NATURAL LANGUAGE PROCESSING IN ASSISTIVE TECHNOLOGIES

INGE GAVAT[1], ANDREEA GRIPARIS[1], SVETLANA SEGARCEANU[2]

Abstract. Natural language processing expanded in the last 30 years, arising to cover a wide range of applications in various domains, among them assistive technologies, helping people with disabilities due to accidents, aging, or genetic heritance, to enhance their social integration and their daily life. The growing need in assistive techniques has led to the involvement in the usual devices of intelligent modules to improve and facilitate their use, natural language processing attaining nowadays a high applicability in this domain. The paper discusses first basic methodologies involved in natural language processing like speech recognition, speech synthesis and text processing. Some possible applications are then presented with accent on experimental realization of intelligent modules for assistive devices in the laboratory of our university.

Key words: assistive devices, natural language processing, speech recognition, speech synthesis, intelligent modules.

1. INTRODUCTION

Natural language is the basic communication mode between humans. It has two main aspects, the oral one, called **speech**, and the written one, called **text**. Speech is produced by the phoniatric system and perceived by the auditory system. Text is produced by writing, and is perceived by reading. Speaking, hearing, writing, reading are all human intelligent behaviours, with specific zones allocated in the human brain and together necessary for normal social interactions. All these behaviours are first learned, then accomplished by rules as reflexes, becoming finally the so-called routines, accomplished automatically.

Assistive technologies are special means helping people with disabilities, acquired congenital by birth, by accidents, or by aging, enhancing lost capabilities for a normal existence. Usually, to enjoy life, a person needs to have good senses, (sight, hearing, touch, smell, taste), moving possibilities, speaking abilities and a healthy, active brain to coordinate all activities.

[1] Department of Applied Electronics and Information Engineering, Politehnica University of Bucharest
[2] Beia Research

Ro. J. Techn. Sci. – Appl. Mechanics, Vol. 68, N^{os} 2–3, P. 129–140, Bucharest, 2023

DOI:10.59277/RJTS-AM.2023.2-3.02

Some well-known aids supporting mobility and walking are the wheelchairs, rollators, crutches, paddles, walking sticks. To enhance hearing is developed a large variety of hearing aids more or less sophisticated and for total loss cochlear, middle ear or brain implants become more and more efficient. To correct sight, eye glasses or magnifying glasses are proposed, and for blinds, from a long time the Braille system is a well-known help, enabling reading and writing. Also often used is the white stick, ensuring protection by walk. To improve pronunciation, many forms of logopedic aids and articulatory modelling are helping people to produce a good understandable speech for a friendly and intelligible communication.

In the last years an intensive interest was accorded to support people with disabilities to gain access to the computer and communication techniques like common users. The World Health Organization (WHO) published a list with standard aids, having to be available, as hardware or software [1].

For the users of the Braille system are proposed among others, Braille displays, keyboards and printers, for sight deficient persons are proposed electronic magnifiers. People with reduced hand mobility could benefit from a special mouse acted by leg, mouth, had, or by breathing.

As standard software aids were recommended natural language-based applications like speech recognition or speech synthesis, today incorporated in the personal digital assistants, available especially for English and other widespread languages.

The paper will continue as follows: Section 1 will treat the usual Natural Language Processing (NLP) techniques like speech recognition, text processing and speech synthesis. Section 2 will be dedicated to possible assistive applications and to some assisting modules realized in the laboratory of the University POLITEHNICA of Bucharest. Conclusions and future work proposals with the Literature index will close the paper.

2. NATURAL LANGUAGE PROCESSNG TECHNIQUES

For the purposes of this paper, the NLP techniques to be presented are addressing automatic speech recognition and understanding (ASRU), namely a speech to text application, text processing (TP). speech synthesis, as text to speech (TTS) application.

2. 1. AUTOMATIC SPEECH RECOGNITION AND UNDERSTADING (ASRU)

Automatic speech recognition and understanding [2] is a NLP technique that transforms speech, the voice signal, into text. The processing flow in ASRU systems is represented in the block diagram drawn in Fig.1 below.

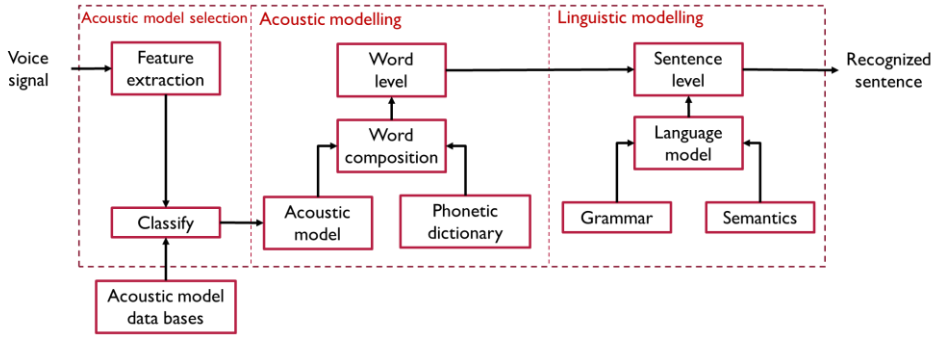

Fig. 1 – The general block diagram of an ASRU system.

The digitally processed voice signal (antialiasing filtered, then sampled and windowed) is subject for feature extraction, a feature vector being calculated for each window. Related vectors characterize related speech parts (phones, triphones, syllables) and for each such part a model is trained off line. The models built for all speech parts are put together into the so-called acoustic model data base, forming a very important knowledge resource for the recognition process.

In the first step of the recognition process, that of the acoustic model selection, the current vector is classified to the model with the best match after a minimum distance, maximum likelihood or maximum probability criterion. The speech part corresponding to this selected model represents one of the constituents for the word to be recognized. This process continues in the second step, of the acoustic modelling, on the word level, until the sequence of speech parts completes a word contained in the next knowledge resource, namely the phonetic dictionary. In the last step, that of the linguistic modelling, the words are ordered to give a correct utterance, based on a language model as knowledge resource. The recognized sentence is finally displayed as text, the recognition process enabling a speech to text conversion.

The components of an ASRU system are detailed further [3].

Acoustic modelling, having as data the features extracted from the acoustic voice signal, is very important in speech recognition. The models are implemented as statistical automata in form of hidden Markov models (HMMs), as connectionist structures in form of artificial neural networks (ANNs), or in hybrid forms, combining HMMs and ANNs. For each speech part (usually named class) of all words contained in the dictionary, a model is trained by machine learning. By training, the parameters of the model are modified until the model fits (learn) the data. The quality of the obtained model can be verified by testing. If the class for that the model was trained is better recognized than other classes of interest (discriminativity), than the model can be accepted as a good model. To obtain such models, huge quantities of spoken material are necessary, registered in standard

formats (wav files), with a large number of speakers, having different ages and educational backgrounds, located in various country sides. The files are organized in data bases, usually called corpora, 60–70% of the material being used for training, the rest of 30–40% for testing.

In Fig. 2 are represented two very simple acoustic model structures, on the one hand a discreet HMM with three states S_i, and five observations o_j (feature vectors) and on the other hand a perceptron with an input, an output and a hidden layer of neurons. The HMM parameters are the transition probabilities from one state in another, a_{i1}, organized in the matrix A, the observation probabilities $b_k\left(o_j\right)$, organized in the matrix B and the initial state probabilities π_i arranged in the matrix $\boldsymbol{\pi}$. The parameters of the perceptron are the weights connecting the neurons from the input layer with them of the hidden layer and the weights connecting the neurons from the hidden layer with them of the output layer. In the training process will be modified according to the data, for the HMM, the matrixes A, B and $\boldsymbol{\pi}$ and for the perceptron all the mentioned weights [4, 5].

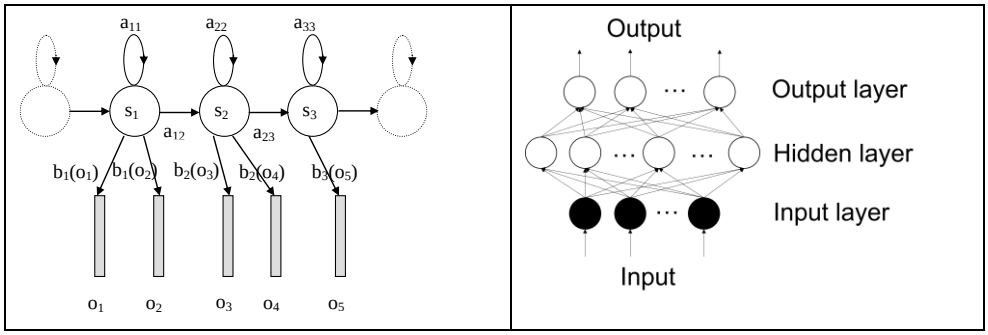

Fig. 2 – Three state HMM and three layered perceptron as acoustic models.

The language model (linguistic modelling) has the role to form from the recognized words a correct and meaningful sentence. The language models can be constructed on two ways. The first way is based on encompassing knowledge as morphology, syntax, semantic, even pragmatic about the natural language, and constructing the **rule based models**, named **grammars**. They have to be elaborated by specialists, especially linguists, by professional text processing. The second way, adopted by nonlinguists, is based on encompassing knowledge about probabilistic relations between the words and sequence of words of a language, to construct the **statistical models** known as **NGram** structures, N representing the the maximal length of the word sequences for those probabilistic dependencies are considered. These probabilistic dependencies are learned processing huge text quantities, extracting estimates of the relative word and word sequences frequencies [6]. The

quality of a statistical model is given by the “perplexity”, that is expressing the possibility of the model to react correctly even in new, stil not learnned situations.

The phonetic dictionary has the role to represent words how they are pronounced, not how they are written. Romanian language is a near phonetic one. In Romanian exist 34 phonemes: vowels, semivowels, consonants, diftongs and triftongs. For the construction of the phonetic dictionary are used standard notations, given by SAMPA (Speech Assesment Methods Phonetic Alphabet), that delivers ASCII 7 bits printable characters of IPA (International Phonetic Alphabet), developped also for Romanian.

Feature extraction is a necessary step in recognition processes in order to obtain a characteristic, essential, reduced pattern of the entities to be recognized for each window of the signal flow. This reduced pattern is in form of a feature vector, having 10–50 components, representing the feature extracted either by **classical methods**, based on fixed algorithms, or by learning from the whole signal, in a many layered reducing process, implemented with ANNs, called **Deep Learning** (DL), and introduced in 2006 by Hinton [8]. Comparing with the classical methods, that reduce the information mechanically, by fixed algorithms, DL, thanks to the layered processing offer the chance to reduce more intelligent and more reliable the information, by discovering and preserving hidden correlations in the signal. This more reliable feature extraction procedure, more complicated that the classical one, became possible only due to the high increase in processing power of the up-to-date computing systems associated with last discoveries in the neurosciences, stating the layered information processing in the brain of mamalians. Widespread today, DL has a considerable contribution to the enhancement of the recognition results [9].

The feature extraction methods applied in speech processing are represented in Fig. 3.

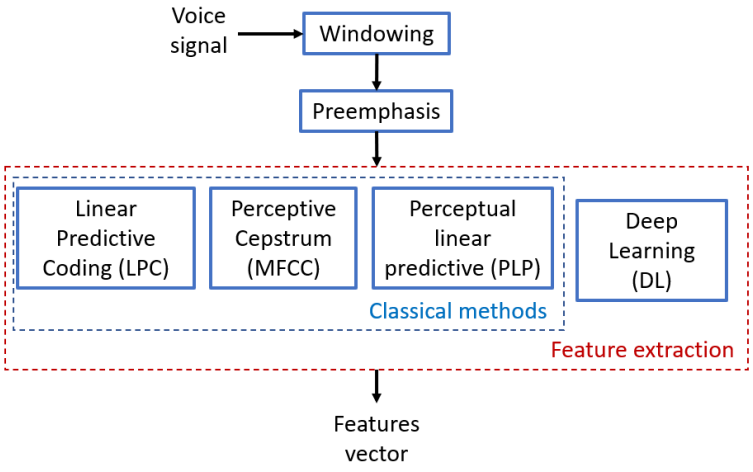

Fig. 3 – Feature extraction methods.

.

The for **classic metods** feature extraction most often used algorithms are further breefly recalled. The first applied algorithm was the **Linear Predictive Coding (LPC)**, furnishing a large variety of features. It was followed by the algorithms for **cepstral and Melcepstral coding**, delivering from the real cepstrum the cepstral coefficients and the very popular perceptually adapted Melfrequency cepstral coefficients (MFCC). An enhanced LPC algorithm, **Perceptual Linear Prediction**, is giving the PLP coefficients, a good and real concurence for the from years applied MFCCs.

The for **Deep Learning** best known structures are **Autoencoders** (AE), **Convolutional Neural Networks** (CNN), **Restricted Boltzmann machines** (RBM), **Deep Believe Nets** (DBN), **Long-short Term Memories** (LSTM). For speech applications are often used CNNs, initially specialized in image processing, now often applied in speech having speech spectrograms as input image.

A speech spectrogram displays the spectrum of an utterance as a diagram with frequency on the Oy axis, time, on the Ox axis, and the magnitude of each spectral component as gray levels or collores code. The spectrogram of the word "paranteze" is represented in Fig. 4, with gray levels, the spectral analysis being performed with law band filters (top of the image), putting in evidence the fundamental frequency in the voiced parts ("*a*", "*e*") of the signal and large band filters (bottom of the image) showing the formants (energy concentrations) for voiced parts, law frequency componets for plosiv sounds ("*p*", "*t*") and high frequency components for fricatives ("*z*"). The linear frequency scale can be replaced by a perceptual scale, proposed by Mel, giving the Melscale spectrogram.

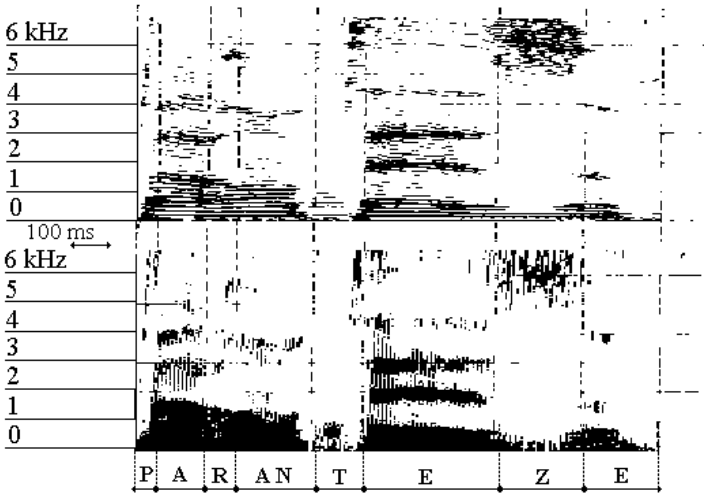

Fig. 4 – Spectrogram of the word "paranteze".

The processing chain to learn speech features with a CNN is represented in Fig. 5.

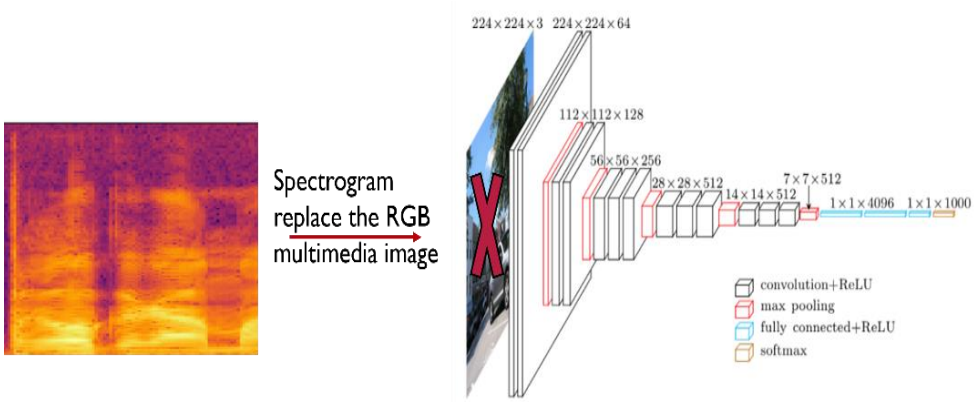

Fig. 5 – CNN applied in speech processing.

The convolutional neural network [10] consists of an input layer, displaying the spectrogram, hidden layers performing convolution (convolutional layers), giving convolution maps, pooling layers reducing the dimensionality, giving reduced feature maps and an output flattened end is performed today giving the finally reduced learned features vector. The activation function of neurons has ReLu (Reduced Linear unit) form for the hidden layers and softmax form for the output layer.

2. 2. TEXT PROCESSING

Text processing is an aspect of NLP with extended applications, supported massively by computers and artificial intelligence programs. In this paper will be presented only text processing steps necessary to build language models and to realize the text to speech synthesis. [11].

A very important, primary procedure is **text analysis**, performed in the first step, by *tokenization*, referring to delimitation of phrases, propositions and parts of speech (POS), namely words. Further is to mark the morphological role of words (nouns, adjectives, verbs, etc), process called POS Tagging and highlight their lemma (root of the word), process called *lemmatisation*. The next step is to detect entity names (proper names, calendar dates, numbers, abbreviations, logos), called *Names Entity Recognition (NER)* and to transform them in clear, readable text. Finally, by *chunking* could be identified nominal groups, as words having sense together.

Alongside, syntactical procedures are also to be applied: *grammatical formalization* to determine the syntactical role of words in propositions, *parsing* to determine the phrase structure, *disambiguation* to determine the real syntactical role of a word.

The necessary text processing steps to build a language model in grammar form are the following: tokenization, POS Tagging, lemmatisation, NER, chunking, grammatical formalization, parsing, disambiguation. To construct statistical models in form of NGrams, the necessary text processing steps are: tokenization, lemmatization, NER, chunking, elimination of the senseless words (prepositions, conjunctions, indefinite articles, indefinite numerals), and then, on this clear text, composed only by meaningful words, statistical methods have to be applied to determine the apparition frequencies for words and for succession of 2 to N words. For both categories of language models, huge quantities of text from literary, journalistic, media kind must be processed.

2. 3. SPEECH SYNTHESIS

Speech synthesis as text to speech (TTS) application, enables the conversion of the text into speech. There are many possibilities to do that, the block diagram of one often applied method is represented in the Fig. 6.

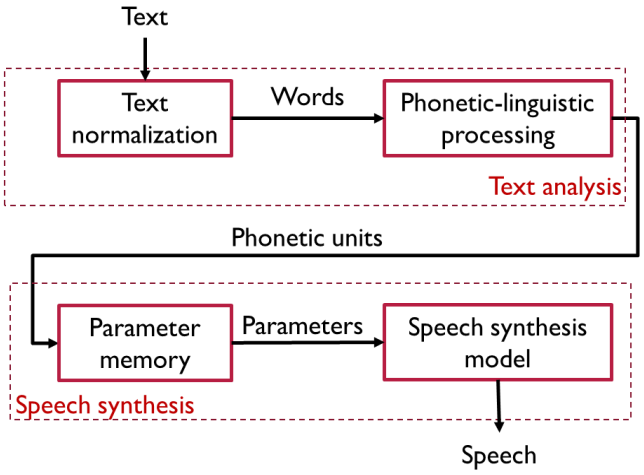

Fig. 6 – Block diagram of a TTS system.

To realize speech synthesis from text [12], first the text analysis must be done how described above in order to obtain a clear text, followed by phonetic transcription, according to the pronunciation and finally the transcription has to be decomposed in speech parts, also called phonetical units (phones, triphones, syllables).

The phonetic units can be described each by features (LPC or PLP coefficients, MFCCs) that extracted from the parameter memory and applied to the speech synthesis model reproduce the corresponding speech part; by their concatenation speech is obtained. There exist also algorithms that can compose speech only putting together speech parts. The speech parts itself and also their features must be extracted from text read by a single person.

3. ASSISTIVE APPLICATIONS BASED ON NATURAL LANGUAGE PROCESSING

In the last years were developed numerous assistive applications based on natural language processing [13], envisaging accessibility support, corrective support or informational support. Some of them will be briefly reviewed below

Accessibility support can be ensured by screen readers, providing audio output for users with visual impairments; by speech dictation software, allowing a text entry alternative to the keyboard, using speech for wheelchair control, enabling handsfree drawing, for persons with reduced hand mobility; speech enabled control of home lighting or door locks, starting and controlling home appliances (TV, radio, wash machine, refrigerator), internet banking or shopping for persons with reduced mobility; text input on smartphones and desktops for persons with hearing problems [14].

Corrective support comprises reading tutors for persons with reading difficulties, logopedic training, and a large number of applications for pronunciation correction or articulatory modelling for persons with speaking difficulties [15].

Informational support consists in giving the possibility to ask questions in an adequate mode, by voice, or by typing and receiving answers by choice on the screen, as text or in a loudspeaker as speech.

Following this trend, two years ago, at the University "Politehnica" of Bucharest, Department for Applied Electronics and Information Engineering was proposed and since then held a lecture for one semester for the master students in "Medical Electronics and Informatics" named "NLP in Assistive Technologies". Some results are already visible in form of master dissertations, from which two will be further presented.

"Speech Synthesis Module for a Text Introduced by the User" is the first one and allows people with speech impairments, who know how to write, the possibility of verbally expressing a wish, a request, or an answer to a question, using concatenative syllable processing. The block diagram is represented in Fig. 7, which reveals the three stages of the module: input, text analysis, and speech generation. In the input stage, the user can either enter a text from the keyboard or use a text file. Once the system as received the text, it will split it into sentences, words, and finally into syllables. Further, the vocal registrations of the detected unique syllables are read from the phonetic syllables dataset and concatenated to create the speech related to the input text.

To increase the word-syllables dictionary, whenever the input text contains a new word, the system uses a web scrapping service to access https://www.silabe.ro/ and find which are the word's syllables.

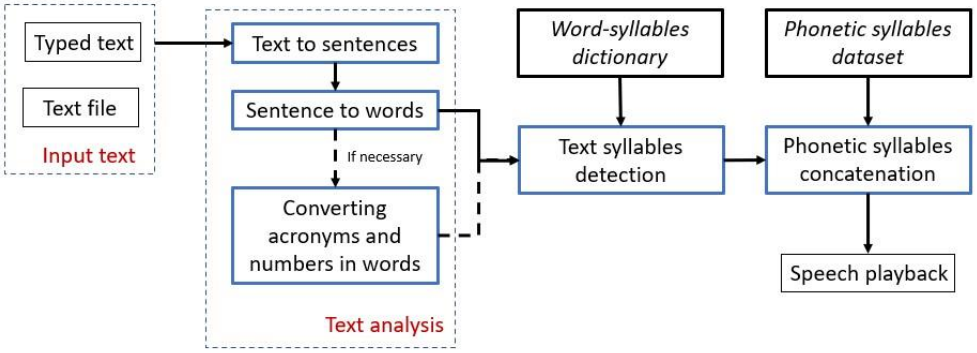

Fig. 7 – Block diagram of the text to speech synthesis module for the text introduced by the user.

"Voice controlled smart mirror for wheelchair" is the second one and enables people in wheelchairs to receive information about weather, media, medical cares by voice recognized with a CNN and a smart mirror device implemented on a Raspberry Pi Board. The block diagram is represented in Fig. 8.

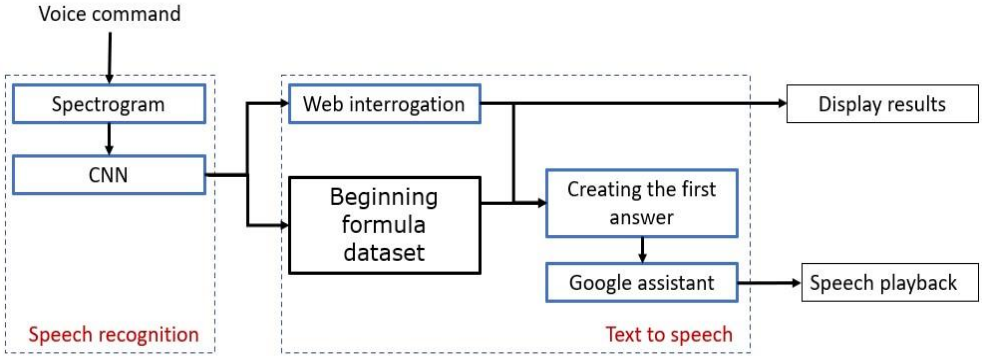

Fig. 8 – Block diagram of the voice controlled smart mirror for wheelchair.

This prototype can recognize a small number of voice commands using their spectrogram, which allows to use a CNN. Since the vocabulary is limited, a CNN architecture having a normalization, two convolutions, a maxpooling, a dropout, a flatten and two fully connected layers, was sufficient to obtain good performances. The confusion matrix of the recognition process is given in Fig. 9 and is showing an acceptable command error rate of 5.5%. After command's recognition, if the query's result is large, the system will display it on screen or say it, if the requested information is short.

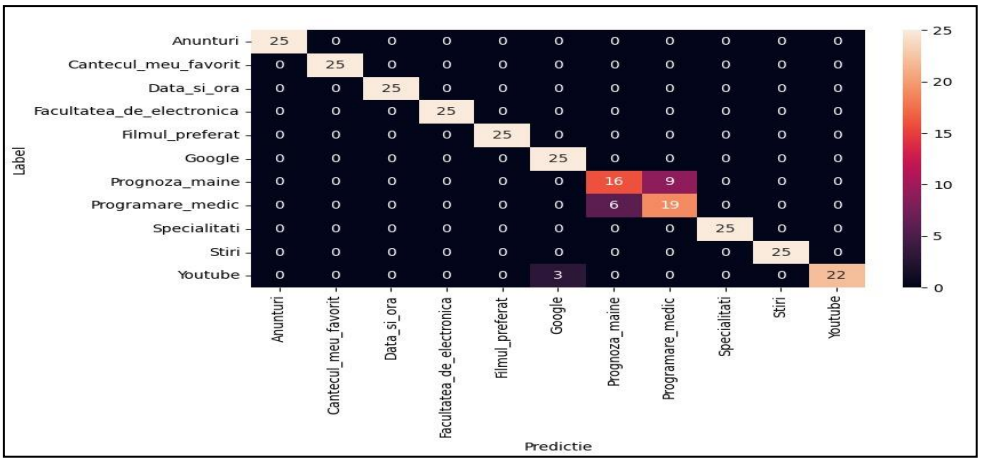

Fig. 9 – Confusion matrix in the recognition task of short commands with a CNN.

4. CONCLUSIONS AND FUTURE WORK

In the last years a special attention was accorded by the WHO to support people with disabilities for a normal life, recommending a large variety of assistive devices, a number of them based on NLP. The paper presents in the first part the most usual building blocks for NLP, namely ASRU, TP, TTS and in the second, some of the assistive systems based on NLP together with two experimental modules realized in the UPB laboratory. In order to transform this module in products, is envisaged to enlarge the collaboration with Beia Research started in the frame of the DAFFCC (Data Analysis for Call Centers) project.

Acknowledgements. This work has been supported in part by UEFISCDI Romania through project DAFCC and funded in part by EU Horizon 2020 research and innovation program under grant agreement No.261 from 01/09/2020.

Received on August, 2023

REFERENCES

1. WHO, *Priority Assistive products list*, available at http://www.who.int/ 2018.
2. DELLER, J. R., PROAKIS, J. G., HANSEN, J. H. L, *Discrete Time Processing of Speech Signals*, McMillan P.C., 1993.
3. HUANG, X., ACERO, A., HON, H. W., *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001
4. JELINEK, F., *Statistical Methods for Speech Recognition*, Cambridge, MA, MIT Press, 1998.

5. WESSEL, F., NEY, H., *Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition*, IEEE Transactions on Speech and Audio Processing, **13**, *1*, pp. 23–31, 2005.
6. VAPNIK, V., *Statistical Learning Theory*, Vol. **3**, Wiley, New York, 1998.
7. GOLD, B., MORGAN, N., *Speech and audio signal processing*, John Wiley and Sons, New York, 2002.
8. HINTON, G. E., *et al., A Fast-Learning Algorithm for Deep Belief Nets*, Neural Computation, vol. **18**, 2006.
9. KAMATH, U., LIU, J. WHITAKER, J., *Deep Learning in NLP and Speech Recognitio*n, Springer, 2019.
10. PATIL, A., RANE, M., *Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition*, Smart Innov. Syst. Technol., **195**, pp. 21–30, 2021.
11. GELBUKH, A., *Computational Linguistics and Intelligent Text Processing*, Springer Science, 2009.
12. TAYLOR, P., *Text to Speech Synthesis*, Cambridge University Press, 2009.
13. JAIN, A., KULKARNI, G., SHAH, V., *Natural Language Processing*, International Journal of Computer Sciences and Engineering, **6**, *1*, pp. 161–167, 2017.
14. PRADHAN, A., MEHTA, K., FINDLATER, L., *Use of voice-controlled intelligent personal assistants by people with disabilities*, International Journal of Computer Sciences and Engineering, January 2018.
15. BASTIDAS, J.P.G., PARRA, O.J.S., MIGUEL, J.E.R., *Language Processing Services in Assistive Technology*, International Journal of Mechanical Engineering and Technology (IJMET), **10**, *07*, pp. 114–128, 2019.