Deep Learning-Based Semantic Information Transmission Through Imperfect CSI

I. Background

The recent progress in the field of artificial intelligence has led to a surge in the demand for efficient data transmission, especially in natural language processing and speech recognition scenarios. The methods for conventional communication systems have been facing troubles in sufficing the demands. This leads us to the concept of semantic communication whose nucleus is taking into account the actual meaning and not the raw data.

To convey the historical difficulty of weighing semantic information, Deep Learning technologies have proven to be efficient. For the task of efficiently exchanging meaningful information the semantic communication systems that are deep learning-based show encouraging performance. The systems that are task-oriented, have gained great prominence particularly in the context of 6G and beyond that as well even when there is a scenario with a close upper cap to the resources.

In the text-based communication systems, in the real world there is more focus lies on exchanging meaningful information methods like hybrid automatic repeat request (HARQ) have been incorporated and they turned out to be reliable as well. The communication systems that are speech and audio-dependent rely upon deep learning-powered systems to extract, transmit, and reconstruct the necessary semantic features while at the receiver end some intelligent tasks can be performed.

Due to the increase in the demand for efficient data transmission, there is a deep learning-enabled semantic communication system proposed named DeepSC-ST which has been carved out specially for speech transmission over a wireless channel. The proposed model also works with dynamic channels while compressing the speech into lower dimensional semantic features which also ensures higher accuracy in tasks such as text recognition and speech reconstruction at the receiver end. The proposed system can reduce the network traffic by a significant amount, by providing the users with speech as well as text information whichever is necessary.

II. Motivation

The motivation behind this study flows from the developing domain of communication technologies, especially in the branch of semantic communications and their prospective effect on speech transmission. The state-of-the-art research in semantic-aware communication systems, data transmission like audio/video and text, spanning text, showcases the leap made in a forward direction in the extraction and transmission of semantic information in an efficacious manner. Although there are advancements in the methods for the extraction of semantic information from the input speech signals and then reconstructing those speech signals again into a textual format at the receiver end. Here a very important area remains under exploration which is the execution of specifically speech-intelligent tasks using the method of semantic communications at the receiver end.

A passing need to optimize the data transmission in the devices was generated with the increase of intelligent devices in the post-Shannon era of communication systems. The issue of limited spectrum sources was still prevalent leading to the narrowing in the traditional communication systems. The proposed out-of-ordinary deep learning-enabled semantic communication system, DeepSC-ST, is built especially for speech transmission over wireless channels. This specific communication system aims to address the limited resource scenarios by compression the speech data into dimension textual semantic features which are then transmitted over the physical channel. And further on at the receiving end, these semantic features are used to estimate the textual output while reducing the traffic over the network significantly so that more users can use the channel to request the text information.

This proposed model restores the textual sequence at the receiver end depending on the user's pre-registered identity, utilizing the available spatial information to reconstruct the sequence as accurately as possible. Putting it in a nutshell, this report is motivated towards narrowing intelligent speech and semantic communications. The proposed DeepSC-ST system turns out as a good possible solution for being a prominent change in the speech communication and transmission era of limited resources.

III. Contribution

- We know that in a real-life scenario, perfect CSI at the transmitter is merely an illusion. A novel semantic communication system, named DeepSC-ST, is proposed for communication scenarios with speech input where there is imperfect CSI at the transmitter, in which a joint semantic-channel coding scheme is developed.

- A demonstration of the DeepSC-ST with operable user interface is built to produce the recognized text and the synthesized speech based on the real human speech input.

IV. Problem Formulation

In Wireless Semantic Communication systems, one of the major problems encountered is automatic speech recognition, semantic encoding-decoding, and transmitting semantic information in a low signal-to-noise(SNR) ratio along with a resource-constrained environment.
Let m be the input speech sample sequence, $m = [m_1, m_2, \ldots, m_Q]$ that is being divided into N frames. These frames are converted into the spectrum through the Hanning window transform, Fast Fourier Transform (FFT), logarithmic operations, and normalization.
Let **w** be the vector of length N of Hanning window transform.

$$\therefore w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n < N$$

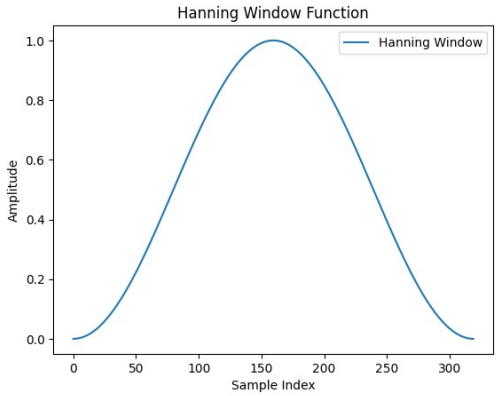

Fig. 1. Hamming Window transform

Let **X** be the vector of length N of Fast Fourier transform.

$$X(k) = \sum_{n=0}^{N-1} w(n) \cdot e^{-j\frac{2\pi}{N}kn}$$

By doing so, the spectrum, $s = [s_1, s_2, \ldots, s_N]$ contains the characteristics of the sample sequence as shown in following Figure 4. The ultimate goal of the speech recognition task

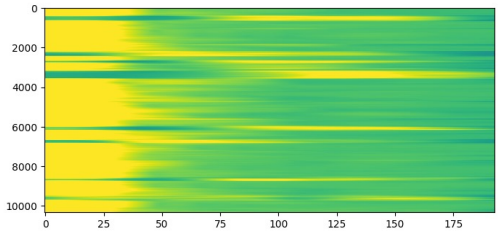

Fig. 2. Spectogram to be given as input in semantic encoder

is to recover the final text transcription, $\widehat{t}$, as close to t as possible. Denote $t = [t_1, t_2, \ldots, t_K]$, where t_k is a token from the token set, t, that could be a character in the alphabet or a word boundary.

Semantic Encoder: Once the preprocessing of the speech signal is done, the semantic encoder encodes the information using CNN, and GRU (Bidirectional RNN), as shown in Figure 2. The input spectra **S**, are first converted into the intermediate features via several CNN modules. Particularly, the number of filters in each CNN module is E_p, $p \ \epsilon[1, 2, ..., P]$, and the output of the last CNN module is $b \ \epsilon R^{B*C_P*D_P*E_P}$ Then, b is fed into **Q** BRNN modules, successively, and produces $d\epsilon R^{B*G_Q*H_Q}$, where the number of GRU units in each BRNN module, $H_q, q \ \epsilon[1, 2, \ldots, Q]$, is consistent. Finally, the text-related semantic features, P, are obtained from d by passing through multiple cascaded dense layers and a softmax layer.

Channel Encoder: Then the channel encoder transforms the semantic information to $U = \mathbf{T}_\beta^C(\mathbf{T}_\alpha^S(s))$ using two dense layers. Then U is reshaped and transmitted via a physical channel as

$$\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{w}$$

where **y** is the channel output with **w** as complex Gaussian noise.
Imperfect CSI: The system model will be

$$\mathbf{y} = \widehat{\mathbf{h}} * \mathbf{x} + \mathbf{n}$$

Let us assume the imperfect CSI to be $\widehat{\mathbf{h}}$ and correlation between **h** and $\widehat{\mathbf{h}}$ be ρ,. i.e.

$$E[h * \widehat{h}] = \rho$$

Assumptions:
1. Imperfect CSI is denoted as $\widehat{h}$, is available at the receiver for detection.
2. $h \sim CN(0, 1)$, $\rho =$ correlation between **h** and $\widehat{\mathbf{h}}$.

Using the Maximum Likelihood detection rule, the detection variable d is as follow:

$$d = Re\{\widehat{h} * y\} = Re\{\widehat{h} * (hx + n)\}$$

as h can be represented in terms of $\widehat{h}$ and random error δ as follow:

$$\widehat{h} = \rho h + \sqrt{1 - \rho^2}\delta$$

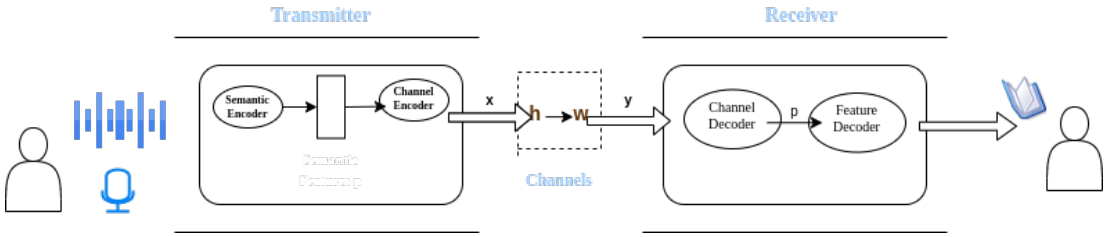

Fig. 3. Deep SC-ST Model Architecture

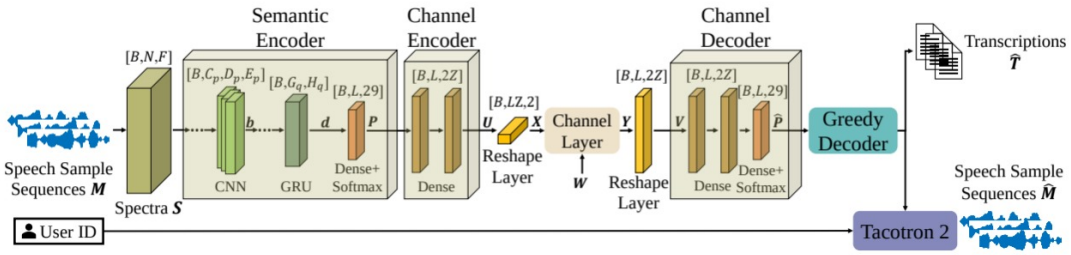

Fig. 4. Deep SC-ST Model Architecture

By further simplifying the detection variable, we get:

$$z = \rho|\widehat{h}|x + Re\{\sqrt{1-\rho^2}\frac{\widehat{h}}{|\widehat{h}|}\delta x + \frac{\widehat{h}}{|\widehat{h}|} * n\}$$

Greedy Decoder: Given input sequence $X = (x_1, x_2, \ldots, x_T)$ and output sequence $Y = (y_1, y_2, \ldots, y_U)$, the probability of generating Y is defined as:

$$P(Y|X) = \prod_{u=1}^{U} P(y_u|y_1, y_2, \ldots, y_{u-1}, X)$$

In a greedy decoder, the model selects each token $\hat{y}_u$ at each decoding step u by choosing the token with the maximum probability:

$$\hat{y}_u = \arg\max_{y_u} P(y_u|y_1, y_2, \ldots, y_{u-1}, X)$$

The overall objective is to maximize the probability of the entire output sequence:

$$\hat{Y} = \arg\max_{Y} P(Y|X)$$

To maximize the posterior probability $p(t \mid s)$, the Connectionist Temporal Classification Loss (CTC loss) is adopted as the loss function for speech recognition task in our system, denoted as

$$\mathcal{L}CTC(\boldsymbol{\theta}) = -\ln\left(\sum A \in \mathfrak{A}(\boldsymbol{s}, \boldsymbol{t})\left(\prod_{l=1}^{L}\widehat{p}_l\left(a_l \mid \boldsymbol{s}, \boldsymbol{\theta}\right)\right)\right)$$

where θ denotes the neural network parameters of the transmitter and the receiver, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}}\right)$.

Moreover, for given prior channel state information (CSI), the neural network parameters, $\boldsymbol{\theta}$, can be updated by the stochastic gradient descent (SGD) algorithm as follows,

$$\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)} - \eta\nabla_{\boldsymbol{\theta}^{(i)}}\mathcal{L}_{CTC}(\boldsymbol{\theta}),$$

where $\eta > 0$ is a learning rate and ∇ indicates the nabla operator.

V. Experiments

A. Dataset

The LJ-speech dataset plays a crucial role in understanding and implementing the DL-enabled semantic communication system. It acts as a basic source of information for training, testing and validating the system's functionality in transmitting speech-related semantic features efficiently.

The dataset employed in this study was obtained from TensorFlow. The dataset consists of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. A transcription is provided for each clip. Clips vary in length from 1 to 10 seconds and have a total length of approximately 24 hours.

We know that we are working in a speech-based semantic communication system. The LJ-speech dataset plays a pivotal role in the order to perform the CNN + GRU (a bidirectional RNN) algorithm.

The dataset was divided into 2 parts. 90 percent was allotted to training and the remaining 10 percent was allotted to validation.

B. Simulations

Algorithm 1 Training algorithm for speech-recognition task

Initialization: Initialize parameters $\theta^{(0)}$, $i = 0$
Input: Speech sample sequences $\boldsymbol{M}$ and transcriptions $\boldsymbol{T}$ from trainset $\mathfrak{S}$, fading channel $\boldsymbol{H}$, noise $\boldsymbol{W}$.
Generate spectra S from sample sequences $\boldsymbol{M}$.
while CTC loss is not converged **do**
  $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}\left(\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{S})\right) \rightarrow \boldsymbol{X}$.
  Transmit $\boldsymbol{X}$ and receive $\boldsymbol{Y}$ via (2). $\mathbf{R}_{\delta}^{\mathcal{S}}(\boldsymbol{Y}) \rightarrow \widehat{\boldsymbol{P}}$.
  Compute loss $\mathcal{L}_{CTC}(\boldsymbol{\theta})$ via (7).
  Update parameters $\boldsymbol{\theta}$ via SGD according to (8).
  $i \leftarrow i + 1$
end while
Output Trained networks $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\delta}^{\mathcal{C}}(\cdot)$.

Algorithm 2 Testing algorithm for speech-recognition task

Input: Speech sample sequences $\boldsymbol{M}$ from test- set, trained networks $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\delta}^{\mathcal{C}}(\cdot)$., testing channel set $\mathcal{H}$, a wide range of SNR regime.
Generate spectra S from sample sequences $\boldsymbol{M}$.
for channel condition $\boldsymbol{H}$ drawn from $\mathcal{H}$ **do**
  for each SNR value **do**
    Generate Gaussian noise $\boldsymbol{W}$ under the SNR value.
    $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}\left(\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{S})\right) \rightarrow \boldsymbol{X}$.
    Transmit $\boldsymbol{X}$ and receive $\boldsymbol{Y}$ via (2).
    $\mathbf{R}_{\delta}^{\mathcal{S}}(\boldsymbol{Y}) \rightarrow \widehat{\boldsymbol{P}}$.
    Decoding $\widehat{\boldsymbol{P}}$ into $\widehat{\boldsymbol{T}}$ via (4).
  end for
end for
Output: Recovered text transcriptions, $\widehat{\boldsymbol{S}}$.

C. Results

The figures 5, 6 and 7 show the relationship between the CTC loss and the number of epochs at different learning rates. With decreasing learning rates i.e. the step size, the corresponding loss at different epochs decreases maximum for learning rate 0.0001.

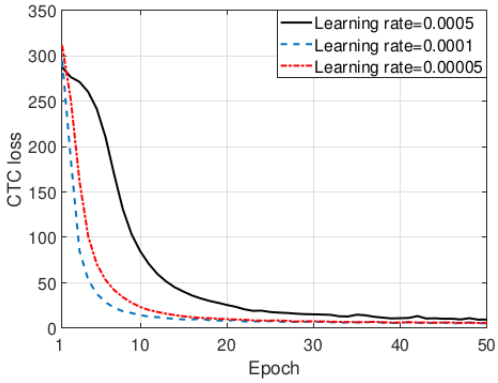

Fig. 5. CTC loss curves: from the paper

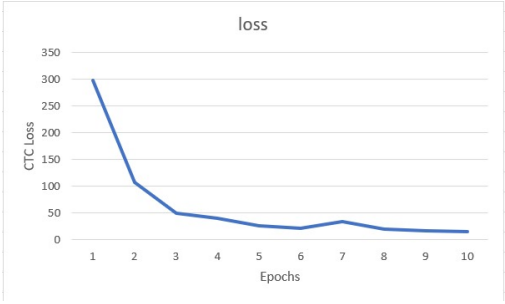

Fig. 6. CTC loss curve: Regenerated with perfect CSI Learning rate= 0.0001

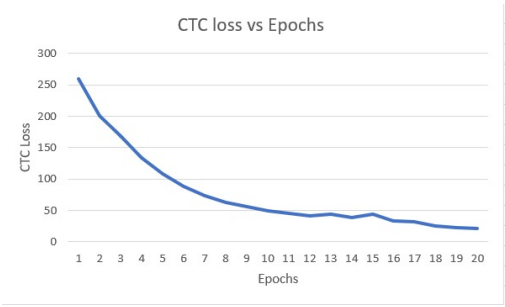

Fig. 7. CTC loss curve: Imperfect CSI with Learning rate= 0.0001

Both of the following curves depict the behaviour of CTC loss as the number of epochs increases at the learning rate of 0.0001 and with the former being for perfect CSI and the later being for imperfect CSI. It is observed that there is a sudden and rapid decrease in the loss incase of perfect CSI as the loss reaches around 10 after 20 epochs whereas in the case of imperfect CSI it reduces to just 20. The convergence occurs speedily in the case of perfect CSI. Both the curves follow negative exponential pattern as the gradually approach 0 moving towards infinity.

In the domain of wireless communication systems, the plot of CTC loss for various CSI depicts the influence of channel state knowledge in the training process. Incase of perfect CSI the highly efficient convergence depicts the ideal scenario where the model benefits from the channel state information. Whereas incase of imperfect CSI it shows a very gradual decline depicting a realistic communication scenario where obtaining accurate information about the channel state is a difficult task due to all the interferences like noise, fading, multi path propogation and dynamic channel conditions.

VI. Conclusion and Future Scope

In this research, we studied DeepSC-ST, a DL-enabled semantic communication system for voice recognition and synthesis tasks. DeepSC-ST uses text-related semantic features to reconstruct the speech and recover text transcription by utilizing text-related semantic features. Specifically, in order to achieve voice recognition, a hybrid semantic-channel coding scheme was built to learn and extract semantic information and reduce the channel effects. The DeepSC-ST performs better than both existing semantic communication

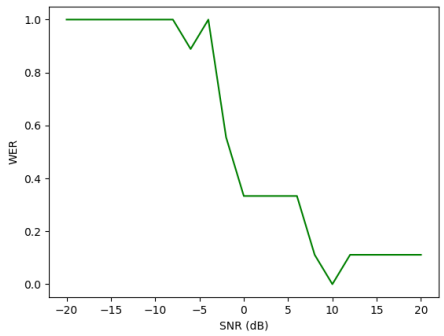

Fig. 8. Word Error Rate(WER) curve for perfect CSI Regenerated at various discrete SNR points (AWGN Channel)

systems and traditional communication systems, particularly in the low SNR region, according to simulation studies. Furthermore, for the proof-of-concept, we developed a software demonstration that accepts actual human speech input. However, we know that in a real-life scenario, assuming perfect channel state information at the transmitter is far from ideal. We usually don't have perfect channel state information at the transmitter. Our proposed DeepSC-ST in case of imperfect channel state information at the transmitter in is envisioned to be a promising candidate for semantic communication systems for speech recognition and speech synthesis tasks. One of the limitations currently is that the DeepSC-ST model works only with audio and text, so in the future we can also expand this model so that it works with other forms of media such as images and videos.

VII. Team Learning

ECE310 is one of the courses where we had to work with the same group throughout the semester. We had varied ideas and opinions and it was sometimes tedious to be on the same page. During the entire course of the project, our professor Dr Dhaval K. Patel and teaching assistant Kashish Shah kept us motivated and always provided help whenever required.

The project helped us to connect the theoritical learnings of the wireless communications course with real-life problems. We learned what are semantic features, learned about its transmission from the transmitter to the receiver. We also learned how to train a model from scratch. We broke the entire problem statement into multiple smaller segments which helped us to steer away from the potential confusion about the jargon.

The biggest learning outcome was working with the same set of people for quite an extensive period. As a whole, project really helped us in team building along with academic learning.

References

[1] Brownlee, J. (2019, August 5). How to prepare univariate time series data for long short-term memory networks. MachineLearningMastery.com. https://machinelearningmastery.com/prepare-univariate-time-series-data-long-short-term-memory-networks/

[2] Deep Joint Source-channel coding for Semantic Communications - arXiv.org. (n.d.-b). https://arxiv.org/pdf/2211.08747.pdf

[3] A demo of Semantic Communication: Rosefinch — IEEE conference ... (n.d.-a). https://ieeexplore.ieee.org/document/10039193

[4] Europarl Parallel corpus. (n.d.). https://www.statmt.org/europarl/

[5] How to read a paper - SIGCOMM. (n.d.-c). http://ccr.sigcomm.org/online/files/p83-keshavA.pdf

[6] IEEE Xplore Full-text PDF: (n.d.-d). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=amp;arnumber=9465776

[7] The LJ speech dataset. Keith Ito. (n.d.). https://keithito.com/LJ-Speech-Dataset/ Qin, Z., Tao, X., Lu, J., Tong, W., amp; Li, G. Y. (2022, June 27). Semantic Communications: Principles and challenges. arXiv.org. https://arxiv.org/abs/2201.01389

[8] Principles and challenges. arXiv.org. https://arxiv.org/abs/2201.01389

[9] Robust semantic communications with masked VQ-Vae enabled ... - arxiv.org. (n.d.-e). https://arxiv.org/pdf/2206.04011.pdf

[10] Understand-before-talk (UBT): A semantic communication ... - IEEE xplore. (n.d.-f). https://ieeexplore.ieee.org/document/9937052/