POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 963 **from** 16-11-2022

Ph.D. THESIS SUMMARY

Alexandru-Lucian GEORGESCU

METODE ȘI TEHNOLOGII DE INTELIGENȚĂ ARTIFICIALĂ APLICATE ÎN TEHNOLOGIA VORBIRII

METHODS AND TECHNOLOGIES OF ARTIFICIAL INTELLIGENCE APPLIED IN SPEECH TECHNOLOGY

THESIS COMMITTEE

Prof. Dr. Eng. Gheorghe BREZEANU Politehnica Univ. of Bucharest	President
Prof. Dr. Eng. Corneliu BURILEANU Politehnica Univ. of Bucharest	PhD Supervisor
Prof. Dr. Eng. Daniela TĂRNICERIU "Gh. Asachi" Technical Univ. of Iași	Referee
Prof. Dr. Eng. Corneliu RUSU Technical Univ. of Cluj-Napoca	Referee
Assoc. Dr. Eng. Horia CUCU Politehnica Univ. of Bucharest	Referee

BUCHAREST 2022

Acknowledgements

First of all, I would like to express my special appreciation and deepest gratitude to the coordinator of my Ph.D. thesis, Prof. Dr. Eng. Corneliu Burileanu. I want to thank him for his guidance over time, for the chance to be part of the research team he leads, for the model offered both as an engineer and as a person.

I also thank the guidance committee, Prof. Dr. Eng. Dragoș Burileanu, Assoc. Dr. Eng. Horia Cucu and Dr. Eng. Dan Oneață, for the chance to evolve and learn from their experience and knowledge. I am highly grateful and I have a special appreciation to Assoc. Dr. Eng. Horia Cucu who put a lot of trust in me, even more than I did sometimes. All the time invested in me had a crucial role in my progress, and this thesis is largely due to him.

I would like to thank all the colleagues from the Speech and Dialogue Laboratory for the collaboration and the ideas exchanged.

Many thanks to The Center for Advanced Research on New Materials, Products and Innovative Processes (CAMPUS) within the Politehnica University of Bucharest for the opportunity to work in a professional research environment.

I want to express my gratitude to the members of Xilinx Research Labs from Dublin, Ireland, especially to Michaela Blott, Alessandro Pappalardo and Lucian Petrică for the opportunity to do a great internship that had a major impact on my evolution.

Of course, I want to thank the members of the thesis committee for agreeing to be reviewers and for the time given to the evaluation of this work.

Last but not least, I want to express my deep gratitude to my family, friends, everyone around me who supported and encouraged me all the time.

Table of contents

Chapter 1

Introduction

1.1 Thesis motivation

SpeeD (Speech and Dialogue Research Laboratory) research group [1] within the Faculty of Electronics, Telecommunications and Information Technology of the Politehnica University of Bucharest has a long experience in the field of voice signal processing. An important part of the works mentioned in the previous section, approaching voice signal processing in the Romanian language, belong to the members of this group. Also, within the laboratory, successful diploma and PhD projects were developed over time.

In the 3rd year of my bachelor's degree, I came into contact with the SpeeD laboratory for the first time. I had the opportunity to do a summer internship, followed by the BSc and the MSc projects. Within them, I got good theoretical knowledge of voice signal processing, especially automatic speech recognition, speaker recognition, keyword detection, working with algorithms and specific toolkits. I built simpler projects, such as speaker verification, automatic speech recognition for isolated words, then systems with limited vocabulary, reaching more complex continuous speech systems.

Thus, I had the opportunity to start a PhD program in this field. My role was basically to carry forward the approach that was state-of-the-art at the time, more precisely the LVCSR system in Romanian presented in [8]. There were already enough ideas and approaches that could be tested in an attempt to improve that existing system. Among these, the most important were: the transition to speech modeling based on neural networks, the expansion of the speech corpora used for training, the expansion of the vocabulary and the use of more complex language models. Even if these things were already addressed in English, in Romanian there was enough space to explore them. At the time, there were not many Romanian LVCSR systems with such increased accuracy and robustness as those for English.

A first advantage was the support provided by colleagues and, of course, by supervisors in the research group. I had a competitive environment, where I could develop my knowledge and skills. I benefited from the existing speech processing know-how, the vast experience of the group members. I also benefited from the resources already existing within the group and the systems developed previously.

A second advantage was the opportunity to be employed full-time in national research projects, two of them being the most important for my development in the field of speech technology: SPIA-VA [18] and ReTeRom-TADARAV [3]. Within them, I dealt with speech and speaker recognition, but also with the automatic annotation of audio data.

Thus, this thesis arose in a natural manner, incorporating the author's concerns from the last years and continuing the joint efforts of the group members, being just another piece added on top of the already existing ones.

1.2 Objectives

The main objective of this thesis was to take advantage of the more performant artificial intelligence paradigms in order to obtain improvements in various applications related to speech technology. In the first place, it was desired to improve an already existing Romanian ASR system, created and improved over many years within our research group. This system ended up being completely replaced by a new one using new technologies based on neural networks. A challenge was to adapt the system so that it would be as robust as possible in front of evaluation sets with increased complexity, which contain spontaneous speech in various scenarios and acoustic environments. In addition to the modern technology used, the robustness of the system was given by training it with increasingly large amounts of audio and text data.

In the context presented above, the main objectives of the thesis are:

a) Overview of the state of the art in speech recognition in terms of automatic speech recognition, speaker recognition and automatic annotation of speech corpora.

b) Design and build speaker recognition systems on large Romanian datasets, these systems serving as a baseline for further research.

c) Enhance an already existing automatic annotation procedure using complementary ASR systems. Obtaining new corpora of Romanian annotated speech using the respective procedure, but also experimenting with new procedures.

d) Design and build automatic speech recognition systems for Romanian, using state-of-the-art algorithms and leveraging the training datasets obtained in the previous step.

1.3 Thesis organization

This thesis is organized in seven chapters, as follows:

Chapter 1 introduces the general concepts of machine learning, deep learning, neural networks. Speech processing tasks and their challenges are stated. A brief history of the worldwide speech processing approaches, but also of those regarding the Romanian language, is presented. The chapter continues with thesis motivation, objectives and thesis organization.

Chapter 2 is a state-of-the-art divided into 3 main directions: automatic speech recognition, automatic speaker recognition and the existent speech corpora in the Romanian language. The chapter introduces the principles of these technologies and systems, emphasizing the architectures based on neural networks. Regarding the Romanian corpora, this chapter summarizes information about the characteristics of each existing dataset.

Chapter 3 describes the entire process of creating speech datasets, exemplifying the manner how two such Romanian datasets were obtained. All the necessary steps are presented, starting from the audio recording and gathering, to data validation, data splitting in corresponding subsets, the summarization of the statistics related to the corpus and up to their release in a standard format, useful especially for speech and speaker recognition tasks.

Chapter 4 deals with the creation of automatic speaker recognition systems, trained on Romanian speech datasets. The experiments aim the tasks of speaker verification and speaker identification, in closed-set and open-set scenarios. The systems are based on two different paradigms: GMM-UBM and UBM-ivectors.

Chapter 5 describes the methodology and the activities regarding the task of automatic annotation of speech corpora. We experimented the creation of these corpora by data-filtering based on the multiple hypotheses method, which involves aligning the transcriptions provided by two complementary ASR systems and considering the common parts to be correct. We used various complementary systems and we compared this method with two other data-filtering methods: approximate transcriptions and confidence scoring. The new obtained datasets were further used in Chapter 6.

Chapter 6 represents the most consistent part of the thesis in terms of effort. This chapter is dedicated to improving our Romanian ASR. The approach was divided into 3 major stages, each one being characterized by improvements brought to the level of a certain component of the system. Briefly, the first stage marked the transition from probabilistic acoustic models to convolutional based neural networks in time domain, as well as expanding the vocabulary and creating more complex language models. The second stage evaluated several types of architectures for acoustic modeling and introduced the lattice rescoring technique with language models based on recurrent neural networks. The third stage retrained the ASR system using various combinations of audio datasets, including those obtained in Chapter 5.

Chapter 7 is reserved for conclusions. This chapter presents the results obtained, the author's contributions, the list of papers where these contributions were published, as well as ideas and directions that can be the subject of future research approaches in the field of speech technology.

Chapter 2

State-of-the-art

This chapter includes the fundamental concepts that underlie the main directions of this thesis: automatic speech recognition, automatic speaker recognition and speech corpora for Romanian. A review of the main approaches is provided, starting from the classic approaches and emphasizing the considered state-of-the-art ones nowadays. The latter proved to be superior in terms of performance, thanks to the implementations based on neural networks, as well as the hardware advance, especially of the GPU computing, which allowed running these very intense computational processes.

Section 2.1 is dedicated to the automatic speech recognition and explains the transition from traditional to end-to-end systems, presenting the main approaches to acoustic and linguistic modeling, voice signal parameterization and the most common architectures of these systems, using neural networks. Section 2.2 deals with the topic of automatic speaker recognition. It starts with a short history of these approaches and then both the probabilistic and the neural based approaches are presented. Section 2.3 summarizes all the annotated speech corpora in Romanian. Their characteristics are presented, such as the type of speech, the size of the corpora or info regarding speakers.

2.1 Automatic speech recognition

This section introduces the basic concepts in automatic speech recognition. It presents the road from traditional ASR to end-to-end ASR. Going forward, it describes the most common speech features which are used in current state-of-the-art implementations. We introduce the main principles in traditional ASR and we present the characteristics of different end-to-end approaches. The common language models and their integration techniques in ASR systems are summarized. Finally, we present in detail a number of 8 state-of-the-art ASR implementations, providing architectural characteristics.

2.2 Automatic speaker recognition

2.2.1 Task definition

Biometrics, the science that deals with the statistical study and the measurement techniques applied to living organisms, has become a very intense subject of study over the last few years, especially in the context of data security, a very sensitive and crucial domain. It can be integrated into any security system that involves verifying or identifying users. Faced with a classic system, based on password or access tokens, which can be easily lost or stolen, biometric data has the property of being based on the anatomical features of the subject.

Speaker recognition is one of the most fashionable biometric technologies, developed as an important branch of digital signal processing along with speech recognition. Because speech is one of the most natural forms of human communication and the voice contains a multitude of speaker-specific parameters that can be extracted quite simply from the vocal signal, avoiding direct interaction with the speaker (non-invasive method), speaker recognition can be found in several areas. Speakers differ a lot from their physical characteristics, such as shapes and dimensions of the vocal tract. In addition, a person's speech can be characterized from a behavioral point of view. More precisely, each individual has his own manner of speaking, depending of his accent, rhythm, intonation, pronunciation and more [21]. Depending on the speakers' intentions, they can be cooperative, wanting their identity to be recognized or non-cooperative. The first situation is found in some applications, such as access control, transaction authentication (telephone banking, e-commerce) or device customization (retrieve personal settings based on speaker identity) [26]. In the second situation, the speaker does not want his identity to be determined, this being a common case in forensics or law enforcement.

When talking about speaker recognition, we usually take into account two different tasks: speaker verification and speaker identification. Speaker verification is supposed to verify if the real speaker identity matches his claimed identity. Speaker identification consists in determining the identity of the speaker, without providing any prior information about his possible identity. If his voice certainly comes from one of the system's known speakers, it can be said that this is a closed-set speaker identification task. Otherwise, if the voice of the person to be identified does not come from any of the known users, this is an open-set identification task.

Automatic speaker recognition systems can also be categorized into text-dependent or text-independent systems [20]. The first category includes systems where the speaker has to say a predefined word or sentence. This is especially common in access control systems where, in order to be verified, the speaker has to utter a password. A problem that may arise in this situation is given by the possibility of fooling the system using an audio record with the legitimate speaker while he utters the password. In the second category, text-independent systems, no constraint is imposed on what needs to be pronounced.

These systems are very useful in forensics. However, from the point of view of accuracy, text-dependent systems are more efficient.

2.3 Speech corpora for Romanian language

This section reviews the existing annotated speech corpora for the Romanian language. Details are provided about them, such as the type of speech, the size of the sets, both in hours and as the number of speakers or number of sentences, as well as the availability of the sets: public or private. Datasets belonging to our research group are presented, as well as datasets created by others.

A summary of the most important Romanian speech corpora is presented in Table 2.1. As the table shows, the largest corpora are those created within our research group during the time, presented in [16], [13], [14] and grouped in the much more recent work [15]. There are also a couple of small corpora for which details are given.

2.4 Chapter conclusions

This theoretical chapter provided an overview over the fundamentals of the main directions of this thesis: automatic speech recognition, automatic speaker recognition and summarization of Romanian speech corpora.

From the point of view of automatic speech recognition, we note that pipeline systems have been replaced by end-to-end systems, where signal parameterization, acoustic and phonetic modeling can take place within the same neural network. Such networks take raw audio signal at the input and provide text at the output, without the need for pre-existing alignments between the audio signal and the corresponding transcripts. Additionally, language models implemented with recurrent neural networks can be used, but they serve more as a rescoring step over the initial transcription. Various types of neural architectures have been explored and analyzed in detail, highlighting their characteristics.

Automatic speaker recognition has also benefited from the widespread development of neural networks, replacing the probabilistic approaches that represented the state of the art for long.

From the point of view of the annotated speech corpora in Romanian, corpora that can be used to train speech and speaker recognition systems, it has been considered and may still be considered that Romanian is a language with such limited resources. Although there are some datasets, not all of them are public and their size relatively small. Compared to English, where there are corpora of thousands or tens of thousands of hours, those in Romanian are of the order of tens or hundreds of hours of speaking. It should be mentioned, however, that in recent years there have been efforts to enrich Romanian audio resources.

Table 2.1 Romanian speech resources

Name & ref.	Type of speech	Domain	Size Utt.	Hrs.	Spkrs.	Avail.
RASC [10]	Read	Wikipedia articles	3k	4.8	N/A	public
RO-GRID [19]	Read	General	4.8k	6.6	12	public
IIT [4]	Read	Literature	N/A	0.8	3	non-public
N/A [6]	Read	Eurom-1 adapted translations	4k	10.0	100	non-public
N/A [24]	Spontaneous	Internet, TV shows	N/A	4.0	12	non-public
RSS [29]	Read	News, literature	4k	4.0	1	public
SWARA [28]	Read	Newspapers	19k	21.0	17	public
MaSS [5]	Read	Bible	N/A	23.1	N/A	public
N/A [31]	Spontaneous	Broadcast news	N/A	31.0	N/A	non-public
N/A [30]	Spontaneous	Banking	N/A	40.0	30	non-public
RoDigits [11]	Read	Spoken digits	15k	38	154	public
RSC-train [16]	Read	News, interviews, literature	133k	95	157	public
RSC-eval [16]	Read	News, interviews, literature	2504	5.5	21	public
SSC-train1 [9]	Spontaneous	Radio & TV broadcasts	53k	27	N/A	non-public
SSC-train2 [13]	Spontaneous	Radio & TV broadcasts	170k	103	N/A	non-public
SSC-train3 [12]	Spontaneous	Radio & TV broadcasts	N/A	42	N/A	non-public
SSC-train4 [14]	Spontaneous	Radio & TV broadcasts	277k	250	N/A	non-public
SSC-eval1 [15]	Spontaneous	Radio & TV broadcasts	3035	3.5	N/A	public
SSC-eval2 [15]	Spontaneous	Radio & TV broadcasts	100	1.5	N/A	public
CoBiLiRo [15]	Spontaneous	Excerpts, interviews	50k	31	N/A	non-public
CoRoLa [15]	Spontaneous + read speech	Various sources: radio, studio recordings, news broadcasts, professional speakers	30k	84	N/A	non-public
CDP-train [15]	Spontaneous	Romanian Parliament (Chamber of Deputies)	1.7M	878	2500	non-public
CDP-eval [15]	Spontaneous	Romanian Parliament (Chamber of Deputies	300	5	N/A	public

Chapter 3

Speech datasets collection

This chapter presents the activity regarding collecting, cleaning and organizing two Romanian speech corpora, which were performed during the PhD studies and for which the author had significant contributions.

Automatic speech recognition requires a high amount of data for training the models. For some languages, including Romanian, the biggest problem is still represented by the availability of acoustic and linguistic resources. Such data are not public or simply do not exist for many of the spoken languages. While this problem does not arise for English, being available corpora that contain thousands and tens of thousands hours of speech, Romanian is considered a low-resourced language, struggling with a shortage of resources that can be used in speech technology systems.

Table 3.1 presents the author's contributions to the speech datasets created by SpeeD Research Lab. Each row corresponds to a dataset, while each column designates a specific operation from the dataset creation process. The *acquisition software* column indicates that the author contributed to creation of the software used for corpus acquisition, while the *acquisition* column refers to the acquisition act itself. The *semi-automatic/ manual verification* column refers to filtering out some recordings based on criteria such as completeness of audio file set, file size in bytes, audio duration, word error rate or other indicators. The *automatic annotation* column indicates that the author has annotated the corpus in an automatic way. The *dataset organization* designates the activity of organizing the corpus into train/dev/evaluation subsets, establishing the number of speakers and the number of utterances for each set, while the *dataset statistics* column indicates that the author performed a corpus analysis in terms of duration, number of words, number of characters and some correlations between them. The last column, *packing/ distribution* indicates that the author packed the corpus in a suitable form for distribution, the corpus being publicly released.

As a summary, during the PhD studies we collected two Romanian speech corpora: RoDigits (further described in section 3.1) - a dataset of Romanian connected digits and Read Speech Corpus (abbreviated as RSC, further described in section 3.2) - a dataset of read speech collected in laboratory environment, without background noise.

Table 3.1 The author's contributions to the speech datasets created by SpeeD Research Lab

Dataset	Acquisition software	Acquisition	Semi-automatic/ Manual verification	Automatic annotation	Dataset organization	Dataset statistics	Packing/ distribution
RoDigits	½	½	✓	n/a	✓	✓	✓
RSC	✗	✗	✓	✗	✓	✓	✓
SSC-1	✗	✗	✗	✗	✗	✗	✗
SSC-2	✗	✗	✗	✗	✗	✗	✗
SSC-3	✓	✓	✓	✓	✓	✓	✓
SSC-4	✓	✓	✓	✓	✓	✓	✓
CDP	✗	✗	✗	✗	✗	✗	✗

3.1 RoDigits

In this section we present the efforts made to collect, process and disseminate the dataset that contains spoken digits in Romanian language: RoDigits.

3.2 Read Speech Corpus (RSC)

This section introduces the largest publicly-available Romanian Read Speech Corpus, called RSC. The recordings represent read utterances from literature, news and interviews. This is considered a core speech corpus used for training our speech recognition systems.

3.3 Chapter conclusions

This chapter described all the steps involved in the process of creating and releasing speech datasets. We went from the recording and collecting steps, going through the stages of correction and organization until exposing them in a standard format so that they could be used further in tasks of automatic speech or speaker recognition. We presented these steps on some practical tasks, regarding the creation of two such corpora in Romanian. Additionally, we provided statistics for these datasets, such as duration, number of words, number of sentences, type of speakers or different data distributions.

Chapter 4

Speaker recognition

Speaker verification and speaker identification experiments represent the main topic of this chapter. They were conducted using the RoDigits corpus, a speech dataset which is several times larger than those used in other similar attempts for Romanian language.

The first section of this chapter acts as a baseline, presenting a GMM-UBM speaker recognition system trained on an older version of the RoDigits copus, containing 31 hours of spoken connected digits by over 120 speakers.

The second section of this chapter revolves around the comparison between GMM-UBM system and UBM-ivectors system with PLDA scorer. For the latter system, the last version of RoDigits corpus was used, together with the RSC-train, SSC-train1 and SSC-train2 speech corpora, summing a total of more than 200 hours. In this section, the approach was more elaborate and for an easier evaluation of the both systems, the EER metric was used for the speaker recognition task and Identification Error Rate for the speaker identification task, respectively.

4.1 The GMM-UBM speaker recognition system

This section presents the first text-independent speaker recognition experiments and results using the RoDigits corpus. Experiments in various scenarios involving speaker verification and speaker identification were performed. As with speech recognition, the main parameters were varied to determine the best configuration.

4.2 The UBM-ivector speaker recognition system

The experiments in this section were performed using RoDigits corpus. The corpus was split into an enrollment set and two test sets (one for the open-set scenario and one for the closed-set scenario). The enrollment set comprises 11,120 utterances: 80 utterances from each of the 139 enrollment speakers. The close test set consists of 2,780 utterances: 20 utterances from the same 139 enrollment speakers. The open test set consists of

1489 utterances: around 100 utterances from each of the other 15 speakers. Some other Romanian corpora were used for these experiments. The first one, RSC-train, is composed by 145k read speech utterances from 157 different speakers, summing a total of 100 hours of speech. The second and the third ones contain read speech from news broadcasts and also spontaneous speech, extracted from radio and TV shows, sometimes affected by background noises. They comprise together around 224k utterances (about 130 hours of speech).

4.3 Chapter conclusions

This chapter acts as a baseline, providing speaker verification and speaker identification experiments using Romanian speech corpora. First, we present a simplistic approach, an initial GMM-UBM system, evaluated in terms of false rejection and false acceptance rates. Then, we pass to a more complex approach, performing a direct comparison between the GMM-UBM system and an UBM-ivectors system, evaluating them in terms of equal error rate. A read speech corpus and a spontaneous speech corpus were used to train the universal voice models. Also, a spoken Romanian connected-digits corpus was used for speaker enrollment and for testing purposes. During the experiments, the number of the enrollment files and the number of Gaussian densities were varied. As this number was higher, the performance was better.

For both speaker verification and speaker identification in the closed-set scenario, especially when many enrollment files are used, both systems perform similarly, sometimes the GMM-UBM obtaining better results. Instead, for open-set speaker verification, the UBM-ivectors system is much better.

For the speaker verification task, the EER was computed for both closed and open-set scenario. The best results were obtained for 80 enrollment files, the EER value being equal to 0.17%. In the closed-set scenario, the speaker identification task obtains competitive results, with errors of less than 1%, while in the open-set scenario the error rates were high.

Chapter 5

Automatic annotation of speech corpora

This chapter deals with the procedure regarding automatic annotation of speech corpora. It starts with a theoretical introduction which acts like an overview, presenting the general aspects of the concept, as well as the main directions and approaches found in the literature. Then, a hands-on approach using one of the methods will be described, explaining exactly how it was applied to obtain new Romanian annotated data in an automatic manner.

As previously presented, the annotated audio data resources in Romanian are limited, which makes it difficult to train artificial intelligence systems, such as automatic speech recognition systems, which require very large amounts of data. On the other hand, the online media is a continuous source of speech: radio and television broadcasts or recordings from public institutions, all of them are easily accessible and rich in real-life, but unannotated speech. At the same time, it is obvious that manual annotation, although considered superior in terms of accuracy, is very time and effort consuming and given the context of machine learning, where thousands of audio data are required, this solution becomes impossible or far too expensive. Also, the human factor has its own limitations: from various reasons, a person can make mistakes, or, if we want to transcribe audio data from a rare language, the availability of a speaker of that language it also becomes a problem. In our research group, we have approached over time several methods of annotating and then filtering the annotated data, such as the confidence scoring method, the multiple hypothesis method or the approximate transcription method. However, the main concern of the author in this thesis is represented by the method of multiple hypotheses, which is also the subject of this chapter.

Therefore, section 5.1 represents a theoretical summary of the automatic annotation methods that have been used over time in various research groups. Section 5.2 describes in detail the particularities of the automatic annotation method using multiple hypotheses. Section 5.3 presents automatic annotation experiments using this method on some raw speech corpora, collected from the Romanian media. Section 5.4 presents comparative experiments between the 3 data-filtering methods for automatic annotation: the multiple hypotheses method, the alternative transcriptions method and the confidence scores method.

5.1 Automatic annotation of speech corpora. Approaches and methodologies

Large quantities of speech are readily available across many different languages. However, only a small fraction of speech is transcribed, while the vast majority is unlabeled. Considering that for the Romanian language there are few sets of audio data accompanied by the corresponding transcription, this section deals with the task of leveraging large quantities of unlabeled speech to improve existing speech recognition systems.

Traditionally, learning from both labeled and unlabeled data is known as semi-supervised learning [7] and, arguably, the most common class of semi-supervised methods is *self-training* [32]: train an initial system on the existing labeled data, then use the system to automatically annotate the unlabeled data and use those new samples to re-train the system. This process can then be repeated for multiple iterations.

Crucially, the predictions of the initial ASR system on unlabeled data might be erroneous and might yield incorrect transcriptions. Retraining with wrong transcriptions can affect the performance of the next model, so we need to filter the predictions and select only those parts of the transcriptions that are considered reliable. Several approaches have been proposed in the community, but most of the works address only a single method. Instead, we investigate three classes of methods, which are based on confidence scoring, multiple ASR hypotheses and approximate transcriptions. Next we discuss relevant prior work for each of these methods.

5.2 Multiple hypotheses annotation method. A hands-on approach

The principle scheme of the multiple hypothesis method is presented in Figure 5.1. The unlabeled, raw speech corpus is transcribed using several existing ASR systems. In the case of this work, we used a number of two systems. As the ASR systems are not perfect and the obtained transcripts contain errors, further processing of the initial transcripts is required. Therefore, the transcripts then go through a process of filtering and selection, provided that the ASR systems are complementary, in other words, the systems must be mistaken differently. Based on this assumption, the identical transcribed parts are considered to be correct. The complementarity of ASR systems can be obtained in several ways, such as: different training data, different types of features extracted from the vocal signal, different architectures of acoustic or linguistic models or different decoding algorithms.

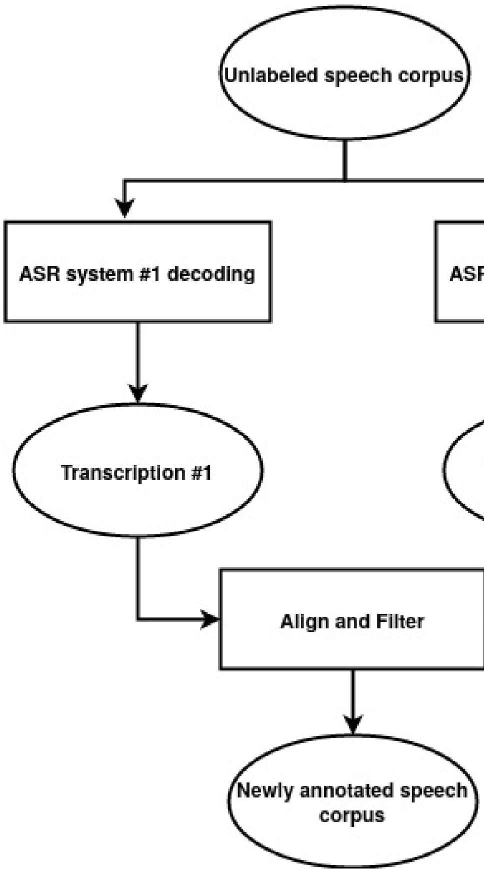

Fig. 5.1 Multiple hypotheses annotation method. Source: [Georgescu, 2018b]

5.3 Automatic annotation experiments for Romanian speech using the multiple hypotheses method

This section presents the practical approach of automatically annotation of speech datasets in Romanian. There are given details of the datasets involved, namely the raw datasets, which are automatically annotated. The initial ASR systems are briefly described, highlighting the elements that distinguish them and make them complementary, to produce different errors. First, their complementarity is evaluated on an evaluation manually annotated dataset and then the same systems are used to annotate the raw datasets. Finally, the results of the annotation procedure are presented and some conclusions are drawn.

5.4 Comparative experiments between the multiple hypotheses method and two other methods

Up to this point, the current chapter has been dedicated to the whole process of automatic annotation of speech datasets using the multiple hypotheses method. We discussed the principles underlying the method, the complementary systems used and the elements by which they differ. We evaluated the method from a qualitative and quantitative point of view and showed that we can obtain data with a correctness of over 95% for more than 50% of the initial, unannotated data. We retrained the initial systems by adding the new automatically obtained data and we were not exactly satisfied with the relative improvements. When we increased the training set by 50%, the improvements were 8% and 12%, and when we more than doubled it, we got relative improvements of 12% and 16%.

We set out to further test whether we could be more efficient using two other methods of automatic annotation, in addition to the method of multiple hypotheses: the approximate transcriptions method and the confidence scores method. The current section compares these three data-filtering approaches in a fair experimental setting and provide answers to questions such as: Which of the filtering approaches is the most useful? Is it beneficial to combine them? Which are the advantages of each approach?

5.5 Chapter conclusions

This chapter was dedicated to the task of automatic annotation of speech corpora. We focused on automatic annotation experiments using the multiple hypotheses data-filtering method. We used several complementary systems, for which we highlighted the differentiators and we tested their complementarity on manually annotated test datasets. In this way, we have shown that the systems, different in some aspects, make different transcription errors. Depending on the test dataset, the accuracy of the selected data is 95% -99% and the amount varies from 20%-30% to 70% compared to the raw corpus. Of course, there is a trade-off between the quantity and quality of automatically annotated data.

Retraining the initial systems by doubling the training data by adding newly annotated data, we demonstrated that the improvements are small compared to the amount of new training data. We performed a detailed analysis of the newly acquired corpus in order to determine whether the automatic annotation method produces unwanted artefacts or not. The conclusion was that the new corpus has similar characteristics with the initial training sets regarding the time duration distribution over the files, the number of words and characters per file and the characters distribution over the whole set. Only after the quantity of newly added data was several times more than the initial data, the improvements in accuracy were significant.

Finally, we evaluated three data-filtering methods in the context of ASR self-training using automatically annotated data: multiple hypotheses, approximate transcriptions and confidence scoring. These methods were used to filter the raw transcripts generated by a seed ASR system for an unlabeled dataset of around 900 hours. All the filtered datasets (confi, 477 h; multi, 555 h; approx, 292 h), turned out to be beneficial in ASR retraining, improving the seed ASR performance by 8.8%, 21.3% and 26.2%, respectively. Although the smallest, the approx dataset brings the most diverse data to the acoustic model training, helping in generalization over degraded speech conditions. On the opposite side, the confi dataset comprises only data that was already confidently transcribed by the ASR seed, bringing little new information. Our empirical evaluation on Romanian speech shows more than 25% relative improvement over the best system so far.

Last but not least, the automatic annotation of audio corpora can be considered a subject that can still be explored more. The trade-off between the correctness and the amount of automatically annotated data still shows room for improvement. Although there are various approaches to filtering the automatically annotated data, an eventual processing of the data that cannot be annotated, the data that are now lost, will represent the key of this task in the future. For sure those difficult data can be the most valuable in training the systems based on machine learning.

Chapter 6

Automatic speech recognition for Romanian language

The current chapter aims to present the evolution of the first large-vocabulary, continuous speech recognition system for Romanian language based on neural networks, this being the main contribution of the author during his PhD studies. This work came as a continuation of the efforts of the SpeeD research group over time.

Improvements to the Romanian ASR system have been gradually brought to different levels and they can be grouped into three major stages, as follows. The first major improvement stage (section 6.1) consists in replacing the ASR CMU Sphinx [22] framework with Kaldi [25], starting to use acoustic models based on neural networks, as well as starting to use the lattice rescoring technique. The second major improvement stage (section 6.2) is dedicated to exploring new neural network architectures for acoustic modeling, as well as introducing language models based on recurrent neural networks. The third and the last major improvement stage (section 6.3) involved training the acoustic and language models using massively extended corpora, as well as some conceptual changes in language modeling.

6.1 The first DNN based approach for Romanian large-vocabulary ASR

This section presents the first major stage in terms of improvements to our large-vocabulary ASR system in Romanian. The content of this section can be considered the beginning, from the point of view of the author's contributions, of a long journey, started in 2017, which aims to update, using state-of-the-art techniques, the baseline Romanian system from 2014. Therefore, the starting point is a system developed in our laboratory prior to this work and described in [9] . What has been used further from the old system to the new system are the training and evaluation data, meaning the voice and text corpora and the phonetic dictionaries. This first stage of improvements has brought several significant changes:

- a new ASR toolkit: from CMU Sphinx to Kaldi
- transition from probabilistic acoustic models, from HMM-GMM framework, to acoustic models based on neural networks: TDNN from Kaldi NNET2 implementation, and then TDNN from Kaldi NNET3 implementation
- new techniques for improving acoustic modeling: e.g: speaker adaptive training - SAT
- additional algorithms for feature processing: we added i-vectors in addition to MFCCs which are standard features
- progress in terms of language modeling: we kept the n-gram probabilistic models, but we extended the vocabulary from 64k words to 200k words and we used 4-gram and 5-gram models, while in the past the maximum order used was 3-gram
- n-gram lattice rescoring technique was used for the first time in the context of Romanian ASR.

6.2 New neural acoustic and linguistic modeling architectures

This section aims to present the second major stage in terms of improvements to the Romanian ASR system. We focused on creating state-of-the-art, TDNN-based acoustic models for Romanian and rescoring ASR results with deep recurrent neural networks (RNNs) using the implementations available in Kaldi NNET3 library, which has long achieved state-of-the-art results on English well-known transcription tasks such as LibriSpeech or TED-LIUM.

Several versions of the TDNN architecture were proposed, implemented and evaluated for acoustic modeling with Kaldi: plain TDNN, CNN-TDNN, TDNN-LSTM and TDNN-LSTM with attention. The various acoustic models were evaluated in conjunction with n-gram and recurrent language models. We report significantly better results over the previous ASR systems for the same Romanian ASR tasks.

6.3 Improved models by using larger speech and language resources. Language model updates

This section presents the third and last major stage in terms of improvements brought to our Romanian ASR system. We collected and used more text and audio data for training the language and acoustic models. Especially in the case of language models, they require periodic updates, because new words constantly appear in the language, such as proper nouns (names of persons or entities), expressions or words that do not exist in the vocabulary of the language model.

Another substantial improvement is represented by finding a solution to the hyphen-related issue in language models. Due to the vocabulary dimension constraint, only the most frequent hyphenated words can be transcribed by the ASR system. Consequently, the ASR system violates the orthographical norms in the case of hyphenated words which are not in the vocabulary of LM. The solution consists in using a more complex hyphen processing procedure in the Natural Language Processing (NLP) application that processes the raw text, before language modeling.

ASR for Romanian language is on an ascending trend of interest for the scientific community. In the last two years several research groups reported valuable results on speech recognition and dialogue tasks for Romanian. In order to facilitate direct comparison with other ASR systems, we provide accuracy results on all the evaluation datasets we have and which we made publicly available, totaling around 15 hours of manually annotated speech. In comparison to our previous best system, we obtained state-of-the-art results for read speech (WER of 1.6%) and significantly better results on spontaneous speech (relative improvement around 40%).

6.4 Chapter conclusions

The improvement of the Romanian ASR system was a long process, carried out gradually over three major stages, which involved varying the systems architecture, varying the training data and a lot of fine-tuning experiments at the level of system parameters. All those improvement stages are illustrated in Figure 6.1, emphasizing the innovative elements at each moment of time, on each of the main development axes of the ASR system: acoustic modeling, linguistic modeling, vocabulary, speech features, speech corpora, text corpora. The figure provides information regarding the progress of the accuracy of the system on read speech, while for spontaneous speech the results differ depending on the nature and difficulty of the task. The detailed description of all these approaches constituted the current chapter.

The first stage involved quite drastic changes compared to the existing system at the time. First of all, we switched the algorithms provided by the CMU Sphinx toolkit to more modern ones, available in the Kaldi toolkit. The use of DNN-based acoustic models instead of HMM-GMM models is one of the most important changes in our ASR system. This change led to obtaining an improvement of 20.7% to 30.8% depending on the type of speech (conversational or read).

The increase of the language model vocabulary size, together with the use of lattice rescoring also triggered new improvements on read speech (27.4% lower WER), but did not bring any improvements on conversational speech. The fact that the WER decreased

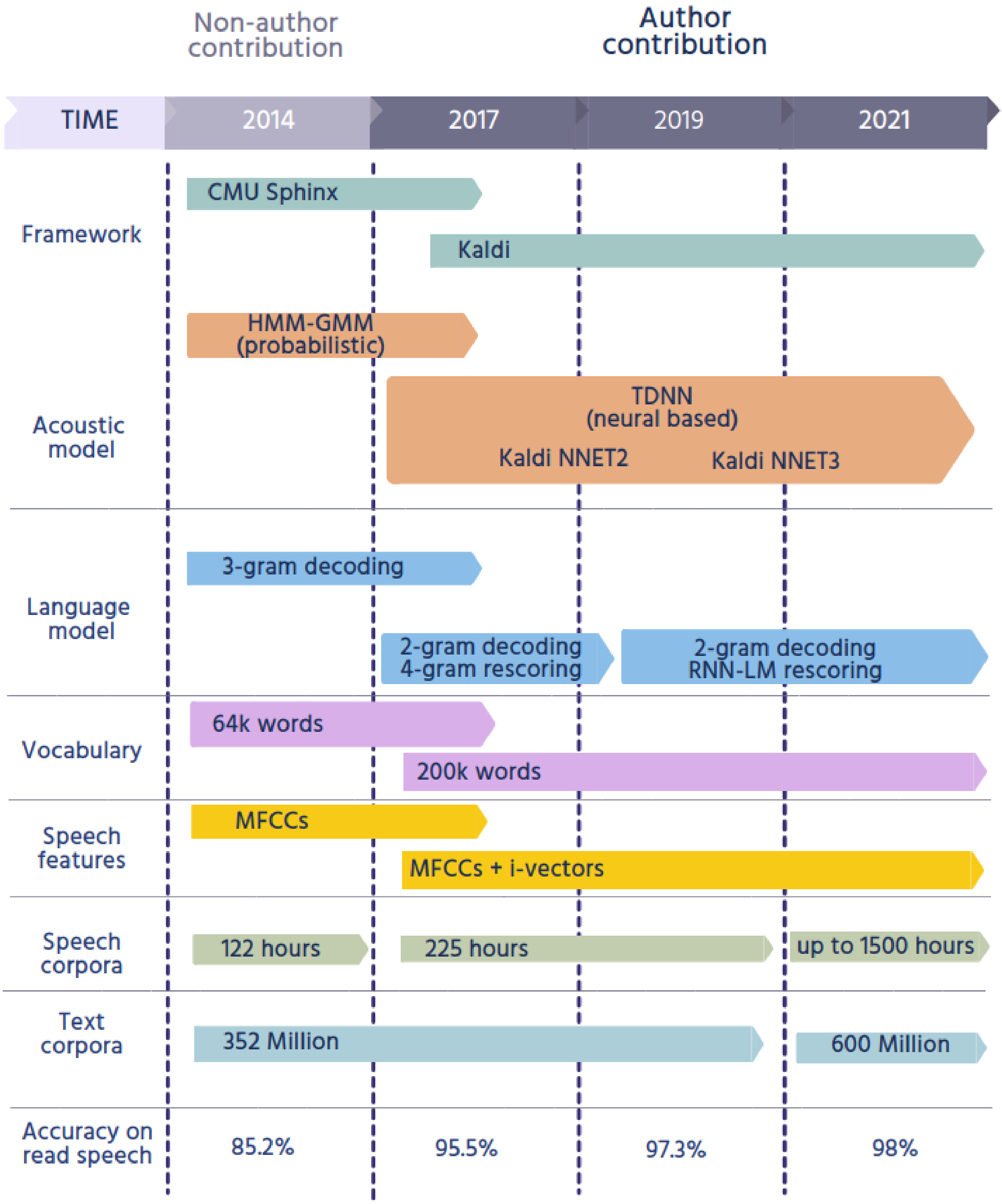

Fig. 6.1 Romanian ASR system evolution. Differences at each component level during those 3 major improvement stages. Comparison with baseline from 2014.

only by 3% on conversational speech when the larger (200k words) language model was used for decoding is very intriguing. The same language model was evaluated very well in terms of OOV rate on conversational speech (83% lower OOV rate). On the linguistic side, we concluded that, given the size and variability of our text corpus, the best results are obtained using a 2-gram LM for ASR decoding and a 4-gram LM for rescoring, both based on a 200k words vocabulary. Overall, we obtained a relative WER improvement of 69.6% on read speech and 48.3% on conversational speech compared to the previous system.

The second stage of major improvements involved testing several adaptations of TDNN for acoustic modeling. We analyzed and used the following architectures: TDNN, CNN-TDNN, TDNN-LSTM and TDNN-LSTM-Attention. For the language modeling, we started to apply lattice rescoring based on recurrent models, in conjunction with n-gram decoding. We compared n-gram rescoring with RNN-LM rescoring, proving that the latter is clearly superior.

It has been shown that the pure TDNN generally obtains better results than the other networks. The n-gram rescoring provides more accurate transcripts over the decoding and that rescoring using an RNN-LM usually performs the best. The overall relative WER improvements of our Romanian ASR system, compared to the system after the first major improvements stage, are 38% on read speech and 17% on spontaneous speech.

The third stage of major improvements brought a significant increase in the speech and text corpora used to train our Romanian ASR system. Thus, we used up to 1500 hours of speech to train the acoustic model, compared to 225 hours in the past, respectively we doubled the number of words used for language modeling, reaching 600 M. We performed various experiments to highlight the contribution of each audio dataset in terms of accuracy. We drew conclusions based on the origin and type of speech corresponding to each set. The accuracy of the state-of-the-art Romanian ASR system is around 98%-99% for read speech, respectively over 90% for spontaneous speech.

Chapter 7

Conclusions

The main objective of this thesis was to use artificial intelligence methods and technologies in order to bring improvements in 3 tasks from the speech technology field: automatic speech recognition, automatic annotation of audio corpora and automatic speaker recognition. The first two tasks have been successfully explored for a long time, while the third task still has plenty of room for exploration. The greatest efforts and the most contributions were made in the areas of speech automatic annotation and training the Romanian ASR systems, these two tasks being interdependent: the evolution of one of the directions also attracted the evolution of the other.

This thesis presented the successive steps taken to train high-performance ASR system for the Romanian language, in parallel with the automatic creation of new speech corpora. To the best of our knowledge, the final ASR system obtains state-of-the-art results.

7.1 Obtained results

Chapter 1 described the main concepts of machine learning, deep learning, neural networks. Speech processing tasks and their challenges were stated. A brief history of the worlwide speech processing approaches, but also of those regarding the Romanian language, was presented. The chapter continued with thesis motivation, objectives and thesis organization.

Chapter 2 presented the state of the art regarding the main directions of the thesis, namely speech and speaker recognition, as well as a summary of the existing Romanian speech datasets. The output of this chapter was a fairly complex overview of the most used systems for these tasks. Very valuable was the detailed low-level analysis of each neural network, describing the role of each component, as well as presenting them through some conceptual diagrams.

Chapter 3 presented a concrete approach of speech dataset collection of two Romanian speech corpora, in order to be used in automatic speech and speaker recognition tasks. Although at first glance it may seem trivial, dataset collection involves much more than the actual recording or procurement of the audio files. The audio went through a semi-automatic validation process, then they were brought to a standard format. The dataset was split in train, development and evaluation subsets. In order to provide a fully useful dataset for machine learning tasks, a detailed analyzes of the datasets have been performed, regarding the number of files, the number of speakers as well as their age and gender distribution, the type of speech, the average number of words per sentence, the average duration of the files. The output of this chapter consists in releasing two Romanian speech datasets, which are also available on the SpeeD's website [2].

Chapter 4 presented two speaker recognition systems, trained on Romanian speech datasets, but based on different paradigms. The systems were directly compared in terms of performance, the output of the chapter being their evaluation results. This chapter can serve as a baseline for Romanian speaker recognition task and invite other researchers to compare their own systems, considering that the used datasets are public, according to the previous chapter.

Chapter 5 presented the automatic annotation based on the multiple hypothesis method, applied on two raw speech corpora, SSC-train3-raw and SSC-train4-raw, comprising 136 hours, respectively 777 hours. The resulting datasets were analyzed, from a qualitative and quantitative point of view, as well as in terms of contribution brought by retraining the ASR systems. We concluded that among the automatic annotation methods studied, the approximate transcription method is the most useful. The output of this chapter was the experiments mentioned above, together with the new training datasets obtained, SSC-train3 and SSC-train4, which contain 42 and 250 hours of annotated speech.

Chapter 6 presented a long effort to improve the Romanian ASR system. The first stage brought the most changes, both at the framework level and at the speech features level, acoustic modeling, language modeling or vocabulary. This first stage was marked especially by the transition from probabilistic systems to neural networks. The second stage involved the exploration of several flavors for acoustic modeling, respectively the use of the lattice rescoring technique based on RNN-LM. The last stage consisted in a significant increase of the audio and text training data. The output of this chapter is represented by the state-of-the-art results of the automatic speech recognition system, with an accuracy of 99% on read speech and around 90%-95% on spontaneous speech, depending on the difficulty of the transcription task. Also, thanks to the fact that the evaluation sets were publicly released [2], the results of this chapter can serve as a baseline for other researchers.

Chapter 7 is reserved for conclusions.

7.2 Original contributions

The personal contributions of the author of this work are found in part in Chapter 2 and Chapter 3, but especially in Chapters 4, 5 and 6. It should be mentioned that these contributions were possible thanks to the members of our research group, their suggestions being fruitful along the way. Also, some methodologies and approaches were developed on top of what already existed within the group. This section summarizes these contributions, indicating the section of the thesis where they appear, as well as the paper number where they were published, according to those in section 7.3, as follows:

a) The systematized exposition of the 8 types of ASR implementations based on neural networks, which specifically consisted in:

   a.1) thorough description of each network, providing details regarding the component blocks and their specific layers and dimensionality

   a.2) detailed graphic representation of each network

   To the best of our knowledge, this is the first attempt to provide such in-depth analysis for these networks. More details are in [Georgescu, 2021c] where we presented the trade-off between ASR performance and hardware requirements of these neural networks.

b) The approach from Chapter 3 of creating and releasing two speech datasets in Romanian, RoDigits and Read Speech Corpus (RSC), in a useful format for automatic speech and speaker recognition tasks. The contributions consisted in:

   b.1) semi-automatic validation of the corpora

   b.2) subsets organization

   b.3) statistics regarding the characteristics of the corpora

   These corpora were also presented in [Georgescu, 2018d] and [Georgescu, 2020]. Both corpora were used in the speaker recognition systems from Chapter 4, as well as in the speech recognition systems from Chapter 6 and Chapter 5.

c) Design and implementation of the speaker recognition systems from Chapter 4, which specifically consisted in:

   c.1) speaker verification and speaker identification tasks in open-set and closed-set scenarios using Romanian datasets

   c.2) fine-tunning of the system parameters

   c.3) comparative experiments between speaker recognition systems based on two different paradigms: GMM-UBM and UBM-iVectors

All these steps were also presented in [Georgescu, 2018d], [Georgescu, 2018a] and [Georgescu, 2018c].

d) Fulfilling the task of the automatic annotation of speech corpora from Chapter 5, which led to obtaining new Romanian annotated audio data of the order of several hundreds of hours:

   d.1) adaptation of the multiple hypotheses data-filtering method using new complementary ASR systems

   d.2) experiments using various complementary ASR systems, evaluating their degree of complementarity

   d.3) design and implementation of the toolkit used for collecting raw audio data from the Internet

   d.4) in-depth analysis of the annotation methodology by inspecting the characteristics of automatically annotated data, which are the subject of the section 5.3

   d.5) comparative experiments between 3 automatic annotation methods, which are the subject of the section 5.4, in order to perform a quantitative and qualitative comparison of the resulting annotated data

   These experiments and their related results were published in [Georgescu, 2018b], [Georgescu, 2019a], [Manolache, 2020a] and [Georgescu, 2021a].

e) Fulfilling the task of the Romanian ASR system improvement, which led to obtaining a new state-of-the-art ASR system within the SpeeD research group. This approach is described in Chapter 6, which represents the author's most consistent contribution:

   e.1) creation of quite large number of ASR systems for Romanian - training and retraining of acoustic models, language models, rescoring models, in different configurations, with different training data

   e.2) extensive fine-tuning experiments of the system parameters

   e.3) evaluation of all the created systems and results interpretation, in order to anticipate the direction of the next series of experiments which can bring more accuracy improvements

   All these successive stages of implementation-improvement-implementation of the Romanian ASR systems were published in [Georgescu, 2017], [Georgescu, 2018b], [Georgescu, 2018d], [Georgescu, 2019a], [Georgescu, 2019b], [Georgescu, 2020], [Georgescu, 2021a], [Georgescu, 2021b]. Also, the tasks that involved training of the new ASR systems from [Manolache, 2020a] and [Manolache, 2020b] are the personal contribution of the author of this thesis.

7.3 List of original publications

7.3.1 Journal papers

[Georgescu, 2018d] **Georgescu A-L.**, Caranica A., Cucu H., Burileanu C., "RoDigits – a Romanian connected-digits speech corpus for automatic speech and speaker recognition," *in University Politehnica of Bucharest Scientific Bulletin*, Series C, vol. 80, issue 3, pp. 45-62, Bucharest, 2018, ISSN: 2286-3540, WOS: 000440896700004, Impact Factor: 0.25, **Q4**.

[Georgescu, 2021c] **Georgescu A-L.**, Pappalardo A., Cucu H., Blott M., "Performance vs. hardware requirements in state-of-the-art automatic speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing 2021*, no. 1 (2021): 28, ISSN: 1687-4722, DOI: 10.1186/s13636-021-00217-4, WOS: 000675403700001, Impact Factor: 2.66, **Q2**.

7.3.2 Conference papers

[Georgescu, 2017] **Georgescu A-L.**, Cucu H., Burileanu C., "SpeeD's DNN Approach to Romanian Speech Recognition," *in the Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, 2017, pp. 1-8, DOI: 10.1109/SPED.2017.7990443, WOS: 000425849600018.

[Georgescu, 2018a] **Georgescu A-L.**, Cucu H., "GMM-UBM modeling for speaker recognition on a Romanian large speech corpora," *in the Proceedings of the 12th Romanian International Conference on Communications (COMM)*, 2018, Bucharest, Romania, pp. 547-551, DOI: 10.1109/ICComm.2018.8453633, WOS: 000449526000104.

[Georgescu, 2018b] **Georgescu A-L.**, Cucu H., "Automatic annotation of speech corpora using complementary GMM and DNN acoustic models," *in the Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP)*, 2018, Athens, Greece, pp. 794-797, DOI: 10.1109/TSP.2018.8441374, WOS: 000454845100178.

[Georgescu, 2018c] **Georgescu A-L.**, Cucu H., Burileanu C., "Comparison of i-vector and GMM-UBM speaker recognition on a Romanian large speech corpus", *in the Proceedings of the 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR-2018)*, Iași, Romania, pp. 25-32, WOS: 000610358400003.

[Georgescu, 2019a] **Georgescu A-L.**, Cucu H., Burileanu C., "Progress on automatic annotation of speech corpora using complementary ASR systems," *in the Proceedings of the 42nd International Conference on Telecommunications and Signal Processing*

(TSP), 2019, Budapest, Hungary, pp. 571-574, DOI: 10.1109/TSP.2019.8769087, WOS: 000493442800124.

[Georgescu, 2019b] **Georgescu A-L.**, Cucu H., Burileanu C., "Kaldi-based DNN architectures for speech recognition in Romanian," *in the Proceedings of the 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Timișoara, Romania, 2019, pp. 1-6, DOI: 10.1109/SPED.2019.8906555, WOS: 000571718700012.

[Oneaţă, 2019] Oneaţă D., **Georgescu A-L.**, Cucu H., Burileanu D., Burileanu C., "Revisiting SincNet: An Evaluation of Feature and Network Hyperparameters for Speaker Recognition," *in the Proceedings of the 28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020, pp. 361-365, DOI: 10.23919/Eusipco47968.2020.9287794, WOS: 000632622300073.

[Manolache, 2020a] Manolache C., **Georgescu A-L.**, Caranica A., Cucu H., "Automatic Annotation of Speech Corpora using Approximate Transcripts," *in the Proceedings of the 43rd International Conference on Telecommunications and Signal Processing (TSP)*, 2020, Milan, Italy, pp. 386-391, DOI: 10.1109/TSP49548.2020.9163405, WOS: 000577106400084.

[Georgescu, 2020] **Georgescu A-L.**, Cucu H., Buzo A., Burileanu C., "RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition", *in the Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France, 2020, pp. 6606-6612, WOS: 000724697202048.

[Manolache, 2020b] Manolache C., **Georgescu A-L.**, Cucu H., Barbu Mititelu V., Burileanu C., "Improved text normalization and language models for SpeeD's Automatic Speech Recognition System", *in the Proceedings of the 13th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, ConsILR 2020, Bucharest, pp. 115-128, Romania, WOS: 000659362800011.

[Georgescu, 2021a] **Georgescu A-L.**, Manolache C., Oneaţă D., Cucu H., Burileanu C., "Data-Filtering Methods for Self-Training of Automatic Speech Recognition Systems.", *in 2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 141-147. IEEE, 2021, DOI: 10.1109/SLT48900.2021.9383577, WOS: 000663633300020.

[Georgescu, 2021b] **Georgescu A-L.**, Cucu H., Burileanu C., "Improvements of SpeeD's Romanian ASR system during ReTeRom project," *in the Proceedings of the 11th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania, 2021, pp. 177-182, DOI: 10.1109/SpeD53181.2021.9587383, WOS: 000786794700032.

7.4 Perspectives for further developments

The author intends to continue the work started in this thesis, regarding various voice signal processing tasks using artificial intelligence.

Thus, from the point of view of automatic speech recognition, although on the English language the author worked with the end-to-end systems studied and presented in [Georgescu, 2021c], on the Romanian language only the hybrid TDNN-HMM networks for acoustic modeling, together with n-gram, respectively RNN-LM for rescoring were used so far. Therefore, there is enough space to explore end-to-end architectures for the Romanian language. This have been only superficially approached by the author until now, within the ReTeRom [Georgescu et al.] research project, where some of the experiments were carried out with DeepSpeech, but the results obtained were not very satisfactory.

Although the transcription of read speech, with clear pronunciations, without noise and good diction, is almost perfect, the spontaneous speech is still a challenge that can be further explored. The accuracy on spontaneous speech from one evaluation set to another is quite different. Therefore, spontaneous speech can present a lot of peculiarities, from background noise, incorrect or incomplete pronunciations, different accents, emotions manifested in speech, all these make the task of transcription very difficult. The only solution for these scenarios is the creation of dedicated acoustic models and language models. Approaches that involve training with polluted data, both natural and artificial, training with data containing speech with accents from various regions of the country, training models with data specific to a certain field of activity could be further explored.

Regarding the automatic speech annotation, we have not yet found a method to process the data that were not kept after the filtering process. Usually, these correspond to difficult areas in the audio signal, areas with strong noise, strong accent or unintelligible speech sometimes. These areas were transcribed incorrectly by the complementary ASR systems and their hypotheses could not be aligned. In this case, the method based on confidence scoring does not work either, because such data are transcribed with a low confidence score, being later rejected at the filtering step. At the moment, the only method that can leverage these data is the method of approximate transcriptions because it involves the alignment between a manual transcription and an automatic transcription, which is mandatory to be correct, even if it has a low confidence score. Using this data to train acoustic models could produce significant improvements, given the different distribution of the data, being different from what the ASR system can already transcribe successfully. This is also the reason why the most useful data for the ASR system retraining were those obtained with the method of approximate transcriptions.

Regarding automatic speaker recognition, this direction has not yet been explored so much by the author, at least compared to the direction of automatic speech recognition. The systems implemented and described in Chapter 4 have more the role of baseline systems, being among the few trained on large amounts of Romanian speech data. The systems used, GMM-UBM and UBM-ivectors, even if they obtain good results at first glance, they are, at least from an architectural point of view, not exactly state-of-the-art. A deeper approach of speaker recognition using neural networks is the one described in [Oneaţă, 20219], where we trained and analyzed an automatic speaker recognition system using SincNet, but the author of this thesis has only a partial contribution in that paper. Therefore, the goal in this direction is to work with more state-of-the-art systems. A good motivation for accomplishing this tasks could be participating in a speaker recognition challenge, such as the VoxCeleb Speaker Recognition Challenge [23] or the NIST Speaker Recognition Challenge [27].

Finally, another hot topic in speech processing using machine learning is that of speech representations. This concept aims to leverage features learned by training a neural network that solves a proxy task, in order to subsequently use these features for the main task, such as automatic speech recognition or speaker recognition. Some types of proxy tasks are: next frame prediction, masked frequency bands/frames prediction, frames order checking. This approach could be of interest in the future, especially for the improvement of the Romanian ASR systems.

References

[1] (2022a). SpeeD's and Dialogue Research Laboratory. https://speed.pub.ro/. [Online; accessed 22-July-2022].

[2] (2022b). SpeeD's speech datasets. https://speed.pub.ro/downloads/speech-datasets/. [Online; accessed 22-July-2022].

[3] (2022). TADARAV. http://tadarav.speed.pub.ro/ro/. [Online; accessed 22-July-2022].

[4] Bibiri, A.-D., Cristea, D., Pistol, L., Scutelnicu, L.-A., and Turculeț, A. (2013). Romanian corpus for speech-to-text alignment. In *Proc. of the 9th International Conference on Linguistic Resources And Tools For Processing The Romanian Language*, pages 151–162.

[5] Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, É. L., and Besacier, L. (2019). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *arXiv preprint arXiv:1907.12895*.

[6] Boldea, M., Munteanu, C., and Doroga, A. (1998). Design, collection and annotation of a romanian speech database. In *Proceedings of the First LREC-Workshop on Speech Database Development for Central and Eastern European Languages*. Citeseer.

[7] Chapelle, O., Schlkopf, B., and Zien, A. (2010). *Semi-Supervised Learning*. The MIT Press, 1st edition.

[8] Cucu, H. (2011). Towards a speaker-independent, large-vocabulary continuous speech recognition system for romanian. *Ph. D. dissertation, PhD Thesis*.

[9] Cucu, H., Buzo, A., Petrică, L., Burileanu, D., and Burileanu, C. (2014). Recent improvements of the speed romanian lvcsr system. In *2014 10th International Conference on Communications (COMM)*, pages 1–4. IEEE.

[10] Dumitrescu, S. D., Boroș, T., and Ion, R. (2014). Crowd-sourced, automatic speech-corpora collection–building the romanian anonymous speech corpus. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, pages 90–94.

[11] Georgescu, A. L., Caranica, A., Cucu, H., and Burileanu, C. (2018). Rodigits-a romanian connected-digits speech corpus for automatic speech and speaker recognition. *University Politehnica of Bucharest Scientific Bulletin (submitted to)*.

[12] Georgescu, A.-L. and Cucu, H. (2018). Automatic annotation of speech corpora using complementary gmm and dnn acoustic models. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–4. IEEE.

[13] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2017). Speed's dnn approach to romanian speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

[14] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2019). Progress on automatic annotation of speech corpora using complementary asr systems. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, pages 571–574. IEEE.

[15] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2021). Improvements of speed's romanian asr system during reterom project. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 177–182. IEEE.

[16] Georgescu, A.-L., Cucu, H., Buzo, A., and Burileanu, C. (2020). Rsc: A romanian read speech corpus for automatic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6606–6612.

[Georgescu et al.] Georgescu, A.-L., Manolache, C., Pop, G., Oneaţă, D., Cucu, H., Burileanu, D., and Burileanu, C. Proiect component tadarav.

[18] Ionescu, B., Ghenescu, M., Răstoceanu, F., Roman, R., and Buric, M. (2020). Artificial intelligence fights crime and terrorism at a new level. *IEEE MultiMedia*, 27(2):55–61.

[19] Kabir, A. and Giurgiu, M. (2011). A romanian corpus for speech perception and automatic speech recognition. In *The 10th International Conference on Signal Processing, Robotics and Automation*, pages 323–327.

[20] Kaur, K. and Jain, N. (2015). Feature extraction and classification for automatic speaker recognition system-a review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(1):1–6.

[21] Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40.

[22] Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., and Wolf, P. (2003). The cmu sphinx-4 speech recognition system. In *Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong*, volume 1, pages 2–5.

[23] Nagrani, A., Chung, J. S., Huh, J., Brown, A., Coto, E., Xie, W., McLaren, M., Reynolds, D. A., and Zisserman, A. (2020). Voxsrc 2020: The second voxceleb speaker recognition challenge. *arXiv preprint arXiv:2012.06867*.

[24] Popescu, V., Petrea, C., Haneş, D., Buzo, A., and Burileanu, C. (2008). Spontaneous speech database for romanian.

[25] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

[26] Reynolds, D. A. (2001). Automatic speaker recognition: Current approaches and future trends. *Speaker Verification: From Research to Reality*, 5:14–15.

[27] Sadjadi, S. O., Greenberg, C., Singer, E., Mason, L., and Reynolds, D. (2022). The 2021 nist speaker recognition evaluation. *arXiv preprint arXiv:2204.10242*.

[28] Stan, A., Dinescu, F., Ţiple, C., Meza, Ş., Orza, B., Chirilă, M., and Giurgiu, M. (2017). The swara speech corpus: A large parallel romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

[29] Stan, A., Yamagishi, J., King, S., and Aylett, M. (2011). The romanian speech synthesis (rss) corpus: Building a high quality hmm-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3):442–450.

[30] Suciu, G., Toma, Ş.-A., and Cheveresan, R. (2017). Towards a continuous speech corpus for banking domain automatic speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

[31] Tarján, B., Mozsolics, T., Balog, A., Halmos, D., Fegyó, T., and Mihajlik, P. (2012). Broadcast news transcription in central-east european languages. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 59–64. IEEE.

[32] Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2):245–284.