WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deep learning based semantic communication system for speech transmission

Ch. Babji Prasad[1], M. Sivaji[2], P. Raj Vardhan Yadav[3], J.Venkata Sai Prasada Rao[4], K.Venu[5]

GMR INSTITUTE OF TECHNOLOGY

ABSTRACT—

A new deep learning-enabled semantic communica- tion system created especially for speech-based human-machine interactions is presented in this paper. The suggested system transmits semantic information instead of conventional signal waveforms, allowing for more reliable and effective communi- cation by utilizing sophisticated deep learning models for both speech recognition and synthesis. In the framework, meaning- rich features are extracted and transmitted from the spoken input by a semantic encoder at the transmitter, and intelligible speech signals are reconstructed from the received semantics by a semantic decoder at the receiver. Numerous tests show that this method, particularly in noisy channel environments, greatly lowers transmission bandwidth while preserving high speech accuracy .

Index Terms—Semantic communication, deep learning, speech recognition, speech synthesis.

Introduction

The goal of traditional communication systems is to main- tain signal fidelity under a variety of channel conditions by transmitting raw data symbols over physical channels. However, these systems frequently lack efficiency, particularly in noisy or bandwidth-constrained environments where the transmitted signals contain unnecessary or redundant data. Semantic communication, a paradigm shift that emphasizes conveying the meaning or intent of information rather than the data itself, has arisen as a solution to this problem. In speech-based human-machine interactions, where maintaining the communicative intent rather than just reproducing sounds is the aim, this method shows special promise.

Recent developments in deep learning have transformed speech synthesis and recognition, allowing machines to com- prehend and produce human language with ever-increasing accuracy. It is feasible to create systems that can extract semantic features from speech, send only the most important information, and create understandable speech at the receiver with less data through putting these capabilities into the com- munication pipeline. This increases robustness to noise and channel impairments in while additionally improving commu- nication efficiency. For speech-based applications, we present a deep learning-enabled semantic communication framework in this paper. The system utilizes an end-to-end neural network architecture, in which a decoder uses speech synthesis models to reconstruct the output after a semantic encoder compresses spoken input into meaning-rich representations. Under dif- ferent channel conditions, experimental results show notable improvements in speech quality and bandwidth efficiency. Our research opens the door to resource-efficient and context-aware intelligent communication systems that have a wide range of uses in virtual assistants, smart devices, and other fields. Recent developments in deep learning have made it possible to produce natural-sounding speech outputs and extract high- level semantic features from spoken language, greatly advanc- ing the fields of speech recognition and synthesis. The system that is suggested in this paper is based on these technologies. We present a deep learning based semantic communication framework that integrates speech synthesis, semantic feature extraction, and automatic speech recognition (ASR) into a single architecture. At the transmitter, deep neural networks transform spoken language into semantic embeddings, sig- nificantly lowering the amount of data that must be sent. By using these embeddings, a corresponding model at the receiver reconstructs the speech, guaranteeing intelligibility and maintaining the communicative intent. The robustness and efficiency of our system are tested under a range of channel conditions. According to experimental findings, even in noisy settings, the suggested approach maintains high speech quality and recognition accuracy while achieving notable bandwidth consumption reductions. These findings show the useful bene- fits of semantic communication, especially in situations where communication that is low latency, real-time, and resource- efficient is crucial. A significant amount of data transmis- sion is required to support massive connectivity because of the intense deployment of intelligent in the post-Shannon communication era. The limited spectrum resources that are accessible, however, lead to congestion in the conventional communication system. Inspired by this, we propose a DL- enabled semantic communication system called DeepSC-ST for speech transmission and serving users with various needs. DeepSC-ST especially compresses the input speech sequence into low-dimensional text-related semantic features broadcast over physical channels. Based on the obtained semantic fea- tures, the text sequence is approximated at the receiver. By doing this, the characteristics- actions of speech signals—that is, the voice of speaker, speech delay, background noise, etc.

Related work

Throughout this project, a critical review of existing liter- ature and technologies was undertaken to inform the design and development of the Deep SC-ST system. Shannon and Weaver's seminal concept of semantic communication, which separates symbol and meaning transmission, was one of the key sources of inspiration. This concept defined the shift from

conventional communication systems, which focus on accu- rate bit-level delivery, to semantic communication systems, which focus on the meaning and intention of the conveyed message. Later applications of this concept, including the Deep SC framework, have demonstrated that deep learning models are capable of successfully encoding and decoding semantic information, particularly in text-based communica- tion environments. They showcased the capacity for reducing bandwidth along with improving performance in low SNR conditions that directly influenced this study's use of the semantic modeling approach. Several advanced neural network models for voice recognition have been studied, including Deep Speech, Wav2Vec 2.0, and Transformer-based ASR architectures. These models can extract high-quality textual features from speech inputs, which is a necessary step in the semantic encoding process. Their precision and robustness under real-world settings influenced the encoder used in Deep SC-ST. Various voice synthesis models were also examined for recreating speech from text on the receiver side. Tacotron, Tacotron 2, and WaveNet provided useful frameworks for producing realistic and expressive voice from textual input.

A. speech recognition

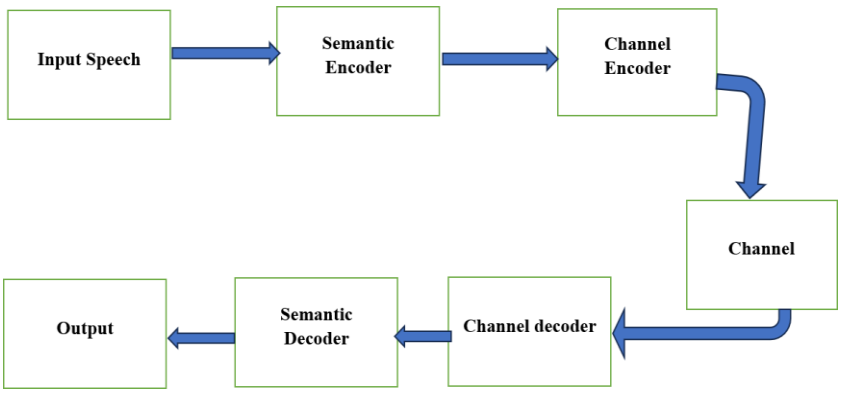
With deep learning, speech recognition has changed dra- matically, allowing machines to comprehend spoken language with previously unheard-of precision. Acoustic models, lan- guage models, and pronunciation dictionaries were the indi- vidual components of the intricate pipelines used by traditional automatic speech recognition (ASR) systems. These pipelines were usually constructed using statistical techniques like hid- den Markov models (HMMs) and Gaussian mixture models (GMMs). These systems were somewhat successful, but they suffered from a lot of hand-engineered features and had trouble with accents, noise, and speaking styles. By introducing end- to-end learning paradigms that use neural networks to directly map raw audio waveforms to text transcripts, deep learning completely changed ASR. Models like DeepSpeech, which used connectionist temporal classification (CTC) and recurrent neural networks (RNNs) to learn temporal dependencies in speech, sparked this change. Later, by more successfully capturing local and sequential audio features, convolutional neural networks (CNNs) and long short-term memory (LSTM) networks further enhanced recognition performance. Semantic communications, in which the objective is to convey the intended meaning rather than precise signals, are made pos- sible by the incorporation of these sophisticated ASR models into larger communication systems. Only the most important parts of the message can be compressed and transmitted by employing deep learning to transform speech into high-level semantic representations. The foundation of next-generation speech communication systems is deep learning-based ASR since it guarantees robustness to noise and improves bandwidth efficiency.

Methodology

Block Diagram

A semantic communication system intended to effectively convey relevant information over a wireless channel is rep-

Fig. 1. Block Diagram



resented by the block diagram. The first step in the process is to run a message input through a semantic encoder. The encoder uses grammatical knowledge, such as part-of-speech (K_p) tagging, to extract important semantic features from the message. The meaning is compressed into a compact representation c rather than being transmitted in its entirety. A channel encoder processes this semantic information and adds redundancy or error correction to get the data ready for trans- mission. As the encoded data x moves over a wireless channel, it could be distorted or noisy, which could cause a different version at the other end. Next, the channel decoder attempts to recover the original semantic features from the noisy version. After receiving these features, the semantic decoder combines the received semantics with contextual information (k) and prior knowledge (K_i) to reconstruct the original message. A predetermined set of codewords provides the prior knowledge, and a language model that aids in more accurate message interpretation provides the context. The reconstructed message is then output by the semantic decoder, hopefully maintaining the original input's

intended meaning. By concentrating on conveying the meaning rather than the precise words, this entire system makes speech or text communication reliable and effective.

System model

.The DeepSC-ST semantic communication system's system model is intended for both speech synthesis and recognition. Using low-dimensional text-related semantic features that are extracted and transmitted from input speech, it recognises the text sequence and uses the user ID and recognised text to reconstruct the speech waveform at the receiver. The system consists of a speech synthesis module, a transmitter with semantic and channel encoders, and a receiver with channel and feature decoders.

Performance metrics

Character Error Rate (CER): This metric measures the accuracy of the recognized text transcription by calculating the number of incorrect characters in the recovered text sequence.

$$CER = \frac{Sc + Dc + Ic}{Nc}$$

- S_c represents the number of character substitutions.
- D_c represents the number of character deletions.
- I_c represents the number of character insertions.
- N_c is the number of characters in the original text.

The number of substitutions (changing one character for an- other), deletions (deleting a character), and insertions (adding a character) required to fix the recovered text is used to compute CER. The CER is then calculated by dividing this sum by the original text's character count.

A lower CER value means that there were fewer transcrip- tion errors and that the recognised text was more accurate.

Word Error Rate (WER): Similar to CER, WER measures the accuracy of the recovered text, but it calculates the number of incorrect words instead of characters.

$$WER = \frac{Sw + Dw + Iw}{Nw}$$

- S_w represents the number of word substitutions.
- $D_w$represents the number of word deletions.
- I_w represents the number of word insertions.
- N_w is the number of words in the original text.

1. Fre´chet Deep Speech Distance (FDSD): This metric quan- titatively evaluates the distribution similarity between the syn- thesized speech and the real speech Higher similarity between the synthesised and real speech is indicated by a lower FDSD value, suggesting that the synthesised speech is of higher quality.
2. Kernel Deep Speech Distance (KDSD): Kernel Deep Speech Distance (KDSD) is a metric that uses kernel functions to assess how similar real and synthesised speech is. Together with Fre´chet Deep Speech Distance (FDSD), it is one of the two quantitative measures that the DeepSC-ST system uses to evaluate the quality of synthesised speech.
3. CTC loss: The goal of the semantic communication system for the speech recognition task is to maximize the posterior probability p, which is equivalent to recovering the text information from the input speech signals. Connectionist temporal classification (CTC) .
4. Epoch:An epoch in deep learning is a single training run of the entire training dataset through the neural network.

Dataset: The collection of information used to train the neural network to carry out its task (speech recognition and synthesis) is known as the training dataset. In this case, the data consists of speech samples and the text that goes with them.

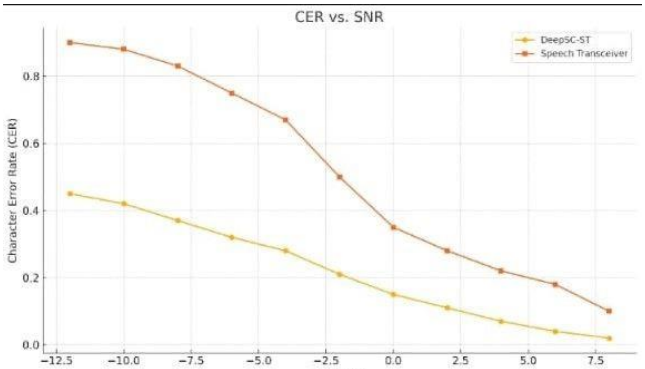
Iteration: Datasets are usually split up into smaller batches because they are frequently very large. One run of a single batch of data across the network is called an iteration.

When the network has processed every batch in the com- plete training dataset, an epoch is said to be finished. Thus, one epoch will comprise 10 iterations if your dataset contains 1000 samples and you choose a batch size of 100.

Results

According to the simulation results, the suggested DeepSC- ST system performs better than both current deep learning (DL)-enabled communication systems and traditional commu- nication systems. When the signal-to-noise ratio (SNR) is low, this benefit is especially noticeable. To put it simply, this

Fig. 2. CER Vs SNR



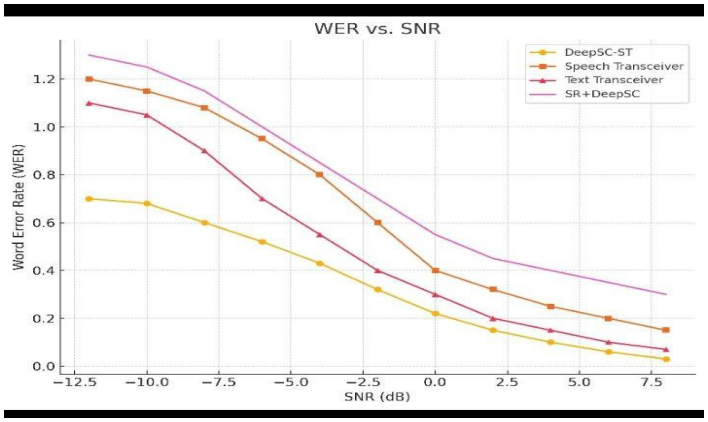


Fig. 3. WER Vs SNR

indicates that DeepSC-ST performs better at speech transmis- sion in difficult situations where the signal strength is lower than the noise. This is an important discovery because noise and interference can deteriorate the quality of the received signal in real-world communication environments. Although the document emphasises this broad trend, it lacks precise numerical data or in-depth comparison graphs to show how much the outperformance has increased. You would probably need to consult the entire research paper or related publications that go into the details of the simulations and the quantitative performance metrics in order to gain a more detailed under- standing.

- Sample input-1: umm.. hello, uh… I wanted to, umm, book a flight to new York for tomorrow morning.
- Sample input-2: hey Alexa. Can you um, set an alarm for, like uh, at 6:30 am tomorrow morning.
- Sample input-3: schedule a meeting at 3:00 pm

A. Numerical comparison

We used 100 speech samples to test each transceiver. With only 20 of the 100 samples successfully encoded, the speech transceiver's encoding efficiency was comparatively low. By successfully encoding 70 samples, on the other hand, the featured transceiver showed better performance. Among the three, the Text transceiver was the most effective, encoding 75 of the 100 speech samples evaluated. As can be seen from this comparison, the Text transceiver is better at processing and encoding speech input than the Feature transceiver, while the Speech transceiver demonstrated the least amount of encoding capabilities.

SNR (dB)	DeepSC-ST (CER)	SpeechTransceiver (CER)
-12.5	0.45	0.88
-10.0	0.42	0.86
-7.5	0.37	0.82
-5.0	0.32	0.70

Fig. 4. comparsion between SNR and CER

	Speech samples	Encoded samples
Speech transceiver	100	20
Feature transceiver	100	70
Text transceiver	100	75

Fig. 5. comparison between original and encoded samples

Conclusion

In this work, the authors create DeepSC-ST, a deep learning- based semantic communication system for speech transmis- sion.

Speech synthesis and recognition serve as the transmis- sion tasks for the DeepSC-ST system. The joint semantic- channel encoder extracts the semantic features related to speech recognition for transmission, and the received semantic features are used to recover the text at the receiver. Without sacrificing performance, this method drastically lowers the volume of data transmitted. A software demonstration is created to provide a proof-of-concept of the DeepSC-ST. The DeepSC-ST system is designed to adapt to dynamic channel environments, and a robust model is identified to cope with different channel conditions. Simulation results show that the proposed DeepSC-ST outperforms conventional and existing DL-enabled communication systems, especially in low signal- to-noise ratio (SNR) regimes. Speech synthesis is carried out at the receiver, where the speech signals are regenerated by feeding the recognised text and speaker information into a neural network module.

REFERENCES

1. F. Xie, B. Gao, F. Jiang, and S. Zhang, "Deepsc: A deep semantic com- munication system for text transmission," IEEE J. Sel. Areas Commun., vol. 39, no. 1, pp. 170–183, Jan. 2021.
2. J. Park, S.-N. Hong, and J. Ha, "R-SC: Reasoning-based semantic communicator robust to noisy channels," IEEE Trans. Veh. Technol., vol. 71, no. 10, pp. 11110–11114, Oct. 2022.
3. Z. Weng and Z. Qin, "Deepsc-s: Deep semantic communication system for speech transmission," in Proc. ICASSP, 2021, pp. 6704–6708.
4. Z. Tong, Y. Shi, and K. B. Letaief, "Federated learning for multi- user audio semantic communications," IEEE J. Sel. Areas Commun., vol. 40, no. 10, pp. 2901–2915, Oct. 2022.
5. H. Liu, Y. Hu, F. Gao, G. Gui, and H. Sari, "Robust semantic communi- cation system against semantic errors," IEEE Trans. Commun., vol. 70, no. 12, pp. 8307–8321, Dec. 2022.
6.
7. Y. Shi, Z. Tong, and K. B. Letaief, "Towards task-oriented com- muni- cations: Semantic transmission for speech," IEEE Trans. Commun., vol. 70, no. 10, pp. 6870–6884, Oct. 2022.
8. S. Zhang, J. Yang, S. Song, and Z. Qin, "Gansc: Generative adversarial networks for semantic communication," IEEE Trans. Cogn. Commun. Netw., vol. 6, no. 4, pp. 1193–1203, Dec. 2022.
9. Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," IEEE Trans. Wireless Commun., to be published.

B. E. Shannon and W. Weaver, The Mathematical Theory of Communi- cation. Urbana, IL, USA: Univ. Illinois Press, 1949.

10. W. Weaver, "Recent contributions to the mathematical theory of com- munication," ETC: A Review of General Semantics, vol. 17, no. 4, pp. 261–280, 1950.
11. H. Ye, G. Y. Li, and B.-H. F. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," IEEE Wireless Commun. Lett., vol. 7, no. 1, pp. 114–117, Feb. 2018.
12. Y. Wu, C.-X. Wang, E. Bastug, M. Bennis, C. Chen, and M. Shikh- Bahaei, "Cell-free massive MIMO for 6G wireless communica- tions," IEEE J. Sel. Areas Commun., vol. 39, no. 3, pp. 931–958, Mar. 2021.
13. M. Giordani, M. Zorzi, and M. Polese, "Toward 6G networks: Use cases and technologies," IEEE Commun. Mag., vol. 58, no. 3, pp. 55–61, Mar. 2020.

C. Salomon, Data Compression: The Complete Reference. Berlin, Ger- many: Springer Science & Business Media, 2007.

14. H. Kim, Y. Lee, J. Kim, I. Lee, and J. Kim, "Semantic communi- cation over noisy channels: Model-based iterative approach," IEEE Trans. Wireless Commun., vol. 19, no. 8, pp. 5317–5331, Aug. 2020.
15. F. Xie, B. Gao, F. Jiang, and S. Zhang, "Deepsc: A deep semantic com- munication system for text transmission," IEEE J. Sel. Areas Commun., vol. 39, no. 1, pp. 170–183, Jan. 2021.
16. F. Jiang, F. Xie, B. Gao, and S. Zhang, "Deepsc-harq: Deep semantic communication systems with hybrid ARQ feedback," IEEE Trans. Cogn. Commun. Netw., vol. 8, no. 1, pp. 509–520, Mar. 2022.
17. H. Yang, F. Jiang, F. Xie, and S. Zhang, "Semantics-native automatic repeat request for deep semantic communication sys- tems," IEEE Trans. Veh. Technol., vol. 71, no. 7, pp. 7990–7995, Jul. 2022.
18. H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Deep learning enabled semantic communication systems," IEEE Trans. Signal Process., vol. 69,
19. pp. 2663–2675, Apr. 2021.
20. Y. Sheng, F. Li, L. Liang, and S. Jin, "A multi-task semantic communi- cation system for natural language processing," in Proc. IEEE 96th Veh. Technol. Conf., 2022, pp. 1–5