computers

MDPI

Article

Multimodal Lip-Reading for Tracheostomy Patients in the Greek Language

Yorghos Voutos [1], Georgios Drakopoulos [1], Georgios Chrysovitsiotis [2], Zoi Zachou [2], Dimitris Kikidis [2], Efthymios Kyrodimos [2] and Themis Exarchos [1,*]

[1] Department of Informatics, Ionian University, 49100 Corfu, Greece; c16vout@ionio.gr (Y.V.); c16drak@ionio.gr (G.D.)

[2] Voice Clinic, Medical School, National and Kapodistrian University of Athens, 11527 Athens, Greece; chrysovi@gmail.com (G.C.); zoizachou@gmail.com (Z.Z.); dimitriskikidis@yahoo.com (D.K.); timkirodimos@hotmail.com (E.K.)

* Correspondence: exarchos@ionio.gr; Tel.: +30-2661087855

Abstract: Voice loss constitutes a crucial disorder which is highly associated with social isolation. The use of multimodal information sources, such as, audiovisual information, is crucial since it can lead to the development of straightforward personalized word prediction models which can reproduce the patient's original voice. In this work we designed a multimodal approach based on audiovisual information from patients before loss-of-voice to develop a system for automated lip-reading in the Greek language. Data pre-processing methods, such as, lip-segmentation and frame-level sampling techniques were used to enhance the quality of the imaging data. Audio information was incorporated in the model to automatically annotate sets of frames as words. Recurrent neural networks were trained on four different video recordings to develop a robust word prediction model. The model was able to correctly identify test words in different time frames with 95% accuracy. To our knowledge, this is the first word prediction model that is trained to recognize words from video recordings in the Greek language.

Keywords: tracheostomy; lip reading; deep learning; multimodal interfaces

check for updates

Citation: Voutos, Y.; Drakopoulos, G.; Chrysovitsiotis, G.; Zachou, Z.; Kikidis, D.; Kyrodimos, E.; Exarchos, T. Multimodal Lip-Reading for Tracheostomy Patients in the Greek Language. *Computers* **2022**, *11*, 34. https://doi.org/10.3390/computers11030034

Academic Editor: Fernando Bobillo

Received: 13 January 2022
Accepted: 25 February 2022
Published: 28 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The human voice is a fundamental characteristic of human communication and expression. It is produced by an organ named larynx which tunes the vocal cords during exhalation. Each person has a unique and recognizable voice, the loss of which constitutes a significant disorder. One of the biggest causes of partial or complete voice loss is tracheostomies. Other causes of voice loss include neurological diseases as well as laryngeal and thyroid cancer. The increase in the age and the number of people with severe disabilities who must bear permanent tracheostomies have contributed to the increase of population experiencing difficulties in voice communication. In the U.S., the number of patients who undergo a tracheostomy is over 100,000 per year [1]. The increase of hospitalization rates in intensive care units and the number of patients at risk of speech loss has increased by 11% over the last 10 years [2] since 40% of the hospitalized patients are tracheostomy candidates. In Greece, the increased smoking rate (more than double the European average) has led to an increase in total number of laryngectomies due to laryngeal cancer (95% in smokers [3]).

In most voice loss cases, the situation does not improve over time and the available voice restoration methods often have poor results. As a result, individuals with partial or total voice loss are socially isolated; often exhibiting derivative disorders such as depression and cognitive function impairment which in turn lowers their quality of life. The currently available surgical solutions for partial speech restoration are often not satisfactory both in terms of quality and understanding, as well as, in the degree of patient adoption. The time bar of voice restoration is shown in Figure 1.

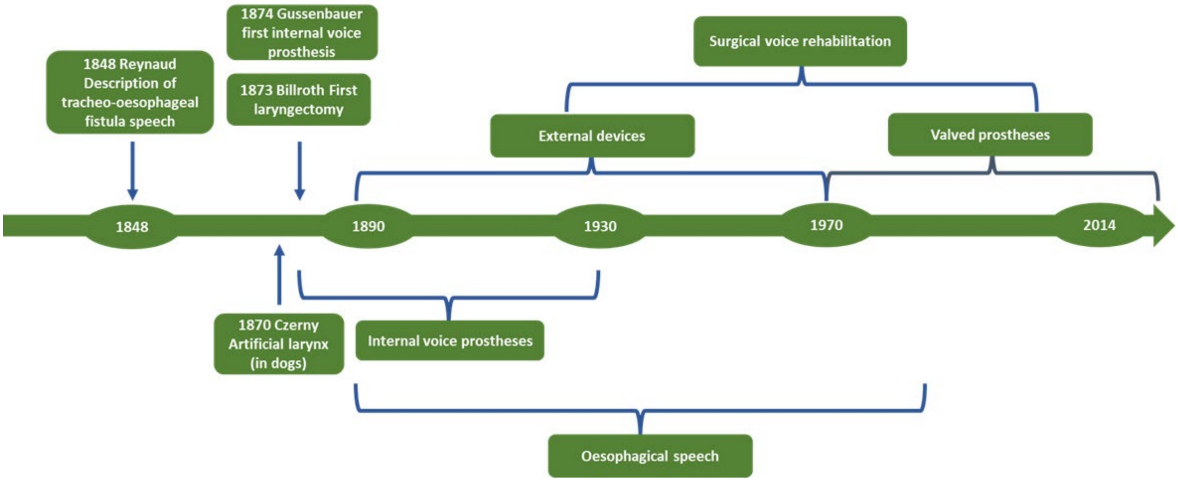

1874 Gussenbauer first internal voice prosthesis
1848 Reynaud Description of tracheo-oesophageal fistula speech
1873 Billroth First laryngectomy
Surgical voice rehabilitation
External devices
Valved prostheses
1848
1890
1930
1970
2014
1870 Czerny Artificial larynx (in dogs)
Internal voice prostheses
Oesophagical speech

Figure 1. The time bar of voice restoration [4].

In some cases, the amount of training required is disproportionate to the resulting voice quality, while in other cases there are significant health complications, while almost always constant patient monitoring and training is required [5].

In the present paper, a system for enabling voice loss patients to communicate with their natural voice tone, is presented. By leveraging deep learning and image processing techniques the proposed algorithm uses frames of the patient's lips to predict what they are saying. Our intention is to develop a high-performance DL model that will be able to capture the patient lips from video recordings to provide word predictions in the Greek language. To our knowledge, this is the first DL-motivated approach to shed light into word prediction based on extracted lips from the Greek language, where there are no models reported. More specifically, the proposed computational pipeline consists of three core stages: (i) the image pre-processing stage, (ii) the training stage, where a straightforward architecture of a recurrent neural network (RNN) is designed and utilized for the training process on a set of annotated lip frames from four different video recordings (9810 frames in total), and (iii) the validation of the proposed RNN on a set of unseen annotated frames. In the image pre-preprocessing stage, the frames are automatically extracted from the video recordings based on the fps (frames per second) ratio. Then, lip segmentation is applied based on a pre-defined mask to isolate the lip area from each individual frame. Frames with bad quality (i.e., motion artifacts, asynchronous or wrong timing) are discarded from further processing. The frames are then automatically annotated based on the provided annotation file. The latter includes the time intervals for each word. The frames are finally split into training and testing frames. The prediction performance of the proposed RNN was favorable, yielding 84% accuracy in the last 500 epochs and 85% in the last 100 epochs during training (a set of 1500 total epochs was used for the training process). To our knowledge, this is the first DL-empowered model that can capture Greek words from extracted lips with increased performance.

The paper is structured as follows: in Section 2 the state of the art in the existing multimodal AI methods for lip reading is presented. The data collection process is described in Section 3. The architecture of the proposed lip-reading system is then presented in Section 4 along with the proposed word prediction model and its high-level deep neural network architecture. A summary of the results of the proposed model is presented next through a case study along with the limitations and the future work which are described in Sections 5 and 6, respectively.

2. Existing Work

Multimedia as a research topic is more than just a combination of diverse data modalities [6], i.e., audio, video, text, graphics, etc. Essentially, it is the amalgamation and the interaction among these divergent media types that lead to the formation of some extremely challenging research opportunities such as that of speech reading and reconstruction [7].

One of the most fundamental problems in multimodality is that of speech reading and reconstruction systems. This problem belongs to one of those research areas which cross domains and exploit the full breadth and depth of multimedia research since it involves not just speechreading but also the synchronization of video with lip movements as well as reconstruction of the audio. Speech is composed by phonemes that are considered to be the smallest distinct and detectable units of sound in a given language [8]. The task of speechreading is made difficult by the fact that often several phonemes correspond to a single viseme, thus producing ambiguity when trying to infer speech from visemes only. In the recent years, Automatic Speech Recognition (ASR) systems have gained significant traction, with systems being deployed widely in cars, mobile phones, homes, refs. [9–12].

However, considering settings such as that of a car on a road, it is not ideal and is plagued by very low signal to noise ratio [13]. This leads to speech recognition systems failing utterly in the absence of usable audio [14]. However, these fallacies with ASR systems can be solved satisfactorily by deploying speech reading and reconstruction systems which can augment the understanding of ASR systems or can even reconstruct the speech for them.

On the other hand, professional lipreaders are required for rendering their services in these cases, which are not only expensive but highly limited. The single solution for all these challenges is a system having speech reading and reconstruction capabilities which can also effectively integrate with ASR systems. Some of the earliest works reported in the field of lipreading are those by Potamianos et al. [15] and Lan et al. [16]. Recent works such as those by Cornu and Milner [17] and Akbari et al. [18] perform the task of lipreading on the GRID database [19]. Previous works [17,18], focused on single view based lipreading, often with hand-picked features.

3. Data Sharing

To evaluate the performance of the word prediction model we acquired four different video recordings from a patient before the tracheostomy operation as described in [20]. Out of four recordings, two had large time duration (163 words) whereas the rest of them had small time duration (122 words). The text that was used for the video recordings was designed by the experts in such a way to reflect all the phonemes in the Greek language. Emphasis was given on the type of the text that was used for each video recording in terms of: the number of words, the average word length, the number of syllables, the average syllable length, the phoneme sets, the total phonemes and the syllabic structure.

A 12 MP camera was used for the video recording process with 1080 p resolution at 30 fps (frames per second). Each video recording was conducted with the patient sitting in front of a white background. The camera was placed at the foreground at 50 cm height. The patients were asked to read the text clearly with a normal pronunciation rate as in their everyday live. An open-source software (Shotcut$^{®}$) was used for video editing to correct motion and vocal artifacts. Then, the video recording was converted into an audio file (.wav file) using Shotcut$^{®}$. Finally, Praat was used to load the audio file and adjust it at a 1.2 s level to generate annotated words for different time durations in the video recording.

More specifically, the word annotation process was done by combining the speech of the patient with the spectrogram in Praat. This was achieved by matching the information from the curves regarding the height, intensity, and the vocal pulses. Then, the text was divided into words and for each word the starting and ending time point was manually annotated by the experts in milliseconds (ms). The time intervals which correspond to silence were annotated as “sil” whereas the speech errors as “err”. This process was applied for each video recording yielding an individual align file.

In this work, four different video recordings (.avi format) of the same patient were used to demonstrate the performance of the word prediction model. The video recordings were obtained upon an agreement fulfilling all the necessary legal and ethical requirements. The data collection process took place in the Voice clinic, Medical School, National and Kapodistrian University of Athens (NKUA), Athens, Greece. The length of the four video recordings were 55 s, 72 s, 51 s, and 69 s, respectively. The total number of frames was 1650, 3360, 1530, and 3270, respectively. The dimension of each frame was 1920 (height) × 1080 (width), with 96 dpi and 34 bit depth.

Upon the data collection process, the experts analyzed the video recordings and created the align files which were used for the frame annotation process. In fact, each video recording has a unique. align file, which includes three fundamental columns. The first column refers to the starting time point of each word (in milliseconds), the second column refers to the ending time point (in milliseconds) and the final column corresponds to the word. In each video recording, the patient was asked to read a different text. Each text was properly designed by experts to include all the available phonemes of the Greek language and thus to enhance the applicability and impact of the proposed method for word prediction.

4. The Proposed Architecture

The proposed architecture is presented in Figure 2 and consists of two fundamental layers that separate the final mobile application (which is located at the top of the architecture) from the analysis and pre-processing of the video recordings (at the bottom): The data pre-processing and deep learning layer (layer 1) and the mobile application layer (layer 2). At the first layer, data pre-processing workflows are applied on the available video recordings (and align files) to extract the video frames, isolate the lip area from each frame, group the frames based on the align files into words. The extracted lip frames from each word are then utilized for the training process where sequence to sequence deep learning algorithms, such as, the recurrent neural networks (RNNs) are trained on a subset of training lip frames and tested on the remaining subset.

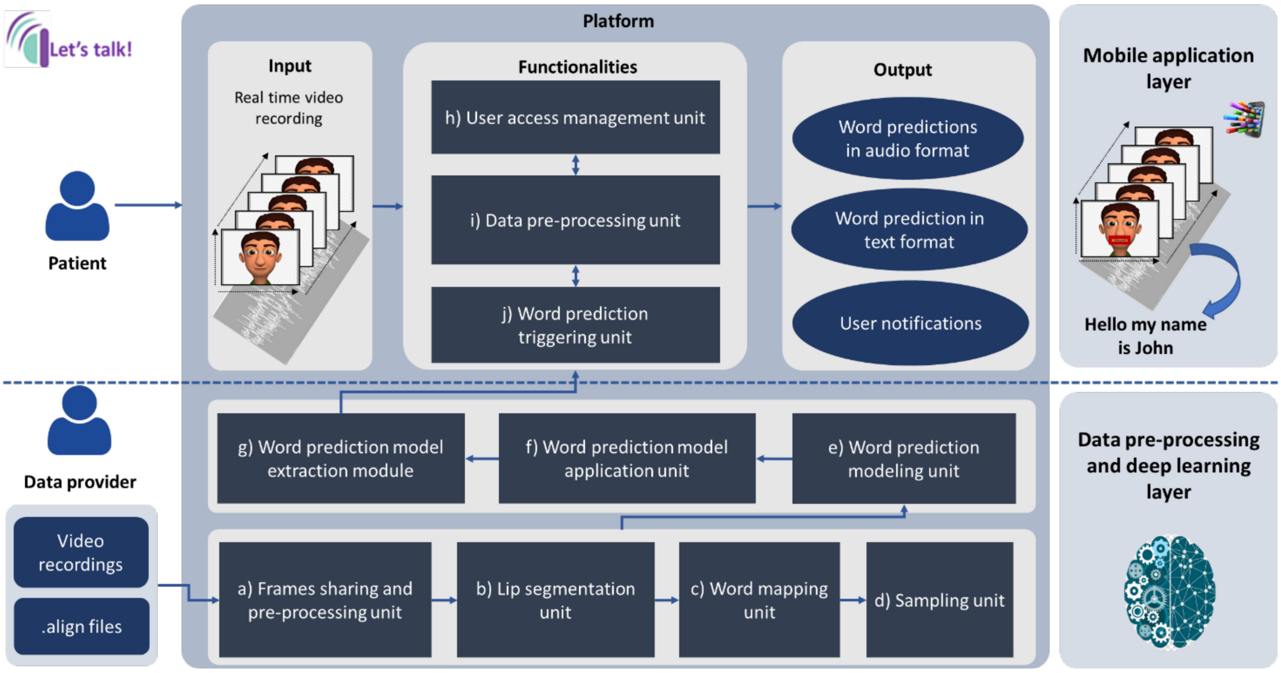

Let's talk!
Platform
Input
Real time video recording
Functionalities
h) User access management unit
i) Data pre-processing unit
j) Word prediction triggering unit
Output
Word predictions in audio format
Word prediction in text format
User notifications
Mobile application layer
Hello my name is John
Patient
Data provider
Video recordings
.align files
g) Word prediction model extraction module
f) Word prediction model application unit
e) Word prediction modeling unit
a) Frames sharing and pre-processing unit
b) Lip segmentation unit
c) Word mapping unit
d) Sampling unit
Data pre-processing and deep learning layer

Figure 2. The Let's talk! architecture.

The data pre-processing and deep learning layer consists of the following units:

(a) the frames sharing and pre-processing unit that is responsible for sharing data and exporting frames per video recording,

(b) the lip segmentation unit of the lip area that is responsible for the automated isolation of the lip area (lip segmentation) per frame and per video recording,

(c) the word mapping unit that is responsible for matching the frames to words taking into account the align file that contains the words that the patient mentions together with the corresponding start and end times per word,

(d) the sampling unit responsible for sub- or sub-sampling of frames to maintain the same number of frames per word in each recording,

(e) the word prediction modeling unit responsible for training Recurrent neural network (RNN) multi-level architectural neural networks by inputting the word frames of each record, where the total number of classes is equal to the number of different words of each recording,

(f) the word prediction model application unit that is responsible for controlling the performance of the individual (or not) model in test subsets using metrics such as accuracy and sensitivity,

(g) the word prediction model export unit that aims to export personalized (or not) models in a format compatible with the mobile application.

The mobile application level: At this level the word prediction is implemented in real time. This level includes the following units:

(a) the user access management unit that is responsible for certifying/verifying the user who wishes to log in to the application,

(b) the data pre-processing unit which is responsible for the extraction of the frames, the division of the lip area as well as the loading of the relevant personalized (or not) model that corresponds to the user of the application,

(c) the word prediction triggering unit for the provision of verbal predictions giving as input the frames of the user's video recording.

Furthermore, the proposed architecture includes two core engines: (i) the data preprocessing engine and (ii) the model design and deployment engine. The former is responsible for extracting the lip area from the frames for each word (and for each video recording) which are in turn used by the model design and deployment engine. More specifically, the frames are automatically extracted from the raw video recordings (based on the fps ratio) and the lip area is segmented from each frame using a pre-defined mask. Then, using the alignment files, the frames are labeled with words. To ensure that that number of frames is the same for each word either upsampling or downsampling methods were used. In the model design and deployment engine, a straightforward sequence-to-sequence neural work (RNN) is trained on the extracted frames for word prediction. An 80/20 train/test split process is finally used to evaluate the performance of the word prediction model.

In addition, prediction performance measures, such as, the accuracy and the word error rate (WER) were used to quantify the performance of the RNN on a test subset. The model design and deployment engine is also responsible for extracting the word prediction model in a compatible format that can be easily integrated into mobile applications.

At the output of the Let's talk! mobile application, the user receives the following: (i) word predictions in audio format, (ii) word predictions in text format, (iii) user notifications regarding the use of the application and its execution time model.

4.1. Data Preprocessing

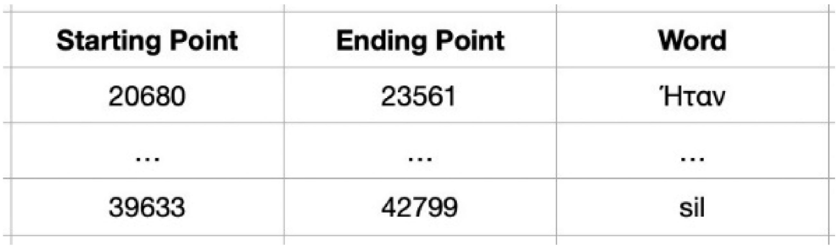
(1) Extraction of frames from the video recordings: The frames are extracted from the video recordings of the patient using "OpenCV" [21]. Since the fps ratio in the video recordings is 30 for a duration 1 min the total number of extracted frames would be 1800 frames. In this word, four different video recordings of the same patient were analyzed with time durations 55 s, 1 min and 12 s, 51 s, 1 min and 9 s, respectively. The

dimension of each extracted frame was 1920 (height) × 1080 (width) with horizontal and vertical resolution set to 96 dpi followed by 34-bit depth.

(2) Isolation of the lip area across the extracted frames: The lip area is segmented from each frame using a 68-point mask with the facial coordinates only, i.e., eyes, eyebrows, nose, lips, etc., as described in [22]. The mask was generated by the analysis of the ibug 300-W dataset [23]. The coordinates that correspond to the lip area were extracted from this mask and the "dlib" library was used to isolate the lip area [23]. The dimensions of the extracted frames with the lip area was 120 (width) × 80 (height). These frames were then placed into folders. Each folder has a unique name/identifier which corresponds to the word. The word (annotation) is located in the third column of the align file). In the case of duplicated words, a unique identifier is added at the end of the word's folder name to separate its frames from the rest. The overall lip segmentation procedure was applied on each individual frame across the folders, upon the frame extraction from the four video recordings.

(3) Word Mapping: according to Section 3, each video recording is accompanied by an alignment file (i.e., a align file) which includes the following information: (i) the starting and the ending timepoint (i.e., time interval) of each word, and (ii) the word that corresponds to this time interval which is recorded by the expert. Then, each word is assigned to a particular group of frames with the lip area to divide the frames into words and train the word prediction model. To do so, the starting and ending timepoints are converted to frame indices by multiplying the time interval with the fps ratio and dividing by a scalar to map the starting and ending time points to starting and ending frames. This scalar depends on the time precision that the experts included in the align files (in some cases division with either 1000 or 10,000 was preferred based on the provided time intervals). Upon the transformation process, the frames are collected with the lip area between the start and end frames for each word. The frames were then classified into folders bearing the names of the words/acts. An automated diagnostic approach was also developed to ensure whether the number of frames per word in each folder is correct. To this end, three cases were identified and corrected where the time intervals in the corresponding align files were erroneously parsed and thus the relevant number of frames was larger than expected. A sample of the align file is shown in Figure 3.

(4) Equally numbered frames per word: A fundamental aspect of the RNN training process lies on the fact that it is based on an equal number of frames per word. To address this need, either upsampling or downsampling is applied on the exported frames to ensure that the number of frames will be stable across each folder. This stable number of frames is defined by the word with the largest number of frames (excluding the "sil" or "err" frames) per video recording. Here, the largest number of frames was equal to 25. Hence, the number of frames across each folder was adjusted to 25.



Starting Point	Ending Point	Word
20680	23561	Ήταν
...	...	...
39633	42799	sil

Figure 3. Sample of the align file structure for the first patient.

More specifically, all the frames (with the lip area, on each folder) are properly adjusted according to the sampling procedure which is described below (the frames of the folders including silence frames or errors are excluded from this procedure, Figure 4):

(1) In the case where the number of frames is greater than the stable number of frames per folder, down-sampling was applied by randomly removing one frame per time until the required number of frames (i.e., 25 frames) is met.

(2) In the case where the number of frames is less than the stable number of frames, "sil" (or silence) frames are sequentially added inside the folder until the number of frames is equal to the stable number of frames. The "sil" frame was selected as the one where the patient's lips remain closed. In addition, the "sil" frame was the same for all video recordings.

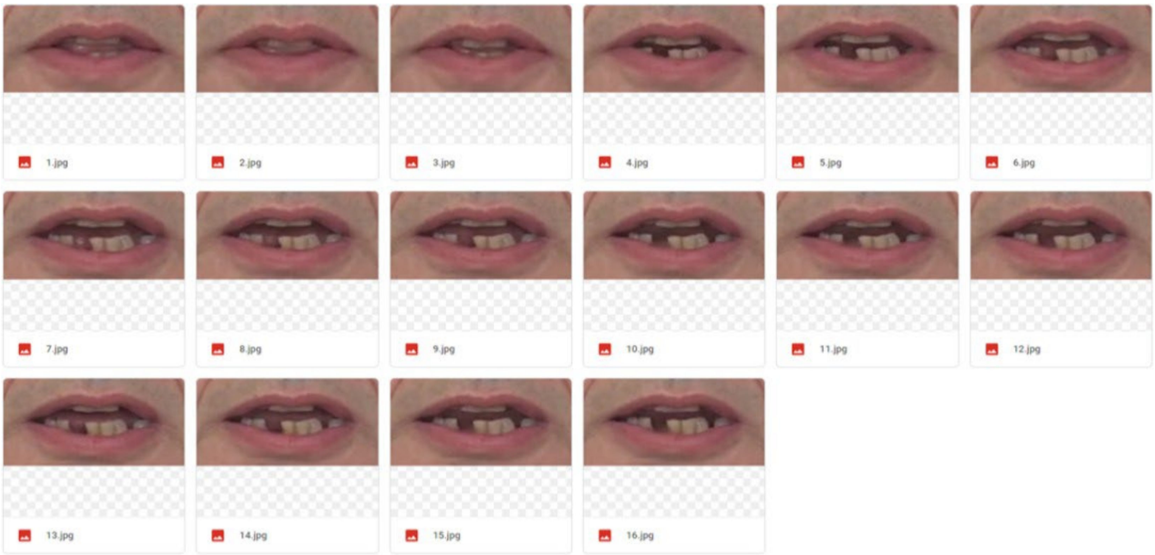


Figure 4. Partitioning frames with lip area.

4.2. Model Development

The input data are properly transformed into a three-dimensional (3D) structure with dimensions, say M × N × K, where M is the number of frames, N is the height of each frame and K is the width of each frame. All frames that are used as input in the model must contain the segmented lip area in pre-defined dimensions as described before. A set of frames is then created by re-configuring the dimensions of the 3D structure into (M × N) × K. For example, in the case where a video recording has 40 words, where the dimensions of the frames (with the lip area) have a height 80 and a width 160, the dimensions of the 3D structure will be 25 × 80 × 160 whereas the dimensions of the re-configured 3D structure would be 2000 × 160.

4.3. Word Prediction Model Architecture

The proposed word prediction model adopts a recurrent neural network (RNN) architecture which belongs to the family of the sequence-to-sequence models. The layers, the output dimensions, and the parameters of the proposed network for word prediction are described in Table 1. In brief, the RNN-based architectures consist of a directed graph, where the time dimension is added in the model to enable the dynamic analysis of the information that travels within the network. This is a key characteristic of the RNNs since they can be used for time sequence analysis with voice recognition applications [24,25]. A widely used RNN-based architecture for voice recognition applications is the long short-term memory (LSTM) one [26] which has been deployed in the current work.

Table 1. Levels and parameters of the proposed model.

Layer (Type)	Output Dimensions	Parameters
LSTM (LSTM)	(None, 200, 256)	427,008
LSTM_1 (LSTM)	(None, 200, 128)	197,120
LSTM_2 (LSTM)	(None, 64)	49,408
Dense (Dense)	(None, 64)	4160
Dense_1 (Dense)	(None, 40)	2600

Here, we adopt the LSTM architecture instead of the other RNN-based ANNs to deal with the vanishing gradient problem [26]. According to the latter, the weights of the gradients either tend to be zero (i.e., disappear) or tend to infinity during the minimization of the gradient of the loss function. To solve this issue, the LSTM makes use of temporary memory vectors (buffers) which allow for the robust calculation of the gradient loss without introducing any effects during the training process.

(1) Neural Network Architecture: The model supports the training process for exported frames with the lip area per word across multiple video recordings. The input frames are then transformed into a suitable data structure to be provided as input to the deep learning neural network. Next, the neural network architecture is designed (input dimensions, number of levels, types of levels, output dimensions). After the design of the network architecture, the hyper-parameter optimization is performed where important internal parameters of the network are defined, such as the activation functions and optimized for the minimization of the loss function.

Finally, the neural network is trained with the input data and the trained model is received at the output. If the frames with the lip area per word from multiple logs are valid, the frames are transformed into datasets, the network architecture is designed, the network parameters are defined and finally the training of the neural network takes place.

If the trained model with the best performance is compatible, then it is converted into a file extension which is compatible with the mobile application layer (Figure 2, architecture of the proposed system) and finally the model is exported. If the model export process is successfully completed then the trained model with the best performance is exported to a safe location that will be accessible from the mobile application.

If the model evaluation is successful, then the in-depth word prediction model with the best performance/predictive ability (i.e., with the highest accuracy and the smallest word error rate) is given at the output of the workflow. The latter is then exported in a proper format (e.g., HDF5) to support its easier integration into mobile applications.

The architecture of the neural network is presented in Table 1 and in Figure 5 and consists of five levels, three of which use 256, 128 and 64 LSTM units (neurons) respectively and finally two dense layers to feed the information taken from the LSTM units in 64 and 40 neurons, respectively, to obtain the information from the 64 LSTM units and extracted to the desired dimension on exit (40 neurons). The model was trained in 1500 epochs with time step 2000 ($N \times M$). In the first LSTM level 256 units were used, with input dimension $(N \times M) \times K = 2000 \times 160$. At the second layer 128 units with an output dimension of 128×1 were used while in the last one, 64 units with an output dimension of 64×1 were used. The values of these are fed to the 1st Dense layer that has 64 units and uses the ReLu (Rectified Linear Unit) activation function, outputting a one-dimensional vector of 64 values (64×1).

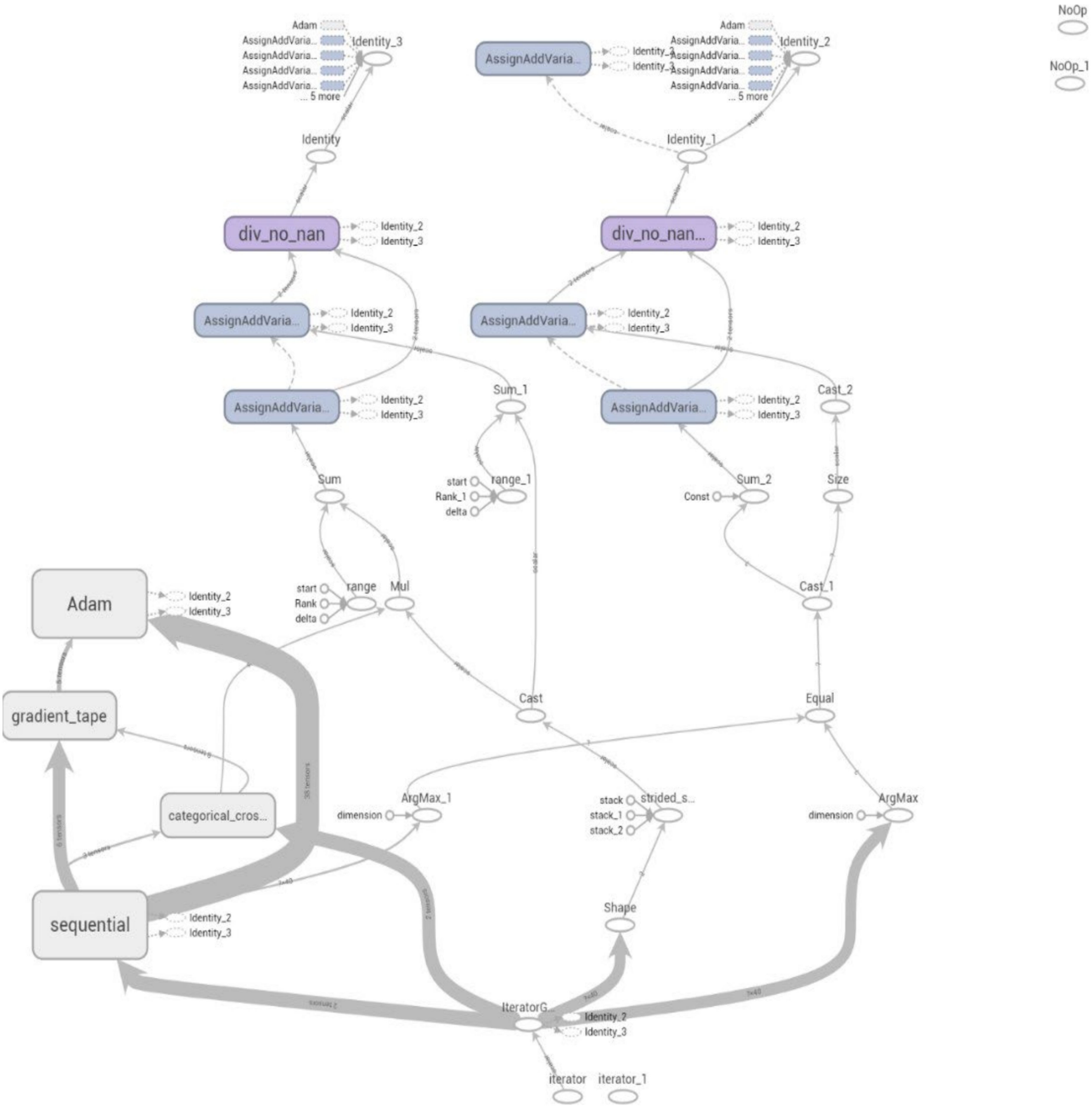


Figure 5. Representation of the proposed deep learning neural network architecture.

At the second dense layer 40 units were used with the activation function softmax to distinguish the information by outputting 40 values each value is a probability of the input belonging to the corresponding class. The class with the highest probability (i.e., the class corresponding to the position with the probability in the vector of 40 values) is taken as the predominant class. The softmax function focuses on the application of the the standard exponential function to each element of the input vector, say $\boldsymbol{x}$, and normalizes these values by dividing by the sum of all these exponentials as in:

$$\sigma(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{n=1}^{N} e^{x_n}}, \tag{1}$$

where N is the size of the input vector.

The adam optimizer was used to utilize exponentially moving averages, computed on the gradient evaluated on a current mini-batch to scale the learning rate as in:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{2}$$

and:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g^2{}_t, \tag{3}$$

where m and v are moving averages, g is gradient on current mini-batch, and the betas are the hyperparameters of the Adam optimization algorithm. In this work, the Adam optimizer was utilized with $\beta_1 = 0.9$, $\beta_2 = 0.999$. For the process of annotation of words, the one-hot encoding by creating an $H \times H$ table, where H is the number of classes (in the in this case 40) where the diagonal elements of the table are set at 1 and the other items the value 0. This assigns vectors to words.

According to Table 1, the total number of network parameters was 680,296 (427,008 in the first layer, 197,120 in the second layer, 49,408 in the third layer, 4160 in the fourth layer and 2600 in the final layer). This allows for the distillation of knowledge from the lip frames but on the other hand it introduces an increased computational burden to the deep learning model. For this reason, the training process took place in an Aurora R12, with the following specifications: 1 × Intel(R) 11th Gen (3.6 GHz to 5 GHz w/Intel® Turbo Boost Max) with eight CPU cores, 16 GB RAM, NVIDIA GeForce RTX 3070 8 GB GDDR6 LHR (5888 CUDA cores) and 1 TB SSD. Upon the finalization of the training process across the 1500 epochs, the resulting deep learning model was exported into a pickle format to enable its easier integration to mobile applications that support the tensorflow lite application. The whole architecture was implemented in a Python environment using "OpenCV" [21], "keras" [27] and "tensorflow" [28].

(2) Evaluation: The evaluation of the word prediction model takes place in two phases. In the first phase, the model is verified in real time during the training process on verification subsets using the accuracy and training loss. In the second phase, the performance of the model is verified in words unknown for the model from different time frames using the word error rate (WER).

5. Results

The accuracy of the model during the training process is shown in Figure 6 which confirms the upward trend of accuracy in increasing the number of epochs giving a maximum accuracy of 84% in the last 500 epochs (85% in the last 100 epochs). However, it is necessary to use more data input frames to avoid overfitting phenomena that affect the performance of the model during training, noting fluctuations in the value of accuracy during the epochs.

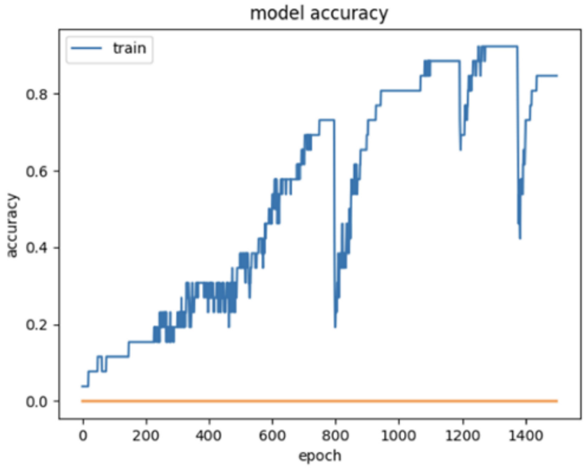


Figure 6. Accuracy of prediction model in training from 1 to 1500 epochs.

The unit accepts as input the trained deep learning model for word prediction. The parameters of the trained model are then loaded (e.g., number of levels, types of levels, activation functions). The subset of the test frames is then determined by sorting a number of frames with the lip area from the available recordings. At this point it is important to note that the subset of test frames consists of word frames that have not been given for model training in the previous stage. In addition, because emphasis will be placed on the development of both personalized and non-personalized prediction models, the subset of test frames consists of word frames in multiple logs. Finally, the predictive capacity of the deep learning neural network is evaluated using performance quantification metrics, such as accuracy, sensitivity, and specificity, among others.

In addition to the training loss and verification study, for external validation, we tested the prediction performance of the model in 20 words in different timeframes. The model was able to correctly identify 19 words giving the highest probability to the corresponding position—class with WER = 0 and accuracy = 95%. This highlights the prediction performance of the model. Furthermore, the proposed model was tested in the context of the 40 words (classes) that were trained to investigate overfitting effects yielding favorable outcomes with stable training loss (average loss equal to 2.38 during validation and 4.54 during training) which highlight the increased resilience of the proposed model against overfitting effects during the training with more than 40 words. We have also compared the performance of the proposed RNN with a conventional CNN network where the performance of the CNN was significantly smaller than the RNN (training accuracy = 78%). The resulting model was able to correctly identify the 16 out of 20 test words giving the highest probability to the corresponding position—class (accuracy = 80%).

If the user′s pre-trained word prediction model as well as the frames with the lip area in real time are valid then the procedure of initializing the word prediction model parameters follows as well as the application of the pre-trained word prediction model in frames with the lip area. If the application of the pre-trained word prediction model is valid then the predicted words are converted to both audio and text format as well as the control of the application usage and the execution time of the model. The predictions are made in real time and are displayed on the user′s screen in the form of text (subtitles) but also with audio. Otherwise, the workflow is terminated.

6. Discussion and Future Work

The presented neural network architecture is modern and is the first neural network implemented to recognize Greek words while having low loss and high percentages of verification accuracy. However, it is necessary to collect more data from frames to be given as input during network training in order to deal with overfitting phenomena. In addition, the proposed architecture has been trained in word frames that correspond to video recordings of patients covering a wide range of phonemes in the Greek language.

Then the requirements of the four users of the system (patient, data provider, data analyst, cloud computing system administrator) are analyzed with the necessary use case diagrams. Next, the system requirements are presented that follow the FURPS standard, which emphasizes the requirements of functionality, usability, reliability, performance and supportability. This work focuses on the description of the system′s architecture which consists of two levels, the data pre-processing and in-depth learning level and the mobile application level. The first level concerns the pre-processing and analysis of the available data (video recordings and align files) for the training of personalized deep learning models and consists of the units for sharing and pre-processing of frames, isolation (partitioning) of the lip area, word mapping, sampling, creating word prediction models, verifying word prediction models, exporting word prediction models. The second level concerns the prediction of words in real time and consists of the units of identification and management of users, preparation of data and sorting of the word prediction model, application of the word prediction model.

Based on the requirements modeling, the basic units of the Let's talk! architecture and the dependencies between them are defined. Next, the individual components of each subsystem are determined, the relationships between them, how they interact, and the hardware components used to run the system, and how they are installed in those components of each subsystem. The design of Let's talk! units will be completed as part of the development and control work of the individual components of the system.

Emphasis is placed on the description of the methodology which provides functions for: (i) pre-processing of image frames, (ii) isolation (segmentation) of the lip area, (iii) mapping of words for matching (iv) sampling to maintain the exact number of frames per word in each recording, (v) creating word prediction models through the training of a multi-layer recurrent neural network (RNN). Finally, emphasis is placed on verifying the word prediction model using the accuracy and exporting it in a mobile-compatible format.

The proposed neural network architecture was designed in a way similar to the LipNet's multi-layer neural network architecture which is currently the state-of-the-art model for speech recognition for the English language. However, the proposed model is not affected by specific speech patterns which are required by LipNet during training (e.g., standard speech frequency) nor on the pre-defined number of neurons (e.g., the proposed model can be easily extended to support more layers). Besides, this model is the first deep learning model that provides robust word predictions for the Greek language.

The current work emphasizes the requirements of the users of the Let's talk! application! as well as the application architecture emphasizing the four main users of the system (patient, data provider, data analyzer, cloud computing system administrator). The two-tier architecture of the system consists of 10 units related to: (i) sharing and pre-processing of frames, (ii) isolation of the lip area, (iii) word mapping, (iv) sampling, (v) creating word prediction models, (vi) verifying word prediction models, (vii) exporting word prediction models, (viii) identifying and managing users, (ix) preparing data and sorting word prediction models, (x) implementing of the word prediction model.

The rationale of Let's talk! is based on the training of multilayer (deep learning neural networks) recurrent neural networks (RNN) with the aim of predicting words from video recordings with high performance in terms of word prediction accuracy and low character error rate (character error rate) and word errors rate. To this end, data providers (clinics) provide video recordings (with audio information and align files) of patients. The frames are extracted from the video recordings automatically and the lip areas are isolated with image partitioning techniques.

The start and end times of the words from the video recordings are recorded by the data providers (clinical) and are automatically assigned to the exported frames with the area of the lips which are grouped by the words. Over- or down-sampling techniques are applied to the grouped frames to maintain a common set of frames per video recording and per patient. Personalized (or non-personalized) deep learning algorithms are trained and verified in grouped frames per word to identify words from the exported frames with the lip area and then stored in a secure location. Models are loaded into the application to provide real-time word predictions in both audio and text format.

Furthermore, the fact that the proposed deep learning model can be exported in a HDF5 (hierarchical data format) or pickle format enhances its potential towards its easier and more effective integration in mobile applications. Besides, since the computational framework that is used for the design and implementation of the proposed LSTM is based on TensorFlow, it can be easily extended to support TensorFlow lite in mobile applications.

In the next phase the neural network architecture will be tested in more unknown words and will be extended to more classes (e.g., 80, 100) but also to frames of more patient words (it is estimated that data will be gathered from 30 patients by the end of 2021). In addition, the neural network will be optimized to be able to manage word prediction from phonemes in case the algorithm is called to recognize an unknown word.

The model will be adjusted to accept as input the user's pre-trained model of word prediction as well as frames with the lip area in real time. The next step concerns the initialization of the model parameters, the application of the model in real time to the data from the input log (in the frames with the lip area), the collection of the predicted words and their conversion both in audio and text format and finally in the application user control and the execution time of the word prediction model.

The available size of the input data will also be increased a fact that will definitely enhance the prediction performance of the proposed word prediction model, as well as further enhance its generalizability. Furthermore, we also plan to include additional text information with more phonemes combinations to enhance the applicability of the proposed word prediction model. Emphasis will also be given on the optimization of the real-time lip segmentation process, where the facial mask introduces a significant computational burden to the overall word prediction process.

In addition, the deep learning neural network architecture is highly sustainable and can be easily generalized as soon as the user provides as input a valid align file and a valid set of frames from a video recording with the same fps and resolution like the original ones. Since the training process is based on a high-performance Aurora R12 the model can be extracted in small computational time. The core component which introduces the highest computational cost lies on the training of the deep neural network.

Additional computational cost is introduced during the lip segmentation process (as mentioned before). Moreover, a main issue during the real-time execution lies on providing synchronized word predictions without any significant delays. To solve this issue, we use temporal buffers according to which the system provides word predictions when a stuck of 25 frames is filled (the number of 25 frames is indicative for the purpose of this work; this number will be replaced by the corresponding largest number of words per video recording). Thus, the word prediction model has been programmed to provide word predictions per 25 frames in order to provide synchronized, real-time word predictions.

The inclusion of additional high-performance equipment is therefore necessary in order to expand the network architecture and at the same time reduce the computational complexity of the training process. The deep learning neural network architecture will be also extended to support "tensorflow lite" so that its real-time execution on mobile applications will be feasible. Since the implementation took place in "keras" using the "tensorflow" package, its complete transformation to "tensorflowlite" is expected to be easier.

Finally, the word prediction model will be extended to provide the word predictions not only to text format but also to audio format as well as to provide notifications regarding the use of the application and the execution time of the model for the connected user. These word predictions will be properly transformed into a personalized audio format, where the unique vocals of each patient will be integrated in the spectrum of the audio information to provide a more realistic effect of the word predictions through the mobile application. The overall goal is to provide not only personalized word predictions but also to develop a generalized word prediction model that can be used for lip reading in the general population. As a future work we will focus on the application of different methods, developed for English language and their customization to Greek words and comparison, in order to identify the optimal method for word identification.

Author Contributions: Y.V., G.D. and T.E. conceived the study, designed the methodology, implemented the deep neural networks and drafted the manuscript. G.C., Z.Z. D.K. and E.K. provided clinical data, performed image annotation and performed the validation. All authors have read and agreed to the published version of the manuscript.

Funding: The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: HFRI-FM17-579), "Let's Talk!: Application of multimodal interfaces for people with loss of voice for speech reproduction with natural voice".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the Medical School of the the National and Kapodistrian University of Athens.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheung, N.H.; Napolitano, L.M. Tracheostomy: Epidemiology, indications, timing, technique, and outcomes. *Respir. Care* **2014**, *59*, 895–915. [CrossRef] [PubMed]
2. Owings, M.F.; Kozak, L.J. *Ambulatory and Inpatient Procedures in the United States*; National Center for Health Statistics, Vital Health Statistics: Hyattsville, MD, USA. Available online: www.cdc.gov/nchs/data/series/sr13/sr13139.pdf (accessed on 10 January 2022).
3. Kikidis, D.; Vlastarakos, P.V.; Manolopoulos, L.; Yiotakis, I. Continuation of smoking after treatment of laryngeal cancer: An independent prognostic factor? *ORL J. Otorhinolaryngol. Relat. Spec.* **2012**, *74*, 250–254. [CrossRef] [PubMed]
4. Lorenz, K.J. Rehabilitation after total laryngectomy—A tribute to the pioneers of voice restoration in the last two centuries. *Front. Med.* **2017**, *4*, 81. [CrossRef] [PubMed]
5. Dwivedi, R.; Jallali, N.; Chisholm, E.; Kazi, R.; Clarke, P.; Rhys-Evans, P.; Elmiyeh, B. Surgical voice restoration after total laryngectomy: An overview. *Indian J. Cancer* **2010**, *47*, 239–247. [CrossRef] [PubMed]
6. Shah, R.; Zimmermann, R. *Multimodal Analysis of User-Generated Multimedia Content*; Springer: Berlin/Heidelberg, Germany, 2017.
7. Shah, R.; Yu, Y.; Zimmermann, R. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In Proceedings of the 22nd ACM International Conference Multimedia ACM, Orlando, FL, USA, 3–7 November 2014; pp. 607–616.
8. Shaywitz, S.E. Dyslexia. *Sci. Am.* **1996**, *275*, 98–104. [CrossRef] [PubMed]
9. Benoit, C.; Lallouache, T.; Mohamadi, T.; Abry, C. A set of French visemes for visual speech synthesis. *Talk. Mach. Theor. Models Des.* **1992**, 485–501.
10. Jachimski, D.; Czyzewski, A.; Ciszewski, T. A comparative study of English viseme recognition methods and algorithms. *Multimed. Tools Appl.* **2017**, *77*, 16495–16532. [CrossRef]
11. Allen, J.R.; West, D.M. How Artificial Intelligence Is Transforming the World. 2018. Available online: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/ (accessed on 10 January 2022).
12. Business Wire. European $1.66 Billion Speech and Voice Recognition Market Analysis 2016–2018 Forecast to 2025—Key Players are Microsoft, Nuance Comms, and iFlytek. Available online: https://www.businesswire.com/news/home/20180417005875/en/European-1.66-Billion-Speech-Voice-Recognition-Market (accessed on 10 January 2022).
13. Kumar, Y.; Aggarwal, M.; Nawal, P.; Satoh, S.I.; Shah, R.R.; Zimmermann, R. Harnessing ai for speech reconstruction using multi-view silent video feed. In Proceedings of the 26th ACM International Conference Multimedia, Seoul, Korea, 22–26 October 2018; pp. 1976–1983.
14. Li, J.; Deng, L.; Haeb-Umbach, R.; Gong, Y. *Robust Automatic Speech Recognition: A Bridge to Practical Applications*; Academic Press: Cambridge, MA, USA, 2015.
15. Potamianos, G.; Neti, C.; Luettin, J.; Matthews, I. Audio-visual automatic speech recognition: An overview. *Issues Vis. Audio Vis. Speech Process.* **2004**, *22*, 23.
16. Lan, Y.; Theobald, B.-J.; Harvey, R.; Ong, E.-J.; Bowden, R. Improving visual features for lip-reading. In Proceedings of the 2010 International Conference on Audio-Visual Speech Processing Hakone, Kanagawa, Japan, 30 September–3 October 2010.
17. Le Cornu, T.; Milner, B. Reconstructing intelligible audio speech from visual speech features. In Proceedings of the Sixteenth Annual Conference International Speech Communication Association, Dresden, Germany, 6 September 2015; pp. 3355–3359.
18. Akbari, H.; Arora, H.; Cao, L.; Mesgarani, N. Lip2AudSpec: Speech reconstruction from silent lip movements video. *arXiv* **2017**, arXiv:1710.09798.
19. Alghamdi, N.; Maddock, S.; Marxer, R.; Barker, J.; Brown, G.J. A corpus of audio-visual Lombard speech with frontal and profile views. *J. Acoust. Soc. Am.* **2018**, *143*, EL523–EL529. [CrossRef] [PubMed]
20. Papathanasiou, I.; Protopapas, A. Voice and speech evaluation protocol in Greek. In Proceedings of the 28th World Congress of the International Association of Logopedics and Phoniatrics (IALP), Athens, Greece, 22–26 August 2010.
21. Beyeler, M. *Machine Learning for OpenCV*; Packt Publishing Ltd.: Birmingham, UK, 2017.
22. Gavras, S.; Baxevanakis, S.; Kikidis, D.; Kyrodimos, E.; Exarchos, T. Towards a Personalized Multimodal System for Natural Voice Reproduction. In Proceedings of the 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Corfu, Greece, 29–30 October 2020; pp. 1–5. [CrossRef]
23. Sagonas, C.; Antonakos, E.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge: Database and results. *Image Vis. Comput.* **2016**, *47*, 3–18. [CrossRef]
24. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
25. Li, J.; Zhao, R.; Hu, H.; Gong, Y. Improving RNN transducer modeling for end-to-end speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 114–121.

26. Shewalkar, A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *J. Artif. Intell. Soft Comput. Res.* **2019**, *9*, 235–245. [CrossRef]
27. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
28. Dillon, J.V.; Langmore, I.; Tran, D.; Brevdo, E.; Vasudevan, S.; Moore, D.; Patton, B.; Alemi, A.; Hoffman, M.; Saurous, R.A. Tensorflow distributions. *arXiv* **2017**, arXiv:1711.10604.