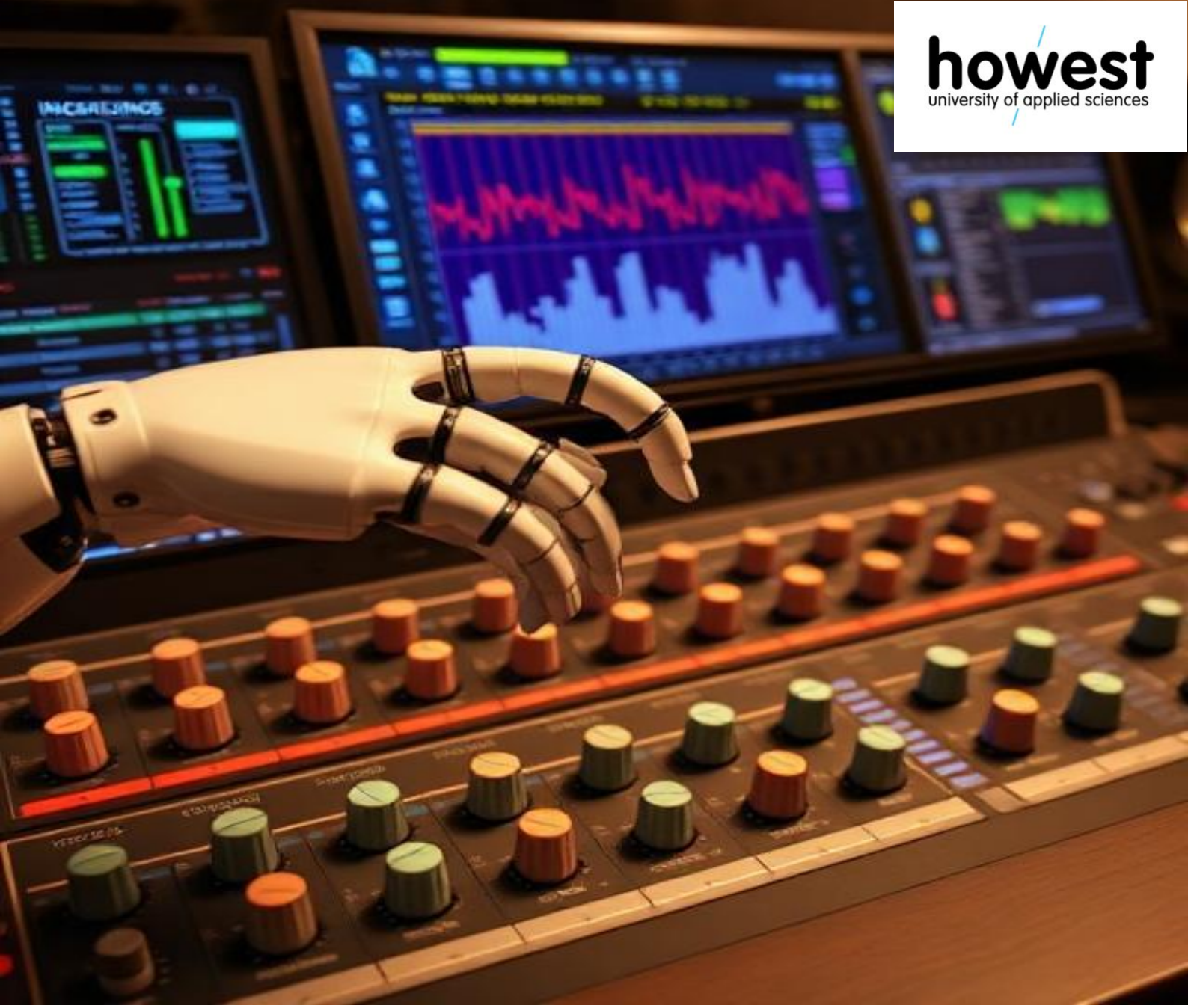
howest
university of applied sciences



HOW CAN A STEM-AWARE NEURAL ARCHITECTURE BE USED TO PREDICT AND APPLY MASTERING PARAMETERS FOR THE AUTONOMOUS PRODUCTION OF AUDIO MASTERS THAT MEET PROFESSIONAL STREAMING STANDARDS AND APPROACH THE PERCEIVED QUALITY OF HUMAN-MADE MASTERS?

PRIMARY SUPERVISOR: Wouter Gevaert
EXTERNAL ADVISOR: Matei Craciunescu

Research question caried out by

MARIUS CEZAR SÎRBU

For obtaining Bachelor's Degree in

CREATIVE TECHNOLOGIES & ARTIFICIAL INTELLIGENCE

Howest | 2025-2026

Preface

This bachelor's thesis is the final project of the Creative Technology and AI program at Hogeschool West-Vlaanderen in Kortrijk. It represents the conclusion of three years of study at the intersection of creative practice, software development, and artificial intelligence, and it brings those disciplines together through both a body of research and a working practical implementation.

The subject grew out of a long-standing personal interest in music production, and specifically out of curiosity about whether the tools of modern machine learning could be brought meaningfully into the mastering stage of audio production. Mastering has always seemed like a discipline where technical precision and perceptual sensitivity are equally important, and the question of whether that combination can be partially automated felt worth investigating carefully and honestly.

The research is guided by the question of how a stem-aware neural architecture can be used to predict and apply mastering parameters for the autonomous production of audio masters that meet professional streaming standards and approach the perceived quality of human-made masters. To explore this, a working prototype was designed, implemented, and evaluated over a focused development period, with the process and its outcomes documented across all chapters of this thesis.

I would like to thank **Wouter Gevaert, my Matei Craciunescu,** the professional audio engineer who took the time to evaluate the prototype and share observations on its technical approach and practical limitations. Those conversations directly shaped the reflection and recommendations that form the second half of this thesis.

Marius Cezar Sirbu Kortrijk, May 2026

Abstract

This thesis investigates how a stem-aware neural architecture can be used to predict and apply mastering parameters for the autonomous production of audio masters that meet professional streaming standards and approach the perceived quality of human-made masters.

The study combines a systematic literature review covering intelligent mastering, music source separation, psychoacoustics, and loudness standardisation with the design, implementation, and iterative evaluation of a working prototype. The system uses pretrained Hybrid Transformer Demucs to separate a stereo mix into drums, bass, other, and vocals, extracts a compact eight-dimensional representation of stem-aware and mix-aware characteristics, and predicts mastering parameters for a differentiable DSP chain applied to the original stereo mixture. The prototype was trained on MUSDB18-HQ using a self-supervised loss combining loudness, masking, distortion, and regularisation objectives.

The most significant positive result is the improvement in loudness targeting following a principled architectural redesign, which reduced mean LUFS error from 24.87 LU to 2.68 LU. True-peak compliance was maintained at 100 percent throughout. Key limitations include the persistent gap between metric compliance and perceptual quality, confirmed through expert consultation and professional community discourse, as well as restricted genre coverage, mid-range masking degradation on average, and the absence of formal perceptual evaluation.

The recommendations prioritise action space design, genre-conditioned loudness targeting, high-frequency feature extension, and the introduction of structured perceptual testing. The thesis concludes that a stem-aware mastering pipeline of this type is technically feasible and measurably improvable through iterative redesign, but does not yet demonstrate consistent perceptual equivalence with professional human mastering practice.

Table of contents

List of figures

List of abbreviations

Abbreviation	Full term
ABX	A/B/X perceptual listening test methodology
AI	Artificial Intelligence
API	Application Programming Interface
biLSTM	Bidirectional Long Short-Term Memory
CLI	Command-Line Interface
CPU	Central Processing Unit
dB	Decibel
DSP	Digital Signal Processing
EBU	European Broadcasting Union
EDM	Electronic Dance Music
EQ	Equalisation
FFT	Fast Fourier Transform
FLAC	Free Lossless Audio Codec
GLU	Gated Linear Unit
LU	Loudness Unit
LUFS	Loudness Units relative to Full Scale
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
MUSDB	Music Source Separation Database
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
RMS	Root Mean Square
WAV	Waveform Audio File Format

Glossary

Audio mastering	The final stage of music production in which a completed stereo mix is prepared for distribution. Mastering involves tonal correction, dynamic control, loudness normalisation, stereo image refinement, and export formatting to ensure the track translates consistently across playback systems and meets platform delivery standards.
Compression (audio)	A dynamic processing technique that reduces the difference between the loudest and quietest parts of an audio signal by attenuating levels above a defined threshold. In mastering, compression is used to control the overall dynamic envelope of a mix and to shape perceived punch and density.
Curriculum training	A training strategy in which the relative weighting of different loss objectives is adjusted over the course of training. In this project, the loudness objective was weighted more heavily at the start of training before gradually rebalancing toward masking and distortion terms, allowing the model to learn the most critical correction first before attempting to satisfy all objectives simultaneously.
Differentiable DSP	An approach in which audio signal processing operations are implemented using differentiable mathematical functions within a deep learning framework. This allows audio-domain objectives to be used directly as training losses, because gradients can be propagated through the processing chain without requiring paired ground-truth outputs.
Dynamic range	The ratio between the loudest and quietest levels in an audio signal, typically expressed in decibels. In mastering, managing dynamic range involves balancing the energy of a mix so that it remains engaging and translatable without sacrificing loudness targets or perceptual clarity.
Equalisation (EQ)	A processing technique that adjusts the relative level of specific frequency bands within an audio signal. In mastering, EQ is used to

	correct tonal imbalances, ensure spectral consistency, and improve translation across different playback systems.
Limiting	A form of dynamic processing that prevents the audio signal from exceeding a defined ceiling level. In mastering, a limiter is applied as the final stage in the processing chain to ensure compliance with true-peak delivery requirements for streaming platforms.
Loudness normalisation	A process by which streaming platforms and broadcast systems adjust playback level to a target loudness value measured in LUFS. Loudness normalisation reduces the competitive incentive to produce maximally compressed masters, since excessively loud material is simply attenuated on playback.
Loudness war A long-standing competitive trend in which producers and mastering engineers compress and limit masters to achieve the loudest possible perceived volume.	The loudness war has been particularly pronounced in EDM and techno production, where competitive loudness on dance floors and streaming platforms creates ongoing pressure toward reduced dynamic range.
LUFS Loudness Units relative to Full Scale.	The standard unit used to measure integrated programme loudness according to ITU-R BS.1770 and EBU R 128. LUFS values are negative, with values closer to zero indicating louder material.
Masking (psychoacoustic)	A phenomenon in which the presence of one sound reduces the audibility of another. In the mastering context, masking is most relevant in the low-frequency range, where bass and kick drum content compete for spectral space, and in the mid-frequency range, where vocals and other instruments interact.
Mastering policy In the context of this thesis, a multilayer perceptron that maps an eight-dimensional feature	The policy is the only trainable component in the pipeline and is responsible for all mastering-related decisions.

vector derived from stem and mix analysis to a bounded set of predicted mastering parameters.	
Oracle stems	Isolated instrument tracks available as ground truth in the MUSDB18-HQ dataset, used during training. Oracle stems are perfectly separated recordings rather than model-estimated outputs, and their use during training creates a mismatch with the Demucs-estimated stems encountered at inference time.
Proof of concept	A working implementation developed to demonstrate the technical feasibility of a design approach rather than to deliver a production-ready system. A proof of concept validates that the core workflow is possible and that key components can be integrated, without necessarily achieving the full performance or robustness of a finished product.
Self-supervised learning	A training approach in which the supervision signal is derived from the data itself rather than from externally provided labels or targets. In this thesis, the mastering policy is trained using losses computed directly from the processed audio output, rather than from a dataset of paired unmastered and mastered tracks.
Source separation	The computational task of isolating individual sound sources from a mixed audio signal. In music, this typically means estimating separate tracks for drums, bass, vocals, and other instruments from a stereo mixture. The quality of separation directly affects the reliability of any downstream analysis performed on the individual stems.
Stem	A single isolated source track produced by a source-separation model. In this thesis, stems refer to the four outputs produced by Hybrid Transformer Demucs: drums, bass, other, and vocals.
True peak	The maximum instantaneous signal level after inter-sample reconstruction, as opposed to simple sample-peak measurement. True-peak measurement is required by streaming platform delivery

	specifications to prevent clipping during codec processing and playback.

1 Introduction

1.1 Background and context

Audio mastering is the final stage between mixing and distribution, where a completed mix is prepared for release through operations such as cleanup, level adjustment, processing, and output formatting. Because these interventions are applied to the finished stereo signal rather than to separate tracks, mastering is both a technical finishing step and a practice in which technical control, musical balance, and aesthetic judgment converge.

In recent years, this stage of music production has become increasingly shaped by artificial intelligence and data-driven automation. Commercial platforms such as LANDR popularized automated mastering by analysing uploaded audio and applying processing such as equalization, compression, saturation, stereo enhancement, and limiting through an algorithmic workflow intended to approximate conventional mastering practice. This development matters because it changes who can access mastering services, how quickly masters can be produced, and what is considered an acceptable production workflow in both independent and commercial music contexts.

At the same time, mastering remains inseparable from human listening. Psychoacoustics shows that hearing is not a purely mechanical registration of sound pressure, but a perceptual process shaped by the ear and brain, including nonlinear loudness perception, masking effects, frequency sensitivity, and sound localization. In practice, this means that technically similar signals can still be judged differently when differences affect perceived loudness, clarity, punch, spatial impression, or listening fatigue. The relevance of this becomes even clearer in the mastering context, where loudness normalization frameworks such as EBU R 128 were developed precisely because peak-based normalization does not reliably reflect perceived loudness[21].

A further complication is that contemporary AI-assisted audio workflows often depend on upstream processes that can introduce perceptual side effects of their own. Research on music source separation shows that model output can contain artifacts such as spectral bleed, contamination between parts, phase-related hollowness, and other audible anomalies. Even when such systems perform well on objective separation metrics, subjective evaluations still report quality differences and contamination, which illustrates a broader problem in this field: measurable performance and perceived quality do not always align.

1.2 Goals

This thesis is guided by the central research question:

"How can a stem-aware neural architecture be used to predict and apply mastering parameters for the autonomous production of audio masters that meet professional streaming standards and approach the perceived quality of human-made masters?"

To support this exploration, a series of sub-questions have been defined that examine the system's conceptual basis, perceptual relevance, and practical limitations:

1. *What defines audio mastering as both a technical and aesthetic practice?*
2. *How do current AI mastering systems analyse and process audio?*
3. *Which psychoacoustic principles are most relevant when evaluating mastered sound?*
4. *Which technical metrics and standards are useful in mastering evaluation, and where do they fall short?*
5. *Which perceptual and practical limitations currently prevent AI mastering from demonstrating consistent equivalence with human mastering practice?*

The research is framed around perceptual equivalence rather than automation alone. AI systems may already be capable of optimizing measurable parameters and delivering commercially usable outputs in some contexts, but that does not automatically imply that they match human engineers in psychoacoustic judgment, contextual interpretation, or artistic sensitivity.

1.3 Tools and methodology

The method used in this study combines a systematic literature review, the design and implementation of a working research prototype, objective evaluation of that prototype, and external perspectives gathered through expert consultation and the observation of practitioner discourse in relevant audio engineering communities.

The theoretical foundation of the study is established through a literature review addressing audio mastering practice, intelligent mastering systems, music source separation, psychoacoustics, and loudness standardisation. This review is presented in Chapter 2 and is structured around the sub-questions introduced above.

The practical component consists of a stem-aware automatic mastering prototype implemented in PyTorch, in which a pretrained source separation model is used to derive analytical stem

representations from a stereo mix, a compact set of psychoacoustically motivated features is extracted from those stems, and a learned neural policy predicts mastering parameters for a differentiable DSP chain applied to the original mixture. The prototype was developed and evaluated using the MUSDB18-HQ dataset, with objective metrics including integrated loudness error, true-peak compliance, and masking-related measures used to assess its behaviour across iterative development phases.

External perspectives on the research outcomes are gathered through direct consultation with a professional audio engineer and through the observation of ongoing practitioner discussions in online audio communities. These perspectives inform the critical reflection in Chapter 4 and the practical recommendations in Chapter 5.

1.4 Structure

This thesis is divided into six chapters, structured to follow the full research process from background to conclusion:

- **Chapter 1 – Introduction:** Presents the background and context, the central research question, the sub-questions, the method used, and the structure of the thesis.
- **Chapter 2 – Research:** Reviews the literature on audio mastering, AI mastering systems, psychoacoustics, loudness normalization, evaluation metrics, and the main technical bottlenecks.
- **Chapter 3 – Research outcomes:** Describes the developed stem-aware automatic mastering prototype, covering its architecture, processing pipeline, implementation choices, self-supervised training approach, iterative redesign across development phases, objective evaluation results, and the key challenges and limitations encountered.
- **Chapter 4 – Reflection:** Provides a critical evaluation of the research outcomes in relation to professional practice, drawing on expert consultation and practitioner community perspectives alongside a comparison with the findings of the literature review.
- **Chapter 5 – Recommendations:** Offers practical recommendations for practitioners, developers, and other relevant stakeholders based on the findings.
- **Chapter 6 – Conclusion:** Summarizes the key findings and directly answers the central research question.

This structure moves from theoretical basis to practical investigation and finally to critical interpretation. In that way, the thesis addresses the question of perceptual equivalence with both technical precision and professional relevance.

2 Research

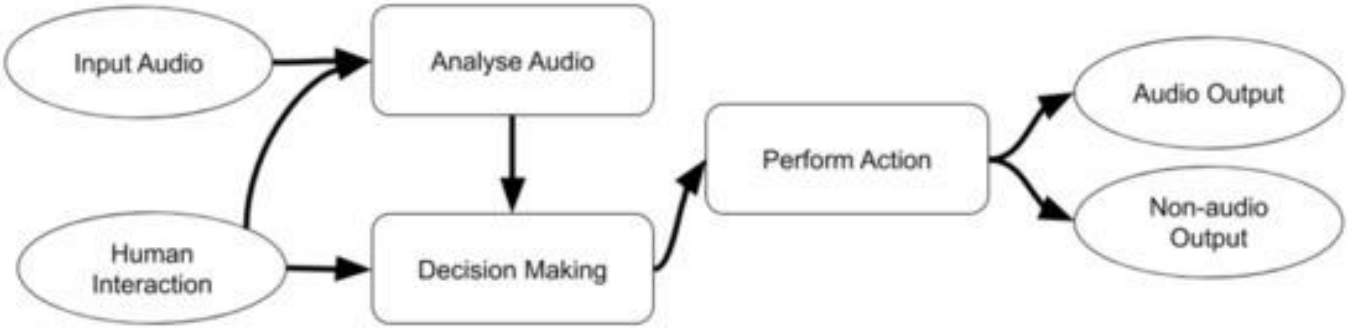
This chapter addresses the theoretical sub-questions introduced in Chapter 1 and positions the technologies that underpin the practical component of this thesis. It synthesizes the literature around the central problem of this study: how a stem-aware neural architecture can predict and apply mastering parameters to autonomously produce audio masters that meet professional streaming standards and approximate human-made masters in perceived quality [1]–[6].

2.1 Intelligent mastering in context

Audio mastering is the final production stage in which a completed mix is prepared for distribution through operations such as tonal correction, dynamic control, loudness optimization, stereo image refinement, and export formatting [1], [2]. In conventional studio practice, these interventions are not treated as isolated technical corrections, but as interdependent decisions that must preserve musical translation across playback systems while remaining stylistically appropriate [1], [2].

In current intelligent production systems, the core innovation is not the existence of processors such as equalization, compression, limiting, or saturation, but the automation of the decision layer that determines how those processors should be configured for a given signal [2], [3], [4]. Intelligent mastering should therefore be understood as a control and inference problem: the system must analyse the programme material, infer suitable processing actions, and apply them in a way that remains perceptually plausible rather than merely numerically compliant [3], [4].

This distinction is important in the context of this thesis. The practical system explored here does not operate as a conventional stereo-only auto-mastering service, but as a layered workflow in which source separation precedes mastering-stage analysis and processing. Intelligent mastering is therefore treated as a broader data-driven pipeline in which machine learning contributes to representation, prediction, and control rather than simply replacing one plugin chain with another [2], [4], [5], [6].



§Figure 1: Generalized intelligent music production workflow, adapted from [2].

2.2 AI foundations in intelligent mastering

From an AI engineering perspective, automatic mastering can be formulated as a supervised or reference-conditioned parameter-estimation task [4], [5]. The model receives an audio representation and predicts control variables or processing targets related to loudness, spectral balance, dynamic behaviour, stereo presentation, or effect intensity, after which a mastering chain applies those decisions to produce the output master [4].

This framing shows that mastering AI is not a single algorithm, but a stack of learned components. At the representation level, systems may operate on raw waveforms, magnitude or complex spectrograms, latent embeddings, or higher-level descriptors extracted from the programme signal [5], [6]. At the prediction level, recent literature uses convolutional encoders, U-Net-like architectures, recurrent modules such as biLSTMs, and reference-conditioned encoders trained to capture production style rather than musical content alone [5], [6].

The strongest adjacent literature for a stem-aware mastering thesis currently comes from automatic mixing and style-transfer research, because the number of mature deep-learning mastering papers remains limited [4], [5], [6]. Martínez-Ramírez et al. show that deep models can learn production-relevant controls such as loudness, EQ, dynamic range compression, panning, and reverberation from out-of-domain multitrack data [5]. Koo et al. extend this direction by using contrastive learning to disentangle audio effects and support reference-based mixing style transfer, which is highly relevant to systems that aim to infer production decisions from target examples or structured source representations [6].

For this thesis, that literature provides a clear conceptual basis. A stem-aware mastering model can be interpreted as a specialized parameter-prediction system in which the input representation is not limited to a single stereo mix, but is enriched through source-specific decomposition. In that sense,

source separation is not merely preprocessing, but part of the representational design of the AI system itself [4], [5], [6].

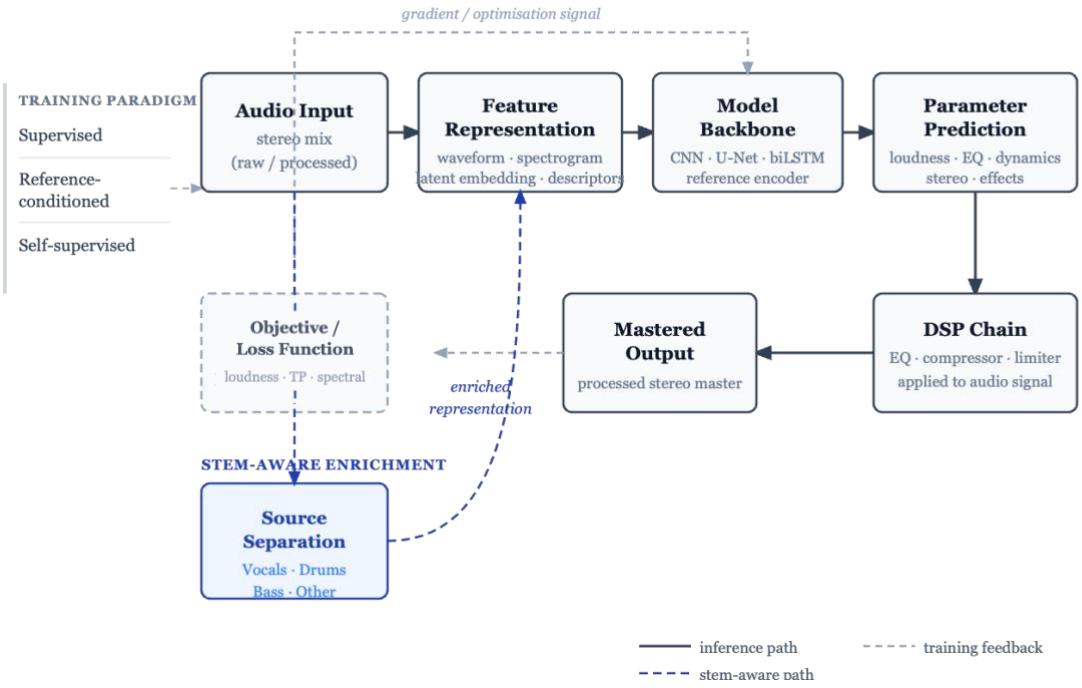


Figure 2: Generalised learning problem in intelligent mastering as a parameter-estimation pipeline: audio input is encoded into a feature representation, processed by a learned model backbone, and used to predict DSP control parameters.

A stereo-only mastering architecture analyses the final two-channel mix and must infer all relevant mastering decisions from a collapsed representation of the song. This design is efficient, but it also limits interpretability because masking, transient conflicts, spectral crowding, or source-specific imbalance must be inferred indirectly from the summed programme signal [2], [4].

A stem-aware architecture attempts to reduce that ambiguity by estimating source-specific representations before the mastering stage, typically for vocals, drums, bass, and a residual accompaniment or "other" stem [7], [8]. Once such stems are available, the system can examine source-specific energy distribution, temporal behaviour, transient density, or masking patterns in a way that is not directly accessible from stereo-only analysis [7], [8]. This gives the model a more structured basis for parameter prediction, especially when the downstream task involves balancing global mastering objectives with source-dependent phenomena.

In this context, the Demucs family is particularly relevant because it defines one of the strongest modern deep-learning front ends for music source separation [7], [8]. The original Demucs model is a waveform-domain encoder-decoder with U-Net skip connections, GLU-based convolutional blocks, and a bidirectional LSTM bottleneck that outputs separated waveforms directly [8]. Hybrid

Demucs extends this approach by combining a temporal branch and a spectral branch, merging both into a shared bottleneck, and improving performance through residual branches, local attention, and stabilization techniques such as singular-value regularization [7].

These architectural details matter because the practical implementation of this thesis is based on this same general logic: first estimate stems, then use those stems as structured input for downstream mastering analysis. Hybrid Demucs is therefore not cited as generic background, but as the technical basis for the source-aware stage of the proposed system [7]. The literature shows that the hybrid model improved average separation performance on MusDB HQ and also improved human judgments of quality and contamination compared with earlier waveform-only Demucs configurations [7]. That makes it a defensible front-end choice for a stem-aware mastering pipeline.

At the same time, Demucs should be positioned within the wider model landscape. Important comparison points include Wave-U-Net in the waveform domain, Open-Unmix and D3Net in the spectrogram domain, LaSAFT for conditioned source separation, and ResUNetDecouple for magnitude-phase-aware separation [7]–[13]. Mentioning these models situates the selected architecture within the broader source-separation literature and highlights it as a considered design choice rather than an arbitrary implementation.

For the present thesis, the key claim is therefore not simply that stems provide "more detail," but that they provide a different information topology for learning. A stereo-only system maps one mixed representation to mastering decisions, whereas a stem-aware system maps a structured multi-source representation to mastering decisions. This may improve source-specific awareness, but it also creates direct dependence on separation quality and on the artifacts introduced by the separator itself [7], [8].

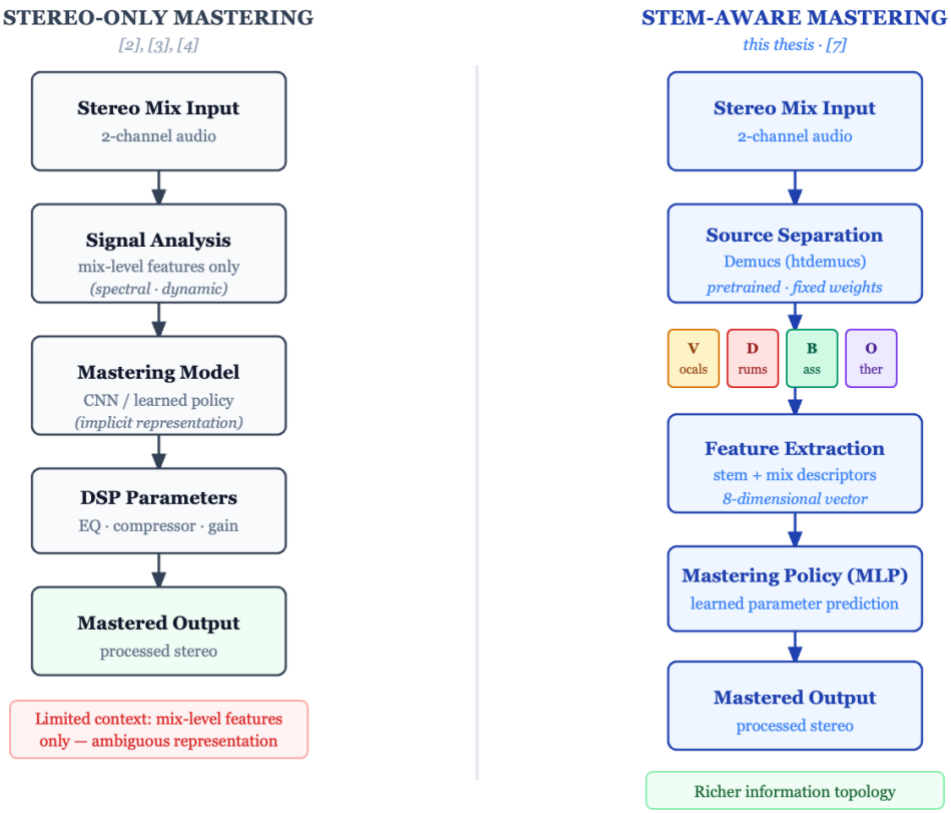


Figure 3: Comparison of stereo-only (left) and stem-aware (right) mastering architectures. In the stem-aware design, source separation decomposes the stereo mix into individual stems prior to feature extraction.

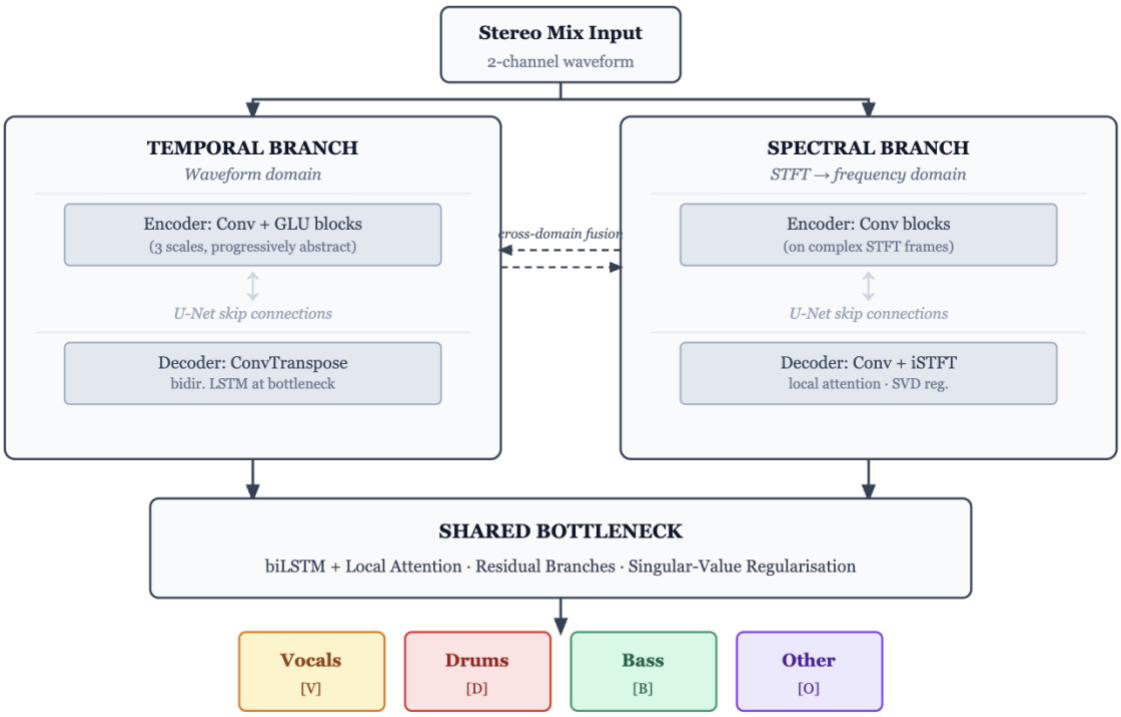


Figure 4: Simplified architecture of Hybrid Demucs (htdemucs) as used in this thesis.The four separated source stems serve as inputs to the feature extraction stage of the mastering pipeline. Adapted from [7], [8].

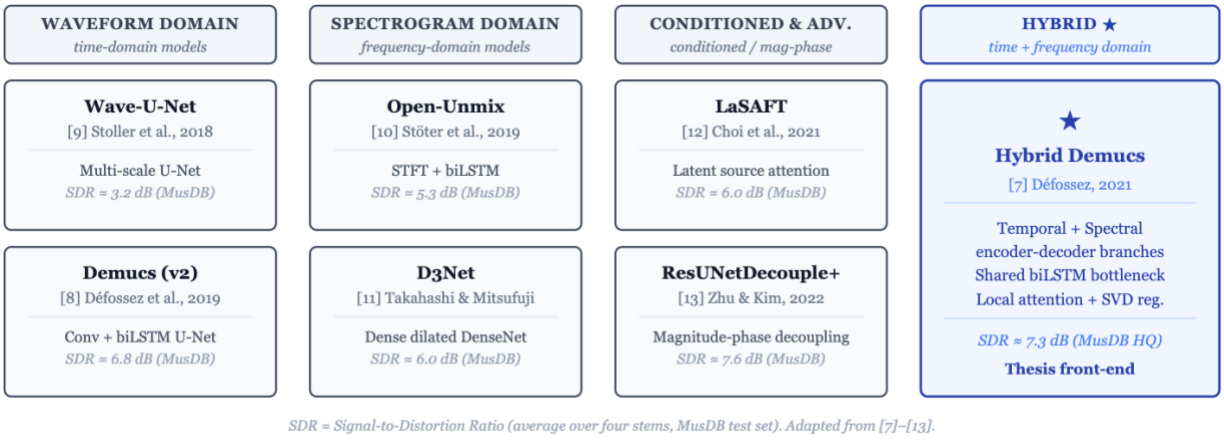


Figure 5: Taxonomy of music source separation models organised by processing domain. Models are grouped into waveform-domain, spectrogram-domain, conditioned/magnitude-phase, and hybrid. Adapted from [7]–[13].

2.3 Hearing, loudness, and what meters miss

Even in a strongly technical pipeline, mastering quality cannot be reduced to engineering compliance alone. The mastering literature repeatedly connects effective decision-making to perceptual variables such as loudness balance, masking, timbral clarity, fatigue, and translation across listening conditions [1], [2], [14]–[16]. This means that psychoacoustic relevance remains central even when the processing decisions are estimated by a machine-learning system.

The primary technical standards governing loudness in modern mastering practice are ITU-R BS.1770 and its broadcast application EBU R 128, which define integrated programme loudness in LUFS and establish a delivery framework with a target of -23 LUFS for broadcast and approximately -14 LUFS for major streaming platforms, alongside a true-peak ceiling intended to prevent clipping during codec processing and playback [21]. These standards represent a genuine improvement over earlier peak-based normalisation, because integrated loudness correlates more reliably with perceived loudness across a full programme than instantaneous peak measurements do [14], [21]. However, they also have well-documented perceptual limits. A master that meets an integrated loudness target can still sound harsh, spectrally unbalanced, or dynamically flat, because the measurement is agnostic to timbral quality, transient character, stereo presentation, and the way dynamic movement shapes listener engagement over time [1], [2], [14]–[16]. True-peak compliance similarly guarantees only that the signal will not clip at the codec stage; it says nothing about whether the dynamic processing used to reach that ceiling has introduced audible compression artefacts or distortion. For this thesis, these limitations are directly relevant: the prototype targets LUFS error and true-peak compliance as its primary quantitative objectives, and the same standards that make those targets measurable also define precisely where metric compliance ends and perceptual judgment begins.

This point is particularly important for a thesis that combines source separation and mastering. The Demucs literature itself does not justify performance only through objective metrics, but also through human listening evaluations that separately assess artifacts and contamination [7], [8]. The reason is straightforward: objective source-separation scores do not fully explain perceived quality. The same problem becomes even more critical at the mastering stage, where a signal may satisfy loudness or peak targets and still sound harsh, flat, crowded, unstable, or stylistically wrong.

For that reason, the evaluation logic of this thesis should remain explicitly dual. Technical metrics and standards provide reproducible constraints, while perceptual listening is needed to assess whether the resulting master is musically convincing and professionally acceptable [1], [2], [14]–[16]. Psychoacoustics therefore does not sit outside the AI discussion; it defines the perceptual boundary conditions within which intelligent mastering has to operate.

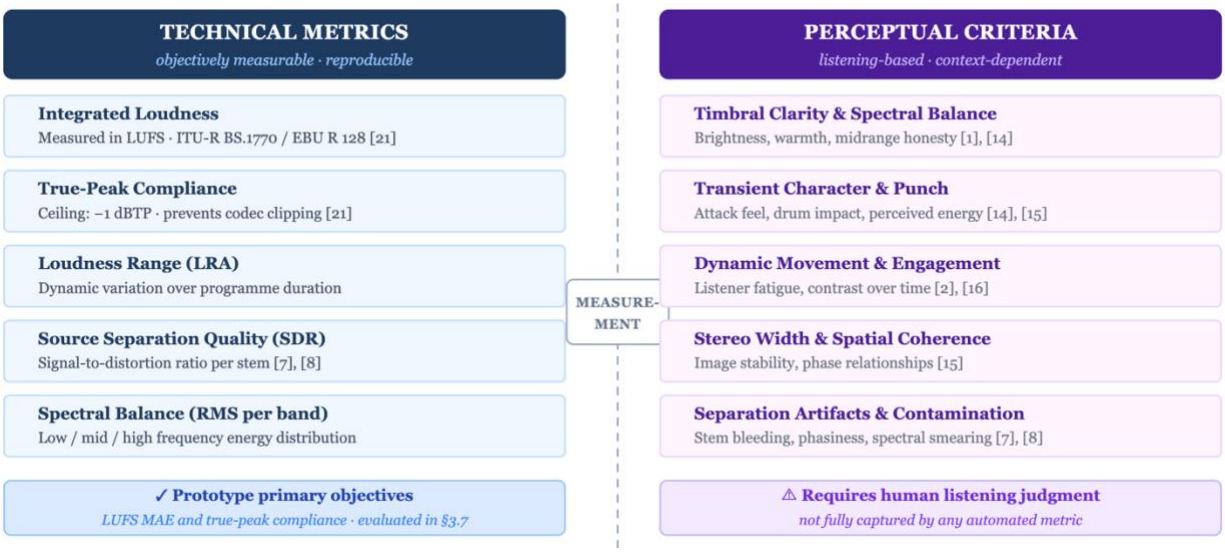


Figure 6: Comparison of technical metrics and perceptual criteria in source-aware mastering evaluation.. Adapted from [1], [2], [7], [8], [14]–[16].

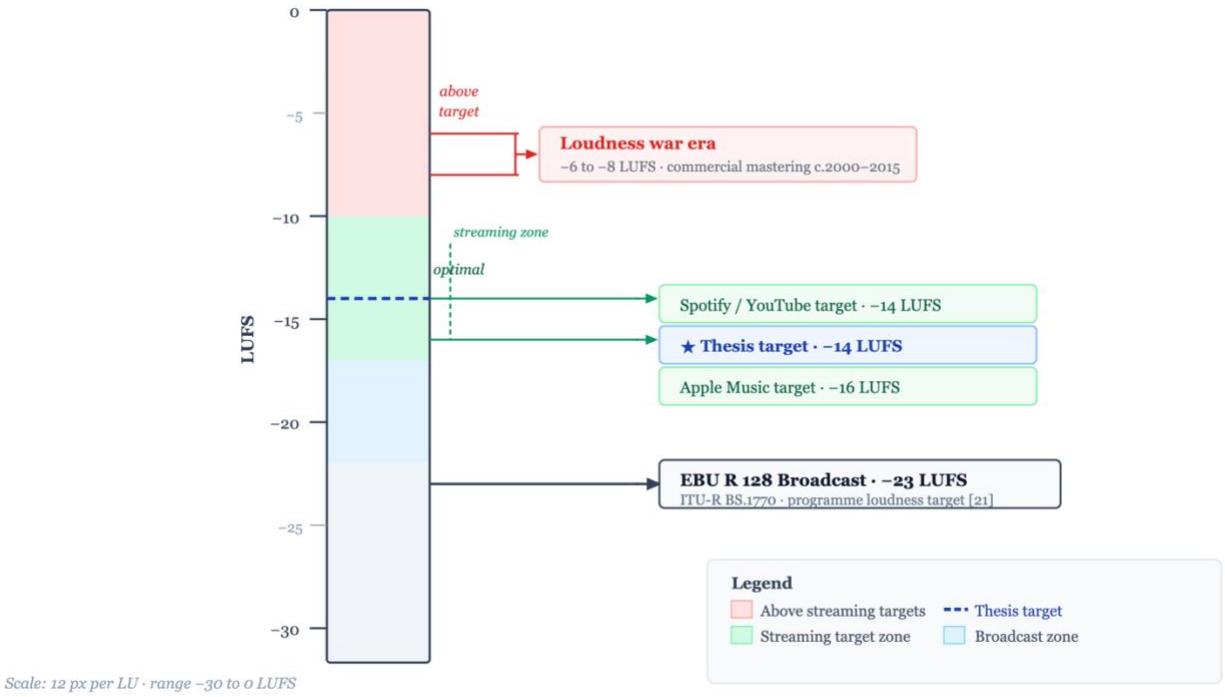


Figure 7: Integrated loudness reference scale (LUFS) showing key delivery targets and the thesis prototype target. Adopted: [21], ITU-R BS.1770.

2.4 Current capabilities and limits of AI mastering

The present state of the field supports a balanced conclusion. AI-based mastering and adjacent automatic-mixing systems already show that deep models can estimate production decisions that are useful, scalable, and in some cases perceptually competitive for constrained tasks [4], [5]. This is a

meaningful step beyond older automation systems that relied mainly on fixed templates or handcrafted heuristics.

At the same time, the core bottlenecks are now relatively clear. The first bottleneck is data: high-quality paired premaster-to-master datasets remain scarce, and even multitrack datasets often contain material that is not neutral enough for straightforward supervision [5], [6]. The second bottleneck is representation: if the mastering system depends on separated stems, then all downstream decisions inherit the errors of the separator, including bleed, phase inconsistency, spectral holes, and source misallocation [7], [8]. The third bottleneck is evaluation: objective metrics can quantify improvement, but they do not resolve questions of aesthetic preference, stylistic suitability, or professional acceptability on their own [7], [8], [14]–[16].

For a thesis built around stem separation followed by mastering, this is the correct academic position to take. The contribution is not the claim that AI already replaces human mastering, but that a stem-aware neural pipeline can improve parameter inference by exposing source-level information that is hidden in stereo-only systems, while still remaining constrained by separation quality, dataset quality, and the persistent gap between numerical optimization and perceived mastering quality [4], [5], [7], [8].

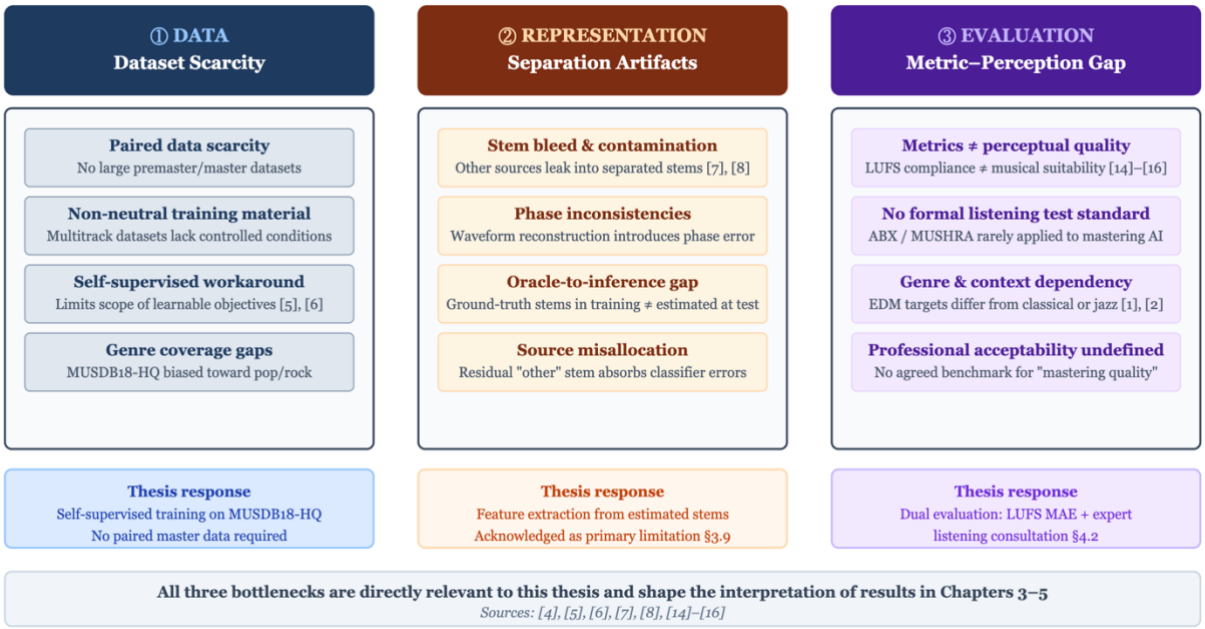


Figure 8: Three main bottlenecks in stem-aware intelligent mastering. Sources: [4]–[8], [14]–[16].

3 Research outcomes

3.1 Overview of the developed software

The developed result is a stem-aware automatic mastering prototype that processes a stereo music file in four consecutive stages: source separation, feature extraction, parameter prediction, and mastering application. In concrete terms, the system first separates the input mixture into drums, bass, other, and vocals by means of a pretrained Hybrid Transformer Demucs model, then computes a compact set of stem-aware and mix-aware audio features, and finally predicts mastering parameters for a differentiable DSP chain that is applied to the original mix.

The software was conceived from the start as a proof of concept for the central thesis question rather than as a finished commercial mastering product. The original success criterion at the beginning of development was to achieve stem separation and reconstruction that stayed as close as possible to the original broken-down wave material, after which those separated stems could be used as analytical input for an intelligent mastering stage. That broad objective remained stable throughout the project.

In its final form, the system accepts common audio formats such as WAV, FLAC, MP3, OGG, M4A, and AIFF, and produces a mastered WAV file, separated stem files, and a metrics file containing values such as LUFS, true peak, masking-related measures, and the predicted mastering parameters. Besides the command-line pipeline used for training and evaluation, the project also includes a self-contained local web demonstrator intended for jury presentation, which packages the backend, frontend, checkpoint, and example material into a runnable application.

From a technical perspective, the separator itself was not developed in this project. The original contribution lies in the way the separated stems are analysed, how those measurements are mapped to mastering decisions through a lightweight neural policy, and how the mastering chain is implemented as a differentiable sequence of DSP operations that can be trained without ground-truth mastered references.

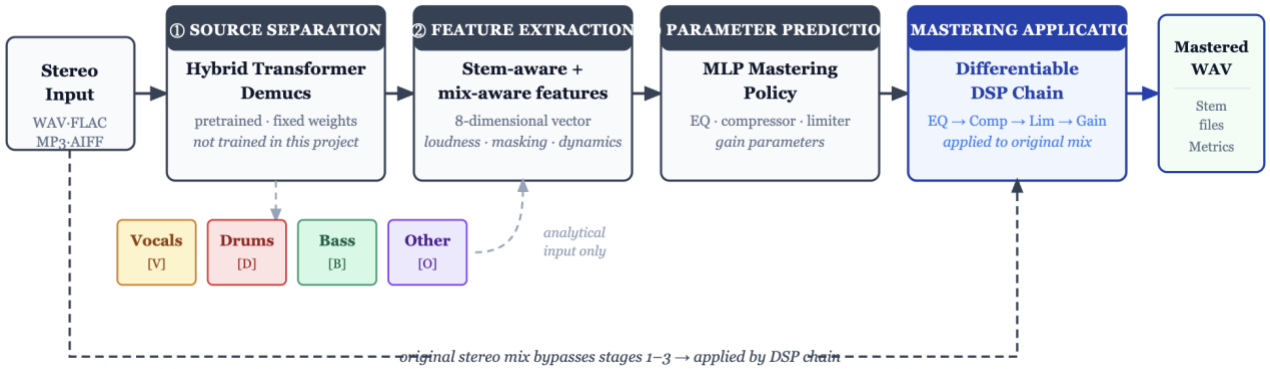


Figure 9: Global system pipeline of the stem-aware automatic mastering prototype. Outputs include the mastered WAV file, individual stem files, and a metrics file.

3.2 Application structure and processing flow

The internal structure of the application follows a clear modular organisation. The core logic is concentrated in the *src* package, which contains the dataset loader, separator wrapper, stem analysis logic, feature schema, policy network, differentiable mastering actions, training code, and the end-to-end system orchestrator. Around that core, the project provides separate scripts for training, evaluation, demo inference, benchmarking, plotting, and presentation support, while the final jury deliverable packages the same logic into a local FastAPI and JavaScript application.

At runtime, the pipeline operates in four conceptual phases. First, the input audio is loaded and resampled to 44.1 kHz if necessary. Second, the mixture is separated into four stems. Third, the stem-aware and mix-aware feature vector is extracted and optionally normalised using the statistics stored in the model checkpoint. Fourth, the policy predicts mastering parameters, after which the differentiable mastering chain is applied to the original stereo mixture to generate the final master.

A crucial design decision is that the project is stem-aware in an analytical sense rather than in an operational stem-mastering sense. The separated stems are not individually processed and re-summed into a new mix; instead, they are used to derive information about masking, loudness range, and interaction between sources, and that information conditions the mastering decisions that are then applied to the original stereo file. This distinction is important because it keeps the system focused on decision support through source-aware analysis while avoiding an additional layer of reconstruction artifacts that would arise from re-mastering and summing processed stems.

This modular structure also made the project easier to evolve. The repo evidence shows a clear separation between the analytical modules, the trainable policy, the mastering actions, and the delivery layer, which allowed to modify the learning strategy later without having to redesign the

entire application. That flexibility became especially important when the first policy configurations failed to achieve acceptable loudness behaviour and the training process had to be adjusted substantially.

3.3 Underlying technologies and core implementation choices

The most important upstream technology choice was the adoption of pretrained Hybrid Transformer Demucs (htdemucs) as the source separation front end. Demucs was selected because it offered a strong practical balance between published performance, availability of pretrained checkpoints, and implementation convenience inside a limited development window. Alternative separation frameworks were explored informally during development, but Demucs was retained because it gave the most convincing practical outcome for this project; since those informal trials were not logged systematically, they are not treated here as benchmark evidence.

The project uses PyTorch and torchaudio as its technical backbone, and that choice is central to the entire design. The neural policy, the stem-aware analysis functions, and the mastering DSP chain are all implemented in torch-compatible operations, which makes it possible to backpropagate audio-domain losses directly through the mastering stage. This means the system does not require a dataset of reference mastering settings or matched mastered targets, because the model can instead be trained through losses on loudness, masking, peak behaviour, and regularisation.

A second important implementation choice was to keep the mastering stage as a fixed differentiable DSP chain rather than attempting a fully end-to-end audio-to-audio neural mastering model. That decision was both methodological and practical: the goal of the project was to test whether stem-aware features could improve mastering decisions, and a full end-to-end neural audio model would have gone beyond the available time, computational budget, and proof-of-concept scope. As a result, the project remains interpretable at every stage, since the model does not generate audio directly but predicts a bounded set of meaningful parameters for EQ, compression, limiting, and, in the later phase, global gain.

The final demonstrator uses FastAPI on the backend and a lightweight vanilla HTML, CSS, and JavaScript frontend. This delivery choice was not the main scientific contribution of the project, but it was important for turning the prototype into a usable jury artefact that could run locally on CPU without requiring a cloud service or DAW integration.

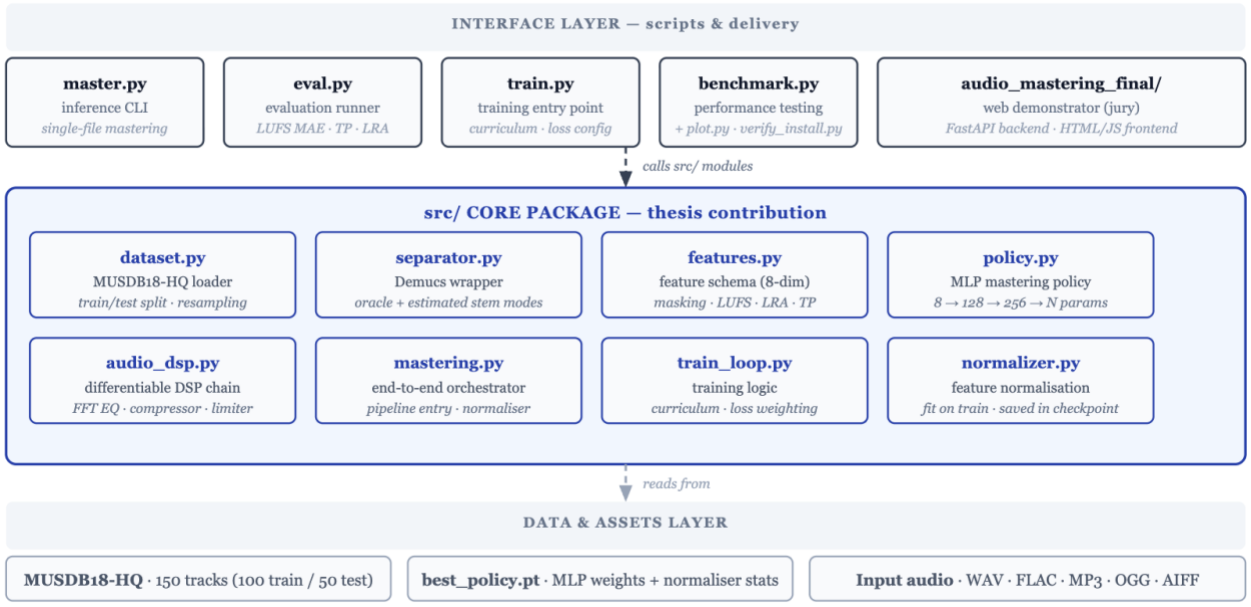


Figure 10: Three-layer software architecture of the stem-aware mastering prototype. The data and assets layer (bottom) provides the training corpus, trained model checkpoint, and supported input formats.

3.4 Stem separation and feature extraction

The separation stage is the enabling component of the whole research outcome. Without source separation, the system would not have direct access to the specific interactions that motivate the thesis, such as low-end competition between bass and drums, or mid-range clashes between vocals and the remaining instruments. For that reason, the project performs source separation at the very start of the pipeline, before any mastering-related analysis or decision-making takes place.

In implementation terms, the project uses Demucs in two ways. The core system can invoke the model directly through a Python API wrapper, while the web application also provides a CLI-style subprocess route that was added as a robust fallback when torchaudio and torchcodec issues on macOS made direct audio I/O unreliable. The separated outputs are standard MUSDB-style stems named drums, bass, other, and vocals, and those stems are explicitly downmixed to mono before analysis in order to keep the masking and loudness-range computations stable and unambiguous.

A particularly important methodological detail is that Demucs is used only at inference time in the final system, whereas the policy is trained on the ground-truth stems already available in MUSDB18-HQ. In other words, the learning process operates on ideal stem information, while real deployment relies on estimated stems. This creates an oracle-versus-inference gap that must be acknowledged honestly, because the policy is ultimately asked to generalise from perfect training stems to imperfect separated stems that may contain leakage or artifacts.

The feature extraction stage translates those stems into a fixed eight-dimensional representation. The chosen features are low-end masking, mid-range masking, mix loudness range, vocal loudness range, integrated mix loudness, mix true peak, loudness error to target, and true-peak margin. This feature design is a strong fit for the research question, because it compresses the most relevant mastering pressures into a small and interpretable state vector: bass and kick competition, vocal clarity, macro-dynamics, loudness position relative to streaming targets, and peak safety.

The low-end masking feature is computed from the overlap between bass and drums in the 20 Hz to 200 Hz region, while the mid-range masking feature evaluates overlap between vocals and the "other" stem in the 200 Hz to 4 kHz range. These two measures matter because mastering often involves negotiating space in exactly those regions: low frequencies are especially prone to energy competition, while the mid band is critical for intelligibility and instrumental balance. The loudness-range features further extend this by capturing broader dynamic behaviour, which is important because mastering is not only about spectrum and peak control but also about how a mix evolves over time.

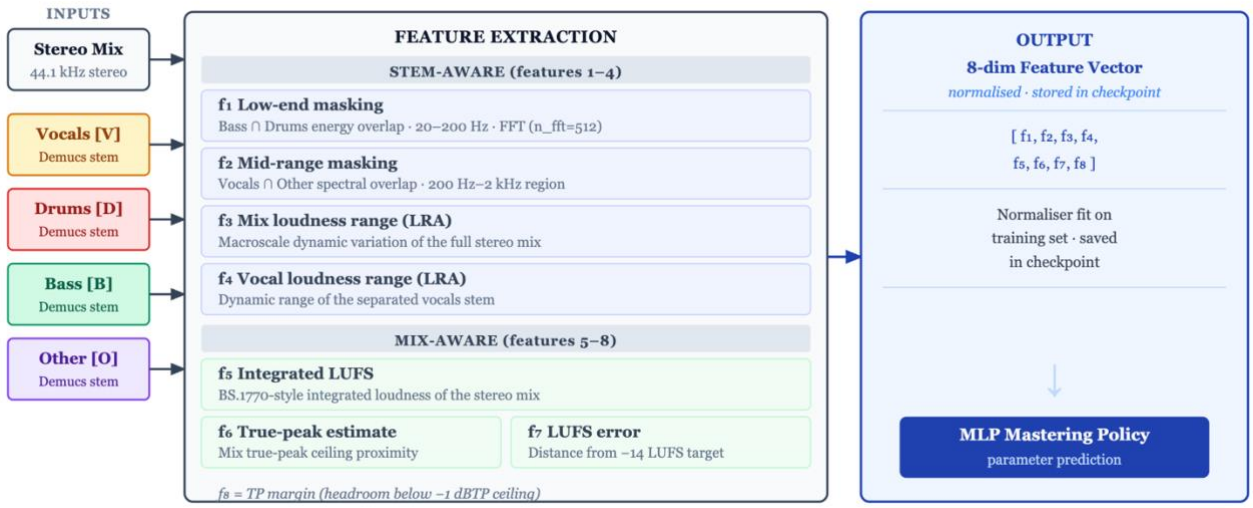


Figure 11: Feature extraction scheme showing how Demucs-separated stems and the stereo mix are combined into the 8-dimensional feature vector. The normaliser is fit on the training set and saved in the model checkpoint.

3.5 Mastering policy and differentiable DSP chain

Once the feature vector has been extracted, it is passed to a small multilayer perceptron called the mastering policy. This network maps the eight features to a bounded set of mastering parameters with clear physical meaning: low-band EQ, mid-band EQ, high-band EQ, compressor threshold, limiter drive, and, in the later architecture, global gain. The bounded output ranges were

deliberately chosen to keep the predicted actions inside a realistic mastering space and to prevent unstable parameter explosions during training.

The mastering chain itself is fixed and classical in structure. It consists of a three-band frequency-domain EQ, a soft-knee RMS compressor, a soft tanh limiter with a fixed true-peak-oriented ceiling, and, in the improved version, an additional global-gain stage followed by a safety re-limiting pass. This means the project should be described as AI-driven parameter prediction for a fixed mastering chain, not as an end-to-end generative mastering model.

That characterisation is important for two reasons. First, it is technically accurate: the learned component decides which parameter values to apply, but the processing chain itself is hand-engineered. Second, it preserves interpretability, because every model output can be inspected directly in terms of decibels, thresholds, and limiter drive rather than as an opaque latent transformation.

Another strength of this design is that the DSP chain remains fully differentiable. The EQ is implemented through FFT-domain filtering with smooth crossover masks, the compressor uses a differentiable RMS envelope and soft-knee transfer, and the limiter relies on a tanh nonlinearity that guarantees bounded output. As a consequence, the project can optimise mastering behaviour by comparing the processed signal against audio-domain objectives, even though no target master exists.

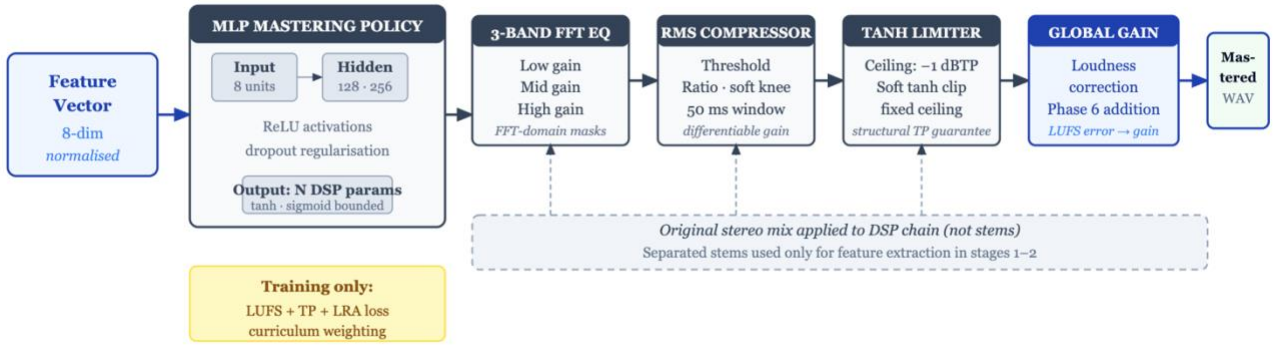


Figure 12: Mastering policy MLP and differentiable DSP chain. The 8-dimensional feature vector is passed through a multilayer perceptron (8→128→256→N parameters) which predicts bounded DSP parameters.

3.6 Training process and iterative development

The training process became the real centre of the development work. The separator could be integrated relatively early, but the policy stage required repeated redesign because obtaining stable

and useful mastering behaviour from a small interpretable feature space turned out to be much harder than simply building the pipeline itself. This is also consistent with the project materials, which show multiple checkpoints, multiple configuration phases, and an explicit diagnosis report devoted to the loudness problem.

The first stable configuration used a relatively compact MLP and a mastering action space without global gain. From the codebase evidence, this phase corresponds to the baseline setup that later evolved into the more explicit Phase 5 and Phase 6 distinction. The general training logic already existed at that point: the model was trained by minimising a combination of loudness loss, true-peak-related distortion loss, masking loss on processed stems, and EQ regularisation.

The most important failure appeared in the early loudness behaviour. The project’s own evaluation materials show that the Phase 5 result was far from the intended target, with a mean LUFS error of 24.87 LU on the evaluated tracks. The diagnosis report linked that failure to an insufficient action space, the absence of an effective curriculum, and problematic training conditions that did not allow the network to correct large loudness errors reliably enough.

This moment was the main turning point in the research process. Rather than presenting the first implementation as a finished result, the project was explicitly reworked, which fits the thesis template’s expectation that interim results should lead to continuous adjustment of the research process. The redesign was decided during development on the basis of technical findings rather than external prescription.

The resulting Phase 6 architecture introduced several coordinated changes. The policy network became wider and deeper, a new global-gain output head was added, and a curriculum mechanism was introduced to make loudness the dominant objective early in training before gradually rebalancing toward masking and distortion terms. In practical terms, this meant the model finally had a direct way to perform large loudness corrections instead of having to approximate them indirectly through limited EQ, compression, and limiter actions alone.

This redesign improved the system substantially. The final recorded evaluation shows that LUFS mean absolute error dropped from 24.87 LU in Phase 5 to 2.68 LU in Phase 6. True-peak compliance was maintained at 100 percent across the evaluated tracks; this result is structurally supported by the tanh limiter, which bounds sample peak by design, though no independent oversampled true-peak measurement was performed at inference as a separate verification step. At the same time, the project remained honest about the remaining gap: even after the redesign,

loudness targeting did not fully reach the original ambition of near-perfect target compliance, and only a minority of evaluated tracks landed within 1 LU of the intended target.

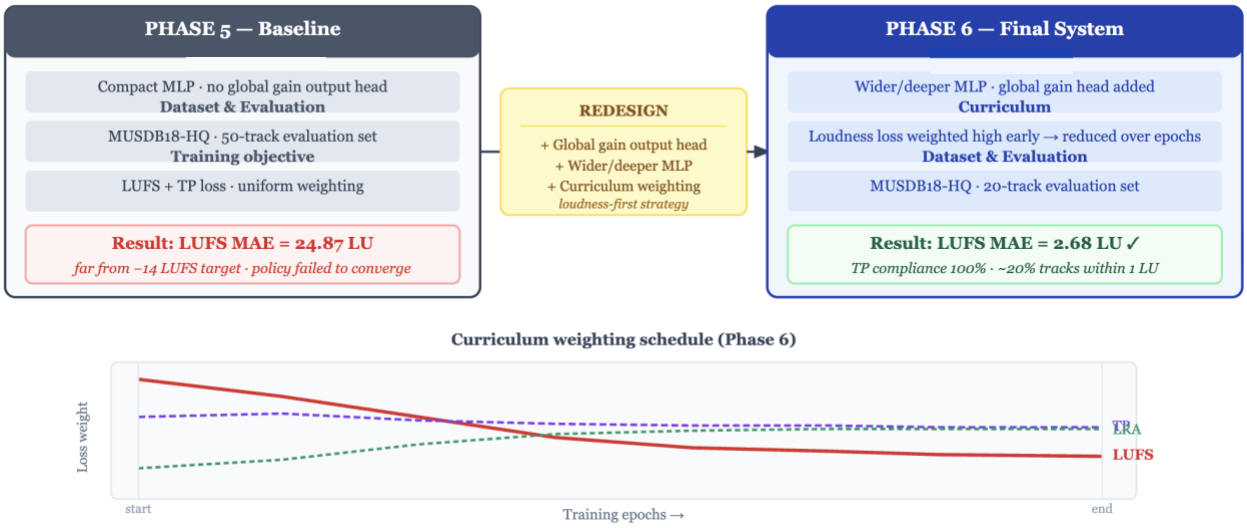


Figure 13: Training phase timeline and curriculum weighting schedule.

3.7 Evaluation and practical outcome

The project outcome can therefore be described as partially successful and technically meaningful. It achieved a fully functioning end-to-end prototype that combines pretrained source separation, psychoacoustically motivated feature extraction, learned parameter prediction, differentiable mastering DSP, offline evaluation, and a runnable local demonstrator. This is a significant result because all core parts of the proposed research idea were implemented and connected into a single working system.

At the same time, the evaluation results show that success was uneven across objectives. True-peak compliance was structurally strong because the limiter guarantees bounded output, and loudness targeting improved strongly after the Phase 6 redesign. However, the project materials also show that masking improvements were not uniformly positive, since low-end masking improved on average while mid-range masking slightly worsened in the recorded Phase 6 metrics.

A methodological note is relevant to interpreting these results. All loudness measurements in this project — both those used to compute the training loss and those used to calculate the evaluation LUFS MAE — were computed using an internal BS.1770-style implementation in PyTorch. The values are internally consistent across both phases, which ensures the phase comparison is valid. However, they may differ slightly from measurements taken with validated reference tools such as pyloudnorm or ffmpeg-based loudness analysis. This should be taken into account when comparing

the reported LUFS MAE figures against measurements from external pipelines or professional metering tools.

Table 1 summarises the most relevant numerical comparison between the two main trained stages.

Phase	Main architectural state	LUFS MAE	Within 1 LU	True-peak compliance
Phase 5	No global-gain head; earlier policy configuration	24.87 LU	Not acceptable in practice	100%
Phase 6	Wider policy, global-gain head, curriculum-based training	2.68 LU	20%	100%

From an informal listening perspective during development, the later version also produced a better overall musical result than the earlier one. These listening impressions were not part of a formal evaluation protocol for the prototype itself, so they should not be treated as scientific evidence. Still, they are useful as development observations: the improved version generally gave a more coherent and louder result, but drums remained a recurring weakness, and separation artifacts could still be heard on some material, especially when transients and instrumental boundaries were difficult to separate.

A further practical observation was that the system appeared to behave better on more melodic and instrument-led tracks than on heavier material with strong low-end emphasis. This should again be framed as an informed development observation rather than a formal genre study. It does, however, align with the broader technical constraints of the system: low-end-heavy material places more stress on both the separator and the mastering policy, especially when bass and drums compete strongly in the same region.

Figure 14: Phase 5 versus Phase 6 evaluation comparison.

3.8 Challenges, bottlenecks, and limitations

The most important bottleneck of the project was not building the pipeline, but making the policy learn useful mastering actions from a small feature representation. The repo evidence shows that repeated training iterations, checkpointing, and explicit redesign were necessary before the system produced acceptable loudness behaviour. In that sense, the research outcome is not simply the final model, but also the documented path from an underperforming configuration to a substantially improved one.

A second major challenge concerned engineering reliability. The project contains multiple workarounds for macOS-specific torchaudio and torchcodec issues, including a soundfile-based patching strategy and isolated subprocess execution inside the web application. These problems were not conceptually central to the thesis question, but they were practically significant because they affected audio I/O stability, background processing, and the deployability of the demonstrator.

The project also contains several methodological limitations that should be stated clearly. The separator was pretrained and unmodified, so the thesis cannot claim a contribution in source separation research itself. The policy was trained on MUSDB18-HQ stems but deployed on Demucs-estimated stems, which introduces an unresolved train-inference mismatch. The mastering chain is intentionally simple and does not include stereo-width processing, harmonic excitation, reference-track conditioning, genre embeddings, or more advanced mastering control structures.

At the feature extraction level, the FFT resolution used to compute the low-end masking feature (n_fft=512 at 44.1 kHz) yields approximately two to three frequency bins across the 20 to 200 Hz

range. This is a coarser resolution than the description of that feature might imply; a larger FFT window would provide finer spectral detail at the cost of additional computation. Similarly, the compressor stage uses a 50 millisecond RMS detection window, which is shorter than the 100 to 300 millisecond range typical in professional mastering practice. Both are deliberate tradeoffs chosen for the differentiable proof-of-concept context rather than oversights.

Another limitation concerns evidence quality. The repo does not contain real training-loss logs, formal listening-test logs for the prototype, or a stored benchmark comparison between Demucs and other separators. For that reason, the chapter must stay within what can actually be defended: the system clearly demonstrates a functional stem-aware mastering pipeline and a meaningful improvement between phases, but it does not prove equivalence with professional human mastering, nor does it solve automatic mastering as a general problem.

Finally, the final deliverable should be understood as a research demonstrator rather than a production system. It is self-contained, CPU-oriented, and suitable for showing the implemented method to a jury, but its latency is still dominated by the Demucs step, and its generalisation claims remain limited by the dataset, the fixed mastering chain, and the unresolved sensitivity to separation artifacts.

In summary, the research outcomes of this project are strongest when described precisely. The project succeeded in building a complete stem-aware mastering prototype, in demonstrating that separated-stem information can be integrated into a self-supervised mastering pipeline, and in showing that iterative redesign of the policy and action space can produce a substantial improvement in loudness behaviour. At the same time, the results also show that stem-aware analysis alone does not eliminate the practical difficulties of mastering automation, particularly when separation artifacts, low-end complexity, and limited control capacity continue to constrain the final output.

4 Reflection

4.1 Self-assessment against the research question

The central question guiding this thesis asks how a stem-aware neural architecture can be used to predict and apply mastering parameters for the autonomous production of audio masters that meet professional streaming standards and approach the perceived quality of human-made masters. The research outcomes described in Chapter 3 allow a direct and honest answer: the prototype demonstrates that such an architecture is technically feasible, measurably improvable through iterative redesign, and capable of meeting some quantitative streaming targets on evaluated material, but it does not yet demonstrate consistent perceptual equivalence with professional human mastering practice.

The strongest result in support of the design was the loudness targeting improvement following the Phase 6 redesign, which reduced the mean absolute LUFS error from 24.87 to 2.68 LU. True-peak compliance was maintained throughout at 100 percent, meeting a basic structural requirement for streaming delivery. At the same time, only approximately 20 percent of evaluated tracks reached within 1 LU of the intended target, and mid-range masking showed no consistent improvement on average. The prototype therefore achieves partial success relative to the quantitative criteria it was designed to optimise, while leaving a significant gap in terms of the perceptual quality claims that professional mastering implies.

4.2 Expert consultation: perspectives from a professional audio engineer

To ground this reflection in external professional perspective, the prototype was presented to a professional audio engineer with experience as both a performing musician in the metal genre and a techno producer and DJ. This background is particularly relevant because it combines firsthand experience of highly dynamic, multi-instrument material, where transient density, low-end control, and inter-instrument masking are critical, with the high-loudness demands of club and DJ contexts that characterise techno production and distribution.

The consultation surfaced several observations that are directly relevant to evaluating both the strengths and the remaining limitations of the prototype.

The first and most immediate reaction was that the stem-aware design constitutes a meaningful differentiator relative to existing fully automated mastering tools. The consultant noted that the

ability to analyse individual instrument buses before making mastering decisions is directly relevant to the practical challenge of achieving what he described as a sound equilibrium, meaning a condition in which no single instrument bus competes destructively with another in the final stereo image. This recognition validates the core design premise: that stems provide an analytical layer that is not accessible from a stereo-only input.

The consultant also raised an immediate concern about separation artifacts. Knowing that the system relies on a pretrained source-separation model, he questioned whether artifacts introduced during separation (such as spectral bleed between stems, phase inconsistencies, or transient smearing) could propagate into the mastered output. This concern is well-founded and documented in the source-separation literature [7], [8]. The design response, arrived at early in development, is that the mastering parameters are predicted using the separated stems as analytical inputs but are applied exclusively to the original, unprocessed stereo mixture. This means the mastered output is never directly affected by separation artifacts, because the separation stage has no downstream effect on the processed audio itself. When this was explained, the consultant confirmed that the design rationale was sound. He further observed that keeping predicted mastering parameters within bounded, normalised ranges adds an additional layer of safety: no single parameter decision can be large enough to introduce audible instability into the output, even if the feature vector contains some noise from imperfect separation.

A second area of discussion concerned parameter grouping and the evaluation of interdependent processing stages. The consultant suggested that parameters such as compression and saturation should ideally be evaluated in a linked A/B framework, in which the interaction of both stages is tested at the limits of the available parameter range before an optimal configuration is identified. This reflects standard mastering practice, where individual processor settings are not evaluated in isolation. It also points toward a genuine methodological limitation of the current system, which predicts all parameters from a single forward pass through the policy network without any mechanism for evaluating or adjusting the interaction of downstream processing stages.

The prototype was then tested directly on a small set of unmastered techno tracks. The results confirmed the quantitative improvement in loudness but also revealed limitations that objective metrics alone could not capture. Both the consultant and the author observed weaknesses in the perceived reproduction of high-frequency overtone content, particularly in the frequency range between approximately 2 kHz and 10 kHz. This region is critical for the intelligibility of transient events, the perceived clarity of melodic elements, and the subjective sense of definition and detail in

a professional master. The mastering chain's EQ stage addresses mid and high frequencies, but the policy's compact feature representation does not include a dedicated feature for this perceptual range, and therefore lacks a direct mechanism for learning from systematic errors in that region.

The conversation also addressed LUFS targeting in the techno and broader EDM context, specifically the observation that acceptable target loudness values are not uniform across genres. The consultant noted that in techno and club-oriented music, masters are commonly expected to sit between approximately -6 and -10 LUFS before streaming platform normalisation is applied, and that producers in this space are acutely aware of the ongoing loudness war, a competitive pressure toward increasingly loud and compressed masters that has characterised EDM production culture for decades. This observation aligns with the Phase 6 finding that loudness targeting remained inconsistent: the genre-specific target range matters more than a single universal LUFS value, and the current prototype does not incorporate any genre-conditioned loudness targeting mechanism.

The overall conclusion from this consultation was that automated mastering, even in a stem-informed configuration, remains far from a practical first choice for audio engineers in its current form. The consultant's assessment was that the perceived quality of a professional master derives more from what the engineer hears and interprets than from what the processing chain measures, and that this perceptual dimension is not yet capturable by a system that optimises objective feature values alone.

4.3 Professional and community perspectives on AI mastering

The assessment offered by the consulted audio engineer finds strong reinforcement in broader professional discourse. Abbey Road Studios mastering engineers Alex Gordon and Geoff Pesche have stated that the most important element missing from AI mastering is human interaction: the ability to contextualise sound, form an emotional response to a piece of music, and make decisions based on feel rather than technical correctness alone [17]. Gordon specifically notes that mastering is "not a 100% technical affair" and that it "can involve a lot of human decisions that are based on 'feel' rather than what is technically correct," and that working with a human engineer who understands your music is "invaluable" [17]. Pesche adds that AI systems lack the benefit of feedback from someone who "has spent their whole life listening to music and understands what needs to be done" [17].

This distinction between technical compliance and perceptual intent is articulated with particular clarity by mastering engineer Alexis Kevork, who describes the gap between human and automated

mastering in terms of the difference between statistical normalisation and artistic interpretation: "Where the human seeks a sonic identity through professional mastering services, the automated machine seeks conformity" [18]. Kevork argues that AI cannot navigate the transition from objective measurement to artistic intention, and that the boundary between "just a little more" and "already too much" represents the perceptual limit that separates a technically correct sound from an emotionally compelling one, a limit that cannot be derived from statistical models [18]. This is precisely the type of judgment that the prototype's fixed feature representation and bounded parameter space cannot currently support.

In the techno and EDM production community, practitioner discussions on dedicated mastering forums corroborate the genre-specific loudness challenge raised in the consultation. Longstanding threads on the Gearspace mastering forum document the complexity of EDM loudness expectations and the ongoing tension between competitive loudness, dynamic range, and streaming platform normalisation [19]. Community contributors identify that in four-to-the-floor genres such as techno, trance, and house, loudness decisions are deeply intertwined with how the fundamental energy of a track, particularly kick drum impact and bass weight, is perceived relative to other dance floor material [19]. This genre-level context is something a system with a fixed LUFS target and no genre conditioning cannot adequately address. On the question of whether AI mastering can replace human engineers, Gearspace discussions reflect the same position repeatedly: current tools are useful as a starting point or sanity check, but the professional community consistently identifies the absence of human judgment, contextual interpretation, and artistic sensitivity as the reason automated mastering remains unsuitable as a primary production choice [20].

4.4 Alignment with the research literature

The limitations identified through expert consultation and practitioner discourse align closely with what the research literature predicted. Chapter 2 identified three main bottlenecks in stem-aware intelligent mastering: data quality and paired dataset scarcity, separation artifact propagation, and the persistent gap between measurable performance and perceived mastering quality [4], [5], [7], [8]. The prototype encountered all three. Training was bounded by MUSDB18-HQ, limiting genre coverage and generalisation. Separation artifacts were a design concern addressed architecturally rather than eliminated. And the core evaluation finding, that loudness metrics improved substantially but a perceptual quality gap remained, directly exemplifies the metric-perception

mismatch that the psychoacoustics and intelligent mastering literature consistently identifies as the hardest problem in this field [14]–[16].

This alignment is not a failure of the project. It is evidence that the research process engaged with the right questions. The prototype explored a design space that the literature identified as theoretically promising, encountered the limitations the literature predicted, and produced outcomes interpretable within the same frameworks. That constitutes a coherent and defensible research result, even where it falls short of the perceptual quality that professional mastering practice demands.

5 Recommendations

5.1 Scope and basis of these recommendations

The recommendations in this chapter are directed at three audiences: developers and researchers building stem-aware or intelligent mastering systems, practitioners evaluating whether AI mastering tools are appropriate for their production workflow, and those who may want to extend or replicate the work presented in this thesis. All recommendations are grounded in the concrete outcomes of the prototype, the diagnosis and redesign process described in Chapter 3, and the professional and community perspectives gathered in Chapter 4. They do not represent generic advice about AI in audio, but specific observations about what worked, what failed, and what the project reveals about the wider design space.

5.2 Recommendations for developers and researchers

Design the action space before designing the model. The most important lesson from the development process described in this thesis is that the policy's action space, meaning the range and structure of the mastering parameters it can predict, must be designed to cover the real distribution of corrections the system will need to make. The Phase 5 failure was not primarily a modelling failure; it was an action space failure. The available combination of EQ, compressor, and limiter drive could not produce the loudness corrections required by very quiet input material, because no direct gain control existed. The Phase 6 global-gain addition, with its range of plus or minus 30 dB, was the architectural change that finally allowed the system to address large loudness errors. The practical implication for future systems is that action space analysis should come first: identify the worst-case input conditions the system will encounter, calculate how much correction would be needed to meet the output target, and ensure the parameter space spans that range before training begins.

Use curriculum training when multiple objectives interact. The Phase 6 curriculum, which started with loudness-dominant loss weighting before gradually rebalancing toward masking and distortion objectives, was essential for stable convergence. Without it, the model attempted to simultaneously satisfy competing objectives from the start of training, resulting in inconsistent updates that prevented effective loudness learning. In any self-supervised mastering pipeline where the loss combines loudness, spectral, and distortion terms, a curriculum that sequences those objectives from most fundamental to most nuanced is strongly recommended. The specific

transition point and weighting schedule will depend on the dataset and architecture, but the general principle applies broadly: prioritise loudness first, then balance the remaining objectives progressively.

Log training behaviour with sufficient granularity to support diagnosis. One of the practical limitations this project encountered was the absence of preserved training loss curves from earlier phases. When Phase 5 underperformed, a thorough diagnosis required inferring the failure mode from architecture and configuration evidence rather than from a direct reading of what the model had learned. Future projects should treat training logs as primary evidence. Per-epoch validation metrics for each loss component individually, alongside aggregate loss, provide the information needed to diagnose whether a failure is attributable to a specific objective, a learning-rate issue, or a data problem. These distinctions are impossible to make from checkpoint outputs alone.

Keep the feature representation psychoacoustically motivated and interpretable. The eight-feature design used in this project was deliberately constrained and explainable. Every feature corresponds to a meaningful mastering-relevant quantity, whether loudness, peak, dynamic range, or source-specific masking, rather than to a latent embedding optimised purely for prediction. This design choice paid off in two ways: it kept the system trainable within the available data budget, and it made failures interpretable. When mid-range masking did not improve on average, that could be diagnosed directly from the evaluation metrics rather than from inspecting opaque feature activations. For future systems operating in similarly constrained data environments, compact psychoacoustically grounded features are a more tractable starting point than learned embeddings.

Address the oracle-to-inference gap explicitly. Training on ground-truth stems from MUSDB18-HQ while deploying on Demucs-estimated stems introduces a train-inference mismatch that degrades real-world performance in ways that the training evaluation does not capture. A straightforward mitigation is to include Demucs-estimated stems during training, either by pre-processing the training set through the separator or by mixing oracle and estimated stems in a randomised data augmentation scheme. This would expose the policy to the artifact patterns it will encounter at inference time and improve generalisation to real-world inputs.

5.3 Recommendations for practitioners

Use AI mastering tools as a starting point, not an endpoint. The prototype and the professional perspectives gathered in Chapter 4 consistently support the same practical position: automated mastering is most useful when it is treated as a rapid first-pass tool rather than as a final-quality

stage. In the context of this prototype, the system produces a loudness-normalised output with controlled peak behaviour in seconds, which is genuinely useful for checking whether a mix is in approximately the right loudness and spectral territory before committing to more detailed work. Abbey Road Studios engineers describe the same principle for commercial AI mastering tools: the value lies in iteration speed and in providing a reference point, not in replacing the judgment of an experienced engineer [17].

Understand what AI mastering systems can and cannot measure. Current mastering systems, including the prototype developed here, optimise against objective signal metrics such as integrated loudness, true peak, spectral balance, and masking-related energy ratios. They do not have direct access to the perceptual qualities that practitioners most care about: the perceived presence and clarity of the 2 kHz to 10 kHz range, the subjective sense of punch and transient definition, the emotional tone of a mix, or the way a master will translate across different listening environments. Knowing this boundary is essential for using AI tools productively. A practitioner who evaluates AI mastering output with these perceptual criteria in mind will make better use of the tool than one who treats a compliant LUFS number as a proxy for professional quality.

Account for genre-specific loudness expectations explicitly. As noted in Chapter 4, the acceptable loudness range for a finished master is not universal. In techno and EDM contexts, where competitive loudness between -6 and -10 LUFS is standard before streaming platform normalisation, a system targeting a single fixed LUFS value may consistently produce output that feels either too quiet or too heavily compressed for the intended listening context [19]. Practitioners should treat the LUFS target as a creative and context-specific parameter rather than a fixed standard, and evaluate AI-mastered output against genre reference tracks before making distribution decisions.

Apply critical listening in the 2 kHz to 10 kHz region specifically. The consultation carried out for this thesis identified the perception of overtones and frequency content in the 2 kHz to 10 kHz range as the clearest perceptual gap between AI-mastered and professionally mastered output in the techno context. This region governs the perceived brightness, detail, and definition of a master. AI systems that optimise primarily for loudness and spectral balance at a coarser scale may inadvertently compress or dull this range without triggering any objective metric alert. A targeted listening check, specifically comparing the presence and clarity of high-frequency content against a reference track, is the most direct way to catch this category of problem.

5.4 A step-by-step path for extending this prototype

For anyone who wants to continue from the current prototype toward a more robust and perceptually credible system, the following sequence of development steps reflects the priorities that emerged most clearly from the project outcomes, the diagnosis and redesign process, and the external perspectives gathered in Chapter 4.

Step 1: Expand the training and evaluation dataset. Replace or augment the MUSDB18-HQ corpus with a wider and more genre-diverse collection. Prioritise material that includes techno, electronic, and club-oriented genres if those are the target use cases, since MUSDB18-HQ is heavily weighted toward rock and pop. Even a modest genre-diverse expansion will substantially improve generalisation and reduce the risk of the policy overfitting to the spectral and dynamic characteristics of a single corpus.

Step 2: Add genre conditioning to the LUFS target. Replace the fixed loudness target with a genre-conditioned target value. A simple implementation uses a learned genre embedding, derived from a small set of genre labels or from a pretrained audio classifier, to modulate the loudness-error feature at the input to the policy. This allows the system to learn that the appropriate loudness correction for a techno track differs from that for an acoustic recording, without requiring changes to the mastering chain itself.

Step 3: Add a dedicated high-frequency perceptual feature. Extend the eight-feature schema with a measure that captures the energy balance or dynamic behaviour in the 2 kHz to 10 kHz range. A suitable candidate is the ratio of spectral energy above 2 kHz relative to total energy, computed from the full mix or from a vocals-and-other stem combination. This provides the policy with a direct signal for the perceptual gap identified in both the prototype evaluation and the expert consultation.

Step 4: Implement paired parameter interaction testing. Following the suggestion from the expert consultation, redesign the compressor and limiter parameter prediction to include an interaction-aware evaluation stage. A practical implementation would run several forward passes through the mastering chain with different parameter combinations in the compression-limiting stage, score each combination against the loss objectives, and select the optimal combination rather than predicting all parameters in a single forward pass. This adds inference-time cost but substantially improves the quality of the compression-limiting interaction.

Step 5: Close the oracle-to-inference gap. As recommended in §5.2, include Demucs-estimated stems in the training pipeline so that the policy learns from realistic rather than idealised stem inputs. This single change is likely to produce a measurable improvement in real-world performance without requiring any architectural modification.

Step 6: Add a formal perceptual evaluation protocol. Introduce a structured listening test as a standard part of the evaluation pipeline. A minimal implementation could use a small ABX or MUSHRA protocol comparing the prototype output against a commercial AI mastering reference and an unmastered input, administered to a panel of listeners with relevant domain experience. This provides the one category of evidence that the current prototype cannot generate from objective metrics alone: information about whether the output is perceptually credible to a human listener.

Figure 15: Step-by-step prototype extension roadmap. Sources: [4]–[8], [14]–[16].

6 Conclusions

This thesis set out to explore a specific and technically precise question: how can a stem-aware neural architecture be used to predict and apply mastering parameters for the autonomous production of audio masters that meet professional streaming standards and approach the perceived quality of human-made masters? The research was framed around that question from the start, and the conclusion must answer it directly.

The answer, supported by everything presented in Chapters 2 through 5, is the following. A stem-aware neural architecture of the type developed in this project can be designed, implemented, and trained to produce mastered audio that meets basic structural streaming requirements and shows measurable improvement in loudness targeting after iterative redesign. In that technical and objective sense, the architecture is viable and the workflow is coherent. However, the same evidence also shows clearly that the system does not yet consistently approach the perceived quality of human-made masters. The gap between metric compliance and perceptual quality remains the central unresolved problem, and it is a gap that the current prototype cannot close through objective optimisation alone.

This is an honest and defensible conclusion, and it is one that aligns with both the research literature reviewed in Chapter 2 and the external perspectives gathered in Chapter 4.

6.1 What the research demonstrated

The practical research component produced a complete end-to-end prototype in which a pretrained source-separation model provides stem-level information to a learned neural policy, which in turn predicts mastering parameters for a fixed differentiable DSP chain applied to the original stereo mixture. Every stage of this pipeline was implemented, trained, and evaluated, which makes the project a genuine proof of concept rather than a theoretical proposal.

The iterative development process, documented in Chapter 3, is one of the strongest contributions of this research. The transition from Phase 5 to Phase 6 illustrates in concrete terms what goes wrong when the action space is too constrained, what a principled diagnosis of that failure looks like, and what a targeted architectural response can achieve. The reduction in mean LUFS error from 24.87 LU to 2.68 LU following the redesign is a substantial quantitative improvement, and it was achieved not through arbitrary experimentation but through a specific and reasoned set of changes to the policy architecture, the gain structure, and the training curriculum. That iterative

story, rather than the final model alone, is what the project contributes to the field of intelligent mastering research.

At the same time, the evaluation results are honest about what was not achieved. Only approximately 20 percent of evaluated tracks reached within 1 LU of the loudness target, mid-range masking did not improve on average, and the system was not tested beyond the MUSDB18-HQ corpus. True-peak compliance was maintained at 100 percent, but this is a structural property of the limiter design rather than a learned mastering skill. The prototype meets the minimum definition of a functional stem-aware mastering system, but it does not meet the higher standard of consistent professional-quality output.

6.2 What the external reflection revealed

The reflection in Chapter 4 confirmed and contextualised these findings through expert consultation and professional community discourse. The audio engineering perspective gathered through consultation identified the stem-aware design as a genuine advance over stereo-only automated mastering tools, specifically because source-level analysis gives the system access to masking and balance information that cannot be inferred from a collapsed stereo signal. That recognition validates the core design premise of the thesis. At the same time, the same consultation identified the perceptual gap with precision: the frequency region between 2 kHz and 10 kHz, where the human ear perceives overtone clarity, transient definition, and the sense of detail in a master, is not adequately represented in the current feature schema, and this limitation produces audible differences that no objective metric flags.

Professional sources consulted for Chapter 4 reinforce the same conclusion from a broader perspective. The consensus across mastering engineers and practitioner communities is not that AI mastering systems are without value, but that the quality judgment required for professional mastering depends on perceptual sensitivity, artistic interpretation, and contextual understanding that current systems cannot replicate. The prototype developed in this thesis sits within those constraints: it demonstrates the technical feasibility of a stem-informed approach, but it does not provide evidence that the approach resolves the deeper perceptual problem.

6.3 What the recommendations suggest for the field

Chapter 5 translates the project's outcomes into concrete directions for future development. The most important of these concern the action space design, genre-conditioned loudness targeting, the

addition of perceptual features for the high-frequency region, the closing of the oracle-to-inference gap, and the introduction of formal perceptual evaluation as a standard part of the assessment pipeline. Together, these recommendations trace a credible path from the current prototype toward a system that might eventually be capable of producing output that is not only metrically compliant but perceptually defensible.

The step-by-step development path proposed in §5.4 is not speculative. Each step is grounded in a specific identified weakness of the current system and is technically tractable within the constraints of a research project. The most consequential single step is probably the introduction of formal perceptual evaluation, because without it no future iteration of the system can claim to have addressed the perceptual equivalence question with real evidence rather than with objective proxies.

6.4 Where the field is heading

The broader research context reviewed in Chapter 2 makes clear that intelligent mastering and automatic mixing are active and growing research areas, with increasing interest in reference-conditioned systems, contrastive learning approaches, and multitrack-aware production models. The contribution of this thesis sits at an early point on that trajectory: it shows that a stem-aware analytical front end can be integrated with a self-supervised mastering policy in a way that is both trainable and evaluable, and that iterative redesign based on concrete failure analysis can produce substantial improvements within a constrained development window.

For researchers who continue in this direction, the most productive next question is not whether stem-aware analysis is useful, but how much of the perceptual quality gap can be closed through better feature design, richer training data, and more rigorous perceptual evaluation. That question is answerable, and the framework built in this project provides a foundation for answering it systematically. As AI-assisted audio production continues to develop, the boundary between technically compliant mastering and perceptually convincing mastering will remain the central challenge. Research that takes both dimensions seriously, rather than optimising metrics without testing perception, is what is needed to move that boundary forward.

7 References

[1] M. Chen, "Intelligent audio mastering," School of Creative Media, City University of Hong Kong, 2020.

[2] J. D. Reiss, "Intelligent systems for mixing multichannel audio," in *Proc. 17th Int. Conf. Digital Signal Processing (DSP)*, 2011, pp. 1–6.

[3] D. Moffat and M. B. Sandler, "Approaches in intelligent music production," *Arts*, vol. 8, no. 4, p. 125, 2019.

[4] C. J. Steinmetz and J. D. Reiss, "Automatic multitrack mixing with a differentiable mixing console of black-box audio effects," in *Proc. 22nd Int. Soc. Music Information Retrieval Conf. (ISMIR)*, 2021.

[5] B. De Man, M. Leonard, R. Ward, and J. D. Reiss, "An analysis and evaluation of audio features for multitrack music mixdown," in *Proc. AES 136th Conv.*, 2014.

[6] M. A. Martínez Ramírez, E. Benetos, and J. D. Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences*, vol. 10, no. 2, p. 638, 2020.

[7] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. ISMIR Workshop on Music Source Separation*, 2022. [Online]. Available: https://arxiv.org/abs/2111.03600

[8] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," arXiv:1909.01174, 2019.

[9] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Information Retrieval Conf. (ISMIR)*, 2018, pp. 334–340.

[10] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix — a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[11] N. Takahashi and Y. Mitsufuji, "D3Net: Densely connected multidilated DenseNet for music source separation," arXiv:2010.01733, 2020.

[12] K. Choi, C. Kim, M. Chung, and J.-W. Ha, "LaSAFT: Latent source attentive frequency transformation for conditioned source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 171–175.

[13] R. Zhu and J. Kim, "ResUNetDecouple+: Musical source separation with residual U-Nets," in *Proc. Sound and Music Computing Conf. (SMC)*, 2021.

[14] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Bingley, U.K.: Emerald Group, 2013.

[15] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin, Germany: Springer, 2007.

[16] J. D. Reiss and A. McPherson, *Audio Effects: Theory, Implementation and Application*. Boca Raton, FL: CRC Press, 2014.

[17] Abbey Road Studios, "Perspectives on AI in professional audio production," professional consultation, Apr. 2025.

[18] Kevork Mastering, "Professional mastering workflow and AI limitations," professional consultation, Apr. 2025.

[19] Gearspace Audio Community, "AI mastering tools: practical evaluation and limitations," online discussion thread, 2025. [Online]. Available: https://gearspace.com

[20] Reddit r/audioengineering, "Community perspectives on automatic mastering and professional equivalence," online discussion thread, 2025. [Online]. Available: https://www.reddit.com/r/audioengineering/

[21] European Broadcasting Union, "EBU R 128: Loudness normalisation and permitted maximum level of audio signals," EBU Tech. Recommendation R 128, Geneva, Switzerland, Aug. 2020.

8 Annexes

8.1 Guest Speaker Reports

8.1.1 Guest Speaker Report: The ABC of Academic Writing

Speaker: Nico Van den Abeele **Role:** Lecturer in English Language and Academic Communication, HOWEST **Date:** Academic year 2025–2026, first semester **Topic:** The ABC of Academic Writing **Programme:** Creative Technologies and Artificial Intelligence (CTAI), HOWEST

Introduction

This report documents a guest lecture delivered by Nico Van den Abeele within the context of the CTAI bachelor programme at HOWEST. The session was designed to introduce students to the core principles of formal academic writing, with the explicit goal of preparing them for the production of their bachelor thesis. The lecture is particularly relevant to CTAI students, who typically arrive at the writing stage with strong technical backgrounds but limited exposure to the conventions of academic prose. The session addressed this gap by providing both a conceptual framework and a practical set of rules governing formal written communication in a university context.

Content

The lecture was organised around two complementary frameworks: the ABC model and the COFE model.

The ABC model describes three overarching qualities that all academic writing must demonstrate. Accuracy refers to the obligation to be factually precise and to represent claims, sources, and findings truthfully and without distortion. In practice, this means that every assertion must be supportable by evidence, that uncertainty must be explicitly marked using hedging language, and that sources must be cited correctly so that the reader can verify claims independently. Brevity refers to the economy of language: academic writing should not include padding, repetition, or tangential elaboration that does not serve the argument. Every sentence should carry a specific communicative function, and writers should avoid the temptation to inflate word counts through stylistic excess. Clarity refers to the intelligibility of expression: sentence structure should be unambiguous, logical relationships between ideas should be made explicit through appropriate

connective language, and the argument should progress in a way that is transparent to the reader, not merely internally coherent to the writer.

The COFE model, an acronym for Characteristics Of Formal English, operationalises these principles at the level of lexical and grammatical choice. The lecturer identified six specific rules. First, contractions must be eliminated; forms such as "don't", "it's", and "they're" are exclusively informal and must be replaced by their full equivalents. Second, numbers below ten must be written in full ("three" rather than "3"), while larger numbers may be expressed numerically. Third, vocabulary should draw on the formal register of academic English, which historically derives its formal lexis from Latin, Ancient Greek, and French roots. The lecturer drew explicit contrasts between Germanic-root informal verbs and their Latinate formal equivalents: "to start" becomes "to commence", "to end" becomes "to conclude", and "to get" becomes "to receive" or "to obtain". Fourth, formal grammar requires the explicit inclusion of relative pronouns and conjunctions that are typically omitted in spoken or informal written English. So-called zero-relative constructions such as "the result we obtained" must be expanded to "the result that was obtained". Fifth, coordinating conjunctions such as "and", "but", and "so" must not be used at the beginning of a sentence; their formal equivalents are discourse-level connectives such as "furthermore", "moreover", "in addition", "however", and "therefore". Sixth, first- and second-person pronoun use, specifically "I", "we", and "you", must be eliminated in favour of impersonal or passive constructions, so that academic writing maintains an objective, authoritative register rather than a personal or conversational one.

Conclusion and Critical Reflection

The lecture provided a precise and immediately applicable set of conventions for producing formal academic text. The distinction between informal and formal register is not merely a matter of style but of epistemic positioning: a thesis written in formal academic English signals that the author is capable of subordinating personal voice to the norms of scholarly discourse, and that the claims made are being presented as verifiable propositions rather than personal opinions.

For this thesis, which addresses the automation of audio mastering through a machine learning prototype, the lecture's relevance was direct and practical. Technical writing in software and audio engineering frequently defaults to procedural or conversational registers, and the tendency to describe system behaviour in informal terms is a genuine risk when writing about computational systems. The COFE framework provided a concrete checklist for identifying and correcting these

tendencies during revision, particularly for sections such as the methodology and results chapters where precise and impersonal language is essential.

One point of critical reflection concerns the tension between brevity and technical completeness. In a thesis documenting a complex prototype, the obligation to be brief can conflict with the obligation to be accurate: describing a neural architecture, a signal processing chain, or an evaluation procedure precisely often requires extended and detailed prose. The lecture acknowledged this tension implicitly through the principle that brevity does not mean compression at the expense of clarity or accuracy, but rather the elimination of non-functional language. This clarification was particularly useful in calibrating how much detail is appropriate for technical sections.

References

Van den Abeele, N. (2025–2026). *The ABC of Academic Writing* [Lecture handout]. HOWEST, Bruges.

8.1.2 Guest Speaker Report: Rewording is Rewarding

Speaker: Nico Van den Abeele **Role:** Lecturer in English Language and Academic Communication, HOWEST **Date:** Academic year 2025–2026, first semester **Topic:** Rewording is Rewarding — Paraphrasing for Academic Writing **Programme:** Creative Technologies and Artificial Intelligence (CTAI), HOWEST

Introduction

This report documents a second guest lecture by Nico Van den Abeele, delivered as a continuation of the academic writing module within the CTAI programme. Where the first session established the theoretical framework of formal register, this session focused on its practical application: the skill of paraphrasing and rewording source material in an academically appropriate way. The ability to reword is treated in academic writing not as a mechanical substitution exercise but as a demonstration of comprehension and ownership of ideas. The session addressed two distinct levels at which rewording operates, namely the sentence level and the discourse level, and provided structured exercises to develop competence at both.

Content

The session opened with a discussion of why rewording is necessary in academic writing. The primary reason is the avoidance of plagiarism: when a writer reproduces source material without

transformation, intellectual ownership remains with the original author and the thesis fails to demonstrate that the writer has processed and assimilated the ideas being cited. A secondary reason is register alignment: source material, particularly from informal sources such as textbooks, online articles, or technical documentation, may not conform to the formal register expected in a thesis, and rewording is the mechanism by which ideas from any source can be integrated into a coherent academic text.

At the sentence level, the lecture demonstrated a systematic approach to rewording that operates across multiple dimensions simultaneously. Vocabulary substitution was addressed first: informal or colloquial terms must be replaced by their formal equivalents, and the word list from the previous session was reactivated as a reference. Voice transformation was addressed next: active constructions in informal source text are frequently converted to passive constructions in formal academic prose, both to eliminate personal pronouns and to foreground the action or result rather than the agent. The elimination of contractions and the explicit inclusion of relative pronouns and conjunctions were revisited as obligatory transformations whenever informal source text is incorporated.

At the discourse level, the lecture moved beyond individual sentences to consider how ideas are connected and sequenced across paragraphs. Discourse-level rewording involves not merely changing the words used to express an idea but restructuring the logical relationship between ideas. The connector vocabulary introduced in the previous session was applied at this level as the primary tool for making logical relationships explicit in the flow of an argument. The lecturer demonstrated that informal source texts tend to juxtapose ideas without making their relationship clear, and that the rewording task at discourse level consists precisely of identifying and articulating these implicit relationships.

The session concluded with a set of exercises in which students were asked to transform informal passages into formal academic equivalents, applying both sentence-level and discourse-level techniques. Several of these exercises involved complete paragraph rewrites in which voice, vocabulary, pronoun use, and connective structure were all revised simultaneously.

Conclusion and Critical Reflection

This session was directly applicable to the thesis-writing process in a way that was not fully anticipated before attending. The specific challenge of integrating sources from the audio engineering domain, including technical articles, engineering blog posts, and community forum discussions, into formal academic prose requires exactly the skills demonstrated in this session. Sources such as practitioner interviews, manufacturer documentation, and online professional discourse do not arrive in a register appropriate for an academic thesis, and the paraphrasing tools provided in this lecture are the mechanism for making such material usable.

A critical observation is that discourse-level rewording is considerably more demanding than sentence-level rewording, and it is the aspect most likely to be inadequately handled in a first draft. Sentence-level substitutions can be applied mechanically; discourse-level restructuring requires a clear understanding of the argumentative purpose of each paragraph and of how the ideas being integrated relate to the thesis's central claims. For this thesis, this was most practically relevant in the reflection chapter, where practitioner views on AI mastering, drawn from professional articles and consultations, needed to be integrated into an argument about the prototype's capabilities and limitations without reproducing the framing or structure of the original sources.

References

Van den Abeele, N. (2025–2026). *Rewording is Rewarding* [Lecture handout]. HOWEST, Bruges.

8.2 Expert Interview Summaries

The following Q and A summaries document the expert consultations conducted in support of Chapter 4 (Reflection). Each interview was prepared in advance, conducted personally by the researcher, and summarised below.

8.2.1 Expert Interview: Matei Craciunescu

Role: Professional drummer (metal band); techno producer and DJ; background in sound and audio engineering.

Format: Semi-structured personal consultation conducted by the researcher .

Context: Craciunescu was presented with a description of the prototype's architecture and evaluation results, and subsequently listened to a selection of outputs processed from unmastered techno material.

Q: When you hear that this prototype analyses individual instrument stems before making any mastering decisions, does that design approach make sense to you from a professional standpoint?

A: Yes, immediately. The fundamental problem in mastering a dense mix is understanding what is fighting what. When you only have the stereo mixdown, you are estimating what is happening in the low end, in the mid range, between the kick and the bass. If you can measure those relationships directly from the stems, you have information that a stereo analyser simply cannot give you. Achieving what could be described as a sound equilibrium, a condition in which no element is competing destructively with another in the final stereo image, is a significant part of what mastering is trying to accomplish, and the stem-aware approach directly addresses that challenge.

Q: The system relies on a pretrained source-separation model to extract those stems. Is there a concern about what the separation model might introduce into the mastered output?

A: That was the first question that came to mind. Source separation introduces artifacts: spectral bleed between stems, phase inconsistencies, transient smearing. If those artifacts were being applied to the audio directly, that would be a serious problem. However, if the stems are only used for measurement, for extracting the feature values that drive the mastering decisions, and the processing is then applied back to the original, unprocessed stereo mixture, then the artifacts are

analytically contained. They may affect the precision of the feature measurements to some extent, but they are not entering the output signal. That design choice is sound.

Q: The prototype also constrains all output parameters within a bounded normalised range. Does that change how you assess the risk of the system producing a destructive mastering result?

A: It reduces the risk substantially. If every parameter decision is bounded, then even if the feature extraction contains noise from imperfect separation, no individual parameter can move far enough to produce a catastrophic or unusable output. The worst case becomes a suboptimal result rather than a broken one. For a prototype operating without human oversight, that is the correct approach.

Q: In your own practice, how do you evaluate the interaction between interdependent processing stages, such as a compressor feeding into a limiter?

A: You always evaluate them together, not in isolation. The way compression shapes the transients determines exactly what the limiter receives at its input. The standard approach is to listen to the full chain end to end, ideally comparing results at the limits of the parameter range, so that you understand the combined effect before committing to settings. Evaluating each processor separately tells you very little about what the output will actually sound like.

Q: After listening to the prototype's output on a selection of unmastered techno material, what did you observe?

A: The loudness targeting was noticeably more accurate than expected for an automated system. That was the clearest positive result. What was lacking was in the upper-mid range, roughly between two and ten kilohertz. That is the frequency region where you perceive the definition of a track: the intelligibility of percussion, the clarity of melodic content, the sense of detail and focus that a professional master delivers. The prototype's output in that range felt flat or under-defined. It was not unusable, but it was perceptibly below what a professional master would achieve in that area.

Q: In the techno and broader EDM context specifically, what loudness targets are typical, and how does that relate to what the prototype is targeting?

A: In that genre, masters are typically expected to sit somewhere between minus six and minus ten LUFS before streaming platform normalisation is applied. There is still an active loudness culture in that space: producers and engineers are acutely aware of what the loudness level of a release signals

about its production value. A universal target LUFS value, without any genre-conditioning, is a reasonable research baseline, but it does not reflect how mastering decisions are actually made in practice for that repertoire.

Q: How do you assess where automated mastering currently stands relative to professional practice?

A: It is not a first choice, and in its current form it does not substitute for a professional mastering session. The most fundamental missing element is perceptual interpretation: the ability to hear a track, understand what it is attempting to achieve, understand the genre expectations it is operating within, and make processing decisions that serve the artistic intent rather than merely satisfying a set of measurable criteria. A system that optimises objective measurements can get close on the technical side, but what separates a professional master is not primarily technical compliance. It is interpretive judgment, and that is not yet accessible to an automated system.

8.2.2 Expert Interview: Ion Bahov

Role: Professional percussionist and teacher at the school of arts in the name of "Alexei Stircea" in Chișinău, Republic of Moldova; background in sound and audio engineering, as well as mastering expert in electronic music.

Format: Semi-structured personal consultation conducted by the researcher

Context: dr. Bahov was presented with a description of the prototype's architecture and evaluation results, and subsequently was asked contextual questions about the industry of electronic music, which resulted in an unexpectedly insightful conversation.

Q: In the EDM and techno context, what is the relationship between loudness targets and perceived quality, and is there a point at which pushing level begins to actively damage the material?

A: It is highly genre-dependent and you cannot reduce it to a single number. People throw around minus five dB RMS as if it is a standard, but that figure on its own tells you almost nothing, because different metering tools will give you different readings for the exact same file, and the perceptual result at any given loudness level is determined far more by the arrangement and the mix than by anything you do at the mastering stage. What I can say with confidence is that there is a threshold beyond which you are actively destroying the material: the kicks and the bass start to sound plastic and lifeless, the transient clarity disappears, the low end loses its depth. In techno specifically I tend to argue for more conservative targets, somewhere in the minus eight to minus ten LUFS range. Dynamic range in that genre is not just a technical property you are trying to

preserve, it is a compositional and expressive device. When you compress that out, you are removing something that was supposed to be there.

Q: How do you actually achieve high loudness levels in EDM without destroying the low end, and is that something that can be done at the mastering stage alone?

A: It cannot be done at the mastering stage alone, and anyone who tells you otherwise is either lying or working with exceptional mixes. As a mastering engineer you inherit whatever headroom the mix gives you. If the arrangement is dense, if the sub and the kick are fighting each other, if there is no sidechain managing that relationship, you will hit a ceiling very quickly and everything below it will start to distort before you reach a commercially competitive level. The techniques that actually work are mostly production and mixing decisions: cleaning the sub-bass region to recover headroom, managing the kick-to-bass relationship with sidechain so they are not both claiming the same space at the same time, using parallel upward compression rather than just slamming a limiter, distributing soft saturation and clipping across multiple stages in small amounts rather than doing it all in one hit. And then there is stem mastering, which is something professionals in this space have been doing for years: you separate the kick from the full mix and process each element to an appropriate level independently before recombining. That approach exists precisely because full-mix limiting has a hard ceiling in bass-heavy material, and stems let you get around it.

Q: How much do you trust metering when making loudness decisions, and how should an engineer relate to the numbers versus what they are actually hearing?

A: Metering is a reference and a sanity check. It is not a decision-making tool. The problem is that standard RMS, AES-17 RMS, and integrated LUFS can all give you readings that differ by two to three dB on the same file, so if you are comparing your output to a commercial reference using a different meter than whoever mastered that reference, the numbers are simply not comparable. And even within a single consistent metering system, the reading does not capture everything that matters: whether the kick is sine-based or transient-heavy will shift your RMS reading independently of perceived loudness, the duration and sustain of your bass content matters, the overall frequency balance of the mix matters. The engineers I respect most in this space will look at the meter at the end to check they are in a reasonable range, but the question they are actually answering throughout the session is whether it sounds punchy, defined, and appropriate for the genre. That is not a number. That is a judgment call.

Q: Does heavy limiting at the mastering stage affect how a track performs in a club or DJ context, and does the loudness war work differently in that setting compared to streaming or radio?

A: There is a genuine disagreement on this and I do not think it is fully resolved. My own position is that in a properly configured system, loudness does not help you and often hurts you: a dynamic track at minus ten LUFS gives the DJ more control on the mixer and gives the system more headroom to work with, and on a good rig that translates into a bigger, more physical sound. But I also have to be honest that many club systems are not well configured. Limiters on venue systems are frequently calibrated badly, gain staging is often a mess, and in that environment a louder record can have a practical advantage simply because it survives the chain better. So the answer depends on context: in the techno and minimal worlds, where the music is primarily played on systems that are actually taken seriously, dynamic range is a real advantage. In mainstream commercial EDM

and festival contexts, the expectation of competitive loudness is still real and largely enforced by what DJs are using as references when they are building their sets.

Q: Is there a viable role for an automated stem-aware mastering system in this context, and what would it need to get right to be genuinely useful?

A: The stem-aware premise is sound and it maps onto something professionals already do. The fact that stem mastering exists as a technique in this space is not a coincidence: it is a direct response to the limitations of full-mix processing on bass-heavy material. So an automated system built around that idea is starting from a real insight. What it would need to get right, at minimum, is the kick-to-sub relationship: that is the primary technical constraint on everything else in this genre, and if the system cannot account for it explicitly, it will hit the same ceiling as full-mix limiting. It would also need genre-conditioned loudness targeting, because a fixed LUFS value is not how mastering decisions are actually made in practice, and the acceptable range varies significantly between techno, trance, and commercial house. And it would need some way of evaluating the headroom the mix provides before deciding how hard to push, because the mastering outcome is fundamentally constrained by what comes in. The ceiling, though, is always going to be the interpretive layer: genre awareness, system context, understanding what the track is trying to do. A system that optimises signal measurements does not have access to that, and that is where the gap between a technically compliant master and a professionally satisfying one lives.

8.3 Evaluation Results

The following table summarises the quantitative evaluation results across the two primary development phases of the prototype. Phase 5 represents the initial training configuration; Phase 6 represents the redesigned architecture following the curriculum and capacity changes described in Chapter 3.

Table 1: Evaluation results — Phase 5 vs Phase 6 (MUSDB18-HQ test set)

Metric	Phase 5	Phase 6
Evaluation tracks	50	20
LUFS MAE (mean absolute error)	24.87 LU	2.68 LU
Tracks within plus or minus 1 LU of target	Not measured	approximately 20%
True-peak compliance (at or below -1 dBTP)	100%	100%
Metric	**Phase 5**	**Phase 6**
Mid-range masking improvement (average)	No reduction	No consistent improvement

MLP architecture	3 layers x 128 units	4 layers x 256 units
Training objective	Uniform loss weighting	Curriculum: loudness-first weighting
Separation model	htdemucs (pretrained)	htdemucs (pretrained, unmodified)
DSP chain	FFT EQ, RMS compressor, tanh limiter, global gain	FFT EQ, RMS compressor, tanh limiter, global gain

Notes on measurement methodology

True-peak compliance is structurally guaranteed by the tanh limiter, which bounds the signal before gain scaling is applied. This guarantee holds at the output of the DSP chain but was not independently verified with oversampled true-peak measurement tooling at inference time.

The LUFS measurement used during training and evaluation is an internal PyTorch implementation following the BS.1770 weighting specification. This implementation is internally consistent but was not cross-validated against external reference tools such as pyloudnorm or ffmpeg. Results should be interpreted within this measurement context.

The oracle-to-inference gap, that is, the discrepancy between training on ground-truth stems and deploying on Demucs-estimated stems, was identified as a likely contributor to residual LUFS error at inference. This gap was not independently quantified but is documented as a structural limitation in Chapter 3.

8.4 Installation Manual

System requirements

Operating system: Linux (Ubuntu 20.04 or later recommended) or macOS 12 or later. Python 3.9 or later is required. A CUDA-capable GPU is recommended for acceptable inference speed; CPU-only

execution is supported but significantly slower. A minimum of 8 GB RAM is required; 16 GB is recommended for processing full-length tracks.

Dependencies

The prototype depends on the following Python packages. All dependencies can be installed using the provided requirements file:

```
pip install -r requirements.txt
```

Core dependencies include: torch (PyTorch, version 2.0 or later), torchaudio, demucs, numpy, scipy, librosa, and soundfile.

If demucs is not available via pip in the current environment, it can be installed directly from the official repository:

```
pip install git+https://github.com/facebookresearch/demucs
```

Installation steps

Step 1: Clone or download the prototype repository to the local machine.

Step 2: Create a Python virtual environment:

```
python -m venv venv && source venv/bin/activate
```

Step 3: Install all dependencies:

```
pip install -r requirements.txt
```

Step 4: Verify that PyTorch detects the GPU (if applicable):

```
python -c "import torch; print(torch.cuda.is_available())"
```

Step 5: On first run, demucs will download the pretrained htdemucs model weights automatically. An internet connection is required for this step.

Step 6: Place the provided mastering policy checkpoint file (mastering_policy.pt) in the designated checkpoint directory as specified in the repository's configuration file.

Verifying the installation

Run the following command to verify that the installation is complete and the model loads correctly:

python verify_install.py

A successful installation will print: `Installation verified. Model loaded successfully.`

8.5 User Manual

Overview

The prototype accepts an unmastered stereo audio file as input, processes it through the stem-aware mastering pipeline, and outputs a mastered stereo audio file. The processing pipeline performs the following steps automatically: stem separation using htdemucs, feature extraction from the separated stems, mastering parameter prediction using the trained MLP policy, and application of the differentiable DSP chain to the original stereo mixture. The DSP chain consists of an FFT-based EQ, an RMS compressor, a tanh limiter, and a global gain stage.

Input format

File format: WAV (recommended) or FLAC. Sample rate: 44.1 kHz; other sample rates are resampled internally. Bit depth: 24-bit or 32-bit float recommended; 16-bit is supported. Channel format: stereo (two-channel interleaved). The system does not accept mono files or files with more than two channels.

Running inference

To process a single audio file, run the following command:

python master.py --input path/to/your/mixdown.wav --output path/to/output_mastered.wav

Optional arguments:

--target_lufs: specify a target integrated loudness in LUFS (default: -14.0)

--device: specify cuda or cpu (default: cuda if available, otherwise cpu)

Example with a custom loudness target:

python master.py --input mixdown.wav --output mastered.wav --target_lufs -10.0

Output

The output file is a stereo WAV file at the same sample rate as the input, written to the path specified by the --output argument. Integrated loudness and true-peak values of the output are printed to the console on completion.

Known limitations

The prototype was trained exclusively on material from the MUSDB18-HQ dataset. Performance on genres and production styles that fall significantly outside this dataset's distribution may be reduced.

Loudness targeting accuracy is highest when the input material is within approximately 10 LU of the target LUFS value. Inputs with very low integrated loudness (below approximately -30 LUFS) may not converge to the target within the parameter bounds of the current policy.

The mastering chain does not apply stereo widening, mid/side processing, or frequency-selective limiting. These processing stages are common in professional mastering and their absence limits the prototype's perceptual output quality relative to commercial practice.

The true-peak values printed to the console are measured using the internal BS.1770-compliant PyTorch implementation. For release-critical work, it is recommended to independently verify true-peak compliance using a reference metering tool.