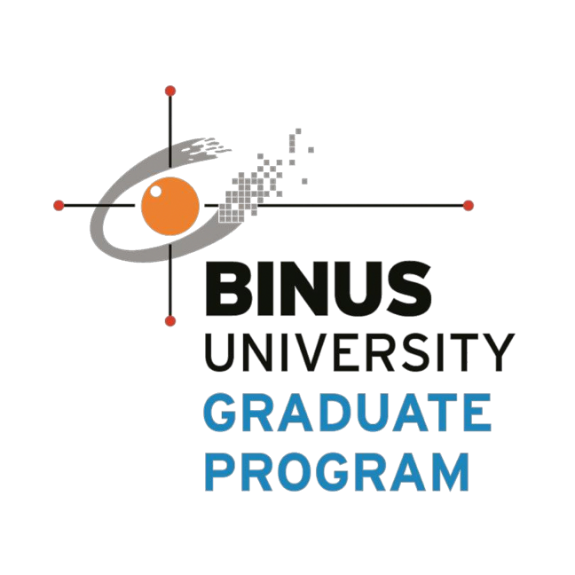
**PROPOSAL**

**Pendekatan *Deep learning attentive recurrent network* dengan *model compression* dan *quantization* untukmengurangi kebisingan agar meningkatkan kejelasan suara manusia**



**RESEARCH**

*SHOFYAN*

*2301980064*

**Program Pascasarjana Ilmu Komputer**

**PROGRAM STUDI TEKNIK INFORMATIKA JENJANG S2**

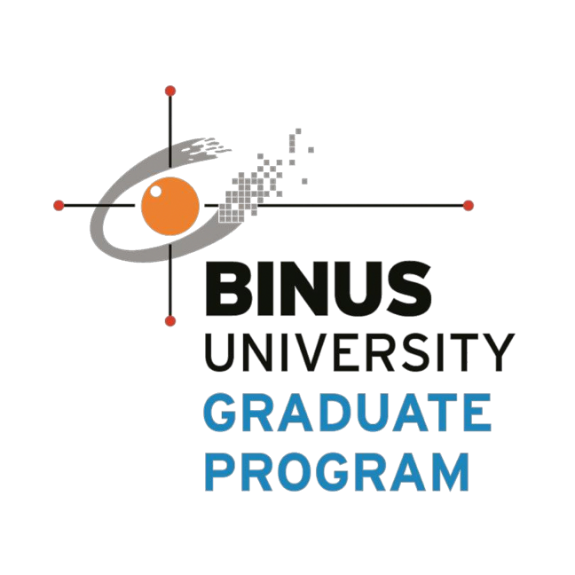
**UNIVERSITAS BINA NUSANTARA**

**JAKARTA**

**2022**

**PROPOSAL**

**Pendekatan *Deep learning attentive recurrent network* dengan *model compression* dan *quantization* untukmengurangi kebisingan agar meningkatkan kejelasan suara manusia**



**RESEARCH**

*SHOFYAN*

*2001847780*

Pembimbing:

**Iman Herwidiana Kartowisastro, Ph.D.**

03-20-2022

# Daftar Isi

[Daftar Isi iii](#_Toc104728500)

[Daftar Gambar v](#_Toc104728501)

[Daftar Persamaan vi](#_Toc104728502)

[BAB 1 PENDAHULUAN 1](#_Toc104728503)

[1.1 Latar Belakang 1](#_Toc104728504)

[1.2 Rumusan Masalah 3](#_Toc104728505)

[1.3 Tujuan dan Manfaat 3](#_Toc104728506)

[1.4 Ruang Lingkup 4](#_Toc104728507)

[BAB 2 LANDASAN TEORI 5](#_Toc104728508)

[2.1 Perkembangan *deep learning* untuk menghilangkan kebisingan dan meningkatkan kejelasan suara manusia 5](#_Toc104728509)

[*2.2* *Digital Signal Processing* 16](#_Toc104728510)

[2.3 Machine Learning 17](#_Toc104728511)

[2.4 Deep Learning 18](#_Toc104728512)

[2.5 Optimizer algorithm 18](#_Toc104728513)

[2.5.1 Gradient Descent 19](#_Toc104728514)

[2.5.2 Momentum 21](#_Toc104728515)

[2.5.3 Adam 21](#_Toc104728516)

[BAB 3 METODOLOGI 23](#_Toc104728517)

[3.1 Metodologi Penelitian 23](#_Toc104728518)

[3.2 Pengumpulan dan Prapemrosesan Dataset 24](#_Toc104728519)

[3.3 Ekstraksi Fitur 25](#_Toc104728520)

[3.4 Pemisahan Data 27](#_Toc104728521)

[3.5 Permodelan *Attentive Recurrent Network* 27](#_Toc104728522)

[3.5.1 Layer Normalization 28](#_Toc104728523)

[3.5.2 Recurrent Neural Network 28](#_Toc104728524)

[3.5.3 Self-attention Block 29](#_Toc104728525)

[3.5.4 Feedforward Block 30](#_Toc104728526)

[3.6 Klasifikasi dan Evaluasi 31](#_Toc104728527)

[DAFTAR PUSTAKA 32](#_Toc104728528)

# Daftar Gambar

[Gambar 3.1 Metode Penelitian 23](#_Toc101738484)

[Gambar 3.2 Diagram ARN. 27](#_Toc101738485)

[Gambar 3.3 Diagram ilustrasi dari LSTM 28](#_Toc101738486)

[Gambar 3.4 *Feedforward Block* 29](#_Toc101738487)

# Daftar Persamaan

[Persamaan 2.1 *Batch gradient descent* 19](#_Toc104728530)

[Persamaan 2.2 *Stochastic gradient descent* 20](#_Toc104728531)

[Persamaan 2.3 *Mini batch gradient descent* 20](#_Toc104728532)

[Persamaan 2.4 Kecepatan momentum 21](#_Toc104728533)

[Persamaan 2.5 Bobot momentum 21](#_Toc104728534)

[Persamaan 2.6 *Mean* 22](#_Toc104728535)

[Persamaan 2.7 *Variance* 22](#_Toc104728536)

[Persamaan 2.8 *Update weight* 22](#_Toc104728537)

[Persamaan 3.1 Sinyal suara bising 25](#_Toc104728538)

[Persamaan 3.2 Peningkatan suara manusia domain waktu 26](#_Toc104728539)

[Persamaan 3.3 Matriks berisi bingkai sinyal 26](#_Toc104728540)

[Persamaan 3.4 Pemrosessan binkai 26](#_Toc104728541)

[Persamaan 3.5 Peningkatan bicara kausal 27](#_Toc104728542)

[Persamaan 3.6 *Layer normalization* 28](#_Toc104728543)

[Persamaan 3.7 Tanh(s) 29](#_Toc104728544)

[Persamaan 3.8 *Correlation score* 30](#_Toc104728545)

[Persamaan 3.9 *Output self attention block* 30](#_Toc104728546)

[Persamaan 3.10 *Softmax* 30](#_Toc104728547)

# PENDAHULUAN

## Latar Belakang

Pandemi *Covid-19* pada akhirnya mengubah cara bekerja yang sebelumnya betatap muka langsung di kantor, kini harus bekerja dari rumah (*Work from Home*), komunikasi menggunakan media *video call* sudah menjadi bagian dari koordinasi dengan rekan kerja. Keadaan ruangan dirumah tidak seperti di kantor yang sudah diatur sedemikian rupa sehingga bebas dari gannguan suara seperti kendaraan yang lewat, anak kecil yang teriak, suara pedagang asongan menjajakan makanan, suara tetangga sedang memperbaiki rumah dan lain lain. Hal ini menimbulkan gangguan suara ketika berkomunikasi dengan rekan kerja. Pada kasus ini kita membutuhkan penghilang kebisingan suara ketika bekomunikasi sehingga pesan yang disampaikan dapat diterima dengan jelas.

Pendekatan statistika dalam menghilangkan kebisingan sudah menjadi masalah yang menarik bagi para peneliti mulai dari metode *Wiener filtering* (Scalart, 1996), *minimum mean-square error short-time spectral amplitude* (MMSE-STSA) *estimator* (Ephraim, 1984), dan *MMSE log-spectral amplitude* (MMSE-LSA) *estimator* (Epraim, 1985). Metode ini sering dibuat degan beberapa asumsi yang mendasar stasioner, dan tidak memiliki kemampuan untuk menangani kebisingan non-stasioner (Zhang, 2019).

Dalam beberapa tahun terakhir, metode peningkatan kualitas suara manusia *supervised* berbasis *deep neural network* (DNN) telah menunjukkan kemajuan besar. Dalam metode peningkatan kualitas suara manusia yang *supervised*, desain target pelatihan memainkan peran penting. Banyak metode DNN telah diadopsi dari pemrosesan gambar, ada perbedaan penting antara domain pada audio. Sampel audio mentah membentuk sinyal deret waktu satu dimensi yang pada dasarnya berbeda dari gambar yang membentuk dua dimensi. Sinyal audio biasanya diubah menjadi representasi waktu-frekuensi dua dimensi untuk diproses, tetapi dua sumbu, waktu dan frekuensi, tidak homogen seperti sumbu horizontal dan vertikal dalam sebuah gambar. Gambar adalah *snapshot* instan dari target dan sering dianalisis secara keseluruhan atau sebagian dengan sedikit batasan urutan; namun sinyal audio harus dipelajari secara berurutan dalam urutan kronologis (Purwins, 2019). Properti ini memunculkan solusi khusus audio yang berdasarkan urutan waktu.

Pendekatan metode *convolutional recurrent neural network* (Tan, 2018) untuk memisahkan kebisingan membutuhkan banyak biaya baik dari sumber daya komputasi maupun penggunaan *memory* (Tan, 2021). Baru-baru ini, model yang disebut *temporal convolutional network* (TCN) (Shaojie, 2018), yang menggunakan pembelajaran residual yang digabungkan dengan unit konvolusi yang melebar, dapat menangkap ketergantungan kontekstual jangka panjang dan mengungguli penelitian , (Jitong, 2017) yang mengusulkan penggunaan *recurrent neural network* (RNN) dengan empat *hidden long shortterm memory* (LSTM). Penggunaan metode TCN ini juga membutuhkan tenaga komputasi yang besar dan juga *memory*  yang banyak. Penelitian kali ini mengusulkan kompresi pada metode TCN tanpa mengurangi kualitas secara signifikan dengan mengkompresi hidden layer TCN

## Rumusan Masalah

Mengacu pada temuan masalah pada latar belakang di atas maka berikut adalah rumusan masalah dalam penelitian ini adalah:

Bagaimana membuat model *deep learning attentive recurrent network* yang lebih efisien pada penggunaan *resource* dengan *model compression* dan *quantization* dalam mengurangi kebisingan dan meningkatkan kualitas suara manusia?

## Tujuan dan Manfaat

Tujuan penelitian ini adalah:

* Mengembangkan sistem pengurangan kebisingan dan peningkatan kualitas suara manusia dengan penggunaan *deep learning attentive recurrent network* dengan *model compression* dan *quantization* agar dapat digunakan pada kasus nyata di kehidupan sehari-hari

Dengan diadakannya penelitian ini, manfaat-manfaat yang didapatkan dijabarkan sebagai berikut:

* Memberikan rancangan dan arsitektur model untuk pengurangan kebisingan dan peningkatan kualitas suara manusia yang dapat diimplementasikan dikehidupan sehari hari.
* Menambah kejelasan suara manusia ketika pertemuan menggunakan *video conference* walaupun ada kebisingan suara.

## Ruang Lingkup

Ruang lingkup penelitian ini adalah:

* Sumber suara berasar dari *monaural* (satu*-*mikropon)
* Suara kebisingan dibuat dengan menggabungkan suara manusia dengan sumber suara kebisingan
* Dataset yang digunakan diambil dari Librispeech (Panayotov, 2015)

# LANDASAN TEORI

## Perkembangan *deep learning* untuk menghilangkan kebisingan dan meningkatkan kejelasan suara manusia

Penggunaan metode deep learning menggunakan *Convolutional Recurrent Neural Network* (CRN) (Ke Tan, 2018) mengungguli dua dasar nilai dari metode long short-term memory (LSTM) untuk pembicara terlatih dan tidak terlatih dalam hal skor short-time objective intelligibility (STOI) dan perceptual evaluation of speech quality (PESQ).

Dilanjut dengan pendekatan *Minimum Mean-Square Error* (MMSE) (Aaron, 2019) yang memanfaatkan *deep learning* mampu menghasilkan ucapan yang disempurnakan yang mencapai skor kualitas dan kejelasan yang lebih tinggi daripada pendekatan pembelajaran mendalam berbasis masking untuk sumber kebisingan non-stasioner di dunia nyata.

Pada tahun 2020 (Hu, 2020) [Yanxin Hu](https://arxiv.org/search/eess?searchtype=author&query=Hu%2C+Y) mengembangkan *Deep complex convolution recurrent network*, (DCCRN). Model DCCRN menggunakan jaringan kompleks untuk pemodelan spektrum bernilai kompleks. Dengan batasan aturan perkalian yang kompleks, DCCRN dapat mencapai kinerja yang lebih baik daripada yang lain (LSTM, CRN) dalam hal PESQ dan  *Mean Opinion Score* (MOS).

Di tahun yang sama (Li, 2020) [Jingdong Li](https://arxiv.org/search/cs?searchtype=author&query=Li%2C+J) mengusulkan *Temporal Convolutional Recurrent Neural Networks* (TCRN) secara signifikan mengungguli baseline LSTM dengan margin yang besar. TCRN secara konsisten lebih baik daripada CRN. TCRN memiliki kemampuan generalisasi kebisingan yang lebih baik daripada *baseline*.

Tahun 2021 Qiquan Zhang (Zhang, 2021) mengusulkan *adaptive attention module* (FAA) untuk meningkatkan model *Temporal Convolutional Network* (TCN) untuk tugas peningkatan suara monaural. Modul FAA terintegrasi ke dalam TCN dan GaTCN untuk melakukan peningkatan suara. Hasil metrik objektif menunjukkan bahwa TCN-FAA dan GaTCN-FAA yang diusulkan secara konsisten mengungguli model dasarnya tanpa modul FAA.

Di tahun yang sama 2021 Ashutosh (Pandey, 2021) mengusulkan *Dense convolutional network* (DCN) yang dilatih pada korpus WSJ0 SI-84 dataset dan dievaluasi pada pembicara WSJ yang tidak terlatih. Peningkatan kemampuan bicara berbasis DNN gagal digeneralisasi menjadi tidak terlatih corpora, dan kinerja yang lebih baik pada corpus yang terlatih tidak tentu mengarah pada kinerja yang lebih baik pada corpora yang tidak terlatih.

Ashutosh (Pandey, 2021) juga mengusulkan *dual-path self-attention recurrent neural network* (DPSARNN)yang diusulkan menambah RNN di DP-RNN dengan perhatian. Menambahkan perhatian di DP-RNN mengarah pada peningkatan yang lebih baik dengan empat kali pergeseran bingkai, menghasilkan model latensi rendah. Sebagai hasilnya, mengembangkan versi real-time dari DP-SARNN yang tidak hanya lebih cepat dari tetapi juga mengungguli pendekatan yang ada.

Tahun 2022 Ruizhe Cao(Cao, 2022) mengusulkan *conformer-based metric generative adversarial network (CMGAN).*  Hasil eksperimen menunjukkan bahwa metode yang diusulkan mengungguli *state of art* saat ini pada set data Voice Bank+DEMAND dengan parameter yang relatif sedikit (1,83 M).

Ashutosh pada tahun 2022 (Pandey, 2022) mengusulkan *self-attending recurrent neural network, or attentive recurrent network (ARN),* Hasil percobaan telah menunjukkan keunggulan ARN atas kompetitif algoritma, seperti RNN, DCCRN, DCN dan DPARN. Arah penelitian selanjutnya akan mengoptimalkan ARN untuk aplikasi dunia nyata dengan menggunakan teknik seperti kompresi model dan kuantisasi . Arah penelitian terkait adalah untuk mengeksplorasi arsitektur DNN yang memiliki jumlah parameter lebih sedikit tetapi memberikan generalisasi lintas korpus yang baik.

|  |  |  |  |
| --- | --- | --- | --- |
| Tahun | Penulis | Metode | Hasil |
| 2018 | Ke Tan, DeLiang Wang | *Convolutional Recurrent Neural Network* (CRN) | Hasil evaluasi menunjukkan bahwa *Convolutional Recurrent Neural Network* yang diusulkan secara konsisten mengungguli dua dasar nilai dari metode *long short-term memory* (LSTM) untuk pembicara terlatih dan tidak terlatih dalam hal skor STOI dan PESQ. |
| 2019 | Aaron Nicolson, Kuldip K. Paliwal | Metode *Deep learning* untuk *Minimum Mean-Square Error* (MMSE) | MMSE yang memanfaatkan *deep learning* mampu menghasilkan ucapan yang disempurnakan yang mencapai skor kualitas dan kejelasan yang lebih tinggi daripada pendekatan pembelajaran mendalam berbasis masking dan pemetaan baru-baru ini, untuk sumber kebisingan non-stasioner di dunia nyata. |
| 2020 | [Yanxin Hu](https://arxiv.org/search/eess?searchtype=author&query=Hu%2C+Y), [Yun Liu](https://arxiv.org/search/eess?searchtype=author&query=Liu%2C+Y), [Shubo Lv](https://arxiv.org/search/eess?searchtype=author&query=Lv%2C+S), [Mengtao Xing](https://arxiv.org/search/eess?searchtype=author&query=Xing%2C+M), [Shimin Zhang](https://arxiv.org/search/eess?searchtype=author&query=Zhang%2C+S), [Yihui Fu](https://arxiv.org/search/eess?searchtype=author&query=Fu%2C+Y), [Jian Wu](https://arxiv.org/search/eess?searchtype=author&query=Wu%2C+J), [Bihong Zhang](https://arxiv.org/search/eess?searchtype=author&query=Zhang%2C+B), [Lei Xie](https://arxiv.org/search/eess?searchtype=author&query=Xie%2C+L) | *Deep complex convolution recurrent network* (DCCRN) | Dalam penelitian ini, (DCCRN). Model DCCRN menggunakan jaringan kompleks untuk pemodelan spektrum bernilai kompleks. Dengan batasan aturan perkalian yang kompleks, DCCRN dapat mencapai kinerja yang lebih baik daripada yang lain dalam hal PESQ dan MOS dalam konfigurasi parameter model yang serupa. |
| 2020 | [Jingdong Li](https://arxiv.org/search/cs?searchtype=author&query=Li%2C+J), [Hui Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C+H), [Xueliang Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C+X), [Changliang Li](https://arxiv.org/search/cs?searchtype=author&query=Li%2C+C) | *Temporal Convolutional Recurrent Neural Networks* (TCRN) | TCRN yang diusulkan secara signifikan mengungguli baseline LSTM dengan margin yang besar. Dan TCRN secara konsisten lebih baik daripada CRN. TCRN memiliki kemampuan generalisasi noise yang lebih baik daripada baseline. |
| 2021 | Qiquan Zhang , Qi Song , Aaron Nicolson, Tian Lan, Haizhou Li | *Temporal Convolutional Network with Frequency Dimension Adaptive* | *adaptive at- tention module* (FAA) untuk meningkatkan model TCN untuk tugas peningkatan suara monaural. Modul FAA terintegrasi ke dalam TCN dan GaTCN untuk melakukan peningkatan suara. Hasil metrik objektif menunjukkan bahwa TCN-FAA dan GaTCN-FAA yang diusulkan secara konsisten mengungguli model dasarnya tanpa modul FAA. Lebih menjanjikan, itu juga menunjukkan kemampuan generalisasi yang sangat baik untuk lingkungan yang bising nyata. |
| 2021 | Ashutosh Pandey, DeLiang Wang | *Dense convolutional network (DCN) with self-attention for speech enhancement in the time domain.* | DCN dilatih pada korpus WSJ0 SI-84 dataset dan dievaluasi pada pembicara WSJ yang tidak terlatih. Peningkatan kemampuan bicara berbasis DNN gagal digeneralisasi menjadi tidak terlatih  corpora, dan kinerja yang lebih baik pada corpus yang terlatih tidak tentu mengarah pada kinerja yang lebih baik pada corpora yang tidak terlatih |
| 2021 | Ashutosh Pandey, DeLiang Wang | *dual-path self-attention recurrent neural network (DP-SARNN) for time-domain speech enhancement* | . DPSARNN yang diusulkan menambah RNN di DP-RNN dengan perhatian. Menambahkan perhatian di DP-RNN mengarah pada peningkatan yang lebih baik dengan empat kali pergeseran bingkai, menghasilkan model latensi rendah. Sebagai hasilnya, mengembangkan versi real-time dari DP-SARNN yang tidak hanya lebih cepat dari tetapi juga mengungguli pendekatan yang ada. |
| 2022 | Ruizhe Cao, Sherif Abdulatif , Bin Yang | *conformer-based metric generative adversarial network (CMGAN) for Speech Enhancement in the time-frequency (TF) domain* | Hasil eksperimen menunjukkan bahwa metode yang diusulkan mengungguli *state of art* saat ini pada set data Voice Bank+DEMAND dengan parameter yang relatif sedikit (1,83 M). |
| 2022 | Ashutosh Pandey; Deliang Wang | *self-attending recurrent neural network, or attentive recurrent network (ARN), for timedomain speech enhancement to improve cross-corpus generalization.* | Hasil percobaan telah menunjukkan keunggulan ARN atas kompetitif algoritma, seperti RNN, DCCRN, DCN dan DPARN. |

## *Digital Signal Processing*

Digital signal processing adalah perhitungan, algoritma, dan teknik yang digunakan untuk memanipulasi sinyal yang didapatkan dari data sensor setelah diubah ke dalam bentuk digital (Smith, 1999). Sinyal yang didapatkan umumnya berasal dari data sensor pada kehidupan nyata, seperti getaran seismik, citra visual, gelombang suara, dan lainlain. Digital signal processing banyak digunakan pada berbagai bidang, salah satunya pada bidang speech recognition. Suara yang didapatkan melalui input analog akan disaring dan diubah menjadi digital audio pada proses analog-to-digital conversion (ADC). Suara yang telah diubah ke dalam bentuk digital akan diproses pada digital signal processing (DSP) dan dilanjutkan dengan mengubah suara bentuk digital tersebut ke dalam bentuk analog pada proses digital-to-analog conversion (DAC). Tidak semua implementasi digital signal processing melewati proses analog-to-digital conversion dan digital-to-analog conversion. Speech recognition tidak membutuhkan proses digital-to-analog conversion karena aplikasi cukup mengenali suara yang didapatkan dan mengubahnya menjadi teks. Text-to-speech membutuhkan proses digital-to-analog conversion untuk dapat mengucapkan kata-kata tersebut melalui alat output yang telah disediakan dan tidak membutuhkan analog-to-digital conversion untuk menerima input dari pengguna.

## Machine Learning

Dalam Kecerdasan Buatan, ada konsep di mana mesin mulai belajar dari data tanpa diprogram. Konsep ini disebut Machine Learning. Machine Learning dapat didefinisikan secara luas sebagai metode komputasi yang menggunakan pengalaman untuk meningkatkan kinerja atau membuat prediksi yang akurat (Mohri dkk., 2013). Ada banyak aplikasi pembelajaran mesin, seperti pengenalan gambar, pengenalan ucapan, pengenalan emosi, prediksi stok, dan banyak lagi. Pembelajaran mesin, berdasarkan teknik, dapat dibagi lagi menjadi dua kategori, yaitu pembelajaran terawasi dan pembelajaran tanpa pengawasan.

Pembelajaran terawasi membangun model berdasarkan data dan label yang ada dan menggunakannya untuk memprediksi label untuk input baru, sedangkan Pembelajaran tanpa pengawasan digunakan untuk menemukan pola tersembunyi dalam kumpulan data. Aplikasi pembelajaran terbimbing adalah klasifikasi dan regresi, di mana keduanya memprediksi sesuatu, sedangkan salah satu aplikasi pembelajaran tanpa pengawasan yang umum adalah pengelompokan, di mana kami menemukan pola/kelompok tertentu dalam kumpulan data.

## Deep Learning

Deep Learning memungkinkan model komputasi yang terdiri dari beberapa lapisan pemrosesan untuk mempelajari representasi data dengan berbagai tingkat abstraksi (Lecun dkk., 2015). Ada perbedaan antara Deep Learning dan Machine Learning, seperti jumlah lapisan dan keberadaan fitur. Machine Learning membuat model berdasarkan data yang diawasi sedangkan Deep Learning menyusun jaringan untuk meniru otak yang dapat mempelajari pola dan mengekstrak fitur dari sejumlah besar data untuk membuat model yang optimal.

Deep Learning menggabungkan kemajuan dalam daya komputasi dan jaringan saraf dengan banyak lapisan untuk mempelajari pola rumit dalam data dalam jumlah besar. Ini adalah perpanjangan dari jaringan saraf klasik dan menggunakan lebih banyak lapisan tersembunyi sehingga algoritma dapat menangani data yang kompleks dengan berbagai struktur (Muniasamy & Alasiry, 2020).

## Optimizer algorithm

Algoritme pengoptimal digunakan oleh jaringan saraf untuk mengontrol perubahan bobot dan kecepatan pembelajaran. Ada beberapa algoritma pengoptimal, seperti Gradient Descent, Momentum, Adam (Adaptive Moment Estimation), dll.

### Gradient Descent

Gradient Descent adalah algoritma optimasi yang paling dasar dan paling banyak digunakan. Algoritma Gradient Descent menghitung arah bobot berdasarkan error untuk mendapatkan fungsi biaya mencapai nilai maksimum/minimum. Gradient Descent memiliki beberapa variasi, seperti Batch Gradient Descent, Mini Batch Gradient Descent, dan Stochastic Gradient Descent.

#### Batch Gradient Descent

Batch Gradient Descent disebut vanilla gradient descent. Batch Gradient Descent memperbarui parameter model setelah semua data pelatihan dievaluasi. Batch Gradient Descent memiliki keunggulan dalam stable error gradient dan stable convergence. Kerugian dari Batch Gradient Descent adalah kemungkinan besar terjebak pada konvergensi yang bukan yang terbaik yang dapat dicapai model.

𝑤𝑡+1 = 𝑤𝑡 −𝑎.∇𝑤 𝐽(𝑤)

Persamaan . *Batch gradient descent*

Pada persamaan (2.1), nilai baru bobot (w) diperoleh dengan mengurangi bobot terakhir dengan nilai gradien, yang diperoleh dengan nilai turunan bobot dikalikan dengan learning rate.

#### Stochastic Gradient Descent

Stochastic Gradient Descent memperbarui parameter model untuk setiap contoh pelatihan. Keuntungan Stochastic Gradient Descent adalah biaya waktu konvergensi yang lebih sedikit dan memori yang lebih sedikit karena tidak perlu menyimpan nilai fungsi kerugian. Kerugiannya adalah perubahan parameter memiliki varians yang tinggi dan biaya komputasi yang tinggi.

𝑤𝑡+1=𝑤 −𝑎.∇𝑤𝐽(𝑥𝑖,𝑦𝑖;𝑤)

Persamaan . *Stochastic gradient descent*

Pada Persamaan 2.2, nilai bobot baru diperoleh dengan mengurangi bobot terakhir dengan turunan bobot pada setiap data dikalikan dengan learning rate.

#### Mini Batch Gradient Descent

Mini Batch Gradient Descent merupakan kombinasi dari keunggulan Stochastic Gradient Descent dan Batch Gradient Descent. Ini membagi data pelatihan menjadi batch yang berbeda dan memperbarui parameter model setelah setiap batch selesai.

𝑤𝑡+1 = 𝑤𝑡 − 𝑎. ∇𝑤 𝐽(𝑥{𝑖:𝑖+𝑏}, 𝑦{𝑖:𝑖+𝑏}; 𝑤)

Persamaan . *Mini batch gradient descent*

Dalam persamaan (2.3), penurunan gradien mini-batch memperoleh bobot baru dari bobot terakhir dikurangi turunan bobot yang dihitung dengan kumpulan data saat ini dikalikan dengan kecepatan pembelajaran.

### Momentum

Momentum diciptakan untuk memperbaiki kelemahan Stochastic Gradient Descent dalam varian tinggi dan melunakkan konvergensi. Dalam persamaannya, variabel tambahan 'γ' disebut momentum saat memperbarui bobot. Dalam menerapkan momentum, ada variabel yang disebut kecepatan (V) yang menentukan kecepatan pertumbuhan berat.

𝑉t+1 =𝛾∗𝑉t + 𝑎.∇w 𝐽(𝑥𝑖,𝑦𝑖;𝑤)

Persamaan . Kecepatan momentum

𝑤 𝑡+1 = 𝑤 t −𝑉 𝑡+1

Persamaan . Bobot momentum

Pada persamaan (2.4), nilai kecepatan diperoleh dengan nilai variabel momentum dikalikan dengan arah dan ditambah dengan turunan bobot pada setiap data. Persamaan (2.5) menunjukkan bobot baru adalah hasil bobot terakhir dikurangi nilai kecepatan.

### Adam

Algoritma Adam Optimizer adalah algoritma yang diusulkan ditujukan untuk optimasi stokastik efisien yang hanya membutuhkan gradien orde pertama dengan sedikit kebutuhan memori (Kingma & Ba, 2015). Ide di balik Adam adalah untuk mengurangi kecepatan sehingga tidak melewati nilai minimum/maksimum. Algoritma Adam Optimizer menghitung 2 Momen M(t) dan V(t) yang masing-masing disebut mean dan varians tidak terpusat. Persamaan 2.6 menunjukkan rumus untuk mean dan persamaan 2.7 menunjukkan rumus untuk varians tidak terpusat.

𝑚𝑡 = 𝛽1 ∗ 𝑚𝑡−1 +(1−𝛽1)∗ ∇𝑤𝐽(𝑤)

Persamaan . *Mean*

𝑣𝑡 = 𝛽2 ∗ 𝑣𝑡−1 +(1−𝛽2)∗ ∇𝑤𝐽(𝑤)2

Persamaan . *Variance*

𝑤𝑡=𝑤𝑡−1− 𝛼∗𝑚𝑡 √𝑣𝑡 + 𝜀

Persamaan . *Update weight*

Dengan Adam, laju pembelajaran pengoptimal berubah selama pelatihan yang disebut peluruhan laju pembelajaran. Nilai peluruhan digunakan dalam mean dan varians tidak terpusat dengan simbol ''. Dengan fitur ini, menjadi banyak pilihan favorit selain SGD tradisional, yang mempertahankan tingkat pembelajaran tunggal untuk semua pembaruan bobot. Rumus untuk mengupdate bobot pada Adam dapat dilihat pada persamaan 2.8.

# METODOLOGI

## Metodologi Penelitian

Pengurangan kebisingan dan meningkatkan kejelasan suara merupakan aspek penting yang dapat digunakan di banyak bidang. Teknologi ini dapat diterapkan di berbagai bidang pekerjaan misalnya pada *video call meeting*. Penelitian tentang Pengurangan kebisingan dan meningkatkan kejelasan suara telah ada sejak lama. Ada banyak model yang ada untuk mendeteksi kebisingan suara. Penelitian ini bertujuan untuk menghasilkan model yang lebih baik dari beberapa model yang ada dari segi efektifitas dan penggunaan ongkos komputasi.

Model yang diusulkan menggunakan *Deep learning* dengan kombinasi ekstraksi tingkat fitur dari dataset yang berbeda karena metode yang diusulkan memiliki kemungkinan untuk mengungguli kinerja dari model yang diusulkan sebelumnya . Garis besar penelitian dapat dilihat pada Gambar 3.1.

Diagram

Description automatically generated

Gambar 3.1 Metode Penelitian

## Pengumpulan dan Prapemrosesan Dataset

Kami mengevaluasi semua model dengan cara independen speaker, noise, dan corpus. Kami menggunakan semua ucapan dari set pelatihan korpus LibriSpeech (Panayotov, 2015) untuk menghasilkan campuran pelatihan. Ini terdiri dari sekitar 280 ribu ucapan pidato lebih dari 2000 pembicara. LibriSpeech telah terbukti menjadi korpus yang efektif untuk generalisasi lintas korpus karena direkam oleh banyak sukarelawan di seluruh dunia, dan karenanya terdiri dari ucapan-ucapan yang direkam dalam kondisi akustik yang berbeda. Ucapan pelatihan yang bising dihasilkan secara online selama pelatihan dengan cara berikut. Untuk setiap sampel dalam kelompok tertentu, kami secara acak mengambil sampel ucapan ucapan, mengekstrak potongan acak 4 detik darinya, dan menambahkan potongan acak kebisingan ke dalamnya dengan *signal to noise ratio* (SNR) acak dari {−5, 4, 3, 2, 1, 0} dB. Pidato sampel digunakan tidak berubah jika durasinya lebih kecil dari 4 detik. Satu set 224 suara non-suara manusia dari perpustakaan efek suara (docbox.etsi.org) digunakan sebagai suara pelatihan.

Semua model dievaluasi pada tiga corpora yang berbeda: WSJ-SI-84 (WSJ) (Paul dkk, 1992), TIMIT (Garofolo, 1993), dan IEEE (Rothauser, 1996), yang tidak digunakan selama pelatihan. Kami menggunakan ucapan satu pembicara pria dan satu pembicara wanita dari IEEE. Perangkat tes WSJ terdiri dari 150 ucapan dari 6 pembicara yang berbeda. Set tes TIMIT terdiri dari 192 ucapan di set tes inti. IEEE Pria dan IEEE Wanita masing-masing terdiri dari 144 ucapan yang dipilih secara acak. Kami menghasilkan ucapan berisik menggunakan empat jenis suara yang berbeda: celoteh, kafetaria, pabrik, dan mesin, tidak ada yang digunakan selama pelatihan. Ucapan uji dihasilkan pada 6 SNR berbeda 5, 2, 0, 2, dan 5 dB

## Ekstraksi Fitur

Sinyal suara bising x didefinisikan sebagai jumlah dari sinyal suara bersih s dan sinyal suara n, digambarkan dengan Persamaan 3.1

x=s+n

Persamaan . Sinyal suara bising

{x,s,n} RM×1, dan M adalah jumlah sampel dalam sinyal suara. Algoritme peningkatan suara manusia bertujuan untuk mendapatkan perkiraan yang mendekati, s, dari s yang diberikan x.Tujuan dari algoritma peningkatan ucapan domain waktu adalah untuk menghitung s langsung dari x daripada menggunakan representasi T-F dari x. Peningkatan suara manusia domain waktu menggunakan DNN dapat dirumuskan sebagai Persamaan 3.2

=fθ(x)

Persamaan . Peningkatan suara manusia domain waktu

di mana fθ menunjukkan fungsi yang diwakili oleh DNN diparametrisasi oleh θ

Secara umum, algoritma peningkatan suara dirancang untuk memproses bingkai sinyal suara. Diberikan sinyal noise x, pertama-tama dipotong menjadi bingkai yang tumpang tindih yang kemudian diproses pada tingkat bingkai oleh model peningkatan suara. Biarkan X RT ×L menunjukkan matriks yang berisi bingkai sinyal x dan xt RL×1 bingkai ke-t. xt didefinisikan sebagai Persamaan 3.3

xt[k]=x[(t−1)·J+k], k=0,···,L−1

Persamaan . Matriks berisi bingkai sinyal

di mana T adalah jumlah bingkai, L adalah panjang bingkai, dan J adalah pergeseran bingkai. T diberikan oleh M , di mana menunjukkan J fungsi langit-langit. x diisi dengan nol jika M tidak habis dibagi J. Pemrosesan bingkai menggunakan DNN dapat didefinisikan sebagai Persamaan 3.4

t = fθ(xt−T1,··· ,xt−1,xt,xt+1,··· ,xt+T2)

Persamaan . Pemrosessan binkai

di mana t dihitung menggunakan xt, T1 frame masa lalu, dan T2 masa depan bingkai.

Algoritme peningkatan ucapan tingkat bingkai dianggap kausal jika estimasi dari s bingkai yang diberikan dihitung menggunakan bingkai yang bising pada waktu yang kurang dari atau sama dengan t. Untuk peningkatan bicara kausal dimodifikasi sebagaimana digambarkan pada Persamaan 3.5

t =fθ(xt−T1,···,xt−1,xt)

Persamaan . Peningkatan bicara kausal

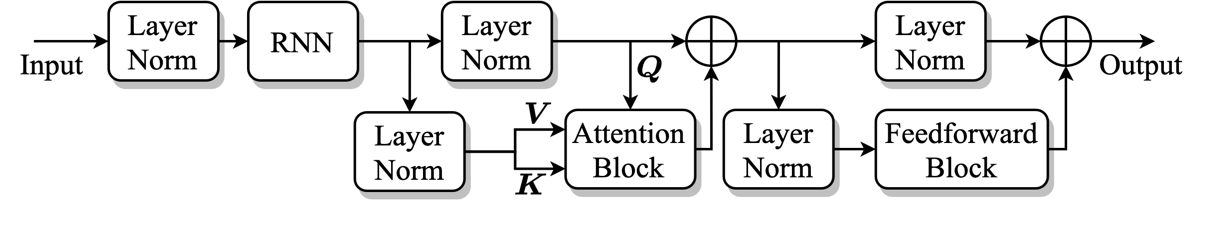
di mana t dihitung menggunakan xt dan T1 melewati frame. Kausalitas adalah persyaratan yang diperlukan untuk peningkatan kemampuan berbicara secara real-time. Lebih lanjut, kami mengamati bahwa algoritme kausal menunjukkan degradasi yang lebih besar pada korpora yang tidak terlatih dibandingkan dengan algoritme non-kausal yang sesuai. Oleh karena itu, kami juga mengembangkan dan membandingkan algoritma kausal.

## Pemisahan Data

Dataset yang digunakan oleh model dipisahkan menjadi set data splitting. Penelitian ini mengikuti metode tradisional dengan membagi dataset menjadi tujuan pelatihan dan pengujian. Langkah pertama dari pemisahan data adalah mengacak dataset untuk membuat data yang lebih merata. Setelah itu, kami membagi kereta dan menguji dengan rasio 8:2, 8 untuk kereta dan 2 untuk validasi untuk menghindari overfitting atau underfitting. Rasio 8:2 diterapkan untuk semua model yang diusulkan.

## Permodelan *Attentive Recurrent Network*

Sebuah diagram blok ARN diberikan pada Gambar 3.2. Blok bangunan ARN adalah normalisasi lapisan, RNN, blok self-attention, dan blok feedforward. Selanjutnya, kami menjelaskan blok bangunan ini satu per satu.



Gambar 3.2 Diagram ARN.

### *Layer Normalization*

*Layer Normalization* adalah teknik normalisasi populer yang digunakan dalam DNN untuk meningkatkan generalisasi dan memfasilitasi pelatihan yang lebih cepat (Ba, 2016). Hal ini diusulkan sebagai alternatif untuk normalisasi *batch* (Ioffe, 2015), yang ditemukan lebih sensitif terhadap pelatihan *batch* sebagaimana digambarkan di Persamaan 3.6

Diagram

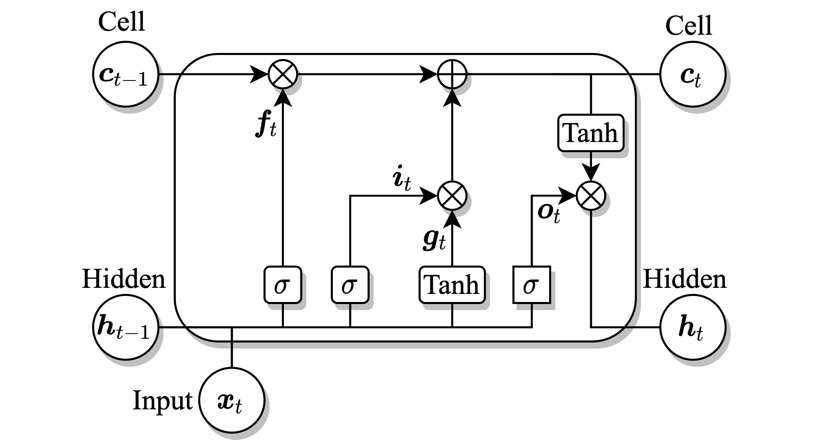
Description automatically generated with medium confidence

Persamaan . *Layer normalization*

dimana xt dan , berturut-turut adalah mean dan varians dari . Simbol dan adalah parameter yang dapat dilatih dengan ukuran yang sama dengan xt, ⊙ menunjukkan perkalian elemen, dan adalah konstanta positif kecil yang digunakan untuk menghindari pembagian dengan nol. Menunjukkan perkalian elemen, dan adalah konstanta positif kecil yang digunakan untuk menghindari pembagian dengan nol.

### Recurrent Neural Network

Kami menggunakan *long short-term memory* (LSTM) RNN di ARN. Sebuah diagram ilustrasi dari LSTM ditunjukkan pada Gambar. 3.3.



Gambar 3.3 Diagram ilustrasi dari LSTM

Diberikan urutan vektor input {x1,··· ,xt−1,xt,xt+1,··· ,xT}, keadaan tersembunyi pada waktu t, ht ,dihitung sebagai Persamaan 3.7

Text

Description automatically generated

Persamaan 3.7 Tanh(s)

Di mana xt, gt, dan ct masing-masing mewakili input, input blok, dan status memori (sel) pada waktu t. Selain itu, ft, dan ot masing-masing adalah gerbang yang dikenal sebagai gerbang input, gerbang lupa dan gerbang keluaran. W dan b menunjukkan bobot dan bias yang dapat dilatih.

### Self-attention Block

Mekanisme *self attention* didefinisikan menggunakan tiga komponen: kunci K RT×R, nilai V RT×S, dan kueri Q RT×R. Pertama, skor korelasi antara pasangan baris dari Q dan K, {Qi,Kj}, dimana i,j {1,···,T}, dihitung dengan menggunakan Persamaan 3.8

W = QKT

Persamaan 3.8 *Correlation score*

Output akhir dari *self attention block* dihitung dengan Persamaan 3.9 dengan lapisan *softmax*  yang di representasikan dengan Persamaan 3.10

Text

Description automatically generated

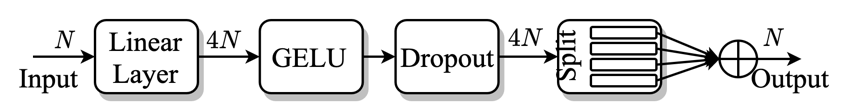
Persamaan 3.9 *Output self attention block*

A = Softmax(W ′)V ′

Persamaan 3.10 *Softmax*

### Feedforward Block

Blok feedforward di ARN ditunjukkan pada Gambar 3.4. Sebuah input yang diberikan dari ukuran N diproyeksikan ke ukuran 4N menggunakan lapisan linier, yang diikuti oleh unit linier kesalahan Gaussian (GELU) [50] dan lapisan putus sekolah [51]. Akhirnya, keluaran berukuran 4N dibagi menjadi empat vektor berukuran N, yang dijumlahkan untuk mendapatkan keluaran akhir.



Gambar 3.4 *Feedforward Block*

Dengan blok bangunan yang dijelaskan, kami sekarang menyajikan aliran pemrosesan ARN yang ditunjukkan pada Gambar. 1. Input ke ARN pertama-tama dinormalisasi dan kemudian diproses menggunakan RNN. Output dari RNN dinormalisasi menggunakan dua normalisasi lapisan paralel. Aliran pertama digunakan sebagai Q dan aliran kedua digunakan sebagai K dan V untuk blok perhatian berikut. Output dari blok perhatian ditambahkan ke Q untuk membentuk koneksi residual. Sekali lagi, output dinormalisasi menggunakan dua normalisasi lapisan paralel. Aliran pertama diproses menggunakan blok *feed forward* dan aliran kedua ditambahkan ke *output blok feed forward* untuk membentuk sambungan sisa.

## Klasifikasi dan Evaluasi

Kami menggunakan *short-time objective intelligibility* (STOI) dan erceptual evaluation of speech quality (PESQ) sebagai metrik evaluasi untuk membandingkan model yang dilatih di Librispeech. STOI memiliki kisaran tipikal [0, 1], yang secara kasar mewakili persen benar. PESQ memiliki rentang [−0.5,4.5], di mana skor yang lebih tinggi menunjukkan kualitas ucapan yang lebih baik. Kedua metrik tersebut biasanya digunakan untuk mengevaluasi algoritma peningkatan bicara

# DAFTAR PUSTAKA

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Cao, R., Abdulatif, S., & Yang, B. (2022). CMGAN: Conformer-based Metric GAN for Speech Enhancement. arXiv preprint arXiv:2203.15149.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109–1121. https://doi.org/10.1109/tassp.1984.1164453

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal* Processing, 33(2), 443–445. https://doi.org/10.1109/tassp.1985.1164550

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon technical report n, 93, 27403.

J. Chen and D. Wang (2017). “Long short-term memory for speaker eneralization in supervised speech separation,” The Journal of the Acoustical Society of merica, vol. 141, no. 6, pp. 4705–4714.

Ke, Y., Li, A., Zheng, C., Peng, R., & Li, X. (2021). Low-complexity artificial noise suppression methods for deep learning-based speech enhancement algorithms. EURASIP Journal on Audio, Speech, and Music Processing, 2021(1), 1-15.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436– 444. https://doi.org/10.1038/nature14539

Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., ... & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264.

Nicolson, A., & Paliwal, K. K. (2019). Deep learning for minimum mean-square error approaches to speech enhancement. Speech Communication, 111, 44–55. https://doi.org/10.1016/j.specom.2019.06.002

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2013). Foundations of Machine Learning second edition. Вестник Казнму, No3(1), с.30.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). 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.

Pandey, A., & Wang, D. (2020). Dual-path Self-Attention RNN for Real-Time Speech Enhancement. arXiv preprint arXiv:2010.12713.

Pandey, A., & Wang, D. (2021). Dense CNN with self-attention for time-domain speech enhancement. IEEE/ACM transactions on audio, speech, and language processing, 29, 1270-1279.

Pandey, A., & Wang, D. (2022). Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Paul, D. B., & Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

Purwins, H., Li, B., Virtanen, T., Schluter, J., Chang, S.-Y., & Sainath, T. (2019). Deep Learning for Audio Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206–219. https://doi.org/10.1109/jstsp.2019.2908700

Scalart, P., & Filho, J. V. (n.d.). Speech enhancement based on a priori signal to noise estimation. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. https://doi.org/10.1109/icassp.1996.543199

S. Bai, J. Z. Kolter, and V. Koltun.( 2018) “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” arXiv preprint arXiv:1803.01271.

Smith, S. W. (1999). Digital signal processing. California Technical Publishing. https://doi.org/10.1109/79.826412

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. IEEE Trans. on Audio and Electroacoustics, 17, 225-246.

Tan, K., & Wang, D. L. (2018). A convolutional recurrent neural network for real-time speech enhancement. *Interspeech 2018*. https://doi.org/10.21437/interspeech.2018-1405

Tan, K., & Wang, D. L. (2021). Compressing deep neural networks for efficient speech enhancement. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp39728.2021.9413536

Zhang, Q., Wang, M., Lu, Y., Zhang, L., & Idrees, M. (2019). A novel fast nonstationary noise tracking approach based on MMSE spectral power estimator. *Digital Signal Processing*, *88*, 41–52. https://doi.org/10.1016/j.dsp.2019.01.019

Zhang, Q., Song, Q., Nicolson, A., Lan, T., & Li, H. (2021). Temporal Convolutional Network with Frequency Dimension Adaptive Attention for Speech Enhancement. Proc. Interspeech 2021, 166-170.