Academic Science Journal

Evaluation Study for Worthwhile Research in Artificial Intelligence Techniques for Tongue Movement's Estimation

Safa Emad Sabri* **and Jamal Mustafa Al-Tuwaijari**

Department of Computer Science, College of Science, University of Diyala, Iraq

*scicompms222301@uodiyala.edu.iq

Received: 24 March 2024 **Accepted: 1 May 2024**

Published: July 2025

DOI: https://dx.doi.org/10.24237/ASJ.03.03.872B

Abstract

The introduction of deep learning has brought about worthy changes in the field of speech processing. By utilizing many processing layers, models have been developed that can estimate tongue motions and extract complex information from speech data. This review provides an overview of the main deep learning models and their applications in the tongue movement estimation function using real-time video sequences. In order to assess the relevant literature, a literature review was performed. All papers published between 2017 and 2023 that discussed methods for using deep learning techniques that were pertinent to this research were considered. After going over each article in detail, 25 of the many found met the inclusion criteria. Relevant articles were found using searches in Google Scholar, IEEE Xplore, and Scopus. This study's findings highlight a significant challenge to improving deep learning network performance: a dataset with real-time video sequences of tongue movements. Such a dataset is essential for developing automatic speech processing and high-accuracy estimation of tongue movements.

Keywords: Tongue Movements, Deep Learning, Real-time video, Speech Processing, Tongue Contour.

Introduction

Examining the most significant current methodological approaches used to estimate tongue movement with the use of AI technology is the goal of this paper. Additionally, it emphasizes the many types of data utilized to derive the most accurate estimation of tongue movement based on face movement. The report additionally elucidates research findings, identifies existing knowledge gaps, and proposes potential avenues for future research [1]. As demonstrated in Table 1, there has been a dearth of reviews and survey studies examining the Deep Learning (DL) models that incorporate estimation of tongue domains, despite the fact that numerous algorithms are employed to construct these models. To fully grasp the complexities of human communication, studies of tongue movement and speech are essential. The utilization of medical ultrasonography systems for capturing tongue movement during speech has grown prevalent due to significant advancements in medical imaging techniques and their remarkable capabilities [1, 2].

According to the research, visual cues can help distinguish between acoustically similar sounds with different articulatory characteristics. Better communication strategies in challenging listening environments and aids for people with hearing loss can result from a deeper understanding of the relationship between these sensory modalities [3].

A valuable articulatory tool that enables the observation of tongue surface movements extending from the base of the tongue to the tip of the tongue is tongue ultrasound. Tongue ultrasound is used in the areas of language teaching and silent speech interfaces, enabling communication through inaudible signals [4].

In terms of convenience and security, ultrasound imaging is unparalleled [5]. But of all the imaging modalities, magnetic resonance imaging (MRI) has the highest resolution and can show more details regarding the craniofacial anatomy, voice tract, and soft tissues [6]. To improve speech analysis, MRI is utilized to acquire images of the vocal tract in real time, in either a 2D or 3D orientation [7, 8] .

Improvements in speech processing systems have been made possible by recent developments in deep learning, particularly in the areas of attention mechanisms [7] and transformers [9]. Transformers make it possible to describe long-range relationships in the input signal, while attention methods let the model zero in on the most relevant parts of the signal. As a result of these advancements, speech processing systems have become much more effective and flexible, opening up new possibilities for use in a wide variety of contexts.

Noninvasive and simple to implement, ultrasound imaging of the tongue (UTI) has found widespread application in studies of speech production and clinical linguistics.

This study contributes to the review of tongue movement estimation and speech extraction by analyzing the efficiency of machine learning and deep learning techniques using criteria such as mean square error (MSE), accuracy, precision, and so on. As a result, a review process must be used to conduct the analysis. ML/DL is a pre-research investigation into the methods utilized to evaluate tongue movements based on specific criteria. The remainder of this review is organized as follows: Section II summarizes several current extant reviews on AI in the realm of speech extraction, tongue movement estimation, and basic information for machine learning or distant learning approaches; the third section describes the research methods used in this study. The fourth and fifth parts provide details on the survey's research methodology and highlight current techniques in this subject. Section VI introduces transfer-learning technology in the subject of tongue movement estimation, followed by a discussion of performance analysis, with an emphasis on current research gaps. Finally, the key findings of the systematic review are provided in the final section.

Related works:

Articulatory information extraction from ultrasound image sequences has been the focus of several prior efforts [10].

C.Wu.et al [11] introduced a 3D convolutional neural network that was trained on a database of unlabeled ultrasound video to predict the upcoming tongue image based on past images. This network is able to do this because it is able to distinguish between different types of tongue tissue.

SahaP.et al. in [12] introduced a method for ultrasonic (US) voice synthesis using a 3D convolutional neural network and a formula-based speech synthesis engine. The application of a unique deep learning architecture is employed to facilitate the mapping of tongue ultrasound (US) pictures obtained from a US probe positioned beneath the chin of a participant. This mapping process transforms the data into a specific format known as ultrasound2formant (U2F).

The improvements that have been made in deep learning and cross-modal mapping have motivated, H. Liu et al. [13] established a connection between these two disparate entities through self-supervised learning. This involves training a deep neural network to forecast tongue movements in a sequence of ultrasound images, relying on a corresponding sequence of lip images. Hence, the model has the capability to utilize temporal alignment information between two routes.

Videos—series of static images—are utilised as input in the study of L. Tóth et al. [14]. Processing multiple neighbouring video frames at once can reveal tongue movement time-course information in this clip. There are numerous time series processing network structures.

Suggested M . Mozaffari et al. [15] RetinaConv is an innovative convolutional module that draws inspiration from human peripheral vision. It extracts features using dilated and standard convolutions. They tested their findings on a difficult tongue ultrasound dataset. Experimental results show that their completely autonomous models can make reliable, real-time predictions on different tongue ultrasound datasets because of their excellent generalization capabilities.

L. Tóth.et al. [14] Employed a video, or a series of images, as input rather than a single still image. Multiple video frames can be processed simultaneously to take advantage of the sequence's additional information regarding the timing of tongue movement. Developing a network architecture for processing time series can be done in a number of ways. Recurrent neural networks are commonly utilized for this kind of information. They are often layered on top of a 2D convolutional neural network (CNN) that aims to process individual frames, much like Long Short Term Memory (LSTM) networks.

By tracking the movement of the tongue as it pronounces consonants, P. Padmini et al. [16] expanded previous work on statistical approaches for vowel shape frequencies and brought them up to date. The tongue-based statistical model of the oral cavity was combined with the larynx model and compared to vocal tract model of speech. The grammatical expression-based algorithm was used to create vowels and consonants for males, females, and nine-year-olds. Using the model in this article to focus on tongue gestures simplifies voice production device creation.

Experiments using an electroencephalogram (EEG) and ultrasound imaging of the tongue, as well as acoustic-to-articulatory inversion, were conducted by T. Csapó.et al.[16] The goal of this study was to draw attention to the fact that EEG has trouble predicting patterns of articulatory movement. They show that EEG input is only adequate for distinguishing between a neutral tongue position and articulated speech by comparing the actual articulatory data with Deep Neural Network DNN-predicted ultrasound, and that melspectrogram-to-ultrasound can also predict articulatory trajectories of the tongue.

L.Tóth.et al.[17] was trying out a Urinary Tract Infection UTI-based SSI network that could be directly adapted to the specific speaker or session. They add a spatial transformer network (STN) module to their network and retrain just that part of the network during the adaptation process to keep from having to retrain the whole thing. The STN uses an affine transformation that it learns from the input images to compensate for camera and speaker misalignment, and to a lesser degree, speaker differences.

Research Methodology

According to the results of surveys, which are presented in Table 1, there is no complete analysis of performance utilizing various AI models with regard to the estimation of tongue movement. This study's primary objective is to examine the accuracy of several artificial intelligence models in order to determine which models are the most accurate in terms of prediction and to demonstrate which models are the best. When it comes to modeling tongue movement estimation, this paper offers a comprehensive evaluation of the performance of several artificial intelligence models, which can serve as a guide for both researchers and practitioners. This assessment took into consideration the most widely used artificial intelligence models, including artificial neural networks (ANNs), deep neural networks (DNNs), classical neural networks (CNNs), and recurrent neural networks (RNNs). As can be seen in Figure (1), this type of systematic review is comprised of multiple steps.

Table 1: Studies Concerning AI Models with Estimating Tongue Movement.

Ref.	Type	Case Study	Category	Year	AI Techniques
[18]	**Search of A Theory**	**Face Motion Measurements**	**Traditional**	**2007**	✗
[19]	**Survey**	**Audio Synthesis And Audio-Visual Multimodal Processing**	**Traditional**	**2021**	✗
[20]	**Review**	**Mouth Interface Technologies**	**Machine Learning**	**2021**	✓
[1]	**Review**	**Tongue Contour**	**Machine Learning**	**2022**	✓
[3]	**Review**	**Speech Processing**	**Machine Learning**	**2023**	✓

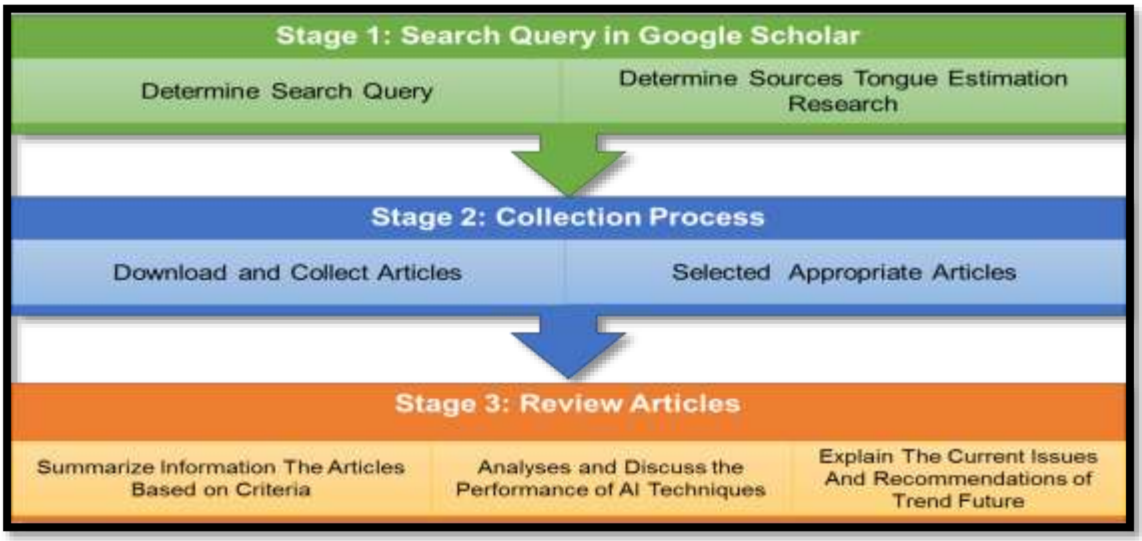
Stage 1: Search Query in Google Scholar
Determine Search Query
Determine Sources Tongue Estimation Research
Stage 2: Collection Process
Download and Collect Articles
Selected Appropriate Articles
Stage 3: Review Articles
Summarize Information The Articles Based on Criteria
Analyses and Discuss the Performance of AI Techniques
Explain The Current Issues And Recommendations of Trend Future

Figure 1: Methodology for Systematic Process

In the first stage, we'll define the words we query and use them to find the important articles; this will involve using artificial intelligence techniques for tongue estimation in the speech process. To find relevant papers, the databases are searched using multiple keywords, such as "artificial intelligence" and "tongue" or "artificial intelligence" and "face motion" or "mouth interface" or "tongue contour" or "ultrasound" or "videos" or "real time" or (at least one word) "prediction," "estimation," or ("forecasting"). In order to find relevant publications, these keywords are the most important guidance on this topic. Thus, the publication status of the extracted research is not constrained. In addition, reputable databases such as IEEE, Springer, and Elsevier are used to determine the source research engine.

In the second stage, which takes place between 2017 and 2023, you'll need to download articles from Google Scholar that propose ML models for the public domain. Third Stage: Read the chosen articles and provide a brief overview; this will include information about the project's type, location or dataset, effectiveness, and AI techniques. New knowledge is generated from this data in order to evaluate the efficacy of existing AI models. Additionally, the present difficulties and potential trends for AI approaches in estimating Tongue Real Time movement are discussed.

Observing the Tongue Movement in Real Time

Whether in or out of a hospital, an ultrasound device is perfect for getting real-time images due to its portability, safety, and convenience. Researchers and clinical linguists have made extensive use of lingual ultrasonography for numerous purposes. Some examples of these uses include research on swallowing, 3D modeling of the tongue, and silent speech interfaces. Object recognition and segmentation are only two of the many medical imaging applications that make use of ultrasonic imaging analysis [1].

The image of the tongue is detected using an ultrasound system. The sonogram can be noisy due to signal noise and some parts of the tongue may be missed in the image in the case of rapid tongue movements during the ultrasound image acquisition Figure (2) displays a view of the tongue contour in the sagittal plane. A bright white concave arc represents the final image of the tongue contour on the ultrasound screen [21].

Real-time imaging of the tongue's movements is something that has been suggested for use in speech therapy by Faber et al. [22]. The brightness of each pixel was determined by employing the PCA-based decomposition method known as EigenTongue. They use a PCA-based tongue contour model that we've dubbed EigenContour in order to get a more compact representation of the annotated tongue surfaces. The third component is an artificial neural network that was created with the purpose of modeling the connections that exist between the two representations. Utilizing single-layer perceptrons, we model the training phase relationships between the EigenTongues and EigenContours parameters. In order to obtain the x and y coordinates of the tongue contour, the segmentation procedure projects the EigenContours parameters onto the basis vector matrix. Segmentation can be done in real time since it is a frame-by-frame operation that doesn't require much computer resources [21, 22].

An approach was proposed by Wen et al. [26] for the purpose of extracting the contour of the tongue from ultrasound recordings in real time. The tongue's outline was divided using U-net and a simplified version of sU-net. They used data from two devices with different training strategies to compare the system's performance. On the other hand, the results of their work demonstrated that the technique that was proposed is extremely competitive in terms of both performance and accuracy. In response to this, they proposed the hypothesis that their deep learning model was only concerned with the spatial information contained within a single image frame, and that it did not take into account the temporal information that pertains to the entire speech that was contained within a video sequence. In addition to this, it advised the utilization of data augmentation in order to improve the training of models by taking into account variances and changing images in order to cope with various scenarios at various granularities [4, 23].

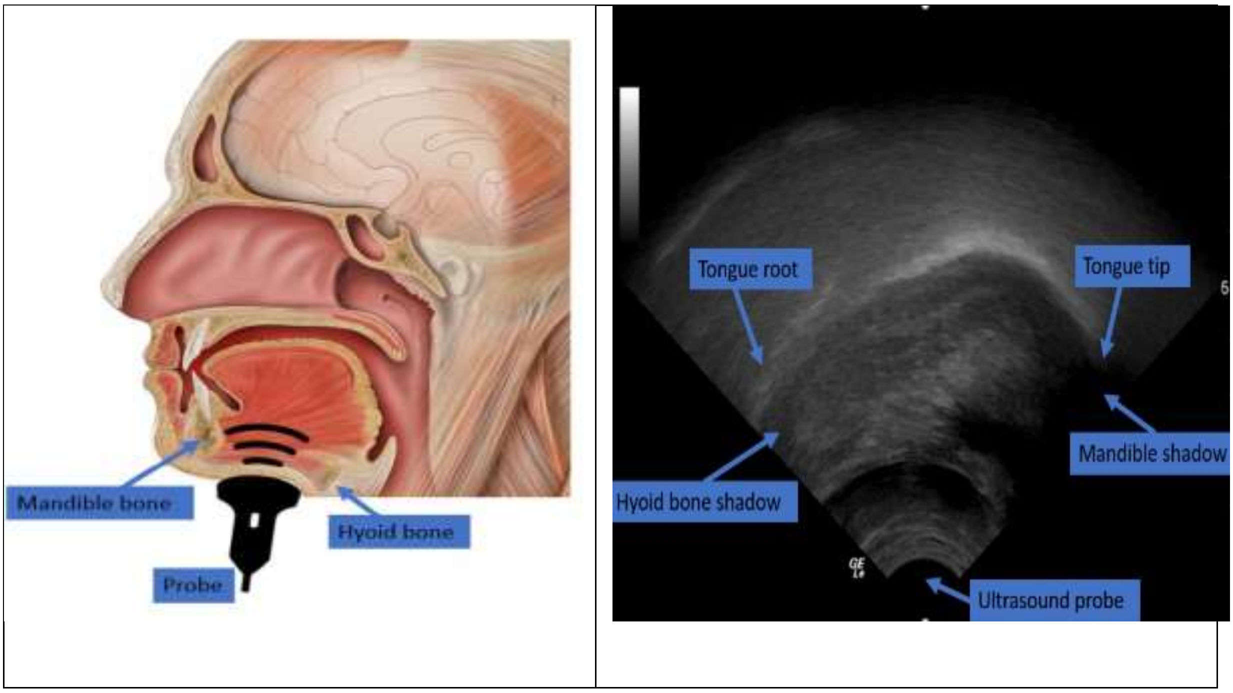

Figure 2: View of Tongue Contour in the Sagittal Plane [1].

Deep Learning Techniques In Real –Time Tongue Movement’s Estimation

Ultrasound tongue contour tracking is one of many computer vision jobs that has aggressively taken advantage of deep learning's progress [26]. Ultrasound is a popular tool for researching articulation and tongue movement in speech due to its attractive features, such as imaging at a reasonably fast frame rate, which enables researchers to see quick and delicate gestures in real-time [27].

1. Convolutional Neural Networks

When it comes to deep neural architectures, convolutional neural networks (CNNs) are a subset that uses alternating convolutional and pooling layers. In place of the more common practice of simply multiplying matrices in one of its layers, this "convolutional neural system" employs a mathematical linear operation known as convolution. Like any conventional multi-layer neural network, a CNN would have a convolutional layer first, followed by a fully connected layer. They are undeniably significant in data science and of the many popular approaches used in computer vision and image recognition frameworks [28].

All the way across the input space, a convolution layer adds filters that process tiny local bits of the input. Moving over the activation map, a pooling layer takes the maximum filter activation inside a specific window and transforms it into low-resolution activations from a convolution layer. Grid-like data processing is facilitated by CNNs, which are variations of fully connected neural networks. Examples of data structures with grid-like properties include time-series data (1D grid) with samples spaced at regular intervals and two-dimensional grid images using pixels [29].

The voice spectrogram is superior to hand-crafted features for capturing speaker characteristics such as vocal tract length variations, varied speaking styles that lead formants to undershoot or overshoot, etc. It has also clarified the frequency domain manifestations of these features. Time and frequency are highly correlated in the spectrogram. For a convolutional neural network (CNN) processing pipeline that needs to maintain locality in both the frequency and time axes, the spectrogram is an ideal input because of these properties. One intriguing use of CNNs is the representation of local correlations in audio sounds. Additionally, convolutional neural networks (CNNs) can efficiently share weights to simplify the model and extract structural information from spectroscopy. Here we'll go over the insides of 1D and 2D convolutional neural networks (CNNs), which are utilized for various speech-processing jobs.

2. Recurrent Neural Networks (RNNs)

Given the intrinsic dynamic nature of the input speech signal, it is only reasonable to contemplate using Recurrent Neural Networks for a variety of speech processing applications. Regular neural networks (RNNs) are able to simulate time-varying (sequential se) patterns that would have been difficult for traditional feedforward neural networks to grasp. At first, RNNs and HMMs worked together; the former would model the sequential data, and the latter would do localized categorization. Nevertheless, the drawbacks of HMMs are often carried over into such a hybrid model. As an example, HMMs necessitate observations of states that are independent of one another and task-specific information. There was a rise in using RNN-based end-to-end systems for sequence transduction applications such as text and speech recognition as a means to avoid the limitations of the hybrid technique [29].

The information derived from time series through the use of concealed states that act as Regular RNNs are an upgrade over feedforward deep neural networks used to represent time series and natural language sequences, both of which produce sequential data. Its memory's temporal dynamics, which are influenced by both past and present states, are captured by its recurrent patterns. Two well-known variants of gated RNNs, the Long Short-Term Memory (LSTM) and the Gated Recurrent Units (GRU), were created to circumvent the vanishing gradient issue and capture both short-term and long-term dependencies. To deal without of the ordinary data, the following imputation-focused models employ several approaches involving traditional forward-directional gated RNN units. Among these methods, you can find estimates for missing values using RNN next-step prediction, higher-order series with latent temporal dynamics generated by deep learning, and exponential decay towards the mean or last value [30].

3. Long Short-Term Memory

In order to get a more accurate extraction of long-term dependencies within the input sequence, it is recommended to utilize a form of the recurrent network known as the Long-Short Term Memory (LSTM). To better manage long-distance interactions between time-related properties, these networks' internal implementations use specific gates [3, 29, 30].

The main point is that every cell should have a recurring edge with a weight of 1. The vanishing gradient problem is thus resolved, since recurrent multiplication by 1 does not diverge nor converge to zero [31].

Also, the feature gates in LSTM blocks are responsible for deciding which bits of data from one stage to be transferred to the next. As a result, the network can figure out when to cut the gradient short. This drastically cuts down on the amount of time required to train for dependencies that persist across time. The forget gate, the input gate, and the output gate are the three gates that comprise an LSTM block. One of the gates is the forget gate [33].

In addition, the hidden state and cell state are the building blocks of an LSTM block. Both the input gate and the forget gate play a role in controlling the updating and forgetting of values, respectively, in a cell. The input and forget gate are used to calculate the new cell state. The output gate is responsible for carrying out calculations in order to ascertain which bits ought to be transmitted to the subsequent node, or the hidden state of the cell. A representation of the LSTM cell's structure may be found in Figure (3) [31].

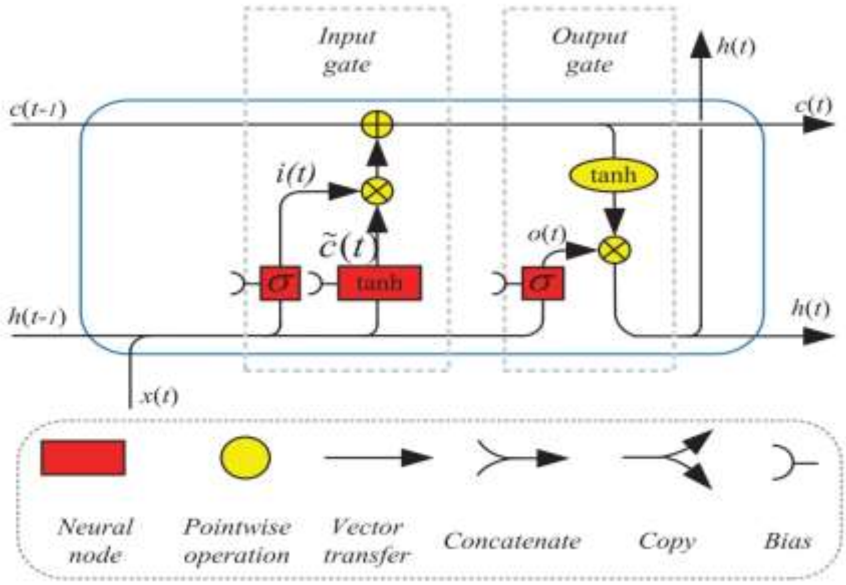

Figure 3: Initial Long Short-Term Memory (LSTM) Structure [31].

4. Transfer Learning with Tongue Movement's Estimation

Methods that are based on deep learning have been increasingly prevalent in recent years, particularly in the areas of silent speech detection and ultrasound image classification of the tongue. However, as supervised learning algorithms, neural networks necessitate a mountain of labeled data, which would be difficult to collect for studies involving ultrasonic speech [34].

One interesting machine learning approach to the aforementioned problem is transfer learning, which centers on transferring knowledge between domains, as shown in Figure (4) [35].

Various concepts and ambiguities appear in the literature on transfer learning. The terms "domain adaptation" and "transfer learning" both describe quite comparable procedures. When it comes to transfer learning, domain adaptation is all about making one or more source domains more conducive to transferring information in order to help a target learner perform better. The goal of domain adaptation is to change the distribution of the source domain to be more like the target domain's distribution. When given the choice between labelled and unlabeled data, there are several discrepancies in the literature over how to characterize the transfer learning process [36].

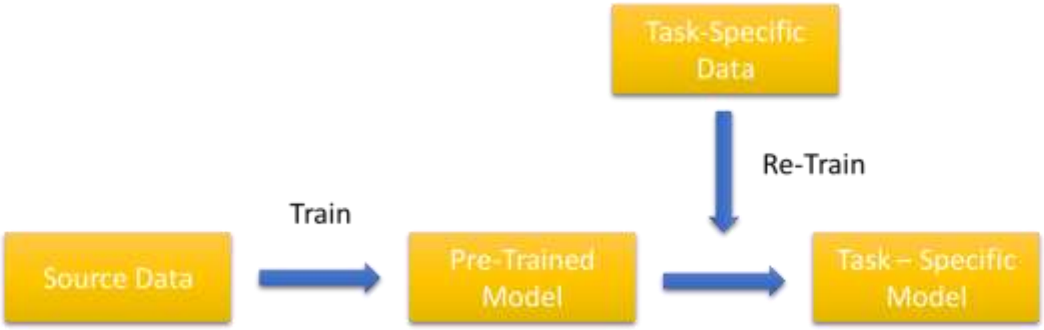

Figure 4: The Original Transfer-Learning Structure [34]**.**

The impact of transfer learning in ultrasound image categorization has received limited attention in ultrasound image processing and has not been thoroughly studied for a significant duration. Using transfer learning—which excels with sparsely labeled data—Feng and Wang [34] investigated the ultrasound tongue classification challenge. Various convolutional neural network (CNN) designs were evaluated for their ability to classify ultrasound images. Regarding the impact of TL on ultrasound picture categorization, it has so far received little attention in the field of ultrasound picture processing.

Teriyaki et al.[37] employed transfer learning to classify lesion categories such as fissure tongue (FT), coated tongue (CT), geographic tongue (GT), and moderate rhomboid glossitis (MRG), as well as normal/normal tongue (NT) pictures, utilising distinct DCNN images. Majority voting was also used for the first time in the literature to increase tongue lesion categorization accuracy.

Zhang, Jing-Xuan, et al. [38] Introduced TaLNet, a model that utilizes transfer learning from text-to-speech to recover audio based on tongue and lip movements. TaLNet utilizes an encoder-decoder framework, where tongue and lip movies are analyzed by specialized encoders that employ three-dimensional (3D) convolutional neural networks (CNNs). The study also suggested merging the concealed outputs from the tongue and lip encoders, along with the speaker code, and inputting them into the decoder to anticipate acoustic characteristics. The process of transfer learning involved initially training the multi-speaker Tacotron 2 model on a substantial text-to-speech (TTS) ensemble. Subsequently, the decoder for this model was transferred to the TaLNet module.

5. Evaluation Measures for Tongue Movement’s Estimation

The accuracy of the estimated tongue movement is assessed in various ways. These methods base their work on retrieved tongue contours, which can be done manually or automatically. The most reliable and conventional approach to comparing findings is to measure the difference between the extracted ground truth contour and the segmented tongue contour using the suggested methods.

- **Mean Sum of Distances (MSD)**

In two steps, the system automatically extracts tongue contours and compares them to ground-truth contours to get the mean sum of distances. The minimal distance between each algorithm-extracted contour element and the nearest ground truth element is established first. The closest point to the ground truth contour that each point can be evaluated is compared to the algorithm-extracted contour. In order to standardize the results, the ground truth element count is split by the total of the minimal distances from these two stages, and automatically extracted contours are used. The MSD formula is given by Equation (1).

$$MSD(u,v) = \frac{1}{m+n}\left(\sum_{i=1}^{n} mini_j\left(\left|v_j - u_i\right|\right) + \sum_{j=1}^{m} mini_j\left(\left|u_i - v_j\right|\right)\right) \quad (1)$$

In which (n) is the ground truth contour length, (m) is the automatically extracted contour length, (v_j) are the ground truth data points for the manually extracted contour, and (u_i) are the datasets for the automatically extracted contour. The closest distances between each point on the contour and the nearest point on the opposite contour are shown by $(mini)$ and $(mini_j)$, respectively [2].

- **Mean Absolute Error (MAE)**

One way to evaluate the amount of error that occurs between paired observations that describe the same phenomenon is to take the average of all absolute faults [39]. The formula is as shown in Equation (2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \quad (2)$$

- **Mean Square Error (MSE)**

It represents the average of the square of the difference between the values that were initially predicted and those that were actually observed [40]. The formula is as shown in Equation (3):

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (3)$$

- **The Structural Similarity Index (SSIM)**

Assuming that HVS is competent at extracting structural information from scenes, the structural similarity index (SSIM) becomes a full-reference image quality assessment (FR-IQA) metric. By making this feature an inherent part of an IQA metric, the authors were able to beat not just MSE-based metrics but also current top-tier perceptual image quality metrics, showing a stronger relationship with the subjective assessment given by humans, like the mean opinion score (MOS) and differential MOS (DMOS), on different IQA datasets. Because of its improved performance, simple mathematical formulation, differentiability, and a high degree of computational parallelization, SSIM has become one of the most popular FR-IQA measures in the scientific community. It has been used as a proxy evaluation for human assessment in various image processing (IP) and computer vision (CV) applications. The next section showcases a multitude of real-life examples utilizing SSIM across several domains [41].

Which evaluates the presence of three distinct types of visual impact caused by variations in luminance l, contrast c, and structure s between two images the formula is as shown in Equation (4):

$$SSIM(y, \hat{y}) = [\iota(y, \hat{y}\,)]^{\alpha}[c(y, \hat{y})]^{\beta}[s(y, \hat{y}\,)]^{\gamma} \tag{4}$$

- **Complex Wavelet Structural Similarity (CW-SSIM)**

This innovative method extends the SSIM approach to the complex wavelet domain, allowing for a picture similarity measurement that is resistant to minute distortions.

An extensively utilized metric for assessing image similarity is the Complex Wavelet Structural Similarity Index (CW-SSIM). However, its application in picture categorization is still in its early stages. The Structural Similarity Index Measure (SSIM) approach for the complex wavelet domain is improved by the CW-SSIM index. The goal was to create a measurement that was unaffected by "non-structural" geometric aberrations. This metric produces a value ranging from 0 (least similar) to 1 (most similar) [42]. The formula is as shown in Equation (5):

$$CWSSIM(y, \hat{y}\,) = \frac{2\left|\sum_{l=1}^{L} w_{y,l}, w_{\hat{y},l}^{*}\right| + K}{\sum_{l=1}^{L}\left|w_{y,l}\right|^{2} + \sum_{l=1}^{L}\left|w_{\hat{y},l}\right|^{2} + K} \tag{5}$$

- **Word Error Rate (WER)**

Among ASR metrics, the word error rate (WER) is now by far the most popular. WER works with the Edit Distance 2 idea. In a sentence, the WER (w) can be determined for any number of insertions (i), deletions (d), and substitutions (s) with a total of N_t tokens as follows the formula is as shown in Equation (6) [43]:

$$W = \frac{(i+d+s)}{N_t} \tag{6}$$

- **R square**:

It is the percentage of $y's$ variance that can be explained by the model's free parameters. It is a measure of the model's likelihood of correctly predicting unseen samples and acts as an indicator of model goodness of fit via the proportion of explained variance [44]. The R-squared (R^2) coefficient of determination is written as in Equation (7):

$$R^2 = 1 - \left(\frac{SSE}{SST}\right) \tag{7}$$

In this context, the sum of squares of the residuals is denoted by SSE, whereas the total sum of squares is denoted by SST. (Worst value = $-\infty$; Best value = $+1$)

6. Analysis and Discussion

Significant advancements in speech synthesis, speaker identification, and speech recognition have been made possible by deep learning algorithms, which have completely transformed speech processing jobs. First, the paper provides a historical context of significant developments in speech processing. Then, it quickly goes over the fundamentals of deep learning and how they might be applied to the prediction of tongue movements. We also illustrate the most current and significant deep learning research, describe the primary tasks of voice processing, and highlight the datasets used for these tasks. In the field of estimating tongue movements by taking advantage of diverse data sets, which may be signals or ultrasound images, video sequences can be employed in real-time.

Choosing the right data collection is one of the most crucial phases in developing a prediction model with artificial intelligence techniques. This research explored the many sorts of data sets that may be employed and how they impact the prediction model's accuracy. Figure (5) depicts the various forms of data sets utilized a combination of the three types which is considered being the most effective in the realm of tongue and lip motions as well as speech processing, according to the majority of research. Real-time video sequences have not been employed in any of the current studies.

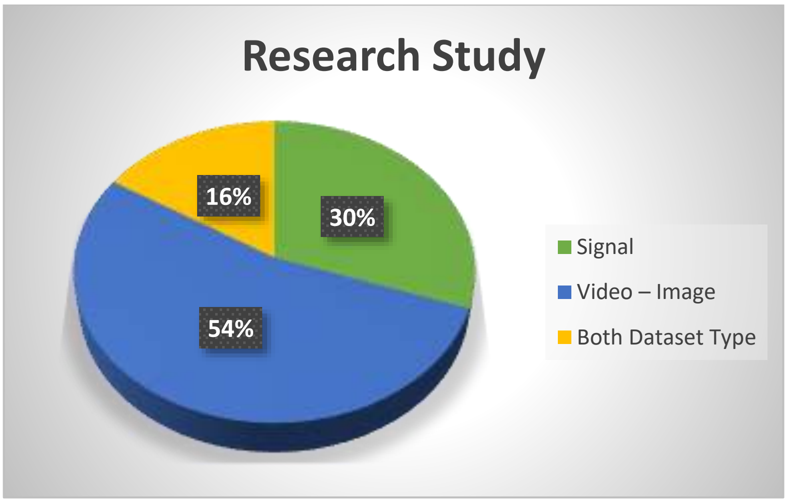

Figure 5: Dataset Type Studies.

Different tools are used to estimate tongue movements with high precision as shown in Figure (6); thus, the decision must be made based on the type of data as well as how accurately these tools handle the process of extracting tongue movements from video sequences in real time. An analysis of papers indicated that artificial intelligence approaches, represented by deep learning algorithms, It is the greatest in this sector; however, it is recommended to integrate many technologies to get the maximum potential accuracy.

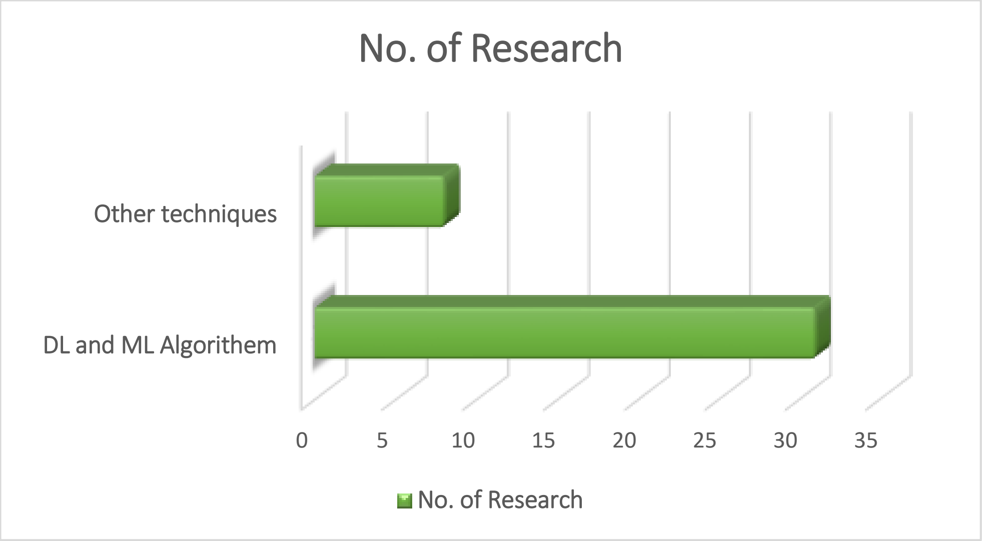

Figure 6: Studies Methodology Tools.

The method for comparing extracted and reference data remains the same, and the provided metrics demonstrate the accuracy of the process, regardless of whether reference data is mechanically or manually extracted. Some measures are relevant for both artificial and conventional intelligence approaches, while others are solely applicable to artificial intelligence techniques. The most regularly utilized measures in recent years were examined, demonstrating how accurate the prediction model is in practice. The Figure (7) demonstrates that the criteria for assessing accuracy are the most widely employed since they are critical in determining the model's appropriateness to work in practice. Similarly, criteria for assessing inaccuracy are likely to be applied to speech processing algorithms.

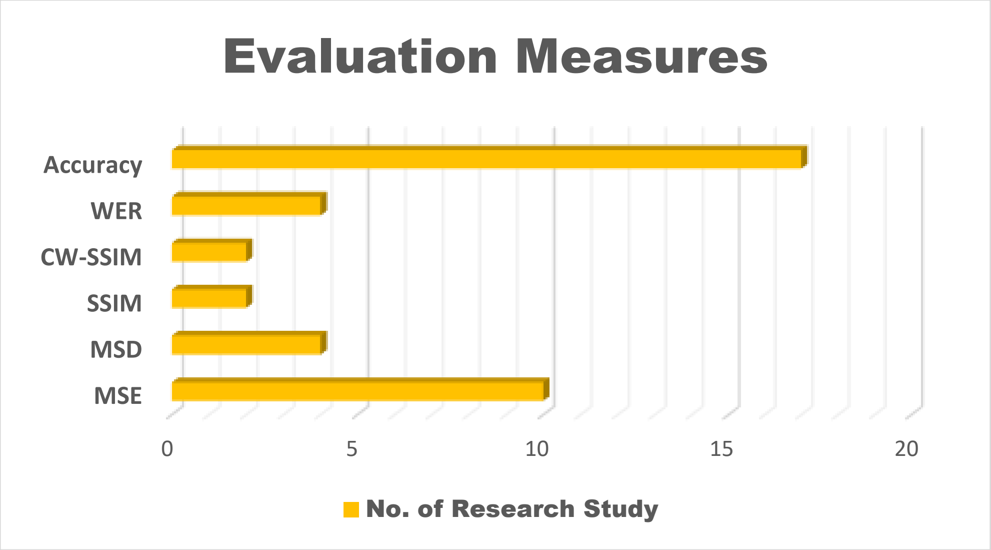

Figure 7: Most Evaluation Measures.

Although there is a dearth of research that makes use of data from real-time video sequences, a meta-analysis of numerous studies indicates that choosing the right data type for estimating tongue movements from lip movement images can be beneficial for ML and DL prediction models. To the significance of measuring tongue motions in anticipating letters and words, particularly for those who have had laryngectomies. During the research, it was discovered that the deep learning approaches employed were nearly identical in the majority of experiments, and no more than one methodology was combined. The investigation also discovered that many earlier studies used ready-made algorithms, as seen in Table 2.

Table 2: Methods of Different AI-techniques for Tongue Movement Estimation**.**

Ref.	Year	Method	Dataset	Dataset Type		Error Rate	Accuracy
				Signal	**Video - Image**		
[45]	2017	A conventional DNN-based TTS (DM-DNN) (AM-DNN)	Articulatory (tongue and lip) and acoustic data	✓	✗	RMSE in log-f0= 0.156 Voiced/Unvoiced Error= 16.01	BAP Distortion= 1.275 MCD Distortion= 5.244
[46]	2017	DNN	tongue movements + face motion	✓	✗	______	(between 13.9 and 33.2%
[47]	2018	PVIRA PCA	speech MRI data	✓	✓	______	______
[48]	2018	Snake Algorithm	The ultrasound image data	✗	✓	______	______
[11]	*2018*	3DCNN	1. "WSJ0" data, 60 frames per second. 2. "TJU" data, 30 frames per second.	✗	✓	MSE for WSJ0 = 21.7 TJU = 32.6 Cross = 154.9	______
[49]	2018	CNN LSTM	GRID audio-visual corpus	✓	✗	______	Correlation =98%
[50]	2018	DNN	ultrasound images data	✗	✓	______	correlation rate of 0.74
[51]	2019	ConvLSTM	TJU datasets	✗	✓	ConvLSTM-10th MSE=4.35 ConvLSTM-11th MSE=61.38	CW-SSIM=0.928 CW-SSIM=0.904
[52]	2019	CNN	The NS test data The Ultrax test data The UltraSpeech test data	✗	✓	MSD = 5.52 (1.65)(Ultrax) MSD = 5.72 (2.88) UltraSpeech	_______
[53]	2019	DCAE	Silent Speech Challenge dataset.	✗	✓	Word Error Rate of 6.17%	_______
[54]	2019	DNN	Permanent magnet localization (PML)	✓	✗	Q3 = 1.8 median error = 1.4 mm.	_______
[20]	2019	CNN LSTM	audio-visual database	✓	✗	______	88.2%
[55]	2019	(DNNs)	PPSD database	✗	✓	______	_______
[56]	2020	SVM	LIBSVM Data	✓	✗	______	Sensitivity (41.5%), Specificity (70.9%), Balanced accuracy (58.0%).

[26]	2020	conv+ReLU +BatchNorm	Iris Net + Tongue Net	×	✓	TongueNet MSE= 4.87 IrisNet MSE= 4.21	______
[57]	2020	DNN adaptation	ultrasound images data	×	✓		95% confidence intervals
[58]	2020	The 3D ResNet18 the LipNet architecture usetemporal augmentation and Dropout	1- (LRW) dataset 2- the GRID audiovisual corpus	✓	✓	WER 2.9% CER 1.2%	85.02% 45.24%
[59]	2020	______	1- Hungarian children' dataset. 2- Ultra Suite dataset .	×	✓	MSC (mean /std) = (178/91) SSIM = (mean/std)(0.28/0.15) CW-SSIM=(mean/std)(0.41/0.01)	______
[60]	2020	CNN BowNet wBowNet	ultrasound tongue images (Dataset I) Seeing Speech project (Dataset II)	×	✓	MSE= 0.01	L (BCE) =0.03 L(Dice)=0.06
[61]	2021	FMLLR	Tal1 + tal80				silent speech is substantially lower than on modal speech((WER))
[62]	2021	Encoder-decoder architecture.	TaL1 + TaL80	✓	✓	WER of 0.5% and 3.5% for TaL1 and TaL80	Accuracy (TaL1) = 98.5% Accuracy(TaL80)= 97.6%
[38]	2021	encoder-decoder architecture	TaL Data	×	✓	(WER)= 36.5	______
[63]	2021	(TAS)	Synthetic and real ultrasound tongue imaging dataset	✓	✓	NRMSE of < 0.15ms	______
[13]	2021	CNN LSTM	Silent Speech Challenge (SSC) data	×	✓	(MSD)= 4.953 (SSIM) = 0.728 (CW-SSIM)= 0.765	______
[64]	2021	DNN-TTS	the UltraSuite-TaL80 database	×	✓	F0 –RMSE = 10 - 54 F0-CORR = 0.2 – 0.7	Mel-Cepstral Distortion, MCD) = 5.5–6.2 dB

						F0-VUV = 6.8 – 26.6	BAF = 0.2 – 0.6
[10]	2021	(CNNs)	real ultrasound tongue imaging dataset	×	✓	Lower MSD =2.243±0.026	______
[65]	2021	1- Encoder and Decoder 2- residual convolution-and-attention (RA) Block	Three public datasets are used 1- DNS Challenge 2- Voice Bank + DEMAND 3- TIMIT Corpus	✓	✓	______	Conv-TasNet SDRi(dB)= 7.57 PESQ2= 2.14 Ours SDRi(dB)= 8.39 PESQ2= 2.14 2.50
[66]	2021	DNN-TTS PCA- (FC-DNNs LSTM)	UltraSuite-TaL80 database	×	✓	FC-DNN, the test error is 2.9, while with LSTM, the test error is 3.1.	______
[14]	2021		The ultrasound data	✓	✓	MSE = 0.315	R2 = 0.683
[67]	2022	MLP LSTM GRU	Tongue Mocap Data	✓	×	prior weight is 0.01	______
[68]	2022	The U-Net model (CNN)	The Natural Scenes Dataset (NSD)	×	✓	______	98.22%
[32]	2022	2D-CNN and 3D-CNN and CONVLSTM	The ultrasound data	______	______	MSE=0.276	______
[69]	2022	AAA v2.18 software	High-speed UTI data were acquired using a Micro machine	✓	×	______	______
[70]	2023	RetinaConv	ultrasound data	×	✓	______	IOU = 98.4 tIOU = 51.5
[71]	2023	FC – DNN	speech signal ultrasound images data	✓	✓	MSE = 0.0055	______
[72]	2023	KD-based SE	TaL80	✓	✓	______	______
[73]	2023	FC-DNN	EEG, ultrasound and speech PPBA database	✓	✓	______	______
[74]	2023	employ a method built on pseudo target generation and domain adversarial training with an iterative training (encoder and the decoder)	TaL (The Tongue and Lip) dataset	×	✓	______	______
[75]	2023	(STN) module	ultrasound images data	×	✓	______	accuracy = 92%
[76]	2023	CNN	UltraSuite-TaL corpus	×	✓	______	______

Conclusion

With differing degrees of effectiveness, many techniques have been applied to estimate tongue motions from signals, ultrasound pictures, or real-time video. Every methodology has benefits and drawbacks. This research introduced artificial intelligence approaches for estimating tongue motions from ultrasound waves, pictures, and video. Because it gives the researcher a comprehensive quantitative and qualitative assessment of methods for computing real-time tongue movements in ultrasound imagery, this review study is significant. The analysis concluded that the best way to achieve more accurate results is to use a combination of AI techniques. Machine learning works well as a method for segmenting the tongue in real time. Conversely, interactive user segmentation tools integrated into traditional training and post-processing procedures can enhance a machine learning model's output.

Source of funding: None

Conflict of interest: None

Ethical clearance: None

References

[1] K. Al-hammuri, F. Gebali, I. Thirumarai Chelvan, and A. Kanan, Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review, Diagnostics, 12(11), (2022), DOI(https://doi.org/10.3390/diagnostics12112811)

[2] C. Laporte, and L. Ménard, Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech, Med. Image Anal., 44, 98–114(2018), DOI(https://doi.org/10.1016/j.media.2017.12.003)

[3] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, A review of deep learning techniques for speech processing, Inf. Fusion, 99, (2023), DOI(https://doi.org/10.1016/j.inffus.2023.101869)

[4] Systematic review of deep learning models in ultrasound tongue imaging for the detection of speech disorders, techrxiv.org, (2023), DOI(https://doi.org/10.36227/techrxiv.22699291.v1)

[5] V. Ramanarayanan, Analysis of speech production real-time MRI, Computer Speech and Language, 52, 1–22(2018), DOI(https://doi.org/10.1016/j.csl.2018.04.002)

[6] Ö. D. Köse, and M. Saraçlar, Multimodal Representations for Synchronized Speech and Real-Time MRI Video Processing, IEEE/ACM Trans. Audio Speech Lang. Process., 29, 1912–1924(2021), DOI(https://doi.org/10.1109/TASLP.2021.3084099)

[7] K. Isaieva, Y. Laprie, A. Houssard, J. Felblinger, and P.-A. Vuissoz, Tracking the tongue contours in rt-MRI films with an autoencoder DNN approach, (2020)

[8] Z. Zhao, Y. Lim, D. Byrd, S. Narayanan, and K. S. Nayak, Improved 3D real-time MRI of speech production, Magnetic Resonance in Medicine, 85(6), 3182–3195(2021), DOI(https://doi.org/10.1002/mrm.28651)

[9] I. P. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Attention Is All You Need, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[10] M. Feng, Y. Wang, K. Xu, H. Wang, and B. Ding, “Improving ultrasound tongue contour extraction using u-net and shape consistency-based regularizer, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2021-June, 6443–6447(2021), DOI(https://doi.org/10.1109/ICASSP39728.2021.9414420)

[11] C. Wu, S. Chen, G. Sheng, P. Roussel, and B. Denby, Predicting Tongue Motion in Unlabeled Ultrasound Video Using 3D Convolutional Neural Networks, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 5764–5768(2018), DOI(https://doi.org/10.1109/ICASSP.2018.8461957)

[12] P. Saha, Y. Liu, B. Gick, and S. Fels, Ultra2Speech - A Deep Learning Framework for Formant Frequency Estimation and Tracking from Ultrasound Tongue Images, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12263 LNCS., 473–482(2020), DOI(https://doi.org/10.1007/978-3-030-59716-0_45)

[13] H. Liu, and J. Zhang, Improving Ultrasound Tongue Image Reconstruction from Lip Images Using Self-supervised Learning and Attention Mechanism, (2021), DOI(https://doi.org/10.48550/arXiv.2106.11769)

[14] L. Tóth, and A. H. Shandiz, 3D Convolutional Neural Networks for Ultrasound-Based Silent Speech Interfaces, pp. 159–169(2020), DOI(https://doi.org/10.1007/978-3-030-

61401-0_16)

[15] M. H. Mozaffari, M. A. R. Ratul, and W. S. Lee, IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision, 2019, [Online]. Available: http://arxiv.org/abs/1911.03972

[16] T. G. Csapó, F. V. Arthur, P. Nagy, and Á. Boncz, Comparison of acoustic-to-articulatory and brain-to-articulatory mapping during speech production using ultrasound tongue imaging and EEG, In: SMM23, Workshop on Speech, Music and Mind, 16–20(2023), DOI(https://doi.org/10.21437/SMM.2023-4)

[17] L. Tóth, A. Honarmandi Shandiz, G. Gosztolya, and T. G. Csapó, Adaptation of Tongue Ultrasound-Based Silent Speech Interfaces Using Spatial Transformer Networks, In: INTERSPEECH 2023, 1169–1173(2023), DOI(https://doi.org/10.21437/Interspeech.2023-1607)

[18] C. Kroos, AUDITORY-VISUAL SPEECH ANALYSIS : IN SEARCH OF A THEORY, no. August, 6–10(2007)

[19] Z. Shi, A Survey on Audio Synthesis and Audio-Visual Multimodal Processing, 2021, [Online]. Available: http://arxiv.org/abs/2108.00443

[20] L. S. Memory, Y. Lu, and H. Li, applied sciences Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based, (2019)

[21] K. Al-Hammuri, Computer vision-based tracking and feature extraction for lingual ultrasound, 2019, [Online]. Available: http://hdl.handle.net/1828/10812

[22] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, A review of deep learning techniques for speech processing, Inf. Fusion, 99, (2023), DOI(https://doi.org/10.1016/j.inffus.2023.101869)

[23] J. M. Porta, J. J. Verbeek, and B. J. A. Kröse, Active appearance-based robot localization using stereo vision, Auton. Robots, 18(1), 59–80(2005), DOI(https://doi.org/10.1023/B:AURO.0000047287.00119.b6)

[24] D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015, pp. 2410–2414(2015),

DOI(https://doi.org/10.21437/interspeech.2015-521)

[25] S. Wen, Automatic Tongue Contour Segmentation using Deep Learning,(2018)

[26] M. H. Mozaffari, N. Yamane, and W. S. Lee, Deep Learning for Automatic Tracking of Tongue Surface in Real-time Ultrasound Videos, Landmarks instead of Contours, In: Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, 2785–2792(2020), DOI(https://doi.org/10.1109/BIBM49941.2020.9313262)

[27] I. Fasel, and J. Berry, Deep belief networks for real-time extraction of tongue contours from ultrasound during speech,In: Proceedings - International Conference on Pattern Recognition, 1493–1496(2010), DOI(https://doi.org/10.1109/ICPR.2010.369)

[28] A. Dhillon, and G. K. Verma, Convolutional neural network: a review of models, methodologies and applications to object detection, Progress in Artificial Intelligence, 9(2), 85–112(2020), DOI(https://doi.org/10.1007/s13748-019-00203-0)

[29] A. Graves, Sequence Transduction with Recurrent Neural Networks Alex. 2012, [Online]. Available: http://arxiv.org/abs/1211.3711

[30] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, A review of irregular time series data handling with gated recurrent neural networks, Neurocomputing, 441, 161–178(2021), DOI(https://doi.org/10.1016/j.neucom.2021.02.046)

[31] D. B. Chklovskii, A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures Yong, 2954, 2925–2954(2017)

[32] A. H. Shandiz, and L. Tóth, Improved Processing of Ultrasound Tongue Videos by Combining ConvLSTM and 3D Convolutional Networks, In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13343 LNAI, 265–274(2022), DOI(https://doi.org/10.1007/978-3-031-08530-7_22)

[33] G. Van Houdt, C. Mosquera, and G. Nápoles, A review on the long short-term memory model, Artif. Intell. Rev., 53(8), 5929–5955(2020), DOI(https://doi.org/10.1007/s10462-020-09838-1)

[34] Y. Feng, and X. Wang, Ultrasound tongue image classification using transfer learning, ACM International Conference Proceeding Series, 38–42(2019), DOI(https://doi.org/10.1145/3379299.3379301)

[35] T. M. K. and D. W. Karl Weiss, A survey of transfer learning | SpringerLink, Journal of Big Data, 3(1), 9(2016), [Online]. Available: http://link.springer.com/article/10.1186/s40537-016-0043-6

[36] H. Daum, Frustratingly easy domain adaptation, ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 256–263(2007)

[37] B. Tiryaki, K. Torenek-Agirman, O. Miloglu, B. Korkmaz, İ. Y. Ozbek, and E. A. Oral, Artificial intelligence in tongue diagnosis: classification of tongue lesions and normal tongue images using deep convolutional neural network, BMC Med. Imaging, 24(1), 2024, DOI(https://doi.org/10.1186/s12880-024-01234-3)

[38] J. X. Zhang, K. Richmond, Z. H. Ling, and L. Dai, TaLNet: Voice Reconstruction from Tongue and Lip Articulation with Transfer Learning from Text-to-Speech Synthesis, Proc. AAAI Conf. Artif. Intell., 35(16), 14402–14410(2021), DOI(https://doi.org/10.1609/aaai.v35i16.17693)

[39] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, Mean Absolute Percentage Error for regression models, Neurocomputing, 192, 38–48(2016), DOI(https://doi.org/10.1016/j.neucom.2015.12.114

[40] D. Chicco, M. J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci., 7, (2021), DOI(https://doi.org/10.7717/PEERJ-CS.623)

[41] A. Gondia, A. Siam, W. El-Dakhakhni, and A. H. Nassar, Machine Learning Algorithms for Construction Projects Delay Risk Prediction, J. Constr. Eng. Manag., 146(1), 04019085(2020), DOI(https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736)

[42] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, Structural similarity index (SSIM) revisited: A data-driven approach, Expert Syst. Appl., 189, (2022), DOI(https://doi.org/10.1016/j.eswa.2021.116087)

[43] P. Shah, Is Word Error Rate a good evaluation metric for Speech Recognition in Indic Languages?, 2022, [Online]. Available: http://arxiv.org/abs/2203.16601

[44] D. Chicco, M. J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci., 7, 1–24(2021), DOI(https://doi.org/10.7717/PEERJ-CS.623)

[45] B. Cao, M. Kim, J. Van Santen, T. Mau, and J. Wang, Integrating articulatory information in deep learning-based text-To-speech synthesis, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017-Augus, 254–258(2017), DOI(https://doi.org/10.21437/Interspeech.2017-1762)

[46] C. Kroos, R. L. Bundgaard-Nielsen, C. T. Best, and M. D. Plumbley, Using deep neural networks to estimate tongue movements from speech face motion, In: 14th International Conference on Auditory-Visual Speech Processing, AVSP 2017, 30–35(2017), DOI(https://doi.org/10.21437/AVSP.2017-7)

[47] J. Woo, Speech Map: a statistical multimodal atlas of 4D tongue motion during speech from tagged and cine MR images, Comput. Methods Biomech. Biomed. Eng. Imaging Vis., 7(4), 361–373(2019), DOI(https://doi.org/10.1080/21681163.2017.1382393)

[48] S. Chen, Y. Zheng, C. Wu, G. Sheng, P. Roussel, and B. Denby, Direct, Near Real Time Animation of a 3D Tongue Model Using Non-Invasive Ultrasound Images, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 4994–4998(2018), DOI(https://doi.org/10.1109/ICASSP.2018.8462096)

[49] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, Lip2Audspec: Speech Reconstruction from Silent Lip Movements Video, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 2516–2520(2018), DOI(https://doi.org/10.1109/ICASSP.2018.8461856)

[50] T. Grosz, G. Gosztolya, L. Toth, T. G. Csapo, and A. Marko, F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April, 291–295(2018), DOI(https://doi.org/10.1109/ICASSP.2018.8461732)

[51] C. Zhao, P. Zhang, J. Zhu, C. Wu, H. Wang, and K. Xu, Predicting Tongue Motion in Unlabeled Ultrasound Videos Using Convolutional Lstm Neural Networks, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May. 5926–5930(2019), DOI(https://doi.org/10.1109/ICASSP.2019.8683081)

[52] J. Zhu, W. Styler, and I. Calloway, A CNN-based tool for automatic tongue contour tracking in ultrasound images, 2019, [Online]. Available: http://arxiv.org/abs/1907.10210

[53] B. Li, K. Xu, D. Feng, H. Mi, H. Wang, and J. Zhu, Denoising Convolutional Autoencoder

Based B-mode Ultrasound Tongue Image Feature Extraction, In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May, 7130–7134(2019), DOI(https://doi.org/10.1109/ICASSP.2019.8682806)

[54] N. Sebkhi, A Deep Neural Network-Based Permanent Magnet Localization for Tongue Tracking, IEEE Sens. J., 19(20), 9324–9331(2019), DOI(https://doi.org/10.1109/JSEN.2019.2923585)

[55] D. Porras, A. Sepulveda-Sepulveda, and T. G. Csapo, DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging, In: Proceedings of the International Joint Conference on Neural Networks, 2019-July,(2019), DOI(https://doi.org/10.1109/IJCNN.2019.8851769)

[56] K. J. Teplansky, Tongue and lip motion patterns in alaryngeal speech, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob, 4576–4580(2020), DOI(https://doi.org/10.21437/Interspeech.2020-2854)

[57] G. Gosztolya, T. Grósz, L. Tóth, A. Markó, and T. G. Csapó, Applying dnn adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces, Acta Polytech. Hungarica, 17(7), 109–124(2020), DOI(https://doi.org/10.12700/APH.17.7.2020.7.6)

[58] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, Can We Read Speech beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition, In: Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, 356–363(2020), DOI(https://doi.org/10.1109/FG47880.2020.00134)

[59] T. G. Csapó, and K. Xu, Quantification of transducer misalignment in ultrasound tongue imaging, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob, 3735–3739(2020), DOI(https://doi.org/10.21437/Interspeech.2020-1672)

[60] M. Hamed Mozaffari and W. S. Lee, Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data, Methods, 179, 26–36(2020), DOI(https://doi.org/10.1016/j.ymeth.2020.05.011)

[61] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, Silent versus modal multi-speaker

speech recognition from ultrasound and video, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 1, 466–470(2021), DOI(https://doi.org/10.21437/Interspeech.2021-23)

[62] M. S. Ribeiro, Tal: A Synchronised Multi-Speaker Corpus of Ultrasound Tongue Imaging, Audio, and Lip Videos, In: 2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings, 1109–1116(2021), DOI(https://doi.org/10.1109/SLT48900.2021.9383619)

[63] P. Padmini, D. Gupta, M. Zakariah, Y. A. Alotaibi, and K. Bhowmick, A simple speech production system based on formant estimation of a tongue articulatory system using human tongue orientation, IEEE Access, 9, (2021), DOI(https://doi.org/10.1109/ACCESS.2020.3048076)

[64] T. G. Csapó, L. Tóth, G. Gosztolya, and A. Markó, Speech Synthesis from Text and Ultrasound Tongue Image-based Articulatory Input, 31–36(2021), DOI(https://doi.org/10.21437/ssw.2021-6)

[65] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, Interactive Speech and Noise Modeling for Speech Enhancement, In: 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 16, 14549–14557(2021), DOI(https://doi.org/10.1609/aaai.v35i16.17710)

[66] T. G. Csapó, Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging, 7–12(2021), DOI(https://doi.org/10.21437/ssw.2021-2)

[67] S. Medina, Speech Driven Tongue Animation, In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022-June, 20374–20384(2022), DOI(https://doi.org/10.1109/CVPR52688.2022.01976)

[68] G. Li, J. Chen, Y. Liu, and J. Wei, wUnet: A new network used for ultrasonic tongue contour extraction, Speech Commun., 141, 68–79(2022), DOI(https://doi.org/10.1016/j.specom.2022.05.004)

[69] L. McKeever, J. Cleland, and J. Delafield-Butt, Using ultrasound tongue imaging to analyse maximum performance tasks in children with Autism: a pilot study, Clin. Linguist. Phonetics, 36(2–3), 127–145(2022), DOI(https://doi.org/10.1080/02699206.2021.1933186)

[70] M. H. Mozaffari, M. A. R. Ratul, and W. S. Lee, IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision, Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.03972

[71] T. G. Csapó, F. V. Arthur, P. Nagy, and Á. Boncz, Towards Ultrasound Tongue Image prediction from EEG during speech production, 1164–1168(2023), DOI(https://doi.org/10.21437/interspeech.2023-40)

[72] R. C. Zheng, Y. Ai, and Z. H. Ling, Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement through Knowledge Distillation, 844–848(2023), DOI(https://doi.org/10.21437/interspeech.2023-780)

[73] T. G. Csapó, F. V. Arthur, P. Nagy, and Á. Boncz, Comparison of acoustic-to-articulatory and brain-to-articulatory mapping during speech production using ultrasound tongue imaging and EEG, Sep. 2023, 16–20(2023), DOI(https://doi.org/10.21437/smm.2023-4)

[74] R. C. Zheng, Y. Ai, and Z. H. Ling, Speech Reconstruction from Silent Tongue and Lip Articulation by Pseudo Target Generation and Domain Adversarial Training, In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2023, 1–5(2023), DOI(https://doi.org/10.1109/ICASSP49357.2023.10096920)

[75] L. Tóth, A. H. Shandiz, G. Gosztolya, and T. G. Csapó, Adaptation of Tongue Ultrasound-Based Silent Speech Interfaces Using Spatial Transformer Networks, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2023-Augus, 1169–1173(2023), DOI(https://doi.org/10.21437/Interspeech.2023-1607)

[76] I. Ibrahimov, G. Gosztolya, and T. G. Csapo, Data Augmentation Methods on Ultrasound Tongue Images for Articulation-to-Speech Synthesis, In: 12th ISCA Speech Synthesis Workshop (SSW2023), Aug. 2023, 230–235(2023), DOI(https://doi.org/10.21437/ssw.2023-36)