Deep Learning System Based on the Separation of Audio Sources to Obtain the Transcription of a Conversation

Nahum Flores, Daniel Angeles, and Sebastian Tuesta
Faculty of System Engineering and Informatic, Universidad Nacional Mayor de San Marcos, Lima, Peru
Email: {nahum.flores, daniel.angeles, sebastian.tuesta}@unmsm.edu.pe

***Abstract*—Podcasting has lately been in the spotlight for being the fastest-growing format, especially during the pandemic. This growth has highlighted the need for making podcasts accessible to diverse audiences, especially those having auditory disabilities. The current transcription methods have been unsatisfactory; therefore, we present an alternative method to transcribe audio files into text by segmenting audio sources. The applied methodology considers the construction of a public audio dataset having a duration of more than 15h. The training model was based on three scenarios in which the duration of the training data was varied to determine the best performance, which was 10.77 in terms of the scale-invariant signal-to-noise ratio. We have simplified podcasting accessibility by making available the source code of each component that we developed.**

***Index Terms*—public dataset, deep learning, audio source separation, speech to text**

I. Introduction

From 2014 to 2018 podcasts grew to 122% in listening time; this has been the most dramatic growth for any type of audio [1]. The number of persons listening to podcasts continues to rise. This increase has been prominent during the pandemic when 37% of Americans (aged 12 years and older) listened to at least one podcast each month; this was a 32% increase over the numbers of 2019. An estimated 125 million people will listen to a podcast each month by 2025 [2].

Spotify, one of the most popular streaming platforms, had more than 345 million monthly active users in the last quarter of 2020 [3]. In its podcasts section, Spotify has more than 1.9 million titles [4]. This huge popularity continues to bring with it an enormous wave of new amateur podcasters who can produce podcasts using the minimum possible tools such as a microphone and a recording device.

The growth of podcasting has raised concerns regarding the relationship between podcasting and its accessibility for a sensory diverse audience, especially persons with hearing and visual impairments. These resources can and should be made accessible to them also [5].

Manuscript received December 14, 2021; revised May 11, 2022.

Many podcast platforms have a section to attach a chapter transcription of a specific program. In many cases, authors themselves publish these transcripts, which can be tedious and time-consuming for them. The alternatives to performing a manual transcription are to use an online paid service for automatic transcription, such as Otter or Temi. However, the audio could include noises or interferences (common in amateur productions) because of the poor quality of the recording devices, or the disturbances present in the recording spaces. In such cases, this automatic transcription software might not work satisfactorily.

Another alternative is to hire a manual transcription service in Spanish [6]. It can cost between 0.75 and 1.75 euros per minute plus taxes, which becomes too expensive for amateur podcast producers. Many effective methods have been proposed to deal with the source-voices separation problem, but none of these solutions have been completely effective because they were designed to manage a particular interfering signal [7]. The separation methods still have limitations when applied to real-world problems.

II. Previous Works

Artificial intelligence (AI) is defined as the field of informatics dedicated to the creation of systems that simulate human intelligence [8], [9]. Machine Learning (ML) is a branch of AI in which a computer system can learn to perform a task (e.g., classify, detect, segment, or predict) from the patterns obtained from its input data [10], [11]. Likewise, Deep Learning (DL) is a branch of ML that incorporates computational algorithms that mimic the biological structure of the brain [12], [13]. One type of DL is a Convolutional Neural Network (CNN) [14], which is a set of processing layers that resemble the physiological processes present in the animal visual cortex.

In 2017, with the release of the public dataset MUSDB18 [15], various groups (ranging from independent researchers to corporate groups) have been intensively investigating the Audio Source Separation (ASS) technology. In three years, models such as Open-Unmix [16], D3NET [17], Wave-U-Net [18], MMDenseLSTM [19], and Demucs [20], have been developed, which have helped achieve increasingly better

doi: 10.12720/jait.13.4.393-397

performances in terms of the Signal-to-Noise Ratio (SNR) (see Table I).

Substantial performance improvement is also seen in a particular case of ASS termed voice separation (VS). VS includes Svoice [21], DPRNN [22], FurcaNeXt [23], CBLDNN-GAT [24], DPCL++ [25], and ADANet [26]. These DL models were trained using the WSJ0 dataset [27]. Currently, Svoice is the model with the best performance (see Table II).

TABLE I. PERFORMANCE OF ASS MODELS

Year	Model	SDR
2020	Demucs	6.8
2020	Open-Unmix	5.3
2020	D3Net	6.7
2020	Spleeter	5.9
2018	Wave-U-Net	3.2
2018	MMDenseLSTM	6.0

TABLE II. PERFORMANCE OF VS MODELS

Year	Model	SI–SNR
2020	Svoice	20.1
2020	DPRNN	18.8
2019	FurcaNeXt	18.4
2018	CBLDNN-GAT	11.0
2018	DPCL++	10.8
2018	ADANet	10.5

In parallel, we have the algorithms that convert audio to text, which when combined with VS gives a solution to the problem of podcasting accessibility. To do this, we relied on the work of Këpuska [28], who chose Google's speech-to-text Application Program Interface (API), to transcribe the audio into text. However, at the start of our research, we faced the problem of nonavailability of a public dataset for VS. Therefore, in this paper, we also present the new public domain dataset DANF-VOICE with which Svoice (the model with the current best performance) will be trained.

III. MATERIALS AND METHODS

A. Dataset

For the construction of the dataset, 258 people participated (121 men and 137 women); they sent three-minute WhatsApp audio clips recorded in noise-free environments. When we finished compiling the audio clips, we selected 189 good quality audio clips (from 91 men and 98 women). The use of multiple sources guaranteed greater similarity to real-life situations.

To complete the dataset, 111 audio clips (from 59 men and 52 women) were extracted from the public discussions and debates of the parliamentarians of Peru, Argentina, Chile, and Spain. Therefore, we used a total of 300 audio clips extending over a duration of 15h. We included both male and female participants between 15 and 70 years of age. Our dataset is available at the following URL: https://github.com/NahumFGz/DANF-VOICE.

B. Preprocessing

To improve the quality of the dataset, each of the audio clips underwent the following six processes:

- **Time trimming**: The audio clips that exceeded three minutes in length were trimmed.
- **Audio normalization**: The audio amplitudes were leveled so that they were in the range from −3 db to −6 db.
- **Frequency clipping**: A low pass filter was applied to eliminate the entire frequency range that corresponded from 0 Hz to 180 Hz. In certain cases, where the voice was too powerful, we trimmed up to 240 Hz.
- **NoiseGate application**: The NoiseGate technique was applied to eliminate all the signals that were below −18 db.
- **Audio conversion**: As the last step, all the audio clips were converted to the WAV format.

After they were standardized in the WAV format, the audios of the participants were combined for the conversation simulations. Each audio file was kept in a folder according to its origin. Finally, we had three folders. One folder contained the audio clips of the first interlocutor. The second folder contained the audio clips of the second interlocutor. The third folder contained the audios in which the conversations were simulated with those having combinations of both voices. The simulated conversations of the DANF-VOICE data set were distributed in 50 conversations for each type of combination (female-female, male-male, and female-male).

C. Train

Three scenarios were considered to determine the efficiency of the models according to the size of the dataset (see Table III). The first scenario had simulated conversations with an audio time range of 20 s to 40 s and a sampling frequency of 4800 hz. In the second scenario, the time range per audio was increased to 25–45 s, and the sampling frequency was decreased to 8000 hz, which was six times lower than the frequency in the previous scenario. In the third scenario, we considered a range of 25–50 s by maintaining a sampling frequency of 8000 hz.

TABLE III. SVOICE TRAINING SCENARIOS

Scenario	Duration range per audio	Number of conversations	Number of audio clips	Overall time
1	20–40 s	150	450	13,501 s
2	25–45 s	150	450	15,897 s
3	25–50 s	150	450	16,812 s

For the training of the DL models, we prepared a set of experiments in which the training parameters were modified to obtain the best performance. In addition, they were based on the approach of the scenarios and were

executed sequentially following the scheme in Fig. 1. Table IV describes the parameters of the seven experiments that we performed.

TABLE IV. FORMULATED EXPERIMENTS

Experiment	Lr	Lr scheduling	Batch size	Scenario
Exp 1	5e-4	Plateau	4	1
Exp 2	1e-3	Step	4	1
Exp 3	5e-4	Step	4	1
Exp 4	7e-4	Step	4	1
Exp 5	5e-4	Plateau	4	2
Exp 6	5e-4	Step	4	2
Exp 7	5e-4	Plateau	4	3

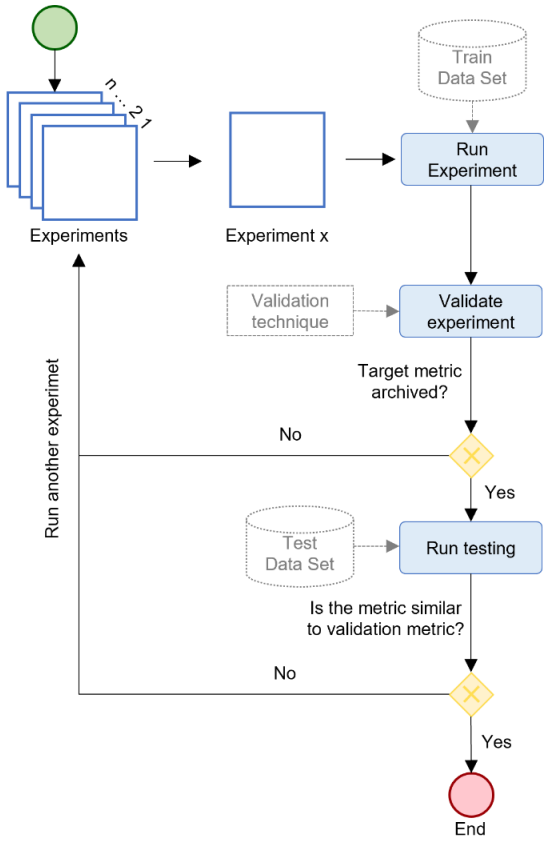

Figure 1. Execution flow for the experiment.

D. Validation

To measure the performance of VS, we considered the metric scale-invariant SNR [29] (SI-SNR); this parameter has been commonly used as the evaluation metric for source separation and is defined as follows:

$$\begin{cases} S_{target} = \dfrac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} = \hat{s} - S_{target} \\ SISNR = 10 \log_{10} \dfrac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \quad (1)$$

E. Results

This section presents the results obtained when training the model with the DANF-VOICE dataset under the three scenarios. In addition, Fig. 2 shows the comparison of the SI–SNR obtained in the training of each epoch. Clearly, experiment 7 gives the best results. Therefore, its loss curve is shown in Fig. 3.

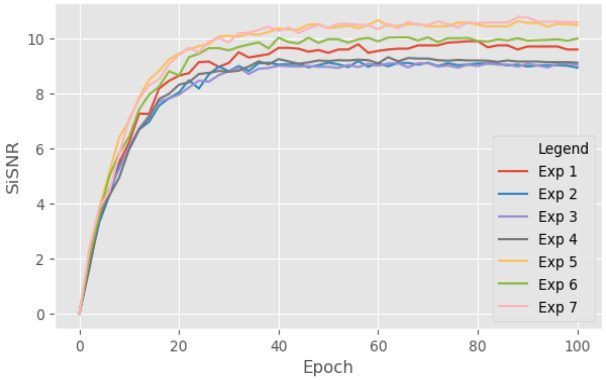

Figure 2. Results of each experiment.

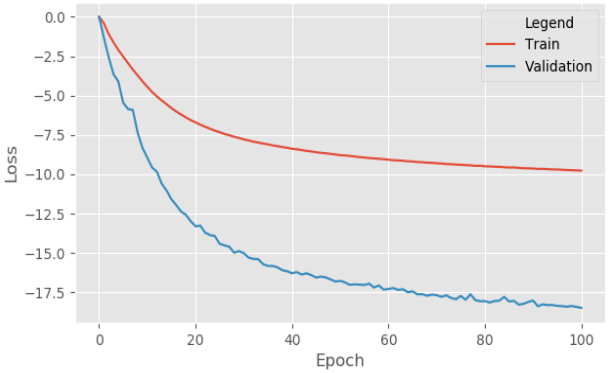

Figure 3. Loss curve for experiment 7.

F. Tool Architecture

For the development of the solution, we used two main components: the frontend part (which is in React) and the backend part (which is in Python and uses the Flask library for API development) (see Fig. 4).

Figure 4. Backend architecture.

G. Implementation

An API solution needs a server where its code is stored and where its execution occurs each time it is called. In the current scenario, it is not necessary to buy or rent a physical server; it is possible to use the existing infrastructure as a service solution in the market. For this solution, we used the Amazon Web Services (AWS) provider and Google Cloud platform, as shown in Fig. 5.

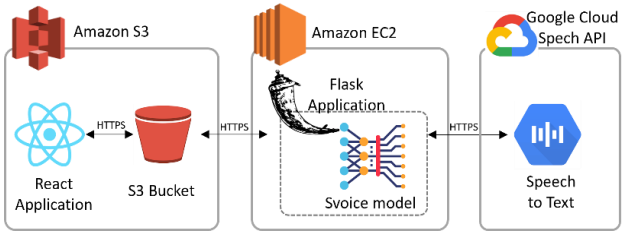

Figure 5. Architecture of the proposed solution.

Fig. 6 shows the page with which the user interacts to use the application (http://aws-podcastext-frontend.s3.us-east-2.amazonaws.com/index.html); this is the main view of the tool.

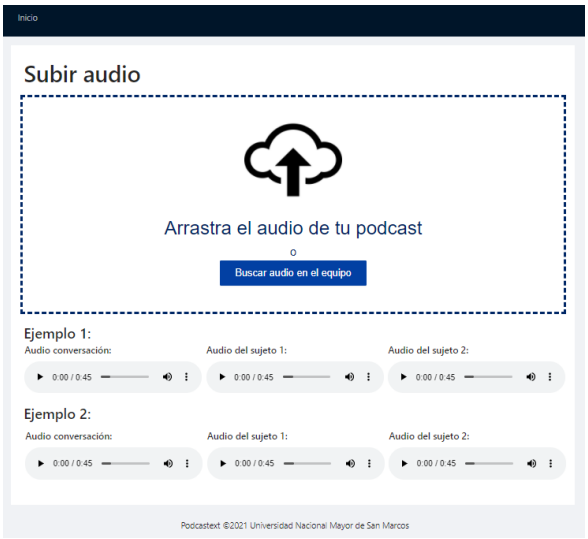

Figure 6. Main view of the tool.

IV. Discussion

A. About the Dataset

A new public dataset is presented in this research. It consists of 300 audio clips; this dataset will allow the elimination of the entry barrier for future research and development of VS models.

B. About the Results

The results are not as good as those obtained for the Svoice trial because we faced hardware limitations. We used an Nvidia GTX 2080ti video card with 11 GB of RAM, which did not allow us to train the model with the entire dataset (only a third part was used). However, the performance improves by increasing the total duration of each dataset in each scenario. Therefore, we can guarantee that the model will increase its performance as we increase the time for which it is trained.

C. About the Solution

By using AWS for the implementation of the solution, anyone worldwide can access the application and test the model performance; they can avoid the tedious process of replicating and launching the local deployment of the model. In addition, the source code of each component that we developed for the solution has been made available; this is the first step to provide podcasting accessibility.

V. Conclusion

In this research, we used a new public dataset to train a deep learning model that obtained 10.77 SI-SNR using only one-third of the dataset. The best model from the 7 experiments was deployed in a Cloud solution so that the demo could be accessed from anywhere in the world.

The biggest challenges during the research were getting volunteers to collect the audios and the hardware limitations during the training, so we had to optimize the preprocessing of each audio as much as possible. In future work, we plan to work on a model that allows the transcription of the audios, so that we can continue contributing to the democratization of Speech to Text and podcasting can reach all people. In addition, we share all the results and the source code of the solution at https://github.com/NahumFGz/svoice_backend_cpu.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Nahum Flores implemented the model and ran experiments; Daniel Angeles built the dataset and developed the frontend, and Sebastian Tuesta deployed the algorithms and backend. All authors contributed to the research and approved the final version of the manuscript.

References

[1] Infinite Dial and Share of Ear. (April 2019). Edison/Triton: 'Consequential Year' for podcasting in new podcast consumer report. *Rain News*. [Online]. Available: https://rainnews.com/edison-triton-consequential-year-for-podcasting-in-new-podcast-consumer-report/

[2] Forbes. (August 2021). As podcasts continue to grow in popularity, ad dollars follow. Fobes. [Online]. Available: https://www.forbes.com/sites/bradadgate/2021/02/11/podcasting-has-become-a-big-business/?sh=24b08b0c2cfb

[3] M. Iqbal. (July 2021). Spotify usage and revenue statistics (2020). [Online]. Available: https://www.businessofapps.com/data/spotify-statistics/

[4] Spotify, "The trends that marked the streaming in 2020," *For the Record*, 2020.

[5] E. B. Pinheiro, "Podcast and accessibility: Theoricak and methodological notes," *Revista GEMInIS*, vol. 11, no. 2, pp. 45-66, 2020.

[6] Mis Transcripciones. (2021). [Online]. Available: https://www.mistranscripciones.es/

[7] H. D. Do, S. T. Tran, and D. T. Chau, "Speech source separation using variational autoencoder and bandpass filter," *IEEE Access*, vol. 8, pp. 156219-156231, 2020.

[8] P. Lakhani, *et al.*, "Machine learning in radiology: Applications beyond image interpretation," *Journal of the American College of Radiology*, vol. 15, no. 2, pp. 350-359, 2018.

[9] S. R. Bohannon, "Artificial intelligence. Fears of an AI pioneer," *Science*, vol. 349, no. 6245, pp. 252-252, 2015.
[10] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson, 2008.
[11] A. Géron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed., O'Reilly Media, Inc., 2017.
[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[13] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. 36-40, 2017.
[14] C. Zheng, T. V. Johnson, A. Garg, and M. V. Boland, "Artificial intelligence in glaucoma," *Current Opinion in Ophthalmology*, vol. 30, no. 2, pp. 97-103, 2019.
[15] Z. Rafii, A. Liutkus, F. R. Stöter, S. Mimilakis, and R. Bittner. (December 2017). The MUSDB18 corpus for music separation. [Online]. Available: https://zenodo.org/record/1117372
[16] F. R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.
[17] Y. Mitsufuji and N. Takahashi, "D3Net: Densely connected multidilated DenseNet for music source separation," arXiv:2010.01733v4, 2021.
[18] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th International Society for Music Information Retrieval Conference*, Paris, 2018.
[19] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th International Workshop on Acoustic Signal Enhancement*, Tokyo, 2018.
[20] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," hal-02379796v2f, 2021.
[21] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," arXiv:2003.01531v4, 2020.
[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," arXiv:1910.06379v2, 2019.
[23] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-End monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. International Conference on Multimedia Modeling*, 2020, pp. 653-665.
[24] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-Based speaker-independent speech separation via generative adversarial training," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 711-715.
[25] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," arXiv preprint arXiv:1607.02173, 2016.
[26] Y. Luo, Z. Chen, and N. Mesgarani, "Speakerindependent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787-796, 2018.
[27] J. Garofolo, D. Graff, D. Paul, and D. Pallett. (2007). CSR-I (WSJ0) complete. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A
[28] V. Këpuska, "Comparing speech recognition systems (Microsoft API, Google API And CMU Sphinx)," *International Journal of Engineering Research and Applications*, vol. 7, no. 3, pp. 20-24, 2017.
[29] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.

Nahum Flores received B.Sc. degree in Software Engineering at the Universidad Nacional Mayor de San Marcos (UNMSM). Currently he is working as a Data Analyst in a payment gateway. He also works as a Researcher in the Artificial Intelligence Group at UNMSM. He is always willing to learn new things with full enthusiasm and passion. He has experience working in Python, Keras, TensorFlow, Sklearn, and Scipy. His research interests include deep learning, image processing, computer vision, and natural language processing.

Daniel Angeles received B.Sc. degree in Software Engineering at the Universidad Nacional Mayor de San Marcos (UNMSM). Currently he is working as a Full Stack Developer. His current research interests are in the area of deep learning related to the investigation of deep generative models applied to music. He has experience working in Python, Keras, TensorFlow, and JavaScript.

Sebastian Tuesta received B.Sc. degree in Computer Science at the Universidad Nacional Mayor de San Marcos (UNMSM). Currently he is working as a Data Engineer. He has experience working in Python, Keras, AWS, and GCP. His research interests include deep learning, image processing, and computer vision.