Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6606–6612
Marseille, 11–16 May 2020

RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, Corneliu Burileanu
Speech and Dialogue Research Laboratory
University Politehnica of Bucharest
Bucharest, Romania
lucian.georgescu@speed.pub.ro, horia.cucu@upb.ro, buzo.andi@gmail.com, corneliu.burileanu@upb.ro

Abstract

Although many efforts have been made in the last decade to enhance the speech and language resources for Romanian, this language is still considered under-resourced. While for many other languages there are large speech corpora available for research and commercial applications, for Romanian language the largest publicly available corpus to date comprises less than 50 hours of speech. In this context, Speech and Dialogue research group releases Read Speech Corpus (RSC) – a Romanian speech corpus developed in-house, comprising 100 hours of speech recordings from 164 different speakers. The paper describes the development of the corpus and presents baseline automatic speech recognition (ASR) results using state-of-the-art ASR technology: Kaldi speech recognition toolkit.

Keywords: speech corpus, automatic speech recognition

1. Introduction

Verbal communication plays a crucial role in our lives. The advance of engineering has brought the possibility of communication between human and machine. Speech technologies have reached some maturity in recent years and intense efforts in the research area led to the emergence of many commercial products. At present, there is a wide range of voice user interfaces, integrated in various applications, such as operating systems, both for personal computers and for smartphones, home and car control systems or other hands-free applications. There are already attempts to create portable translators – pocket-size devices helping people who speak different languages to understand each other in real time.

The algorithms involved in all these applications are able to perform both automatic speech recognition and speech synthesis and they are activated based on the spoken hotword detection mechanism. The very high performances that they achieve are thanks to the use of deep neural networks together with graphic cards for intense computation. Most of the best algorithms can be found in open source toolkits, along with the data preparation and system training recipes.

Although obtaining a speech recognition system seems quite easy, the biggest problem is still represented by the availability of acoustic and linguistic resources. Large amounts of data are required to train such systems and they are not public or simply do not exist for many of the spoken languages. While this problem does not arise for English (for this language there are corpora that contain thousands and tens of thousands hours of speech), Romanian language is still struggling with a shortage of resources that can be used in speech technology systems.

The most important speech resources in Romanian are presented in Table 1. We can notice that not all of them are public and the amount of data has the order of hours or tens of hours, in few cases they sum up to several hundred hours of speech.

One of the first Romanian speech corpus contains translations of English EUROM-1 corpus, totalling 10 hours recorded by 100 speakers [Boldea et al.1998]. An utterance level manually annotated corpus is SWARA [Stan et al.2017], which contains 21 hours of speech from 100 individuals. The Romanian subset of MaSS corpus [Boito et al.2019] comprises 23 hours of read speech from New Testament. Another manually annotated speech corpus representing 31 hours of broadcast news is presented in [Tarján et al.2012]. In [Suciu et al.2017] is presented a 40 hours corpus of banking call center scenarios, which was recorded by 30 speakers. One of the largest corpora is presented in [Mititelu et al.2014], but this is cvasi-public, in the sense that the corpus could be interrogated through a web interface, but the raw material, speech and transcriptions, cannot downloaded.

Over the time, our research group has made a sustained effort to create new acoustic resources. In [Georgescu et al.2018] we presented speech and speaker recognition experiments using a newly available to the scientific community speech corpus, RoDigits. It contains 38 hours of spoken digits from 154 speakers. Then, there were several attempts to automatically annotate speech corpora that were acquired from the media, especially spontaneous speech from radio and TV broadcasts, sometimes affected by background noise. In this way we obtained the corpora presented in [Cucu et al.2014,Georgescu et al.2017,Georgescu and Cucu2018, Georgescu et al.2019].

This paper introduces the largest publicly-available Romanian **R**ead **S**peech **C**orpus, called RSC. The recordings represent read utterances from literature, news and interviews. This is a core speech corpus used for training our speech recognition systems, presented in a series of previous papers.

The remainder of this paper is organized as follows. Section 2 describes the steps required for RSC corpus development, as well as in-depth details and statistics about it. Section 3 offers some baseline results of automatic speech

Name	Speech Type	Domain	Size [hours]	Availability
RASC [Dumitrescu et al.2014]	Read	Wikipedia Articles	4.8	Public
RoDigits [Georgescu et al.2018]	Read	Spoken Digits	38.0	Public
RO-GRID [Kabir and Giurgiu2011]	Read	General	6.6	Public
IIT [Bibiri et al.2013]	Read	Literature	0.8	Non-Public
N/A [Boldea et al.1998]	Read	Eurom-1 Adapted Translations	10.0	Non-Public
N/A [Popescu et al.]	Spont.	Internet, TV	4.0	Non-Public
RSS [Stan et al.2011]	Spont.	Internet, TV	4.0	Public
SWARA [Stan et al.2017]	Read	Newspapers	21.0	Public
MaSS [Boito et al.2019]	Read	Bible	23.1	Public
N/A [Tarján et al.2012]	Spont.	Broadcast News	31.0	Non-Public
N/A [Suciu et al.2017]	Spont.	Banking	40.0	Non-Public
SSC-train1 [Cucu et al.2014]	Spont.	Radio and TV	27.5	Non-Public
SSC-train2 [Georgescu et al.2017]	Spont.	Radio and TV	103.0	Non-Public
SSC-train3 [Georgescu and Cucu2018]	Spont.	Radio and TV	49.5	Non-Public
SSC-train4 [Georgescu et al.2019]	Spont.	Radio and TV	280.0	Non-Public
CoRoLa [Mititelu et al.2014]	N/A	Various Sources	152.0	Cvasi-Public

Table 1: Romanian speech corpora

recognition, when the training step was done based on the RSC corpus. Details are provided about the neural network used in acoustic modeling, as well as information on text resources and language modeling. Finally, we draw some conclusions in section 4.

2. Corpus Development

2.1. Recording the Corpus

The RSC was collected by Speech & Dialogue (SpeeD)[1] Research Laboratory from University Politehnica of Bucharest. The recordings were made under different conditions, various microphones and various audio recording systems, using an online audio recording application developed within our team. In order to use the application, the user have to provide his credentials into the log in form. Then, if she or he is not enrolled as a speaker in the system, a new speaker profile should be created by filling some personal data such as first name, last name, email, mother language, gender, age and eventually some unusual speech characteristics if they exist. The corresponding form for this step is presented in Figure 1.

Before starting the actual recordings, the user must perform two calibration recordings: one that does not contain speech, to capture the background noise, and another that contains speech. These are needed to determine if the environmental conditions are appropriate for the recording session, otherwise one of the following cases may be encountered: low signal-to-noise ration, unplugged microphone, uncalibrated microphone and so on. After selecting the recordings menu, the user must select his own speaker from the speakers list and the group of phrases for which he wants to make the recordings. Once the phrase group has been selected, the first phrase in the group will appear on the screen. The user can browse through the phrases list and each phrase has a status, which represents whether it has already been recorded or not. The first category also offers the playback option, so that the user can verify that his recording is correct. The recording session screen is presented in Figure 2.

The speech recorder application can be found online here[2]. All the audio files have some common features: all of them were sampled at 16 KHz, 1 channel, with 16 bits per sample precision. The encoding is 16-bit Signed Integer PCM.

2.2. Corpus description

The speakers involved in RSC recordings were mainly students and staff of Faculty of Electronics, Telecommunications and Information Technology from University "Politehnica" of Bucharest. The corpus consists of 136,120 audio files collected from 164 Romanian native speakers, 107 male speakers and 57 female speakers. They range in age from 19 to 60 years, with an average around the age of 24. A percentage of 81% from the total number of speakers are between 21-25. More details on gender and age distribution can be found in Figure 3.

The phrases and sentences to be recorded were chosen from various sources in Romanian language, mainly literature and online news. The length of the phrases varies from a single word to as many as 50 words. Many recordings comprise a single word, such as *păsări* - birds, *steag* - flag, *găştile* - gangs. These words were selected to cover the list of all possible syllables in Romanian.

[1] https://speed.pub.ro/

[2] https://speed.pub.ro/speech-recorder/

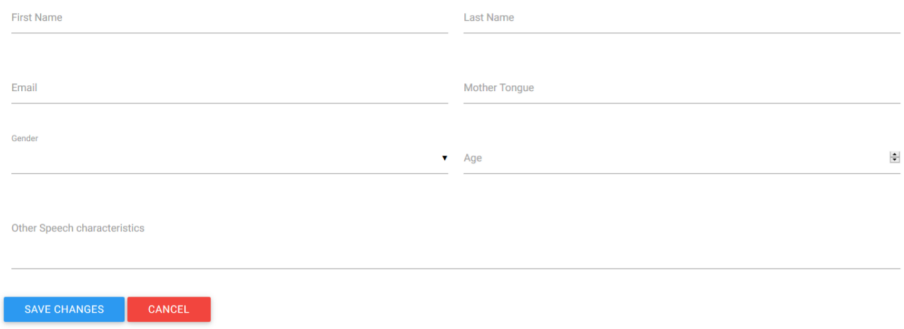

Figure 1: Enrolling a new speaker.

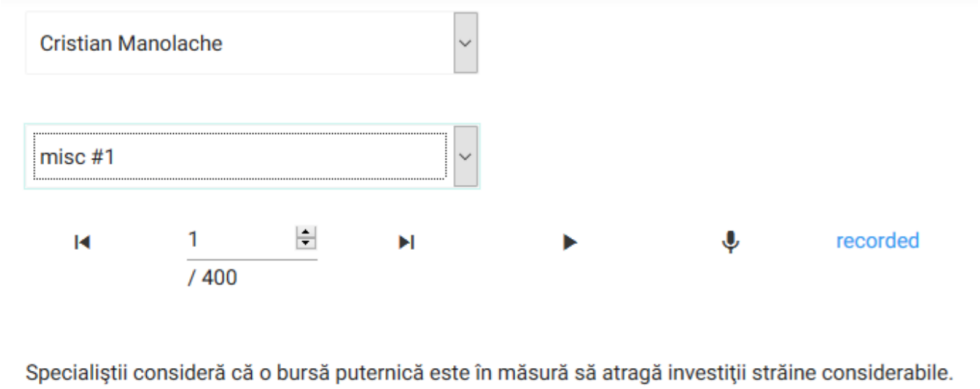

Figure 2: Recording a sentence.

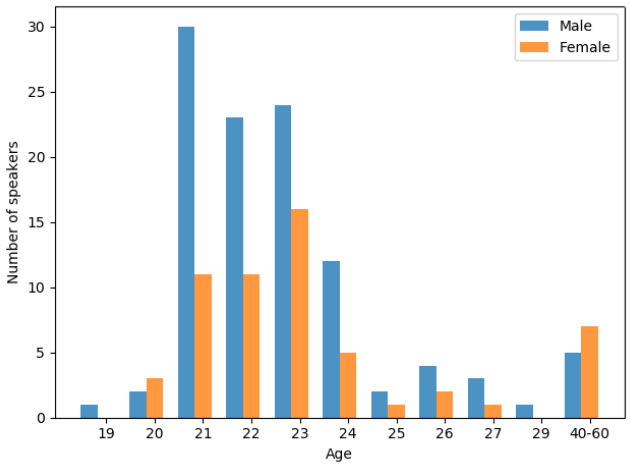

Figure 3: Gender and age distribution across the corpus.

Other sentences were selected from Romanian novels or international novels translated to Romanian. For example, these sentences:

- *Vameşul a privit sabia pe care soţia mea o purta cu ea şi ne-a întrebat ce aveam de gând să facem cu ea.* - The customs officer looked at the sword my wife was carrying with her and asked us what we were going to do with it.,
- *Minciuna a avut succes.* - The lie was successful.

were selected from the novel "The Alchemist" by Paulo Coelho, while these sentences:

- *I-am spus ce vroiam, şi el a întins mâna şi i-am pus în palmă mica monedă galbenă.* - I told him what I wanted, and he reached out and put the little yellow coin in his palm.
- *Se apropia sfârşitul anului şi dascălul examina toată clasa.* - The end of the year was approaching and the teacher examined the whole class.

were selected from the novel "Viaţa ca o pradă" by Marin Preda.

Finally, many sentences, the longest ones, were selected from online news. Here are a few examples:

- *Modul în care poate fi făcut acest lucru ţine de fiecare guvern.* - The way how this can be done is up to each government.
- *O furtună puternică a smuls copaci din rădăcini şi a inundat străzi.* - A powerful storm ripped trees from the roots and flooded the streets.
- *Preşedintele partidului, Emil Boc, a precizat că moţiunea simplă va fi depusă la Camera Deputaţilor săptămâna viitoare.* - The party's president, Emil Boc, said that the simple motion will be tabled in the Chamber of Deputies next week.

Information on the corresponding transcripts, such as the number of sentences, the number of words, the average length of sentences can be found in Table 2.
In general, there are between 130 and 11,000 audio files per speaker. The total size of the database is around 100 hours. The average length of an utterance is 2.6 seconds. RSC is split into training and evaluation sets, as follows. The training set (RSC-train) contains 133,616 files from 156 speakers, 103 male and 53 female. The evaluation set (RSC-eval) comprises 2,504 files from 21 speakers, 10 male and 11 female. More details regarding speech from the both sets are presented in Table 3, such as the number of speakers, their gender, number of utterances, duration and average duration per speaker. It should be mentioned that 13 speakers are found in both the evaluation and training sets. The above overlap occurs only in terms of speakers, not in terms of utterances. The corpus has a total size of 8.3 GB.

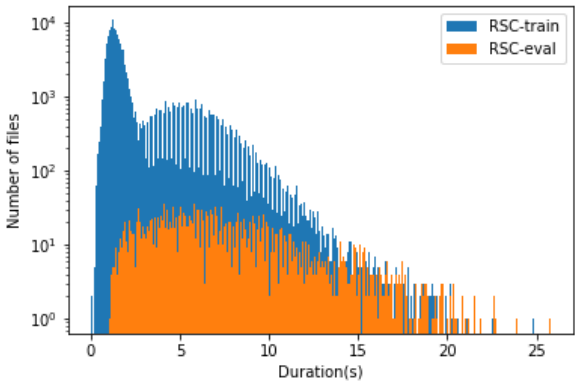

Figure 4: Files duration histogram.

In terms of file distribution depending on their duration, many files in the training set, approximately 100k, last between 0-2.5 seconds, these being the files with isolated word utterances. At the opposite pole, less than 200 files have a duration of more than 15 seconds, these being the long utterances from online media. The evaluation set has a more balanced distribution, most of the files last between 1.5-17 seconds, but there are also files longer than 25 seconds. These statistics are graphically represented in Figure 4.
In a close correlation with the duration of the sentences is their size in terms of number of words. The isolated word utterances constitute almost 10k sentences from the training set, while 42 sentences contain more than 30 words. The evaluation set is more balanced, almost all sentences contain between 1 and 30 words, 12 sentences exceeding this threshold. Figure 5 presents the histogram of sentence length, expressed as the number of words.
Regarding the total duration of recordings per speaker, a total of 140 speakers from the training set recorded individually for up to 2,500 seconds, which means about 40 minutes. Another 12 speakers recorded between 2,500-15,000 seconds, which means a maximum of 4 hours. There are also 4 speakers who made much more recordings, one of them having a maximum over 26,000 seconds, which

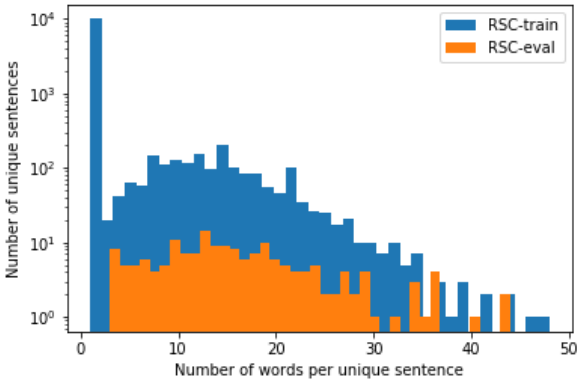

Figure 5: Words per sentence histogram.

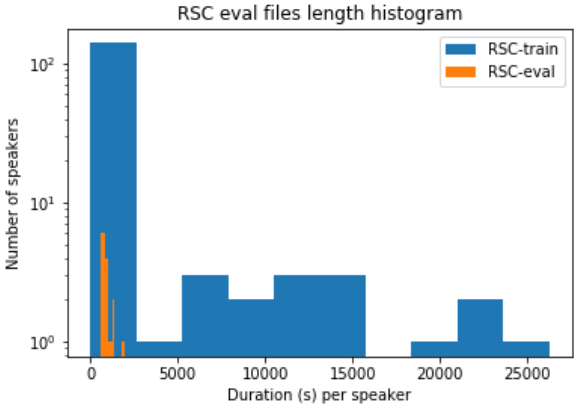

Figure 6: Duration per speaker histogram.

means over 7 hours. The evaluation set is smaller, so all the speakers recorded up to a maximum of 1,500 seconds, about 25 minutes. Figure 6 illustrates this distribution of the total duration of recordings per speaker.

2.3. Availability of the Corpus

The corpus is publicly-available online under the Creative Commons BY-NC-ND 3.0 license and it can be found at the *Download* section on SpeeD's laboratory website. The archive contains a directory called *wav*, which contains in turn other directories name by the ID of each speaker, each of them comprising the speaker's recordings. Another directory in the root of the archive is *text*, containing 6 text files in Kaldi's data format, 3 for each subset: (i) *set.text*, a two-column correspondence between file IDs and the transcriptions, (ii) *set.utt2spk*, a two-column correspondence

	RSC-train	**RSC-eval**
Distinct sentences	11,752	172
Avg. words per sentence	3	16
Total words	532,550	42,012
Distinct words	35,342	2,839

Table 2: Corresponding transcriptions info

	RSC-train	RSC-eval
Speakers	156	21
Male Speakers	103	10
Female Speakers	53	11
Duration	94 h, 46 m	5 h, 18 m
Avg. duration of utterances	2.5 s	7.6 s
Utterances	133,616	2,504
Avg. utterances per speaker	856	119

Table 3: Speech info

between file IDs and speaker IDs and (iii) *set.wav.scp*, one column containing the relative path to each audio file, where *set* can be *train* or *eval*. The content of all these files is plain text.

3. Automatic Speech Recognition using RSC

This section presents the automatic speech recognition system which was trained and evaluated on RSC. We are describing it from the acoustic and linguistic components point of view, as well as the main stages in the training process.

3.1. Phonetic Dictionary and Language Models

The language resources we used for these experiments are two text corpora: one of 315M words gathered from news websites and another containing 40M words from meetings transcripts. These corpora were used to derive n-gram language models using the SRI-LM toolkit and then interpolated using a weight of 0.5 with the same toolkit. We used these probabilistic, n-gram language models to perform speech decoding and language rescoring. For language rescoring we also used an RNN-based language model [Mikolov et al.2010] with 5-word history.

Following previous experiments [Georgescu et al.2017], it has been shown that the optimal ratio between the performance and the hardware resources involved, as well as the real time factor, is obtained when the language model from speech decoding is a 2-gram, while the language rescoring can be done with a probabilistic 4-gram LM or an RNN-LM. The neural network used to train the RNN-LM comprises 3 *Time-Delay* blocks (TDNN) [Peddinti et al.2015], composed by affine, ReLU and renorm layers. Two LSTM layers are interleaved between these blocks. The embedding size of the network is 800.

The phonetic model consists of a dictionary with 200k words (the same as the lexicon of the language model) and comprises the phonetic transcription of each word.

3.2. Acoustic Models

From the point of view of acoustic modeling, we used the best recipe available in Kaldi TED-LIUM system at the moment *(local/chain/tuning/run_tdnn_1g.sh)*. Kaldi supports training of both HMM-GMM and HMM-TDNN models, the latter being trained on top of the alignments obtained with the first one. The HMM-GMM training is performed using Mel-Frequency Cestral Coefficients (MFCCs) as features, 13-dimensional vectors from each acoustic frame of 25 ms and 10 ms shift. Over them is applied the cepstral mean and variance normalization (CMVN) technique. The

Acoustic Model	Language Model	Task	WER [%]
Tri1	2-gram	dec.	10.28
	4-gram	rescr.	7.93
	RNN-LM	rescr.	6.69
Tri2 [LDA+MLLT]	2-gram	dec.	9.25
	4-gram	rescr.	7.26
	RNN-LM	rescr.	6.46
Tri3 [SAT]	2-gram	dec.	7.49
	4-gram	rescr.	5.94
	RNN-LM	rescr.	5.34
TDNN NNET3	2-gram	dec.	4.49
	4-gram	rescr.	3.57
	RNN-LM	rescr.	3.27

Table 4: RSC-eval results using various models

acoustic models are obtained in an iterative way, each being trained based on the forced alignments obtained with the previous one. The first system, called *mono*, comprises context-independent phoneme models. It is followed by 3 systems with context-dependent phonemes, *tri1, tri2, tri3*, over which optimization techniques such as Linear Discriminant Analysis (LDA), Maximum Linear Likelihood Transform (MLLT) [Gales1998] or Speaker Adaptive Training (SAT) [Povey et al.2008] are applied in turn.

Therefore, we used for acoustic modeling a hybrid HMM-DNN model. The neural network component, a Factored Time-Delay Neural Network (TDNN-F) [Povey et al.2018], is depicted in Figure 7. Figure 8 describes the features vector from the network input. A number of *N* consecutive 40-dimensional MFCC vectors extracted from each sliding window, together with an 100-dimensional i-vector [Dehak et al.2010] obtained from a chunk of 1.5 seconds, are combined in order to form the network input speech features vector. Usually, the *N* factor is equal to 3, resulting in 220 features.

The core of this DNN architecture is represented by the TDNN block, which is repeated for a couple of times. Such a block is composed of several layers, usually a linear layer, an affine layer, a layer that introduces the ReLU activation function, a batch normalization layer, a dropout layer, and the last layer in the block is a summation layer which adds a certain percentage from the previous block output to the current block output. This technique works as a residual connection. The linear and affine layers receive at the input a temporal context usually formed from the previous output at the current time point, together with previous or next time frames. Typically, the linear layer receives input from *t-3 or t-1* and *t*, while the affine layer receives input from *t* and *t + 1 or t + 3*. The TDNN-F network has two output heads, both used in training, but only one for the inference. The main difference between them is the loss function, cross-entropy versus Lattice-Free Maximum Mutual Information (LF-MMI) [Povey et al.2016]. The 3560-dimensional output of the network represents the posterior probabilities of the acoustic states. The size of this parameter is chosen in a previous step, when a phonemes clustering method is applied.

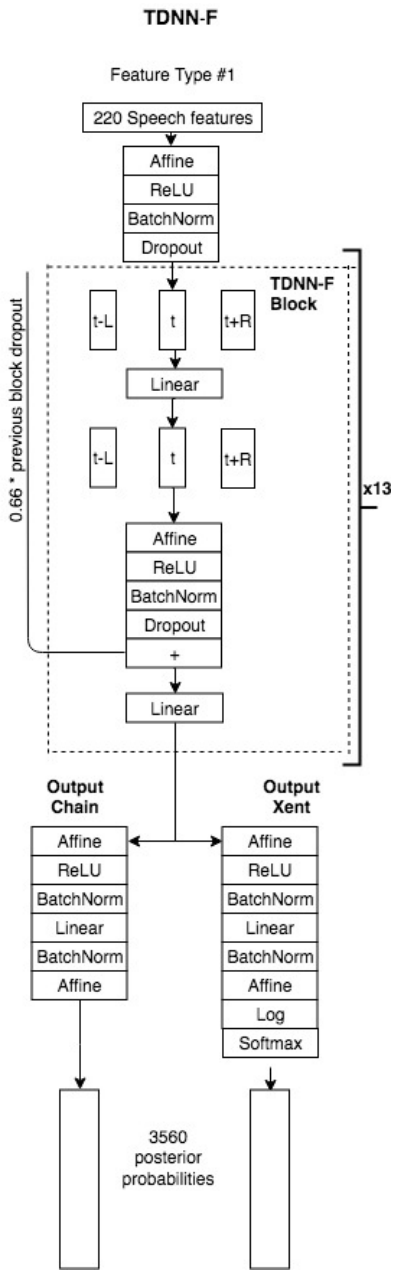

Figure 7: Acoustic model neural network

Table 4 presents some baseline results obtained from training with RSC-train and decoding on RSC-eval. Three HMM-GMM acoustic models and one TDNN-F acoustic model were used, in combination with a 2-gram language model used for decoding, respectively a 4-gram model for rescoring and a RNN language model with 5 words history, also used for rescoring.

As it was obvious, with the use of a more complex acoustic model, as well as the rescoring technique, the results are getting better and better. The maximum performance is obtained from the acoustic model trained with TDNN-F, together with rescoring performed with RNN-LM, the word level error rate (WER[%]) in this case being equal to 3.27%.

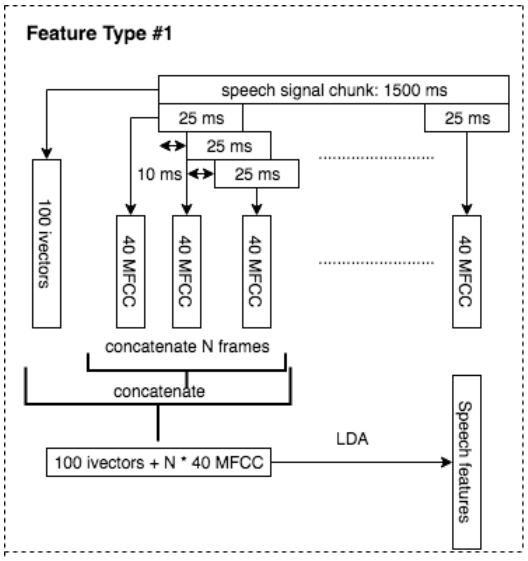

Figure 8: Neural network features

4. Conclusion

This paper presented a Romanian Read Speech Corpus (RSC) which is publicly-available on our laboratory's website as an archive containing both the audio files and the corresponding transcription. We described the recording procedure, as well as we provided an in-depth analysis regarding the structure of the corpus, in terms of files, duration, speakers and number of words.

The train set has been used to train both probabilistic and neural networks acoustic models. These, along with several language models were applied when decoding the evaluation set. The purpose was to provide a baseline for automatic speech recognition in Romanian, the best word error rate being equal to 3.7%.

5. Acknowledgements

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818, within PNCDI III.

6. Bibliographical References

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Gales, M. J. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

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

Povey, D., Kuo, H.-K. J., and Soltau, H. (2008). Fast speaker adaptive training for speech recognition. In *Ninth Annual Conference of the International Speech Communication Association*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747.

7. Language Resource References

Bibiri, Anca-Diana and Cristea, Dan and Pistol, Laura and Scutelnicu, Liviu-Andrei and Turculeț, Adrian. (2013). *Romanian Corpus For Speech-To-Text Alignment*.

Boito, Marcely Zanon and Havard, William N and Garnerin, Mahault and Ferrand, Éric Le and Besacier, Laurent. (2019). *Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible*.

Boldea, Marian and Munteanu, Cosmin and Doroga, Alin. (1998). *Design, Collection and Annotation of a Romanian Speech Database*.

Cucu, Horia and Buzo, Andi and Petrică, Lucian and Burileanu, Dragoş and Burileanu, Corneliu. (2014). *Recent improvements of the SpeeD Romanian LVCSR system*.

Dumitrescu, Stefan Daniel and Boroș, Tiberiu and Ion, Radu. (2014). *Crowd-Sourced, Automatic Speech-Corpora Collection–Building the Romanian Anonymous Speech Corpus*.

Georgescu, Alexandru-Lucian and Cucu, Horia. (2018). *Automatic Annotation of Speech Corpora Using Complementary GMM and DNN Acoustic Models*.

Georgescu, Alexandru-Lucian and Cucu, Horia and Burileanu, Corneliu. (2017). *SpeeD's DNN approach to Romanian speech recognition*.

Georgescu, Alexandru Lucian and Caranica, Alexandru and Cucu, Horia and Burileanu, Corneliu. (2018). *RoDigits-a Romanian Connected-Digits Speech Corpus for Automatic Speech and Speaker Recognition*.

Georgescu, Alexandru-Lucian and Cucu, Horia and Burileanu, Corneliu. (2019). *Progress on automatic annotation of speech corpora using complementary ASR systems*.

Kabir, Ahsanul and Giurgiu, Mircea. (2011). *A romanian corpus for speech perception and automatic speech recognition*.

Mititelu, Verginica Barbu and Irimia, Elena and Tufis, Dan. (2014). *CoRoLa—The Reference Corpus of Contemporary Romanian Language.*

Popescu, Vladimir and Petrea, Cristina and Haneş, Diana and Buzo, Andi and Burileanu, Corneliu.). *Spontaneous Speech Database for Romanian*.

Stan, Adriana and Yamagishi, Junichi and King, Simon and Aylett, Matthew. (2011). *The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate*. Elsevier.

Stan, Adriana and Dinescu, Florina, and Tiple, Cristina and Meza, Serban and Orza, Bogdan and Chirila, Magdalena and Giurgiu, Mircea. (2017). *The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset*.

Suciu, George and Toma, Ştefan-Adrian and Cheveresan, Romulus. (2017). *Towards a continuous speech corpus for banking domain automatic speech recognition*.

Tarján, Balázs and Mozsolics, T and Balog, A and Halmos, D and Fegyó, Tibor and Mihajlik, Péter. (2012). *Broadcast news transcription in Central-East European languages*.