RODIGITS - A ROMANIAN CONNECTED-DIGITS SPEECH CORPUS FOR AUTOMATIC SPEECH AND SPEAKER RECOGNITION

Alexandru Lucian GEORGESCU[1], Alexandru CARANICA[2], Horia CUCU[3], Corneliu BURILEANU[4]

This paper introduces a new Romanian speech corpus, called RoDigits. The corpus comprises spoken connected-digits speech data from 154 speakers. It has an approximate duration of 37.5 hours and it is available online, under a Creative Commons license, on SpeeD's laboratory website: https://speed.pub.ro/downloads. *We present all the steps that included the corpus recording, cleaning and its semi-automatic validation. The corpus was used to perform speech recognition experiments using both the HMM-GMM framework and neural networks. Speaker recognition field has also been approached; speaker verification and speaker identification experiments were performed using the GMM-UBM framework.*

Keywords: speech corpora, connected-digits, speech recognition, speaker recognition

1. Introduction

Recent years have brought major advances in audio signal processing, especially in the development of automatic speech recognition and speaker recognition systems. Although methods and algorithms are continually improved, the biggest problem with these systems is the lack of speech databases, especially for languages with limited resources, as is the case with Romanian language. As far as we know, there are only a few annotated speech corpora for Romanian. According to [1], the Romanian language is in the second weakest group of 5 possible in terms of support for text and speech resources. However, some Romanian speech databases exist and part of them are publicly distributed.

[1] Master student, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, e-mail: lucian.georgescu@speed.pub.ro
[2] Researcher, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, e-mail: alexandru.caranica@speed.pub.ro
[3] Associate professor, Faculty og Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, e-mail: horia.cucu@upb.ro
[4] Professor, Faculty of Electronics, Telecommunications and Information Technology, University POLITEHNICA of Bucharest, e-mail: corneliu.burileanu@upb.ro

A summary of the most important Romanian speech corpora is presented in *Table 1*. As the table shows, the largest corpora are the ones presented in [8 – 11] and there are also a couple of small corpora for which details are given.

The acquisition and annotation of one of the first Romanian continuous speech corpora is presented in [5], more than 10 hours being recorded by 100 speakers. This database has a similar structure to that of the EUROM-1 English corpus.

The SWARA corpus [8] contains 21 hours of high-quality speech from 17 speakers and manually annotated at the utterance level and semi-automatic at phoneme level. The main purpose of the corpus is to give persons with speech deficiency or surgical aphonia the possibility to use a synthesized voice as close as possible to theirs.

Some automatic speech recognition systems created using online broadcasted news in several Eastern European languages are presented in [9]. The Romanian acoustic model was trained using a manually annotated 31 hours speech corpus.

A 40 hours corpus of recorded conversations from 30 speakers, based on 25 scenarios related to banking call centers is presented in [10].

In [11] we presented the latest updates on expanding our speech databases used for training and evaluation of our automatic speech recognition systems. RSC is a read speech corpus recorded in a clean noise environment, summing up to about 100 hours, while SSC is a spontaneous speech corpus summing up to about 135 hours, taken from radio and TV shows broadcasts, some of them being affected by background noise.

Table 1

Romanian speech resources

			Size			
Name & ref.	**Type of speech**	**Domain**	**#utt**	**#hrs**	**#spkrs**	**Avail.**
RASC [2]	Read	Wikipedia articles	3k	4.8	N/A	public
RO-GRID [3]	Read	General	4.8k	6.6	12	public
IIT [4]	Read	Literature	N/A	0.8	3	non-public
n/a [5]	Read	Eurom-1 adapted translations	4k	10.0	100	non-public
n/a [6]	Spont.	Internet, TV shows	N/A	4.0	12	non-public
RSS [7]	Read	News, Literature	4k	4.0	1	public
SWARA [8]	Read	Newspapers	19k	21.0	17	public
n/a [9]	Spont.	Broadcast news	n/a	31.0	N/A	non-public
n/a [10]	Spont.	Banking	N/A	40.0	30	non-public
RSC [11]	Read	News, Interviews, Literature	147k	105	157	non-public
SSC [11]	Spont.	Radio and TV broadcasts	227k	135	N/A	non-public

Therefore, as it can be seen from this summary (see *Table 1*), there is not a wide range of Romanian annotated speech resources and only some of them are freely available. Other languages, such as English, have a much larger amount of data. For example, Switchboard [21] is a conversational telephone corpus and it comprises around 300 hours of speech. Librispeech [22], a large-scale freely available read English speech corpus, has a duration of over 1000 hours.

The purpose of this paper is to present a new Romanian speech corpus, called RoDigits (Romanian digits). This is composed of audio clips recorded by multiple speakers, each clip consisting of a sequence of spoken connected-digits. A first unpublished version of this database was used in [12] to create a speech recognition system for connected-digits.

Spoken digits are very common in speech recognition systems. These are the basis for many applications in the telephony industry, such as phone dialing, interactive voice response systems or data entry. One of the most popular databases for such a task is TIDigits [13], an English corpus collected by Texas Instruments to design and evaluate algorithms for speaker-independent digits sequences recognition. The recordings were made by 326 speakers, belonging to both genres and in a wide age range, each speaking approximately 77 digit sequences. Also speaker recognition systems are based on spoken digits. RSR2015 (Part III) [14] is an English corpus specifically designed to train and test automatic text-dependent speaker verification systems. This corpus was collected by Institute for Infocomm Research (I2R) and it consists in 35 hours audio data from 300 speakers recorded using portable devices.

Our corpus presented in this paper, RoDigits, has a similar magnitude to the other two corpora quoted above. The number of speakers is smaller, but comparable in size order: 154 speakers. The total duration is similar to that of the RSR corpus, approximately 38 hours.

This paper is organized as follows. Section 2 provides details about the development stages for the RoDigits corpus and its final form. Sections 3 and 4 present baseline speech and speaker recognition experiments performed solely on the newly created corpus. Section 5 is reserved for the conclusions and summarization of the work.

2. RoDigits Corpus Development

This section describes all the steps taken in creating the corpus. Details about the recording stage, cleaning and validating the corpus and characteristics about the final corpus are given.

2.1. Recording the corpus

The RoDigits corpus was collected by the SpeeD (Speech & Dialogue Research Laboratory) group from University Politehnica of Bucharest. The

recordings were made over a fairly long time interval, between March 2012 and April 2017. Initially, the database contained records from 173 speakers, 100 male speakers and 73 female speakers. Then, following the validation process described in Section 2.2, the recordings of some speakers were partially or completely removed. Also, speakers who did not record the complete set of 100 utterances were ignored for the moment and their files are not part of the final corpus. The speakers' ages vary between 20 and 45 years old, with an average around the age of 23, most of them being students of Faculty of Electronics, Telecommunications and Information Technology of the above mentioned university. Their native language is Romanian, except for one person (an Albanian native). The recording environment varied: some speakers made the recordings in the lab using workstations and Sennheiser head-mounted microphones, while others recorded their speech at home using various acquisition hardware. All files have some common features. For example, all were resampled at 16 KHz, 1 channel, with a precision of 16 bits per sample. All of them are encoded as 16-bit Signed Integer PCM. The average length for a recording session is about an hour. It is worth to mention that some speakers have interrupted the session and continued it later, at another time of the day or even another day. Because of this there may be differences even in the recordings of a single speaker, these being induced by the different recording environment or by the speaker's mood at the various recording times.

The recordings were made through a web application developed by SpeeD group, which is available online and can be accessed via a laptop or PC. After the application is opened, the user must allow it to capture the signal from the microphone. Then it follows a two-step microphone calibration step. The first step consists in recording the background noise. During this time the user must not speak and there should not be any other sources of audio in the background. Step two involves the recording of a test utterance. Based on these two recordings, the signal-to-noise ratio is computed. By comparing it with a predefined threshold, the speaker recordings are accepted or rejected. If they are rejected, the speaker receives recommendations regarding speech loudness or the speech pauses required at the beginning and end of the utterance.

Each speaker had to record a group of 100 audio clips, each clip consisting of a randomly generated sequence of 12 Romanian digits. The application displays the digits sequence on the screen and by pressing a button in the graphical interface, the audio capture switches on and the speaker can start to utter. When the utterance is over, the stop button is pressed. If the recording is accepted, the following sequence of digits will be displayed. The user has the opportunity to listen to the recordings already made and to repeat the recording process for those considered erroneous. Errors can occur for multiple reasons: mispronunciations, hesitations, stuttering, etc. Moreover, there can be

environment-induced errors such as reverberations or the occurrence of an unexpected background noise. Users were advised to speak as natural and clear as possible in quiet conditions.

2.2. Semi-automatic validation of the corpus

Validation was a necessary step in making the final corpus because, as stated in Section 2.1, even if the speakers received instructions on how to record the audio clips, not all the recordings were correct. Manually validating all the recordings (almost 40 hours of speech) by comparing the utterances with the transcriptions would have been a tedious task. Consequently, we decided to approach it semi-automatically by excluding incorrect files, based on several criteria, as follows: completeness of audio file set, file size in bytes, audio duration, WER. The exclusion process was automated as much as possible.

First, the speakers were automatically sorted by the number of recorded audio clips. The goal was to keep, as far as possible, in the final corpus only complete sets of 100 recordings per speaker. The audio files of the speakers who recorded less than 50 clips were excluded. 155 audio files from 15 speakers were excluded in this step.

Next, the audio files were sorted in ascending order by file size in bytes. Zero-sized audio clips were directly removed. A human operator listened to the non-zero-sized audio clips to empirically determine a minimum size for valid clips. Using this procedure we discovered a speaker who had only zero-sized recordings. Other seven audio clips from four different speakers were identified as being incomplete (containing the pronunciation of only one, two or three digits).

The audio clips were also sorted in ascending order by their time duration. A human operator listened to the shortest audio files to empirically determine a minimum duration for valid audio clips (clips comprising the complete digit sequence). Using this procedure, another four audio clips were identified as incomplete and excluded from the corpus.

Finally, another method of validation involved decoding all recordings using an automatic speech recognition (ASR) system and comparing the resulted transcription with the presumed one. Specifically, the following steps were taken:

- the recordings were automatically transcribed;
- the resulted transcripts were aligned with the reference transcripts;
- the number of correctly transcribed words (NCW) was computed for each recording;
- the recordings were sorted by NCW in ascending order,
- a human operator listened to all the recordings for which NCW < 7 (the ASR transcribed correctly less than 7 words out of 12 possible).

The acoustic model was a general one, trained on a corpus containing both read and spontaneous speech on various topics. More details about it can be found

in [11]. The first used language model was a generic n-gram language model created on the basis of very large text corpus. However, the experiments showed that this induces quite a lot of errors and it is not indicated for an audio corpus validation operation. To reduce the probability of errors, a rule-based grammar model, described in Section 3.1, was chosen to restrict the output of the ASR system to a sequence of digits.

By following the above procedure, several incomplete audio clips were discovered and excluded from the corpus. Moreover, using this procedure, complete, but noisy audio clips were discovered. The exclusion decision in these cases was based on the perception of the human operator who listened to the recordings: only recordings in which the words could be distinguished and identified were kept.

2.3. Splitting the corpus

After the exclusion of several recordings, during the validation procedure discussed above, the final corpus comprised 15,389 recordings (99 recordings from 11 speakers + 100 recordings from 143 speakers).

The final corpus was divided into training set, development set and evaluation set. The training and development sets contain recordings from the same speakers, representing 90% of all speakers, and their choice was a random process. In order to optimize future work with these sets, it was taken into account that the speakers in these two sets must have complete records, 100 files each one. The training set contains 80 files from each speaker, those with IDs between 1-50 and 71-100. The development set contain the remaining 20 files out of the total of 100 files, those with IDs between 51 and 70. The evaluation set contains the entire recordings set from the remaining 10% of the speakers, each with 99 or 100 audio recordings, respectively.

2.4. Characteristics of the final corpus

The final corpus has a size of 3.15 GB and contains a total of 15,389 audio files from 154 speakers, 86 male and 68 females. The total duration of the corpus is about 37.5 hours. As presented in section 2.3, the corpus consists of 3 sets: training, development and evaluation.

The train set consists of 11,120 audio files from 139 speakers, 76 male and 63 females. The total duration of this set is about 27 hours. The development set consists of 2,780 files from the same group of speakers as in the case of the training set. More than 6 hours of speech make up this set. The evaluation set includes 1,489 files from 15 speakers, 10 male and 5 female. The total duration of the set is over 3 hours. *Table 2* summarizes the characteristics of the entire corpus, as well as of each set.

Table 2

Corpus characteristics

	Set			
Characteristic	**Train**	**Dev**	**Eval**	**Total**
# of speakers	139	139	15	154
Male speakers	76	76	10	86
Female speakers	63	63	5	68
# of files	11120	2780	1489	15389
Total duration	27 h 2 m 18 s	6 h 47 m 24 s	3 h 43 m 24 s	37 h 33 m 07 s
Mean duration / file	8.75 s	8.79 s	9.00 s	8.78 s

2.5. Availability of the corpus

The corpus is available online and can be downloaded under the Creative Commons BY-NC-ND 3.0 license from SpeeD's laboratory website [15]. The archive contains both audio files (grouped into folders named after the speaker ID) and their transcriptions. Moreover, it includes several metadata files: list of training files, list of development files, list of evaluation files, list of all speakers IDs and a phonetic dictionary containing phonetic transcripts for Romanian digits.

3. RoDigits Speech Recognition System

As stated before, RoDigits speech corpus can be used to evaluate automatic speech recognition (ASR) and speaker recognition systems. This section presents a basic ASR system for connected-digits, trained and evaluated solely on RoDigits. General information about the phonetic, language and acoustic models are provided in section 3.1, while section 3.2 presents the various experiments performed for finding the optimal ASR main parameters and results related to system performance evaluation.

3.1. Phonetic dictionary and grammar

The phonetic dictionary is a component of an ASR system. It contains all the words that can be transcribed by the system, accompanied by their phonetic transcription. It links the other two ASR components: the acoustic model, which estimates the probabilities of phonemes occurrence, and the language model, which models how likely word sequences are.

In this case, the phonetic dictionary consists of 10 lines, corresponding to the words in Romanian language that designate the digits from 0 to 9.

As mentioned above, the language model has the role of estimating how likely a group of words is, to establish the sequence of words in a sentence. Generally, two types of language models are used: probabilistic models and rule-based models.

In our case, we use a rule-based (grammar) language model, because the vocabulary is small, and we can define the exact constraints regarding the allowed

sequences of words. The grammar was defined as a word-loop model: any word may follow any other and the allowed sequences of words may have any length.

3.2. Kaldi HMM-GMM and TDNN acoustic models

Kaldi [16] is currently one of the most popular open-source toolkits. It provides support for automatic speech recognition and also speaker recognition. Kaldi contains state-of-the-art algorithms focused on voice signal parameterization, HMM-GMM based acoustic models, neural network acoustic models, offline or real-time voice signal decoding.

The first stage in such a system consists in features extraction from the vocal signal. Kaldi uses MFCC (Mel Frequency Cepstral Coefficients) or PLP (Perceptual Linear Prediction) coefficients, on which several types of transforms are applied, such as CMVN (Cepstral Mean and Variance Normalization), LDA (Linear Discriminant Analysis), MLLT (Maximum Likelihood Linear Transform) and more.

Acoustic modeling using the HMM-GMM framework implies training iteratively several models, each one based on phone forced alignments obtained from the previous model. First, a context-independent model for phones (monophones) is created. Next, based on the forced alignments for phones obtained using this model, a more complex, context-dependent, model for phones is trained. Going further, each training iteration aims at training a more complex model, by applying various speaker-independent or speaker-adaptive transforms to the features or using more sophisticated training algorithms.

Using the training set of our corpus, we trained HMM-GMM acoustic models, varying some of the most important parameters (the number of HMM states referred as leaves and the total number of Gaussian densities), in order to determine the best combination of these values. The decoding step was done using both the development set and the evaluation set. *Table 3* and *Table 4* show the results obtained using the TRI3 MMI acoustic model, which models triphones and applies the Maximum Mutual Information (MMI) training technique.

The HMM-GMM models were used as a starting point for training HMM-DNN acoustic models based on time delay neural networks (TDNN) [17]. This type of neural networks model temporal dependencies between acoustic events. A sub-sampling technique is applied, which assumes that activations for neighbor frames are correlated, then they are not considered for consecutive frames, but for spliced frames. This fact brings the advantage that the execution time is comparable to feed-forward networks.

Table 3

Kaldi GMM-based ASR results on RoDigits Dev Set

WER[%]	# GMs		
# leaves	**20k**	**50k**	**100k**
50	0.39	0.42	0.49
75	0.31	0.33	0.31
100	0.31	0.30	0.31
125	**0.28**	0.30	0.36
150	0.31	0.28	0.34

Table 4

Kaldi GMM-based ASR results on RoDigits Eval Set

WER[%]	# GMs		
# leaves	**20k**	**50k**	**100k**
50	1.56	1.90	2.15
75	1.23	1.32	1.97
100	1.11	1.27	1.42
125	**0.85**	0.91	1.36
150	1.26	1.00	1.37

The results tables show that the WERs reported for the models with 20k or 50k Gaussian densities and between 75 and 150 leaves, are very close. The best results are obtained for 20k Gaussian densities and 125 leaves. The models which have 100k Gaussian densities exhibit a significantly higher WER. Nevertheless, on the evaluation set (which contains unknown speakers), the error rate is significantly higher than on the development set (which contains known speakers). Because the best results were obtained for 125 leaves and 20k Gaussian densities, we performed further experiments keeping these values as constants.

Table 5 presents average WER and standard deviation for both development set and evaluation set, as a comparison between HMM-GMM and TDNN acoustic models.

Table 5

Comparison between Kaldi GMM-based and DNN-based ASRs on RoDigits Dev Set and Eval Set

	WER[%] on Dev Set		WER[%] on Eval Set	
Acoustic Model Type	**average**	**std**	**average**	**std**
Tri 1	0.70	1.10	3.30	4.30
Tri 2	0.60	1.00	3.00	3.60
Tri 3 [LDA + MLLT + SAT]	0.50	0.80	1.70	2.20
Tri 3 [MMI]	0.30	0.40	0.90	0.90
TDNN	0.30	0.50	1.20	1.10

The acoustic models based on neural networks were saved after each epoch. Following the experiments, the results are somewhat similar using models obtained after the first 3 training epochs. The best model is the one obtained after 4 epochs. The number of hidden layers of the network has also been varied. At the beginning, the tests were performed using 3 hidden layers, but the presented results are obtained with a 6 hidden layer network.

The standard deviation has very low values on development set, this information providing the certainty that the error is below 1% for almost all speakers. Not the same thing can be said for the evaluation set. The higher standard deviation values correspond to the fact that there are speakers for which

the system transcribes worse. *Table 6* presents the number of speakers for each WER interval using the best system, TRI3 MMI.

Table 6

Number of speakers vs. WER values for TRI3 MMI

	# of speakers	
WER values	**Dev Set**	**Eval Set**
WER = 0%	87	3
0% < WER < 1%	41	6
1% < WER < 2%	10	4
2% < WER	1	2

The average sentence error rate (SER) for the best system is 2.90% on the development set and 8.30 on the evaluation set.

In conclusion, for a simple task with a small 27-hour training database, the use of neural networks is not justified because the results are not improved compared to those obtained using the HMM-GMM framework.

3.3. CMU Sphinx HMM-GMM acoustic models

The CMU Sphinx Toolkit [19] is used to implement a HMM-GMM based ASR architecture, for the initial validation of the corpus. CMU Sphinx, also called Sphinx in short, is the general term to describe a group of speech recognition systems developed at Carnegie Mellon University. These include a series of speech recognizers (Sphinx 2 - 4) and an acoustic model trainer (SphinxTrain). The code is available open source for download and use [20]. The libraries and sample code can be used for both research and commercial purposes.

Acoustic models used in this paper are 5-state HMMs, with output probabilities modeled with GMMs. The traditional MFCC features were used as baseline audio features, plus temporal derivatives (13 MFCC + Δ + $\Delta\Delta$). The number of Gaussian mixtures (GMs) per senone state were varied, to adapt the acoustic model setup to the size and variability of the training speech database. No audio enhancements or noise reduction algorithms were used. Phonemes were modelled in a context-dependent manner. To study the effects of increasing or decreasing the number of senones and Gaussian mixtures per senone, they were varied, according to Tables *Table **7***, *Table **8***, *Table **9*** and *Table **10***. These values where chosen in correspondence with Tables *Table **3*** and *Table **4***, so we can directly make a comparison between the performance of both CMU Sphinx and Kaldi acoustic models, but also to adapt to the size of the corpus.

Final results were compared in terms of WER (Tables *Table **7*** and *Table **9***) and SER (Tables *Table **8*** and *Table **10***). The effects of increasing and decreasing the number of tied-states (senones) and number of Gaussians per senone, for each setup, can also be observed in the corresponding tables.

For the RoDigits corpus, with the current setup, it seems that 512 GMMs with 125 senone states offer the best results for the development dataset, which includes speakers used for training also. With unseen speakers, in the evaluation dataset, 100 senones with 256 GMMs offer the best results, for both SER and WER metrics.

These results can be explained by the fact that the more senones a model has, the more precisely it discriminates among sounds. On the other hand, if a high number of senones is set (more than necessary), the model might not be universal enough to recognize unseen speech. WER will be higher on new data, so it is important also not over-train the models, given the current size of the digits corpus, and the limited numbers of trained phonemes (only 19 independent phonemes where trained, contained in the digits dictionary for CMU Sphinx). In general, the little the mismatch (be it speaker, environment, encoding, etc.) between the training and the evaluation data, the better the results.

Table 7

CMU Sphinx Dev Set results (WER)

WER [%]	**# GMs / senone**				
senones	**32**	**64**	**128**	**256**	**512**
50	4.5	3.8	3.3	2.9	2.7
75	2.8	2.5	2.3	2.0	1.9
100	2.4	2.0	1.8	1.6	**1.5**
125	2.6	2.3	1.9	1.7	**1.5**
150	2.5	2.2	2.1	2.0	2.2

Table 8

CMU Sphinx Dev Set results (SER)

SER [%]	**# GMs / senone**				
senones	**32**	**64**	**128**	**256**	**512**
50	37.5	32.9	30.1	27.1	25.9
75	25.8	24.5	22.5	20.0	18.6
100	22.5	20.4	18.7	16.8	16.2
125	23.5	21.9	19.0	17.1	15.7
150	23.2	21.5	20.2	20.0	21.1

Table 9

CMU Sphinx Eval Set results (WER)

WER [%]	**# GMs / senone**				
senones	**32**	**64**	**128**	**256**	**512**
50	6.8	6.2	6.1	6.0	6.2
75	5.1	4.9	4.9	5.2	5.5
100	4.2	3.9	**3.7**	**3.7**	3.9
125	4.1	4.4	4.0	3.9	3.9
150	4.9	4.8	4.8	5.2	5.5

Table 10

CMU Sphinx Eval Set results (SER)

SER [%]	**# GMs / senone**				
senones	**32**	**64**	**128**	**256**	**512**
50	47.2	43.0	40.9	40.5	40.6
75	36.7	32.0	34.1	35.5	36.2
100	31.1	29.2	28.3	27.6	27.4
125	31.6	31.4	29.8	29.1	28.9
150	32.7	31.5	30.8	33.2	33.6

4. RoDigits Text-Independent Speaker Recognition System

Speaker recognition is another field of interest in speech processing. 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.

4.1. Alize GMM-UBM speaker models

Alize [18] is an open-source toolkit which provides algorithms specialized in features extraction, speaker model training and methods to determine their identity. The experiments presented in this article use 19-dimensional vectors of cepstral coefficients (MFCC). Speaker modeling is achieved based on GMM-UBM framework.

Any speaker recognition system consists of two major stages. The first stage is the training stage, sometimes called the enrollment stage. This phase involves the creation of a general speech model, comprising voice samples from many speakers. The model obtained will be a model that characterizes speech in general terms, without being specific to a particular speaker. This model is called Universal Background Model (UBM). Also in this stage, the creation of individual models takes place. Starting from the UBM, a model that characterizes the speech of a particular speaker will be derived. The vocal features of each speaker are modeled using a mixture of Gaussian densities (GMM). The speaker adaptation stage is repeated for all speakers that need to be enrolled in the system. The second stage is the test stage, which supposes matching a speech sample against speaker models and against the UBM, gaining a similarity score on which decisions are made. The standard performance figures for speaker recognition systems are the false rejection rate (FRR) and false acceptance rate (FAR).

4.2. Speaker verification scenario

Speaker verification is the task of verifying if the claimed speaker identity matches his real identity. Speech samples received from a speaker who claims to have a certain identity are compared against the UBM and against the personal model (GMM) of the claimed speaker. If the distance to the individual model is smaller than the distance from the universal model, the system decides that the verified speaker is the claimed speaker, otherwise he is an impostor. In this scenario, a false rejection error occurs if the system erroneously decides that the claimed identity is not the real speaker identity. A false acceptance error occurs if the system erroneously decides that a falsely claimed identity is the real speaker identity.

The UBM was trained using the training and development corpus sets, as defined in Section 2.4. A number of 139 speakers, 100 files from each one, were used to train the universal model. The number of Gaussian densities was varied between 16 and 512, more and more complex UBMs being obtained. Individual GMMs were trained for the same 139 speakers by deriving UBM. Thus, between 10 and 80 files were used as enrollment files for each speaker in the training set.

These enrollment files had the role of adapting the UBM to each known speaker, resulting in individual GMMs.

The development set files were used to compute the false rejection rates. Thus, as many as 20 files from 139 speakers were tested against a pair made up of its own GMM and the UBM, totaling 2.780 tests.

Both the development set and the evaluation set were used to compute the false acceptance rates. Twenty files from each of the 139 speakers in development set were tested against each existing GMM and UBM pair, except for their own GMM. A total of 383.640 tests were performed. Also, the evaluation corpus was used. The evaluation set contains 100, respectively 99 files from 15 speakers, which are not part of the training and development set. These files are tested against each GMM and UBM pair, resulting a total of 206.971 false acceptance tests.

Based on the obtained results, a number of observations can be made and some conclusions can be drawn. As a general observation, simple models are characterized by the fact that more speakers are modeled by the same Gaussian density within the UBM. For a number of 16 Gaussian densities and 139 speakers, there are about 9 speakers modeled by a single Gaussian. Instead, for the most complex model, trained with 512 Gaussian densities, about 4 densities from UBM are assigned to each speaker.

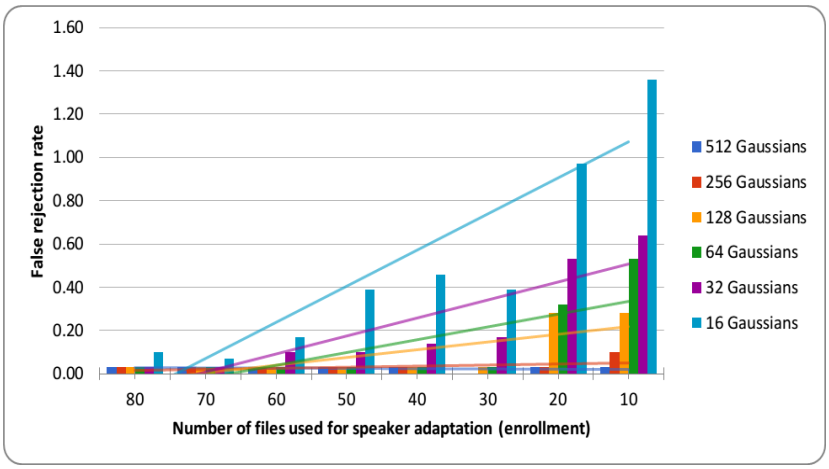

Fig. 1. False rejection rates in speaker verification scenario

As can be seen in Fig. 1, using simple models, trained with few Gaussian densities, the false rejection rate decreases as the number of enrollment data increases. This happens because the speaker-specific GMM, derived from a simple UBM, must have sufficient adaptation data so as to obtain a better score than the UBM. This is because the Gaussian densities in UBM are not specific to the speaker: a single Gaussian density models about 9 speakers. Instead, using

complex models, the false rejection rates do not vary depending on the amount of enrollment data. This is possible because the speaker-specific GMM, derived from a complex UBM, does not need very much enrollment data: the UBM already contains Gaussian densities specific to that speaker. Generally, a more complex model is better. For 139 speakers there is a need for a sufficiently complex model that has enough parameters to model all the variability of the database. In the case of 512 Gaussian densities per model, 4 Gaussian densities model a speaker, getting the best result: 0.03% FRR, regardless of the number of adaptation files.

As can be seen in Fig. 2, few data for adaptation of a simple UBM, trained with few Gaussian densities, lead to lower false acceptance errors than few adaptation data of a more complex UBM. Complex UBMs model relatively well all speakers with different Gaussian densities, having more Gaussian densities for the same speaker. Consequently, GMMs derived from it and adapted with few data will be not speaker-specific enough. Thus, a speaker who claims to be someone else could get a better score on that other person GMM than on the UBM. Performance is constant for 64 Gaussian densities, regardless of the number of adaptation files. The best results were obtained for 512 Gaussian densities and 80 adaptation files, with a value of about 4% FAR.

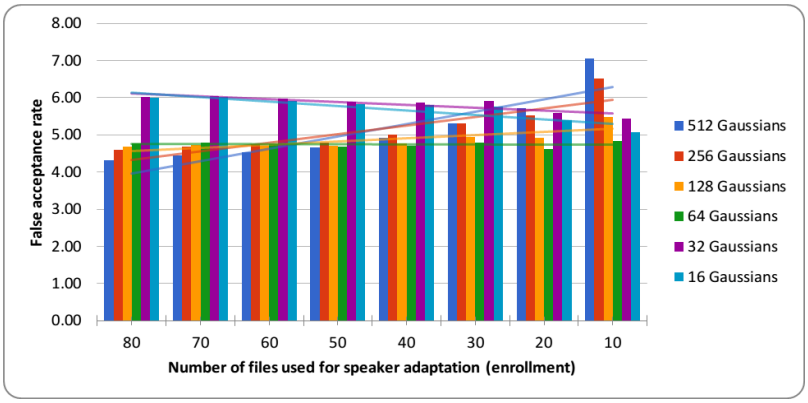

Fig. 2. False acceptance rates in speaker verification scenario

4.3. Close-set speaker identification scenario

Speaker identification consists in determining the identity of the speaker, without providing any priori information about his possible identity. Speech samples provided at the input of the system are compared in turn against each individual model and against the universal model. If the speech sample from the tested speaker does not belong to any speaker whose individual model exists in the system, the best score should be obtained on the universal model. Instead, if there is an individual model for the tested speaker, the best score should be obtained on his own model. In this scenario, an identification error occurs if the

system erroneously decides that the voice of a known speaker, enrolled in the system, but without knowing his claimed identity, belongs to another enrolled speaker. Also, an identification error is considered the situation when the system erroneously decides that the voice of a non-enrolled speaker, belongs to an enrolled speaker.

The training of UBMs and individual models took place in the same way as in the speaker verification scenario as described in Section 4.2. Experiments were performed in closed set scenarios, where the tested speaker is always one of those enrolled in the system, and in open set scenarios, where the tested speaker may be one outside the system, which has not been enrolled, so does not have an own GMM.

Fig. 3 presents the identification error rate in closed-set scenario. The results were obtained on the development set (known speakers) files. Twenty files from 139 speakers were tested against each existing GMM and against the UBM, with a total of 386.420 tests. The speakers from the development set have their own GMM models in the system.

The error rate decreases with the number of adaptation files. This fact is very pronounced for simple models, but it is also true for complex models.

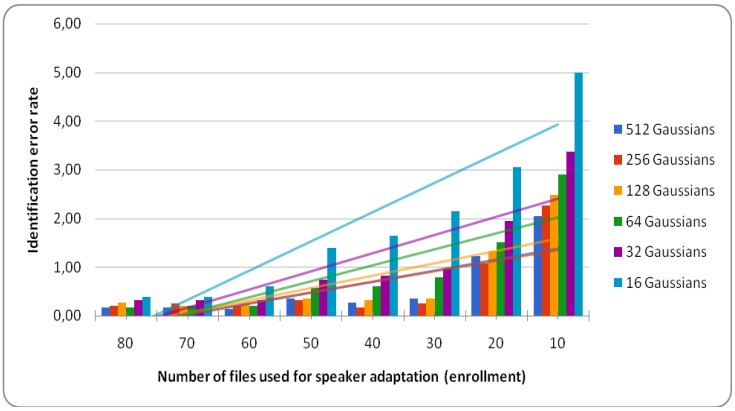

Fig. 3. Identification error in closed-set in speaker identification scenario

Open set scenario experiments (using the evaluation set) were also performed. A total of 99, respectively 100 files from 15 non-enrolled speakers were tested against each existing GMM and against the UBM, summing up a total of 206.971 tests. Speakers being unknown, the best matching should be done against the UBM and not against any of the GMMs for the enrolled speakers. Unfortunately, the results are not good in this scenario, the error rate being greater than 90%, regardless of the number of Gaussian densities and the number of enrollment files.

5. Conclusions

This paper introduced a Romanian corpus of spoken digits, called RoDigits. This corpus, along with the reference transcripts, is available online under the Creative Commons BY-NC-ND 3.0 license.

The first part of the paper presented data about speakers, details about corpus acquisition, modalities to clean and validate it. The corpus was divided into 3 sets: training, development and testing. They are suitable for speech and speech recognition tasks.

The corpus was used to create a connected-digits speech recognition system. The acoustic models were trained using both the HMM-GMM framework and neural networks. Parameters such as the number of Gaussian densities, the number of leaves, respectively the number of hidden layers and the number of epochs have been varied. The lowest WER achieved was 0.28%. The neural network models did not provide better results because the small size of the training set. HMM-GMM models are sufficient for such a low-vocabulary recognition task and few training data.

Finally, the corpus was used in speaker recognition experiments using the GMM-UBM framework. Using it, speech data from more speakers was modeled, obtaining universal models, as well as individual models, specific to particular speakers. The number of Gaussian densities and the number of files used to enroll speakers in the system were varied. Speaker verification experiments were performed, resulting in almost 0% false rejection rates, or around 5% false acceptance rates. Close-set speaker identification experiments have indicated errors below 1%, while open-set errors are quite high and will be further investigated.

R E F E R E N C E S

[1] *G. Rehm, H. Uszkoreit, I. Dagan, V. Goetcherian, M. U. Dogan and T. Váradi*, An update and extension of the META-NET Study “Europe’s Languages in the digital age”, (2014): 1-8.

[2] *Ș. D. Dumitrescu, T. Boroș and R. Ion*, 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 (2014): 90-94.

[3] *A. Kabir and M. Giurgiu*, A romanian corpus for speech perception and automatic speech recognition, The 10th International Conference on Signal Processing, Robotics and Automation. 2011.

[4] *A. D. Bibiri, D. Cristea, L. Pistol, L. A. Scutelnicu and A. Turculeț*, Romanian Corpus For Speech-To-Text Alignment, In Proc. of the 9th International Conference on Linguistic Resources And Tools For Processing The Romanian Language (pp. 151-162).

[5] *M. Boldea, C. Munteanu and A. Doroga*, 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.

[6] *V. Popescu, C. Petrea, D. Haneș, A. Buzo and C. Burileanu*, Spontaneous Speech Database for Romanian Language, 2008.

[7] *A. Stan, J. Yamagishi, S. King and M. Aylett*, The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-beased speech synthesis system using a high sampling rate, In Speech Communication 53.3 (2011): 442-450.

[8] *A. Stan, F. Dinescu, C. Ţiple, Ș. Meza, B. Orza, M. Chirilă and M. Giurgiu*, The SWARA speech corpus: A large parallel Romanian read speech dataset, In Speech Technology and Human-Computer Dialogue (SpeD), 2017 International Conference on (pp. 1-6). IEEE.

[9] *B. Tarjan, T. Mozsolics, A. Balog, D. Halmos, T. Fegyo and P. Mihajlik*, Broadcast news transcription in Central-East European languages, In Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on (pp. 59-64). IEEE.

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

[11] *A. L. Georgescu, H. Cucu and C. Burileanu*, SpeeD's DNN approach to Romanian speech recognition, In Speech Technology and Human-Computer Dialogue (SpeD), 2017 International Conference on (pp. 1-8). IEEE.

[12] *A. Caranica, H. Cucu, A. Buzo and C. Burileanu*, On the Design of an Automatic Speech Recognition System for Romanian Language, Control Engineering and Applied Informatics. 18. 65-76.

[13] *R.G. Leonard*, A database for speaker-independent digit recognition, Proc. ICASSP, 1984, 42.11.1-42.11.4.

[14] *A. Larcher, K. A. Lee, B. Ma and H. Li*, Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication, 2014, 60: 56-77.

[15] Speech and Dialogue Research laboratory. Download section: https://speed.pub.ro/downloads (accessed at 26.01.2018)

[16] *D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel and J Silovsky*, The Kaldi speech recognition toolkit, In IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.

[17] *V. Peddinti, D. Povey and S. Khudanpur*, A time delay neural network architecture for efficient modeling of long temporal contexts, In: Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[18] *A. Larcher, J. F. Bonastre, B. G. Fauve, K. A. Lee, C. Lévy, H. Li and J.Y Parfait*, ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition, In Interspeech. 2013. p. 2768-2772.

[19] *P. Lamere, P. Kwok, E. Gouvêa, B. Raj, R. Singh, W. Walker, M. Warmuth, P. Wolf,* The CMU SPHINX-4 Speech Recognition System, Proc. of ICASSP 2003.

[20] *CMU Sphinx GitHub repository*: https://github.com/cmusphinx

[21] *J. J. Godfrey, E. C. Holliman and J. McDaniel*, SWITCHBOARD: Telephone speech corpus for research and development, In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on (Vol. 1, pp. 517-520). IEEE.

[22] *V. Panayotov, G. Chen, D. Povey and S. Khudanpur*, Librispeech: an ASR corpus based on public domain audio books, In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 5206-5210). IEEE.