Timbre Transfer of the Human Singing Voice for Effective Automatic Music Transcription

Bonface Martin Oywa, Student ID 16520516.

University of East London, UNICAF.

Module Code UEL-CN-7000-91461.

3rd December 2025.

Acknowledgements

I would like to thank my supervisor, Queen Manthimba Phiri, for her ongoing guidance, patience, and encouragement during this project. Her experience and careful feedback were invaluable.

I am also sincerely grateful to UNICAF for awarding me the scholarship that made these studies possible. Their support has been crucial in enabling me to pursue this degree and complete this dissertation. My thanks extend to the faculty, staff, and fellow students of the programme for creating an engaging and supportive learning environment, and for the many discussions that helped clarify and challenge my thinking along the way.

I am especially grateful to my family and friends for their everlasting love, support, and understanding. Their faith in me has supported and inspired me throughout my journey.

This work relies heavily on the VocalSet (Annotated) dataset and on open-source tools and models, including DDSP, MT3, Basic Pitch, and the wider Python and Music Information Retrieval ecosystems. I would like to thank the creators and maintainers of these resources for making their work publicly available and for fostering an open, collaborative research culture.

Finally, I acknowledge the use of the AI-based tool https://chatgpt.com/ during the brainstorming, coding support, and draft-refinement stages of this project. This use took place during the preparation of the dissertation (July - December 2025). The tool was used to support ideation, clarify wording, and check the coherence of arguments; all methodological choices, analyses, and the final written text were made by me in line with the university's academic integrity guidelines. The prompts used included, among others:
- "Validate contents in Literature Review, Methodology, and Results and Discussion sections using this marking criteria rubric."
- "What features in the VocalSet (Annotated) dataset other than Start/End Times and MIDI code can be used to generate MIDI files?"

Table of Contents

List of Tables

List of Figures

List of Nomenclature

1. F0 (Fundamental frequency) – The fundamental frequency of a (quasi-)periodic signal; the inverse of the signal's period.
2. Onset – The time instant when a note or event starts.
3. Offset – The time instant when a note or event ends.
4. Pitch – The perceived musical "height" of a sound (e.g., C4, A4 = 440 Hz); the perceptual correlate of F0.
5. Velocity (MIDI) – A value from 1–127 attached to MIDI note-on (and sometimes note-off) events, roughly corresponding to how hard a note is played and perceived loudness/attack (not tempo).
6. Precision – The proportion of predicted items that are actually true.
7. Recall – The proportion of true items in the dataset that were correctly predicted.
8. F-measure / F1 score – The harmonic mean of precision and recall, used as a single summary of detection performance.
9. TER (Transcription Error Rate) - Measure the accuracy of predicted notes against reference or ground truth notes. Lower values equal better transcription.
10. Overlap Ratio – A duration-weighted measure of agreement: the fraction of reference note time that overlaps with correctly predicted note time.
11. Error structure – The pattern and types of errors a model makes during prediction (e.g., false positives vs. false negatives, onset vs. offset errors, pitch stability issues), describing how it fails rather than just how often.
12. AMT - Automatic Music Transcription.
13. MIR - Music Information Retrieval.

1. Introduction

1.0 Background Statement

Automatic Music Transcription (AMT) systems have made significant progress in recent years; however, they continue to struggle with accurately transcribing the human singing voice. This limitation arises from three main challenges: first, the complex harmonic structures and variability of human vocals make direct transcription difficult; second, most existing systems are designed for direct audio-to-score transcription without intermediate processing; and third, many models are biased towards instrumental recordings, often being tuned for piano or other instrument-like sounds. To address these issues, this research explores the timbre transfer of the human singing voice into instrument-like representations before transcription. This approach has the potential to simplify the spectral structure of vocals, reduce the reliance on large-scale data collection or model fine-tuning, and ultimately improve the efficiency and accuracy of downstream transcription models.

1.1 Research Statement

This dissertation investigates whether converting human singing voice into instrument-like sounds through timbre transfer can improve the performance of AMT systems, and under what conditions such a strategy is actually beneficial. Starting from the observation that state-of-the-art AMT models are typically optimised for instrument recordings and struggle with the continuous pitch motion, unstable harmonics, and expressive variability of vocals, the project explores the hypothesis that standardised timbres (e.g., trumpet) provide cleaner and more model-friendly harmonic structures than raw human voice. By inserting a timbre transfer stage, implemented with differentiable synthesis tools such as DDSP, before existing AMT models, this work empirically examines the trade-off between timbre fidelity and transcription accuracy, and identifies which target timbres and similarity measures are most predictive of improved or degraded performance.

Beyond methodological interest, the research is motivated by the broader impact that robust singing-voice transcription could have across Music Information Retrieval (MIR) applications. Reliable AMT for vocals enable commercial tools for karaoke scoring, melody search, and content indexing in large music catalogues; support accessible interfaces where users "sing to search" or "sing to compose" instead of relying on instrumental proficiency; and underpin assistive technologies for music education, such as apps that give automatic feedback on intonation, phrasing, and note accuracy to singers. For creators and producers, improved vocal AMT would simplify workflows for generating lead sheets, MIDI guides, and arrangement drafts directly from sung ideas. By focusing on a lightweight, timbre-transfer-based front end that leverages existing open-source models rather than training new systems from scratch, this dissertation aims to deliver insights and tools that are practically deployable in low-resource settings and can be integrated into real-world MIR pipelines.

1.2 Research Questions

1. Effectiveness of Timbre Transfer. To what extent does converting the human singing voice into instrument-like timbres improve the accuracy of downstream AMT compared to direct vocal transcription?
2. Optimal Timbre Representation. Which target timbre (e.g., piano, violin, humming, or synthetic sinusoidal) provides the most reliable transcription performance across different evaluation metrics such as note accuracy, onset/offset detection, and transcription error rate?
3. Trade-off Between Timbre Fidelity and Transcription Accuracy. What relationship exists between the perceptual fidelity of the transferred timbre (i.e., its naturalness and closeness to the target instrument) and the effectiveness of transcription models? Does improving timbre quality always correlate with better transcription results, or is there a point of diminishing returns?

1.3 Research Objectives

1. Design and implement a signal processing or neural pipeline capable of converting a human singing voice into instrument-like timbres using timbre transfer techniques.
2. Integrate the timbre transfer module with existing AMT models to create a complete transcription pipeline.
3. Benchmark the performance of the proposed pipeline against baseline AMT systems that operate directly on raw singing voice, using MIR and Machine Learning evaluation metrics such as note-level precision, recall, F1 score, Onset F1, Onset-Offset F1, and Transcription Error Rate (TER).
4. Visualise and document the resulting transcriptions in symbolic formats (e.g., MIDI, score representation) to qualitatively assess differences between approaches.

1.4 Project Scope

This project focuses on testing the hypothesis that timbre transfer of human vocals into instrument-like sounds can improve the accuracy and efficiency of downstream AMT systems. The scope is limited to the use of existing open-source frameworks and pre-trained models, meaning no deep learning models will be trained from scratch. Instead, the research will adapt and integrate tools such as Differentiable Digital Signal Processing (DDSP) for timbre transfer and transcription models like MT3 and Basic Pitch.

The project will involve the design of a pipeline that converts raw singing voice recordings into standardised instrument timbres before applying transcription. The evaluation benchmarks this pipeline against direct vocal-to-MIDI transcription, with performance measured using metrics such as note accuracy, onset/offset detection, and overall transcription quality.

The scope does not extend to building new AMT architectures, large-scale data collection, or training deep models on custom datasets. Rather, the focus is on leveraging and combining existing open-source resources to experimentally validate the effectiveness of timbre transfer in enhancing AMT for human singing voice. The expected outcome is a comparative analysis demonstrating whether timbre transfer provides measurable benefits and a proof-of-concept system that could inform future AMT research and applications.

1.5 Research Limitations

This study is constrained by several limitations that may influence its outcomes. First, no models will be trained from scratch, and the research will rely entirely on pre-trained open-source frameworks such as DDSP, MT3, and Basic Pitch. As a result, the effectiveness of the proposed pipeline is highly dependent on the performance, design choices, and training data of these existing models. This limitation may restrict the generalizability of the findings across different vocal ranges or recording conditions not well represented in the pre-trained models.

Second, while techniques such as audio source separation could be employed to isolate vocals from polyphonic music, this study will not explore them. Instead, it will primarily operate on monophonic or pre-isolated vocal recordings. This restriction narrows the scope of the research and means that the system may not directly extend to real-world scenarios involving complex mixtures of voice and accompaniment, i.e., polyphonic recordings.

Finally, given the time and resource constraints of this research, the evaluation will be limited to a comparative analysis of timbre-transfer-based vocals against raw vocal transcription using selected benchmark datasets. Broader testing, such as user studies or large-scale evaluations across diverse datasets, will fall outside the scope of this project.

2. Literature Review

2.0 Introduction

AMT is a core process in MIR, intended to transform an audio input into symbolic representations such as a piano roll, MIDI, or staff notation. Typical applications include music education, content-based retrieval, musicological analysis, and creative tools that allow performers and producers to move fluidly between audio and symbolic domains (Schedl et al., 2014). Over the past decade, deep learning has driven substantial progress, particularly for piano and other instrument-specific settings, yet robust, general-purpose AMT remains an open challenge (Benetos et al., 2019).

Within AMT, the human singing voice is a particularly difficult source to transcribe accurately. While monophonic instrumental transcription is often treated as close to "solved," polyphonic music and expressive vocals still expose fundamental limitations in current models and datasets (Benetos et al., 2019). This section reviews three strands of work relevant to this dissertation: (i) AMT in the broader MIR landscape, (ii) specialised approaches to singing-voice AMT, and (iii) timbre transfer and differentiable synthesis methods such as DDSP. The review concludes by identifying the gap that motivates the proposed timbre-transfer-based pipeline for vocal AMT.

2.1 AMT in MIR

The extended definition of AMT is the process of converting an acoustic signal into discrete musical events characterised by pitch, onset and offset times, duration, and, in some cases, dynamics and instrument identity. Classic taxonomies distinguish four main levels of transcription: frame-level (continuous f_0 or pitch-class activation), note-level (discrete note events), stream-level (voice or instrument lines), and notation-level (fully typeset scores). This dissertation focuses on note-level AMT, aligning with modern models that predict note events via onset/offset detection and pitch estimation (Benetos et al., 2019).

Early AMT systems relied heavily on signal processing and probabilistic models such as Non-negative Matrix Factorisation (NMF), Hidden Markov Models (HMMs), and spectral decomposition, often tailored to specific instruments (Sigtia et al., 2015). More recent work has shifted toward deep neural networks, including convolutional and recurrent architectures that operate on spectrograms or constant-Q transforms. A major milestone in this evolution is Onsets and Frames, which jointly predicts onset and frame activations for piano and achieves large gains in note-with-offset F1 on the MAPS dataset by explicitly modelling onsets and conditioning frame predictions on them (Hawthorne et al., 2017).

Beyond piano, there has been a move toward instrument-agnostic and multi-instrument AMT. For example, Bittner et al. propose a lightweight polyphonic system that jointly estimates onsets, multipitch, and note activations while generalising across a wide range of instruments, including vocals; this model underpins Basic Pitch, designed as a fast, open-source AMT system suitable for real-time use (Bittner et al., 2022). In parallel, Multi-Task Multitrack Music Transcription (MT3) frames AMT as a sequence-to-sequence problem: a Transformer encoder–decoder maps spectrograms to tokenised MIDI-like events, jointly learning from multiple datasets and instruments in a unified framework. MT3 demonstrates strong performance on piano (e.g., MAESTRO) and improved results for low-resource instruments by sharing parameters across tasks (Gardner et al., 2021).

Recent surveys highlight two important trends in the field. First, deep learning has narrowed the gap with human performance for specific instrument settings, but generalisation to new timbres, recording conditions, and repertoires remains difficult. Second, most benchmark results are reported on instrument-focused datasets (piano, string ensembles) with relatively controlled timbres, leaving the behaviour of these systems on expressive singing voice underexplored. These limitations motivated our approach to manipulate the input representation via timbre transfer to better fit existing AMT models (Benetos et al., 2019).

2.2 AMT for the Singing Voice

Compared to instruments, the singing voice exhibits greater variability in harmonic structure, dynamics, and articulation, which complicates both low-level pitch estimation and high-level note segmentation. Vocal timbre varies by gender, age, and physiology; even within a single singer, differences in vowel, diction, style, and technique can produce rapidly shifting spectral envelopes. Loudness is highly dynamic and expressive, undermining onset detection methods that rely on simple energy changes. Techniques such as vibrato, portamento, and glissando introduce continuous pitch motion and blurred note boundaries that contradict the discrete-note assumptions baked into many AMT models (Nishikimi et al., 2021).

Early singing transcription systems typically decomposed the task into f_0 tracking plus post-hoc quantisation to notes. For instance, heuristic or HMM-based methods such as Tony were designed to estimate note-level pitch contours from monophonic singing, but often struggled with ornamentation and expressive timing (Mauch et al., 2015). More recent work has introduced more structured models that explicitly encode musical and linguistic constraints. Nishikimi et al. (2021) propose a CRNN–HSMM hybrid model for audio-to-music-score transcription, combining a convolutional recurrent neural network acoustic model with a semi-Markov language model over keys, pitches, and onset times. This hybrid approach leverages both local spectral cues and higher-level score regularities, and further shows that applying singing voice separation before transcription can improve performance in mixture conditions.

Recent work also explores fairness, robustness, and multimodality in singing AMT. Some studies report that models tend to perform better on certain singer groups (e.g., female voices) and propose methods to mitigate these biases (Gu et al., 2023). Others jointly model audio and visual cues (e.g., lip motion) to improve note detection in challenging acoustic conditions, or integrate AMT systems into end-to-end lyric transcription and singing synthesis pipelines (Gu, Zeng, Zhang, et al., 2023). Overall, the singing-voice literature converges on three key observations: note-level AMT for vocals is substantially harder than for instruments; dedicated architectures and datasets can significantly improve performance; and separation or robust front-end processing is often necessary in polyphonic settings.

In contrast to these specialised systems, this dissertation deliberately uses instrument-agnostic AMT models, Basic Pitch and MT3, trained primarily on instrumental data, and investigates whether preprocessing the audio via timbre transfer can make singing voice more compatible with these architectures.

2.3 Datasets for Vocal AMT

Progress in singing-voice transcription has been tightly linked to the development of high-quality datasets. MIR-1K, NUS-48E, TONAS, and related corpora have supported early research in vocal f_0 estimation and melody extraction, but they lacked fully aligned note-level ground truth in MIDI form. More recently, VocalSet was introduced as a 10.1-hour dataset of monophonic, a cappella singing by 20 professional singers, providing extensive coverage of standard and extended vocal techniques across all five vowels (Wilkins et al., 2018). Faghih and Timoney extend this resource with Annotated-VocalSet, adding detailed onset/offset annotations, MIDI pitch estimates, and higher-level descriptors, thereby enabling note-level AMT evaluation without the need for manual labelling (Faghih & Timoney, 2022).

Other datasets, such as the Choral Singing Dataset, provide framewise f_0 and note annotations for choir recordings. These corpora differ in the degree of polyphony, stylistic variability, and annotation detail. The present work is positioned firmly in the monophonic regime; therefore, this study leveraged Annotated-VocalSet's precise note-level labels to isolate the effect of timbre transfer without confounding factors from accompaniment or complex mixtures.

2.4 Timbre Transfer and Differentiable Digital Signal Processing (DDSP)

Timbre transfer refers to transforming an audio signal from one timbral domain (e.g., human voice) into another (e.g., violin, trumpet) while preserving core musical features such as pitch, rhythm, and dynamics (Engel et al., 2020). Early neural approaches to musical timbre transfer adapted image style-transfer techniques to frequency representations. For example, TimbreTron uses a CycleGAN operating on constant-Q spectrograms, followed by a WaveNet vocoder to resynthesise high-quality audio; human listening tests confirm that the method can recognisably transfer timbre while largely preserving musical content (Huang et al., 2019). This was not used since its weights were not publicly available at the time of writing.

A different but complementary line of work is Differentiable Digital Signal Processing (DDSP). Engel et al. (2020) introduce DDSP as a library that integrates classic signal-processing elements, harmonic oscillators, filtered noise generators, and reverb, with neural networks in an end-to-end differentiable framework. For monophonic sources such as solo instruments or vocals, DDSP-based autoencoders learn to map conditioning features (f_0, loudness, timbre embeddings) to synthesis parameters, enabling interpretable control of pitch and loudness and high-fidelity resynthesis without large autoregressive models. Demonstrations include transforming a singing voice into a violin timbre by re-synthesising extracted pitch and loudness trajectories with a violin-trained decoder.

Timbre has also been studied explicitly in the context of AMT. Hernández-Oliván et al. (2021) compare deep learning methods for timbre analysis in polyphonic AMT and show that note transcription accuracy is highly sensitive to onset strength and timbral envelope, particularly for models such as Onsets and Frames that were optimised on piano data. Their results suggest that varying timbre can significantly affect model performance, and that models trained on piano do not automatically generalise to instruments with very different temporal and spectral characteristics. However, in most of this literature, timbre transfer is treated as an end in itself (for creative synthesis) or as a factor in AMT model design, rather than as an explicit preprocessing stage for vocal transcription.

These recent timbre transfer works therefore, further highlight a disconnect between perceptual quality and downstream MIR utility: models that score highly on timbre similarity or naturalness do not necessarily preserve the spectral structure needed for reliable symbolic inference. This raises the question of how “good” a timbre transfer needs to be for AMT: is higher perceptual fidelity always beneficial, or can it introduce expressive variability that confuses models tuned for more regular, instrument-like inputs?

In summary, the literature reviewed above reveals several key points:

1. Current AMT systems are highly effective in constrained, instrument-centric settings but remain fragile when confronted with new timbres and expressive performance styles. Architectures such as Onsets and Frames, Basic Pitch, and MT3 achieve strong performance on piano and related datasets, yet their behaviour on challenging vocal material is less well documented.
2. Singing-voice AMT has motivated specialised architectures and datasets, but these typically require new models or complex training regimes. CRNN–HSMM hybrids, hierarchical classification networks, and end-to-end EfficientNet-based systems trained on datasets like MIR-ST500 show that dedicated designs can handle vocal idiosyncrasies more effectively than generic AMT systems. However, such systems are comparatively heavyweight and not always available as plug-and-play tools.
3. Timbre transfer and DDSP provide powerful, interpretable ways to reshape audio, but their impact on downstream AMT has not been systematically studied. Most prior work evaluates timbre transfer using perceptual listening tests, spectral similarity metrics (e.g., MFCC distances), or embedding-based measures, rather than task-aware metrics such as note-level F1, onset/offset accuracy, or transcription error rate.

Taken together, these observations expose a clear gap: there is little systematic evidence on whether converting singing voice into instrument-like timbres via neural timbre transfer actually helps or hinders state-of-the-art, instrument-optimised AMT models. Existing work tends either to design new architectures specifically for vocals or to develop timbre transfer for creative or perceptual goals without considering its effect on symbolic transcription.

This dissertation addresses that gap by:

1. Applying DDSP-based timbre transfer to monophonic vocal recordings from Annotated-VocalSet, generating instrument-like renderings (e.g., trumpet, tenor saxophone, violin, flute).
2. Feeding both raw and timbre-transfer-based audio into two widely used AMT models, Basic Pitch and MT3, and evaluating note-level and onset/offset metrics using the MIR-Eval toolkit.
3. Analysing how different timbres and timbre similarity measures relate to changes in transcription performance, thereby probing the non-trivial relationship between perceptual timbre fidelity and symbolic accuracy.

By positioning timbre transfer as a task-aware pre-processing step rather than a purely perceptual goal, the study tests a widely held but rarely examined assumption: that making vocals “more instrument-like” will automatically make them easier to transcribe. The empirical findings presented in later chapters challenge this assumption and provide new insights into the interplay between timbre, model design, and vocal AMT.

3. Methodology

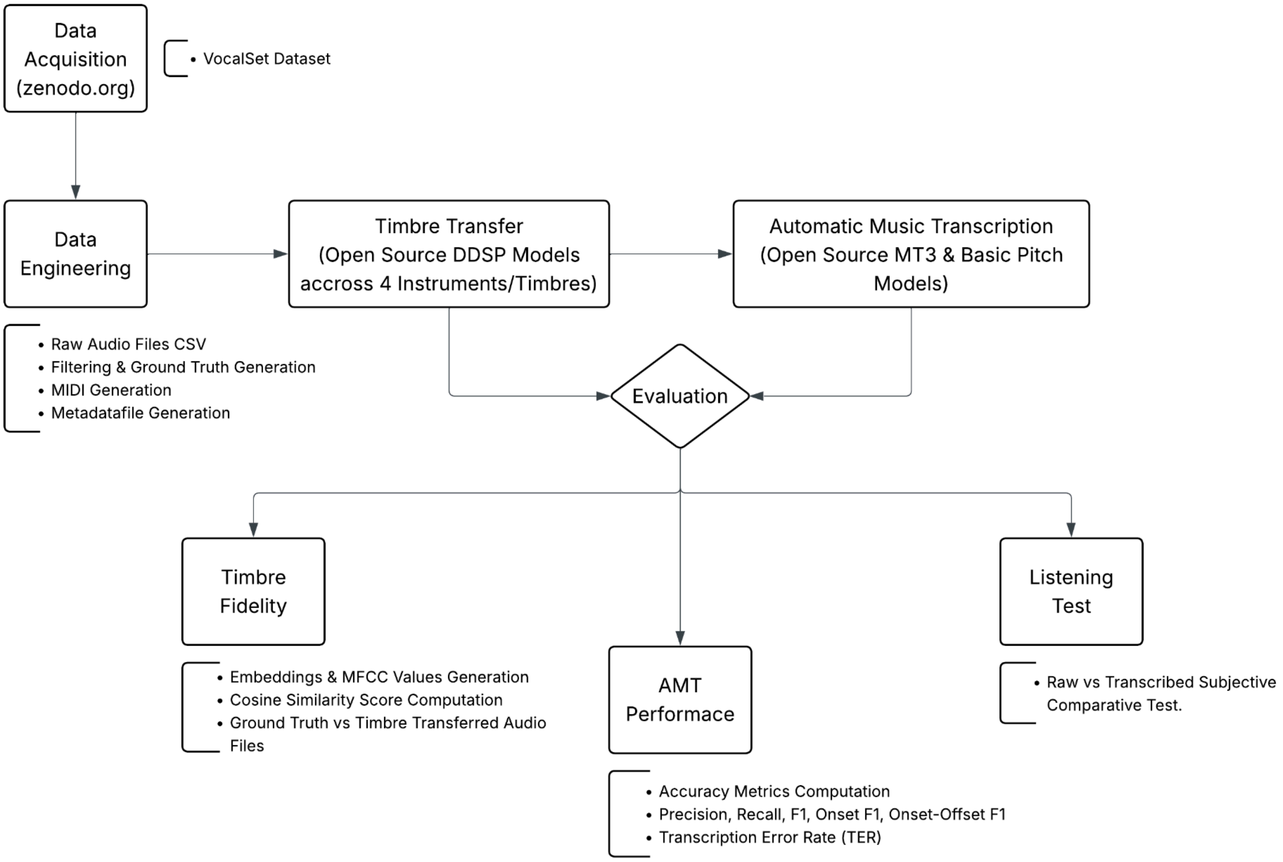

Data Acquisition (zenodo.org)
• VocalSet Dataset
Data Engineering
• Raw Audio Files CSV
• Filtering & Ground Truth Generation
• MIDI Generation
• Metadatafile Generation
Timbre Transfer (Open Source DDSP Models accross 4 Instruments/Timbres)
Automatic Music Transcription (Open Source MT3 & Basic Pitch Models)
Evaluation
Timbre Fidelity
• Embeddings & MFCC Values Generation
• Cosine Similarity Score Computation
• Ground Truth vs Timbre Transferred Audio Files
AMT Performace
• Accuracy Metrics Computation
• Precision, Recall, F1, Onset F1, Onset-Offset F1
• Transcription Error Rate (TER)
Listening Test
• Raw vs Transcribed Subjective Comparative Test.

Figure 3.0: Singing Voice Unified Pipeline Architecture

3.0 Single Example Excerpt Pipeline End-to-End: A Preview

Below is a preview “walk-through” of what would happen to one example excerpt, m6_row_straight.wav, as it moves through the full pipeline shown in Figure 3.0 above. The detailed rationale for each step will be given in the following subsections of this chapter.

1. Data Engineering (building the dataset):
    a. From the VocalSet (Annotated) dataset, all files whose name contains the keyword row_straight are selected. One of these is the audio file m6_row_straight.
    b. The MIDI version is created using pretty_midi and stored in the MIDI folder (dataset/midi/m6_row_straight.mid).
    c. The corresponding WAV file is copied into a local dataset folder (dataset/wav/m6_row_straight.wav) so that all audio used in the experiments is in a consistent location.
    d. A row is added to metadata.csv:

   - the excerpt name (m6_row_straight),
   - its MIDI file path, and
   - its WAV file path.
2. Timbre Transfer (DDSP models). Using the information in metadata.csv, the timbre transfer step:
   a. Loads dataset/wav/m6_row_straight.wav at 16 kHz (the default sampling rate for the DDSP models).
   b. Extracts the audio features required by DDSP, such as loudness (in dB), fundamental frequency (F0), and pitch.
   c. Feeds these features into four pretrained DDSP models (violin, flute, trumpet, tenor saxophone) to resynthesise the same excerpt with each of these timbres.
   d. Saves the resulting audio in separate folders, e.g., tt_dataset/violin/m6_row_straight.wav, tt_dataset/flute/m6_row_straight.wav, and so on.
3. AMT. Next, both the original and timbre-transfer-based audio are passed through the AMT models (Basic Pitch and MT3):
   a. The original excerpt dataset/wav/m6_row_straight.wav is transcribed by each AMT model. The resulting MIDI files are stored as “baseline” transcriptions, e.g., amt_dataset/baseline/basic_pitch/m6_row_straight.mid.
   b. Each timbre-transfer-based version (violin, flute, trumpet, tenor saxophone) creates the “converted” transcriptions that are stored in paths of the form amt_dataset/converted/{amt_model_name}/{ddsp_model_name}/m6_row_straight.mid.
4. Evaluation. The evaluation stage combines three types of analysis:
   a. Timbre Fidelity. The original audio dataset/wav/m6_row_straight.wav and each transferred version (e.g., tt_dataset/violin/m6_row_straight.wav) are compared using similarity measures. This yields a numerical score describing how similar the generated timbre is to the target instrument.
   b. AMT Performance. The MIDI files produced by the AMT models (both baseline and converted) are compared against the ground-truth MIDI dataset/midi/m6_row_straight.mid. For each model, I compute standard transcription metrics: precision, recall, F1, onset-only and Onset-Offset F1 scores, and transcription error rate (TER). For example, Basic Pitch may obtain scores similar to: [0.428, 0.55, 0.483, 0.452, 0.400, 0.518, 0.354, 1.185] for these metrics, respectively.
   c. Listening Test. Finally, a randomised listening test is conducted using the MuseScore tool and VLC. Participants hear the converted MIDI files (e.g., amt_dataset/converted/{amt_model_name}/{ddsp_model_name}/m6_row_straight.mid) and judge whether they sound similar (“Yes”/“No”) to the original recording dataset/wav/m6_row_straight.wav on VLC. Figure 3.1 below shows how MuseScore renders the MIDI file.

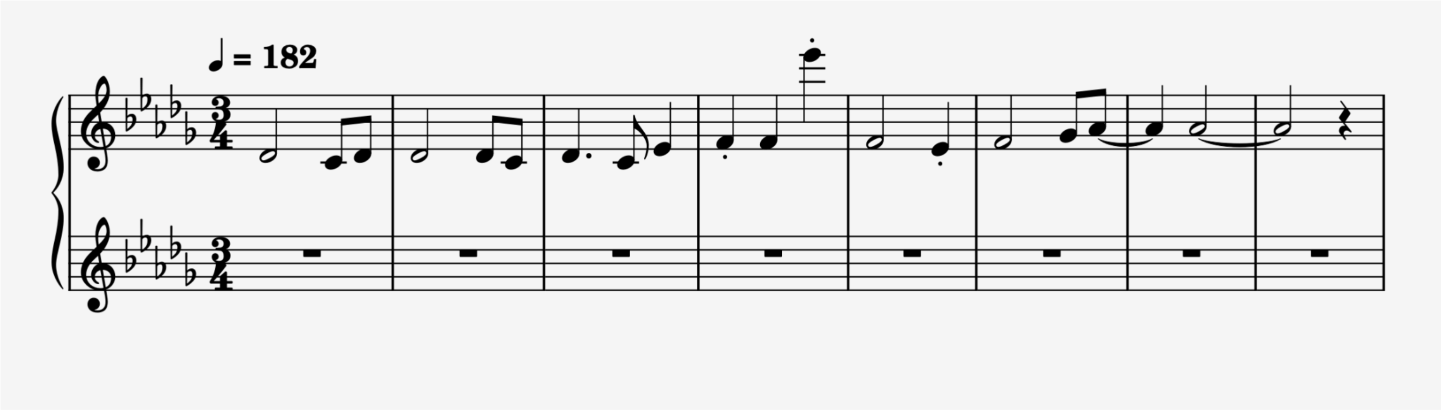


Figure 3.1: m6_row_straight.mid on MuseScore

This single example excerpt illustrates how every file in the dataset would move through the full pipeline shown in the architecture diagram.

3.1 Dataset Selection and Collection

The objective for the selection and collection process was to find datasets that could provide audio with human singing voices and their associated MIDI versions. The rationale behind this was that the singing voices would be required for the timbre transfer pipeline, while the associated MIDI versions were to be used as potential ground truths during the evaluation of our hypothesis.

Datasets that were originally identified included the DALI, MIR-1K, and NUS-48, whose search criteria were singing audio files. They fell short as access to DALI was declined due to academic affiliation, as I did not have access to a school email, while the rest lacked a means to gather associated MIDI versions of the audio.

Two further datasets were later identified through additional exploration, recommendations, and a literature search on AMT for the singing voice, using the presence of aligned MIDI files as an extra criterion:

1. The Choral Singing Dataset: Contained Frame-wise f0 annotations, note annotations and semi-synchronised MIDI files for each choir section (Cuesta et al., 2019). This was later discarded for only containing vibrato stylistic variations, which would introduce too much noise.
2. VocalSet - A Singing Voice Dataset: Monophonic recorded audio of experienced performers across standard and extended vocal styles on all five vowels.VocalSet aims to improve the state of existing singing voice datasets and singing voice research by capturing not only a range of vowels but also a diverse set of voices (Wilkins et al., 2018).
3. Annotated-VocalSet: An extension of VocalSet that offered comprehensive annotations for 10.1 hours of monophonic recordings featuring 20 experienced vocalists (9 male, 11 female) performing standard and extended vocal techniques across all five vowels. This was the key dataset for the investigation (Faghih & Timoney, 2022).

Both datasets were obtained from zenodo.org, constructed and developed by researchers to facilitate participation in Open Science (UNESCO, 2021).

3.2 Dataset Preprocessing

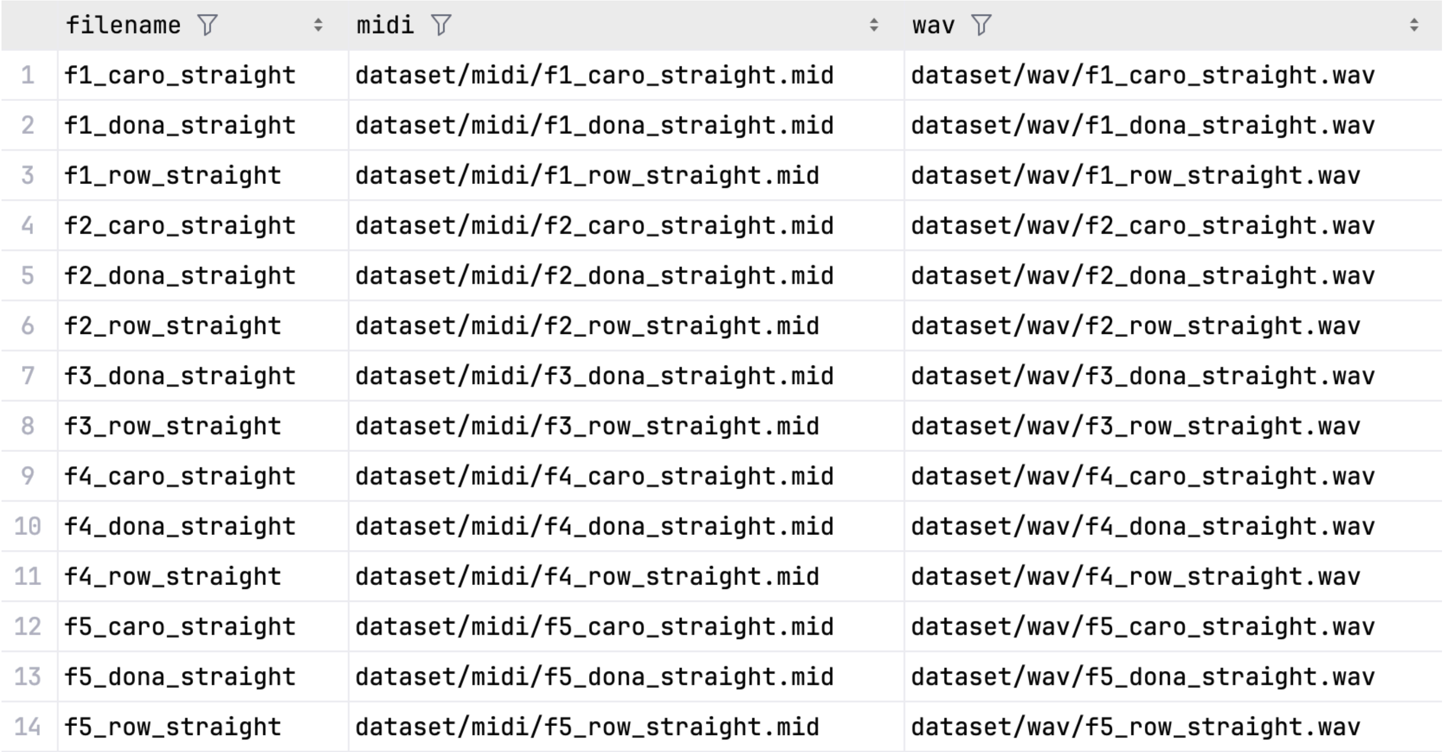
In addition to the raw audio recordings in WAV format, the Annotated-VocalSet provides annotation files in both raw and extended CSV formats. The extended files contain higher-level musical features, including onset/offset times, note durations, estimated and ground-truth MIDI pitch codes, and associated lyrics. These extended annotations effectively describe the sung notes in symbolic form, but they are not distributed in standard MIDI format.

For our experiments, I relied exclusively on the extended CSV files, which provide reliable temporal and pitch annotations. I selected extended_2, the version of the dataset in which note onsets occur immediately after the preceding offsets. This choice ensured clean, non-overlapping note boundaries, making it well-suited for generating MIDI files intended for playback and for visualisation tasks (Faghih & Timoney, 2022).

The dataset was further filtered to retain only files belonging to the straight category of excerpts (“caro,” “dona,” and “row”), which correspond to a natural singing voice without stylistic variations. This step removed experimental vocalisations and ensured that the training data reflected conventional melodic singing.

Using the Python library pretty_midi (Raffel, 2022), I converted the selected extended CSV annotations into MIDI files. Each note was created using the annotated start time, end time, and estimated MIDI pitch. Only rows labelled as Sound in the Type column were included, thereby excluding rests and transitions that do not correspond to discrete musical events. A fixed velocity of 100 was applied to all notes, although the annotations also provide amplitude values that could be mapped to velocity in future work.

To align symbolic and audio representations, I also collected the corresponding .wav recordings for each excerpt. Finally, I generated a metadata file (metadata.csv) that records the filenames and the paths to the paired .wav and ground truth .mid files. This metadata table served as the entry point for subsequent training and evaluation, ensuring consistent linkage between audio and symbolic representations of the singing voice.



	filename	midi	wav
1	f1_caro_straight	dataset/midi/f1_caro_straight.mid	dataset/wav/f1_caro_straight.wav
2	f1_dona_straight	dataset/midi/f1_dona_straight.mid	dataset/wav/f1_dona_straight.wav
3	f1_row_straight	dataset/midi/f1_row_straight.mid	dataset/wav/f1_row_straight.wav
4	f2_caro_straight	dataset/midi/f2_caro_straight.mid	dataset/wav/f2_caro_straight.wav
5	f2_dona_straight	dataset/midi/f2_dona_straight.mid	dataset/wav/f2_dona_straight.wav
6	f2_row_straight	dataset/midi/f2_row_straight.mid	dataset/wav/f2_row_straight.wav
7	f3_dona_straight	dataset/midi/f3_dona_straight.mid	dataset/wav/f3_dona_straight.wav
8	f3_row_straight	dataset/midi/f3_row_straight.mid	dataset/wav/f3_row_straight.wav
9	f4_caro_straight	dataset/midi/f4_caro_straight.mid	dataset/wav/f4_caro_straight.wav
10	f4_dona_straight	dataset/midi/f4_dona_straight.mid	dataset/wav/f4_dona_straight.wav
11	f4_row_straight	dataset/midi/f4_row_straight.mid	dataset/wav/f4_row_straight.wav
12	f5_caro_straight	dataset/midi/f5_caro_straight.mid	dataset/wav/f5_caro_straight.wav
13	f5_dona_straight	dataset/midi/f5_dona_straight.mid	dataset/wav/f5_dona_straight.wav
14	f5_row_straight	dataset/midi/f5_row_straight.mid	dataset/wav/f5_row_straight.wav

Table 3.0: Sample Metadata

3.3 Timbre Transfer Pipeline

The timbre transfer (tt) experiments were designed to explore the capability of neural synthesis models to re-render vocal audio with the timbral characteristics of different instruments. Specifically, I evaluated Differentiable Digital Signal Processing (DDSP) models in conjunction with MT3 and Basic Pitch transcription systems to study end-to-end conversion from singing voice to instrument-like renderings. The goal was to assess the quality, realism, and generalisation behaviour of pre-trained DDSP models when applied to raw, monophonic singing recordings from the dataset.

The experimental environment was implemented locally after extensive adaptation of DDSP's original Colab (https://bit.ly/4oxDpt5). Package installation required a forked version of the official repository to address compatibility issues and dependency mismatches on macOS. All Colab-specific utilities, such as file upload/download, IPython widgets, and Google Cloud Storage mounts, were removed or replaced with local equivalents to enable standalone execution. The play() audio rendering function from the Colab notebook was reimplemented as a save_wav() function to export synthesised outputs as .wav files for later analysis.

Model checkpoints and dataset statistics (e.g., dataset_statistics.pkl) were manually downloaded from Google Cloud Storage since remote fetching through the original notebook failed in the local environment. The instrument models were stored under ../pretrained/{ddsp_model_name}/, containing both the operative configuration files (operative_config-0.gin) and checkpoint weights (ckpt-*). The DDSP package was verified to run successfully on macOS, while Linux installations consistently failed due to TensorFlow compilation conflicts.

The timbre transfer process consisted of five main stages:

1. Audio loading: Raw vocal recordings were retrieved from the dataset metadata CSV (Fig. Sample Metadata in the previous section). Each file was read as a float waveform using audio_bytes_to_np() from DDSP utilities.
2. Feature extraction: Spectral and temporal conditioning features (fundamental frequency (f_0), loudness/velocity, and confidence) were computed using DDSP's feature extraction pipeline. These features were truncated or padded to match the temporal resolution of the target model.
3. Pitch and loudness adjustment: The extracted f_0 and loudness curves were preprocessed to match the target instrument's register using dataset statistics. Automatic tuning and loudness normalisation were applied through detect_notes(), fit_quantile_transform(), and auto_tune() functions, ensuring consistent pitch range alignment.
4. Synthesis: A pre-trained DDSP Autoencoder instrument model was loaded and used to synthesise audio conditioned on the modified feature representations. The synthesis step reconstructed audio from its parametric components (harmonic and filtered noise synthesisers).
5. Output storage: The generated waveforms were saved to disk using the adapted save_wav() function for subsequent evaluation and listening tests. These were saved in tt_dataset/{instrument_model_name}/.

On average, each timbre-transfer-based operation took approximately 1 minute 32 seconds per 33-second recording, with runtime dominated by spectral feature extraction and model inference. All experiments were executed sequentially to prevent TensorFlow memory conflicts in the DDSP session graph.

Four pre-trained DDSP instrument models were tested: Violin, Flute, Trumpet, and Tenor Saxophone. MelGAN was not used in this study, as its pre-trained models were not publicly released.

The DDSP framework successfully reconstructed instrument-like audio from monophonic singing inputs with limited artefacts. The generated timbres reflected the spectral envelope of the target instruments, though residual vocal formant structure was occasionally audible. Despite these limitations, the pipeline provided an efficient, modular approach for local timbre transfer without dependence on external APIs or cloud environments.

3.4 AMT Pipeline

The AMT pipeline was implemented to generate symbolic note representations (MIDI files) from both the raw vocal recordings and their timbre-transfer-based versions. The resulting transcriptions form the foundation for quantitative evaluation of transcription accuracy across different domains, specifically, to assess whether raw or timbre-transfer-based audio can be transcribed more accurately relative to ground-truth annotations. Two state-of-the-art AMT systems were employed: Basic Pitch and MT3.

Both AMT models convert audio signals into symbolic note sequences, yet they differ substantially in design philosophy and input representations.

1. Basic Pitch: A lightweight convolutional recurrent network trained primarily on polyphonic datasets. It jointly predicts note activation, pitch, and onsets using harmonic CQT features (Bittner et al., 2022).
2. MT3: A large-scale sequence-to-sequence model that transcribes audio to symbolic notes through a transformer-based encoder–decoder architecture. MT3 uses a multi-task tokenisation scheme to model polyphonic, multi-instrument recordings and has demonstrated state-of-the-art performance on the MAESTRO dataset (Gardner et al., 2021).

In this work, both models were repurposed for monophonic singing transcription and evaluated on the VocalSet-derived data. The pipeline was implemented using the open-source versions of Basic Pitch (basic_pitch.inference) and MT3 (via the custom InferenceModel wrapper).

The pipeline automates transcription for all dataset audio files and outputs per-song MIDI representations for both models. The workflow consists of the following steps:

1. Audio loading: Audio were either loaded from the metadata CSV file or paths containing converted audios. This was supported by the load_audio_files() function.
2. Model initialisation: The Basic Pitch model is loaded from the ICASSP 2022 checkpoint using the default spectral front-end, while MT3 is initialised using locally stored transformer checkpoints (checkpoints/mt3/). Both models operated at a 16 kHz sample rate.
3. Automatic transcription:
    a. For Basic Pitch, the predict() function computes spectral features, detects note events, and directly exports the resulting note_seq.NoteSequence to a MIDI file.
    b. For MT3, the input audio is preprocessed with the helper upload_audio() function, passed through the transformer model, and converted into a MIDI sequence using note_seq.sequence_proto_to_midi_file().
4. Output organisation: Each model’s transcription outputs are saved separately to ensure clear traceability:
    a. Baselines which were generated from original vocal audio files: amt_dataset/baselines/basic_pitch/ and amt_dataset/baselines/MT3/.
    b. Converted or Timbre Transferred were associated with the model used to generate them: amt_dataset/converted/basic_pitch/{ddsp_model_name}/ and amt_dataset/converted/MT3/{ddsp_model_name}/.

On average, each transcription operation took approximately 20 seconds per instrument for Basic Pitch on macOS, while ~13 minutes per instrument in MT3 running on Ubuntu. Similar to the previous section, some of the toolings only ran exclusively on different systems.

3.5 Evaluation

3.5.1 MIR Evaluation

To objectively assess the performance of the AMT systems, I employed the MIR-Eval library, a standardised Python framework for evaluating MIR algorithms. This toolkit provides reproducible and transparent implementations of evaluation metrics widely adopted in music transcription research. The framework was used to compute Precision, Recall, F1-score, Onset F1, and Onset–Offset F1 for all transcription outputs (Raffel et al., 2014), with some of the computed values being extended to provide the Transcription Error Rate (TER), which measures the proportion of incorrectly transcribed notes (insertions and deletions) relative to the total number of reference notes.

The evaluation process compared the predicted MIDI transcriptions from each AMT model against the ground-truth annotated MIDI files provided by the Annotated-VocalSet dataset. Separate evaluations were conducted for:

1. Original vocal recordings, transcribed using Basic Pitch and MT3; and
2. Timbre-transfer-based audio, synthesised using DDSP across four instruments: violin, flute, trumpet, and tenor saxophone.

Each evaluation yielded model-specific performance metrics for every song, followed by aggregated averages representing overall system accuracy.

Tolerance settings were as follows: Onsets were considered correct if the predicted onset occurred within 50 ms of the reference onset, and Offsets were considered correct if they fell within 20 % of the ground-truth note duration, following MIR-Eval's standard transcription metric definitions.

The evaluation followed a consistent four-stage process across all models and datasets:

1. MIDI Extraction: Each MIDI file was parsed using the pretty_midi library to extract note start times, end times, and pitch values for all instruments in the file.
2. Pairwise Matching: Predicted and ground-truth MIDI files were matched by filename across the two directories, dataset/midi/ and the respective model output directories, amt_dataset/baselines/basic_pitch/, amt_dataset/baselines/MT3/, and amt_dataset/converted/* versions.
3. Metric Computation: The note-level and onset-level metrics were computed using MIR-Eval. The function evaluate_midi_pair() calculated the metrics for each file pair, while batch_evaluate() aggregated and stored results in .csv format.
4. Summary and Visualisation: The average results for each model and dataset were printed to the console and saved for later visualisation in tables and comparative plots.

This evaluation methodology provides a transparent, quantitative, and reproducible assessment of AMT system performance. The combination of MIR-Eval metrics, dataset-wide averaging, and comprehensive visual analysis ensured that model comparisons are fair, statistically interpretable, and aligned with established MIR community standards.

Note that the formula for TER = (FP+FN)/N_ref ×100%, where FP and FN are the numbers of false positive and false negative note events, respectively, and N_ref is the number of reference notes, following error-rate style metrics used in AMT and related tasks (Vogel et al., 2005). A lower TER indicates more accurate transcription, and TER acts as an error-based complement to the F1-score: because both metrics are derived from the same confusion matrix, systems with high F1 typically exhibit low TER when evaluated on the same note events. In this study, I interpret TER values such that 0% corresponds to perfect (error-free) transcription, values < 30% indicate fairly accurate transcriptions (generally associated with good F1-scores), values around 100% indicate as many transcription errors as reference notes, and values > 100% indicate performance worse than a naive "random guessing" behaviour due to an excess of spurious notes. For plotting purposes, I omitted the ×100% scaling so that TER values remained on the same numerical scale as the other metrics. A random seed was included for reproducibility.

3.5.2 Timbre Fidelity Evaluation

Timbre fidelity was evaluated to quantify how acoustically and perceptually similar the timbre-transfer-based (converted) audio samples were to the original singing voice recordings. In this context, fidelity reflects how closely the converted instrument preserves the spectral and temporal characteristics of the human vocals. Two complementary techniques were employed to capture both signal-level and perceptual similarities between paired audio files: Mel-Frequency Cepstral Coefficients (MFCC) (*Mel-Frequency Cepstrum*, 2019) and deep audio embeddings (Cramer et al., 2019).

MFCCs were extracted using the Librosa library to represent spectral envelopes and timbral textures, while deep embeddings were generated using OpenL3 (marl, 2023), a model pre-trained on large-scale music and environmental sound datasets to align with perceptual similarity. Both feature types were aggregated across time, and cosine similarity was computed between each converted and original audio pair to yield two fidelity measures: MFCC Similarity and Embedding Similarity. Higher similarity values indicated stronger preservation of the source timbre characteristics.

The implementation was automated via a Python script (timbre_fidelity.py), which loaded each pair of WAV files, computed MFCC and embedding features, and recorded similarity scores in a results file. A batch script (run_timbre_fidelity_batch.sh) was developed to iterate over the instruments, each representing a distinct timbral domain. This approach allowed scalable and reproducible evaluation across all DDSP-generated datasets, with each instrument taking approximately 15 minutes to process due to OpenL3 embedding computation.

The evaluation pipeline consisted of:

1. Data Loading: Aligning converted and original WAV pairs.
2. Feature Extraction: Computing MFCCs and OpenL3 embeddings.
3. Cosine Similarity Computation: Quantifying timbre closeness per pair.

4. Batch Execution and Aggregation: Collecting mean similarity scores per instrument for comparative analysis.
5. Summary and Visualisations: Average MFCC and Embedding Similarity scores were compared across instruments and correlated with AMT performance metrics to investigate whether higher timbre fidelity corresponded to improved transcription accuracy.

3.5.3 Listening Evaluation

A small subjective qualitative listening test was conducted to evaluate the perceptual similarity between the original vocal recordings and their corresponding timbre-transfer-based and transcribed versions. This was conducted as a simple sanity check. The goal was to determine whether timbre transfer and AMT preserved the melodic “note flow” of the original performance. The procedure followed the Subjective Quality Evaluation guidelines by Bäckström et al. (2022).

A Python randomisation script (listening_test.py) was implemented to automate the selection of five pairs comprising: 1, an original singing-voice sample, and 2, a transcribed timbre-transfer-based version rendered through DDSP models generated by either MT3 or Basic Pitch. Both the instrument and transcription model were randomly assigned to ensure unbiased representation. The script produced two output files: a .txt listing of selected trios and a structured .csv file (listening_test.csv) containing columns for Singing Voice, Transcribed, Instrument, Similarity Response, and Notes.

Each pair was auditioned using MuseScore and VLC, with MIDI transcriptions rendered for playback alongside the corresponding original audio recordings. I compared each original and converted pair and answered a single question: *“Do the two versions sound musically similar in terms of note flow and structure?”*
Responses were binary (Yes/No), with optional notes regarding pitch, rhythm, or expressiveness.

The test emphasised comparative similarity rather than absolute quality to assess whether the transcription preserved the musical structure.

3.6 Software and Hardware Requirements

Due to challenges encountered when installing the required libraries and tools, particularly DDSP running only on macOS and MT3 only on Ubuntu, the table below summarises the software and hardware requirements for each system.

macOS	Software	Python 3.11/3.8
		Anaconda Package Manager

		Libraries and Packages: requirements-macOS.txt file
	Hardware	M1 Pro Unified Chip
		16GB RAM
Ubuntu	Software	Followed a similar format to macOS
	Hardware	NVIDIA RTX 3060 GPU
		Intel i3-12000F CPU
		16GB RAM
Core Repository	https://github.com/martinoywa/singing-voice-amt/	

Table 3.1: Software and Hardware Requirements

4. Results and Discussion

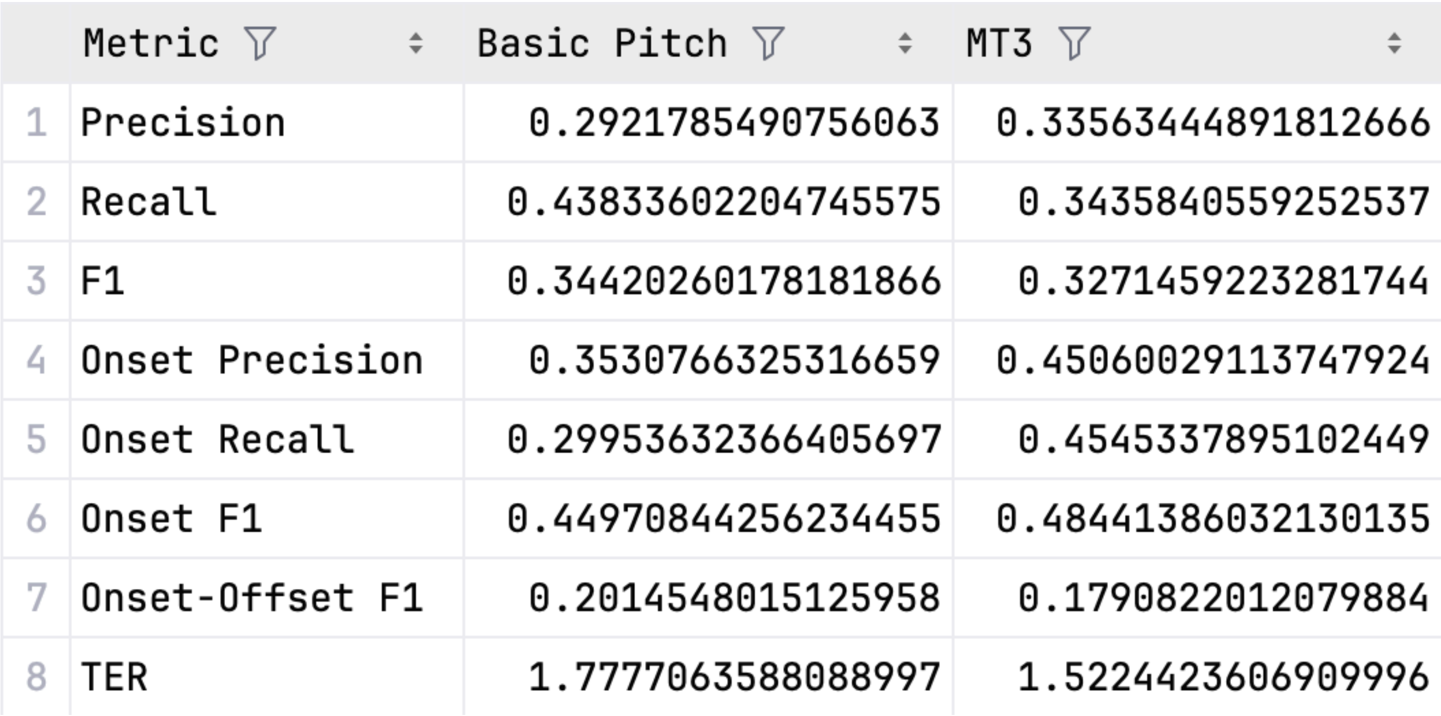
This section will be organised based on the order of our research questions, and will be split into results and discussion subsections.

Effectiveness of Timbre Transfer. To what extent does converting the human singing voice into instrument-like timbres improve the accuracy of downstream AMT compared to direct vocal transcription?

4.0 Results

4.0.0 Baseline Performance Report

As shown in Table 4.0 and Figure 4.0 below, overall, both Basic Pitch and MT3 models underperformed, achieving note-level F1 scores of F1 (0.34, 0.33) and Onset-Offset F1 (0.20, 0.18), respectively. This pattern is also reflected in both Precision and Recall scores, with Basic Pitch out-performing MT3 in Recall (0.44 vs 0.34), while the inverse in Precision (0.29 vs 0.34) and Transcription Error Rate - TER (1.78 vs 1.52).



	Metric	Basic Pitch	MT3
1	Precision	0.2921785490756063	0.33563444891812666
2	Recall	0.43833602204745575	0.3435840559252537
3	F1	0.34420260178181866	0.3271459223281744
4	Onset Precision	0.3530766325316659	0.45060029113747924
5	Onset Recall	0.29953632366405697	0.4545337895102449
6	Onset F1	0.44970844256234455	0.48441386032130135
7	Onset-Offset F1	0.2014548015125958	0.1790822012079884
8	TER	1.7777063588088997	1.5224423606909996

Table 4.0: Average Baseline Performance Metrics Across Models

For Onset predictions, where I detect the beginning of note-events, MT3 shows more consistent performance, which outperforms Basic Pitch across all metrics with scores of Onset Precision (0.45), Onset Recall (0.45), and Onset F1 (0.48).

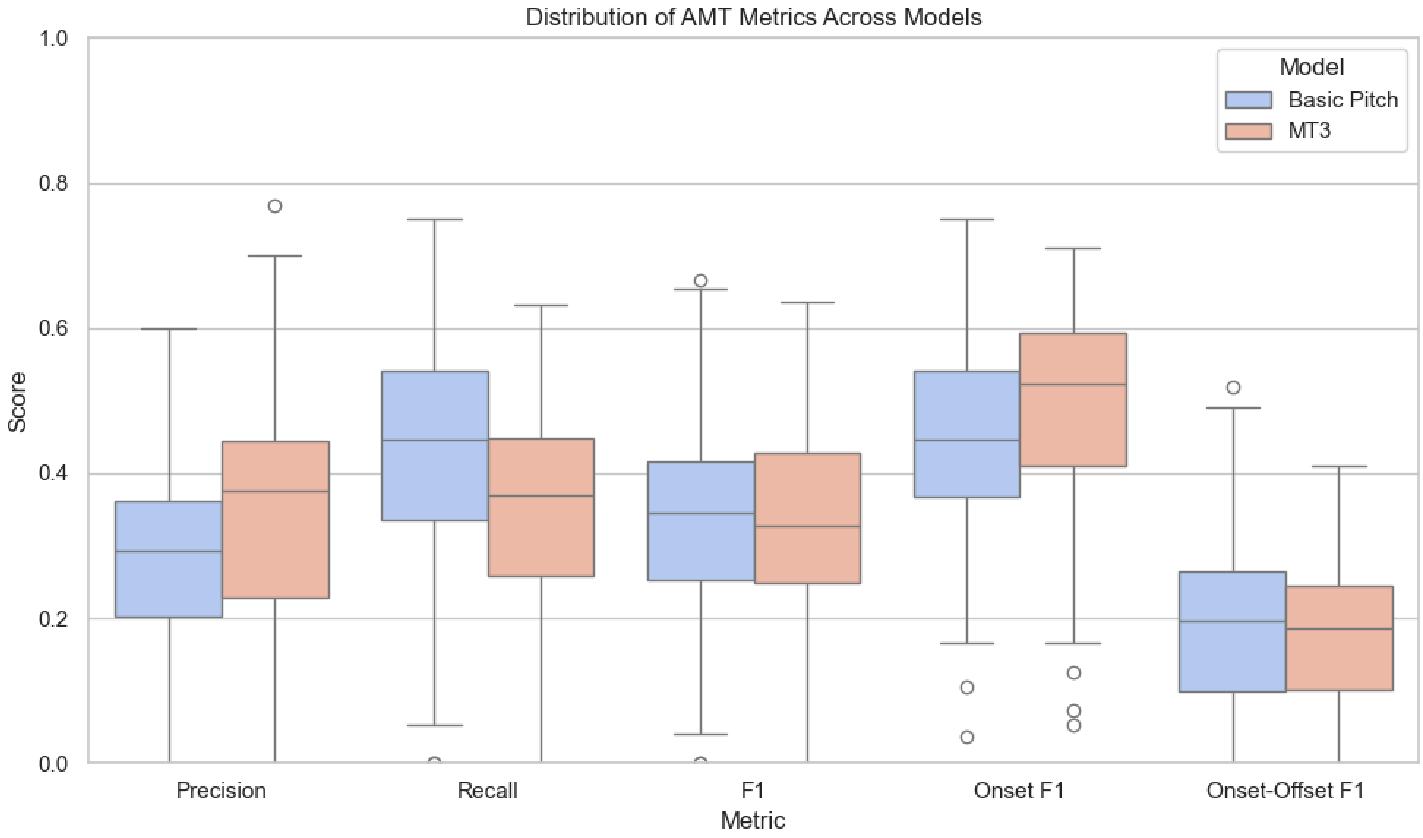


Figure 4.0: Distribution of Baseline Performance Metrics Across Models

For Onset-Offset F1, where I measure sustain, which refers to overlap or note duration accuracy, not just Onset, I discovered a slightly inconsistent correlation between it and Onset-F1, where Basic Pitch Corr (0.73) significantly outperformed MT3 Corr (0.43), indicating that Basic Pitch Onset-Offset F1's correlation to its Onset F1 was significantly stronger than MT3. See Figure 4.2 below.

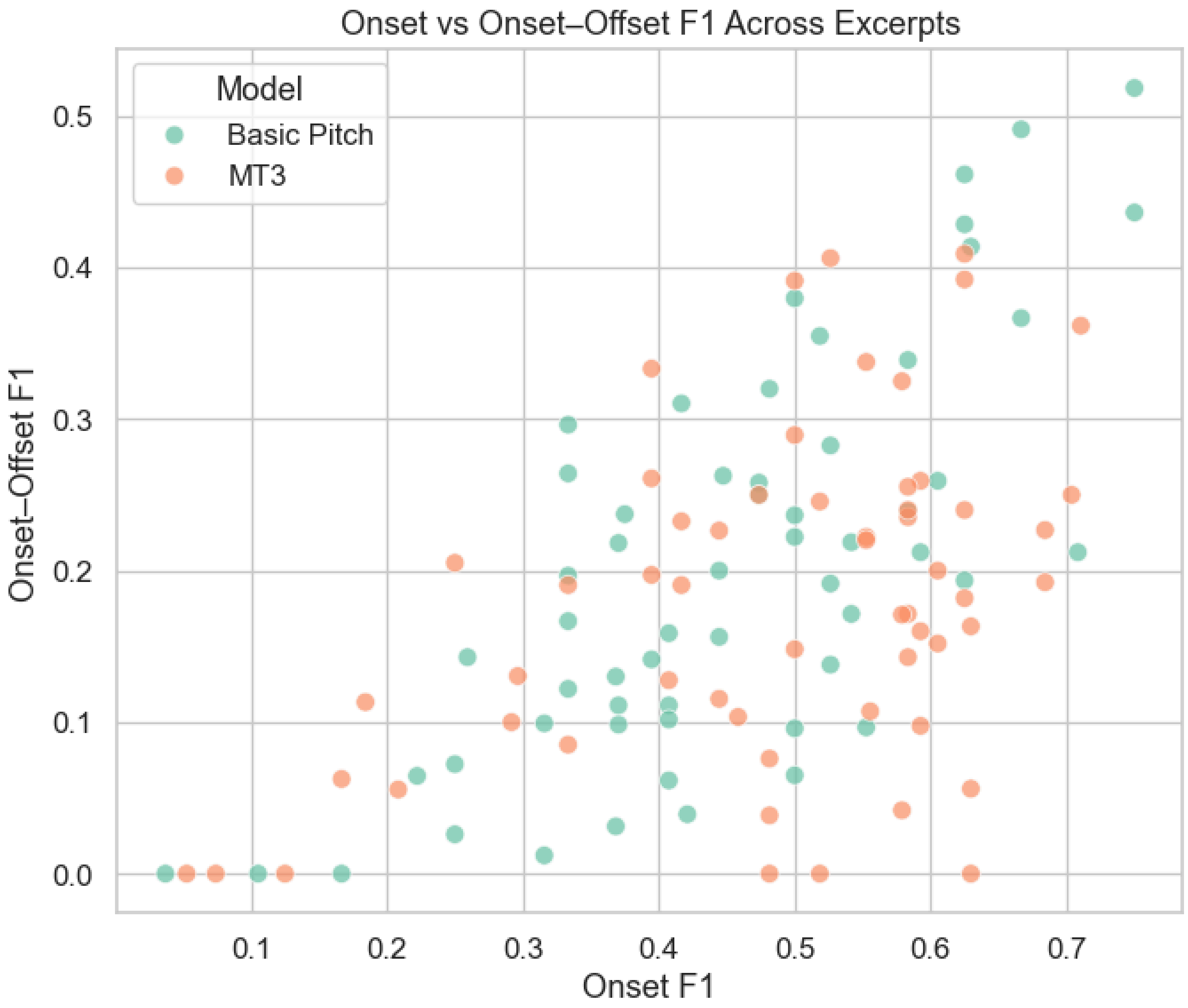


Figure 4.1: Onset vs Onset-Offset F1 Across Excerpts

There was also variability observed across test excerpts, as shown in Figure 4.3 below, with MT3 producing consistently lower TERs compared to Basic Pitch. This pattern was also more consistent among the female-sung excerpts compared to the male-sung ones, a bias also reported by Gu, Zeng, and Wang (2023).

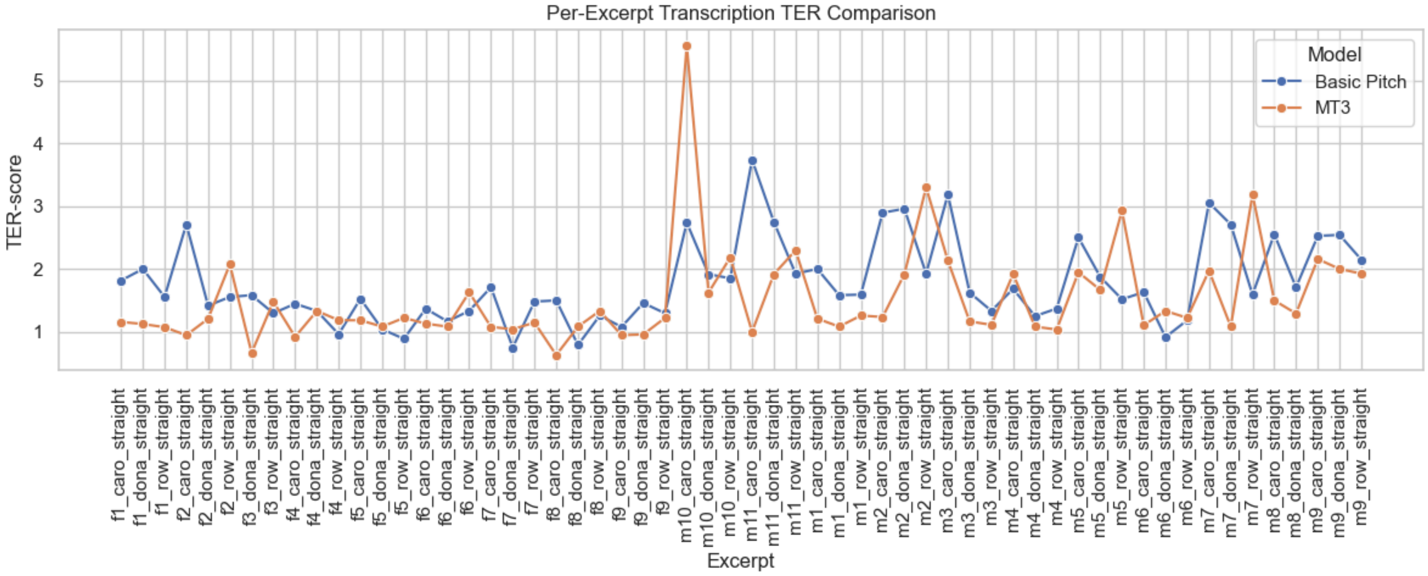
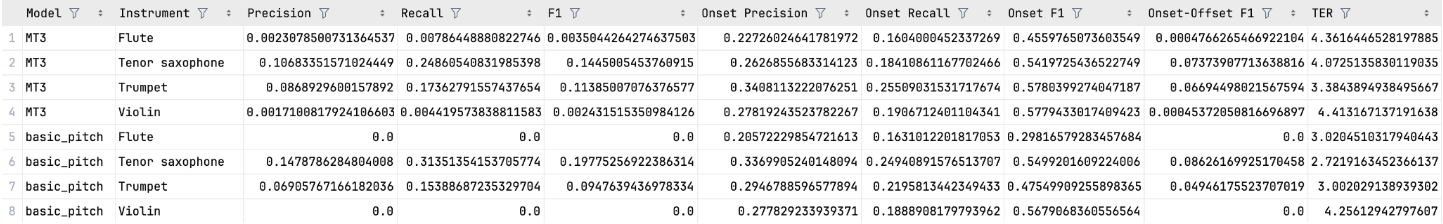


Figure 4.2: Per-Excerpt Baseline AMT TER Comparison

Effectiveness of Timbre Transfer & Optimal Timbre Representation. Which target timbre provides the most reliable transcription performance across different evaluation metrics such as note accuracy, onset/offset detection, and transcription error rate?

4.0.1 Converted Performance Results

This section focuses on presenting converted audio outputs and partial performance tie-in to the pros and cons of the Differentiable Digital Signal Processing (DDSP) technique, which allows for preserved pitch and rhythmic structure from the original vocal recordings while replacing the spectral envelope with instrument-specific harmonic profiles (Engel et al., 2020). This allowed an assessment of how the Basic Pitch and MT3 systems interpret singing melodies when presented as more regular, instrument-like audio signals.



	Model	Instrument	Precision	Recall	F1	Onset Precision	Onset Recall	Onset F1	Onset-Offset F1	TER
1	MT3	Flute	0.0023078500731364537	0.00786448880822746	0.0035044264274637503	0.22726024641781972	0.1604000452337269	0.4559765073603549	0.0004766265466922104	4.3616446528197885
2	MT3	Tenor saxophone	0.10683351571024449	0.24860540831985398	0.1445005453760915	0.2626855683314123	0.18410861167702466	0.5419725436522749	0.07373907713638816	4.0725135830119035
3	MT3	Trumpet	0.0868929600157892	0.17362791557437654	0.11385007076376577	0.3408113222076251	0.25509031531717674	0.5780399274047187	0.06694498021567594	3.3843894938495667
4	MT3	Violin	0.0017100817924106603	0.004419573883811583	0.002431515350984126	0.27819243523782267	0.1906712401104341	0.5779433017409423	0.00045372050816696897	4.413167137191638
5	basic_pitch	Flute	0.0	0.0	0.0	0.20572229854721613	0.1631012201817053	0.29816579283457684	0.0	3.0204510317940443
6	basic_pitch	Tenor saxophone	0.1478786284804008	0.31351354153705774	0.19775256922386314	0.3369905240148094	0.24940891576513707	0.5499201609224006	0.08626169925170458	2.7219163452366137
7	basic_pitch	Trumpet	0.06905767166182036	0.15388687235329704	0.0947639436978334	0.2946788596577894	0.2195813442349433	0.47549909255898365	0.04946175523707019	3.002029138939302
8	basic_pitch	Violin	0.0	0.0	0.0	0.277829233939371	0.1888908179793962	0.5679068360556564	0.0	4.25612942797607

Table 4.1: Average Converted Performance Metrics Across Models and Instruments

As shown in Table 4.1 above and Figure 4.4 below, across the converted dataset, both models displayed measurable transcription activity, though with varied success depending on the target timbre. The Flute and Violin timbres exhibited the lowest measurable note-level scores of F1, Precision, Recall and Onset-Offset F1, with Basic Pitch failing to detect any valid note events (0.00) and MT3 achieving marginal values between (0.00045 and 0.0035). In contrast, the Tenor Saxophone and Trumpet timbres were able to be detected by both models across the same metrics, with the Tenor Saxophone exhibiting the highest Basic Pitch and MT3 metrics across the board: (F1 = (0.20, 0.14), Precision = (0.15, 0.11), Recall = (0.31, 0.25), Onset-Offset F1 = (0.09, 0.07)), respectively.

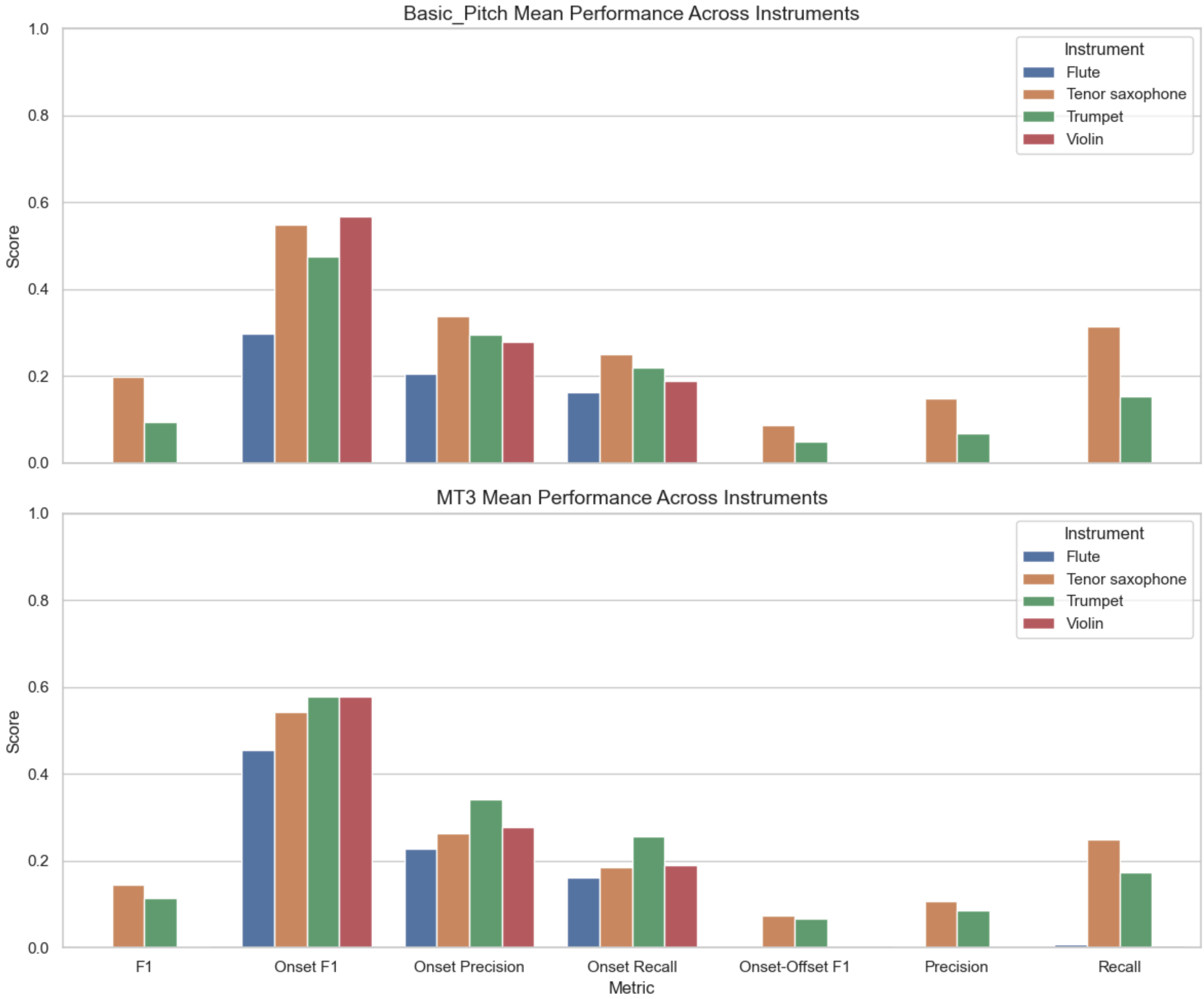
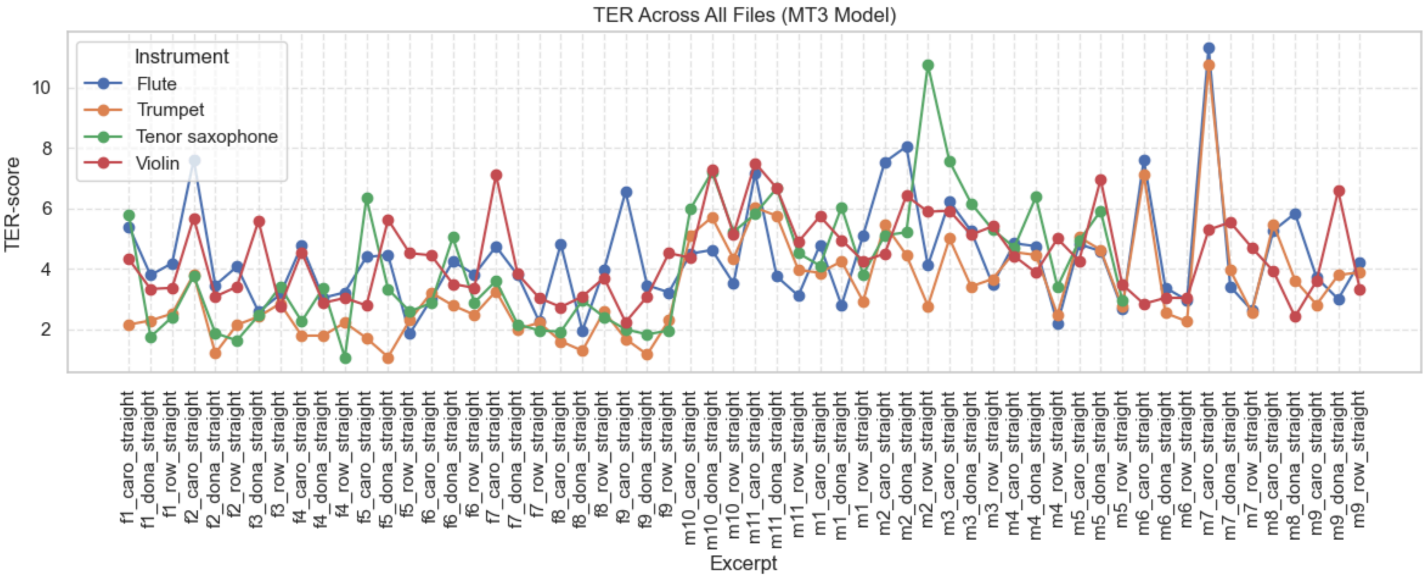


Figure 4.3: Average Converted Performance Metrics Across Models

In terms of Onset predictions, MT3 consistently outperforms Basic Pitch in Onset F1 across all timbres apart from a (0.01) difference in the Tenor Saxophone timbres with Onset F1 (0.54, 0.55) respectively. A similar trend was also experienced in the Onset Recall and Onset Precision scores.

As shown in Table 4.1, in terms of TER, Basic Pitch Transcriptions exhibited the lowest values across ¾ of its timbres, i.e., Tenor Saxophone (2.72), Trumpet (3.00), and, surprisingly, the Flute (3.20). As for MT3, the Trumpet (3.38) outperformed all its timbres, followed by Tenor Saxophone (4.07).

Variability across excerpts, as shown in Figure 4.5 below, was also experienced, where Basic Pitch Transcriptions achieve much lower lows in TER compared to MT3 ones. Across the timbres, the Trumpet consistently achieve lower TER scores across the excerpts, while in Basic Pitch, the Tenor Saxophone takes the lead, which was also experienced above.



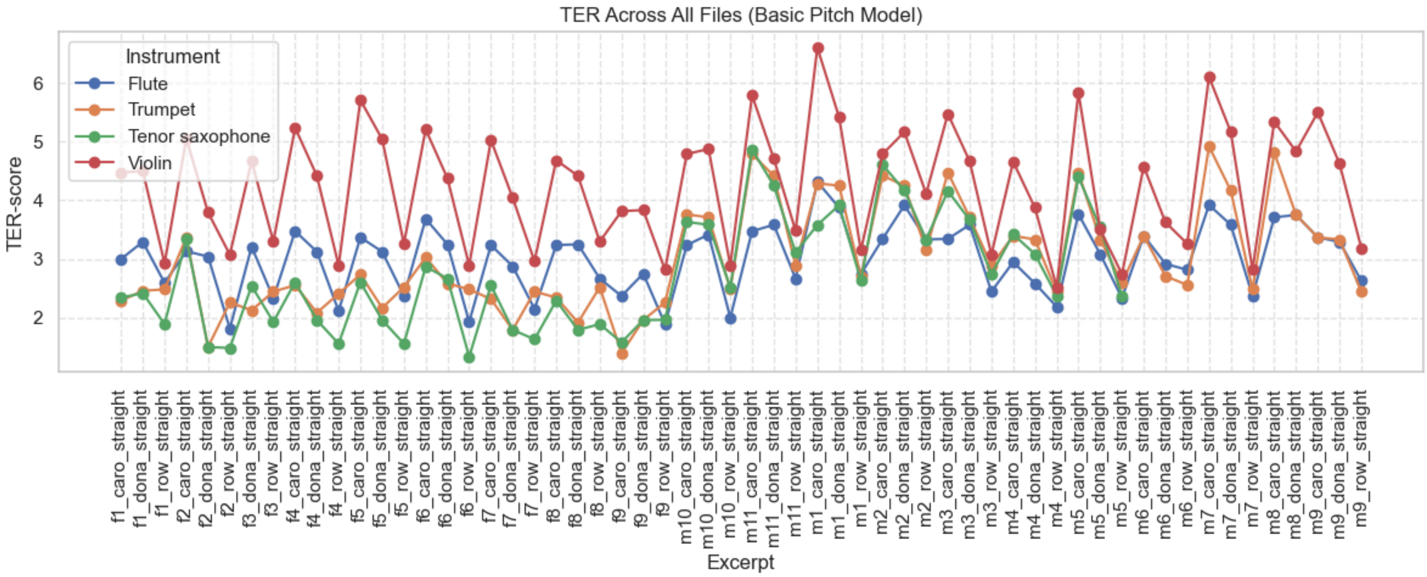
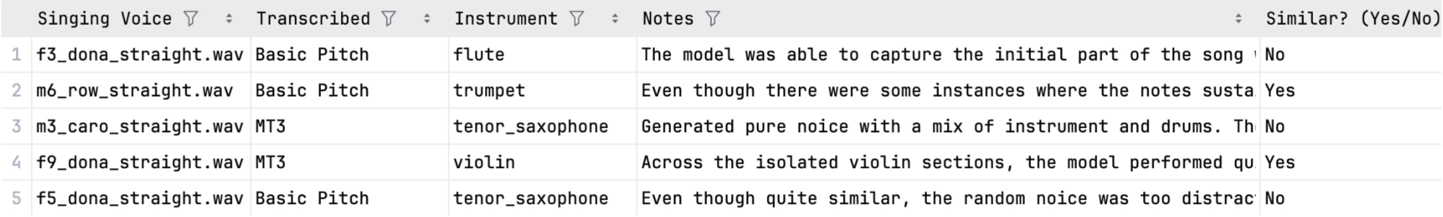


Figure 4.4: Per-Excerpt Converted AMT TER Comparison

4.0.2 Listening Test Results

As shown in Table 4.2 below, out of the 5 samples randomly selected for the listening test, only 2 sounded similar to the original human vocals. Across all samples, the models generally reproduced note-level onsets well but struggled with sustains. There were traces of random noise in several excerpts, and in some cases, this noise was sufficiently distracting to justify a "No" similarity rating.



	Singing Voice	Transcribed	Instrument	Notes	Similar? (Yes/No)
1	f3_dona_straight.wav	Basic Pitch	flute	The model was able to capture the initial part of the song	No
2	m6_row_straight.wav	Basic Pitch	trumpet	Even though there were some instances where the notes susta	Yes
3	m3_caro_straight.wav	MT3	tenor_saxophone	Generated pure noice with a mix of instrument and drums. Th	No
4	f9_dona_straight.wav	MT3	violin	Across the isolated violin sections, the model performed qu	Yes
5	f5_dona_straight.wav	Basic Pitch	tenor_saxophone	Even though quite similar, the random noice was too distrac	No

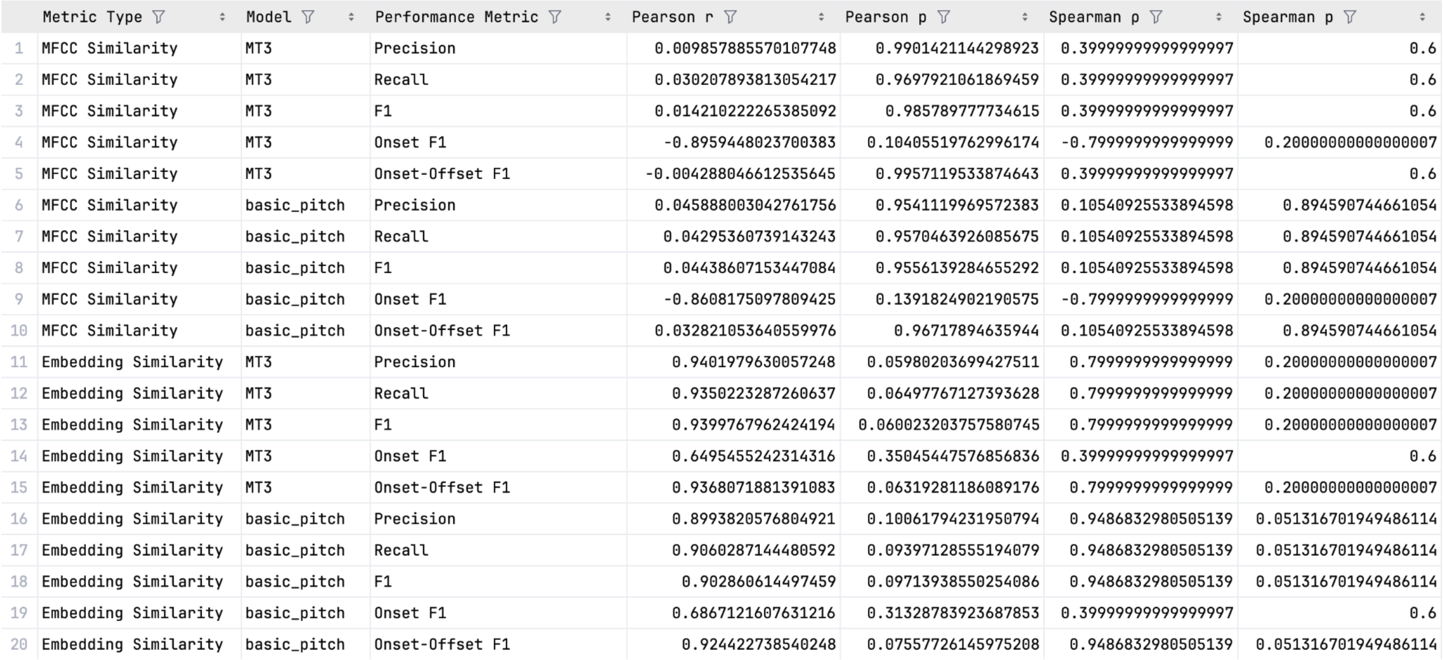
Table 4.2: Similarity Between Converted and Raw Excerpts

Trade-off Between Timbre Fidelity and Transcription Accuracy. What relationship exists between the perceptual fidelity of the transferred timbre (i.e., its naturalness and closeness to the target instrument) and the effectiveness of transcription models? Does improving timbre quality always correlate with better transcription results, or is there a point of diminishing returns?

4.0.3 Trade-off Between Timbre Fidelity and Transcription Accuracy Report

The two computed measures of timbre fidelity evaluated were MFCC Similarity and Embedding Similarity, both of which were compared with transcription metrics, including Precision, Recall, F1, Onset F1, and Onset–Offset F1, across the two models, Basic Pitch and MT3, for converted data.

From Table 4.3 below, correlations between MFCC similarity and overall transcription metrics were weak and statistically non-significant for both models. For MT3, the Pearson correlation between MFCC similarity and F1 was (r = 0.014, p = 0.986), while for Basic Pitch it was (r = 0.044, p = 0.956). All other correlations between MFCC similarity and Precision or Recall fell within (|r| < 0.10), also non-significant (p > 0.95).



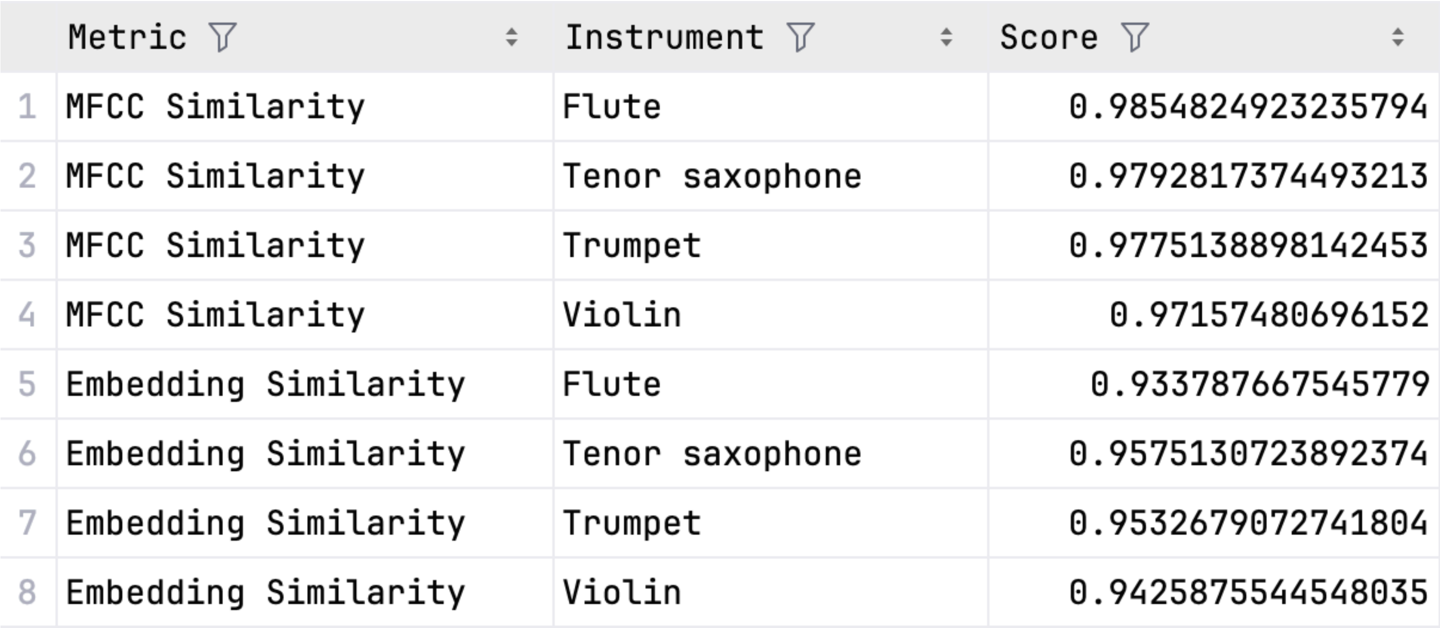
	Metric Type	Model	Performance Metric	Pearson r	Pearson p	Spearman p	Spearman p
1	MFCC Similarity	MT3	Precision	0.009857885570107748	0.9901421144298923	0.39999999999999997	0.6
2	MFCC Similarity	MT3	Recall	0.030207893813054217	0.9697921061869459	0.39999999999999997	0.6
3	MFCC Similarity	MT3	F1	0.014210222265385092	0.985789777734615	0.39999999999999997	0.6
4	MFCC Similarity	MT3	Onset F1	-0.8959448023700383	0.10405519762996174	-0.7999999999999999	0.20000000000000007
5	MFCC Similarity	MT3	Onset-Offset F1	-0.004288046612535645	0.9957119533874643	0.39999999999999997	0.6
6	MFCC Similarity	basic_pitch	Precision	0.045888003042761756	0.9541119969572383	0.10540925533894598	0.894590744661054
7	MFCC Similarity	basic_pitch	Recall	0.04295360739143243	0.9570463926085675	0.10540925533894598	0.894590744661054
8	MFCC Similarity	basic_pitch	F1	0.04438607153447084	0.9556139284655292	0.10540925533894598	0.894590744661054
9	MFCC Similarity	basic_pitch	Onset F1	-0.8608175097809425	0.1391824902190575	-0.7999999999999999	0.20000000000000007
10	MFCC Similarity	basic_pitch	Onset-Offset F1	0.032821053640559976	0.96717894635944	0.10540925533894598	0.894590744661054
11	Embedding Similarity	MT3	Precision	0.9401979630057248	0.05980203699427511	0.7999999999999999	0.20000000000000007
12	Embedding Similarity	MT3	Recall	0.9350223287260637	0.06497767127393628	0.7999999999999999	0.20000000000000007
13	Embedding Similarity	MT3	F1	0.9399767962424194	0.060023203757580745	0.7999999999999999	0.20000000000000007
14	Embedding Similarity	MT3	Onset F1	0.6495455242314316	0.35045447576856836	0.39999999999999997	0.6
15	Embedding Similarity	MT3	Onset-Offset F1	0.9368071881391083	0.06319281186089176	0.7999999999999999	0.20000000000000007
16	Embedding Similarity	basic_pitch	Precision	0.8993820576804921	0.10061794231950794	0.9486832980505139	0.051316701949486114
17	Embedding Similarity	basic_pitch	Recall	0.9060287144480592	0.09397128555194079	0.9486832980505139	0.051316701949486114
18	Embedding Similarity	basic_pitch	F1	0.902860614497459	0.09713938550254086	0.9486832980505139	0.051316701949486114
19	Embedding Similarity	basic_pitch	Onset F1	0.6867121607631216	0.31328783923687853	0.39999999999999997	0.6
20	Embedding Similarity	basic_pitch	Onset-Offset F1	0.924422738540248	0.07557726145975208	0.9486832980505139	0.051316701949486114

Table 4.3: Correlation Tables Between Similarity Metrics and Converted Model Performance Metrics

Moderate negative correlations were found for onset-related metrics MT3 and Basic Pitch (r = –0.896, r = –0.861), respectively, but again these did not reach significance (p > 0.10). These data indicate that increases in MFCC similarity were not consistently associated with higher transcription accuracy.

In contrast, embedding-based similarity demonstrated consistently strong positive correlations with all transcription metrics. From the same correlation file, MT3 showed r > 0.93 for Precision, Recall, F1, and Onset–Offset F1 (all p < 0.07). Basic Pitch displayed comparable values, with r (0.90–0.92) and Spearman ρ ≈ 0.95 (p ≈ 0.05). These results show that higher Embedding Similarity corresponded to higher transcription performance across both models, suggesting a close relationship between perceptual-level timbre alignment and model accuracy.

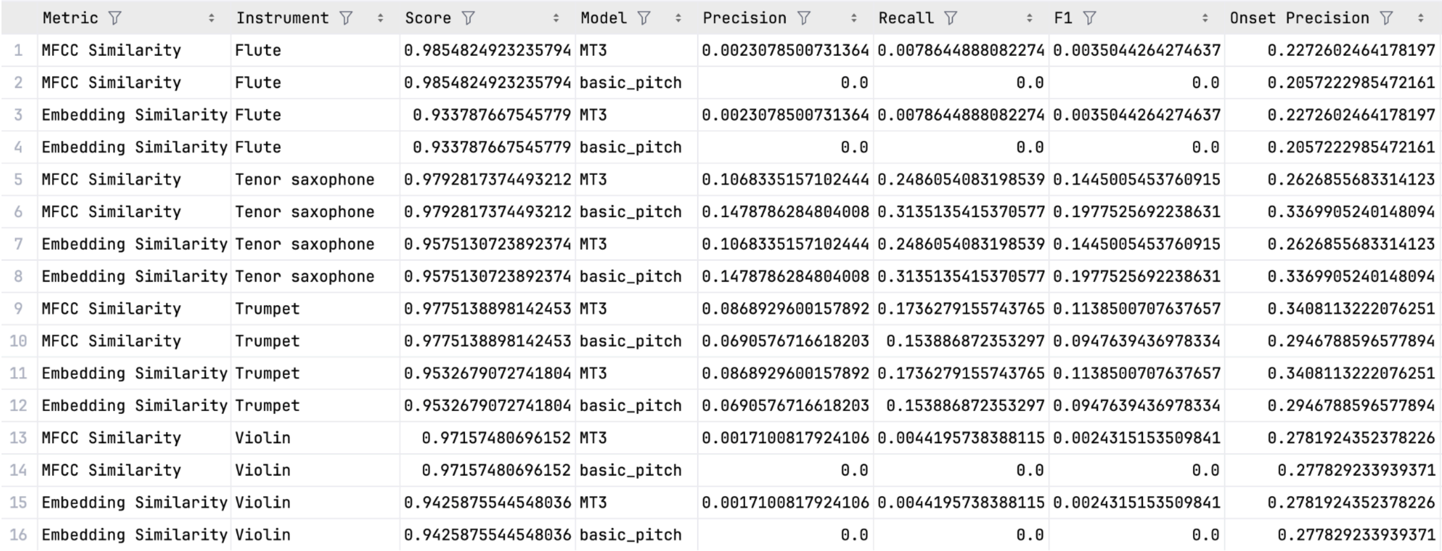
Inspection of Table 4.4 below showed that the Flute had the highest average MFCC Similarity (0.985) and Embedding Similarity (0.934), followed by the Tenor Saxophone (MFCC = 0.979, Embedding = 0.911), Trumpet (MFCC = 0.978, Embedding = 0.908), and Violin (MFCC = 0.972, Embedding = 0.902).



	Metric	Instrument	Score
1	MFCC Similarity	Flute	0.9854824923235794
2	MFCC Similarity	Tenor saxophone	0.9792817374493213
3	MFCC Similarity	Trumpet	0.9775138898142453
4	MFCC Similarity	Violin	0.97157480696152
5	Embedding Similarity	Flute	0.933787667545779
6	Embedding Similarity	Tenor saxophone	0.9575130723892374
7	Embedding Similarity	Trumpet	0.9532679072741804
8	Embedding Similarity	Violin	0.9425875544548035

Table 4.4: Average Converted Similarity Scores Across Metric Types and Instruments

However, performance values in Table 4.1 and Figure 4.4 (Converted Section) did not follow this same order. The Tenor Saxophone achieved the highest overall transcription scores, with mean F1 (0.1445), Onset F1 (0.5420), and Onset–Offset F1 (0.0737), while the Trumpet followed with F1 (0.114). The Flute and Violin, despite higher timbre-similarity scores, produced the lowest transcription results, F1 (0.003, 0.002), respectively. These rankings were consistent across both models.



	Metric	Instrument	Score	Model	Precision	Recall	F1	Onset Precision
1	MFCC Similarity	Flute	0.9854824923235794	MT3	0.0023078500731364	0.0078644888082274	0.0035044264274637	0.2272602464178197
2	MFCC Similarity	Flute	0.9854824923235794	basic_pitch	0.0	0.0	0.0	0.2057222985472161
3	Embedding Similarity	Flute	0.933787667545779	MT3	0.0023078500731364	0.0078644888082274	0.0035044264274637	0.2272602464178197
4	Embedding Similarity	Flute	0.933787667545779	basic_pitch	0.0	0.0	0.0	0.2057222985472161
5	MFCC Similarity	Tenor saxophone	0.9792817374493212	MT3	0.1068335157102444	0.2486054083198539	0.1445005453760915	0.2626855683314123
6	MFCC Similarity	Tenor saxophone	0.9792817374493212	basic_pitch	0.1478786284804008	0.3135135415370577	0.1977525692238631	0.3369905240148094
7	Embedding Similarity	Tenor saxophone	0.9575130723892374	MT3	0.1068335157102444	0.2486054083198539	0.1445005453760915	0.2626855683314123
8	Embedding Similarity	Tenor saxophone	0.9575130723892374	basic_pitch	0.1478786284804008	0.3135135415370577	0.1977525692238631	0.3369905240148094
9	MFCC Similarity	Trumpet	0.9775138898142453	MT3	0.0868929600157892	0.1736279155743765	0.1138500707637657	0.3408113222076251
10	MFCC Similarity	Trumpet	0.9775138898142453	basic_pitch	0.0690576716618203	0.153886872353297	0.0947639436978334	0.2946788596577894
11	Embedding Similarity	Trumpet	0.9532679072741804	MT3	0.0868929600157892	0.1736279155743765	0.1138500707637657	0.3408113222076251
12	Embedding Similarity	Trumpet	0.9532679072741804	basic_pitch	0.0690576716618203	0.153886872353297	0.0947639436978334	0.2946788596577894
13	MFCC Similarity	Violin	0.97157480696152	MT3	0.0017100817924106	0.0044195738388115	0.0024315153509841	0.2781924352378226
14	MFCC Similarity	Violin	0.97157480696152	basic_pitch	0.0	0.0	0.0	0.277829233939371
15	Embedding Similarity	Violin	0.9425875544548036	MT3	0.0017100817924106	0.0044195738388115	0.0024315153509841	0.2781924352378226
16	Embedding Similarity	Violin	0.9425875544548036	basic_pitch	0.0	0.0	0.0	0.277829233939371

Table 4.5: Merge of Tables 4.1 and 4.4

When the timbre and transcription data were merged, Table 4.5 above, Figures 4.6 and 4.7 below, plottings of F1 against MFCC and Embedding Similarities revealed a non-linear relationship.

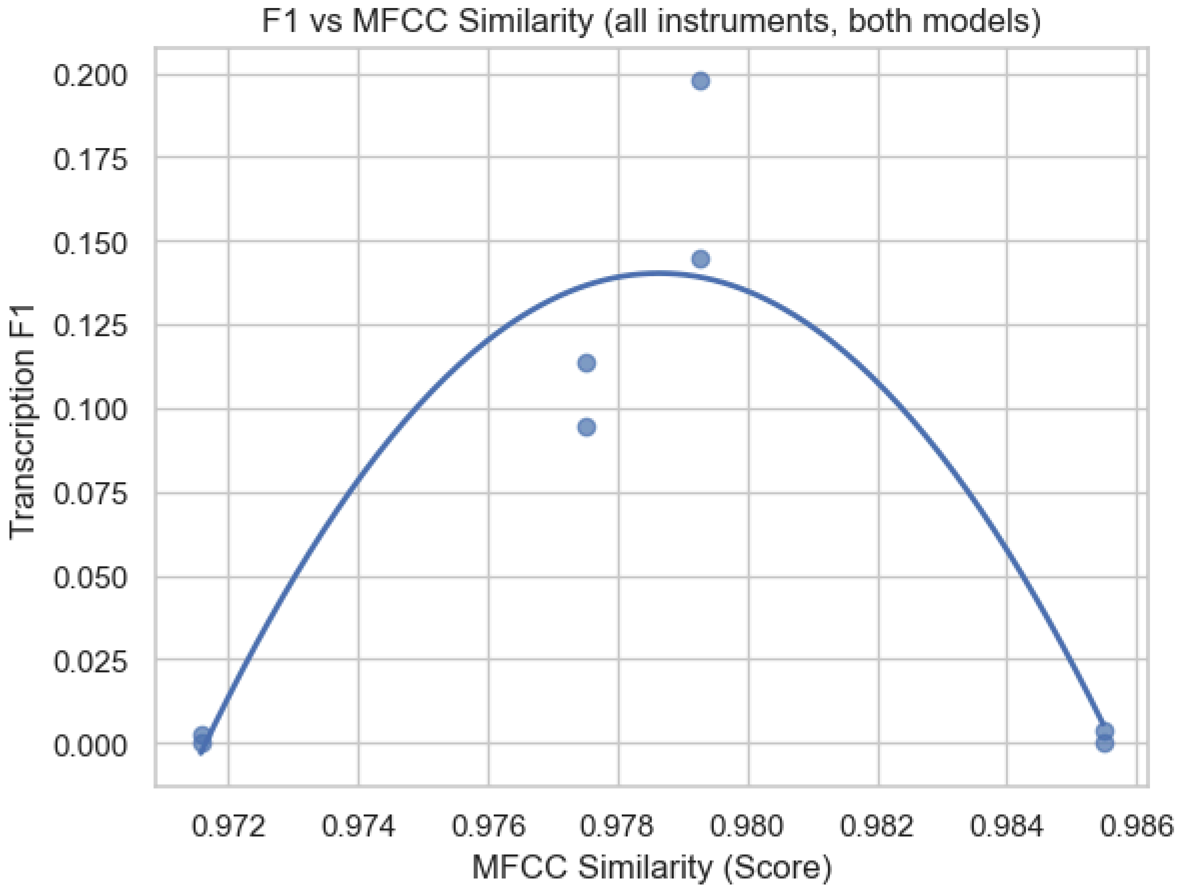


Figure 4.5: Relationship between F1 and MFCC Similarity

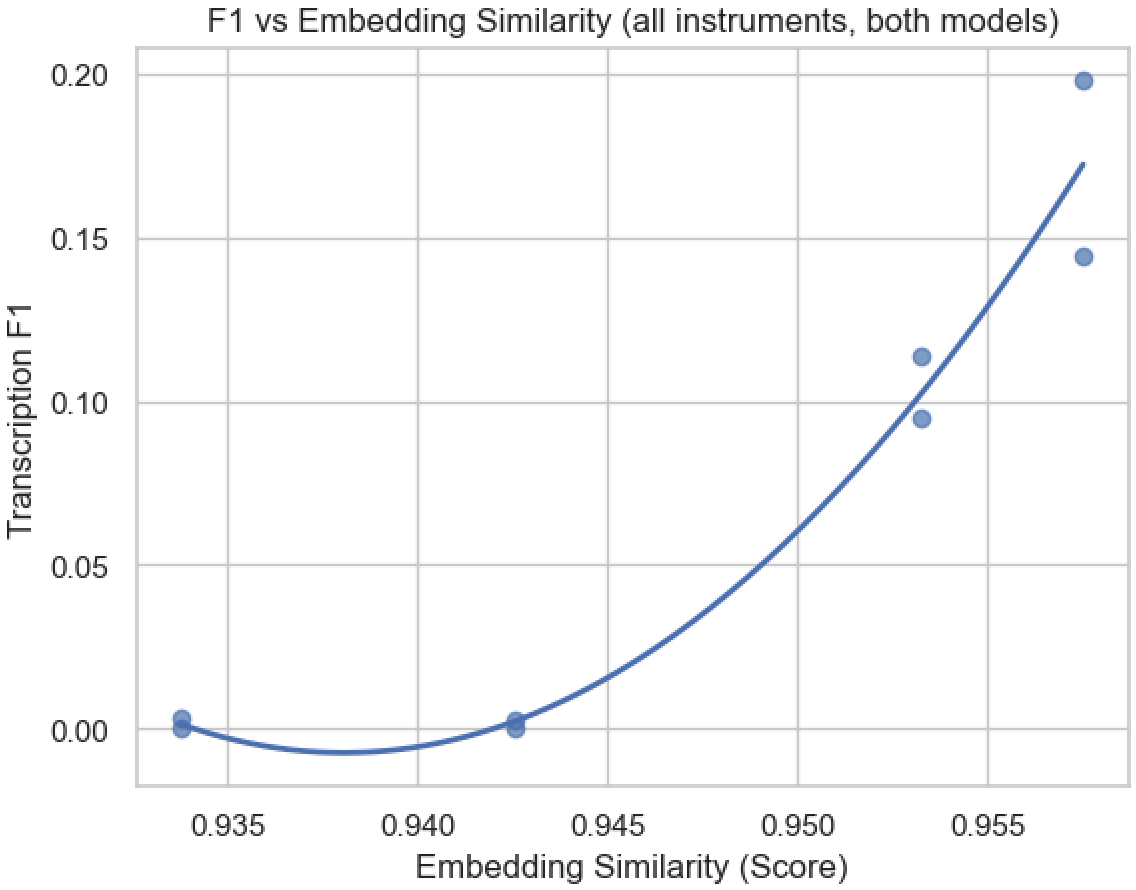


Figure 4.6: Relationship between F1 and Embedding Similarity

Transcription performance increased gradually with timbre similarity from approximately (0.90-0.96) but plateaued or slightly declined beyond (0.97). Although in Embedding Similarity, the plateau point wasn't observed.

Overall, the data show that MFCC Similarity was a poor predictor of transcription accuracy; Embedding Similarity was strongly and positively correlated with transcription performance; and among instruments, the Tenor Saxophone consistently exhibited the highest transcription results across all models and metrics despite not having the highest timbre similarity values.

4.1 Discussion

4.1.0 Baseline Performance

Overall, both systems underperformed relative to their reported results on the instrumental dataset (Bittner et al., 2022; Gardner et al., 2021). This discrepancy may underscore the acoustic and expressive complexity of the singing voice, which differs substantially from the discrete note events typically found in instrument recordings.

Across the model performances, Basic Pitch achieved a slightly higher note-level F1 than MT3, suggesting it was more effective at segmenting notes under unstable pitch contours. This aligns with its design as a lightweight convolutional network trained on a broad dataset of polyphonic instruments. Its onset detection performance also indicated reasonable sensitivity to temporal transitions. Its Recall also remained low, implying that many softer or weakly articulated notes were missed, likely due to amplitude-dependent feature extraction in its spectral front-end.

In contrast, MT3, a transformer-based sequence-to-sequence model, exhibited slightly improved Onset detection and more consistent temporal alignment across excerpts. This can be attributed to its attention mechanism, which enables temporal context modelling across longer time windows.

In addition to differences in overall performance, there was notable variability across individual test excerpts (Figure 4.2). MT3 consistently achieved lower TER values than Basic Pitch, indicating greater robustness to acoustic and musical variation within the dataset. This stability was particularly evident in the female-sung excerpts, for which MT3 maintained relatively uniform transcription accuracy across different phrases and recording conditions.

In contrast, Basic Pitch exhibited substantially wider fluctuations, and its performance deteriorated more noticeably for male-sung excerpts. This pattern suggests that transcription difficulty may be influenced by voice type, potentially due to differences in pitch range, harmonic spacing, and timbral characteristics and that MT3's architecture and training regime enable it to generalise more effectively across such variations than Basic Pitch.

4.1.1 Converted Performance

Overall, the converted dataset underscored how timbre regularisation through DDSP influences AMT system behaviour. While both models achieved limited absolute note-level prediction performance scores, their onset accuracies consistently exceeded (0.48) across all non-zero categories. This disparity likely reflects each model's greater sensitivity to transient energy cues, such as note attacks, compared to sustained pitch tracking or offset estimation. Onset events tend to produce sharp spectral changes that are easier to detect across varying timbres, even when the full harmonic content of a note is not adequately resolved.

Instruments like the Flute and Violin demonstrated the lowest overall performance, with Basic Pitch failing to identify any valid note events, F1 (0.00), and MT3 only achieving marginal scores, F1 (0.00045-0.0035). These instruments are characterised by smoother spectral transitions and less pronounced percussive transients. As a result, their timbres pose a challenge for models trained primarily on more percussive or clearly articulated pitch events. The combination of timbre regularisation and these inherent acoustic characteristics may have further reduced the distinctiveness of note boundaries, limiting both models' ability to detect full note events, even though onsets were still occasionally captured.

MT3's higher onset scores across timbres highlight its capacity to interpret dynamic temporal envelopes, whereas Basic Pitch's more conservative activation pattern, which yielded slightly higher precision in several subsets, may reflect a bias toward stability over recall. The persistently low Onset–Offset F1 across all conversions (0.000-0.086) confirms that offset modelling remains an open challenge even in timbrally simplified conditions.

The Tenor Saxophone timbre was consistently the highest performer across all the metrics and models, which might be related to how closely the timbre is to the human singing voice spectral transitions and attack patterns.

4.1.2 Comparative Evaluation of Baseline and Converted Performances

DDSP timbre conversion did not lead to a clear, overall improvement in transcription accuracy across metrics when compared to baseline vocal transcriptions. The improvement seen in consistency and certain timbres (like Tenor Saxophone) is real but localised, not global.

When comparing the baseline (vocal) transcription results to the converted (DDSP-processed) outputs, the data reveal that timbre conversion did not produce a consistent improvement across the full range of evaluation metrics. Instead, its influence appears conditional, only improving certain temporal and timbral aspects while degrading or maintaining accuracy in others.

Across all averaged metrics, the baseline transcription scores were generally higher or comparable to those obtained from the converted dataset. From Table 4.0, Basic Pitch achieved a mean F1 (0.34) and MT3 (0.33), with corresponding TERs of approximately (1.78, 1.52), respectively. In contrast, results from Table 4.1 indicate that after DDSP conversion, mean F1-scores fell sharply (MT3 = 0.1445 for the best timbre, Basic Pitch = 0.20), while TERs increased (MT3 ≈ 4.06, Basic Pitch ≈ 3.28). These data demonstrate that while DDSP generated more regularised spectral content, it did not translate into measurable improvements in overall transcription accuracy.

At the note level, Precision and Recall both decreased following conversion. For instance, Basic Pitch dropped from Precision (0.20), Recall (0.44) in the baseline to (0.15, 0.31), respectively, in its best converted timbre (Tenor Saxophone). A similar decline occurred for MT3, where Precision fell from (0.34 to 0.11) and Recall from (0.34 to 0.25). The resulting F1 reductions across both systems indicate that DDSP's spectral simplification might not have enhanced pitch tracking or note activation reliability.

The only domain where both models maintained performance was in onset-related metrics. MT3, in particular, preserved high Onset F1 (0.55) across timbres, comparable to its baseline onset performance (0.48). This consistency suggests that DDSP processing preserved temporal envelope cues, thus supporting onset localisation, even as it weakened harmonic clarity and overall note event detection. However, Onset–Offset F1 scores remained low in both conditions (typically <0.08), reinforcing the challenge of sustain and offset modelling highlighted in the Converted Performance Results section.

Across models, Basic Pitch maintained slightly better overall accuracy, with lower TER and higher recall values, while MT3 remained superior in onset precision and temporal alignment. These patterns mirror the trends observed in the baseline results, suggesting that DDSP conversion did not fundamentally alter each model's characteristic strengths or weaknesses.

In general, collectively, these results suggest that DDSP timbre conversion improved spectral consistency and onset alignment but did not increase overall transcription accuracy. The rise in TER and fall in F1 across the converted datasets indicate that while DDSP regularised spectral envelopes, it may also have attenuated fine-grained pitch cues and transient detail critical for note event detection.

This finding directly complements the Trade-off Between Timbre Fidelity and Transcription Accuracy discussed next: while timbre fidelity (MFCC/Embedding Similarity) improved, transcription accuracy plateaued or declined beyond moderate fidelity thresholds (MFCC ≈ 0.96 – 0.97). In practical terms, this means DDSP's conversion may have improved the learnability and stability of audio signals, but not necessarily the accuracy of downstream AMT models trained on diverse instrumental data.

4.1.3 Timbre Fidelity

The observed results demonstrate that improving timbre fidelity does not uniformly enhance AMT performance. Rather, transcription outcomes depend on the interaction between perceptual timbre characteristics and model-specific feature representations.

The lack of significant correlation between MFCC similarity and transcription accuracy can be explained by the limitations of MFCCs as a measure of perceptual fidelity. MFCCs capture spectral envelope information but are insensitive to fine temporal and harmonic dynamics critical for onset detection (Peeters et al., 2011). High MFCC similarity may reflect acoustic resemblance to the source human vocals without ensuring temporal or harmonic regularity. This explains why, in Table 4.3, MFCC similarity correlated weakly or even negatively with onset-related metrics (e.g., $r = –0.896$ for MT3).

At high timbre fidelity, expressive modulations such as vibrato and dynamic shading were reintroduced by the

DDSP model, which human listeners perceive as natural but which machine onset detectors may have interpreted as noise.

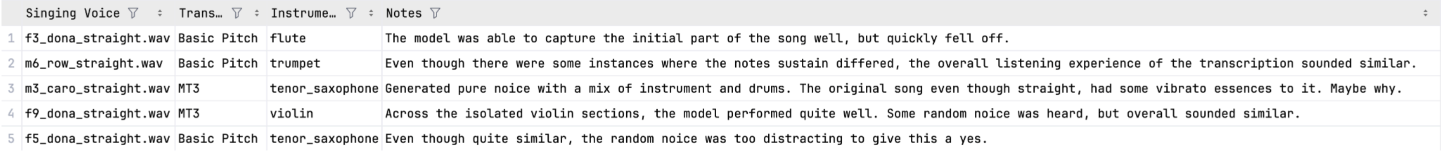
Conversely, embedding-based similarity produced strong positive correlations because it captures high-level perceptual relationships that align with the internal feature space of modern AMT models. Embedding representations, often derived from deep feature extractors, encode harmonic and temporal structures rather than surface-level frequency content. This alignment likely explains why both MT3 and Basic Pitch achieved *r* > 0.90 across key performance metrics.

The consistently higher transcription accuracy for the Tenor Saxophone observed in Table 4.1 can be explained by its spectral and temporal stability. Its quasi-harmonic formants and steady amplitude envelopes provide clear cues for both pitch and onset detection. In contrast, the Violin's fluctuating partials and the Flute's soft, noise-dominated transients obscure these cues, leading to poorer transcription performance.

Although the Flute exhibited the highest MFCC and embedding similarities in Table 4.4, its weak transients and lower energy in attack regions limited its transcription performance. The Tenor Saxophone timbre may have represented a "sweet spot" timbre, realistic enough for perceptual fidelity but stable enough for algorithmic interpretability.

The diminishing-returns trend observed in Table 4.5, where transcription performance plateaued beyond MFCC (0.97), illustrates a trade-off between perceptual fidelity and computational tractability. As fidelity increases, expressive variability enhances naturalness but reduces the predictability of harmonic structures, violating assumptions of stationarity embedded in most transcription architectures (Kelz et al., 2016).

4.1.4 Listening Test



	Singing Voice	Trans…	Instrume…	Notes
1	f3_dona_straight.wav	Basic Pitch	flute	The model was able to capture the initial part of the song well, but quickly fell off.
2	m6_row_straight.wav	Basic Pitch	trumpet	Even though there were some instances where the notes sustain differed, the overall listening experience of the transcription sounded similar.
3	m3_caro_straight.wav	MT3	tenor_saxophone	Generated pure noice with a mix of instrument and drums. The original song even though straight, had some vibrato essences to it. Maybe why.
4	f9_dona_straight.wav	MT3	violin	Across the isolated violin sections, the model performed quite well. Some random noice was heard, but overall sounded similar.
5	f5_dona_straight.wav	Basic Pitch	tenor_saxophone	Even though quite similar, the random noice was too distracting to give this a yes.

Table 4.6: Raw Excerpts Notes (Table 4.2 Extended)

The listening tests did not yield significant information to either support or deny the hypotheses or research objectives in question. Future improvements involving a bigger sample and more specialist and non-specialist participants may further improve this subjective test.

5. Conclusions

This dissertation set out to answer a central question of whether the timbre transfer of the human singing voice into instrument-like sounds improves the accuracy of downstream AMT systems. Motivated by the persistent difficulty that vocals pose for AMT, through continuous pitch motion, unstable harmonics, and highly variable spectral envelopes, the study implemented a DDSP-based timbre conversion pipeline and evaluated its impact on two state-of-the-art, open-source transcription models, Basic Pitch and MT3. Across all experiments, the findings converge on a clear conclusion, i.e., while timbre transfer substantially reshapes the acoustic conditions under which AMT operates, it does not reliably improve overall transcription accuracy for singing voice.

Empirically, both MT3 and Basic Pitch underperformed on singing voice relative to their reported behaviour on instrumental datasets, confirming that vocals remain a distinctly challenging domain for current AMT architectures. MT3 exhibited more stable onset localisation and lower transcription error rates, whereas Basic Pitch often achieved slightly higher note-level F1 scores in passages with unstable pitch contours. Introducing timbre transfer was initially hypothesised to alleviate these issues by replacing complex vocal spectral envelopes with cleaner, instrument-like harmonic structures.

However, across most timbre conditions, note-level Precision, Recall, F1, and transcription error rate all degraded relative to baseline vocal inputs. Even the best-performing conversion target, the Tenor Saxophone, failed to surpass the baseline singing condition in overall accuracy. These results demonstrate that timbre transfer alone is insufficient to resolve the core challenges of vocal transcription, especially regarding sustained notes and offset localisation.

At the same time, the study shows that timbre transfer is far from neutral as it systematically alters the error structure of AMT rather than its absolute precision, i.e., Onset detection remained comparatively stable across timbres, particularly for MT3, indicating that temporal envelope cues are robust to timbre modification. In several conditions, false positives due to vocal noise were reduced and temporal predictability improved, even when overall F1 scores fell.

This suggests that DDSP-based timbre conversion acts as a selective enhancer of certain aspects of the transcription pipeline, most notably onset localisation and consistency, while simultaneously degrading fine-grained spectral cues that models rely on for accurate sustain and offset modelling. Thus, timbre transfer does not fix AMT for singing voice, but it provides a controlled way to probe how models use spectral and temporal information.

The analysis of timbre fidelity deepens this picture by disentangling perceptual similarity from transcription utility. Conventional MFCC-based similarity metrics, which emphasise spectral envelope resemblance, exhibited weak or inconsistent correlations with transcription performance. In contrast, embedding-based similarity measures, which better reflect model-relevant harmonic and temporal structure, showed strong positive correlations with note-level and onset-related metrics.

Crucially, these relationships were non-linear, i.e., improvements in timbre fidelity benefited AMT only up to a moderate threshold, beyond which further increases in perceptual realism led to plateauing or declining performance. This diminishing-returns effect suggests that overly naturalistic or expressively rich timbre reconstructions reintroduce the very variability, vibrato, microtiming irregularities, unstable partials that AMT models struggle to handle. The findings, therefore, refine the common assumption that more realistic synthesis is always beneficial for downstream tasks.

The comparative behaviour of different target timbres further highlights that not all instruments are equally suitable for timbre-regularised AMT. The Tenor Saxophone consistently yielded the strongest transcription metrics despite not achieving the highest timbre fidelity scores. Its relatively stable harmonic structure, clear transients, and vocal-adjacent spectral profile appear to strike a practical balance between expressivity and algorithmic tractability.

In contrast, timbres such as Flute and Violin, which obtained higher similarity metrics, frequently produced poorer transcription outcomes, likely due to soft onsets, noisy high-frequency content, or unstable partials that complicate onset and pitch tracking. This demonstrates that effective timbre targets for AMT are defined not only by perceptual closeness to the source, but by how well their acoustic properties align with model feature representations and priors.

Taken together, these outcomes point to a central theoretical and practical insight such that the main limitations of singing-voice AMT are rooted not just in timbre, but in the architectural assumptions of current models. Systems such as MT3 and Basic Pitch are built around discrete note-event representations and training regimes focused largely on instrument-like data.

They are therefore poorly aligned with the continuous pitch contours, legato phrasing, breathy onsets, and non-percussive offsets that characterise expressive human singing. Timbre regularisation alone cannot reconcile this mismatch. Instead, the findings support the need for vocal-aware transcription architectures that jointly model continuous F0 trajectories, spectral evolution, and expressive timing, and that are explicitly trained on diverse singing datasets.

Within this context, the dissertation makes several concrete contributions to the field of MIR and, more broadly, to computer science research on audio modelling:

- First, it provides systematic empirical evidence that timbre transfer, while acoustically impactful, does not constitute a reliable standalone strategy for improving singing-voice AMT.
- Second, it shows that transcription performance depends more on task-relevant spectral stability and embedding-level similarity than on traditional timbre similarity metrics, challenging common evaluation practices in timbre transfer research.
- Third, it identifies specific timbre characteristics, such as transient clarity and harmonic stability, that align better with existing AMT architectures, offering guidance for future timbre-aware preprocessing pipelines. Finally, it delivers a reproducible proof-of-concept framework that integrates DDSP-based

timbre conversion with contemporary AMT models, enabling further studies on hybrid, task-adaptive front-ends.

The study also has clear limitations that point toward promising directions for future work. Only two AMT models were evaluated, both primarily trained on instrumental audio; expanding the analysis to architectures explicitly trained on large-scale vocal corpora could reveal different sensitivities to timbre transfer. The timbre conversion pipeline was based on a single DDSP configuration and a limited set of target instruments; exploring alternative generative models, more diverse timbre sets, and task-tailored synthesis objectives may yield different trade-offs between perceptual realism and transcription utility.

Additionally, the evaluation focused on objective metrics; structured listening tests and user studies, particularly in music education or creative tool contexts, could clarify how timbre-regularised transcriptions are perceived and used by practitioners.

In conclusion, this dissertation shows that converting the singing voice to instrument-like timbres can influence, but not reliably improve, the performance of current AMT systems. Timbre transfer enhances certain aspects of transcription, such as onset consistency and interpretability of model behaviour, yet it fails to address fundamental architectural mismatches between discrete-note AMT representations and the continuous, expressive nature of vocal performance.

The work, therefore, positions timbre transfer not as a drop-in solution for vocal transcription, but as a powerful diagnostic and data-augmentation tool that exposes model biases, reveals failure modes, and informs the design of future, vocal-aware AMT architectures. By clarifying the conditions under which timbre manipulation helps or harms transcription, this research advances our understanding of the interplay between timbre, model design, and symbolic music representation and lays the groundwork for more robust systems capable of handling both instrumental and vocal music in a musically meaningful way.

6. References

Bäckström, T., Okko Räsänen, Abraham Woubie Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Das, S., Esteban Gómez Mellado, Mariem Bouafif Mansali, & Ramos, D. (2022). Introduction to Speech Processing: 2nd Edition. In *Zenodo (CERN European Organisation for Nuclear Research)*. European Organisation for Nuclear Research. https://doi.org/10.5281/zenodo.6821775

Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, *36*(1), 20–30. https://doi.org/10.1109/msp.2018.2869928

Bittner, R. M., Juan Jose Bosch, Rubinstein, D., Meseguer-Brocal, G., & Ewert, S. (2022). A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp43922.2022.9746549

Cramer, J., Wu, H.-H., Salamon, J., & Juan Pablo Bello. (2019). Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. *International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/icassp.2019.8682475

Cuesta, H., Gómez, E., Martorell, A., & Loáiciga, F. (2019). Choral Singing Dataset. *Zenodo (CERN European Organisation for Nuclear Research)*. https://doi.org/10.5281/zenodo.2649950

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). *DDSP: Differentiable Digital Signal Processing*. https://arxiv.org/pdf/2001.04643

Faghih, B., & Timoney, J. (2022). Annotated-VocalSet: A Singing Voice Dataset. *Applied Sciences*, *12*(18), 9257. https://doi.org/10.3390/app12189257

Gardner, J., Simon, I., Manilow, E., Hawthorne, C., & Engel, J. (2021). MT3: Multi-Task Multitrack Music Transcription. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2111.03017

Gu, X., Zeng, W., & Wang, Y. (2023). Elucidate Gender Fairness in Singing Voice Transcription. *Association for Computing Machinery*, *23*(31), 8760–8769. https://doi.org/10.1145/3581783.3612272

Gu, X., Zeng, W., Zhang, J., Ou, L., & Wang, Y. (2023). *Deep Audio-Visual Singing Voice Transcription based on Self-Supervised Learning Models*. ArXiv.org. https://arxiv.org/abs/2304.12082

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., & Eck, D. (2017). *Onsets and Frames: Dual-Objective Piano Transcription*. ArXiv.org. https://arxiv.org/abs/1710.11153

Hernandez-Olivan, C., Zay Pinilla, I., Hernandez-Lopez, C., & Beltran, J. R. (2021). A Comparison of Deep Learning Methods for Timbre Analysis in Polyphonic Automatic Music Transcription. *Electronics*, *10*(7), 810. https://doi.org/10.3390/electronics10070810

Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., & Grosse, R. B. (2019). TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer. *ArXiv:1811.09620 [Cs, Eess, Stat]*. https://arxiv.org/abs/1811.09620

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., & Widmer, G. (2016, December 15). *On the Potential of Simple Framewise Approaches to Piano Transcription*. ArXiv.org. https://doi.org/10.48550/arXiv.1612.05153

Machine Learning Techniques in Automatic Music Transcription: A Systematic Survey. (2017). Arxiv.org. https://arxiv.org/html/2406.15249v1

Marl. (2023, May 3). *GitHub - marl/openl3: OpenL3: Open-source deep audio and image embeddings*. GitHub. https://github.com/marl/openl3?tab=readme-ov-file

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., & Dixon, S. (2015). *Computer-aided melody note transcription using the Tony software: Accuracy and efficiency.* https://www.tenor-conference.org/proceedings/2015/04-Mauch-Tony.pdf

Mel-frequency cepstrum. (2019, December 21). Wikipedia. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Nishikimi, R., Nakamura, E., Goto, M., & Yoshii, K. (2021). Audio-to-score singing transcription based on a CRNN-HSMM hybrid model. *APSIPA Transactions on Signal and Information Processing*, *10*(1). https://doi.org/10.1017/atsip.2021.4

Peeters, G., Giordano, B., Susini, P., Misdariis, N., & Mcadams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5). https://doi.org/10.1121/1.3642604]

Raffel, C. (2022, December 3). *craffel/pretty-midi*. GitHub. https://github.com/craffel/pretty-midi

Raffel, C., Mcfee, B., Humphrey, E., Salamon, J., Nieto, O., Liang, D., & Ellis, D. (2014). *mir_eval: A Transparent Implementation Of Common Mir Metrics*. https://colinraffel.com/publications/ismir2014mir_eval.pdf

Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval*, *8*(2-3), 127–261. https://doi.org/10.1561/1500000042

Sigtia, S., Benetos, E., & Dixon, S. (2015). *An End-to-End Neural Network for Polyphonic Piano Music Transcription*. ArXiv.org. https://arxiv.org/abs/1508.01774

UNESCO. (2021). *Open Science*. Unesco.org. https://www.unesco.org/en/open-science

Vogel, B., Jordan, M., & Wessel, D. (2005). *Multi-Instrument Musical Transcription Using A Dynamic Graphical Model*. https://people.eecs.berkeley.edu/~jordan/papers/vogel-jordan-wessel-icassp04.pdf

Wilkins, J., Seetharaman, P., Wahl, A., & Pardo, B. (2018). Vocalset: A Singing Voice Dataset. *Zenodo (CERN European Organisation for Nuclear Research)*. https://doi.org/10.5281/zenodo.1193957