MIDI-VALLE: Improving Expressive Piano Performance Synthesis Through Neural Codec Language Modelling

Jingjing Tang[1], Xin Wang[2], Zhe Zhang[2], Geraint Wiggins[13], Junichi Yamagishi[2], György Fazekas[1]
[1]Centre for Digital Music, Queen Mary University of London, UK [2]National Institute of Informatics, Japan
[3]Vrije Universiteit Brussel, Belgium

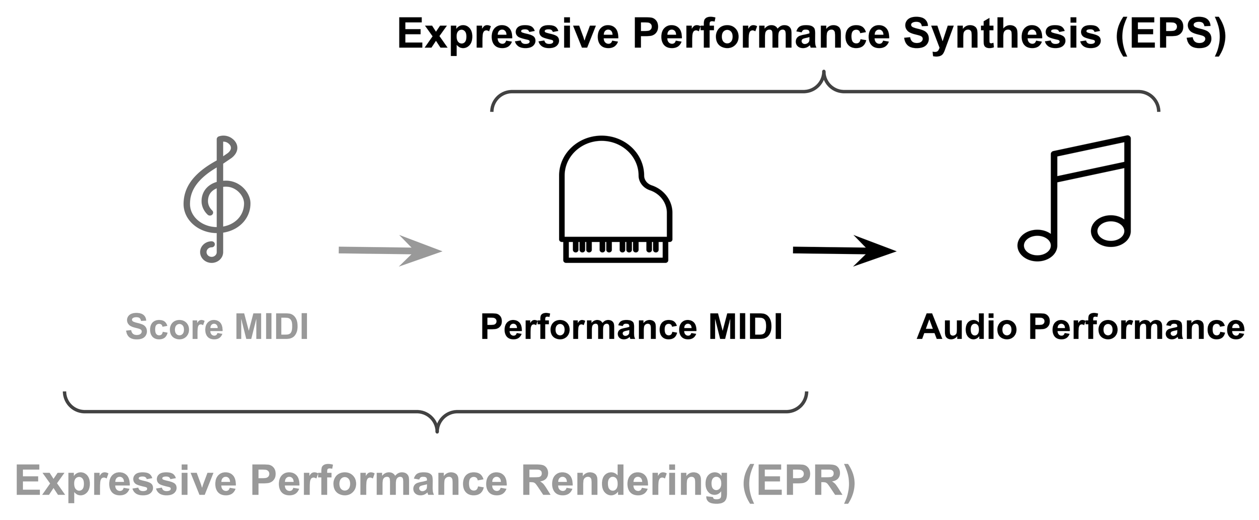
AI + MUSIC

Expressive Performance Synthesis

Objective: Synthesise **expressive piano performance audio** from **performance MIDI.**

Challenges

- **Generalisation**: Difficulty in handling unseen timbres, styles, and acoustic environments.
- **Control**: No fine control over output acoustic characteristics.
- **Integration**: Differences in how EPR (e.g. tokenization; feature encoding) and EPS models (piano-rolls) represent MIDI cause inconsistencies, reducing synthesis quality.



Scan for Demo

Scan for Paper

Why VALL-E[1] for EPS?

- Employs **EnCodec**[2] to compress and tokenise audio, allowing training on larger, more diverse datasets to improve generalisation
- **Token-based discrete modelling** aligns symbolic MIDI and audio representations more consistently
- Supports **zero-shot** adaptation by conditioning on short audio prompts, enabling control over acoustical conditions
- Proven effectiveness in **high-fidelity** and expressive text-to-speech synthesis.

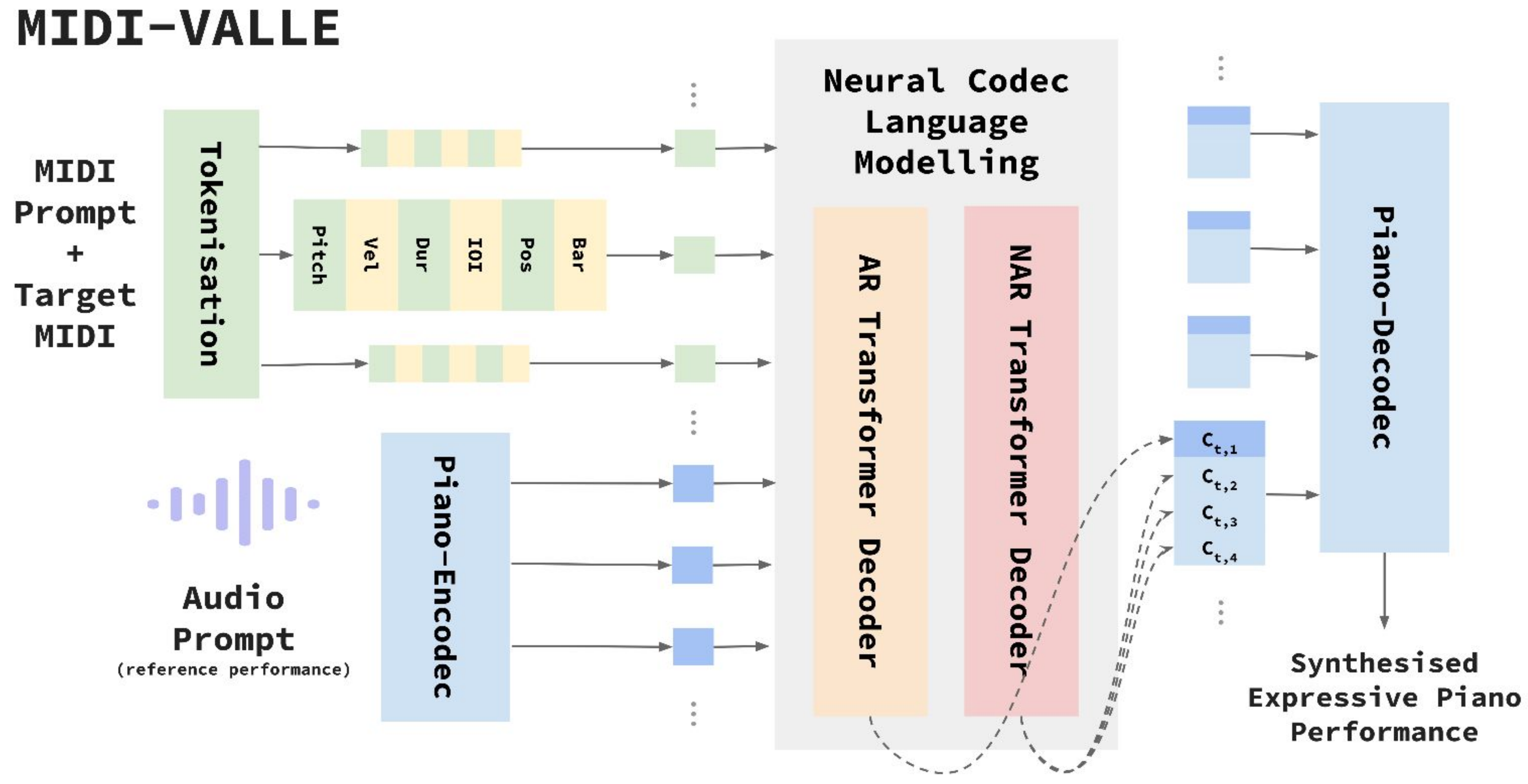
Proposed Method

Audio and MIDI Tokenisation

- 🎵 **Audio** : we **fine-tuned EnCodec** from MusicGen[3] on ATEPP dataset to create **Piano-Encodec**. It uses residual vector quantisation to produce discrete audio tokens as 4 codebooks.
- 🎹 **MIDI** : we uses the Octuple[4] MIDI representation, encoding each note with features and includes **IOI tokens** to explicitly model expressive onset timing.

Model Architecture

MIDI-VALLE



Evaluation and Results

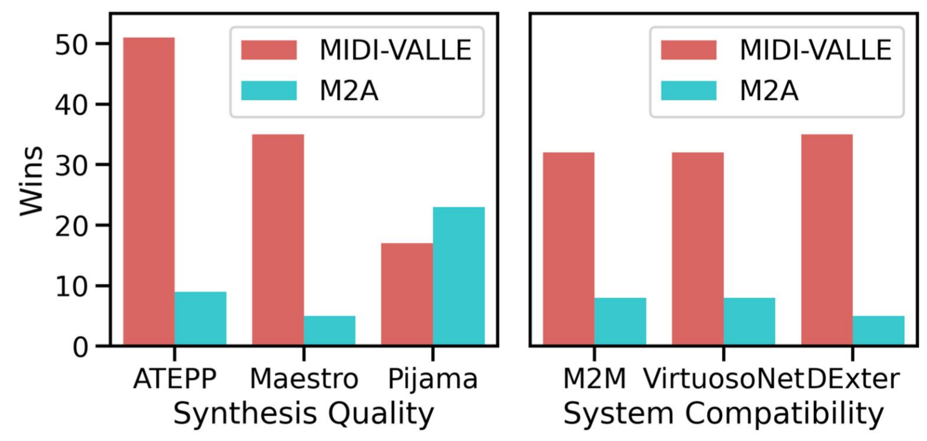
Model	Dataset	FAD ↓	Spec. ↓	Chroma ↓
Encodec [21]	ATEPP	–	0.304 ± .005	0.478 ± .011
	ATEPP	0.685	0.123 ± .002	0.140 ± .002
Piano-Enc.	Maestro	0.984	0.135 ± .002	0.139 ± .001
	Pijama	1.133	0.143 ± .003	0.137 ± .001

- Piano-Encodec fine-tuned on ATEPP dramatically reduces spectrogram distortion and chroma distortion compared with Encodec.
- The model generalises well, achieving similarly strong reconstruction quality on Maestro and Pijama despite being trained only on ATEPP.

- MIDI-VALLE outperforms M2A on ATEPP and Maestro, reducing FAD by over 75% and showing closer alignment to reconstructed audio.
- While both models face challenges on Pijama, MIDI-VALLE still achieves lower FAD, suggesting better timbral preservation despite higher harmonic distortions.
- MIDI-VALLE's lower FAD against reconstructions than against ground truth highlights its closer fit to quantised embeddings than to raw performance audio.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
ATEPP				
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
Maestro				
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
Pijama				
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

GT - Ground Truth
RC - Reconstruction with Piano-Encoder
MV - Generation from MIDI-VALLE



- Listening tests show MIDI-VALLE is preferred over M2A in synthesis quality for ATEPP and Maestro and in system compatibility overall, though M2A is favoured for jazz in Pijama.

Conclusion

- We presented MIDI-VALLE, an EPS model based on neural codec language modelling, that achieves high-quality, expressive synthesis output.
- Future work will explore generalisation across more musical genres and examine the effects of model size and codebook design, and compare MIDI-VALLE with physical modelling and alternative audio codec approaches.

References

[1] Chen, Sanyuan, et al. "Neural codec language models are zero-shot text to speech synthesizers." IEEE Transactions on Audio, Speech and Language Processing (2025).
[2] Défossez, Alexandre, et al. "High Fidelity Neural Audio Compression." Transactions on Machine Learning Research.(2023)
[3] Copet, Jade, et al. "Simple and controllable music generation." Advances in Neural Information Processing Systems 36 (2023): 47704-47720.
[4] Zhu, Hongyuan, et al. "MusicBERT: A self-supervised learning of music representation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

This work was supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music [grant number EP/S022694/1] and the National Institute of Informatics, Japan. J. Tang is a research student jointly funded by the China Scholarship Council [grant number 202008440382] and Queen Mary University of London. G. Wiggins received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen". We thank the reviewers for their valuable feedback, which helped improve the quality of this work.

UK Research and Innovation

ISMIR 2025

Queen Mary University of London