applied sciences

MDPI

Article

Frame-Based Phone Classification Using EMG Signals †

Inge Salomons *,‡, Eder del Blanco ‡, Eva Navas, Inma Hernáez and Xabier de Zuazo

HiTZ Basque Center for Language Technology, University of the Basque Country, Ingeniero Torres Quevedo Plaza, 1, 48013 Bilbao, Spain; eder.delblanco@ehu.eus (E.d.B.); eva.navas@ehu.eus (E.N.); inma.hernaez@ehu.eus (I.H.); xabier.dezuazo@ehu.eus (X.d.Z.)

* Correspondence: inge.salomons@ehu.eus

† This paper is an extended version of our paper published in Proceedings of IberSPEECH 2022.

‡ These authors contributed equally to this work.

Abstract: This paper evaluates the impact of inter-speaker and inter-session variability on the development of a silent speech interface (SSI) based on electromyographic (EMG) signals from the facial muscles. The final goal of the SSI is to provide a communication tool for Spanish-speaking laryngectomees by generating audible speech from voiceless articulation. However, before moving on to such a complex task, a simpler phone classification task in different modalities regarding speaker and session dependency is performed for this study. These experiments consist of processing the recorded utterances into phone-labeled segments and predicting the phonetic labels using only features obtained from the EMG signals. We evaluate and compare the performance of each model considering the classification accuracy. Results show that the models are able to predict the phonetic label best when they are trained and tested using data from the same session. The accuracy drops drastically when the model is tested with data from a different session, although it improves when more data are added to the training data. Similarly, when the same model is tested on a session from a different speaker, the accuracy decreases. This suggests that using larger amounts of data could help to reduce the impact of inter-session variability, but more research is required to understand if this approach would suffice to account for inter-speaker variability as well.

Keywords: EMG signals; phone classification; silent speech interfaces; human–computer interaction; speech processing

Citation: Salomons, I.; del Blanco, E.; Navas, E.; Hernáez, I.; de Zuazo, X. Frame-Based Phone Classification Using EMG Signals. *Appl. Sci.* **2023**, *13*, 7746. https://doi.org/10.3390/app13137746

Academic Editors: Francesc Alías, Zoraida Callejas Carrión, António Joaquim da Silva Teixeira and José Luis Pérez Córdoba

Received: 31 May 2023
Revised: 27 June 2023
Accepted: 29 June 2023
Published: 30 June 2023

1. Introduction

This paper presents a study on classifying phones (speech sounds) using electromyographic (EMG) signals obtained from the recently developed Spanish ReSSInt-EMG database. This database is part of the ReSSInt project [1], which aims to restore speech for laryngectomees using an EMG-based silent speech interface (SSI). Laryngectomees are individuals whose larynx (voice box) has been surgically removed, and as a result, they are no longer able to produce speech naturally and thus depend on alternative methods to communicate verbally. There exist three main options for voice restoration after laryngectomy, namely esophageal, tracheoesophageal, and electrolaryngeal speech. However, each of these alternative speaking methods has some limitations [2].

For this reason, important research efforts are dedicated to developing technological solutions to overcome those limitations. Technological approaches to restore speech for laryngectomees include personalized text-to-speech systems, voice conversion, bionic voices, lean-AI approaches, and SSIs, among others [3].

The ReSSInt project of which the current study is part of aims to create a database and research the potential of developing an SSI for Spanish laryngectomees. Most of the works and databases related to SSIs [4–6] have been developed for English, and there are some for other languages [7–11]. However, none of these works focus on Spanish, and therefore, this project intends to narrow that gap.

SSIs aim to convert non-acoustic biosignals into text or acoustic speech [12,13]. Biosignals refer to the product of chemical, electrical, physical, and biological processes taking place during speech production, such as neural activity, articulator motor control, muscle activity, articulatory gestures, the vibration of the vocal folds, and pulmonary activity. Technologies to capture these biosignals include vocal tract imaging [14], magnetic tracing [15], electroencephalogram [16], and EMG [17,18]. The conversion from these silent biosignals to audible speech can be done directly—using some machine-learning algorithms that model the relationship between the feature vectors extracted from the biosignals and the acoustic signals [5,19]—or indirectly—by first producing the related text [20–22] and then using a text-to-speech (TTS) model to generate synthetic speech.

The non-acoustic biosignals that are used in this work are EMG signals or, more specifically, surface (i.e., non-invasive) EMG [23]. Electromyography is a technique used to measure and record the electrical activity of muscles. When a muscle is active, it produces an electrical signal, called an action potential, that can be detected by an electrode placed on the skin over the muscle. Since for this study we are interested in speech, we target muscles in the face and the neck.

In order to develop an EMG-to-speech SSI, a large database of EMG and speech data is required. The main idea is to obtain a model trained on large amounts of parallel EMG and speech data. To ensure the generalization capabilities of the models, it is important to use a diverse and representative dataset for training. However, the process of acquiring the data is complex and presents a number of difficulties.

Two prominent challenges in the development of these interfaces are the dependency of the trained models on the session (session dependency) and on the speaker (speaker dependency). Session dependency arises from the variations observed in the obtained EMG signals when electrodes are positioned differently on the subject's face. Speaker dependency is due to differences in the way of speaking from person to person. Additionally, an important issue arises from inadequate adhesion of the electrodes to the skin, leading to the detachment of electrodes over time and the generation of noisy signals. As a consequence, long sessions are difficult to carry out, thus limiting the amount of data available per session.

EMG signals have been previously used to perform phone classification [24,25], syllable identification [9], word recognition [11,26], continuous speech recognition [27], speaker recognition [28,29], and direct speech generation [22,30–32]. In this study, we perform a set of phone classification experiments using data from different speakers and sessions. Classifying phones offers a straightforward means of gaining valuable insights into the information conveyed by each muscle involved in the speech production process, making it an advantageous task for studying a setup performance [33].

This work is an extension of the study presented in [34], which describes a set of experiments designed to validate the acquisition setup of the newly developed ReSSInt-EMG database. Using data from nine recording sessions, in our previous study, we compared the performance of the new database with that of a comparable subset extracted from the well-known EMG-UKA Trial Corpus [4]. The results of the phone classification experiments performed on both databases reassured us of the established data acquisition procedures. In this paper, we extend the experiments and analysis to newly acquired data and analyze the speaker and session dependency of the results while at the same time improving the classification and feature-reduction methods.

This paper is structured as follows: Section 2 describes the data acquisition setup, including the recording procedure and the electrode setup, as well as the ReSSInt-EMG database, feature extraction method, and the phone classification experiments. The results of the experiments are described in Section 3, which are then interpreted and discussed in Section 4.

2. Materials and Methods

This section provides a thorough description of our study's methodology. Specifically, we detail the materials and procedures used to record the database and provide comprehensive information on its contents. We also describe the methodology employed to calculate the features extracted from the EMG signals. Finally, we describe the classification experiments conducted in this research.

2.1. Acquisition Setup

This section describes in detail the devices used to record the database, the methodology employed to identify reference points on participants' faces, and our approach to mitigating inter-session variability. Additionally, we provide comprehensive information regarding the set of tracked muscles and outline the procedure used to select them.

2.1.1. Recording Procedure

Each session is recorded in a soundproof room using a silent computer in an attempt to reduce interference with the audio and EMG signals as much as possible. The EMG signals are recorded with a Quattrocento bio-electrical amplifier at a sampling frequency of 2048 Hz, and the voice is captured with a Neumann TLM103 (diaphragm) microphone with a sampling frequency of 16 kHz.

For the acquisition and synchronization of the audio and EMG signals, we use publicly available software (https://github.com/cognitive-systems-lab/EMG-GUI, accessed on 1 March 2022), which also includes a user interface. Additionally, a camera captures a video of the facial movements, which is meant to provide supplementary data and allow multi-modal experiments in the future, such as automatic lip reading. For this paper, the video data are not considered. See Figure 1 for a photo of the complete acquisition setup.

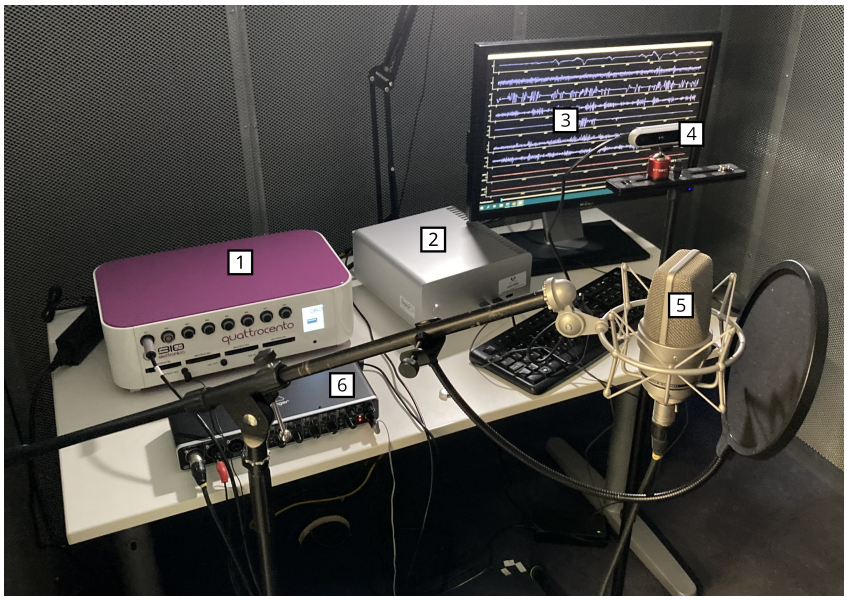

Figure 1. Acquisition setup: (1) electromyographic (EMG) signal amplifier; (2) silent computer; (3) computer screen; (4) camera; (5) microphone; and (6) audio interface.

In order to reduce inter-session variability in audio and video as much as possible, the positions of the subject, microphone, and video camera are kept constant for all sessions. Furthermore, a personalized 3D mask (Figure 2) is used to ensure that the electrode locations remain constant throughout all sessions. Prior to the first session with each speaker, we locate the positions of the electrodes using reference points and a measuring tape. To give an example, to locate the risorius, or laughing muscle, we position the first electrode adjacent to the corner of the mouth and place the second electrode in the direction of the

earlobe on the same side of the face. We mark three points: one on each outer side of both electrodes and one in the middle. This process is repeated for all eight electrode pairs, resulting in a total of 24 reference points. A 3D-printing professional then creates a 3D scan of the face and prints a mask with holes corresponding to the reference points. During subsequent sessions, we draw the points again on the subject's face using the holes and place the electrodes accordingly.

Prior to each recording session, speakers are instructed to articulate their speech slightly more than usual. A supervisor is always present in the room to ensure the correct pronunciation of the utterances.

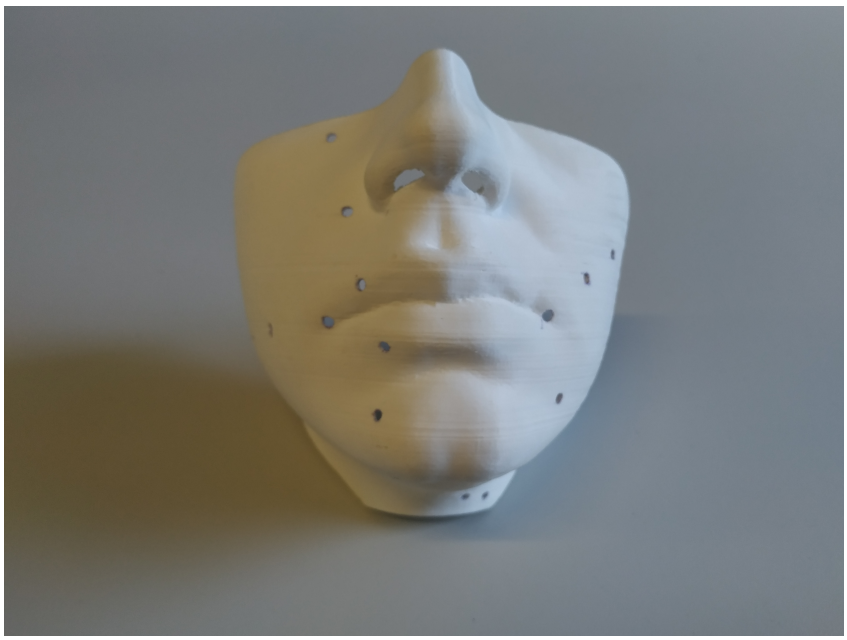

Figure 2. A personalized 3D mask. The holes are used as reference points to find the positions of the electrodes on the subject's face.

2.1.2. Electrode Setup

Previous studies have employed various approaches in determining the optimal electrode setup, such as targeting muscles specifically [31,35–38], analyzing anatomical regions [20], and looking for patterns in a high-density electrode setup [39]. Knowing that an activation potential travels along the muscle as a wave, the most appropriate way to use bipolar acquisition is to place the two electrodes longitudinally over the muscle. We decided to target muscles individually and performed a pilot study that consisted of targeting all superficial muscles in the face and neck to find the muscles that were most useful for the task. The final setup (see Figure 3) slightly differs from those used in the previously mentioned studies. These are the targeted muscles (using one channel each):

1. Levator labii superioris (channel 1)
2. Masseter (channel 2)
3. Risorius (channel 3)
4. Depressor labii inferioris (channel 4)
5. Zygomaticus major (channel 5)
6. Depressor anguli oris (channel 6)
7. Anterior belly of the digastric (channel 7)
8. Stylohyoid (channel 8)

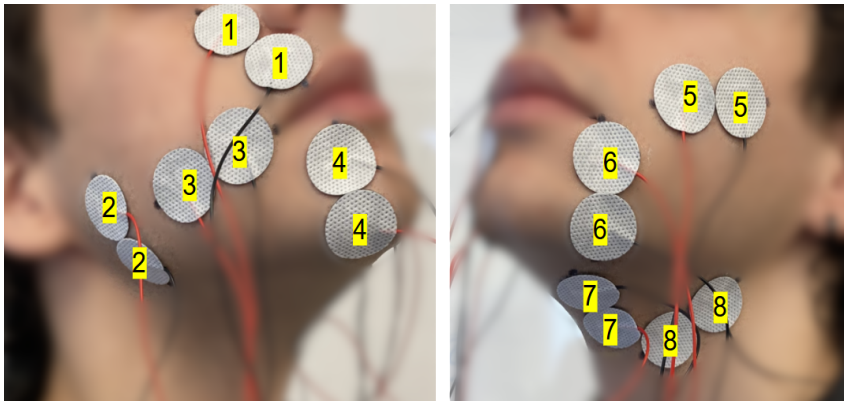

Figure 3. Electrode setup for the ReSSInt-EMG database showing the eight bipolar electrode pairs (eight channels), each targeting a different muscle. The numbers correspond to the channel that each electrode pair captures (see Section 2.1.2).

2.2. The ReSSInt-EMG Database

Table 1 shows the details of the currently recorded sessions of the ReSSInt-EMG database, namely, 16 sessions in total from 6 different speakers. Note that the acquisition process is still ongoing and that the final database will be larger. The complete database also includes data from laryngectomees, since they are our final target users. However, the data from laryngectomees cannot be used for the phone classification experiments in this study since aligned audio signals are required in order to obtain labeled segments.

Table 1. Speaker and session information for the ReSSInt-EMG database. The duration is expressed in the format of mm:ss and is limited to the portion of each session that includes the sentences.

Speaker	Gender	Age	Session	Duration	Train	Test
001	M	29	101	16:51	13:28	03:23
			102	17:32	14:04	03:28
			103	17:00	13:48	03:12
			104	19:22	15:14	04:08
002	F	29	101	25:25	20:20	05:05
			102	30:34	24:27	06:07
			103	22:36	18:17	04:19
			104	27:06	21:18	05:48
003	M	51	101	24:38	19:50	04:48
			102	21:43	17:27	04:16
004	F	46	101	26:04	20:46	05:18
			102	24:09	19:17	04:52
005	M	45	101	23:39	18:56	04:43
			102	22:31	18:00	04:31
006	F	61	101	32:57	26:21	06:36
			102	29:01	23:21	05:40

In each recording session, three different kinds of items are recorded, namely: non-sense words including vowel–consonant–vowel (VCV) structures, isolated words, and sentences. The sentences are taken from the Sharvard Corpus [40] and from a text corpus called Ahosyn that was developed to record TTS databases [41] (see Table 2).

For the current experiments, we only used the signals corresponding to the Sharvard and Ahosyn sentences and not the VCV combinations or isolated words. The number of Ahosyn sentences for each session is smaller than the number of Sharvard sentences because they are generally longer.

Each session is split into 80% training and 20% testing data. During the recording process, utterances are presented in a unique and random order for each session. To ensure consistency, we assigned the final 20% of each set of sentences as the testing set prior to the experiment. This approach ensures that the time of recording within each session is unrelated to the train–test split, and the utterances designated as the testing set remain constant for each speaker.

Table 2. Corpus information for ReSSInt-EMG sessions.

Session	Corpus
all	110 VCV combinations 100 isolated words Sharvard sentences 1-100
101	Sharvard sentences 101-400
102	Sharvard sentences 401-700
103	Ahosyn sentences 1-150
104	Ahosyn sentences 151-300

Each utterance is segmented at the phone level using the Montreal Forced Aligner [42]. The phonetic dictionary was created using the Aholab transcriber, which uses the Spanish SAMPA phone set, comprising 29 phones. Initial and final silences were removed, while short pauses between words were considered in the classification experiments.

2.3. Feature Extraction

After removing the direct-current offsets from the EMG signals and normalizing them, five time-domain (TD) features are calculated as proposed in [38]. Similar parameters with small variations have also been used in [25,43]. The procedure to obtain these TD features is described here for clarity purposes.

First, the signal ($x[n]$) is separated into two components: a low-frequency signal ($w[n]$) and a high-frequency signal ($p[n]$). To obtain the low-frequency signal, $w[n]$, a double average of $x[n]$ is calculated using a nine-point window:

$$w[n] = \frac{1}{9}\sum_{k=-4}^{4} v[n+k], \quad \text{where } v[n] = \frac{1}{9}\sum_{k=-4}^{4} x[n+k] \tag{1}$$

Having calculated $w[n]$, we can then obtain the high-frequency signal $p[n]$ by subtracting $w[n]$ from $x[n]$:

$$p[n] = x[n] - w[n]. \tag{2}$$

A rectified version $r[n]$ of the high-frequency signal is also obtained, given by:

$$r[n] = \begin{cases} p[n], & \text{if } p[n] \geq 0 \\ -p[n] & \text{if } p[n] < 0 \end{cases} \tag{3}$$

Once $w[n]$, $p[n]$, and $r[n]$ are obtained, the set of five time-domain features of a frame is defined as follows:

$$TD0 = [\bar{w}, \bar{r}, P_w, P_r, z] \tag{4}$$

where:

$$\bar{w} = \frac{1}{N}\sum_{n=0}^{N-1} w[n], \quad \bar{r} = \frac{1}{N}\sum_{n=0}^{N-1} r[n] \tag{5}$$

$$P_w = \frac{1}{N}\sum_{n=0}^{N-1} |w[n]|^2, \quad P_r = \frac{1}{N}\sum_{n=0}^{N-1} |r[n]|^2 \tag{6}$$

$$z = \sum_{n=1}^{N-1} g(p[n]p[n-1]), \quad \text{where } g(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \end{cases} \tag{7}$$

To provide the classifier with temporal context, a stacking filter concatenates the features of $2k+1$ adjacent frames. Specifically, the stacked feature vector of the j-th frame, denoted by $S(f_j, k)$, is given by:

$$S(f_j, k) = [f_{j-k}, f_{j-k+1}, \ldots, f_j, \ldots, f_{j+k-1}, f_{j+k}] \tag{8}$$

Here, j is the index of the central frame (i.e., the frame intended to be classified). A stacking filter of $k = 15$ is chosen, combining a total of 31 frames.

Finally the stacked TD0 vectors from all eight channels are combined into a single array, which serves as the input for the classifier.

We used a window with a duration of 25 ms and a frame-shift of 5 ms to extract the EMG features. Since five TD features are calculated for each of the EMG channels, the length of the parameter vector assigned to each frame is calculated as

$$M \cdot 5 \cdot (2k+1), \tag{9}$$

which results in 1240 features for a width of the stacking filter of $k = 15$ and $M = 8$ channels.

To reduce the dimension of the parameter vector, we apply linear discriminant analysis (LDA) [44], as in [25,43]. To select the optimum dimension, we analyzed the effect of the number of features on the frame-based phone classification accuracy. Figure 4 shows the average validation accuracy per number of LDA features for the first session of each speaker. Based on this graph, we chose to use 21 LDA features because the average accuracy reaches a plateau at that value. Choosing a higher number of features would result in a more complex model and a longer training time. The classifier used to search for the optimal LDA value was a neural network with a batch size of 128 and 20 epochs.

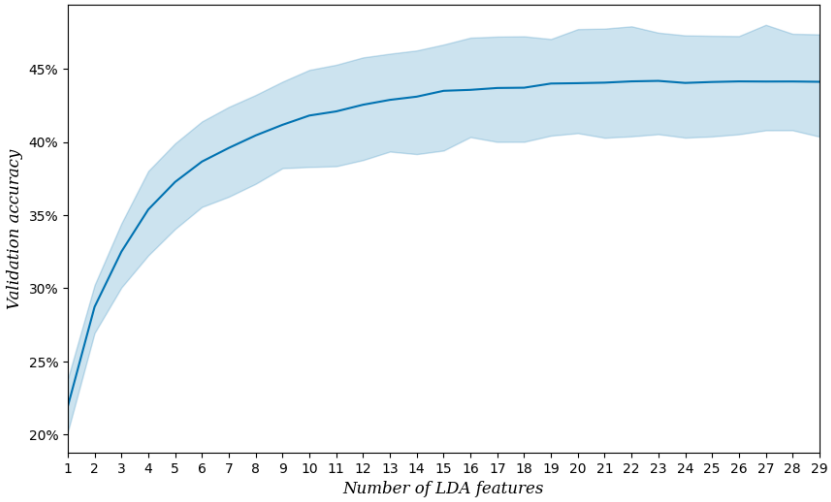

Figure 4. Validation accuracy per number of Linear Discriminant Analysis (LDA) features averaged over Session 101 of all speakers. Classification method: neural network with batch size of 128 and 20 epochs. The solid line represents the average accuracy, and the area above and below the line shows the standard deviation range.

2.4. Experiments

This section describes the experimental part of the study, namely, the classifier used and its configuration, the manner in which we considered speaker and session dependency for the experiments, and how we applied cross-validation.

2.4.1. Classification Method

The classifier used for the experimental part of this study is a feed-forward neural network with one hidden layer using a batch size of 256 and 100 training epochs. We chose these parameters based on a hyper-parameter search by tracking the validation accuracy during 250 training epochs for three batch sizes: 64, 128, and 256. We repeated this for Session 101 of all six speakers and averaged the results (see Figure 5). We chose 100 training epochs because at that point the performance reaches a plateau, and we chose the largest batch size because there is no difference between the three batch sizes, and a larger batch size means lower training time. The network has an input dimension equal to the number of features (21 nodes) and a dense layer with twice as many nodes as features (42 nodes in total) using a rectified linear units activation function [45]. The output layer has the same number of nodes as the number of classes (30, which includes 29 phones and a silence) and uses a *softmax* activation function [46]. Furthermore, a categorical cross-entropy loss function and the *Adam* optimizer [47] with a learning rate of 0.001 are applied.

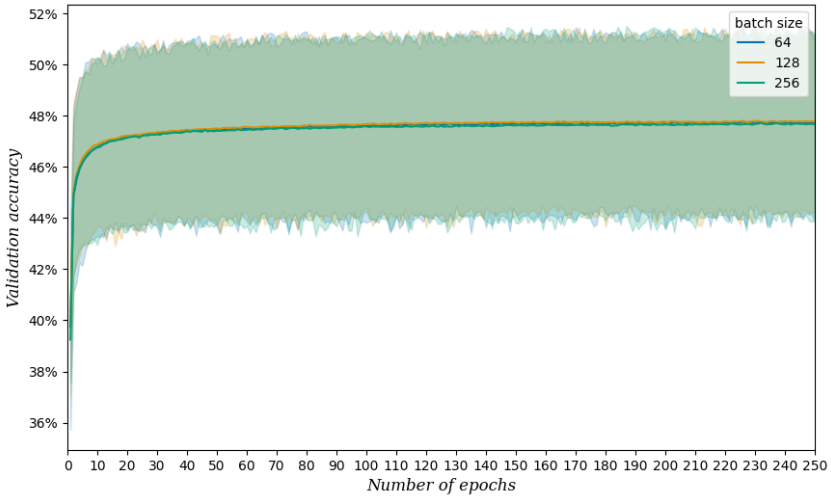

Figure 5. Validation accuracy per number of epochs and three batch sizes averaged over Session 101 of all speakers. The solid line represents the average accuracy, and the area above and below the line shows the standard deviation range.

2.4.2. Speaker and Session Dependency

Our study involves three separate rounds of experiments, each varying in terms of speaker and session dependencies. The first round of experiments was both speaker-dependent and also session-dependent, which means that the training and testing data were taken from the same session. Additionally, we performed a second round of experiments in which the data were speaker-dependent but session-independent. This means that the training data came from a different session or different sessions than the testing data, but that all sessions were recorded by the same speaker. This method allows for the evaluation of the effect of increasing the amount of data from the same speaker on the performance of the model as well as the impact of inter-session variability on the accuracy. In the third round, we used speaker-independent data by training the model using data from multiple sessions of one speaker and testing it using data from another speaker. The testing session contains a session-specific corpus that was not included in the sessions used to train the model, making the experiment both speaker-independent and session-independent. The goal is to assess the potential to create a model that can be applied to new speakers without the need for adaptation by training it only on data from the actual database.

2.4.3. Cross-Validation

We used five-fold cross-validation to obtain the validation accuracy. This means that five different classifiers are trained, each time leaving out a different fold that functions

as the validation set. The obtained results were then averaged. The testing accuracy was obtained after a new classifier was trained using all the training data and then tested on the unseen testing set.

3. Results

In this section, we show the results of the experiments, first from those in the session-dependent mode and then from the ones we performed in the session-independent mode, which are both speaker-dependent. Lastly, we also show the results from the speaker-independent experiments.

3.1. Speaker-Dependent, Session-Dependent Classification

Table 3 shows the results for the session-dependent experiments, for which the model was trained and tested with data from the same speaker and the same session. Some speakers show higher classification accuracy (Speakers 001 and 005) than other speakers. Speaker 006 has the worst results, in particular for Session 102. After reviewing the data from sessions with relatively lower results, we realized that some channels presented recording issues, probably due to the detachment of the electrodes. Specifically, in Sessions 003-102, 004-102, 005-101, and 006-102, we observed problems in the recordings, with some ill-defined signals. Surprisingly, Table 3 does not show this problem for Speaker 004, but as we will see next, it does affect the following speaker-independent experiments.

Table 3. Speaker-dependent, session-dependent classification results.

Speaker	Session	Validation Accuracy	Testing Accuracy
001	101	50.48 ± 1.01	46.42
	102	49.12 ± 0.86	47.15
	103	45.80 ± 0.66	45.53
	104	50.41 ± 1.05	50.54
002	101	43.71 ± 0.48	42.61
	102	42.80 ± 0.96	42.52
	103	38.76 ± 1.35	38.05
	104	39.39 ± 0.77	39.64
003	101	46.73 ± 1.12	45.27
	102	42.41 ± 1.07	39.45
004	101	43.22 ± 1.50	38.44
	102	41.29 ± 1.37	39.62
005	101	43.61 ± 1.56	41.19
	102	51.45 ± 0.54	50.40
006	101	35.92 ± 1.17	35.27
	102	28.39 ± 1.31	24.72
Average		43.34 ± 5.80	41.68 ± 6.14

3.2. Speaker-Dependent, Session-Independent Classification

To evaluate session-independent classification, we first used the models from the previous section (session-dependent experiments) and tested them using data from another session. We then trained new models using a variable number of sessions from the same speaker to analyze the impact of additional data on the classification accuracy. The results (see Table 4) show that the testing accuracy decreases in a session-independent scenario. This decrease in testing accuracy is not the same for every speaker. However, when additional sessions are included in the set of training data, the testing accuracy increases. Nevertheless, it is always lower than the testing accuracy obtained with session-dependent classification.

On the other hand, contrary to the testing accuracy, the validation accuracy decreases as more sessions are added. This is an indication of less over-fitting, as it shows better generalization capability.

The effect of some lower-quality signals in a few sessions (mentioned in Section 3.1) is challenging to assess in these experiments because both training and testing sessions include some of these defective signals. For instance, both experiments for Speaker 004 show very low results because Session 004-102 is used either for training or testing.

Table 4. Speaker-dependent, session-independent classification results.

Speaker	Training Session(s)	Testing Session	Validation Accuracy	Testing Accuracy
001	102	101	49.12 ± 0.86	23.40
	102,103		45.08 ± 0.89	27.89
	102,103,104		42.50 ± 0.74	30.41
	101	102	50.48 ± 1.01	19.57
	101,103		46.85 ± 1.07	22.11
	101,103,104		43.81 ± 0.70	24.54
	101	103	50.48 ± 1.01	14.19
	101,102		48.16 ± 1.14	18.09
	101,102,104		45.00 ± 0.34	18.25
	101	104	50.48 ± 1.01	15.86
	101,102		48.16 ± 1.14	22.38
	101,102,103		44.49 ± 0.37	24.93
002	102	101	42.80 ± 0.96	10.00
	102,103		39.69 ± 0.53	18.32
	102,103,104		37.90 ± 0.61	21.93
	101	102	43.71 ± 0.48	20.90
	101,103		41.23 ± 1.09	23.81
	101,103,104		37.79 ± 0.80	24.19
	101	103	43.71 ± 0.48	17.79
	101,102		42.46 ± 1.02	18.03
	101,102,104		39.63 ± 0.53	16.73
	101	104	43.71 ± 0.48	19.01
	101,102		42.46 ± 1.02	20.84
	101,102,103		39.42 ± 0.51	22.92
003	102	101	42.41 ± 1.07	20.66
	101	102	46.73 ± 1.12	15.05
004	102	101	41.29 ± 1.37	10.95
	101	102	43.22 ± 1.50	8.63
005	102	101	51.45 ± 0.54	11.83
	101	102	43.61 ± 1.56	23.61
006	102	101	28.39 ± 1.31	16.02
	101	102	35.92 ± 1.17	8.30

3.3. Speaker-Independent, Session-Independent Classification

To evaluate session-independent classification, we employed the models trained with three sessions from the session-independent experiments and tested them with the remaining session from each of the other speakers. The results, presented in Table 5, indicate that classification accuracy varies greatly despite all models being trained on similar amounts of data.

As explained above (Section 3.1), Session 004-102 contains ill-defined signals, probably due to detaching of the electrodes. This explains the bad results when this session is used to test any model. The same can be said for Sessions 003-102, 005-101, and 006-102. A comparison of speaker-independent experiments to speaker-dependent, session-independent experiments reveals a substantial decrease in the accuracy compared to the former.

Table 5. Speaker-independent, session-independent classification results.

Training Speaker	Training Sessions	Testing Session	Testing Speaker	Testing Accuracy
001	102,103,104	101	002	19.47
			003	14.47
			004	12.08
			005	9.33
			006	8.41
	101,103,104	102	002	18.28
			003	15.10
			004	6.91
			005	19.90
			006	8.51
	101,102,104	103	002	8.36
	101,102,103	104	002	10.47
002	102,103,104	101	001	14.07
			003	15.71
			004	14.78
			005	8.11
			006	10.53
	101,103,104	102	001	15.95
			003	18.09
			004	10.26
			005	16.90
			006	7.21
	101,102,104	103	001	16.79
	101,102,103	104	001	20.43

4. Discussion and Conclusions

This paper presents the results of phone classification experiments conducted on the new ReSSInt-EMG database. Compared to previous work [34], we revisited the linear discriminant analysis (LDA) reduction procedure, which resulted in changing the number of LDA features from 28 to 21. The change in the number of features used to train the model helped to reduce the training time and the complexity of the model, but the obtained accuracy remained similar. Furthermore, we included new sessions from the speakers that were already part of the database and recordings from two new speakers and extended the experiments with different modalities regarding speaker and session dependency. To accommodate the increased complexity of these experiments, we also used a neural network as a classification method instead of a bagging classifier.

The session-dependent classification results show varying outcomes not only across speakers but also across multiple sessions from the same speaker. Furthermore, the session-independent results indicate a substantial decrease in testing accuracy when the model is applied to data from sessions not included in the training phase, with the magnitude of this effect differing between the two speakers.

The decrease in testing accuracy observed when training with data from a session different from the one used to test the model is likely due to inter-session variability, which can be attributed to several factors. First, despite the use of a 3D mask, variations in electrode placement can occur between sessions. Second, the physical or mental state of the speaker may lead to slight differences in articulation between sessions, as each is recorded on a different day. For instance, a person may articulate differently when feeling exhausted, resulting in less articulation effort. Third, environmental conditions such as temperature and humidity can affect the speaker's state and the contact between the electrodes and the skin. High temperatures may cause increased sweating and decreased motivation. These

factors can impact the recorded EMG signals, resulting in each session being recorded under unique circumstances. Consequently, a model that can identify patterns in the EMG signals of one session may struggle to recognize those same patterns in signals from a different session.

Interestingly, when additional session data are added to the training data, testing accuracy increases. Given a corresponding decrease in validation accuracy, we believe that the improvement is due to enhanced diversity and representation of the data, allowing the model to better generalize beyond the training data. These results suggest that developing an EMG-based SSI with sufficient performance for real-world applications requires a large and diverse database. While using a larger set of training data may potentially slow down the experiments and require additional resources, we firmly believe that it is crucial to leverage as much training data as possible, provided that sufficient processing capabilities are available and the addition of new data leads to improved model performance. Our rationale stems from the fact that an SSI system suitable for real-world applications requires extensive preparation to handle unseen data.

The speaker-independent classification results demonstrate a substantial decrease in model accuracy when trained with data from other speakers, even when the amount of training data is comparable to the speaker-dependent, session-independent models. This suggests that the differences between speakers' data are substantial, making it challenging for the model to generalize to a different speaker. These differences can be attributed to various factors, such as differences in speakers' physiognomy, articulation manner, or speaking pace. These findings suggest that using an SSI trained on a different speaker presents extra difficulty. Further experiments are needed to investigate whether training the model with a more extensive database from a single speaker or with data from multiple speakers can enhance speaker-independent performance.

It is important to note that during four sessions, we observed a deviation in the signals of one channel, which cast uncertainty on its quality. The classification accuracy of these sessions is indeed lower compared to the other sessions by the same speaker. The most likely cause for these signal deviations is the detachment of electrodes in this channel during recording or the use of defective electrodes or cables. Acquiring EMG data is a sensitive technique and can result in variations in EMG signals depending on the speaker and recording conditions.

Considering all of our findings, we plan to record more data from fewer speakers for future studies to address the issue of inter-session variability. We believe that this strategy will allow us to collect a more diverse range of data and enhance the performance of the EMG-based SSI. Furthermore, we intend to undertake more complex tasks, such as direct speech generation from EMG signals, to achieve our ultimate goal of developing an EMG-based SSI for Spanish-speaking laryngectomees.

Author Contributions: Conceptualization, E.N. and I.H.; methodology, I.S., E.d.B., I.H. and E.N.; software, I.S., E.d.B. and X.d.Z.; validation, I.S. and E.d.B.; formal analysis, I.S. and E.d.B.; investigation, I.S. and E.d.B.; resources, I.H. and E.N.; data curation, E.d.B., I.S. and X.d.Z.; writing—original draft preparation, I.S. and E.d.B.; writing—review and editing, I.S., E.d.B., I.H., E.N. and X.d.Z.; visualization, I.S.; supervision, I.H. and E.N.; project administration, I.H. and E.N.; funding acquisition, I.H. and E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agencia Estatal de Investigación grant number ref.PID2019-108040RB-C21/AEI/10.13039/501100011033.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee CEISH of the UPV/EHU (project code M10_2021_269, act 142/2021, approved on the 23rd of September 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The ReSSInt-EMG database will be made publicly available once collection of the recordings is finished.

Acknowledgments: We would like to thank all the participants for helping us build the database, knowing that without them, this study would not have been possible.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EMG	Electromyography
LDA	Linear Discriminant Analysis
SSI	Silent Speech Interface
TD	Time-Domain
TTS	Text-to-Speech
VCV	Vowel–Consonant–Vowel

References

1. Hernaez, I.; Gonzalez Lopez, J.A.; Navas, E.; Pérez Córdoba, J.L.; Saratxaga, I.; Olivares, G.; Sanchez de la Fuente, J.; Galdón, A.; Garcia, V.; Castillo, J.d.; et al. ReSSInt project: Voice restoration using Silent Speech Interfaces. In Proceedings of the IberSPEECH 2022, ISCA, Granada, Spain, 14–16 November 2022 ; pp. 226–230. [CrossRef]
2. Tang, C.G.; Sinclair, C.F. Voice Restoration after Total Laryngectomy. *Otolaryngol. Clin. N. Am.* **2015**, *48*, 687–702. [CrossRef] [PubMed]
3. Zieliński, K.; Rączaszek-Leonardi, J. A Complex Human-Machine Coordination Problem: Essential Constraints on Interaction Control in Bionic Communication Systems. In Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–8. [CrossRef]
4. Wand, M.; Janke, M.; Schultz, T. The EMG-UKA corpus for electromyographic speech processing. In Proceedings of the Interspeech 2014, Singapore, 14–18 September 2014; pp. 1593–1597. [CrossRef]
5. Gaddy, D.; Klein, D. Digital voicing of silent speech. *arXiv* **2020**, arXiv:2010.02960.
6. Diener, L.; Roustay Vishkasougheh, M.; Schultz, T. CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020.
7. Freitas, J.; Teixeira, A.; Dias, J. Multimodal corpora for silent speech interaction. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 4507–4511.
8. Safie, S.I.; Yusof, M.I.; Rahim, R.; Taib, A. EMG database for silent speech Ruqyah recitation. In Proceedings of the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), Kuala Lumpur, Malaysia, 4–8 December 2016; pp. 712–715. [CrossRef]
9. Lopez-Larraz, E.; Mozos, O.M.; Antelis, J.M.; Minguez, J. Syllable-based speech recognition using EMG. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 4699–4702. [CrossRef]
10. Ma, S.; Jin, D.; Zhang, M.; Zhang, B.; Wang, Y.; Li, G.; Yang, M. Silent Speech Recognition Based on Surface Electromyography. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 4497–4501. [CrossRef]
11. Lee, K.S. EMG-Based Speech Recognition Using Hidden Markov Models with Global Control Variables. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 930–940. [CrossRef] [PubMed]
12. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]
13. Gonzalez-Lopez, J.A.; Gomez-Alanis, A.; Martin Donas, J.M.; Perez-Cordoba, J.L.; Gomez, A.M. Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access* **2020**, *8*, 177995–178021. [CrossRef]
14. Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453. [CrossRef]
15. Gonzalez, J.A.; Cheah, L.A.; Gomez, A.M.; Green, P.D.; Gilbert, J.M.; Ell, S.R.; Moore, R.K.; Holdsworth, E. Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2362–2374. [CrossRef]
16. Anumanchipalli, G.K.; Chartier, J.; Chang, E.F. Speech Synthesis from Neural Decoding of Spoken Sentences. *Nature* **2019**, *568*, 493–498. [CrossRef] [PubMed]
17. Toth, A.R.; Wand, M.; Schultz, T. Synthesizing speech from electromyography using voice transformation techniques. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.

18. Janke, M.; Wand, M.; Nakamura, K.; Schultz, T. Further investigations on EMG-to-speech conversion. In Proceedings of the ICASSP, Kyoto, Japan, 25–30 March 2012; pp. 365–368. [CrossRef]
19. Li, H.; Lin, H.; Wang, Y.; Wang, H.; Zhang, M.; Gao, H.; Ai, Q.; Luo, Z.; Li, G. Sequence-to-Sequence Voice Reconstruction for Silent Speech in a Tonal Language. *Brain Sci.* **2022**, *12*, 818. [CrossRef] [PubMed]
20. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2386–2398. [CrossRef] [PubMed]
21. Wu, J.; Zhao, T.; Zhang, Y.; Xie, L.; Yan, Y.; Yin, E. Parallel-inception CNN approach for facial sEMG based silent speech recognition. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Online, 1–5 November 2021; pp. 554–557.
22. Gaddy, D. Voicing Silent Speech. Ph.D. Thesis, University of California, Berkeley, CA, USA, 2022.
23. De Luca, C.J. *Surface Electromyography: Detection and Recording*; Technical Report; DelSys Incorporated: Natick, MA, USA , 2002.
24. Zhou, Q.; Jiang, N.; Hudgins, B. Improved phoneme-based myoelectric speech recognition. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 2016–2023. [CrossRef] [PubMed]
25. Wand, M.; Schultz, T. Analysis of phone confusion in EMG-based speech recognition. In Proceedings of the ICASSP, Prague, Czech Republic, 22–27 May 2011; pp. 757–760.
26. Wand, M.; Schultz, T. Session-independent EMG-based Speech Recognition. In Proceedings of the Biosignals, Rome, Italy, 26–29 January 2011; pp. 295–300.
27. Wand, M.; Schmidhuber, J. Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 3032–3036.
28. Diener, L.; Amiriparian, S.; Botelho, C.; Scheck, K.; Küster, D.; Trancoso, I.; Schuller, B.W.; Schultz, T. Towards Silent Paralinguistics: Deriving Speaking Mode and Speaker ID from Electromyographic Signals. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3117–3121. [CrossRef]
29. Khan, M.U.; Choudry, Z.A.; Aziz, S.; Naqvi, S.Z.H.; Aymin, A.; Imtiaz, M.A. Biometric Authentication based on EMG Signals of Speech. In Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey, 12–13 June 2020; pp. 1–5. [CrossRef]
30. Zahner, M.; Janke, M.; Wand, M.; Schultz, T. Conversion from facial myoelectric signals to speech: A unit selection approach. In Proceedings of the Interspeech 2014, Singapore, 14–18 September 2014; pp. 1184–1188.
31. Diener, L.; Janke, M.; Schultz, T. Direct conversion from facial myoelectric signals to speech using Deep Neural Networks. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–7. [CrossRef]
32. Janke, M.; Diener, L. EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2375–2385. [CrossRef]
33. Salomons, I.; del Blanco, E.; Navas, E.; Hernáez, I. Accepted for publication—Spanish Phone Confusion Analysis for EMG-Based Silent Speech Interfaces. In Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH), Dublin, Ireland, 20–24 August 2023.
34. Del Blanco, E.; Salomons, I.; Navas, E.; Hernáez, I. Phone classification using electromyographic signals. In Proceedings of IberSPEECH 2022, ISCA, Granada, Spain, 14–16 November 2022; pp. 31–35. [CrossRef]
35. Chan, A.D.C.; Englehart, K.; Hudgins, B.; Lovely, D.F. Myo-Electric Signals to Augment Speech Recognition. *Med. Biol. Eng. Comput.* **2001**, *39*, 500–504. [CrossRef] [PubMed]
36. Maier-Hein, L.; Metze, F.; Schultz, T.; Waibel, A. Session independent non-audible speech recognition using surface electromyography. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Cancun, Mexico, 27 November–1 December 2005; pp. 331–336. [CrossRef]
37. Schultz, T.; Wand, M. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Commun.* **2010**, *52*, 341–353. [CrossRef]
38. Jou, S.C.; Schultz, T.; Walliczek, M.; Kraft, F.; Waibel, A. Towards continuous speech recognition using surface electromyography. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
39. Zhu, M.; Zhang, H.; Wang, X.; Wang, X.; Yang, Z.; Wang, C.; Samuel, O.W.; Chen, S.; Li, G. Towards Optimizing Electrode Configurations for Silent Speech Recognition Based on High-Density Surface Electromyography. *J. Neural Eng.* **2021**, *18*, 016005. [CrossRef] [PubMed]
40. Aubanel, V.; Lecumberri, M.L.G.; Cooke, M. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *Int. J. Audiol.* **2014**, *53*, 633–638. [CrossRef] [PubMed]
41. Sainz, I.; Erro, D.; Navas, E.; Hernáez, I.; Sanchez, J.; Saratxaga, I.; Odriozola, I. Versatile Speech Databases for High Quality Synthesis for Basque. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012.
42. McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; Volume 2017, pp. 498–502.

43. Wand, M. *Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling*; KIT Scientific Publishing: Karlsruhe, Germany, 2015. [CrossRef]
44. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
45. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
46. Bridle, J. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In Proceedings of the 2nd International Conference on Neural Information Processing System, Denver, CO, USA, 27–30 November 1989; Volume 2.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.