The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset

Adriana Stan,* Florina Dinescu,† Cristina Ţiple,† Şerban Meza,*
Bogdan Orza,* Magdalena Chirilă† and Mircea Giurgiu*
*Communications Department, Technical University of Cluj-Napoca, Romania
†Department of Otorhinolaryngology, Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania
{adriana.stan, serban.meza, bogdan.orza, mircea.giurgiu}@com.utcluj.ro,
{salajan.florina, uzun.cristina, magdalena.chirila}@umfcluj.ro

***Abstract*—This paper introduces one of the largest Romanian speech datasets freely available for both academic and commercial use. The dataset comprises speech data recorded over the last year from 12 speakers, along with 5 other speakers previously recorded in a separate environment. The data was manually segmented at utterance-level and semi-automatically labelled at phone-level. The resulting corpus amounts to approximately 21 hours of high-quality read speech data, split into over 19,000 utterances. The speakers read between 921 and 1493 utterances each. 880 utterances are common to all speakers and add up to over 16 hours of parallel data.**

We present the steps of performing the recordings and data segmentation, as well as a first use of this corpus in the context of synthetic voice development.

Keywords— speech corpus, Romanian, phone-level annotation, read data, speech synthesis

I. Introduction

Building speech-enabled applications has lately become a fast growing field. Aside from the mainstream languages, more and more research groups and companies focus on enabling the speech interaction in any language [1]. However, an essential prerequisite to building such applications is the availability of speech data, and in the best case scenario, the availability of manually annotated speech data. Also, with the increase of voice cloning research, parallel speech corpora are of high interest. Mapping the characteristics of speaker A to speaker B when only a limited amount of adaptation data is available, is another important issue in the speech processing field.

SWARA - Mobile System for Rehabilitative Vocal Assistance of Surgical Aphonia[1] is a national project funded by the Romanian Ministry of Education with the main objective of enabling speech impaired persons, and especially those with surgical aphonia, to use a fast, personalised text-to-speech synthesis system. This objective is achieved through two main directions: 1) a mobile-friendly *fast text-input method* combined with a prototype lip reading system; and 2) a *customised synthetic voice* which resembles the patient's vocal identity as much as possible. If for the fast text input method, adapting a general Romanian language model can be done iteratively and incrementally through the constant use of the application, for the custom synthesis voice, the online adaptation is not possible. Therefore the availability of a large speaker database from which the patient can choose a voice and prosodic characteristics which are similar to his or her identity, is essential.

In comparison to other European languages, Romanian speech datasets are scarce, and have mostly been developed with a particular restrictive purpose in mind. For example, The Romanian Anonymous Speech Corpus (RASC) [2] contains around 3000 utterances collected from various speakers, and recorded by the speakers with their personal equipment and uploaded into an online platform. A spontaneous speech corpus with over 4 hours of data recorded from 12 speakers is presented in [3]. The recordings contain internet broadcast Romanian TV shows, and are sampled at 8kHz. [4] is the Romanian version of the GRID corpus, and includes 400 utterances recorded from 12 speakers manually labelled at phone-level. In [5] the authors built the Romanian version of the EUROM_1 database, achieving a dataset of over 10 hours of speech data from 100 speakers. The IIT corpus presented in [6] contains approximately 45 minutes of read speech sampled at 22kHz collected from 3 female speakers. The largest multi-speaker Romanian corpus available for research purposes is RSC [7]. It contains over 100 hours of speech data uttered by 157 speakers. The speakers used their personal recording equipment via an online application. The corpus has been successfully employed in the development of a large vocabulary automatic speech recognition system, and its results are at the moment one of the best for the Romanian language.

The largest single speaker labelled speech corpus available at the moment is the Romanian Speech Synthesis (RSS) corpus [8].[2] It includes one main speaker with over three and half hours of read data, and a second speaker recorded for over 1 hour and 45 minutes. Both datasets are recorded in a professional studio, and are semi-automatically labelled at

[1]http://speech.utcluj.ro/swara/

[2]http://romaniantts.com/new/db2.php

978-1-5090-6497-7/17/$31.00 2017 IEEE

TABLE I
Sample of the recording prompts along with their English translation.

RO	*Suntem una dintre cele mai vechi familii din Eforie.*
EN	We are one of the oldest families in Eforie.
RO	*Se poate face asta, dar depinde cum o faci, fiecare are modul lui de a vedea lucrurile.*
EN	This can be done, but it depends on how you do it, because everybody has his own way of looking at things.
RO	*Guvernul Britanic a comandat şaizeci de milioane de doze.*
EN	The British Government ordered sixty million doses.
RO	*De pildă, aş face un brand din brânza şi gemurile locale.*
EN	For example, I would brand the local cheese and jams.

phone-level. The corpus is freely available for both academic and commercial purposes.

Starting from this overview of available resources, we concluded that the development of a large dataset of Romanian speech data is essential not only for the objectives of our project, but also for the entire research community. However, such a task is hard to achieve, mainly due to the fact that people are in general quite weary when it comes to recording their voices. Also, the annotation and cleaning of the recorded data is time consuming and requires manual input. As a result, in our previous work [9], we automatically built a speech corpus extracted from an audiobook, labelled at sentence level. The orthographic transcription's accuracy is around 93% at sentence level, and 99.5% at word level. However, when building a text-to-speech synthesis system from this data, we noticed that even such a small error rate caused artefacts in the system's output speech. Methods to correct these errors were also investigated [10], but the results were unfortunately not much better.

Therefore, in order to obtain a general-purpose, high quality speech corpus for Romanian, we had to resume to performing studio recordings and manually check and annotate the data. Having this task set, we also wanted to ensure that the resulting data is of use to more than just our project's objectives. Therefore the recordings contain multiple speakers, reading the same set of random newspaper sentences, in a controlled environment.

The paper is organised as follows: Section II describes the recording procedure with its technical details. Section III introduces the segmentation processes, both at utterance and at phone-level. In Section IV the resulting corpus is presented along with some statistics and a brief overview of the synthetic speech voice development. Conclusions are drawn in Section V.

II. Recording Process

Voice-enabled applications require high quality speech data, so that all the details of the uttering process are accurately captured. This ensures that, independent of the application

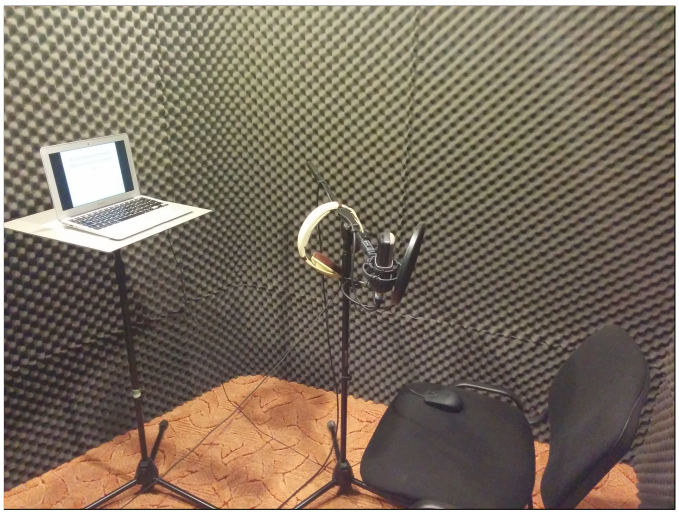

Fig. 1. The recording booth setup for the SWARA corpus.

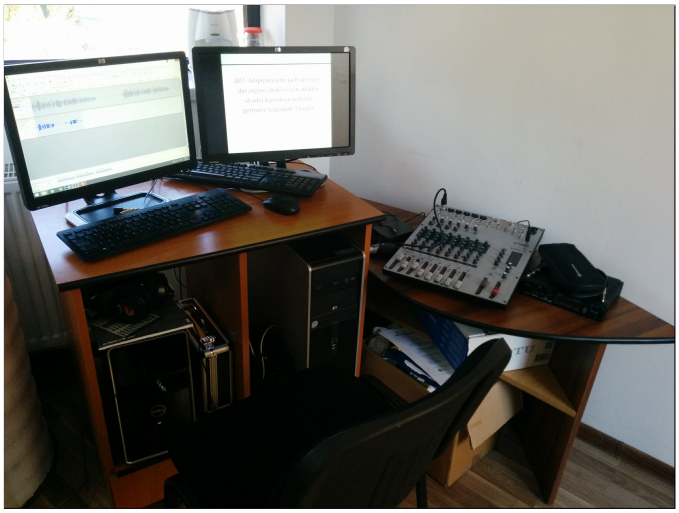

Fig. 2. The monitoring panel for the recording process of the SWARA corpus.

in which the data is used, the analysis, pre-processing and final results obtained from it are not hindered by this step. Therefore, the recordings for our corpus were performed in a studio environment with high-end equipment. Such a studio is already available within our research facility, and consists of a recording booth and an external monitoring station. Figures 1 and 2 show the setup for the recording booth and monitoring panel during the recording process.

The recording booth is a completely isolated sound-proof room, which contains only the microphone, a set of communication headphones and a prompt monitor. The microphone used for recording is a AKG C214 large-diaphragm, condenser, cardioid microphone, placed on shock mount, and additionally protected by a foam windscreen and a pop filter. The prompts are presented to the speaker one at a time via a laptop which is not connected to a power outlet. Due to the lack of reverberation in the booth, which usually causes additional vocal effort, the speaker was able to hear himself

through a pair of headphones. The headphones were also used to communicate with the person standing at the monitoring panel.

The external monitoring panel is used to oversee the recordings. The monitoring person controls both the technical details of the recordings, such as levels, duration or quality, as well as the correspondence between the prompts and what the speaker utters. In case the speaker deviates from the provided text, or if he stops, repeats words, or uses an unnatural pronunciation, the monitoring person would ask him to read that sentence again. The acquisition of the speech data was done through a MOTU UltraLite MK3 sound card interfaced with a computer which ran Audacity as a recording software. The analog-to-digital conversion was performed at 48kHz sampling rate, with a 16 bit depth. An additional Yamaha MW12c digital mixer was used to split the talk-back and recording streams.

The most important requirement for the speakers was to use a natural prosody and read as fluently as possible, without stops and repetitions. The monitoring person would only intervene if they notice any inconsistencies between the sentence and what the speaker read, otherwise the recordings were not interrupted. A short break of 10-15 minutes was made after each 30 minute session, thus ensuring that the speakers do not alter their voice quality.

The set of prompts prepared for the recordings consist of approximately 1000 out-of context sentences, randomly selected from newspaper articles. The set corresponds to the *rnd1*, *rnd2* subsets of the RSS speech corpus. All the numeric values, abbreviations and acronyms from the sentences were expanded to ensure a unitary reading of these tokens. Some samples from the prompts are presented in Table I. This subset of prompts has already proven its efficiency in building synthetic speech voices [8], and was therefore considered to be an appropriate choice for developing the already available Romanian speech resources.

All speakers were volunteers, informed of the purpose of the recordings. An agreement for the data release and future use was signed by each of them.

Aside from the audio acquisition, we were also interested in recording video of the speakers for multimodal speech recognition and lip reading. However, if for the audio recordings the speakers had fewer objections, and some of them were actually enthusiastic about donating their voice, having videos taken of them was not as appealing. This happened even though we assured them that we would only preserve the mouth region of the images. Out of all the volunteers, only two speakers agreed to this. As a result, we also recorded a set of 45 minutes of synchronised video and audio data which are yet to be processed and released.

III. Data Segmentation

A common practice when distributing speech corpora is to deliver them as a set of individual utterances accompanied by their corresponding orthographic transcripts. In some cases their phonetic transcription and other types of annotations are also provided. The way in which the utterance segmentation and phonetic annotations are obtained differs. In some cases, the recordings are performed one sentence at a time, and therefore the utterances are already segmented. However, if no automatic system is used to enable such a type of recording, interfering with the speaker can cause additional discomfort and be time consuming. In terms of the phonetic annotation, most of the available corpora provide semi-automatic annotations, with very few of them having manual labels. This is due to the fact that labelling a large quantity of speech data is laborious, and requires several experts to perform the same task.

For our corpus we opted to record the data without any technically related interruptions, and then to manually segment them at utterance-level and to label at phone-level with a semi-automatic method. The following sections describe these processes into more detail.

A. Utterance-level segmentation

During the recording process, the speakers were not interrupted unless they read a sentence incorrectly, and the outside monitor asked them to repeat it. As a consequence, the result was a set of long sessions of raw speech, which except for the correct utterances, also contained long pauses, dialogues between the speaker and outside monitor, and repeated or erroneous speech segments. Hence, the first task for creating the end result of this dataset was to segment and clean the data at utterance level.

Due mainly to the fact that the utterances' correctness is hard to determine automatically, this segmentation process was carried out manually. Each speaker segmented their own data, and two other individuals re-checked the result. This process was carried out using the Wavesurfer[3] software. The main task of the labellers was to mark the correct utterances using one label, and everything else with a different one. Aside from this, they were also asked to try and maintain at least a 200 ms silence segment at the beginning and end of each utterance. Figure 3 shows a sample of the utterance-level labelling step.

The result of the recording process, followed by this step should have been the complete text subsets *rnd1* and *rnd2* of the RSS corpus. However, despite our best efforts, not all speakers read all the sentences. This was mainly caused by the fact that the speakers either stuttered or repeated a word within the sentence, or sometimes skipped some of the provided prompts altogether.

A different case was when the speaker partially misread the text. In this situation, if small differences between the script and what the speaker read were found, the labeller marked those utterances, but also provided the correct transcript for them.

The output of this first segmentation process is a set of utterances with their orthographic transcripts. As a post-processing step, each utterance was individually normalised to 0dBFS. To enable the use of the corpus for parallel speech

[3]https://sourceforge.net/projects/wavesurfer/

experiments, the data which is common to all the speakers is marked as such, and consists the main part of our released corpus.

B. Phone-level transcription and annotation

Because in most speech-enabled applications the phone is considered to be the smallest and most relevant speech unit, providing the phone-level segmentation of a speech corpus is important. This also ensures a consistency of the results obtained using the speech data across various research groups or methods.

However, manually labelling a large speech corpus at phone-level is a highly time consuming task, and in general, the accurate labelling of phonetic boundaries is not required in the majority of applications. Therefore, semi-automatic or fully automatic methods are used to provide this type of annotation [11]. For the automatic phonetic alignment, the most common method is the use of iteratively trained probabilistic models, such as those based on Hidden Markov Models (HMM) [12] or Deep Neural Networks (DNN) [13]. The results of this type of alignment are widely accepted as a valid input to most applications.

To obtain the phone-level annotation of a speech segment a two step process is performed: first, a grapheme-to-phoneme converter is used to translate the orthographic transcript of the text into its constituent phones; and second, an automatic alignment method is applied in order to determine the phonetic boundaries within the speech data.

For the phonetic transcription of the prompts, we used an open-source Romanian text-processor also developed in the SWARA project. Its phonetic transcriber module is based on simple decisions trees and achieves an accuracy of around 96%. This module was used to transcribe the entire set of sentences from the corpus, including those which were misread and afterwards corrected.

For the semi-automatic labelling we opted for an HMM-based acoustic model training method. The alignment accuracy of this type of acoustic models is around 93% [14] when measured at a 20 ms margin. The models were built using HTK,[4] and have a 5 state left-right configuration. The re-estimation was performed 8 times without any state tying and from a flat start. This means that no manually labelled data was used to train the initial models. However, an initial rough time alignment estimate of the phone boundaries was provided. This estimate is obtained by assigning to each phone a time stamp which is obtained by equal division of the length of the utterance to the number of phones contained within. All the data from the speech corpus (approx. 21 hours) was used to train the models, but no speaker adaptation was performed. In Figure 4 we show an example of the result of the automatic alignment versus a manual reference. It can be noticed that the majority of the phonetic borders are accurately determined. The objective evaluation of the alignment accuracy is beyond the scope of this paper.

[4]http://htk.eng.cam.ac.uk

TABLE II
The contents of the SWARA Corpus. Speakers recorded in different conditions than the ones described in this paper are marked with a star (*).

No.	Speaker ID	Sex	Duration	No. of utterances
1.	*BAS	F	1h 34' 30"	1493
2.	CAU	F	1h 11' 35"	996
3.	*DCS	F	1h 50' 01"	1493
4.	DDM	F	1h 09' 18"	996
5.	*EME	F	1h 53' 36"	1493
6.	FDS	M	0h 57' 21"	996
7.	HTM	F	1h 06' 27"	981
8.	IPS	M	0h 58' 08"	996
9.	PCS	M	1h 08' 03"	996
10.	PMM	F	1h 01' 53"	921
11.	*PSS	M	1h 27' 45"	1486
12.	RMS	M	1h 08' 56"	996
13.	*SAM	F	1h 43' 31"	1493
14.	SDS	M	1h 01' 28"	996
15.	SGS	M	0h 55' 22"	994
16.	TIM	F	1h 09' 27"	973
17.	TSS	M	1h 01' 54"	996

IV. Results

The results of the steps described in Sections II and III represent the contents of the SWARA speech corpus. This section introduces several statistics of it, and also briefly presents the process of building synthetic voices for all the speakers, as well as an eigen voice created from the parallel corpus data.

A. The contents of the SWARA Corpus

During the SWARA project and following the procedures described above, we managed to record a number of 12 volunteer speakers with a total recording time of around 13 hours. However, the RSS corpus also contains two speakers which recorded the same set of prompts. And we also previously recorded 3 other speakers uttering the *rnd1*, *rnd2* and *rnd3* subsets of the RSS corpus. The latter three speakers uttered the data in a TV studio using the same recording equipment as the one used for SWARA. The only difference is the presence of a slight reverberation due to the large recording room. The segmentation process for this data was the same as the one described in Section III.

As there is no such thing as too much data, we combined all these recordings into the released SWARA Corpus. A full overview of its contents is presented in Table II. The speakers which were recorded in different conditions than the ones presented in this paper are marked with a star (*). There are a total of 17 speakers in the final version of the corpus, 9 females and 8 males. Their ages vary between 20 and 35 years old, and some of them have mild regional accents. None of them reported having speech or hearing impairments.

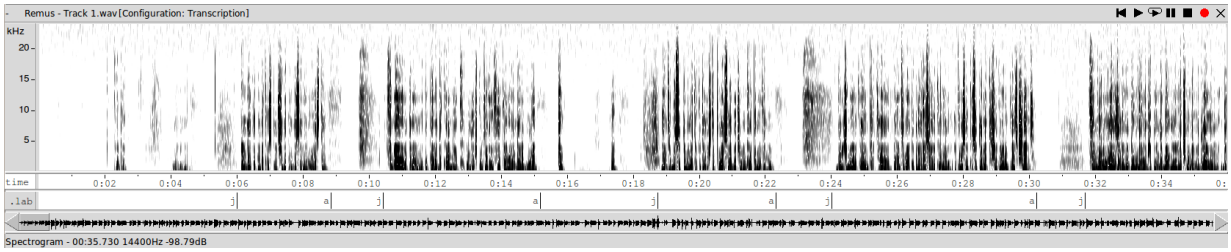

Fig. 3. Sample of the utterance-level annotation process. The correct utterances are labelled with "a" (audio), and everything else is labelled "j" (junk).

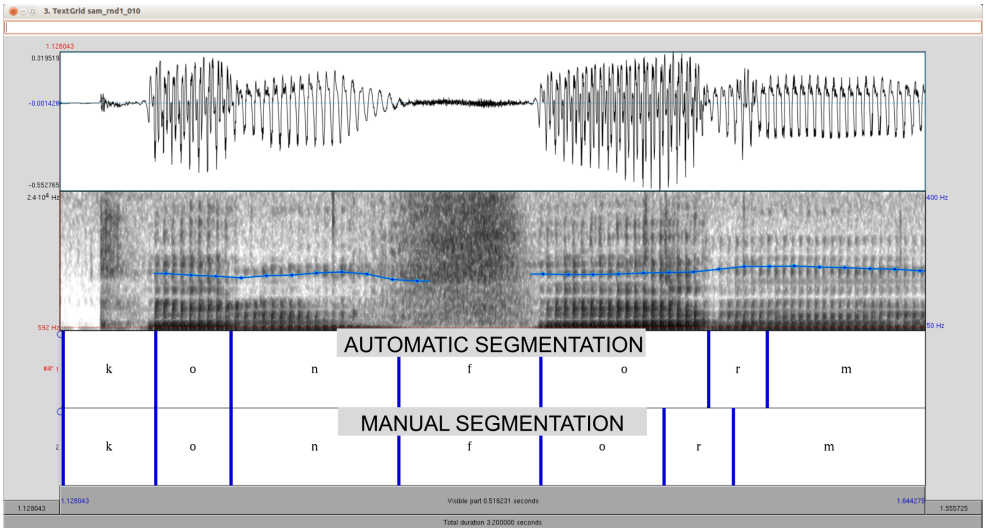

Fig. 4. Sample of the automatic phone-level annotation versus a manual reference, presented in Praat.

However, speaker SDS has a slightly jerky speech, which translates into an uncommon intonation pattern.

The total duration of the corpus is 21 hours and 19 minutes, including silence, and split into 19,292 utterances. The number of utterances recorded by each speaker varies between 921 and 1493. The 1493 set of utterances was read only by the speakers which were previously recorded. The extra utterances represent the *rnd3* subset of the RSS corpus. The newly recorded speakers did not read this extra set due to time limitations. Out of the utterances read by all the speakers, 880 were correctly read by all of them. This data adds up to over 16 hours of parallel speech data, and is marked as such in the release. The corpus is available at http://www.speech.utcluj.ro/swarasc/ under a CC-by-SA 4.0[5] licence, which stipulates that you can copy, distribute, remix and transform the corpus, as long as you give appropriate credit and you distribute your contributions under the same license.

B. Synthetic speech voices

To test the usability of the SWARA corpus, we developed a set of synthetic speech voices. One speaker-dependent voice was built for each speaker in the corpus. A separate one, called an eigen-voice [15] was built using only the subset of data which contains the utterances common to all the speakers. These synthetic voices were developed using the HMM-based Speech Synthesis System (HTS) [16].

The conversion of the orthographic transcripts into HTS-format labels which contain a substantial amount contextual information, was extracted with the help of our SWARA text processing tool. The phone-level alignment obtained in the segmentation process was also used as an input to the system.

However, the contents of the prompts selected for recordings have already proven their efficiency in the HTS system. Therefore, we also wanted to test them in the context of DNN-based speech synthesis systems. For this task we selected the Merlin toolkit [17] and the WORLD vocoder [18], and built a voice from the SAM speaker data. Although no listening tests have been performed so far, the quality of the synthetic voice is similar to the one built using the HTS system. This means that the corpus could be used in both HMM and DNN-based speech synthesis systems.

Samples of all the voices built from the SWARA Corpus are available at: http://www.speech.utcluj.ro/swarasc/samples/.

V. Conclusions

This paper introduced the SWARA Speech Corpus, which is one of the largest speech resources for the Romanian language. It contains over 21 hours of high-quality read speech

[5]https://creativecommons.org/licenses/by-sa/4.0/

data collected from 17 different speakers. 16 hours of the data is composed of 880 utterances read by each of the 17 speakers. This makes it easier to develop and test parallel speech applications, such as waveform-based adaptation or style transplantation.

The corpus was recorded in a professional environment with high-end equipment. The prompts read by the speakers are random sentences selected from newspaper articles. All the data was manually segmented at utterance-level, and semi-automatically labelled at phone-level.

The corpus is released under a CC-by-SA 4.0 licence to enable further advancements in the Romanian speech-enabled applications. A first use of this corpus was presented in the paper, through the development of synthetic speech voices.

As future work, we would like to investigate the use of the SWARA Corpus in automatic speech recognition systems, as well as speaker adaptation for speech synthesis systems.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Romanian Ministry of Education, under the grant agreement PN-II-PT-PCCA-2013-4 Nº 6/2014 (SWARA).

REFERENCES

[1] "KTH and Wikipedia develop first crowdsourced speech engine," www.kth.se/en/forskning/artiklar/kth-hjalper-wikipedia-borja-prata-1.631897, accessed: 2016-03-30.

[2] S. D. Dumitrescu, T. Boros, and R. Ion, "Crowd-sourced, automatic speech-corpora collection - building the Romanian Anonymous Speech Corpus," in *Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL2014)*, Reykjavik, Iceland, May 2014, pp. 90–94.

[3] V. Popescu, C. Petrea, D. Hanes, A. Buzo, and C. Burileanu, "Spontaneous Speech Database for Romanian Language," in *Proc. of 5th European Conference on Intelligent Systems and Technologies*, 2008, pp. 78–86.

[4] A. Kabir and M. Giurgiu, "A Romanian Corpus for Speech Perception and Automatic Speech Recognition," in *Proceeding of 10th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications*, 2011, pp. 323–326.

[5] M. Boldea, C. Munteanu, and A. Doroga, "Design, Collection, and Annotation of a Romanian Speech Database," in *Proceedings of 1st Conference on Language, Resources and Evaluation*, 1998.

[6] A.-D. Bibiri, D. Cristea, L. Pistol, L. A. Scutelnicu, and A. Turculet, "Romanian Corpus For Speech-To-Text Alignment," in *Proc. of the 9th International Conference on Linguistic Resources And Tools For Processing The Romanian Language*, 2013, pp. 151–162.

[7] H. Cucu, A. Buzo, L. Petric, D. Burileanu, and C. Burileanu, "Recent improvements of the SpeeD Romanian LVCSR system," in *Proc. of The 10th International Conference on Communications (COMM)*, May 2014, pp. 1–4.

[8] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. 53, no. 3, pp. 442–450, 2011.

[9] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision," in *Proc. Interspeech*, August 2013, pp. 2331–2335.

[10] A. Moldovan, A. Stan, and M. Giurgiu, "Improving Sentence-level Alignment of Speech with Imperfect Transcripts using Utterance Concatenation and VAD," in *Proc. of IEEE ICCP*, Cluj-Napoca, Romania, September 2016, pp. 171–174.

[11] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[12] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models." *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[13] O. Kalinli, "Combination of auditory attention features with phone posteriors for better automatic phoneme segmentation." in *INTERSPEECH*. ISCA, 2013, pp. 2302–2305.

[14] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, no. 4, pp. 352–368, Apr. 2009.

[15] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*, 2002, pp. 1269–1272.

[16] H. Zen, T. Nose, J. Yamagishi, S. Sako, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*, Aug. 2007, pp. 294–299.

[17] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, USA, 2016, pp. 218–223.

[18] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.