Neural Network Based Carrier Frequency Offset Estimation From Speech Transmitted Over High Frequency Channels

Jens Heitkaemper, Joerg Schmalenstroeer, Reinhold Haeb-Umbach
Department of Communications Engineering, Paderborn University, Germany
{jensheit, schmalen, haeb}@nt.uni-paderborn.de

***Abstract*—The intelligibility of demodulated audio signals from analog high frequency transmissions, e.g., using single-sideband (SSB) modulation, can be severely degraded by channel distortions and/or a mismatch between modulation and demodulation carrier frequency. In this work a neural network (NN)-based approach for carrier frequency offset (CFO) estimation from demodulated SSB signals is proposed, whereby a task specific architecture is presented. Additionally, a simulation framework for SSB signals is introduced and utilized for training the NNs. The CFO estimator is combined with a speech enhancement network to investigate its influence on the enhancement performance. The NN-based system is compared to a recently proposed pitch tracking based approach on publicly available data from real high frequency transmissions. Experiments show that the NN exhibits good CFO estimation properties and results in significant improvements in speech intelligibility, especially when combined with a noise reduction network.**

***Index Terms*—speech enhancement, carrier frequency offset estimation, single-sideband transmissions**

I. Introduction

With carrier frequency offset (CFO) we denote the frequency difference between the modulation oscillator in the transmitter, which mixes the baseband signal to the transmit carrier frequency and the demodulation oscillator of the receiver which mixes it back to baseband. A CFO causes a shift in the frequency spectrum in the demodulated audio signal. This shift significantly affects the speech quality [1] and leads to a degraded listening experience. It has been termed "chipmunk-like" speech [2] or "duck-speaking" [3] speech. Even small CFOs of 5 Hz already result in a noticeable degradation [2] and CFOs above 10 Hz lead to a reduced intelligibility [4].

The main cause of CFOs is an inaccurate knowledge of the modulation frequency. But even with perfect knowledge of the modulation frequency, a significant CFO can be caused by the temperature sensitvitiy of the crystal oscillators of the receiver due [5, p. 92], or by the Doppler shift due to a moving receiver/transmitter [5, p. 91]. For example, a temperature-induced inaccuracy of 25 ppm in the demodulation frequency for a transmission a 7 MHz carrier frequency results in a CFO of about 175 Hz!

For analog single-sideband (SSB) transmissions there exist several approaches to estimate the CFO which exploit the spectral properties of speech signals [2, 3, 6, 7]. Some systems focus on the pitch and harmonics of the received speech signal to derive the CFO [6, 7]. Others use the third-order harmonics of the signal [2] or a combination of a rough preliminary estimation with a subsequent neural network (NN)-based fine-tuning step [3].

In this work a NN-based CFO estimator is presented. It consists of two stages, a masking stage to extract areas in the signal with high energy and a full-band classifier that retrieves a final CFO estimate. Thereby, the mask is estimated from the signal spectrum that is segmented along the time axis, and the segments are independently processed by sub-band layers (SBLs). Here, either a convolutional neural network (CNN) or a recurrent neural network (RNN) architecture is employed.

Furthermore, a software framework is designed to simulate SSB signals exhibiting different CFOs. It artificially adds additive distortions to the signals that have been gathered from real recordings of high frequency transmissions. As shown by the experiments, these simulations are realistic enough that a NN trained solely on simulated data achieves competitive results on real recordings.

In our evaluation, the proposed network is compared to the recently proposed "RAKE" estimator [7] for CFO estimation. Both approaches operate on segments, for which a NN-based speech activity detection (SAD) estimator has detected speech activity [8]. Additionally, the influence of the CFO estimation and correction on the performance of a downstream noise reduction unit is examined. The noise reduction network is based on the ConvNet architecture [9] and has shown strong noise reduction performance on the CHiME-4 database [10].

The paper is structured as follows. In Section II the NN-based CFO estimator is introduced. Afterwards, the simulation framework is described in Section III. Section IV gives a short introduction to the noise reduction network configuration and training. The evaluation results are given in Section V and the paper ends with a short conclusion in Section VI.

II. NN-Based CFO Estimation

A NN for CFO estimation should be robust against non-stationary noise even at low signal to noise ratio (SNR). To achieve that, the network is designed as a two-stage model consisting of a masking and a classification layer. The first stage calculates an attention mask that is multiplied with the observed signal. Thereby, frequencies with activity in the input

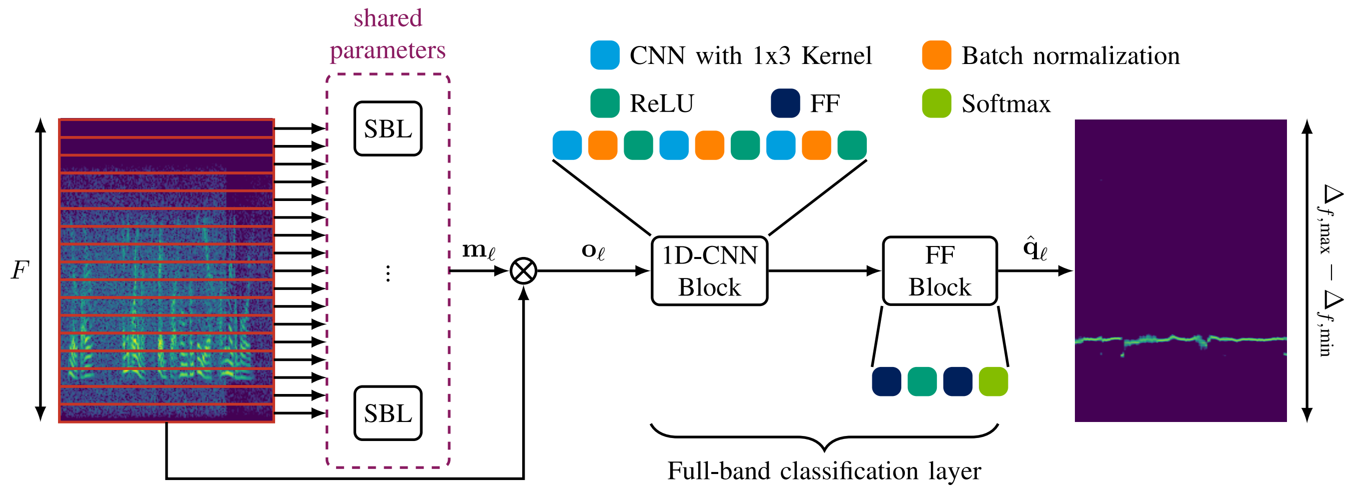

Fig. 1. Block diagram of the NN-based carrier frequency difference estimator.

vector are emphasized, such that the second stage can focus on frequencies with high energy and is less affected by additive noise.

The mask estimation is performed separately for each of $N_{\tilde{F}}$ sub-band layers (SBLs) using a network with shared parameters. Each layer processes a sub-band vector $\tilde{\mathbf{y}}_\ell \in \mathcal{R}^{\tilde{F}}$ that is extracted from the input vector $\mathbf{y}_\ell \in \mathcal{R}^F$ with F denoting the number of frequency bins, $\tilde{F} = \lfloor F/N_{\tilde{F}} \rfloor$ the sub-band size and ℓ the frame index. The sub-bands do not overlap, and possibly remaining frequency bins $[\tilde{F} \cdot N_{\tilde{F}}, \dots, F]$ near the Nyquist frequency are dropped. Estimating masks per sub-band allows the SBLs to highlight speech and suppress noise regardless of which sub-band contained the observed noise pattern during training. Finally, the attention vector $\mathbf{m}_\ell$, which has the same size as the input vector $\mathbf{y}_\ell$, is constructed from the sub-band mask vectors $\tilde{\mathbf{m}}_\ell$ estimated by the SBLs.

The vector $\mathbf{m}_\ell$ is multiplied with the input vector resulting in $\mathbf{o}_\ell = \mathbf{m}_\ell \circ \mathbf{y}_\ell$, with $\circ$ denoting the Hadamard product. This masked vector $\mathbf{o}_\ell$ is further processed by the second stage of the network (”*full-band classification layer*”). This layer takes advantage of the full frequency range to predict the CFO. The output vector $\hat{\mathbf{q}}_\ell$ has the size $\Delta_{f,\text{max}} - \Delta_{f,\text{min}}$, where the frequency bin indices $\Delta_{f,\text{min}}$ and $\Delta_{f,\text{max}}$ refer the lowest and the highest considered CFO value.

Two possible architectures are compared for the SBLs, a multi-layer RNN and a 1D-CNN block consisting of three CNN layers each including a batch normalization [11]. Both architectures use a Sigmoid output activation function that limits the value range of the SBL output masks $\tilde{\mathbf{m}}_\ell$ to $[0, 1]$.

For the classification layer a 1D-CNN block with a subsequent feed forward (FF) block is chosen. The 1D-CNN block consists of three CNN layers with batch normalization and a rectified linear unit (ReLU) activation function. Each CNN layer has a 1×3 kernel and a decreasing number of channels. For the FF block two linear layers are used with a ReLU and a softmax activation function, respectively. A block diagram of the described architecture is depicted in Figure 1.

Binary cross entropy (BCE) is chosen as the loss function since the discrete number of possible CFOs states a typical classification task. All NNs require adequate training data for the task at hand. Therefore, the next section introduces a framework to simulate SSB signals with a CFO.

III. CFO Simulation

Recording a database of real SSB transmissions with various CFOs is expensive and time consuming. Therefore, we decided to simulate the entire training data for the CFO estimation network and use real recordings only for evaluation purposes.

The impact of a CFO on the demodulated signal is two-fold. First, because of its spectral shift part of the speech spectrum is suppressed by the filters applied to the signal during modulation and demodulation. This leads to the fact that a part of the signal is lost. Second, the remaining speech signal spectrum is shifted on the frequency axis.

To simulate these effects, signals from the clean training set of the LibriSpeech database [12] are utilized in a resampling based procedure, which is computationally much more efficient than simulating the whole modulation and demodulation process.

According to the International Telecommunication Union (ITU) regulations, the bandwidth of a SSB transmission is limited to 2.7 kHz. Therefore, the clean signal is band-limited with a low-pass filter and it is interpolated by a factor of four to a higher sampling rate to obtain the required bandwidth for the simulation of a shift along the frequency axis. Subsequently, the interpolated signal is shifted to a simulated carrier frequency via modulation and processed by a band-pass filter to remove one of the side bands. Which to remove depends on whether lower sideband (LSB) or upper sideband (USB) transmission is simulated. The signal is then filtered with a band pass whose center frequency is chosen depending on the CFO to simulate. Afterwards, the signal is demodulated with a carrier frequency including the CFO, low-pass filtered and decimated by a factor of four. The resulting signal shows the CFO induced frequency

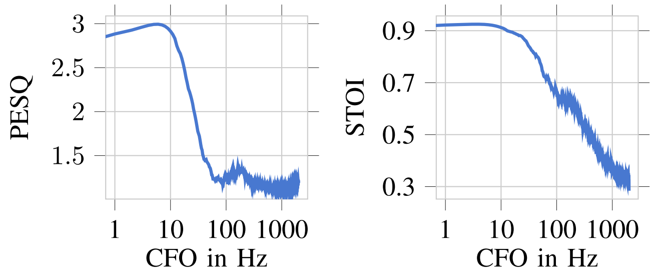

Fig. 2. PESQ and STOI values for different CFOs.

shift and bandwidth loss. In a final step, real recordings of a high frequency link without an active transmission are added to the simulated signal with a random SNR to represent channel noise.

A similar procedure is used to correct an estimated CFO. In this case the CFO is considered and compensated during the modulation of the signal to the simulated carrier frequency.

IV. Noise Reduction Network

The CFO estimation and correction is followed by a speech enhancement system that aims at reducing the channel noise. Here, the ConvNet [9] architecture that was proposed in [10] is used. It is trained on the simulated SSB signals described in the last section with a CFO of zero using the scale invariant signal to distortion ratio (SI-SDR) [13] as the loss function. As suggested in [10], the network is trained to estimate both the clean, noisefree speech and the noise. Here, the target speech signal is the LibriSpeech clean signal downsampled to $8\,\mathrm{kHz}$ and band-limited to $2.7\,\mathrm{kHz}$. The network's encoder and decoder use a large window size of $64\,\mathrm{ms}$ and a shift of $16\,\mathrm{ms}$, as first experiments have shown that these parameters leads to the best results on real SSB transmissions.

V. Evaluation

The performance of the NN-based CFO estimator is evaluated against the cumulative distribution function (CDF) of the CFO estimation errors. Its impact on the following speech enhancement system is measured using the perceptual evaluation of speech quality (PESQ) [14] and short-time objective intelligibility (STOI) [15] metrics of the enhanced signals. Since both metrics have been developed for signals without a CFO, it needs to be checked if they allow a valid evaluation for our scenario with CFO correction. To do so, we calculated the average PESQ and STOI values on a set of 5000 CFO distorted signals. All signals are simulated as described in Section III with channel noise added with an SNR uniformly drawn at random between 25 and $30\,\mathrm{dB}$. From the graphs in Figure 2 a clear correlation can be seen between the CFO and the STOI value, if the CFO is above $10\,\mathrm{Hz}$. For the PESQ value a correlation with the CFO is visible if the CFO is in the range between $10\,\mathrm{Hz}$ and $100\,\mathrm{Hz}$.

A. Database

While training was carried out on simulated data as described, the evaluation experiments are conducted on real SSB recordings, which are publicly available from [7][1]. The evaluation data consists of utterances from the LibriSpeech database [12] that were transmitted using LSB modulation over high frequency links at $7.06\,\mathrm{MHz} - 7.063\,\mathrm{MHz}$ and $3.6\,\mathrm{MHz} - 3.62\,\mathrm{MHz}$ and recorded by Kiwi-software defined radio devices [16]. We followed the steps described in [17] to create the database, choosing modulation and demodulation frequencies appropriately for each transmission to generate CFOs from the following set: $[0\,\mathrm{Hz}, 100\,\mathrm{Hz}, 300\,\mathrm{Hz}, 500\,\mathrm{Hz}, 1000\,\mathrm{Hz}]$. Here, only positive CFOs are considered since negative shifts lead to the loss of the pitch and its first harmonics, so that the speech signal cannot be reconstructed easily. As the database [7] had been recorded without human supervision, about $1\,\%$ of the recordings include concurrent speakers at neighboring frequencies. These recordings are not considered in the following experiments in order to focus on the actual task, but this somewhat limits the comparability with the error rates reported in [7].

B. Baseline

The "RAKE" algorithm for CFO estimation presented in [7] is used as a baseline system. It estimates the CFOs by tracking the pitch and the harmonics in the speech signal. Since the approach is limited to positive CFOs, only the range $\Delta_{f,\min} = 0\,\mathrm{Hz}$ to $\Delta_{f,\max} = 1500\,\mathrm{Hz}$ is considered.

C. NN parameters of the CFO estimator

All networks are trained on simulated signals as described in Section III with a random SNR in the range from $-5\,\mathrm{dB}$ to $10\,\mathrm{dB}$ and a CFO between $-100\,\mathrm{Hz}$ and $2000\,\mathrm{Hz}$. The short time Fourier transform (STFT) magnitudes of the input signals are chosen as features with a window length of $32\,\mathrm{ms}$, an overlap of $22\,\mathrm{ms}$ and a size of 4096. Since the network operates in the frequency domain, its estimation accuracy is limited by the frequency resolution. For the given STFT size, the network has an inherent limit to its estimation accuracy of about $1.95\,\mathrm{Hz}$.

The classifier consists of a 1D-CNN block, which includes CNN layers with 128, 256 and 512 channels. For the FF block 512 and 2100 units are chosen.

Two SBL architectures are investigated, each employing approximately $5\,\mathrm{M}$ parameters. The CNN consists of three layers with 256, 256 and $\tilde{F} = 128$ channels. For the RNN three layers with 256, 256, 128 units are chosen. All networks are trained for $100\,\mathrm{k}$ iterations with the ADAM optimizer [18] and a learning rate of 0.001.

During evaluation the CFO estimate $\hat{\Delta}_f$ for an utterance is calculated with

$$\hat{\Delta}_f = \underset{i}{\operatorname{argmax}} \sum_{\ell=1}^{L} \hat{p}_{\ell,i}, \tag{1}$$

[1] https://zenodo.org/record/4485559

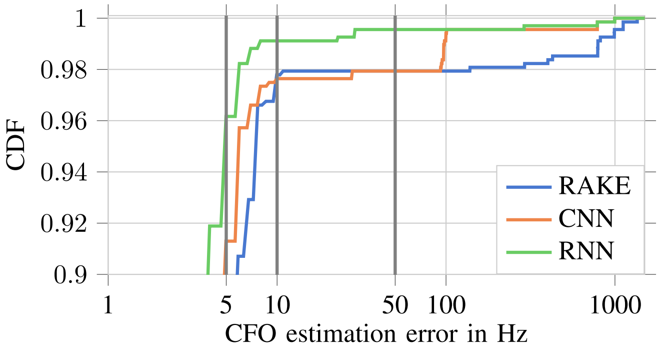

Fig. 3. CDF of the CFO estimation error for the NN-based systems compared with the RAKE algorithm.

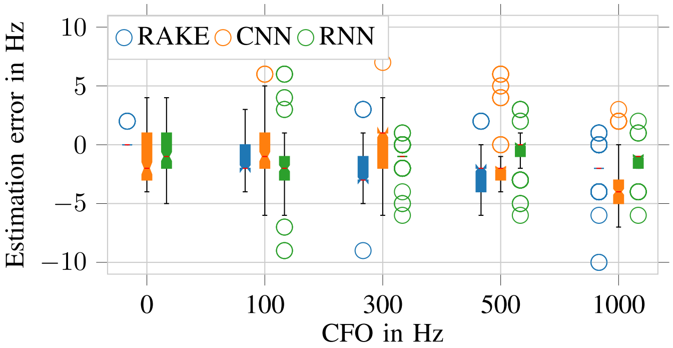

Fig. 5. CFO estimation error for the five possible CFOs for the NN-based systems compared with the RAKE algorithm.

where $\hat{p}_{\ell,i}$ is the i-th value of the NN output vector $\hat{\mathbf{p}}_{\ell}$. The sum over the time dimension reduces the influence of small errors in the estimation, e.g., in case of short silence segments.

D. CFO Estimation

In the following experiments, the NN-based estimators are compared to the RAKE algorithm. First, the SAD from [8] is applied to the signal to identify segments with active speakers. Then all active segments in a record are concatenated before the CFO estimation is performed. Each recording includes at least 10 s of activity. To handle signals including CFOs the SAD had been retrained with appropriate data.

As displayed in Figure 3, the RNN-based estimator clearly outperforms the RAKE algorithm, while the CNN is on par with the RAKE algorithm. All approaches provide in 98 % of the cases an error below 10 Hz, which is considered acceptable because of the negligible intelligibility loss [4].

A detailed analysis showed that the errors above 10 Hz of the RAKE algorithm are mostly originating from an interfering narrow-band signal that is active on periodically repeating frequencies, which is misinterpreted as speech harmonics (see Figure 4). The NN-based systems are not distracted by these interfering narrow-band signals since similar distortions are seen during training.

In Figure 5 the estimation error is plotted for the five possible CFOs in the evaluation set. For all three estimators, an increasing CFO does not lead to a higher estimation error, which indicates an independence of the estimation error from the true CFO value, assuming positive CFOs.

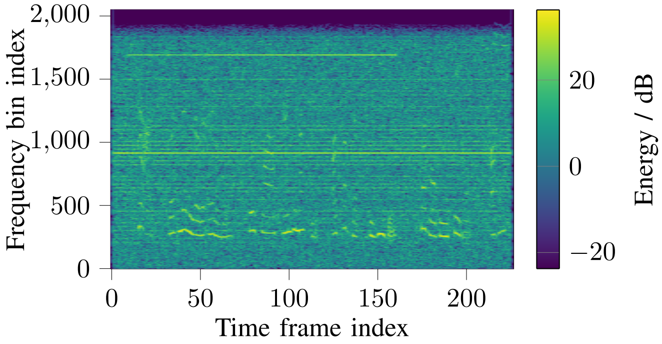

Fig. 4. Excerpt of the spectrogram for a signal that led to an error in the CFO estimation with the RAKE algorithm.

E. Speech enhancement

Although the NN-based CFO estimators outperform the RAKE algorithm, it remains to be examined whether the difference in accuracy has a significant impact on the speech quality after applying a denoising system to the CFO compensated signal. Therefore, the noise reduction system described in Section IV is applied to recordings with and without CFO compensation and the corresponding results are given in Table I.

Both the NN-based estimators and the RAKE achieve similar enhancement metrics as an oracle CFO compensation. However, the NN-based estimators do not result in better enhancement performance than the RAKE CFO estimator. This can be explained by highly distorted signals, where the noise reduction fails even with oracle CFO information. The relatively small performance difference between the NN-based and the RAKE CFO estimator is then insignificant. Note, that all CFO estimation systems lead to a large improvement in both enhancement metrics compared to the noise reduction without a CFO compensation.

VI. Conclusions

In this paper we have presented a NN-based CFO estimator and a simulation framework to generate artificial training data. The network, solely trained on simulated signals, achieves a CFO estimation accuracy with a remaining error of less than 10 Hz in about 99 % of the cases on real SSB recordings, and

TABLE I
Speech enhancement results on real HF recordings for CFO correction and denoising.

CFO Estimation	Not Denoised		Denoised	
& Compensation	PESQ	STOI	PESQ	STOI
None	1.28	0.40	1.47	0.31
RAKE	1.69	0.60	2.31	0.75
CNN	1.67	0.60	2.29	0.75
RNN	1.69	0.60	2.33	0.75
Oracle	1.69	0.61	2.43	0.78

is superior to a signal-characteristic-based method used for comparison.

However, this improvement in CFO estimation is not transformed into substantial gains in speech quality. All investigated methods achieve comparable enhancement results in terms of PESQ and STOI. Overall, the experiment with a subsequent noise reduction system demonstrated the importance of CFO correction to achieve good enhancement performance.

Acknowledgment

The authors would like to thank Plath GmbH in Hamburg (Germany) for funding part of the research presented in this contribution. Computational resources were provided by the Paderborn Center for Parallel Computing.

References

[1] P. F. Assmann, S. Dembling, and T. M. Nearey, "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proc. Interspeech 2006*, 2006, pp. paper 1710–Tue1BuP.10.

[2] P. Clark, S. H. Mallidi, A. Jansen, and H. Hermansky, "Frequency offset correction in speech without detecting pitch," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7020–7024.

[3] H. Xing and J. H. L. Hansen, "Single sideband frequency offset estimation and correction for quality enhancement and speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 124–136, 2017.

[4] T. Gülzow, U. Heute, and H. J. Kolb, "SSB-carrier mismatch detection from speech characteristics: Extension beyond the range of uniqueness," in *2002 11th European Signal Processing Conference*, 2002, pp. 1–4.

[5] United States. Department of the Army and United States. Department of the Air Force, *Fundamentals of Single-sideband Communication*, ser. Air Force TO, United States. Department of the Army and United States. Department of the Air Force, Ed. Departments of the Army and the Air Force, 1961. [Online]. Available: https://books.google.de/books?id=mcEXAAAAYAAJ

[6] S. Ganapathy and J. Pelecanos, "Enhancing frequency shifted speech signals in single side-band communication," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1231–1234, 2013.

[7] J. Schmalenstroeer, J. Heitkaemper, J. Ullmann, and R. Haeb-Umbach, "Open range pitch tracking for carrier frequency difference estimation from hf transmitted speech," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1–5.

[8] J. Heitkaemper, J. Schmalenstroeer, and R. Haeb-Umbach, "Statistical and Neural Network Based Speech Activity Detection in Non-Stationary Acoustic Environments," in *Proc. Interspeech 2020*, 2020, pp. 2597–2601.

[9] 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. PP, pp. 1–1, May 2019.

[10] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "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)*, 2020, pp. 7009–7013.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[13] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," *CoRR*, vol. abs/1809.07454, 2018.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[16] J. Seamons, *Kiwi-SDR*, 2021 (accessed July 21, 2021), http://kiwisdr.com.

[17] J. Heitkaemper, J. Schmalenstroeer, J. Ullmann, V. Ion, and R. Haeb-Umbach, "A Database for Research on Detection and Enhancement of Speech Transmitted over HF links," *arXiv e-prints*, p. arXiv:2106.02472, Jun. 2021.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.