IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Fourier Attention: The Attention Mechanism as a Frequency Analyzer

Michail Raptakis[1,2], *Yannis Pantazis*[2]

[1]Computer Science Department, University of Crete, Greece
[2]Institute of Applied and Computational Mathematics, FORTH, Greece
mrap@csd.uoc.gr, pantazis@iacm.forth.gr

Abstract

Frequency estimation is a fundamental step in many speech applications and predominantly relies on variations of Fourier analysis. In this paper, we demonstrate that frequency estimation can be performed by repurposing the attention mechanism in a non-parametric setup. The proposed approach enables the iterative identification and selection of the most dominant periodic components through direct projections of the input signal to sine and cosine functions (Fourier modes). We explore three different selection criteria and investigate their behavior across a wide range of hyperparameters. We compare our method against the Quadratically Interpolated FFT (QIFFT) in the problem of sinusoidal parameter estimation under two scenarios: multi-sinusoid signals with additive white noise and recorded speech. In both cases, our findings show that the proposed Fourier attention method outperforms QIFFT in frequency estimation, with an average of 29.95% lower frequency RMSE in noisy conditions.

Index Terms: attention mechanism, transformer, speech analysis, frequency estimation, sinusoidal model, FATT, QIFFT

1. Introduction

Frequency analysis plays a pivotal role in a multitude of applications across signal processing, communications, electrical engineering, and data science, among other disciplines. Traditionally, the Fourier transform, particularly the FFT, has served as the cornerstone for extracting frequency domain information, whether via spectrogram analyses or direct examination of frequency and amplitude responses [1]. In the realm of speech processing, such analysis is vital for tasks like speech recognition, speaker identification, and emotion detection, to name a few, where understanding the spectral content of speech signals can significantly enhance accuracy and performance [2].

Building on the importance of frequency analysis in speech and signal processing, sinusoidal parameter estimation is a well-known technique for analyzing and modeling time-series data. At their core, sinusoidal models decompose signals into their sine and cosine components, offering a powerful and versatile framework for manipulating these data streams. Despite years of development and numerous successful sinusoidal models [3], this area continues to attract notable research interest [4, 5, 6]. The continuous evolution of sinusoidal modeling reflects the need for higher precision in demanding applications ranging from telecommunication systems like radar and sonar [7, 8], to assistive technologies including hearing aids [9, 10].

In response to these challenges, this work introduces a novel method for frequency analysis, leveraging the powerful representation capabilities of transformer-based deep learning models. These models utilize the attention layer [11] as a critical component to model long and complex dependencies in sequential data [12, 13, 14]. A key feature of the attention mechanism, which we will fully exploit, is its ability to transform the input domain into an arbitrary output domain. Indeed, through learned weights, the input is mapped to *queries*, which dynamically highlight the input's most pertinent features, as identified by *keys*. These emphasized components are then integrated with *values*, resulting in an output that could belong to a completely different representation space.

In this work, we suggest a specific setup where the attention mechanism is explicitly used as a frequency analyzer without any trainable parameters. In our setting, the single query is the waveform itself, the keys correspond to the Fourier modes (i.e., sine and cosine functions) while the values represent the frequencies of the corresponding keys. Variations of softmax will pick the most prominent periodic component resulting in three versions of attention-based frequency estimation.

We assess the performance of the three proposed variations of our approach across a broad range of hyperparameters using single-component sinusoidal signals. During this analysis, we show that the maximum estimator outperforms the other two (average and median). Subsequently, we combine our best estimator with Least Squares, apply it to the problem of sinusoidal parameter estimation and compare it with the conventional QIFFT [15], revealing that our method surpasses QIFFT both in terms of frequency estimation accuracy and overall signal reconstruction. We validate these results through experiments where both synthetic multi-sinusoidal signals under white noise environments and recorded speech showcase the comparative advantages of the proposed estimation approach. Our implementation in Python is publicly available through the following repository: https://github.com/mikerapt/FATT

2. Fourier Attention Mechanism

In this section, we describe the standard 1D self-attention layer in more detail, and then our proposed modified version that is specifically designed for the extraction of frequencies.

2.1. Self-Attention

Within the transformer model, an attention layer has three primary matrices: the query matrix $Q \in \mathbb{R}^{m\times k}$, the key matrix $K \in \mathbb{R}^{n\times k}$, and the value matrix $V \in \mathbb{R}^{n\times v}$. The query matrix is multiplied by the transpose of the key matrix. Then, this product is normalized with respect to the square root of the input length, and each row is passed through a softmax function to produce a probability vector. The resulting matrix from this operation, denoted $S \in \mathbb{R}^{m\times n}$ and referred to as the attention matrix, is then multiplied by the value matrix V to generate the output O of the attention layer:

10.21437/IberSPEECH.2024-24

$$O \triangleq S \cdot V \tag{1}$$

$$\triangleq \text{softmax}\left(\frac{Q \cdot K^{\intercal}}{\sqrt{n}}\right) \cdot V. \tag{2}$$

In the case of 1D self-attention, an input sequence $x(t) \in \mathbb{R}^{1\times n}$ is projected into Q, K and V using separate trainable weights $W_Q \in \mathbb{R}^{n\times k}, W_K \in \mathbb{R}^{n\times k}, W_V \in \mathbb{R}^{n\times v}$, that is: $Q = x\cdot W_Q$, $K = x \cdot W_K$ and $V = x \cdot W_V$.

2.2. Fourier Attention (FATT)

Having eq. (1) as our basis, we set the query weight matrix W_Q to be equal to the identity matrix I, thus making the query matrix Q to be the input sequence $x(t) \in \mathbb{R}^{1\times n}$. Next, the FATT key matrix is the concatenation of two sub-matrices:

$$K = \begin{bmatrix}\hat{K}\\ \tilde{K}\end{bmatrix} \in \mathbb{R}^{2\ell\times n}, \text{ where} \tag{3}$$

$$\hat{K}_{j,i} = \sin\left(\omega_j t_i\right), \text{ and} \tag{4}$$

$$\tilde{K}_{j,i} = \cos\left(\omega_j t_i\right), \tag{5}$$

with $\omega_j = 2\pi f_j/f_s,\ j \in \{1,2,\ldots,\ell\},\ i \in \{1,2,\ldots,n\}$. The reason for having both sine and cosine functions is to also incorporate the phase information. FATT's value matrix is defined as the following frequency vector:

$$V = [f_1, f_2, \ldots, f_\ell]^{\intercal} \in \mathbb{R}^{\ell\times 1}, \tag{6}$$

where the frequencies are equally spaced with a step $\Delta f = f_{j+1} - f_j > 0$, which defines the frequency resolution. Our corresponding $Q \cdot K^{\intercal}/\sqrt{n}$ part is the concatenation of $\hat{T}, \tilde{T}$:

$$\hat{T} = Q \cdot \hat{K}^{\intercal} \cdot n^a, \tag{7}$$

$$\tilde{T} = Q \cdot \tilde{K}^{\intercal} \cdot n^a, \tag{8}$$

where $a \in \mathbb{R}$ is the normalization parameter. These operations are to essentially model types of Fourier sine and cosine projections. The attention matrix in this case will be[1]:

$$S = \text{softmax}\left(T\right), \text{with} \tag{9}$$

$$T_{i,j} = \sqrt{\hat{T}_{i,j}^2 + \tilde{T}_{i,j}^2}. \tag{10}$$

Finally, we investigate three different frequency estimators for FATT that come from the average, maximum, and median of S:

$$\hat{f}_{avg} \triangleq O_{\text{avg}} = S \cdot V, \tag{11}$$

$$\hat{f}_{\max} \triangleq O_{\max} = f_{j_{\max}}, \tag{12}$$

$$\hat{f}_{\text{med}} \triangleq O_{\text{med}} = f_{j_{\text{med}}}, \text{ where} \tag{13}$$

$$j_{\max} = \underset{j\in\{1,\ldots,\ell\}}{\text{argmax}}\, S_j, \text{ and} \tag{14}$$

$$j_{\text{med}} = \underset{j\in\{1,\ldots,\ell\}}{\text{argmed}}\, S_j. \tag{15}$$

This formulation provides a single scalar value to represent the estimated frequency at all three output scenarios.

[1]The maximum is often subtracted for numerical stability without affecting the softmax values, i.e., $S = \text{softmax}(T - \max\left(T\right))$.

2.3. FATT with Least Squares (FATT-LS)

To extend the proposed frequency estimator for multicomponent signals, we apply a least squares solution to estimate the coefficient of the sinusoidal components as in [16, 17]. Given a signal $x(t) \in \mathbb{R}^n$ and frequencies f_k for $k = 1, ..., K$, we compute the amplitudes α_k, β_k, such that the sum of squares of the following error function (residual) is minimized:

$$e(t) = x(t) - \sum_{k=1}^{K}\left(\alpha_k \sin\left(\omega_k t\right) + \beta_k \cos\left(\omega_k t\right)\right), \tag{16}$$

where $\omega_k = 2\pi f_k/f_s$. The estimated amplitude pairs $\hat{\alpha}_k, \hat{\beta}_k$ are then analytically computed as the minimizers of $\sum_t e^2(t)$.

2.4. Sinusoidal Parameter Estimation with FATT-LS

The proposed sinusoidal parameter estimation algorithm designates a fixed number of iterations, K, corresponding to the number of sinusoidal components that are assumed to exist within an input signal $x(t)$. Initially, the residual signal $e(t)$ is initialized with the input $x(t)$. At each iteration, FATT estimates a new frequency value from $e(t)$, utilizing either the average, maximum, or median estimator as defined in Section 2.2. Subsequently, a least squares approach is applied in order to estimate all amplitude pairs, $\hat{\alpha}_k, \hat{\beta}_k$, up to the current iteration. Utilizing these amplitudes and the frequencies identified thus far, we assemble the aggregate sinusoidal summation of eq. (16). This iterative process continues, at each step using the new residual signal $e(t)$, that is, by subtracting the latest sinusoidal sum from the original input signal $x(t)$. The algorithmic procedure described is outlined below in pseudocode format:

Algorithm 1 Sinusoidal Parameter Estimation with FATT-LS

Require: Input signal $x(t) \in \mathbb{R}^n$, sampling rate f_s, number of sinusoidal components K, frequency resolution Δf, normalization parameter a, estimator type (e.g., "max")

Ensure: Estimated frequencies $\hat{f}_k$ and amplitude pairs $\hat{\alpha}_k, \hat{\beta}_k$ for all $k = 1, \ldots, K$ sinusoidal components of $x(t)$.

Initialize $\hat{f} \leftarrow \emptyset,\ e(t) \leftarrow x(t)$
for $k = 1$ to K **do**
  $\hat{f} \leftarrow \hat{f} \cup \text{FATT}\left(e(t), f_s, \Delta f, a, \text{type}\right)$
  $\hat{\alpha}, \hat{\beta} \leftarrow \text{LeastSquares}\left(x(t), \hat{f}\right)$
  $\hat{\omega} \leftarrow 2\pi\hat{f}/f_s$
  $s(t) \leftarrow \sum_{i=1}^{k}\left(\hat{\alpha}_i \sin\left(\hat{\omega}_i t\right) + \hat{\beta}_i \cos\left(\hat{\omega}_i t\right)\right)$
  $e(t) \leftarrow x(t) - s(t)$
end for

3. Results

Initially, we examine the impact of various hyperparameter values for FATT's three frequency estimators, as detailed in Section 3.1. Subsequently, leveraging the least squares method described in Section 2.3, we compare FATT-LS with QIFFT in the context of sinusoidal analysis-synthesis, utilizing both synthetic and real-world data. For the synthetic scenario, we add Gaussian noise to multi-component signals, as described in Section 3.2. For the real-world application, we analyze and resynthesize recorded speech signals, as detailed in Section 3.3.

3.1. Single Sinusoid

To enhance our understanding of the operational behavior behind the proposed FATT mechanism, we investigate its three principal hyperparameters: (i) frequency step Δf, (ii) input length in samples n, and (iii) normalization parameter a. This examination spans across a wide range of values and is done for all three FATT estimators. In each of the three graphs presented in Figure 1, the Mean Absolute Error (MAE) is derived from an average over 10^4 frequency predictions for each point. This error quantifies the difference between the frequencies estimated by FATT and those of a sinusoid with constant amplitude $A = 1$, phase $\phi = 0$, and frequency selected uniformly at random $f \sim \mathcal{U}(\epsilon, f_s/2 - \epsilon)$, where $\epsilon = 1$ Hz and $f_s = 1$ kHz.

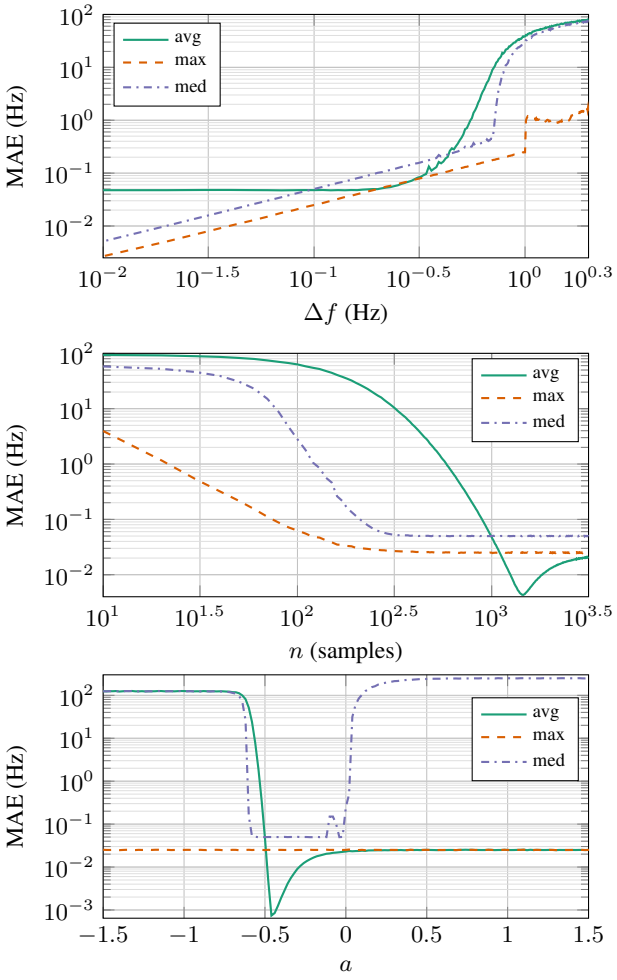

Figure 1: *Single sinusoid case: MAE (log scale) for FATT's three frequency estimators over a wide range of possible hyperparameter values.*

The upper plot in Figure 1 quantifies the MAE across a spectrum of frequency steps for 1-second long input signals, while a is fixed at -0.5, mirroring the default setting in self-attention. As anticipated, employing a finer-grained frequency matrix enhances the estimation accuracy, most notably for the median and maximum frequency estimators. The error for the average estimator, on the other hand, is not reduced for frequency steps below a threshold.

The middle plot in Figure 1 illustrates the MAE in relation to the signal length as measured in samples n, maintaining the same sampling rate f_s and normalization parameter a as the previous plot, while fixing $\Delta f = 0.1$. Notably, the maximum estimator requires the fewest samples to accurately estimate the actual frequency, unlike the other estimators which converge at larger input lengths. For lengths longer than approximately 3 seconds, the average estimator achieves the smallest MAE.

The lower plot in Figure 1, arguably the most revealing, illustrates the MAE across different normalization constants a, for inputs of 1 second in duration, with the same sampling rate f_s and frequency resolution $\Delta f = 0.1$. The maximum estimator remains entirely invariant with respect to the values of a, in stark contrast to the average estimator, which exhibits a sharp global minimum near $a = -0.5$. This observation further motivates the adoption of $a = -0.5$ as the default value in standard self-attention. For values larger than $a = 0$, the average estimator coincides with the maximum estimator, an outcome anticipated by the inherent properties of the softmax function. Notably, the median estimator demonstrates the least favorable performance across the entire range of a values.

3.2. Multi-Sinusoid with Noise

We further assess the efficacy of the proposed frequency analyzer on multi-component data within an additive white noise setting, comparing its frequency estimation accuracy against that of QIFFT. Owing to its demonstrated efficiency in handling any signal length and its robustness to the normalization exponent, as evidenced in the preceding subsection, we compare QIFFT with the maximum estimator for FATT, setting $\Delta f = 0.1$ and $a = -0.5$ for all subsequent applications. To facilitate complete signal reconstruction, we integrate FATT with a least squares approach, detailed in Section 2.3, enabling amplitude and phase estimation. All constructed synthetic signals are 1-second long in duration and are formulated as follows:

$$x(t) = \sum_{k=1}^{K} A_k \sin\left(\omega_k t + \phi_k\right) + w(t), \tag{17}$$

where $w(t)$ signifies the white noise term, $K = 30$, and the rest of the parameters are chosen uniformly at random, i.e., $A_k \sim \mathcal{U}(1, 1.1)$, $\omega_k = 2\pi f_k / f_s$, $f_k \sim \mathcal{U}(\epsilon, f_s/2 - \epsilon)$ with $\epsilon = 1$ Hz, $f_s = 1$ kHz, and $\phi_k \sim \mathcal{U}(-\pi, \pi)$. The Root Mean Square Error (RMSE) is averaged over 10^4 estimations for each Signal-to-Noise Ratio (SNR) level. RMSE is used to evaluate both frequency estimation and signal reconstruction accuracy. To ensure a fair comparison, both methods were configured to operate with identical frequency resolution settings. Specifically, the QIFFT bins were set to be twice the size of FATT-LS's frequency matrix. Each estimated frequency is paired with the closest ground-truth frequency using the Hungarian algorithm [18]. The results are illustrated in Figure 2.

The upper plot in Figure 2 clearly demonstrates that FATT-LS surpasses QIFFT in accurately estimating the true frequencies across the entire range of SNR values considered in this study. Specifically, FATT-LS achieves an average of approximately 29.95% lower frequency RMSE over the entire SNR range. Regarding signal reconstruction error, both methods demonstrate comparable performance, with QIFFT slightly underperforming compared to our proposed approach, as seen in the lower plot of Figure 2. Averaged across the entire SNR range, FATT-LS was measured to have close to 10.85% lower signal reconstruction RMSE than QIFFT.

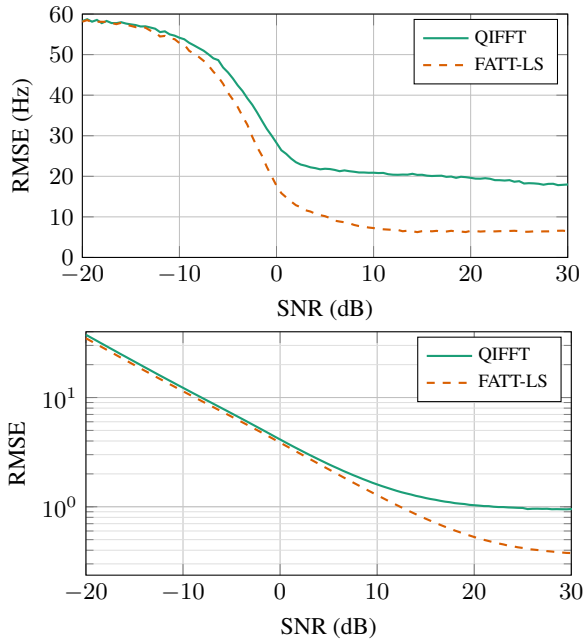

Figure 2: *Synthetic multi-sinusoid case with additive Gaussian noise: The upper plot shows frequency error (linear scale), and the lower shows signal reconstruction error (log scale).*

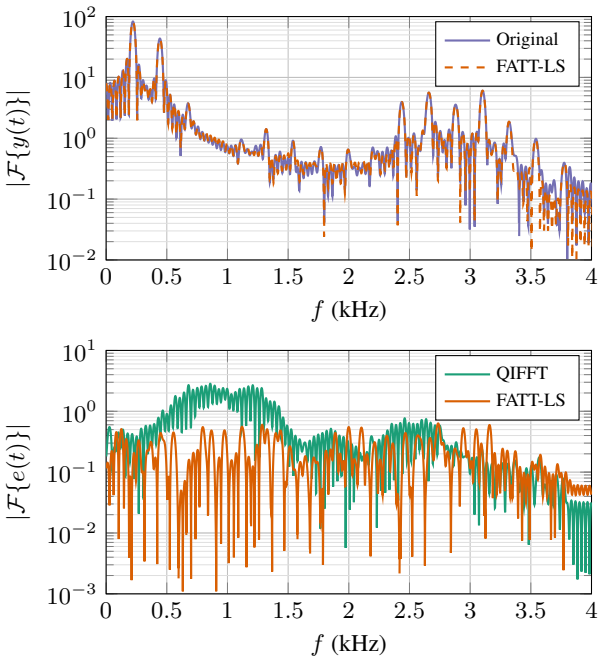

Figure 3: *Speech vowel case: Amplitude responses (log scale) of the reconstruction (top) and of the residual signals (bottom).*

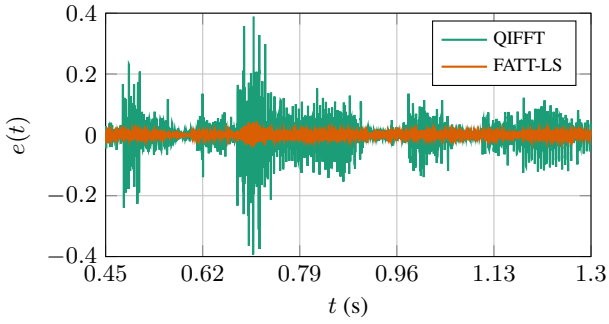

Figure 4: *Speech sentence case: The residual errors in time domain for the word "comparatively", synthesized via OLA.*

3.3. Speech Signals

Our final experiment evaluates the reconstruction abilities of FATT-LS using the maximum estimator and compares it to QIFFT on actual speech signals. Both methods are given the same frequency resolution and use the same hyperparameters as described in Section 3.2, with $K = 40$ sinusoids. The upper plot in Figure 3 displays the amplitude responses after reconstructing a 30 ms frame of a sustained /ɛ/ vowel from a female speaker. The lower plot shows the amplitude response of the residual signals for both methods. Figure 4 displays a segment of the residual signals following the reconstruction of the second sample in the LJ-Speech dataset [19] using the Overlap and Add (OLA) method with a 20 ms window and a 10 ms overlap for both FATT-LS and QIFFT.

The upper plot of Figure 3 shows that FATT-LS captures the most significant frequency peaks in terms of amplitude, while the lower plot demonstrates that FATT-LS's residual error is significantly smaller than that of QIFFT over the same signal frame for most of the spectrum. These discrepancies are further highlighted by the time-domain residual signals presented in Figure 4. The inaccuracies in QIFFT accumulate, resulting in a much larger residual signal that starkly contrasts with the precision offered by FATT-LS. Importantly, these disparities cannot be attributed to the quasi-stationary characteristics of the input signals. If this were a contributing factor, FATT-LS would have been similarly affected, since both methods assume stationarity.

4. Conclusions and Future Work

In this study, we introduced a novel application of the attention mechanism, termed Fourier Attention, or FATT, for frequency estimation and tested it with both synthetic and real-world signals for sinusoidal analysis and synthesis. Our method, when paired with a simple least squares approach, demonstrates superior estimation accuracy compared to the conventional QIFFT in both multi-component scenarios across a wide range of SNRs and in actual speech recordings. The adaptability and robustness of FATT against noise further highlight its potential as a versatile tool for various signal processing tasks. This initial exploration yields promising results and lays the foundation for future research and development of the FATT mechanism.

Indeed, future directions include expanding the FATT framework to accommodate less densely spaced frequencies, thereby broadening its application scope. This could be achieved by making the frequency matrix a trainable parameter, which would enhance the model's adaptability to specific signal characteristics. Furthermore, integrating this mechanism into deep learning architectures opens exciting possibilities for more complex tasks such as Text-to-Speech synthesis, speech recognition, music analysis, and bioacoustic signal processing. Additionally, further research aimed at optimizing FATT's computational efficiency and memory usage is crucial to ensure its viability for large-scale and real-time applications.

5. Acknowledgements

The work of Michail Raptakis and Yannis Pantazis was supported by the Hellenic Foundation for Research and Innovation (HFRI) through the "Second Call for HFRI Research Projects to support Faculty Members and Researchers" under Project 4753.

6. References

[1] A. V. Oppenheim, "Discrete-Time Signal Processing," *Prentice Hall*, vol. 3, pp. 804–809, 1999.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education India, 2002.

[3] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Digital Audio Effects (DAFx) Conference*, 2002, pp. 51–58.

[4] B. Hayes, C. Saitis, and G. Fazekas, "Sinusoidal frequency estimation by gradient descent," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] D. Kundu, "Robust parameter estimation in two-channel sinusoidal frequency model," *Franklin Open*, vol. 6, p. 100059, 2024.

[6] S. L. Kiser, M. Rébillat, M. Guskov, and N. Ranc, "Real-time sinusoidal parameter estimation for damage growth monitoring during ultrasonic very high cycle fatigue tests," *Mechanical Systems and Signal Processing*, vol. 182, p. 109544, 2023.

[7] S. M. Kay, *Modern spectral estimation*. Prentice Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1988.

[8] S. L. Marple Jr, *Digital spectral analysis*. Courier Dover Publications, 2019.

[9] S. Tokgoz and I. M. Panahi, "Robust three-microphone speech source localization using randomized singular value decomposition," *IEEE Access*, vol. 9, pp. 800–811, 2021.

[10] A. Ganguly and I. Panahi, "Non-uniform microphone arrays for robust speech source localization for smartphone-assisted hearing aid devices," *Journal of signal processing systems*, vol. 90, pp. 1415–1435, 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," vol. 30, 2017.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[14] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[15] M. Abe and J. O. Smith III, "Design criteria for the quadratically interpolated fft method (i): Bias due to interpolation," Stanford University, Tech. Rep. STAN-M-114, 2004. [Online]. Available: https://ccrma.stanford.edu/STANM/stanms/stanm114/stanm114.pdf

[16] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, 1996.

[17] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative estimation of sinusoidal signal parameters," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 461–464, 2010.

[18] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017, accessed: September 23, 2024.