ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **21**, Number 4, 2018, 460–474

Decoding communication: a deep learning approach to voice-based intention detection

Eduard FRANŢI[1, 2], Monica DASCALU[2, 3], Ioan ISPAS[2, 4], Ana Voichita TEBEANU[5], Zoltan ELTETO[2], Silvia BRANEA[6], and Voichita DRAGOMIR[3]

[1]IMT Bucharest, 126A, Erou Iancu Nicolae Street, Romania
[2]ICIA 13 September Street, Bucharest, Romania
[3]*Politehnica* University of Bucharest, Faculty of Electronics, Telecommunication and Information Technology, 313, Splaiul Independentei Street, Bucharest, Romania
[4]*Politehnica* University of Bucharest, Doctoral School of the Faculty of Electronics, Telecommunication and Information Technology, 313, Splaiul Independentei Street, Bucharest, Romania
[5]*Politehnica* University of Bucharest, Teacher Training Department, 313, Splaiul Independentei Street, Bucharest, Romania
[6]University of Bucharest, Faculty of Journalism and Communication Studies, Communication and Public Relations Department, 1-3, Iuliu Maniu Street, Bucharest, Romania

Abstract. This paper presents an original method of intention detection that can open a new direction of research in voice-based affective computing. A deep learning approach was used to detect the consistency between real and expressed intentions of a speaker (or the inconsistency, that is related to deceiving – or manipulative – intention), as reflected in their voice. The labeling and triangulation of results imply a qualitative research method, critical discourse analysis, and require expert evaluation. The method was implemented in a software platform integrated with the neural network programming frame. The deep learning architecture selected is based on similar models used by the authors in affective computing applications. The experimental research applied the proposed method for a famous historical case: US President Richard Nixon's audio speeches from the 'Watergate affair'. A labeled data base of 2758 files (2 seconds audio fragments) was generated, based on publicly available voice recordings of President Nixon. These files were used for training and tests and an accuracy of over 94% was obtained.

1. Introduction

This paper introduces a new direction of research in affective computing: intention detection. Although there are different psychological models of intentions and there is not, to our knowledge, a definitive explanation of intentions relating to emotions, thoughts and different physiologic parameters, there is scientific evidence that intentions are reflected in communication processes, for instance in speech (both at the textual level and means of expression, including voice). Since a theoretical model of the parameters for intention identification in the human voice

has not yet been built, we have used classical text analysis methods to replace the 'missing link' of intentionality.

To a certain point, our approach resembles voice-based emotion detection, as it exploits the very broad skills of deep learning architectures (like convolutional neural networks, CNN) to extract information from the human voice. Voice-based affective computing is based on the fact that someone's emotions trigger certain psychological and physiological changes within, which also change the speaker's voice [1]. CNN based solutions in the frame of voice-based affective computing reported in scientific literature include lie detection [2] and emotion detection [3]. So far there are no research results reported in the scientific literature - at our knowledge - regarding intention detection, in the psychological sense of intention. A project for intention detection based on voice, text, and image processing [4] announced by Microsoft in 2008 seems to be focused on human-computer interaction and aims to predict the users' intentions for the enhancement of computer applications. Noguchi et al. [5] describe an indoor activity support system that is able to detect the user's intention to move.

The research in intention detection is based on the hypothesis that intentions are in a certain manner impregnated in the human voice. This work hypothesis was based on theories and experimental results in psychology, semiotics and rhetoric, that analyzed the connections between voice, content of discourse and intentions of the speaker [6].

The paper is organized as follows: the conceptual model for intention detection and the interdisciplinary theoretic background of the research are presented in section 2. Two complementary approaches were used: voice analysis and text analysis. Fundamental ideas from psychology are introduced, together with the method selected for text analysis.

Section 3 presents the application of our method in a particular case study (US president Richard Nixon and the Watergate affair). This particular case was selected due to the availability of the data – voice recordings, historic documentation, and previous scientific analysis, which allow us to construct and verify the model and the experiments.

Section 4 explains the research methodology (regarding, mainly, the generation of the database and labeling of the files). Special software was developed for the labeling of the audio and text files and to integrate the CNN programming with the database generation. The software is presented in Section 5. The details of the CNN structure are given in Section 6.

Section 7 describes the experiments performed with the audio recordings and transcripts of President Nixon's speeches and conversations related to the Watergate situation. The paper ends with some conclusions and further directions of research.

2. Voice-based and text-based approach to intention: theoretical background

The theoretical background for intention detection in this research is given by psychology (that explains the relation between the intentions of the speaker, their speech and vocal characteristics) and linguistics, semiotics and rhetoric theories that correlate the intentions with the content of the speech. Thus, we have two separate means of analysis (Figure 1) that permit both the labelling of intentions and the triangulation of results.

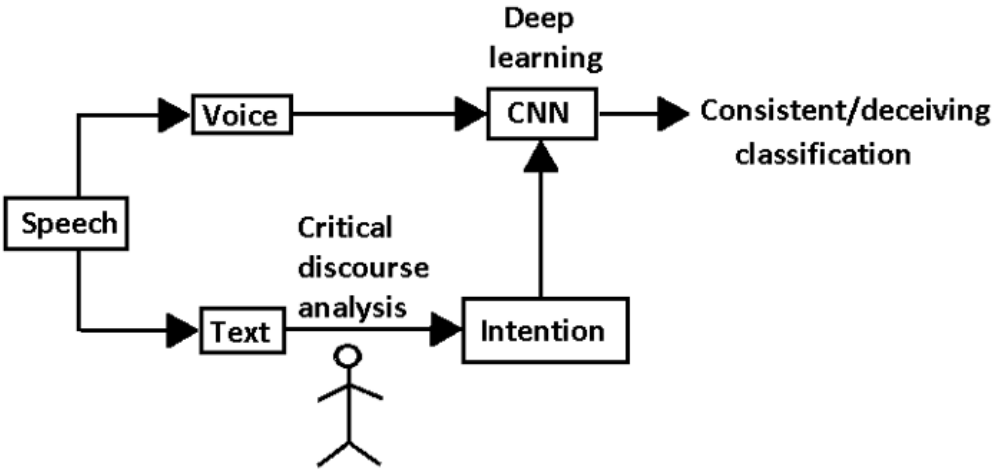

Fig. 1. Conceptual map of intention detection; the voice analysis is automatic, while the text analysis needs the human (expert) evaluation.

Several theories in psychology state that someone's intentions influence the behavior, the gestures, the mimic, the voice and the words one uses to express ideas and feelings [7]. Freud wrote about the connection between voice and intention: *the suppression of the speaker's intention to say something is the indispensable condition for the occurrence of a slip of the tongue* [8]. Freud also demonstrated that mistakes in everyday life (such as forgetting familiar names or reading mistakes) are not accidental, but events that show some psychological inner conflicts, and also reveal hidden intentions [9].

The intonation of the voice gives many clues about one's psychic and emotional state: *"The intonation rich in inflections is characteristic of individuals with a rich affective background and at the same time tends, consciously or less consciously, to impress the (affective) interlocutors. Instead, the flat, monotonous, intonation of infatuation may denote either a poor affective background, or certain difficulties or inhibitions in social behavior such as the inability of exteriorizing their feelings, difficulties in establishing contacts with people due to shyness, etc."* [10].

According to Scherer: *"In social interaction, the outcome of a particular act usually depends partly on the reactions of the significant others in the encounter. Thus, the projection of how the other(s) will react to the different courses of action open to the organism largely determines its behavioral choice. The signaling of the emotion communicates the emoting organism's evaluative reaction to a stimulus event or act and thus narrows the possible range of behavioral intentions that are likely to be inferred by observers. In addition, because of the action tendency component of the emotion, the momentary behavioral intention will be expressed even more clearly"* [11].

The definitions of intention in psychology literature underlines that every intention implies the existence of a purpose and a plan. Cohen and Levesque define intention as follows: *"Intention is choice plus motivating engagement"* [12]. For Moses and Malle [13], intention's specificity is a purpose and a plan to perform the required actions, which will lead to a desired outcome. The most important difference between intention and desire is that desire is focused on the outcome, without any representation on how to obtain it. McTaggart explains: *"An intention was directed at the intender's own actions; it required some sort of reasoning; it required a commitment to do*

the intended deed. Intention implied purposefulness: an understanding of a plan of action and a planned satisfactory result" [14].

Therefore, intention is more complex than desire: *"Intentions have a stronger connection to goals or outcomes than desires because they imply a commitment, and embrace at least some form of partial planning to achieve the goals or outcomes, while desires do not. Intentions to act are directly linked to a multitude of activities and outcomes related to the choice of means for action implementation, impediments to action, temptations to perform other actions, or consider other goals, cues for retrieval of the intention at a future point in time, and so forth"* [15].

Semiotics' perspective, on the other hand, looks at the content of the discourse and states that all of a person's intentions, influence not only their voice but also the words they choose to express their ideas. For instance: *"intention is a decisive factor in what constitutes the message. What hits our tympanum will be a message only if it is something pre-conscious as a signal for a listener. The intention of the transmitter may be specified or unspecified (whether or not notified), whether conscious or not, but identifiable by analyzing the message. The message is what the transmitter puts inside by whatever means"* [6].

Any idea can be expressed in many ways, but speakers will choose those words that are consistent with their intentions, even when they are secret. Rascanu writes: *"In most cases, to express one thing it is necessary to choose between several verbal expressions, similar in meaning, but with different nuances. The choice is never accidental because, without the individual's will, it is largely determined by attitude to the reality taken into account: accepting or rejecting, minimizing or overestimating, etc."* [16].

Sometimes the intentions of a speaker are oriented towards a specific goal that is not revealed to the audience and, more so, expect a particular action from the audience. In order to obtain a specific action or reaction, emotions play an important role. Although ideas can convince listeners at the theoretical level, emotions will be the ones that will always make them move to action: *"The speakers who know how to impress the crowds appeal to their feelings and never to their reason"* [17]. Words are important, but the voice that deliver those words will be the one that will awaken intense emotions, related to the intentions of the speaker. About the power of voice, Cicero wrote: *"Nothing seems more beautiful than captivating the attention of a gathering by the power of the word, inciting the minds of the listeners and determining their will in one sense or another. Is there anything stronger and more magnificent than that a single man can only change the word of the crowd, shake the judges' consciences and the authority of the state?"* [18].

The real intentions behind a discourse may be revealed by the content of speech and the way the speaker delivered it: the quality of the voice (stuttering, interjections, voice tremor), inappropriate tone (shouted / whispered), volume and speed (too loud / too slow), or speech-reading errors. About identifying hidden intentions, Codoban writes: *"the spoken or written words tell us, over what they communicate directly, and what is thought or felt without being told. The interference between language and thought produces in expressions formulated in words a kind of resonance of connotations, a kind of message over a message, a metamorphosis or a metalanguage. "Metalanguage" is not entirely an adequate term that proposes the interpretation of words and expressions to guess the speaker's intentions and ideas, or how we go beyond language to discover what people think"* [19].

In linguistics, communication sciences, rhetoric and other scientific branches that imply text analysis, both quantitative and qualitative methods of research were developed. Among several research methods available, the methodological approach of critical discourse analysis was se-

lected because of its potential in identifying the speaker's real intentions. According to prominent researchers in the field, like Fairclough [20] and van Dijk [21], critical discourse analysis may reveal the ideology behind the discourse. In spite of the controversies regarding the concept of ideology, there is a consensus that textual and contextual details of a discourse reveal much more (about the agenda of the speaker) than the factual information as intended by the speaker/author of the discourse [22]. Another advantage of the method is the fact that it is a qualitative method and it may be applied for small samples of data.

So far, there is no automatic tool for such a qualitative method – human analysis is required and thus the analysis is subject to interpretation. The analyst's role is not to establish the veracity of a discourse, but to identify the small details that reveal more information regarding the agenda of the author of the discourse (and this is related to the intentions).

To conclude, the research results and theoretical models from psychology, and communication sciences justify the hypothesis that both the human voice and the textual means of expression are influenced in a specific way by intentions. However, there is no taxonomy of intentions (similar to the taxonomy of emotions), hence the classification task in this direction misses a theoretical foundation. Broad classification in good/bad intentions, or altruistic/egotistic intentions is subject to relativity of perception and interpretation. A more approachable question is whether the declared intentions of a speaker are coherent with the real intentions, which in our opinion is a more refined question than the true/false dichotomy.

The object of our experiments belongs to public communication, as there are several domains where this question is particularly interesting: politics, advertising, justice. There is always some suspicion regarding, for instance, the veracity of politicians, as 'popular wisdom' states that the real intentions of a political discourse are often hidden by dogmatic rhetoric. According to practitioners of critical discourse analysis, this possible inconsistency can be scientifically deduced.

In order to look for intention-related information in the voice, we have selected a famous case that can already be seen in a historical perspective. It was already thoroughly documented and studied from several scientific angles and therefore it offers a solid basis for comparison of the 'intention imprint' in the text and in the voice. Since the intention of deception is established already [23], we have used the critical discourse analysis to select the parts of the discourse richer in intention-related information.

3. Case study for deceiving intentions: US President Richard Nixon and the 'Watergate scandal'

The Watergate scandal took place in the United States between 1972 and 1974 and was triggered by the arrest of five people who entered the offices of the Democratic Party's National Committee (from the Watergate building) to install listening devices. In the investigations that followed in the US Senate and Supreme Court, it turned out that the Watergate incident was part of a secret plan of spying on the National Committee of the Democratic Party for the presidential election in 1973. US President Richard Nixon publicly denied any connection to the case, but the investigations documented his involvement in the plan behind. The main evidence consisted of some audio records in which Nixon gave the order to stop the investigations in that case. As a consequence, Nixon resigned as president of the US, on August 9, 1974.

Paul Ekman wrote about Nixon as follows: "*Former president Nixon is probably the public*

official who has been most often condemned for lying. He was the first president to resign, but it was not simply because he had lied. Nor was he forced to resign because people working for the White House were caught at the Watergate office and apartment complex in June of 1972 attempting to break in to the Democratic party headquarters. It was the cover-up he directed and the lies he told to maintain it. Audiotapes of conversations in the White House, later made public, revealed Nixon to say at the time, 'I don't give a shit what happens, I want you to stonewall it, let them plead the Fifth Amendment, or anything else, if it'll save it—save the plan'. The cover-up did succeed for nearly a year until one of the men convicted for the Watergate break-in, James McCord, told the judge that the burglary was part of a larger conspiracy. Then it came out that Nixon had audio- taped all conversations in the Oval Office. Despite Nixon's attempt to suppress the most damaging information on those tapes, there was enough evidence for the House Judiciary Committee to bring articles of impeachment. When the Supreme Court ordered Nixon to turn over the tapes to the grand jury, Nixon resigned on August 9, 1974" [24].

Therefore, the Nixon-Watergate case offers the context and material of analysis for a documented situation were declared intentions of a discourse are not similar with the real intentions (and the intention to deceive intention is a reasonable presumption). What we are proposing for now is to search for information in the voice that delivered the deceiving message, using the voice recordings publicly available on the internet.

4. Methodology of the research

The interdisciplinary research background shows that the intentions are reflected in the voice and in the content of communication. In absence of a theoretical taxonomy of the intentions, the experimental part of the research aims to find an automatic method to verify the consistency of the declared intentions of a speaker and his/her real intentions. The main objective of this research is to retrieve the information in the voice regarding deceiving intentions (inconsistency between declared and real intentions).

So far, we have no reasonable clue on what variables of the voice contain this information, therefore we decided to use a deep learning architecture, that is able to access the information contained in large data sets even in the absence of a theoretical model. We have used a CNN architecture that was developed for other affective computing applications, adjusting the number of classes to 2 (corresponding to 'deceiving' or 'honest' intentions). The details of the CNN architecture are presented in Section 6.

The methodological approach follows the following scheme (Figure 2):

- Primary database with recordings and translations;
- Primary labeling of the recordings as deceiving/neutral based on contextual analysis and historical documentation (particularly for the official speech regarding the Watergate scandal, the critical discourse analysis was used for selection of fragments of the speech denser in targeted information);
- Selection and labeling of significant recordings for deceiving fragments (where the inconsistency of the declared versus real intention is revealed through critical discourse analysis of the text) as explained in Section 5;
- The experimental set of recordings for deep learning method is divided in training and test subsets;

– Training and test of the CNN (see sections 6 and 7);
– Further experiments with the trained CNN (section 7).

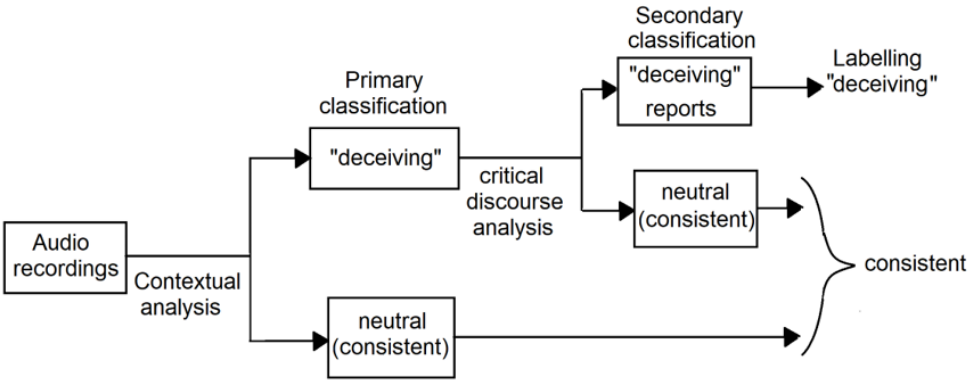

Fig. 2. The methodological scheme for database labeling.

Note that after primary and secondary classification (Figure 2), the files are labeled as 'deceiving' and 'neutral', or, in our experiments, 'manipulative' and 'non-manipulative'. The significance of this labeling is that the so-called 'deceiving' recordings are characterized by the inconsistency between the declared intention of the discourse (as given by the general content examination) and the real intention (as revealed by context analysis and critical discourse analysis).

5. Experimental software platform

This section presents the software tool used for data analysis and labelling. The GUI (graphic user interface) of this software is reproduced in Figure 3. The software was specially conceived to help the operator select audio fragments corresponding to selected text fragments. Expert human analysis of the text is required and can be done prior the audio selection.

The workspace has two main windows, for text (transcript) and audio waveform. If the user selects a fragment in the text window, the corresponding waveform is selected and labeled. There are several functions available. The operator can listen to the selected fragment to decide if the labeling is significant; the pauses between words can be deleted etc.

In order to obtain a consistent and large database, necessary for the training of the neural network, the audio recordings are further divided in 2 seconds pieces with the corresponding labeling. Two classes of files are generated. They are labeled in our experiment M (manipulative content, corresponding to deceiving intentions) and NM (*non-manipulative*). For different experiments, different classes and tags may be defined. The number of files created for each class is displayed in the lower left corner of the working space.

The same software will do the random selection of files for training and test, based on the percentage of files used for training vs. test. The software is integrated with the CNN programming and training (see next section). Additional processing operation on the labeled files is performed

by the software. In our experiments, since the classification is binary, we used the spectrogram of the processed waveforms.

When the database is completed, the 'start training' button in the lower part of the screen starts the transfer of the processed audio recordings to the CNN software. The processed files (here, the spectrograms) are the input of the neural network in the phases of training and test (see next section). Once the trained network is available, the user can use the same software for the classification of new files using the trained CNN, or check the classification for the existing files. The workspace is similar (Figure 4). Our experiments (section 7) proved the right classification for other recordings of the same speaker (here, President Nixon) that were not used in the phases of training and test of the CNN. The experiments are supervised by qualified experts in critical discourse analysis that check the classification.

The user can divide the new text in fragments or use the implicit in 2 seconds segmentation of audio files.

The 'prediction' offers a 'score' of deceiving (manipulation) for each file (in the range 0–100%). These refer to the content of a whole file. The CNN classifies 2 seconds fragments that cover the whole file. Files with a percentage of over 50% are included in the M category. All 2-second audio fragments that have been categorized by the CNN in the M category are marked together with the corresponding text fragments and the operator can hear them by pressing the play button. In Figure 4, two audio fragments were detected as belonging to 'Manipulative' (deceiving intention) class.

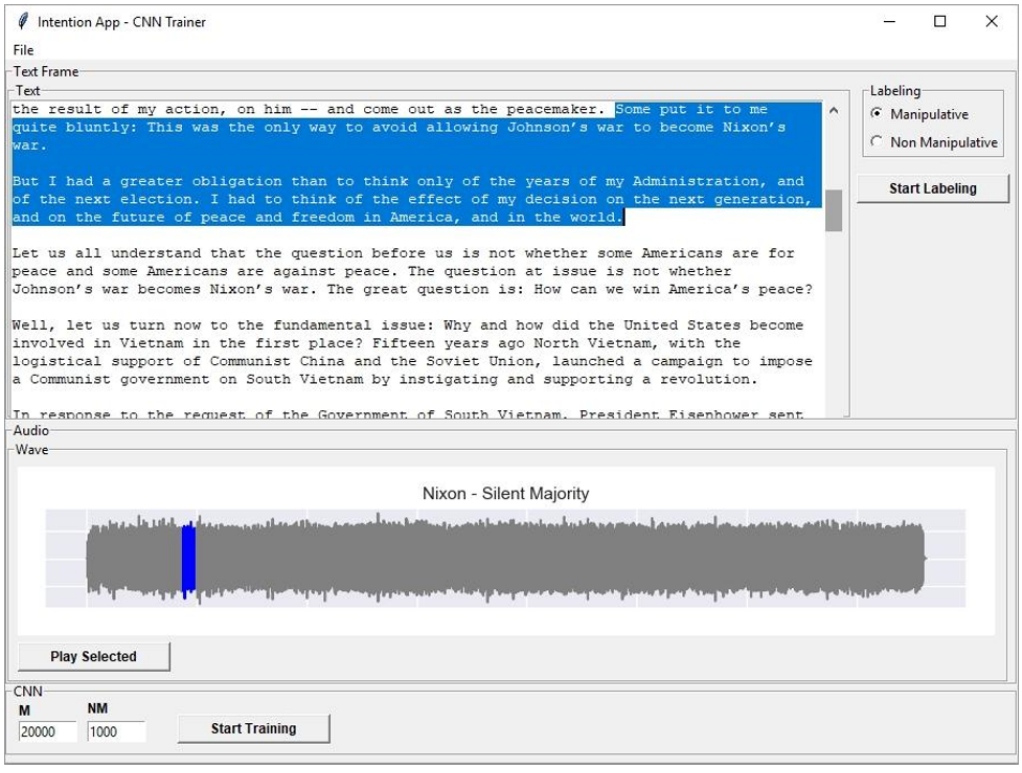

Fig. 3. GUI of the software platform for text/audio selection and labelling.

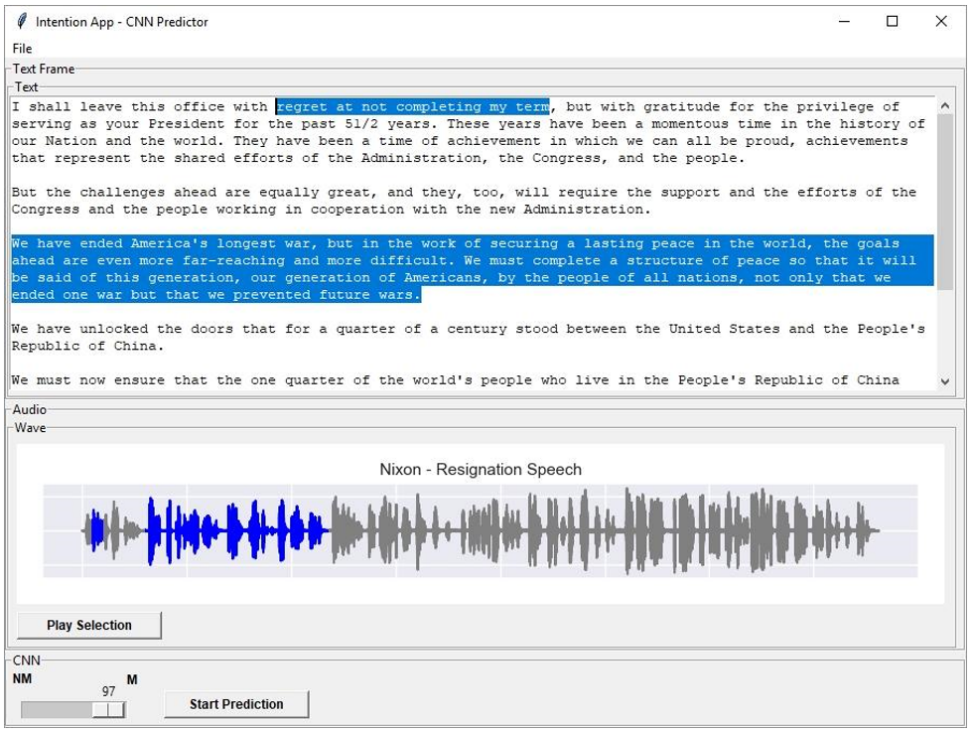

Fig. 4. Prediction functionality of the experimental software.

6. CNN structure

The deep learning architecture used was verified in affective computing applications. [3], [25] The CNN was implemented in Python, using the TensorFlow back-end, with the Keras framework [26], adapting the model from [27] which is an image recognition application. The structure of the CNN contains:

– 2 convolutional layers; 16 filters (5 x 5) on the first layer and 32 filters (3 x 3) on the second layer. On these two layers there are a total of 5056 trained parameters

– 2 dense layers with 256 neurons and 16 neurons and the ReLU activation function. On these two layersm, there are a total of 23,073,040 trained parameters.

– 1 final layer with 1 neuron and 'sigmoid' activation function. On this layer, there are a total of 17 trained parameters.

The description of the CNN (written in Python) follows.

```
VALIDATION_SPLIT=0.2 # how many files are reserved for VALIDATION
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=VALIDATION_SPLIT)
print("Training:",X_train.shape, Y_train.shape," Testing:", X_test.shape, Y_test.shape)

Training: (1104, 1025, 44, 1) (1104,)  Testing: (276, 1025, 44, 1) (276,)

N_DENSE = 128
model = Sequential()
# Layer 1
model.add(Conv2D(16, (5,5), padding="same", input_shape= ( N_LINES , NSTD_COLS, 1
),activation='relu') )
model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))
model.add(Dropout(0.25))
# Layer 2
model.add(Conv2D(32, (3,3), padding="same", input_shape= ( N_LINES , NSTD_COLS, 1
),activation='relu') )
model.add(MaxPooling2D(pool_size=(2, 2), strides=(2, 2)))
model.add(Dropout(0.25))

model.add(Flatten())
model.add(Dense(N_DENSE, activation='relu'))
model.add(Dropout(0.25))
model.add(Dense(16, activation='relu'))
model.add(Dropout(0.25))
model.add(Dense(1, activation='sigmoid'))
model.summary()
# Compile model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

The network structure is summarized (in Table 1).

Table 1. Structure of the CNN, after compilation

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 1025, 44, 16)	416
max_pooling2d_1 (MaxPooling2	(None, 512, 22, 16)	0
dropout_1 (Dropout)	(None, 512, 22, 16)	0
conv2d_2 (Conv2D)	(None, 512, 22, 32)	4640
max_pooling2d_2(MaxPooling2)	(None, 256, 11, 32)	0
dropout_2 (Dropout)	(None, 256, 11, 32	0
flatten_1 (Flatten)	(None, 90112)	0
dense_1 (Dense)	(None, 128)	11534464
dropout_3 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 16)	2064
dropout_4 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 1)	17
Total params: 11,541,601 Trainable params: 11,541,601 Non-trainable params: 0		

– The operator presses the CNN auto-training button and the software platform performs the following steps:

1. A neural network is created (all network parameters are set - these are on the order of 10^7).
2. The CNN is tested by random separation from the input spectra of 20% for both categories (N and NM). If the test accuracy is less than 90%, steps 1 and 2 of the training step are automatically resumed. Experiments with Nixon's audio files produced an accuracy of over 95%.
3. Finally, the CNN network will be saved for classifying other audio files of the same speaker.

– Once a valid CNN network (with over 90% value) is obtained, it can be used to make predictions about the content (level) of intention of manipulation in other speeches of the same politician.

7. Experimental results

The first step of the experiments was the creation of the primary database. Several audio recordings that are available on the Internet were used:
– President Nixon's official speech about the Watergate scandal;
– The president's private telephone conversations on the same topic, in which details of the case were more or less openly addressed (these recordings were made public as a decision of the Supreme Court of US);
– Recordings of President Nixon's discussions on other subjects (that have no similar stake) prior the Watergate scandal.

All files were downloaded from public websites (watergate.info and youtube.com [28]).

The secondary database, that was used for training and testing of the CNN, consists of 2 seconds audio-files labeled in two categories (M and NM). The labeling requires the evaluation of discourse analysis experts, that can be done directly by the expert in the working frame of the software platform, or can be introduces by the non-qualified operator based on the prior evaluation. The software platform was used for the generation of the labeled 2 seconds files. Each of these 2 seconds audio files has been normalized (brought to the same volume in decibels and speech pauses).

The resulting database consists of 2758 files. The platform randomly divides the files in training and test subsets. We have used 80% of files for training, respectively 20% of files for tests (2206 training files, 552 test files).

The purpose of CNN networking was to classify manipulative / non-manipulative audio sequences, starting from the spectrogram values of the ready-to-use examples. The training was repeated 30 times and an average accuracy of over 94% was obtained in 15 epochs. The figure below describes how the training error (Loss function) varied according to the number of epochs for which CNN was trained. The training results are also presented in Table 2.

Table 2. Training results report

```
BATCH_SIZE = 15
NB_EPOCH = 15
VERBOSE = 1
model.fit(X_train, Y_train, batch_size=BATCH_SIZE, epochs=NB_EPOCH, verbose=VERBOSE)
≠Final evaluation of the model
print("Testimg:")
scores = model.evaluate(X_test, Y_test, verbose=1)
print("\n%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))

Epoch 1/15
1104/1104 [==============================] - 55s - loss: 6.1950 - acc: 0.6114
Epoch 2/15
1104/1104 [==============================] - 55s - loss: 6.1950 - acc: 0.6114
Epoch 3/15
1104/1104 [==============================] - 56s - loss: 6.1950 - acc: 0.6114
Epoch 4/15
1104/1104 [==============================] - 56s - loss: 6.1950 - acc: 0.6114
Epoch 5/15
1104/1104 [==============================] - 57s - loss: 6.1950 - acc: 0.6114
Epoch 6/15
1104/1104 [==============================] - 58s - loss: 6.1950 - acc: 0.6114
Epoch 7/15
1104/1104 [==============================] - 53s - loss: 6.1950 - acc: 0.6114
Epoch 8/15
1104/1104 [==============================] - 52s - loss: 6.2076 - acc: 0.6105
Epoch 9/15
1104/1104 [==============================] - 52s - loss: 6.1664 - acc: 0.6132
Epoch 10/15
1104/1104 [==============================] - 52s - loss: 6.2203 - acc: 0.6096
Epoch 11/15
1104/1104 [==============================] - 52s - loss: 6.1562 - acc: 0.6105
Epoch 12/15
1104/1104 [==============================] - 56s - loss: 3.4999 - acc: 0.6024
Epoch 13/15
1104/1104 [==============================] - 54s - loss: 0.5920 - acc: 0.6322
Epoch 14/15
1104/1104 [==============================] - 54s - loss: 0.2457 - acc: 0.9149
Epoch 15/15
1104/1104 [==============================] - 55s - loss: 0.1429 - acc: 0.9393
Testimg:
276/276 [==============================] - 3s

acc: 94.57%
```

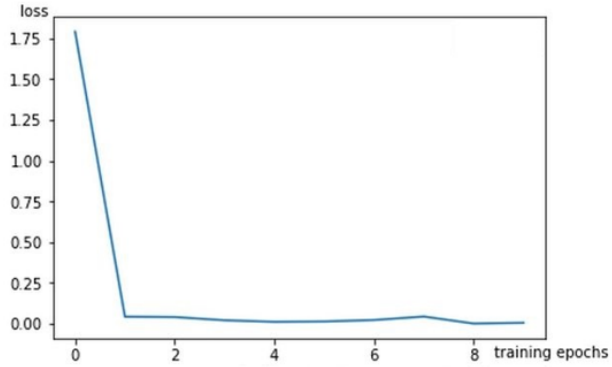

Fig. 5. Training loss variation related to the number of training epochs.

The last phase of the experiments realized so far implied the classification of other recordings of the same speaker (President Nixon), recordings or fragments that were not used in the training and testing sets. We have called this phase 'prediction' (see the operation details in Section 5). The verification of the performances is subject to expert evaluation of the transcript of the files and has not been yet formalized or evaluated as a numerical figure.

8. Conclusions

A new method of identifying deceiving intentions in human voice was presented in this article. This classification is conceptually different than the true/false (or lie detection) approach, since a deceiving intention may be detected even in the absence of lie. The scientific interdisciplinary background of the research confirms the possible imprint of the intentions in the human voice, but offers no details or model referring of how the voice is related to the intentions. The deep learning approach was selected because of the ability of CNN to retrieve information from data in such cases. The method is also based on the use of textual and contextual analysis, using the analytical methodology of critical discourse analysis in order to label the voice samples and requiring expert evaluation of the training and test voice samples. For the integrated application of the two methods (text analysis and deep learning), a software platform was developed. The method was successfully applied in experiments focused on a well-known historical case (US President Richard Nixon and the Watergate affair). The results obtained showed the effectiveness of this new method for detection of inconsistent or deceiving intentions and recommended further research to identify different types of intentions in the human voice and develop a model and taxonomy of intentions.

Acknowledgements. The work reported in this paper was partly supported by the European Project ERANET-FLAG RoboCom++ FLAG-ERA JTC.

References

[1] KRISHNAMURTHY Gangeshwar, MAJUMDER Navonil, PORIA Soujanya, and CAMBRIA Erik, *Deep Learning Approach for Multimodal Deception Detection*, in proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing, March 18 to 24, 2018 Hanoi, Vietnam.

[2] MUCKENHIRN Hannah, MAGIMAI-DOSS Mathew, and MARCEL Sebastien, *End-to-end convolutional neural network-based voice presentation attack detection*, in Proc. of International Joint Conference on Biometrics, Denver, CO, USA, 1–4 October 2017.

[3] FRANTI E., ISPAS I., DRAGOMIR V., DASCALU M., ELTETO Z., STOICA I. C., *Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots*, in ROMJIST **20**(3), pp. 222-–240, 2017.

[4] *Understand User's Intent from Speech and Text*, December 17, 2008, https://www.microsoft.com/en-us/research/project/understand-users-intent-from-speech-and-text/

[5] K. NOGUCHI, P. SOMWONG, T. MATSUBARA, Y. NAKAUCHI, *Human intention detection and activity support system for ubiquitous autonomy*, Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium, 16-20 July 2003, DOI: 10.1109/CIRA.2003.1222300.

[6] D. BORŢUN, *Semiotică. Limbaj Şi Comunicare*, 06 June 2008, Available: https://bibliotecalibera.files.wordpress.com/2008/06/semiotica.pdf.

[7] ICEK Ajzen, *Attitudes, Personality and Behavior*, New York: Open University Press, 2005.

[8] FREUD S., *Parapraxes (1916), Complete Psychological Works of Sigmund Freud*, **15**, p. 66, 1976.

[9] FREUD S., *Psihopatologia vieții cotidiene*, București: Editura Trei, 2006.

[10] OLTEANU G. , VOICU C., PAUN C., PLETEA C. și LAZAR E., *Ascultarea persoanelor în cadrul anchetei judiciare*, AIT Laboratories, 2008.

[11] SCHERER K. R., *What Is the Function of Emotions?, in The Nature of Emotion: Fundamental Questions* (editors P. Eckman and R. J. Davidson), Oxford Univ.Press, 1994

[12] PHILIP R. COHEN, HECTOR J., *Levesque, Intention is choice with commitment, Artificial Intelligence*, **42**(2–3), pp. 213–261, 1990.

[13] LOUIS J. MOSES, BERTRAM F. *Malle, Intentions and intentionality: Foundations of Social Cognitio*n, Cambridge, Mass.: MIT Press, 2001.

[14] MCTAGGART L., T*he Intention Experiment: Using Your Thoughts to Change Your Life and the World*, Published January 1st 2007 by Free Press.

[15] MARCO Perugini, RICHARD P. Bagozzi, *The distinction between desires and intentions*, European Journal of Social Psychology, , **34**, pp. 69–84, 2004.

[16] R. RăşCANU, Psihologie si comunicare, Editura Universitatii din Bucuresti, 2001.

[17] LE BON G., *Psihologia mulțimilor*, ANIMA , 1991.

[18] CICERO, Opere alese, 1973.

[19] CODOBAN A., *Gesturi, Vorbe si Minciuni, Mic tratat de semiotică gestuală extinsă și aplicată*, Cluj Napoca: Eikon, 2014.

[20] FAIRCLOUGH N., *Analysing Discourse: textual analysis for social research*, London: Routledge, 2003.

[21] T. A. VAN DIJK, *Principles of critical discourse analysis*, Discourse & Society, vol. **4**(2), 249–283, 1993.

[22] SILVERMAN D., *Interpretarea datelor calitative. Metode de analiză a comunicării, textului și interacțiunii*, Iași: Polirom, 2001.

[23] Ervin Sam, U.S. Senator, *et. al.*, *Final Report of the Watergate Committee*, https:// archive.org/ stream/FinalReportOfTheSenateSelectCommitteeOnPresidentialCampaignActivities/

[24] Final+Report+of+the+Senate+Select+Committee+on+Presidential+Campaign+Activities_djvu.txt

[25] EKMAN P., *Telling Lies: Clues to Deceit in the Marketplace*, Politics, and Marriage, 2009, W.W. Northon & Company, New York.

[26] Eduard FRANTI, Ioan ISPAS, Monica DASCALU, *Testing the Universal Baby Language Hypothesis - Automatic Infant Speech Recognition with CNNs*, 2018 41st International Conference on Telecommunications and Signal Processing (TSP 2018), Athens, Greece 4-6 July 2018, pp.424–428.

[27] KERAS: The Python Deep Learning library, Available at: https://keras.io.

[28] ANAGNOSTOPOULOS Christos-Nikolaos, THEODOROS Iliou, IOANNIS Giannoukos, Features and classifiers for emotion recognition from speech a survey from 2000 to 2011, 2012, Springer Science+Business Signal Processing (TSP 2018), Athens, Greece 4-6 July 2018, pp.424–428.

[29] The Watergate Scandal: Timeline and Background, https://www.youtube.com/watch?v=IHnmriyXYeg