Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions

Mariusz Kleć
Polish-Japanese Academy of Information Technology
Warsaw, Poland *mklec@pjwstk.edu.pl*

Krzysztof Szklanny
Polish-Japanese Academy of Information Technology
Warsaw, Poland *kszklanny@pjwstk.edu.pl*

Alicja Wieczorkowska
Polish-Japanese Academy of Information Technology
Warsaw, Poland *alicja@poljap.edu.pl*

Abstract

This paper presents a solution for generating corpora of simulated Polish speech recordings in complex acoustic environments. The proposed method introduces a layer of unpredictable sound events, in addition to the acoustic scene noise and reverberation, making the solution unique. We generated a corpus comprising over 277 hours of training examples and over 5.5 hours for testing purposes using publicly available data sources. Next, we trained several Conv-TasNet networks on the generated data to enhance single speech and separate two speakers from complex noise. The results of the experiments indicated the potential of the generated corpora for solving these tasks. Researchers can use publicly available code to create their corpora tailored to the Polish language and solve various speech-related tasks.

Keywords: speech denoising, speech separation, speech enhancement.

1. Introduction

Speech recordings can be a valuable source of information in various fields of science, such as linguistics, history, psychology, sociology, and medicine, to name a few. However, non-professional microphones, various acoustic environments, and casual settings can significantly affect the intelligibility of speech. The audio files obtained this way are often noisy, with reverberation and random sound events like car horns or dog barking. Additionally, when two or more people are involved in the conversation, their utterances sometimes occur concurrently (i.e. crosstalk occurs). These issues greatly influence the performance of Automatic Speech Recognition (ASR) services, which require one person to speak at a time and a signal of relatively high signal-to-noise ratio (SNR) to transcribe speech accurately. Therefore, developing speech enhancement and speech separation methods are key preprocessing steps for ASR, speech corpus analysis, and real-time communication.

The speech enhancement can be accomplished by speech denoising [34], increasing the resolution of the signal [21], or by conditional speech synthesis [1]. However, this task can be challenging due to the complex and dynamic nature of the acoustic environment. Various disturbances such as non-stationary noise, reverberation, and other acoustic phenomena and unpredictable sound events may complicate the denoising process further. Recent research has revealed that Deep Learning (DL) techniques are more effective in speech denoising than conventional methods, such as spectral subtraction [41], Wiener filtering [37], and non-negative matrix factorization [14]. The DL-based denoising techniques comprise models based on Wave-U-Net

[7, 46], Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks [12, 8], Generative Adversarial Networks (GAN) [29, 36], Transformers [17], and recently also models that do not require clean data for training [15].

Researchers have also used speech separation techniques to denoise speech signals, as mentioned in [9, 16, 20]. Speech separation is the separation of concurrent speakers in monophonic audio recordings. To this end, the researchers in most cases employ time-domain models, which are more accurate than models based on time-frequency representations. The latter require estimating phase information, which can lead to distortions and inaccuracies during signal reconstruction [38]. Among the successful time-domain models, the Conv-TasNet neural network [23] has proven to be very effective in speech separation, outperforming other time-frequency methods. Another effective model is the Hybrid Tasnet network [43], which integrates the time and frequency domains to improve separation performance. In another paper [22] the model leverages Recurrent Neural Network for utterance-level sequence modelling. Wavesplit [44] is another end-to-end speaker separation model that achieves very high efficiency in various speech separation tasks, including clean mixtures of 2 speakers from WSJ0-2Mix dataset [10], and in noisy and reverberated settings from WHAMR dataset [24]. The authors of [44] obtained 22.2 dB on WSJ0-2Mix and 13.2 dB of Signal-to-Distortion Ratio (SI-SDR) on WHAMR. Moreover, a recent study [38] investigated the use of transformer architecture called SepFormer for the speech separation task, yielding promising results of 22.3 dB of SI-SDR on the WSJ0-2Mix. Additionally, the MossFormer2 model [45] currently achieves the best speech separation results for the Libri2Mix [4] and WSJ0-2Mix [10] datasets, achieving 24.1 dB and 21.7 dB of SI-SDR, respectively.

1.1. Contribution

Real-life speech recordings are often made during spontaneous situations and in uncontrolled acoustic environments. As a result, they can contain a lot of noise, which seriously affects speech intelligibility and may prevent further use of such recordings. The speech enhancement methods are constantly being developed, but most of them use English data sources and try to remove non-stationary noise and reverberation [44, 45]. Our approach is unique in that we build the speech corpora explicitly using the Polish language and introduce unpredictable sound events as an additional layer of noise, in addition to the acoustic scene signals and reverberation. The script we have published allows generating any number of such simulated real-world noisy speech recordings based on publicly available data sources. The corresponding components of speech recordings, such as clean speech, events, acoustic scenes and reverberation, are saved in separate files that prepare the created corpora for training deep models of Polish speech enhancement in noisy, reverberated and disturbed acoustic environments. Additionally, adding a phase-inverted version of one of the corresponding components to a simulated speech cancels that component from the file. This feature makes the created corpora higly customizable and easily adjustable for various other problems like scene and event recognition or dereverberation.

Using the script, we created a corpus containing over 277 hours of training examples and over 5.5 hours for testing purposes. Using the data generated by the script, we trained three models. The first enhances single speech by separating it from complex noise. The second model is used for speech separation when two speakers occur concurrently against a background noise. The third model separates two speakers, which occurs without any noise. We evaluated our models using the prepared test set and the Libri2Mix [4]. We also compared the performance of our models with other pre-trained solutions.

2. Corpus for Polish Speech Enhancement

Deep learning models for speech enhancement require a large number of noisy recordings and their corresponding clean speech signals for training. However, obtaining clean speech signals can be difficult and expensive, often requiring access to professional recording studios. To address this problem, we have developed a script [1] that generates noisy speech corpora from the corresponding clean speech signals, which makes it ready to be used for training deep models. In the script, we use publicly available high quality data sources to create the corpora. The speech data sources are in Polish, and each generated audio file comprises a mix of seven distinct layers, as shown in Figure 1.

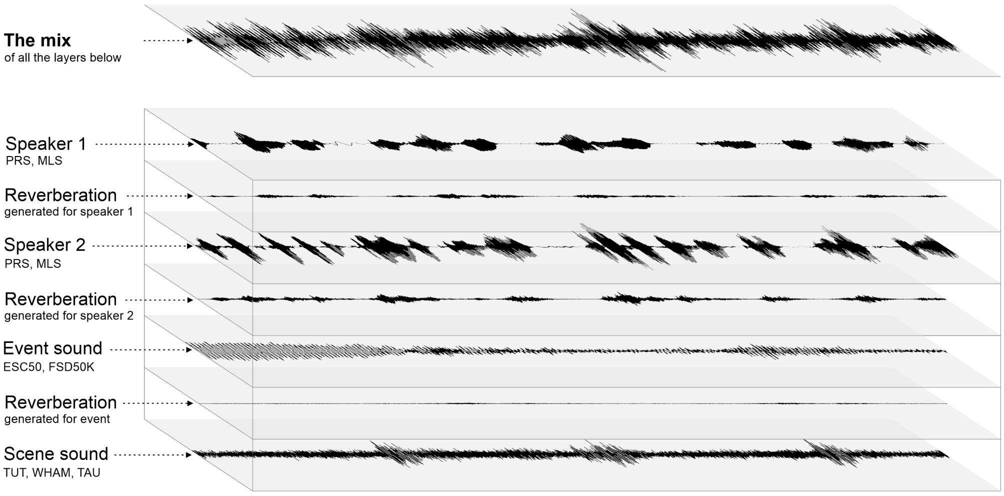

Fig. 1. The structure of a simulated speech recording generated by our script. The first four layers consist of two speech signals and the corresponding reverberation signals, based on room impulse responses. The speech signals are sourced from the Polish Read Speech corpus (PRS) [18] and the Polish section of The Multilingual LibriSpeech (MLS) [32]. Layers five and six contain additional sound events and their corresponding reverberation. The recordings of sound events are sourced from the ESC50 [31] and FSD50K [5] datasets. The final layer of the mix comprises the sound of the environmental scene taken from TUT Acoustic Scenes 2017 [26], TAU Urban Acoustic Scenes 2019 [25], and the WHAM [42] datasets.

Each layer of the mix is saved as a separate audio file in its designated folder. This allows for easy cancellation of a component from the mix by adding its phase-inverted version. It is possible to adjust the created corpora to a specific task using this principle. For instance, one speaker can be removed from the mix by adding its phase-inverted version. The same should be done with the reverberation for this speaker in such a case to remove this speaker completely. Therefore, it is easy to obtain the corpus with one speaker instead of the two without the need to generate the corpus again. Other possibilities are also feasible; exemplary ideas are presented in Table 1.

Our goal was to develop a solution that is accessible to everyone to stimulate the research in this field. To achieve this, each layer of the mix required to contain publicly available, high-quality data sources. In our quest for data sources to use in particular layers of the mixes, we provide an overview of the most popular data sources in the following section.

[1] `https://github.com/mklec/PolSMSE`

Table 1. Possible corpus adjustments to make it suitable for training different speech enhancement or speaker separation problems. This can be achieved by cancelling a given component from the mix without generating the new corpus from scratch. Each column shows the layers left in a mix to solve a particular problem. For instance, one can remove one speaker while leaving another, or cancel out all reverberation. Other examples include cancelling the background scene while retaining only the sound events, with or without reverberation, etc.

Layers of the mix	Speaker separation							Speech enhancement				
Speaker 1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaker 2	✓	✓	✓	✓	✓	✓	✓					
Scene sound	✓			✓		✓		✓	✓	✓		
Event sound	✓	✓		✓	✓			✓	✓		✓	
Reverb for speaker 1	✓	✓	✓					✓				✓
Reverb for speaker 2	✓	✓	✓									
Reverb for event	✓	✓						✓				

2.1. Review of Existing Data Sources

Many data sources are released when scientific challenges and workshops are organized, but most of them are in English and only occasionally provide sources of noisy and corresponding clean data. For example, the CHiME [2] challenge frequently releases new datasets. The CHiME-5 [3] provides conversational speech recordings in everyday home environments, but is intended to advance ASR performance rather than improve speech enhancement. On the other hand, the Deep Noise Suppression challenge (DNS) [34] focuses on improving overall speech quality when recorded in a challenging background. The dataset from this challenge contains over 10000 hours of noise and over 1000 hours of clean speech signals for training the noise suppression models and a representative test set of real-world scenarios consisting of both synthetic and real recordings.

The Wall Street Journal (WSJ0) corpus provides a speech source for creating other corpora. WSJ0-2Mix [10] takes two speakers and mixes them at speech-to-speech ratios (SSR) between 0 and 5 dB. It provides 30 hours of training examples. However, WSJ0 is not open-sourced. Therefore, the Libri2Mix [4] dataset is an alternative to WSJ0-2Mix. It is based on LibriSpeech [28], and consists of mixes of two speakers combined with ambient noise sampled from the WHAM [42] dataset. It provides 212 hours of training examples. The WHAMR dataset [24] extends the WSJ0-2Mix by noise from the WHAM and reverberation and mixes with the loudest speaker at SNR between -6 and 3 dB. It provides 58 hours of noisy and corresponding clean speech examples for training.

Other corpora can also provide valuable speech sources; however, their recordings often contain distortions and reverberation and are made with poor-quality microphones. Therefore, their usage for speech enhancement should include a selection of the best quality candidates. One example of such a data source is VoxCeleb [13], containing over 100000 utterances from more than 6000 speakers. Another example is the multi-language Common Voice dataset [2], which also contains the Polish subset of 177 hours of spontaneous speech. Librivox project [3] is another example of a multi-language source of speech data. It contains recordings of volunteers reading over 10000 public-domain books in various languages, also in Polish. The Polish clean and high-quality recordings are available from the Polish Read Speech corpus (PRS) [18]. It provides 56 hours of recordings featuring phonetically rich words and sentences spoken by 317 speakers in an acoustically treated recording studio. The Multilingual LibriSpeech (MLS) [32] is a multilingual dataset of read books (audiobooks), and the Polish section contains 137 hours

[2] `https://www.chimechallenge.org/`

[3] https://librivox.org/

of clean speech from 16 speakers reading 25 books. Finally, there are over 140 corpora available in the Common Language Resources and Technology Infrastructure (CLARIN) from the Polish language [11, 30].

As for the environmental scene and noise data sources, Audioset [6] provides a collection of about 2 million human-labeled 10s sound clips extracted from YouTube videos, which belong to about 600 audio classes. The TUT Acoustic Scenes 2017 dataset [26] includes 52 hours of audio recordings from 15 locations, such as homes, city centres, forest paths, grocery stores, metro stations, and more. The WHAM dataset [42] provides 80 hours of background noise from urban environments, such as restaurants, bars, cafes, and parks. The TAU Urban Acoustic Scenes dataset 2019, introduced in [25], features 40 hours of audio recordings from various urban acoustic scenes such as pedestrian streets, trams, and airports.

The sound events can also be downloaded from the ESC50 [31], which comprises 2000 recordings, each five seconds long, classified into 50 semantic classes such as clapping, vacuum cleaners, fireworks, and more. The FSD50K [5] contains over 50000 audio clips, amounting to more than 100 hours of audio, organized into 200 classes drawn from the AudioSet Ontology. Examples of these classes include various musical instruments, splashes, zippers, telephones, and many others.

2.2. The Corpus Creation

We considered factors such as license, accessibility, language, and data quality when selecting the sources for our solution. Ultimately, we chose two spoken Polish recordings sources, PRS and MLS, and obtained the sounds of real-world environments from the TUT, WHAM, and TAU datasets. Sound events were sourced from the ESC50 and FSD50K datasets. Finally, we created simulated recordings of noisy speech by combining signals from the aforementioned sources and applying reverberation, using the following formula.

$$y(t) = \sum_{i=1}^{C=2} (s_i(t) + r_i(t)) + b(t) + e(t) + r_e(t) \tag{1}$$

where $y(t)$ represents the simulated speech recording in the time domain with a maximum of C speakers, $C = 2$ in our case. The signal $s_i(t)$ represents the ith speaker in the mix, which is mixed with their corresponding room response $r_i(t)$, generated earlier as reverberation. The signal $b(t)$ denotes the scene's ambient sound, and $e(t)$ represents a non-speech sound event, along with its corresponding reverberation $r_e(t)$. In this context, the background noise refers to the sum of four components: $r_i(t)$, $b(t)$, $e(t)$, and $r_e(t)$.

First, we divided all files from different sources into three subsets: training, validation, and testing, according to the instructions provided by each data source. This ensures that no mix component in the training subset is used for testing or validation. Next, we excluded files with speech-related events, such as whispering or singing, from the event sources, to avoid conflicts with speech layers. Figure 2 shows the remaining event classes used for creating our corpus. The two speakers were mixed at a speech-to-speech ratio that is randomly selected from -5 to 5 decibels, rounded to the nearest whole decibel. The resulting mix includes a randomly selected 4-second fragment from the given source files, making the final mixes always different.

In order to recreate the actual recording's conditions as much as possible we took into account several characteristics of them. To ensure that the selected sound events are suitable for a particular scene, we created a matrix that maps scene classes to possible event classes. This step helps prevent the random selection of mismatched events, such as the sound of a cow in an airport, which would be absurd. The matrix provides guidelines for the event and specific scene classes when mixed with the script. Additionally, reverberation was generated in Matlab only when the scene class represents an indoor category. We manually defined a dictionary with possible reverb parameters range for these classes to ensure that the characteristics of generated room reflections are suited to the particular scene class. The reverb parameters were randomly selected each time, but only from such predefined range. This approach allows us to avoid generating unsuitable reverbs, such as long reverbs for the library class, which typically has a short decay time. We controlled parameters like decay time, reflection diffusion, strength of high-frequency damping, and early reflection time in this manner. The same values of these parameters were applied to the speech and event sources, except with different early reflection times for events, as these two sound sources usually have different placements in the recording room, affecting the time when the microphone captures their first reflection.

We created a training subset of 250000 files of $y(t)$ using our script. The testing and validation subsets contain 5000 files each. The duration of each file is 4 seconds, and the mix layers are saved in separate files, at 8000 Hz and 16 bits. This setting is justified by using the same values in [23]. However, generating the files encoded in 16 kHz is also possible by the script. The training subset provides over 277 hours (1 million seconds) of continuous noisy speech along with the corresponding clean sources and other layers for training purposes. This corpus is called PolSMSE-2-Noisy and is intended to separate speakers into two channels when they speak simultaneously over a noisy background, i.e., to separate noisy signals (SN). Therefore it contains recordings for all layers shown in Fig. 1. Next, we used the phase-inversion phenomenon to cancel out one speaker and their reverberation, creating another corpus called PolSMSE-1-Noisy. This corpus is intended to enhance single speech (ES), containing only one speech signal over the noisy background. The noisy background refers to all the other layers of the mix, except the speech.

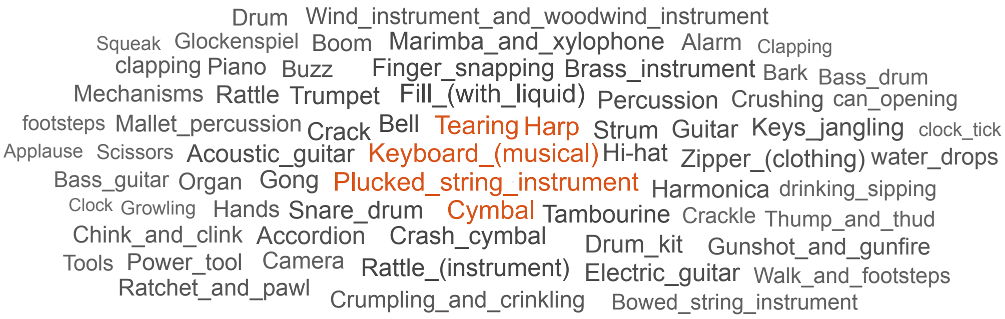

Fig. 2. The word cloud showing the classes of events used in the created corpus. For the sake of clarity, the least common events have been excluded. Four of the most frequent event classes have been highlighted in red. The most common event classes are related to musical instruments.

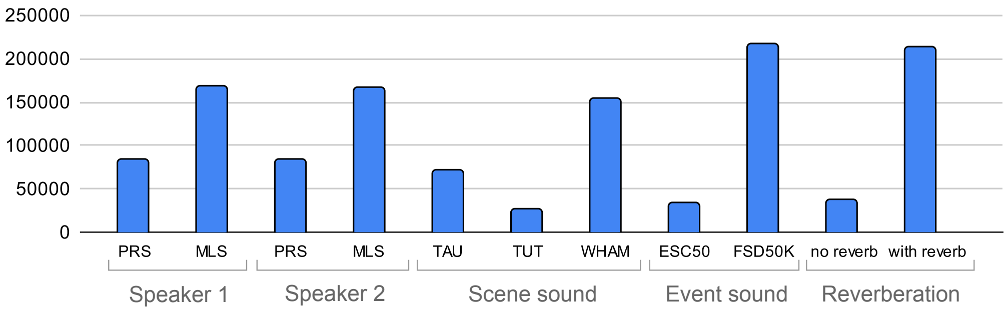

Fig. 3. The histogram illustrates each data source's contribution to creating layers of the PolSMSE corpus. The y-axis denotes the number of training examples that contain a given data source.

3. Experiments

Our research aimed to illustrate the feasibility of developing a corpus according to the ideas presented in Section 2.2 and utilizing it in practical experiments involving speech separation and enhancement through deep learning. To the best of our knowledge, no other speech corpora exhibit such a complex and noisy recording environment. The published script enables the recreation of the testing subset, ensuring the comparison of the results and hopefully achieving state-of-the-art results with our testing subset in future work.

We utilized the Conv-TasNet architecture initially designed for speech separation [23]. Conv-TasNet uses a linear encoder to generate a representation optimized for learning masks for the speakers. Next, the mask is applied to the encoder output, which is then inverted back to the waveforms, representing the speech of separated speakers. The network finds the masks using a temporal convolutional network (TCN) consisting of stacked 1-D dilated convolutional blocks. We reduced the number of filters in the encoder and decoder from 512 to 256, which decreased the model's capacity, but significantly accelerated the calculations. Other hyperparameters of the network are: the length of the input filters equal to 20 samples, and 32 convolutional blocks in the TCN. The loss functions and training procedure followed [23], with the initial learning rate set to 1e-3 and Adam used as the optimizer. If the validation set's accuracy did not improve over two consecutive epochs, the learning rate was halved. The models were implemented in Matlab and trained from scratch for fifteen epochs on a single GPU GeForce RTX 3080.

We trained three models using PolSMSE to estimate $s_i(t)$ from $y(t)$ (see Equation 1). The first model was trained to enhance the speech (ES) using PolSMSE-1-Noisy, containing a single speaker in complex background noise. The second model was trained to separate two speakers conversing in a noisy environment and events (SN) using PolSMSE-2-Noisy, containing all the layers of speech and noise depicted in Figure 1. The third model was trained to separate two clean speeches (SC) using a mixture of two clean speakers without any noise, events, or reverberation. This version of the dataset will be referred to as PolSMSE-2-Clean. Figure 4 illustrates the objective of training the models based on the provided input data and desired output.

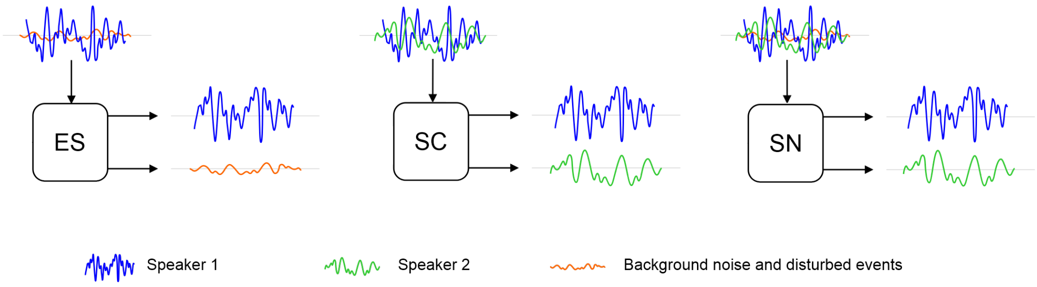

Fig. 4. The inputs and target outputs we utilized to train the experimental models. These models were created to address the following problems: improving the quality of single-speaker speech (ES), separating speech signals with noisy backgrounds (SN), and isolating clean speech signals without any interference (SC).

In evaluating the models' performance, we used the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [19]. Higher SI-SDR values indicate better separation quality. Besides, we used other pre-trained models and datasets to provide a more comprehensive evaluation. The first pre-trained model, SepFormer [38], was trained for speaker separation using Libri2Mix-Clean from SpeechBrain [33]. It achieved a 20.6 dB SI-SNR with this dataset. Additionally, the second pre-trained model, referred to as CTNoar [40], is the Conv-TasNet network pre-trained with 2-speaker mixtures from the WSJ0-2mix datasets, containing clean speech signals without noise. It achieved a 14.6 dB SI-SNR with this dataset.

4. Results and Discussion

Table 2. The dB values of SI-SDR representing the performance of three Conv-TasNet models trained using the PolSMSE and tested using both the PolSMSE and LibriMix testing subsets. These datasets come in two versions: clean and noisy. The clean version comprises a mixture of clean speech signals devoid of noise, events, or reverberation. The first model aims to separate two speakers in their respective channels when they are present amidst complex noise (SN). The second model is designed to separate two clean speech signals (SC). The third model was trained to isolate single speech from noise (ES). The table also includes other pre-trained networks: SepFormer, pre-trained with Libri2Mix-Clean from SpeechBrain [33], and another Conv-TasNet (CTNoar) [40], trained with clean WSJ0-2mix. Best results are shown in bold.

Speech separation				Single speech enhancement		
Testing subset	SN	SC	SepF	Testing subset	ES	CTNoar
PolSMSE-2-Noisy	**1.69**	-3.55	-0.48	PolSMSE-1-Noisy	**10.24**	-1.39
PolSMSE-2-Clean	5.28	6.64	**18.25**	Libri1Mix-Noisy	**11.97**	-0.64
Libri2Mix-Noisy	4.59	1.73	**6.84**	-	-	-
Libri2Mix-Clean	6.71	7.85	**20.56**	-	-	-

The preliminary results summarized in Table 2 suggest that the ES model effectively enhances single speech by isolating it from complex noise, disturbing sound events, and reverberation, even with a short training time limited to fifteen epochs. The results show almost 12 dB of SI-SDR for Libri1Mix-Noisy and over 10 dB for PolSMSE-1-Noisy, underscoring the proposed solution's potential. It is important to note that these datasets contain speech mixed with noise. However, the Libri1Mix-Noisy contains English speech without reverberation or disturbing sound events, explaining slightly better results in this case.

Despite the short training time, the SN model, trained with PolSMSE-2-Noisy, outperforms SepFormer, achieving an SI-SDR of 1.69 dB compared to -0.48 dB. These results indicate that the created corpus offers valuable data for addressing the speaker separation problem, especially in challenging and complex noise environments. The difference in the results also highlights the limitations of models trained with clean speech signals, as SepFormer was trained with Libri2Mix-Clean. The reported SI-SDR result of 20.56 dB for SepFormer with Libri2Mix-Clean aligns with previous research results [39].

In our evaluation of single speech enhancement, we compared our model to the CTNoar, which had been originally trained on a mix of 2 and 3 speech signals to separate one signal and put the others in a separate channel. The hypothesis was that the CTNoar model could identify one speaker and separate the noise into a distinct channel when tested with a noisy dataset. However, the results showed negative SI-SDR values for PolSMSE-1-Noisy and Libri1Mix-Noisy, indicating that this model was ineffective at enhancing single speech in a noisy environment. Our ES model successfully addresses this issue, achieving 10.24 dB for PolSMSE-1-Noisy and 11.97 dB for Libri1Mix-Noisy.

The research detailed in [16] demonstrated that a single-channel time-domain denoising technique could reduce the word error rate (WER) by 30%. These findings motivate us to enhance our models in the future by integrating more data sources, generating additional training examples, increasing the model's capacity, and extending the model training time. Furthermore, the speech data sources should include transcriptions to help other researchers evaluate their models with ASR in terms of WER. Currently, only the MLS data source contains the transcription. Another potential source is the newly released Polish speech corpus discussed in [30], which offers transcriptions and conversational Polish speech, complementing the supervised speech recordings used in the current study.

Turning off specific training layers, as presented in Table 1, may also help address other problems, such as event recognition or dereverberation in complex and non-stationary background noise. Additionally, our future research will further explore this specific feature of the proposed solution, by investigating the effect of particular noise layers on the performance of speech separation, enhancement, and recognition tasks. The corpus can also be utilized to train voice activity detection, which is crucial for effectively operating in unpredictable and noisy environments, for example, recognizing speech in cars or voice-controlled machines [27, 35]. Accurately distinguishing between noise and speech is essential in such scenarios.

5. Conclusion

This paper addresses the challenges of enhancing and separating speech from noisy, reverberant, and disrupted backgrounds, especially in the context of Polish speech recordings. The publicly available, customizable, and scalable corpora generated by the proposed data generation script can be valuable for training deep models and facilitating further research. We hope the proposed solution will contribute to developing new models that leverage the layers' interrelation in the data and potentially set a new state-of-the-art for separating Polish speech from complex noise with disruptions and reverberation.

References

1. AlBadawy, E.A., Gibiansky, A., He, Q., Wu, J., Chang, M.C., Lyu, S.: Vocbench: A neural vocoder benchmark for speech synthesis. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 881–885. IEEE (2022)
2. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4218–4222 (2020)
3. Barker, J., Watanabe, S., Vincent, E., Trmal, J.: The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. In: Interspeech 2018-19th Annual Conference of the International Speech Communication Association (2018)
4. Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., Vincent, E.: Librimix: An open-source dataset for generalizable speech separation (2020)
5. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, pp. 829–852 (2021)
6. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
7. Guimarães, H.R., Nagano, H., Silva, D.W.: Monaural speech enhancement through deep wave-u-net. Expert Systems with Applications 158, pp. 113582 (2020)
8. Hao, X., Su, X., Horaud, R., Li, X.: Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6633–6637. IEEE (2021)
9. Hasumi, T., Kobayashi, T., Ogawa, T.: Investigation of network architecture for single-channel end-to-end denoising. In: 2020 28th European Signal Processing Conference (EUSIPCO). pp. 441–445. IEEE (2021)
10. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative

embeddings for segmentation and separation. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 31–35. IEEE (2016)

11. Hinrichs, E., Krauwer, S.: The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) pp. 1525–1531 (May 2014), http://dspace.library.uu.nl/handle/1874/307981
12. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L.: Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264 (2020)
13. Huh, J., Brown, A., Jung, J.w., Son Chung, J., Nagrani, A., Garcia-Romero, D., Zisserman, A.: Voxsrc 2022: The fourth voxceleb speaker recognition challenge. arXiv e-prints pp. arXiv–2302 (2023)
14. Kagami, H., Kameoka, H., Yukawa, M.: Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 31–35. IEEE (2018)
15. Kashyap, M.M., Tambwekar, A., Manohara, K., Natarajan, S.: Speech denoising without clean training data: A noise2noise approach. arXiv preprint arXiv:2104.03838 (2021)
16. Kinoshita, K., Ochiai, T., Delcroix, M., Nakatani, T.: Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7009–7013. IEEE (2020)
17. Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y., Takeuchi, D.: Speech enhancement using self-adaptation and multi-head self-attention. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 181–185. IEEE (2020)
18. Koržinek, D., Marasek, K., Brocki, Ł.: Polish read speech corpus for speech tools and services. In: CLARIN Annual Conference 2016 in Aix-en-Provence, France (2016)
19. Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: Sdr–half-baked or well done? In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 626–630. IEEE (2019)
20. Lee, D., Kim, S., Choi, J.W.: Inter-channel conv-tasnet for multichannel speech enhancement. arXiv preprint arXiv:2111.04312 (2021)
21. Lim, T.Y., Yeh, R.A., Xu, Y., Do, M.N., Hasegawa-Johnson, M.: Time-frequency networks for audio super-resolution. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 646–650. IEEE (2018)
22. Luo, Y., Chen, Z., Yoshioka, T.: Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 46–50. IEEE (2020)
23. Luo, Y., Mesgarani, N.: Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing 27(8), pp. 1256–1266 (2019)
24. Maciejewski, M., Wichern, G., McQuinn, E., Le Roux, J.: Whamr!: Noisy and reverberant single-channel speech separation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 696–700. IEEE (2020)
25. Mesaros, A., Heittola, T., Virtanen, T.: A multi-device dataset for urban acoustic scene classification. In: Scenes and Events 2018 Workshop (DCASE2018). p. 9

26. Mesaros, A., Heittola, T., Virtanen, T.: Acoustic scene classification: an overview of dcase 2017 challenge entries. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). pp. 411–415. IEEE (2018)
27. Mihalache, S., Burileanu, D.: Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection. Sensors 22(3), pp. 1228 (2022)
28. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
29. Pascual, S., Bonafonte, A., Serrà, J.: Segan: Speech enhancement generative adversarial network. Interspeech 2017 (2017)
30. Pęzik, P., Karasińska, S., Cichosz, A., Jałowiecki, Ł., Kaczyński, K., Krawentek, M., Walkusz, K., Wilk, P., Kleć, M., Szklanny, K., et al.: Spokesbiz–an open corpus of conversational polish. arXiv e-prints pp. arXiv–2312 (2023)
31. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)
32. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R.: Mls: A large-scale multilingual dataset for speech research (2020)
33. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al.: Speechbrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624 (2021)
34. Reddy, C.K., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., Matusevych, S., Aichner, R., Aazami, A., Braun, S., et al.: The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results (2020)
35. Rho, D., Park, J., Ko, J.: Nas-vad: Neural architecture search for voice activity detection. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 2022, pp. 3754–3758. International Speech Communication Association (2022)
36. Satheesh, A., Muthu-Manivannan, K.: Denoising speech signals with hifi-coulomb-gans. Journal of Student Research 11(3) (2022)
37. Scalart, P., et al.: Speech enhancement based on a priori signal to noise estimation. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. vol. 2, pp. 629–632. IEEE (1996)
38. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 21–25. IEEE (2021)
39. Subakan, C., Ravanelli, M., Cornell, S., Grondin, F., Bronzi, M.: Exploring self-attention mechanisms for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)
40. Takahashi, N., Parthasaarathy, S., Goswami, N., Mitsufuji, Y.: Recursive speech separation for unknown number of speakers. Interspeech 2019 (2019)
41. Upadhyay, N., Karmakar, A.: Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. Procedia Computer Science 54, pp. 574–584 (2015)
42. Wichern, G., Antognini, J., Flynn, M., Zhu, L.R., McQuinn, E., Crow, D., Manilow, E., Roux, J.L.: Wham!: Extending speech separation to noisy environments. arXiv preprint arXiv:1907.01160 (2019)
43. Yang, G.P., Tuan, C.I., Lee, H.Y., Lee, L.s.: Improved speech separation with time-and-frequency cross-domain joint embedding and clustering. arXiv preprint arXiv:1904.07845 (2019)

44. Zeghidour, N., Grangier, D.: Wavesplit: End-to-end speech separation by speaker clustering. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, pp. 2840–2849 (2021)
45. Zhao, S., Ma, Y., Ni, C., Zhang, C., Wang, H., Nguyen, T.H., Zhou, K., Yip, J.Q., Ng, D., Ma, B.: Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10356–10360. IEEE (2024)
46. Zhao, S., Nguyen, T.H., Ma, B.: Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6648–6652. IEEE (2021)