IFF-AR: Conditional Autoregressive Audio Generation via Instantaneous Frequency Modelling

October 7, 2025

Valerio Santini

Abstract

This report presents **IFF-AR**, an autoregressive audio generation model in the frequency domain, based on the joint prediction of log-magnitude and instantaneous frequency (IF). The architecture combines a random convolutional encoder (TCN) with a magnitude decoding head *MagHead* and a conditioned *normalising flow* for phase modelling. The loss function acts on the three domains—spectral, phase, and temporal—ensuring perceptual and structural consistency. This formulation allows for consistent time-frequency generation without iterative phase reconstruction.

1. Introduction

In recent years, deep generative models have made remarkable progress in sound synthesis, although direct phase modelling remains a challenge. Most existing approaches reconstruct phase iteratively, for example using Griffin-Lim (Griffin & Lim, 1984). This project proposes **IFF-AR**, an autoregressive model that generates audio directly in the frequency domain by jointly predicting *log-magnitude* and *instant frequency* (Takamichi et al., 2018). The architecture integrates a temporal convolutional encoder, a lightweight decoder for magnitude, and a conditional normalising flow (Dinh et al., 2016) for phase modelling. The phase is then treated as a learnable probabilistic variable conditioned on spectral energy, avoiding the rigid separation between magnitude and phase. Experimental results show that the proposed formulation produces stable and consistent spectrograms, highlighting the potential of conditioned flows for explicit phase modelling.

Email: Valerio Santini <santini.2176222@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

2. Related Works

Early autoregressive models operated in the time domain, such as **WaveNet** (Van Den Oord et al., 2016), which synthesizes audio sample-by-sample. In the spectral domain, models like **MelNet** (Vasquez & Lewis, 2019) showed that predicting spectrogram sequences autoregressively reduces temporal length and complexity. **GANSynth** (Engel et al., 2019) introduced the use of instantaneous frequency (IF) as a phase representation, demonstrating audio generation with GANs. Diffusion models redefine the state of the art: **DiffWave** (Kong et al., 2020) provides a non-autoregressive pipeline for waveform synthesis; **AudioLDM** (Liu et al., 2023) applies latent diffusion on CLAP embeddings; **FlashAudio** (Liu et al., 2024) uses rectified flows for efficient one-step generation. Autoregressive Transformers, e.g. **Next-Scale** (Qiu et al., 2024) and **SongGen** (Liu et al., 2025), focus on sequence compression and controllable generation. The phase has traditionally been handled with iterative algorithms like Griffin-Lim, which require many iterations and often produce artefacts. Neural approaches have sought to improve this: (Takamichi et al., 2018) model phase with a von Mises DNN, while (Masuyama et al., 2023) show that predicting phase differences between adjacent frames is more stable and suitable for causal reconstruction. However, few approaches integrate this representation into a conditional autoregressive framework, simultaneously predicting magnitude and IF. This is the main novelty of this project.

3. Method

IFF-AR model architecture consists of four main modules, which are necessary to perceive the general context, predict magnitude and phase. During preprocessing, the audio is converted into a time-frequency representation using Short-Time Fourier Transform (STFT). This transformation produces complex coefficients, from which it is possible to extract two sets of features: *Log-magnitude*, log-scaled spectral power, standardised over the training set; *Instantaneous frequency (IF)* calculated as the phase difference between consecutive frames computed per-frequency as $\mathrm{IF}_t(f) = \Delta\phi_t(f)/\Delta t$, describing phase dynamics di-

rectly without absolute-angle ambiguity. IF is normalised using the median and median absolute deviation (MAD). Each example is split into context–target windows for autoregressive prediction. **TCNContextEncoder** is implemented as a 1D random convolutional network with dilations, inspired by WaveNet (Van Den Oord et al., 2016), but instead of working at the audio sample level it operates directly at the frame level. This choice allows long-range dependencies in the time-frequency domain to be modelled effectively, avoiding Transformers and keeping computation low through convolutional parallelism. The encoder receives the log-magnitude and instantaneous frequency sequence of the context as input and produces a compact embedding that will subsequently be used both by the MagHead module and in the normalising flow conditioned for phase modelling. **MagHead** is a decoding head that uses embeddings provided by the TCNContexEncoder to predict the log-magnitudes of STFT frames, normalised as in the preprocessing stage. It consists of a light sequence of fully-connected layers and 1D convolutions with small kernels that project the embedding onto the frequency dimension; the architecture is deliberately light to reduce overfitting and to focus the model's capacity on the flow for the phase. Furthermore, MagHead not only produces magnitudes, but also provides a condition for flow, in this way, the phase (IF) is modelled consistently with spectral energy: this condition is commonly avoided in models that treat magnitude and phase separately (Nugraha et al., 2019; Dai et al., 2025). **IFConditionalFLow** uses a conditional normalising flow to learn the distribution of future IFs explicitly and consistently with the predicted magnitudes, similar to the RealNVP design (Dinh et al., 2016). Each layer divides the input vector into two parts, x_a and x_b, and calculates, through small conditional networks, the shift and scale by applying an invertible transformation shown in the appendix in equation 1. Each coupling layer receives the context embedding (from the TCN) and the predicted magnitude (from the MagHead) as conditioning; the transformations depend not only on the IF input but also on the context. This choice represents an innovative element with respect to the literature: the phase is not treated as a residual or reconstructed iteratively (Griffin & Lim, 1984), nor estimated separately from conditional networks only on magnitudes (Takamichi et al., 2018; Masuyama et al., 2023), but explicitly modelled as a conditional probability distribution. Finally, to complete the architecture, it is necessary to reconstruct the absolute phase of the STFT coefficients to recover the time-domain signal via: $\phi_t(f) = \phi_{t-1}(f) + \Delta\phi_t(f)$ where $\Delta\phi_t(f)$ is the predicted phase. Next, the complex coefficients of the spectrogram are reconstructed as: $X_t = \hat{M}_t(f) \cdot e^{i\phi_t(f)}$ where $\hat{M}_t(f)$ are the magnitudes predicted by MagHead. Finally, the Inverse Short-Time Fourier Transform (iSTFT) is applied to obtain the reconstructed audio signal: $\hat{x} = iSTFT\{X_t(f)\}$.

The loss function was designed to deal with three different domains: spectral, phase and time. Many architectures, such as WaveNet (Van Den Oord et al., 2016) and MelNet (Vasquez & Lewis, 2019), do not include perceptual or structural terms such as Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) or Multi-Resolution Short-Time Fourier Transform Loss (MRSTFT), which enhance the perceptual quality and consistency of the generated spectrum. The total loss is a weighted combination of these terms and is optimized using the Adam optimizer with a learning rate of 10^{-3} and early stopping on the validation loss.

4. Experimental Results

The NSynth dataset (Engel et al., 2017), segmented into context windows L and target windows K, was used for training. STFT was applied with FFT = 1024, hop = 256, and Hann window. Training was performed using teacher forcing, with autoregressive predictions packaged on the actual target sequences. The model was trained for 5 epochs using Adam optimizer with learning rate 10^{-4}, batch size 16 and teacher forcing for autoregressive rollouts. The evaluation was performed on 2500 validation sample. Log-magnitude and IF metrics, in Table 1, show good stability, meaning that the network has learned a consistent time-frequency domain structure. The negative SI-SDR indicates that the audio reconstruction is still far from the reference, but the phaseGT and magGT values show that the spectral patterns and phase dynamics are captured in a meaningful way. Furthermore, the average IF error is consistent with the normalised scale (in the range [-1, 1]), indicating a good approximation of the phase derivatives. MagPred (right) in image 1 reproduces the general energy pattern of the targets, although it tends to smooth spectral transitions and lose weak harmonics. Energy trends confirm temporal consistency despite absence of absolute phase supervision. The smoothing effect stems from the limited ability of the flow to model high-variance IF, a behaviour resulting from the short context taken into consideration. The model manages to correctly predict the energy trend and IF dynamics. The results show convergence towards consistent representations even under limited training conditions. The conditioned architecture IF-Flow + MagHead is stable, and the reconstruction is consistent despite low SI-SDR.

5. Conclusions

IFF-AR is an autoregressive model in the time-frequency domain that combines log-magnitude and instantaneous frequency prediction in a conditional framework. The introduction of a normalising flow for explicit phase modelling, conditioned on both context and spectral energy,

represents an innovative element compared to existing models. Despite computational limitations and a reduced number of training epochs, the results show that the model is capable of learning a consistent and stable representation of the spectral structure and phase dynamics. The multi-domain loss function contributes to convergence towards consistent spectrograms, suggesting that direct modelling of instantaneous frequency may be a promising avenue for improving perceptual quality in spectral autoregressive models.

6. Appendix

Evaluation metrics. This table reports the quantitative evaluation computed on the NSynth validation set. The combination of spectral (LSD, SpecConv), phase (IF MAE/RMSE), and perceptual (SI-SDR) metrics allows assessing both reconstruction accuracy and perceptual consistency. Although the SI-SDR is low due to limited training, the stability of spectral metrics confirms coherent magnitude–phase modelling.

Metric	Value	Metric	Value
LSD (dB)	3.36	IF MAE	1.39
SpecConv	6.28	IF RMSE	1.73
Complex MSE	180.37	SI-SDR (dB)	-15.21
LogMag MAE	44.77	SI-SDR (phase GT)	-9.67
LogMag RMSE	49.33	SI-SDR (mag GT)	-4.81

Table 1. Quantitative evaluation on the NSynth validation set.

Predicted vs target spectrogram. The predicted log-magnitude spectrogram reproduces the main energy structure of the target while smoothing high-frequency components. This behaviour suggests that the conditional flow captures overall temporal–spectral dynamics even with reduced variance in instantaneous frequency prediction.

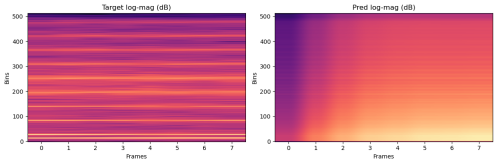

Figure 1. Predicted (left) and target (right) log-magnitude spectrogram.

Coupling transformation in IFConditionalFlow. Each coupling layer transforms part of the input (x_b) using scale and translation parameters (s, t) conditioned on both the complementary partition (x_a) and the context c. This formulation follows the RealNVP paradigm and ensures invertibility while allowing context-dependent modelling of phase dynamics.

$$\begin{cases} y_a = x_a \\ y_b = x_b \cdot \exp(s(x_a, c)) + t(x_a, c) \end{cases} \tag{1}$$

References

Dai, L., Li, A., Lei, T., Yu, M., Li, X., and Zheng, C. Rethinking the joint estimation of magnitude and phase for time-frequency domain neural vocoders. *arXiv preprint arXiv:2509.18806*, 2025.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., and Norouzi, M. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

Griffin, D. and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Liu, H., Wang, J., Huang, R., Liu, Y., Lu, H., Zhao, Z., and Xue, W. Flashaudio: Rectified flows for fast and high-fidelity text-to-audio generation. *arXiv preprint arXiv:2410.12266*, 2024.

Liu, Z., Ding, S., Zhang, Z., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., and Wang, J. Songgen: A single stage auto-regressive transformer for text-to-song generation. *arXiv preprint arXiv:2502.13128*, 2025.

Masuyama, Y., Yatabe, K., Nagatomo, K., and Oikawa, Y. Online phase reconstruction via dnn-based phase differences estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:163–176, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2022.3221041. URL http://dx.doi.org/10.1109/TASLP.2022.3221041.

Nugraha, A. A., Sekiguchi, K., and Yoshii, K. A deep generative model of speech complex spectrograms. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 905–909. IEEE, 2019.

Qiu, K., Li, X., Chen, H., Sun, J., Wang, J., Lin, Z., Savvides, M., and Raj, B. Efficient autoregressive audio modeling via next-scale prediction. *arXiv preprint arXiv:2408.09027*, 2024.

Takamichi, S., Saito, Y., Takamune, N., Kitamura, D., and Saruwatari, H. Phase reconstruction from amplitude spectrograms based on von-mises-distribution deep neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 286–290. IEEE, 2018.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12:1, 2016.

Vasquez, S. and Lewis, M. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*, 2019.