterances were evaluated with an accuracy of 55.40%.

7. CONCLUSION

In our study, we have demonstrated the adaptability of a two-stream neural network originally designed for synchronizing lip videos with audio. We have successfully repurposed this model to synchronize UTI data with audio. By leveraging the correlation between these modalities, our model learns cross-model embeddings, enabling the accurate determination of synchronization offsets. Importantly, our model demonstrates robust generalization capabilities on unseen data, accurately synchronizing the majority of test utterances. However, it is most effective for utterances that exhibit natural speech variations and less suitable for those containing isolated phones, except for stop consonants. In future directions, we plan to integrate our model and synchronization offset prediction process into speech therapy software. Additionally, we aim to utilize the learned embeddings for other tasks, such as active speaker detection.

8. REFERENCES

1. Suzhen Wang, Licheng Li, Yu Ding, Xin Yu, "One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning", 2021.
   https://arxiv.org/abs/2112.02749
2. Fei Yin, Yong Zhang, Xiaodong Cun, "StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN", 2022.
   https://arxiv.org/abs/2203.04036
3. Shrikanth Narayanan, Krishna Nayak, Erik Bresch, " Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research, 2014.
   https://asa.scitation.org/doi/10.1121/1.4890284
4. Shrikanth Narayanan, Krishna Nayak, Erik Bresch, "Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans", 2006.
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC
5. K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild", 2020.
   https://arxiv.org/abs/2008.10010
6. 6. Aciel Eshky, Manuel Sam Ribeiro, "Synchronising audio and ultrasound by learning cross-modal embeddings", 2019.
   https://arxiv.org/abs/1907.00758
7. Aciel Eshky, Joanne Cleland, "Automatic audiovisual synchronization for ultrasound tongue imaging ", 2021.
   https://arxiv.org/abs/2105.15162