Detecting Deception from Gaze and Speech Using a Multimodal Attention LSTM-Based Framework

Ascensión Gallardo-Antolín [1,*] and Juan M. Montero [2]

1   Department of Signal Theory and Communications, Universidad Carlos III de Madrid,
     Avda. de la Universidad, 30, Leganés, 28911 Madrid, Spain
2   Speech Technology Group, ETSIT, Universidad Politécnica de Madrid, Avda. de la Complutense, 30,
     28040 Madrid, Spain; juanmanuel.montero@upm.es
*   Correspondence: gallardo@ing.uc3m.es

Abstract: The automatic detection of deceptive behaviors has recently attracted the attention of the research community due to the variety of areas where it can play a crucial role, such as security or criminology. This work is focused on the development of an automatic deception detection system based on gaze and speech features. The first contribution of our research on this topic is the use of attention Long Short-Term Memory (LSTM) networks for single-modal systems with frame-level features as input. In the second contribution, we propose a multimodal system that combines the gaze and speech modalities into the LSTM architecture using two different combination strategies: Late Fusion and Attention-Pooling Fusion. The proposed models are evaluated over the Bag-of-Lies dataset, a multimodal database recorded in real conditions. On the one hand, results show that attentional LSTM networks are able to adequately model the gaze and speech feature sequences, outperforming a reference Support Vector Machine (SVM)-based system with compact features. On the other hand, both combination strategies produce better results than the single-modal systems and the multimodal reference system, suggesting that gaze and speech modalities carry complementary information for the task of deception detection that can be effectively exploited by using LSTMs.

Keywords: deception detection; multimodal; gaze; speech; LSTM; attention; fusion

1. Introduction

Deception is a kind of human behavior that can be defined as the intentional attempt to produce in the receiver a false belief. Nowadays, many applications where the automatic detection of deception plays a crucial role have arisen, mainly in areas related to security, criminology, ref. [1] social media [2] or computer-mediated communications, such as chats or social networks [3]. For this reason, this topic has recently attracted the attention of researchers, although it is still not studied enough.

Automatic Deception Detection (ADD) systems commonly rely on verbal and/or nonverbal cues [4]. Verbal-based ADD systems use features related to the linguistic content of the communication, typically derived from text analysis. In contrast, nonverbal-based approaches utilize paralinguistic information contained in different modalities, such as gait, body gestures, facial expressions, gaze or speech. In this work, we focus on this latter type of ADD systems, in particular, on those based on gaze or speech features and their combination.

Eyes are one of the most expressive elements of the human face, and for that it is argued that their attributes and movements (such as blinks, visual and non-visual saccades and gaze direction), and eye interactions can provide useful cues for deception detection.

With respect to eye attributes, several research works have shown that the pupil dilation is a reliable indicator of lying [5,6]. Regarding eye dynamics, the analysis of eye micro-movements can be used for the discrimination of deceptive behaviours. In this sense, an increase of the number of eye blinks performed over a certain period seems

to be correlated to deception [7], although other works have pointed out to inconsistent findings about the relationship between blink rate and lie-telling [6]. In addition, it has been observed that the rate of non-visual saccades (spontaneous eye movements not related to a specific visual stimulus) increases when the subject is lying as this kind of movements are related to long-term memory search and this process is more intensive when telling lies [8]. Visual saccades (eyes movements induced by the human visual attention mechanism in presence of a particular visual stimuli) and gaze direction are also related to deception [5,9]. Finally, eye gazing behaviours can reflect the cognitive load and/or emotional status of a person and the nature of his/her interactions to other individuals, and therefore, they can also contain useful information about deception. In fact, individuals tend to avert the gaze when lying whereas they are prone to prolong the time they look at their interlocutors during truth-telling [10].

Most of the eye-based deception detection systems use features based on these findings. For example, the aspect eye ratio and an estimation of the gaze direction have been proposed as features for the detection of deceitful answers [9,11]. Gupta et al. [12] extracts a set of features (fixations, number of eye blinks and mean and standard deviation of the pupil size) from the raw gaze data recorded with an eye-tracker. Other research works propose a set of features consisting in measures of facial and eye movements (blinks, shifts, etc.) obtained by using computer vision techniques over videos [13]. For performing the classification between truths and lies, these systems usually consider traditional machine-learning algorithms, such as Support Vector Machines (SVM) [13], Random Forest (RF) [11–13] or simple Multilayer Perceptron [12,13] with the aforementioned hand-crafted features as input. Note that these techniques are not able to deal with variable length sequences, and therefore, they work on compact representations of the frame-level gaze features, typically statistical functionals (average, extreme values, etc.), that are fixed-length vectors.

In nonverbal communications, voice acoustics also conveys useful information about the speaker's truthful or deceptive behavior. Several studies have shown that certain prosodic characteristics change when lying. In fact, in deceptive speech the speaking rate and fundamental frequency usually increase in comparison to truth utterances [14], and the energy suffers variations [15]. Besides, it seems that liars and honest speakers make a different use of silent and filled pauses [15]. Finally, other subtle changes affect the sound pressure and vocal organs [16] modifying the segmental characteristics of speech.

From these observations, existing speech-based ADD systems are focused on different types of frame-level acoustic-prosodic features, such as pitch, energy, phoneme duration [17], zero crossing rate, spectral centroid, spectral bandwidth, spectral rolloff, chroma frequencies, Mel frequency cepstrum coefficients [18] and its derivatives [12,19,20], glottal flow, percentage of voiced/unvoiced frames and harmonic model and phase distortion [21], among others. For the classification module, most of these works adopt traditional machine-learning methods that are fed with compact representations (usually, statistical functionals) of these hand-crafted acoustic characteristics. Among these methods, it is worth mentioning the Ripper rule-induction [17], K-Nearest Neighbour (KNN) [12], RF [12,19] and SVM classifiers [19]. As in the case of the gaze modality, these classification algorithms are not able to exploit the dynamic information contained in the sequences of frame-level features. For overcoming this issue, recently, it has been proposed a speech-based system based on Long Short-Term Memory (LSTM) networks and frame-level features that outperforms a SVM-based system with statistical functionals, showing the importance of the speech dynamics for this task [20].

Due to the diversity of cues for deception detection, recent works have proposed multimodal ADD systems where two or more modalities of different nature are combined. For example, Abouelenien et al. reports a system based on decision tree classifiers and the fusion of linguistic, thermal and physiological characteristics (heart and respiration rate, blood volume pulse and skin conductance) that outperforms the corresponding single-modal systems on an opinion task [22]. Wu et al. develops a multimodal approach utilizing motion, audio and textual channels with a Gaussian Mixture Model classifier for real-life

trial videos [19]. Rill-García et al. combines visual, audio and textual modalities in a SVM-based system that is evaluated over two different problems: an opinion task and a real court trial dataset [21]. Finally, Gupta et al. fuses different modalities (electroencephalogram (EEG), gaze, video and audio) and classifiers, achieving better results than the individual channels over the Bag-of-Lies dataset, which has been specially collected for the development of multimodal deception detection systems [12].

The aim of this work is to deepen in several aspects of ADD that have not been sufficiently studied until now. In particular, we focus on the effective use of two non-intrusive modalities, speech and gaze data, by exploiting the information contained in their temporal variations and exploring different strategies for their fusion. This type of combination has not been much explored due to the lack of appropriate datasets. A special emphasis is placed on dealing with recordings collected in real scenarios in order to gain insight in the applicability of the developed systems in practical situations. In this paper, we use the aforementioned Bag-of-Lies database [12] that meets the required characteristics.

As earlier mentioned, one of the weakness of most of the previous research works is the use of traditional machine learning methods that are not able to properly model the dynamics of sequences (as is the case of eye-tracker and speech signals) and require compact feature representations (typically, statistics) as input, which could lead to a loss of relevant information contained in the temporal evolution of the frame-level features.

For overcoming this issue, we propose the use of Attention Long Short-Term Memory networks [23,24] that have the ability of modeling the long-term dynamics of raw gaze and speech feature sequences. The main reason for this choice is the dramatic improvements achieved by LSTMs in other tasks involving the processing of temporal sequences, such as audio and speech-related applications. Moreover, the incorporation of an attention mechanism in the LSTM framework generally improves the performance of these techniques, as it tries to learn the structure of the temporal sequences by modeling the particular relevance of each frame to the task under consideration [25]. Recently, these models have been successfully utilized for, among others, acoustic event detection [26], acoustic scene classification [27], automatic speech recognition [28], speech emotion recognition [29,30], cognitive load classification from speech [31,32] or speech intelligibility level classification [33,34]. To the best of our knowledge, the use of attentional LSTMs has not been previously explored in the literature neither for ADD systems based on eye-tracker signals nor based on speech signals. In addition, we propose two different alternatives for the fusion of both kind of modalities inside the LSTM architecture, namely Late Fusion and Attention-Pooling Fusion. Finally, we carry out the comparison between the performance of the single-modal and multimodal LSTM-based ADD systems running on frame-level features and the corresponding reference systems based on SVM and compact features as input.

This research aims to develop an ADD system working with speech and gaze data collected in real conditions, that is supported on two main hypotheses. The first one states that the dynamics of gaze and speech feature sequences convey important information for deception detection that can be exploited by using attention LSTM networks. The second one refers to the complementarity of the gaze and speech modalities, that can be effectively combined in the LSTM framework.

2. Materials and Methods

2.1. Dataset

In this work, we used and adhered to the terms of this license the Bag-of-Lies dataset [12], which is a multimodal deception database composed of video, audio, gaze and EEG recordings from 35 subjects (25 male and 10 female), although EEG signals are only available for 22 users. As requested by the authors of the database, data from one of the female participants (user 12) were not used. So, in total there are 315 recordings, of which 157 corresponded to lies and 158 to truths.

The database was collected in realistic conditions. Participants were shown 6–10 images and requested to describe them in such a way that they were free to lie or tell the truth. Video and audio were recorded with a conventional smartphone, and therefore, the audio recordings present a significant amount of noise. Gaze data was captured by using a Gazepoint GP3 Eye Tracker (Gazepoint, Vancouver, BC, Canada), whereas a 14-Channel Emotiv EPOC+EEG headset (Emotiv, San Francisco, CA, USA) was used for the recording of EEG signals.

In our experiments, we have only used the gaze and audio (speech) modalities. Gaze data is sampled at 60 Hz and follows the GazePoint Open Gaze API specification [35], providing 26 channels that encode different information related to eye tracking. Audio was extracted from the video recordings by using the FFmpeg software [36] and downsampled to 16 KHz. As some of these audio files contain speech from both, the person in charge of the recordings and the participant, the first voice was eliminated by cropping the initial seconds of the audio recordings.

2.2. Feature Extraction

2.2.1. Gaze Features

From the 26 channels provided by the Gazepoint GP3 Eye Tracker, we selected four of them, Horizontal Fixation Point-Of-Gaze (FPOGX), Vertical Fixation Point-Of-Gaze (FPOGY), Left Pupil Diameter (LPD) and Right Pupil Diameter (RPD), that corresponded to the horizontal and vertical coordinates of the user's fixation point-of-gaze and the diameter of the left and right eye pupil in pixels. The choice of these channels was due to the fact that they contain information about fixations, saccades and pupil size, that are indicators of deception behaviors [5,6,8,9]. Due to the high correlation between the left and right eye diameters, instead of using LPD and RPD channels independently, we computed their average, that is called Mean Pupil Diameter (MPD). Therefore, gaze features were derived from the following three signals: FPOGX, FPOGY and MPD.

The gaze feature extraction process is depicted in Figure 1a. The first step consisted of dividing the three gaze signals into segments of 3 s length with an overlap of 2.5 s seconds (it meant a shift of 0.5 s between consecutive windows). Incomplete sequences (i.e., segments shorter than 3 s) were padded with masked values in order to not to be used in further computations. This step was required due to the small number of recordings, as this data augmentation process allowed the obtaining of more reliable and representative models. A similar technique has been used for depression detection from speech [37] or Parkinson's Disease detection from drawing movements [38]. After this process, the total number of segments was 5258, distributed in 2736 lies and 2522 truths. Note that from now on, we differentiated two levels: segment level that referred to each of the segments of 3 s length and turn level that corresponded to the full recordings.

For the LSTM-based systems, the gaze features X_G consisted of a temporal sequence with dimension $N_G \times T_G$, where N_G is the number of the eye tracking signals considered and T_G is the number of elements composing each gaze sequence. In this case, $N_G = 3$ (FPOGX, FPOGY and MPD) and $T_G = 180$ as it was the length corresponding to a segment of 3 s sampled at 60 Hz.

In the case of the SVM-based systems, as this kind of traditional machine-learning techniques does not allow the adequate modeling of temporal signals, a summarized representation of the frame-level gaze features was required. For doing that, several statistical functionals (in particular, mean, standard deviation, skewness, kurtosis, max and min) were obtained from the raw gaze sequences X_G, yielding a compact feature vector X_G^C whose dimension is $N_G \times N_F$, where N_F is the number of statistics computed (in this case, $N_F = 6$).

2.2.2. Speech Features

Figure 1b shows the feature extraction process for speech signals. As in the case of the gaze modality, firstly speech signals were divided into segments of 3 s length with an

overlap of 2.5 s, obtaining the same number of chunks. Again, masked values were used for padding shorter segments than 3 s.

For the LSTM-based systems, the speech features consisted of log-mel spectrograms that were extensively used in a variety of speech and audio-based applications, such as depression detection [37], speech intelligibility level classification [33,34] or environmental sound classification [39]. For each audio segment, log-mel spectrograms were computed by using the Python's package LibROSA [40] as follows: first, the speech signal was divided into short analysis Hamming windows (frames) of 20 ms length with an overlap between adjacent windows of 10 ms. Then, the magnitude of the Short-Time Fourier Transform (STFT) of each window was computed and mapped to the mel-frequency spacing [18] by using an auditory filter bank composed of mel-scaled filters, and later converted to a logarithmic scale. The mel scale is a frequency transformation that emulates the non-equal sensitivity of the human hearing at different frequencies. This way, each speech segment was represented by a temporal sequence X_S with dimension $N_S \times T_S$, where N_S is the number of mel filters and T_S is the number of speech frames (note that each acoustic frame corresponds to 10 ms). In this case, $N_S = 32$ and $T_S = 300$ as it is the number of frames contained in a 3 s segment.

In the case of the SVM-based systems, a compact representation of the log-mel spectrograms was used, that consisted of the following six statistics: mean, standard deviation, skewness, kurtosis, max and min. This way, the final speech feature set X_S^C was obtained, whose dimension was $N_S \times N_F$, where N_F is the number of statistics computed (in this case, $N_F = 6$, as in the case of the gaze modality).
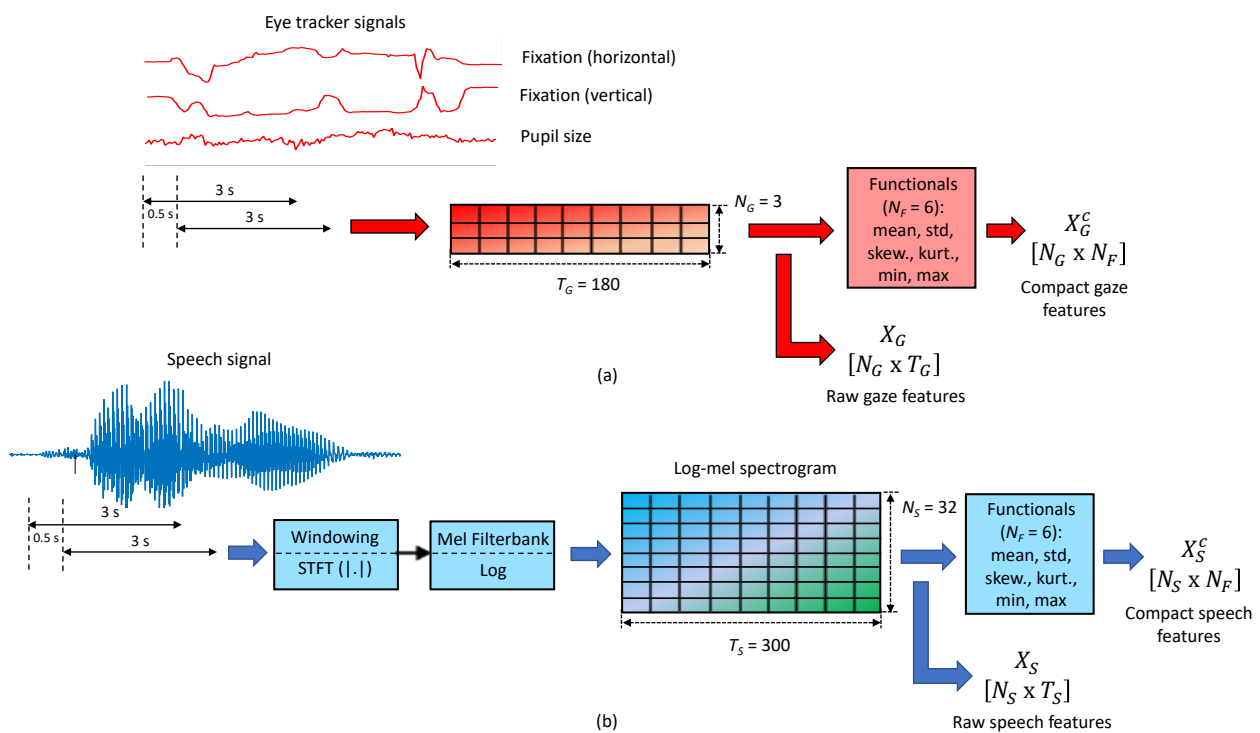

Figure 1. Feature extraction process for (**a**) Eye tracker signals (gaze); (**b**) Audio signals (speech).

2.3. Deception Detection Systems

2.3.1. Single-Modal SVM-Based Add System

The reference system was based on an SVM with Gaussian kernel [41] that was fed with the normalized gaze or speech compact representations described in Sections 2.2.1 and 2.2.2, respectively. This kind of binary classification technique discriminated between the different classes by using a set of hyperplanes that satisfied the maximum separation criterion.

2.3.2. Single-Modal Attention LSTM-Based Add System

As the dynamics of gaze (FPOGX, FPOGY and MPD) and speech features (log-mel spectrograms) contain relevant information about the deception behaviour of a subject, our proposal is to develop a common framework for both modalities based on attention Long Short-Term Memory networks that are known to have the ability to learn long-term dependencies by storing information from the past in their memory blocks [23,24].

LSTMs perform a sequence-to-sequence learning where an input sequence of length T, $X = \{x_1, \ldots, x_T\}$ is transformed to an output sequence $Y = \{y_1, \ldots, y_T\}$ of the same length. ADD is a many-to-one problem in the sense that a single label (truth or lie) must be assigned to the whole input segment. For this reason, usually the LSTM sequence Y is summarized in a single value Z, which can be seen as a segment level representation, prior to be fed to the classifier itself by means of a certain pooling mechanism [29,42]. In this work, we adopted the attention pooling strategy as it has been successfully used in other classification problems involving the modeling of temporal sequences [26–34]. The hypothesis behind this approach was that certain LSTM frames contained more cues about the task under consideration than other ones. Therefore, frames conveying more useful information for the discrimination between truths and lies should be emphasized and their contribution to the Z computation should be higher, whereas non-relevant frames should be diminished or even ignored. This way, the segment level representation Z is computed as the weighted arithmetic mean of the output LSTM frames as follows,

$$Z = \sum_{t=1}^{T} \alpha_t y_t \tag{1}$$

where $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ is the attention weight vector, and the value of each weight is related to the importance of the corresponding frame for deception detection.

The weights were obtained following the technique proposed, among others, by [30], that was adequate for scenarios with scarce training data, as is our case. According to this approach, the un-normalized attention weights, $\bar{\alpha} = \{\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_T\}$, are calculated through the following Equation (2).

$$\bar{\alpha}_t = u^{\text{tr}} y_t, \tag{2}$$

where u is the attention parameter vector and the superscript *tr* denotes a transpose operation. The attention parameters and the LSTM outputs were jointly learnt during the training of the network. Finally, these weights were normalized for guaranteeing that their sum across all the frames of the sequence is equal to one. For doing this, a softmax transformation was applied, yielding a set of normalized weights $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$, as indicated in Equation (3).

$$\alpha_t = \frac{e^{\bar{\alpha}_t}}{\sum_{k=1}^{T} e^{\bar{\alpha}_k}} \tag{3}$$

Figure 2 depicts the block diagram of the attention LSTM system proposed for ADD that used either gaze features or log-mel spectrograms as input. In both cases, mean and standard deviation normalization were applied at segment level. As can be observed, it was composed of two main stages: LSTM and Fully Connected modules.

The LSTM module consisted of a Masking layer that makes sure the dummy values of the padded sequences were not used in further computations (see Sections 2.2.1 and 2.2.2) and an attention LSTM layer with 32 hidden units and a dropout of 0.25 in order to avoid overfitting in the training process. As well, the attention parameter vector u had a dimension of 32. The Fully Connected module was composed of a dense layer of 50 neurons, and final dense layer with 1 node that was activated by a sigmoid function for performing the binary classification.

The only difference between the gaze and speech-based systems was the type of feature vectors used as input and the length of the LSTM input/output sequences T. In the

case of the gaze modality, the input X is the gaze feature vector (FPOGX, FPOGY and MPD) X_G and the LSTM length is $T = T_G = 180$, whereas in the case of speech, the network is fed with the speech features (log-mel spectrograms) $X = X_S$ and the LSTM length is set to $T = T_S = 300$. In the same way, the components of the attention parameter vector u were initialized to $1/T$, i.e., to $1/180$ and $1/300$ for gaze and speech modalities, respectively.

Both, reference and LSTM-based systems work at segment level. In other words, they produce a prediction of deception/honesty for each 3-s segment. As our objective was to classify each turn as truth or lie, it was necessary to obtain the predictions at turn level. For doing that, we followed the method proposed in, among others, ref. [37] that consisted of a majority voting strategy. This way, the final label assigned to each turn was the most frequent prediction value of the segments belonging to this turn.
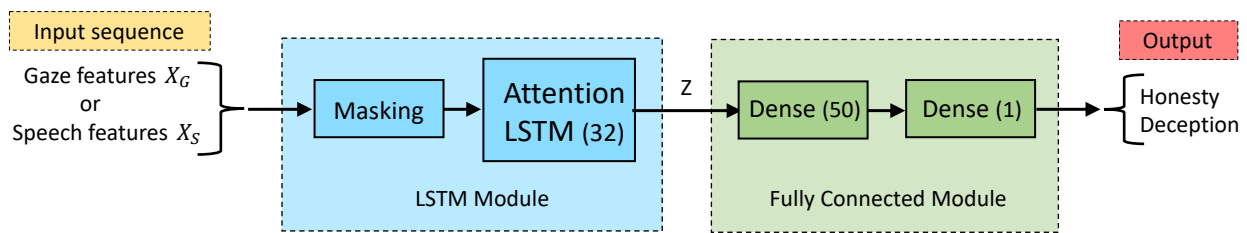

Figure 2. LSTM-based architecture for deception detection when using either gaze or speech features as input sequences.

2.4. Multimodal Systems

For the reference system, two strategies for combining gaze and speech modalities into a SVM framework are developed. In the first case, *Late Fusion*, the combination is done at decision level. More precisely, the scores produced by each of the single-modal SVM systems are transformed to probabilities by means of a softmax operation, and then averaged for producing the final score. In the second case, *Early Fusion*, both modalities are fused at feature level by concatenating the corresponding compact representations into a single feature vector.

For the LSTM-based systems, we also explored two combination techniques for the fusion of gaze and speech modalities, namely, Late Fusion and Attention-Pooling (AP) Fusion. As can be observed in Figure 3a, the late fusion strategy consisted of the combination of the two single systems at decision level by averaging their outputs. In the AP fusion, as shown in Figure 3b, the outputs of the Attention LSTM layers of the individual systems were combined by using a dense layer of 100 nodes followed by a final fully-connected layer of 1 neuron and sigmoid activation.
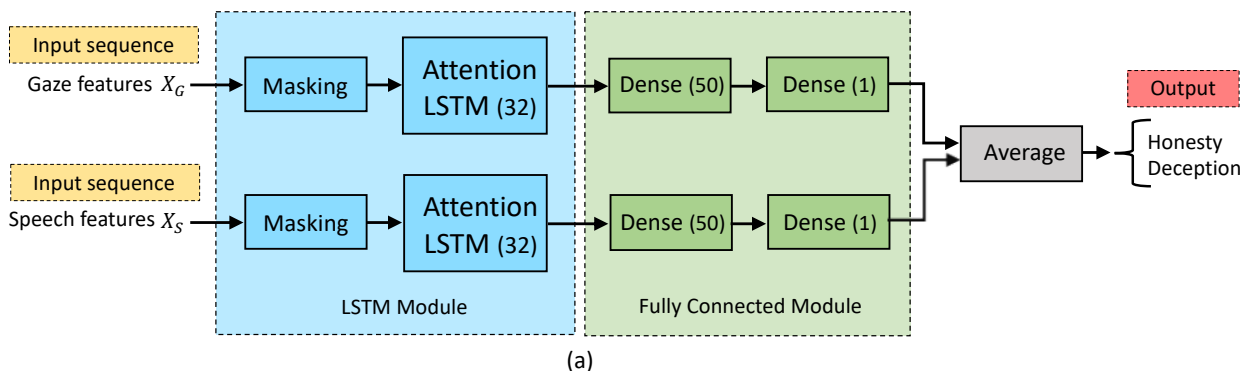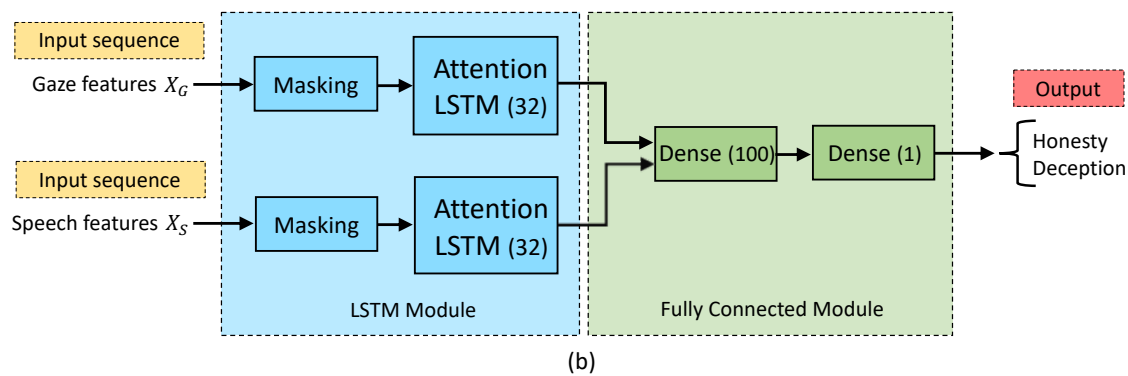

(a)

Figure 3. *Cont.*

(b)

Figure 3. Fusion strategies for the combination of gaze and speech modalities for deception detection. (**a**) Late Fusion: combination at decision level; (**b**) AP Fusion: combination at attention-pooling level.

2.5. Experimental Protocol

On the one hand, the reference system was developed using the MATLAB Statistics and Machine Learning toolbox. The optimal hyperparameters were obtained by minimizing a cross validation loss by means of a Bayesian optimizer.

On the other hand, all the LSTM-based systems were implemented with the Python's packages Tensorflow [43] and Keras [44]. In all cases, the architectures were trained during a maximum of 50 epochs using the Adam optimizer with an initial learning rate of 0.0002, a batch size of 64 and a binary cross-entropy loss function.

Regarding the experimental protocol, a subject-wise three-fold cross validation was used, according to [12]. Specifically, the database was split into three balanced groups. In each fold, two groups were used for training, and the remaining group was divided into two parts, one was kept for validation and the other one for test. The experiments were repeated three times rotating the training, validation and test sets, averaging the results afterwards. In all cases, training, validation and test sets were disjoint, in such a way that all gaze or speech recordings from the same subject were included in only one of these sets. We adopted this subject-independent configuration in order to prevent the system to learn the participant's identity instead of his/her honest or deceptive behaviour.

2.6. Assessment Measures

Following [20], the developed systems were assessed in terms of the average deception accuracy (DACC), average honesty accuracy (HACC) and average identification rate or accuracy (ACC) at segment and turn levels, that were defined as,

$$DACC = \frac{TN}{TN + FP} \tag{4}$$

$$HACC = \frac{TP}{TP + FN} \tag{5}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

where TP and TN are respectively, the number of honest and deceptive segments/turns correctly recognized, FN is the number of truths classified as lies and FP is the number of lies classified as truths.

As DACC and HACC were not independent in the sense that there was a tradeoff between them, for an overall analysis of the systems, ACC and the Area-Under-the-Curve (AUC) at segment and turn levels are also reported. As a complementary information, results are also presented in terms of precision and recall in Appendix A.

In all cases, each experiment was run 10 times and therefore, results shown in the tables and figures contained in Section 3 are given as the average accuracies and AUCs across these 10 subexperiments.

3. Results

3.1. Results with a Single Modality

Table 1 contains the DACC, HACC, ACC and AUC at segment and turn level achieved by the SVM and LSTM-based systems with the two modalities under consideration: gaze and speech. Additionally, for a better visualization of the results, Figure 4a shows the mean and standard deviation of the AUC at segment level obtained by all the systems evaluated in this work, whereas Figures 4b contains the same information at turn level.

Firstly, it can be observed that LSTM-based systems outperformed the reference SVM-based ones regardless of the input features for all metrics. The only exception was the performance of the speech modality in terms of deception accuracy for the LSTM system that was worse than for the reference one. However, in this case, the honesty accuracy was considerably better for the LSTM-based system. The differences in performance between SVM and LSTM-based systems were especially noticeable for the gaze modality. In fact, in terms of AUC, LSTM with the speech modality achieved a relative increase with respect to SVM and speech of 8.69% and 8.10% at segment and turn level, respectively, whereas with LSTM and gaze a relative increase was obtained with respect to the corresponding baseline of 26.39% and 33.77% at segment and turn level, respectively.

Secondly, focusing on LSTM-based systems, it can be observed that at segment level, gaze features were more able to correctly detect deception than honesty. In contrast, the speech modality performed considerably better for detecting truths than lies. The performance at turn level presented the same trends.

Finally, regarding the general comparison between the modalities in terms of AUC in the LSTM framework, gaze outperformed speech, especially at turn level, where this first modality achieved a 22.83% relative increase in AUC with respect to the second one.

Table 1. Average deception accuracy (DACC) [%], honesty accuracy (HACC) [%], accuracy (ACC) [%] and Area-Under-the-Curve (AUC) at segment and turn level achieved by the SVM-based reference system and the LSTM-based system with either gaze or speech features as input.

Modality	System	Segment Level				Turn Level			
		DACC	HACC	ACC	AUC	DACC	HACC	ACC	AUC
Gaze	SVM	62.17	48.35	55.67	55.67	65.74	51.13	58.83	57.86
	LSTM	63.45	60.11	61.88	67.37	67.78	61.86	64.98	72.09
Speech	SVM	60.80	59.77	60.31	59.61	53.24	63.71	58.20	60.64
	LSTM	53.21	75.92	63.89	63.12	44.44	80.31	61.41	63.83

3.2. Results with the Multimodal Fusion

Table 2 contains the results attained by the systems where the two types of modalities, gaze and speech, are combined. In particular, the corresponding DACC, HACC, ACC and AUC metrics for the different fusion strategies proposed in this work are reported at both, segment and turn level. In terms of ACC and AUC, in the case of the SVM-based models, both combination strategies outperformed the corresponding systems with a single modality, Late Fusion being slightly better than Early Fusion. However, these results were significantly worse than those attained by any of the fused LSTM-based systems.

Analyzing the behaviour of the multimodal LSTM-based systems, it can be observed that both combination approaches (Late Fusion and AP Fusion) produced better results in terms of ACC and AUC than the LSTM-based systems with a single modality. Moreover, the combination at Attention-Pooling level seemed to be more beneficial than the fusion at decision level. Specifically, the best system (AP Fusion) achieved a relative increase in

AUC of 5.18% with respect to Late Fusion, 18.63% with respect to Gaze and 28.01% with respect to Speech at segment level. At turn level, the respective relative increases in AUC were 8.91%, 19.81% and 38.13%.

Table 2. Average deception accuracy (DACC) [%], honesty accuracy (HACC) [%], accuracy (ACC) [%] and Area-Under-the-Curve (AUC) at segment and turn level achieved by the SVM-based reference system and the LSTM-based system with the different fusion strategies.

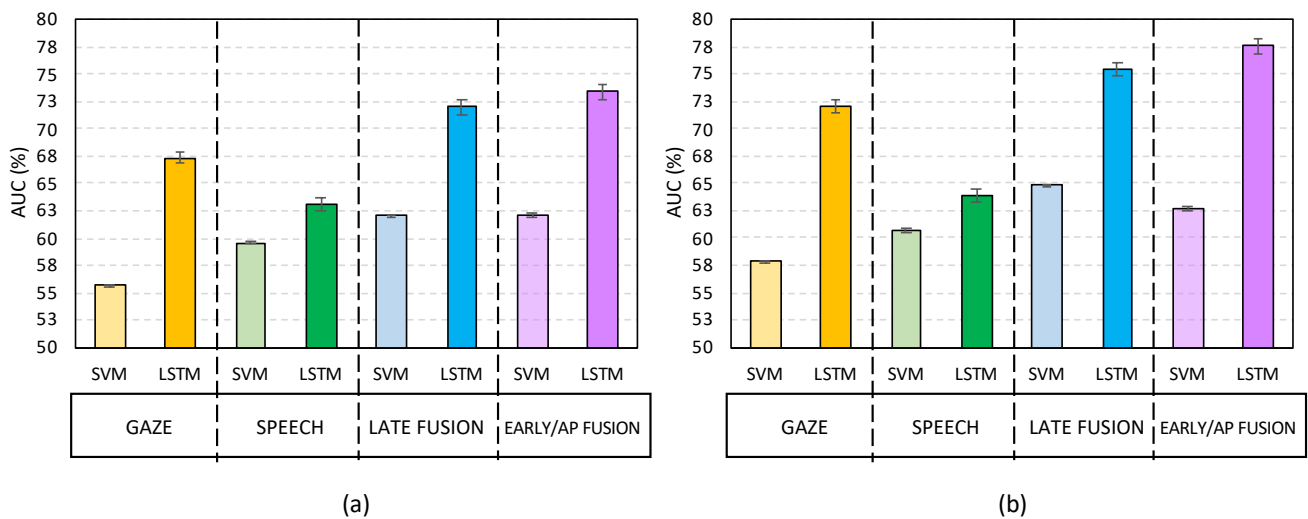|		Segment Level				Turn Level			
Modality	**System**	**DACC**	**HACC**	**ACC**	**AUC**	**DACC**	**HACC**	**ACC**	**AUC**
Late Fusion	SVM	60.59	61.86	61.19	62.04	50.93	72.37	61.07	64.84
Late Fusion	LSTM	62.50	72.82	67.35	72.00	57.75	78.87	67.74	75.43
Early Fusion	SVM	58.97	63.23	60.97	62.12	48.98	71.03	59.41	62.75
AP Fusion	LSTM	66.58	70.99	68.66	73.45	68.06	73.32	70.55	77.62

(a)

(b)

Figure 4. Mean and standard deviation of AUC [%] for both, SVM-based and LSTM-based systems, at (**a**) segment level and (**b**) turn level.

4. Discussion

Results corroborate the two main hypotheses of this paper. The first hypothesis focuses on the importance of the modeling of the dynamics of gaze and speech feature sequences in the development of ADD systems, whereas the second one points out the complementary nature of the gaze and speech modalities for this task.

Regarding the first hypothesis, the achieved results suggest that the temporal evolution of both, gaze and speech features, contain relevant information for the task of deception detection and for that, it is required to choose classification techniques able to properly model it, as is the case of LSTM networks. The success of the LSTM-based systems for this task also rely on the attention mechanism, that is able to automatically determine which frames of the gaze and speech sequences are most significant for the discrimination between truths and lies. As a consequence, it seems crucial to accentuate the contribution of the more relevant frames for the task regardless the modality used.

The behavior of the attention mechanism is illustrated in the following examples. Figure 5a shows an instance of the gaze modality, where from top to bottom are depicted the FPOGX, FPGOY, MPD feature sequences and the corresponding gaze Attention-Pooling weights corresponding to a recording segment of 3 s labeled as deception. As a general observation, the gaze weights present a significant variation with time according to the importance that the attention procedure gives to each temporal frame. As can be observed,

large weights are assigned to the segment from 6.3 to 6.8 s where FPOGX and FPOGY present a large deviation from the mean value, which means that the user made a sharp eye movement. According to [5,8,9,13], this kind of gaze dynamics are indicators of deception. In this slice, MPD shows a high degree of fluctuation, very likely due to the limitations of the eye-tracker for measuring the pupil size during an abrupt fixation change.

For the same 3 s segment, Figure 5b depicts from top to bottom the waveform, log-mel spectrogram and Attention-Pooling weights for the speech modality. Again, the weight value of a specific frame is related to its relevance to the task according to the attention mechanism. In this case, as can be expected, larger weights are assigned to high energy speech frames. However, in silence regions, like the slice from 6.4 to 7.5 s, weights are smaller but greater than zero. This fact supports the hypothesis that pause characteristics are cues of lie detection, as pointed in [15].
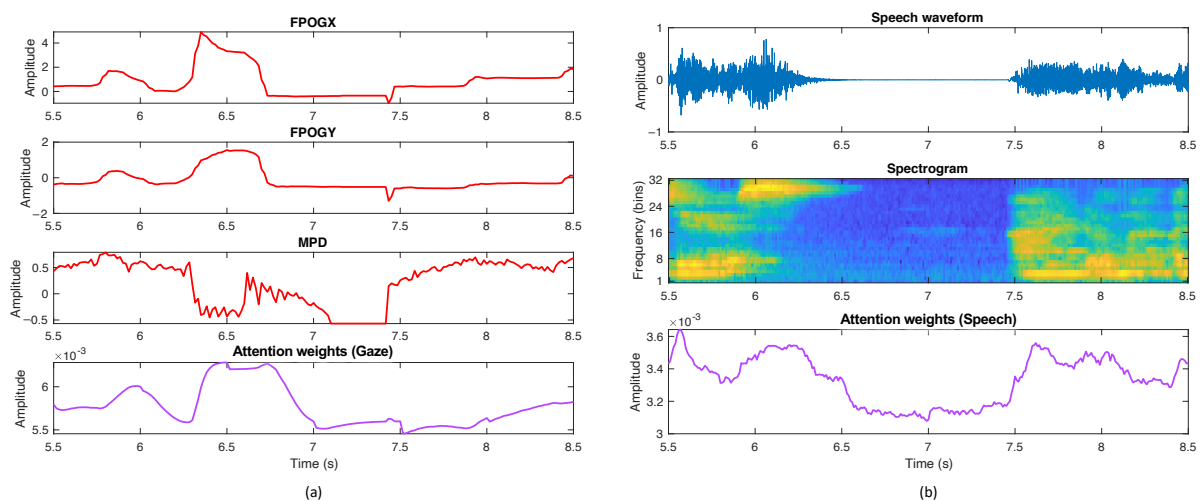

Figure 5. Attention-Pooling weights for a recording segment of 3 *s* for gaze and speech features. (**a**) Gaze modality, from top to bottom: FPOGX, FPGOY, MPD feature sequences, gaze attention weights; (**b**) Speech modality, from top to bottom: waveform, log-mel spectrogram and speech attention weights. This segment belongs to the turn #7 of user #34 that has been labeled as deception.

Regarding the second hypothesis, results achieved by the fused systems show that gaze and speech carry complementary information about deceptive or honest behaviour. In order to highlight this fact, we have obtained the confusion matrices at turn level produced by the single-modality systems and the multimodal system with the AP Fusion strategy. These confusion matrices are depicted in, respectively, Figure 6a–c, where the rows correspond to the correct class, the columns to the hypothesized one and the values in them are computed as the average over the test recordings belonging to the first fold.

As can be observed, for the gaze modality, deception is less confusable than honesty, i.e., it presents a higher identification rate, whereas there is about 36% of truths that are misclassified as lies. In contrast, for the speech modality, the class presenting the best performance is honesty. In this case, more than 55% of lies are incorrectly assigned to the truth class. This distinctive behaviour supports the hypothesis that both modalities carry complementary information, so their fusion is likely to outperform the single-modal systems, as is ratified by the confusion matrix of the multimodal system represented in Figure 6c, and the results shown in Table 2 and Figure 4.

Finally, as for the comparison of the proposed system with previous studies over the same database, Gupta et al. [12] obtained an accuracy at turn level of 57.11% with RF as classifier for the gaze modality and of 56.22% with KNN for the speech modality using hand-crafted compact features in both cases. In contrast, our attention LSTM-based system achieves identification rates of 64.98% and 61.41% for gaze and speech, respectively. This shows that the modeling of the temporal evolution of the features is essential for the

ADD task. In the same way, the multimodal system proposed in [12] consisting of the combination of gaze and audio modalities produced an accuracy of 59.42%, whereas our best system (gaze and speech combined with the AP Fusion strategy) obtains a result of 70.55%, that is significantly better.
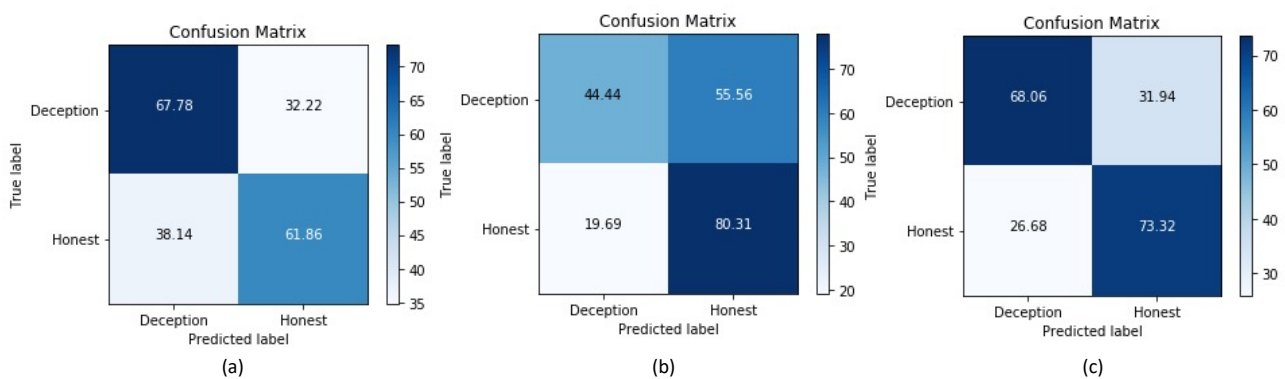

Figure 6. Confusion matrices at turn level [%] for the LSTM-based system and three different cases: (**a**) Gaze; (**b**) Speech; (**c**) Multimodal system with AP fusion.

Limitations and Future Research

The automatic detection of deception is a very challenging task for several reasons. Firstly, it is inherently a hard task even for humans, because deceptive behavior is often subtle as liars try to hide that they are not telling the truth. Secondly, there is not a general consensus about what are the physical or behavioral characteristics that must be observed for detecting if a person is lying, and besides, the way people deceive varies across different subjects. Thirdly, there is a lack of training data due to the difficulty of collecting databases in realistic conditions. Fourthly, in real scenarios the presence of noise can deteriorate the quality of the recordings, leading to the system performance degradation. In fact, in this particular database, audio is very noisy and this could be the reason for the worse performance of this modality in comparison to the gaze one, as eye-tracking data has been recorded in better environmental conditions (see Table 1).

All these factors determine that the performance of the ADD systems are far to be optimal, opening new research lines for addressing these issues. In particular, for future work, we plan to extend our research towards the exploration of suitable data augmentation techniques for alleviating the problem of scarce training data, the use of advanced denoising techniques for improving the quality of the audio recordings and the incorporation of the visual modality.

5. Conclusions

This paper deals with the development of an automatic deception detection system based on gaze and speech features. We present two main contributions. In the first one, we explore the use of attention LSTM networks for single-modal ADD systems with either frame-level gaze features (fixations and pupil diameter) or frame-level speech features (log-mel spectrograms) as input. In the second contribution, we present a multimodal ADD system where two fusion strategies for the combination of the gaze and speech modalities into the attention LSTM architecture have been tested, namely Late Fusion and Attention-Pooling Fusion.

The proposed systems have been assessed over the Bag-of-Lies dataset, a multimodal database for deception detection recorded in real conditions. First, it can be observed that attention LSTM-based systems significantly outperform the performance of the reference systems that are based on traditional machine learning methods (in particular, SVM) running over compact feature representations. These results show that attention LSTM networks are able to properly modeling the dynamics of raw gaze and speech feature sequences that contains relevant cues of deception. Second, both combination strategies,

Late and AP Fusion, achieve better results in terms of accuracy and Area-Under-the-Curve than the individual systems. In particular, our best system (Attention LSTM + AP Fusion) achieves a relative increase in AUC at turn level of 19.81% with respect to Attention LSTM + Gaze, 38.13% with respect to Attention LSTM + Speech and 36.35% with respect to the best multimodal SVM-based system (SVM + Late Fusion).

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
ADD	Automatic Deception Detection
AUC	Area-Under-the-Curve
AP	Attention Pooling
DACC	Deception Accuracy
EEG	Electroencephalogram
FPOGX	Horizontal Fixation Point-Of-Gaze
FPOGY	Vertical Fixation Point-Of-Gaze
HACC	Honesty Accuracy
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MPD	Mean Pupil Diameter
RF	Random Forest
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine

Appendix A

Depending on the application for which the ADD system is intended to be used, it could be desirable for models to be as much precise as possible in detecting truths or lies. For this reason, we have complemented the results shown in Section 3 by analysing the

LSTM-based systems performance in terms of deception precision (DPREC), deception recall (DREC), honesty precision (HPREC) and honesty recall (HREC). DPREC and HPREC as computed as,

$$DPREC = \frac{TN}{TN + FN} \tag{A1}$$

$$HPREC = \frac{TP}{TP + FP} \tag{A2}$$

Note that DREC and HREC match the values of DACC and HACC and are computed by using, respectively, Equations (4) and (5). All these precision and recall metrics are contained in Table A1.

On the one hand, for the single-modal systems, it can be observed that the speech modality is able to detect deceptions more precisely than gaze, whereas the honesty precision is better when using eye-tracking data, especially at turn level.

On the other hand, in general, the two proposed fused systems outperform the corresponding individual systems in terms of DPREC and HPREC and are more precise in detecting lies than truths. In particular, at turn level, the best system (*Attention LSTM + AP Fusion*) achieves a relative increase in DPREC of 22.82% and 7.93% with respect to *Attention LSTM + Gaze* and *Attention LSTM + Speech*, respectively, whereas the relative improvement in terms of HPREC is of 16.47% and 29.70% with respect to the same single-modal systems.

Table A1. Average deception precision (DPREC) [%], deception recall (DREC) [%], honesty precision (HPREC) [%] and honesty recall (HREC) [%] at segment and turn level achieved by the single-modal and multimodal LSTM-based systems.

Modality	System	Segment Level				Turn Level			
		DPREC	DREC	HPREC	HREC	DPREC	DREC	HPREC	HREC
Gaze	LSTM	64.18	63.45	59.38	60.11	66.44	67.78	63.44	61.86
Speech	LSTM	71.52	53.21	59.08	75.92	71.87	44.44	56.56	80.31
Late Fusion	LSTM	72.20	62.50	63.33	72.82	75.32	57.75	62.71	78.87
AP Fusion	LSTM	72.15	66.58	67.42	70.99	74.01	68.06	69.46	73.32

References

1. Meservy, T.; Jensen, M.; Kruse, J.; Burgoon, J.; Nunamaker, J.; Twitchell, D.; Tsechpenakis, G.; Metaxas, D. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intell. Syst.* **2005**, *20*, 36–43. [CrossRef]
2. Tsikerdekis, M.; Zeadally, S. Online Deception in Social Media. *Commun. ACM* **2014**, *57*, 72–80. [CrossRef]
3. Efthymiou, A.E. Modeling Human-Human Dialogues for Deception Detection. Master's Thesis, University of Amsterdam, Amsterdam, The Netherlands, 2019.
4. Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Xiao, Y.; Linton, C.; Burzo, M. Verbal and Nonverbal Clues for Real-life Deception Detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 2336–2346. [CrossRef]
5. Wang, J.T.Y.; Spezio, M.; Camerer, C.F. Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games. *Am. Econ. Rev.* **2010**, *100*, 984–1007. [CrossRef]
6. Pak, J.; Zhou, L. Eye Movements as Deception Indicators in Online Video Chatting. In Proceedings of the AMCIS 2011 Proceedings, Detroit, MI, USA, 4–8 August 2011.
7. Fukuda, K. Eye blinks: New indices for the detection of deception. *Int. J. Psychophysiol.* **2001**, *40*, 239–245. [CrossRef]
8. Vrij, A.; Oliveira, J.; Hammond, A.; Ehrlichman, H. Saccadic eye movement rate as a cue to deceit. *J. Appl. Res. Mem. Cogn.* **2015**, *4*, 15–19. [CrossRef]
9. Borza, D.; Itu, R.; Danescu, R. In the Eye of the Deceiver: Analyzing Eye Movements as a Cue to Deception. *J. Imaging* **2018**, *4*. [CrossRef]
10. Pak, J.; Zhou, L. Eye Gazing Behaviors in Online Deception. In Proceedings of the AMCIS 2013 Proceedings, Chicago, IL, USA, 15–17 August 2013.

11. Belavadi, V.; Zhou, Y.; Bakdash, J.Z.; Kantarcioglu, M.; Krawczyk, D.C.; Nguyen, L.; Rakic, J.; Thuriasingham, B. MultiModal Deception Detection: Accuracy, Applicability and Generalizability*. In Proceedings of the 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, GA, USA, 28–31 October 2020; pp. 99–106. [CrossRef]

12. Gupta, V.; Agarwal, M.; Arora, M.; Chakraborty, T.; Singh, R.; Vatsa, M. Bag-of-Lies: A Multimodal Dataset for Deception Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 83–90. [CrossRef]

13. Khan, W.; Crockett, K.; O'Shea, J.; Hussain, A.; Khan, B.M. Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection. *Expert Syst. Appl.* **2021**, *169*, 114341. [CrossRef]

14. DePaulo, B.M.; Lindsay, J.J.; Malone, B.E.; Muhlenbruck, L.; Charlton, K.; Cooper, H. Cues to deception. *Psychol. Bull.* **2003**, *129*, 74–118. [CrossRef]

15. Benus, S.; Enos, F.; Hirschberg, J.; Shriberg, E. Pauses in deceptive Speech. In Proceedings of the ISCA 3rd International Conference on Speech Prosody, Dresden, Germany, 2–5 May 2006.

16. Kirchhübel, C. The Acoustic and Temporal Characteristics of Deceptive Speech. Ph.D. Thesis, Department of Electronics, University of York, York, UK, 2013.

17. Hirschberg, J.B.; Benus, S.; Brenier, J.M.; Enos, F.; Friedman, S.; Gilman, S.; Girand, C.; Graciarena, M.; Kathol, A.; Michaelis, L.; et al. Distinguishing deceptive from non-deceptive speech. In Proceedings of the Interspeech 2005, Lisbon, Portugal, 4–8 September 2005. [CrossRef]

18. Mermelstein, P. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* **1976**, *116*, 374–388.

19. Wu, Z.; Singh, B.; Davis, L.S.; Subrahmanian, V.S. Deception detection in videos. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

20. Xie, Y.; Liang, R.; Tao, H.; Zhu, Y.; Zhao, L. Convolutional Bidirectional Long Short-Term Memory for Deception Detection With Acoustic Features. *IEEE Access* **2018**, *6*, 76527–76534. [CrossRef]

21. Rill-García, R.; Escalante, H.J.; Villaseñor-Pineda, L.; Reyes-Meza, V. High-Level Features for Multimodal Deception Detection in Videos. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1565–1573. [CrossRef]

22. Abouelenien, M.; Pérez-Rosas, V.; Mihalcea, R.; Burzo, M. Detecting Deceptive Behavior via Integration of Discriminative Features From Multiple Modalities. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1042–1055. [CrossRef]

23. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

24. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning Precise Timing with LSTM Recurrent Networks. *J. Mach. Learn. Res.* **2003**, *3*, 115–143. [CrossRef]

25. Zacarias-Morales, N.; Pancardo, P.; Hernández-Nolasco, J.A.; Garcia-Constantino, M. Attention-Inspired Artificial Neural Networks for Speech Processing: A Systematic Review. *Symmetry* **2021**, *13*, 214. [CrossRef]

26. Kao, C.C.; Sun, M.; Wang, W.; Wang, C. A Comparison of Pooling Methods on LSTM Models for Rare Acoustic Event Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020. [CrossRef]

27. Guo, J.; Xu, N.; Li, L.J.; Alwan, A. Attention based CLDNNs for short-duration acoustic scene classification. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017. [CrossRef]

28. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 577–585.

29. Huang, C.W.; Narayanan, S.S. Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016. [CrossRef]

30. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231. [CrossRef]

31. Gallardo-Antolín, A.; Montero, J.M. A Saliency-Based Attention LSTM Model for Cognitive Load Classification from Speech. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 216–220. [CrossRef]

32. Gallardo-Antolín, A.; Montero, J.M. External Attention LSTM Models for Cognitive Load Classification from Speech. *Lect. Notes Comput. Sci.* **2019**, *11816*, 139–150. [CrossRef]

33. Fernández-Díaz, M.; Gallardo-Antolín, A. An attention Long Short-Term Memory based system for automatic classification of speech intelligibility. *Eng. Appl. Artif. Intell.* **2020**, *96*, 103976. [CrossRef]

34. Gallardo-Antolín, A.; Montero, J.M. On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification. *Neurocomputing* **2021**, *456*, 49–60. [CrossRef]

35. Open Gaze API by Gazepoint. 2010. Available online: https://www.gazept.com/dl/Gazepoint_API_v2.0.pdf (accessed on 5 July 2021).

36. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, *2006*, 10.

37. Vázquez-Romero, A.; Gallardo-Antolín, A. Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks. *Entropy* **2020**, *22*, 688. [CrossRef]

38. Gil-Martín, M.; Montero, J.M.; San-Segundo, R. Parkinson's Disease Detection from Drawing Movements Using Convolutional Neural Networks. *Electronics* **2019**, *8*, 907. [CrossRef]

39. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6. [CrossRef]

40. McFee, B.; Lostanlen, V.; McVicar, M.; Metsai, A.; Balke, S.; Thomé, C.; Raffel, C.; Malek, A.; Lee, D.; Zalkow, F.; et al. LibROSA/LibROSA: 0.7.2. 2020. Available online: https://librosa.org (accessed on 5 July 2021).

41. Vapnik, V.; Chervonenkis, A.Y. A note on one class of perceptrons. *Autom. Remote Control* **1964**, *25*, 61–68.

42. Huang, C.; Narayanan, S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the ICME 2017, Hong Kong, China, 10–14 July 2017; pp. 583–588.

43. Abadi, M. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org (accessed on 5 July 2021).

44. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 5 July 2021).