Universiteit
Leiden
The Netherlands

Bachelor Computer Science

Gradual Soft Parameter Sharing with Attention
for Music Transcription

Yaell Brouwer

Supervisors:
Dr. E.M. Bakker & Prof. Dr. M.S. Lew

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

30/04/2026

Abstract

Multimodal music transcription, specifically the fusion of Optical Music Recognition (OMR) and Audio-to-Score (A2S), is a relatively new research field that continues to improve through recent developments in deep learning. However, the scarcity in annotated datasets and the inability to retrain a Deep Neural Network (DNN) means that the benefits of deep learning can not be optimally applied to multimodal music transcription. To alleviate the demands on the scarce annotated datasets and make use of several recent developments in deep learning, we propose a novel DNN with Multi-Task Learning (MTL). Specifically, we propose the novel GradSPSA architecture with gradually soft parameter sharing adapted for an encoder and for unidirectional usage in combination with dense connections and attention. Here, the dense connections and attention are used to lessen the observed performance gap between the two tasks. The novel GradSPSA model is evaluated on nine unique scenarios and showed an average improvement in Symbol Error Rate (SER) of 37.2% and up to 86.2% for an individual scenario. Here, GradSPSA is trained on a 36% subset, and validated and evaluated on 12% subsets of the cleansed Camera-PrIMuS dataset compared to the baseline, where the baseline is trained on a 60% subset, and validated and evaluated on 20% subsets of the cleansed Camera-PrIMuS dataset.

Contents

1 Introduction

Multimodal music transcription is a relatively new research field, because most music transcription studies focusing on unimodal models. However, combining multiple modalities, such as sheet music and audio recordings, has the potential to improve multimodal music transcription models [17]. Here, sheet music, also known as a score, is the input for Optical Music Recognition (OMR), while audio recordings are the input for Audio-to-Score (A2S) and the closely related Automatic Music Transcription (AMT) [4, 49]. Furthermore, through developments in deep learning, the field of music transcription has continued to improve [7, 34]. Multi-Task Learning (MTL), a subfield of deep learning, has been used in unimodal OMR [44], A2S [36], and AMT [38], and in multimodal audio-visual models [67, 2, 12]. However, as far as we know it has not been used in multimodal music transcription.

Due to the scarcity in annotated A2S and AMT datasets [6], MTL could tip the scale in favor of deep multimodal music transcription due to its ability to share knowledge and reduce computational resources needed [65]. Based on [4], [40], and [64], a novel model is proposed which adapts the state-of-the-art multimodal late fusion model with gradual soft parameter sharing, a MTL technique, in combination with an interaction block and dense connections, where attention and the dense connections are used to lessen the performance gap observed between OMR and A2S [3, 4].

In short, a deep multimodal music transcription model named Gradual Soft Parameter Sharing with Attention (GradSPSA) is introduced and investigated in this work to gain insight in the effect of gradual soft parameter sharing on the state-of-the-art multimodal music transcription model [4]. The proposed model, given a score and its corresponding audio recording that is preprocessed into a spectrogram, is able to transcribe music. The main contributions of this work are:

1. A novel DNN named GradSPSA is proposed based on the work of [4], [40], and [64], where the model is improved by using gradual soft parameter sharing, attention, and dense connections. Specifically, we added dense connections and an interaction block to the less performing task, and adapted gradual soft parameter sharing for an encoder and for unidirectional usage.

2. The novel GradSPSA is trained on a 36% subset, and validated and evaluated on 12% subsets of the cleansed Camera-PrIMuS dataset, and can outperform the baseline, where the baseline is trained on a 60% subset, and validated and evaluated on 20% subsets of the cleansed Camera-PrIMuS dataset.

The rest of the paper is organized as follows: Section 2 discusses the related work; Section 3 describes the methodology; Section 4 describes the experiments; Section 5 discusses the results of the experiments; Section 6 is the discussion of the results in a bigger context; Section 7 is the conclusion of this thesis and describes the future work.

2 Related Work

In the current music transcription literature, most studies focus on unimodal OMR, A2S, and AMT.

In OMR, several state-of-the-art developments rely on machine learning techniques such as deep learning or neural networks (NNs)[7]. Depending on the used technique, different approaches are used. The holistic approach, such as in [10], is used within OMR to transcribe entire sections of music

at once, and is often based on Connectionist Temporal Classification (CTC) or Image-to-Sequence [7]. A challenge for this approach is the transcription of more complex music such as homophonic, pianoform, or polyphonic music instead of monophonic music [54].

On the other hand, the pipeline-based approach, such as in [5], transcribes the music in multiple stages [7]. In some pipeline-based studies Hidden Markov Models (HMM) are used to skip the segmentation and staff line removal part of the pipeline [54]. Nonetheless, music object detection remains a challenge for this approach [7].

The end-to-end approach merges the pipeline-based and the holistic approach [7]. In [45], an end-to-end transformer architecture is used but due to the scarcity of training data a Recurrent Neural Network (RNN) is preferred over the transformer architecture.

Furthermore, handwritten scores remain a challenge regardless of the used approach due to the tilted or curved lines, skewed images, and degradation of the paper [7, 54].

The other task of the proposed model is Audio-to-Score (A2S), and is closely related to Automatic Music Transcription (AMT). In [49], the distinction between AMT and A2S is characterized by the model's output. Traditional AMT output is intended as an intermediary step and is non-human readable, e.g. MIDI [49]. On the other hand, A2S is intended to have a more human readable output, e.g. MusicXML [49]. Notable A2S papers are the holistic-based model in [36], and the end-to-end models in [48] and [47].

Since the field of A2S is relatively new, AMT is also included in this section [56]. Previous AMT studies relied on Non-negative Matrix Factorization (NMF) and neural networks [26]. These neural networks are often used in polyphonic music, a type of relatively complex music, but other approaches are also explored in papers such as [38], which uses Multi-Task Learning (MTL) techniques in combination with Deep Neural Networks (DNNs) [26]. In order to use a MTL technique, signal sources separation is often required before the AMT technique is applied [26]. An example of such signal processing method is NMF, which is simple and fast [26]. However, DNNs can achieve a higher accuracy on certain instruments [26], can represent complex manifolds in a robust and relatively efficient way [6], and can be trained in an end-to-end fashion [6]. Therefore, it is one of the most popular approaches in recent developments [34].

In AMT and A2S DNNs, Long Short-Term Memory (LSTM) layers are used to compactly model the spectral changes in a note, which is necessary due to note decay across input frames [6]. These spectral properties in the form of spectrograms are the preferred input for models such as the at the time state-of-the-art methods Onsets and Frames [22, 26]. However, the scarcity of annotated datasets, the unavailability of datasets for many types of instruments, and the inability to re-train, finetune, or adapt (D)NNs make NMFs still so popular [6].

Besides unimodal methods, there are also methods that combine OMR and A2S into a multimodal fusion. Here, the fusion can be either on an early fusion or late fusion level[17].

The late fusion model in [4] compares multiple methods, but only the Minimum Bayes Risk (MBR) and confusion networks (CN) methods showed notable improvements with respect to the unimodal system. In [3], the early fusion model reduces the error rate for the less-performing task with the usage of the OMR architecture and performed less overall with the A2S architecture. However, these early fusion models cannot outperform the unimodal frameworks [3]. Furthermore, the late fusion model in [3] is not deemed worth the effort by the authors of the paper due to the only slight improvement under the condition that the unimodal and multimodal framework do not differ greatly in performance.

Papers such as [3], [4], and [17] are all part of the MultiScore project, which aims to make

contributions to OMR, A2S, and AMT with the end goal to develop neural models that combine OMR and A2S/AMT in an end-to-end fashion [8].

Audio-visual models use the same modalities as the proposed model and have been used in (automatic) speech recognition[1]. In [37], the conformer model has a Word Error Rate (WER) of 1.2% and performs well with a high level of noise, which is in essence a low Signal-to-Noise Ratio (SNR). Both SNR and training time influence the WER, where a less noisy audio, i.e. a high SNR, and longer training time both decrease the WER [1, 37]. Another notable paper is [41] due to its classification rate of 98% and its ability to outperform the unimodal audio model with a low SNR. The model in [19] can represent the audio-visual features in an unified manner and aligns them better than modal-independent models, but loses this ability gradually when the model complexity increases.

Multimodal audio-visual architectures are also used in other fields, such as speech enhancement [24], emotion recognition [66], and sentiment analysis [2]. In [66], the model uses a deep belief network to learn complex feature representations from low complexity features in order to fuse the audio-visual features. Another notable audio-visual transformer model with an additional textual task is trained and fine-tuned in two phases. The parameters of the text modality are frozen in order to protect it from its performance gap with the audio and visual modalities during attention [64]. These are then used in the second phase to fine-tune multiple tasks [64]. Combining these tasks is possible due to the transformer's ability to jointly learn from different representation subspaces [61, 64]. Attention is used in another way in the hybrid CTC-attention model in [42], namely to find an alignment between the input features and the output sequences, while CTC is used to prevent non-sequential alignments. These attention blocks often replace a RNN layer [61].

Multi-Task Learning is mentioned in both [3] and [26], multimodal and unimodal respectively, as a viable research approach for a music transcription due to its ability to train the model in an end-to-end manner. On top of this, it also has the ability to share knowledge, reduce computation resources needed, and reduce overfitting [65]. Specifically its ability to share knowledge helps with the demands on the scarce datasets [6].

In deep MTL, the architectures are generally divided into two categories: hard and soft parameter sharing architectures [50]. In hard parameters sharing architectures, the shallow layers share identical parameters, while the deeper layers have their own heads with their own parameters [65]. On the other hand, soft parameter sharing architectures have separate shallow layers and uses feature propagation techniques such as fusion, aggregation, and attention [65].

Although MTL has not yet been used in multimodal music transcription models, it has been used in unimodal models [35]. Notable examples of MTL in music transcription are the hard parameter sharing Cerberus [38] and the soft parameter sharing improved version of the Onsets and Frames methods [29].

Soft parameter sharing is preferred over hard parameter sharing for tasks with looser connections between modalities [62]. In [51], loosely related tasks are described as heterogeneous tasks, i.e. tasks that only share a part of their input variables, while more related tasks are homogeneous tasks. In [28], the connection between tasks is defined as more related if the tasks have a common ancestor.

Furthermore, soft parameter sharing is also used in other research fields, such as active speaker detection [67] and sound event detection [32].

OMR and A2S need a different type of input data, sheet music and audio recordings, respectively. This means that most multimodal research resort to using multiple or custom datasets [14]. For unimodal OMR research, commonly used datasets are Camera-PrIMuS [9], GrandStaff [46], and

DoReMi [53]. Unimodal AMT and A2S research commonly use the MAESTRO [23] and the MAPS [15] dataset. Here, the MAESTRO dataset is preferred over MAPS due to the MAPS dataset including synthesized audio, while MAESTRO consists of live performances [26]. Besides unimodal datasets, there are also a few suitable multimodal datasets, such as MUSCAT [18] and Camera-PrIMuS [9]. MusicNet is another multimodal dataset, however the audio and score do not align completely accurately [26].

The late multimodal fusion model in [4] is the only multimodal music transcription architecture as far as we know. Therefore, it forms the basis of this work. We use the Camera-PrIMuS dataset since this is the dataset used in the baseline [4]. Furthermore, we looked into MTL specifically since it is mentioned in both [3] and [26] as a viable research approach for music transcription. Here, specifically the gradually soft parameter sharing from [40] is used since soft parameter sharing is suitable for the baseline's architecture. Lastly, the interaction block from [64] is used in this work to lessen the performance gap observed in the preceding work of the baseline [3].

3 Methodology

Overview of the baseline

The complete architecture of the baseline [4] is shown in Figure 1a, and consists of four components: both DNNs, word graph creation, and late multimodal fusion. Before a given musical score and the corresponding audio recording are fed into their respective DNN to learn the task's representation, both are preprocessed. Then these are fed into their respective DNN, followed by each modality of the input pair being fed into the word graph creation module, after which the two word graphs are fused together in the late multimodal fusion module.

Dataset cleansing & adaptation

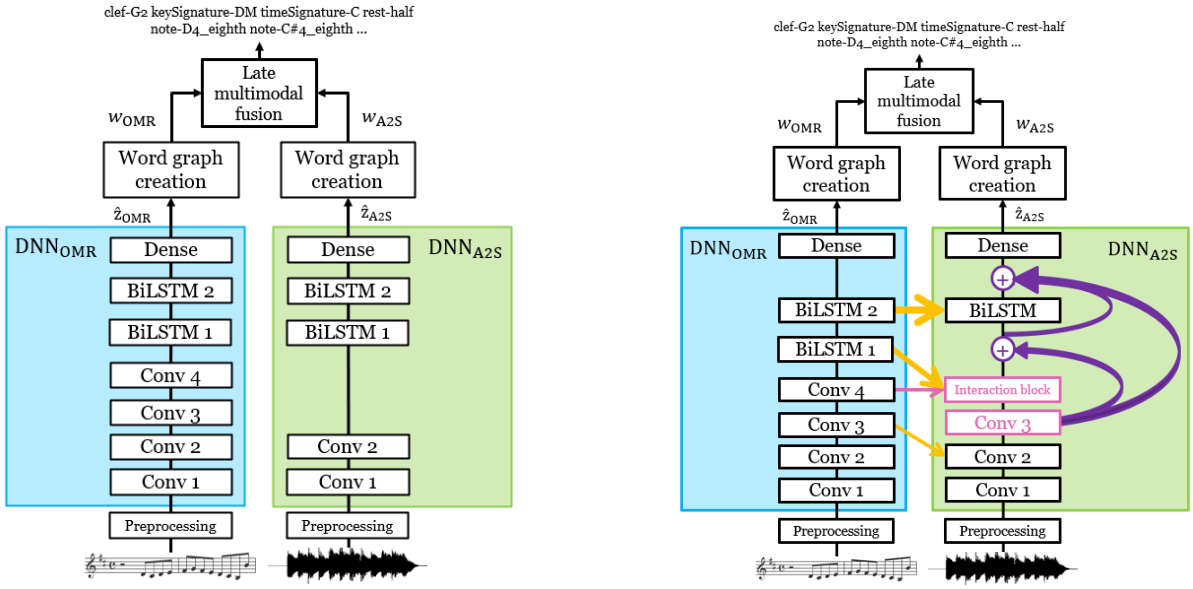
In order to obtain sample pairs that can be fed into the DNNs, the used dataset first needs to be cleansed and adapted due to the Camera Printed Images of Music Staves (Camera-PrIMuS) dataset being originally meant for unimodal OMR [9, 4]. The cleansing process as described in [4] primarily removes samples containing long rests, since these will span a large number of frames in the audio recordings, but contribute minimally to the score image [4]. Following this, the adaptation process converts the semantic files present in the dataset into MIDI with the converter the authors of the Camera-PrIMuS dataset provided [11]. Then, the obtained MIDI files are synthesized with the FluidSynth software [39] and a piano timbre [57] with a sampling rate of 22,050 Hz [4]. This results in 22,285 sample pairs divided into 60% training partition, 20% validation partition, and 20% test partition [4] as shown in Table 1. Furthermore, signal sources separation, which is normally necessary for MTL, is not needed due to the monophonic nature of the dataset [9]. While it is common in multimodal research to use a custom or adapted dataset [14], the cleansing process as described in [4] is not a common approach.

Cleansed Camera-PrIMuS dataset		
Partition	Sample pairs	
	Full dataset	60% of the dataset used in our experiments
Train (60%)	13,371	8,023
Validation (20%)	4,457	2,674
Test (20%)	4,457	2,674
Total	22,285	13,371

Table 1: The cleansed Camera-PrIMuS dataset has train, validation, and test partitions corresponding to 60%, 20%, and 20%, respectively, of the dataset.

Then, the resulting dataset samples are further preprocessed. For A2S, the audio features are extracted by obtaining a Constant-Q Transform (CQT) representation [52] with 512 samples between successive columns, 120 frequency bins, and 24 bins per octave [4]. This representation is transformed into a spectrogram, scaled to a height of 256 pixels while keeping the aspect ratio, and converted to grayscale [4].

For OMR, the image features are scaled to a height of 64 pixels while keeping the aspect ratio, and converted to grayscale [4].



(a) The baseline architecture

(b) The proposed GradSPSA architecture

Figure 1: The complete baseline and GradSPSA architectures compared to each other, where Conv represents a convolutional block, BiLSTM represents the bidirectional Long Short-Term Memory unit, Dense is a fully-connected layer, interaction block represents the transformer layers Multi-Head Attention, Add & Norm and Feed-Forward, $\oplus$ and the purple arrows represent dense connections, and $\hat{z}_{\text{OMR}}$ and $\hat{z}_{\text{A2S}}$ represent the estimated output sequences for DNN_{OMR} and DNN_{A2S}, respectively. In Figure 1b, the orange arrows represent unidirectional gradual soft parameter sharing and its thickness denotes the strength of the penalty for the difference in representation.

Baseline DNN_{OMR}

For DNN_{OMR}, the preprocessed input moves first through two convolutional blocks each consisting of a 2D convolutional layer with 64 filters, kernel size 5, and stride 1, followed by batch normalization, a leaky rectified linear unit (LeakyReLU) activation, and a 2D max pooling layer. Here, the first 2D max pooling layer has pooling size (2, 2) and stride (2, 2), while the second 2D max pooling has pooling size (2, 1) and stride (2, 1). Then, it moves through two convolutional blocks each consisting of a 2D convolutional layer with 128 filters, kernel size 3, and stride 1, followed by batch normalization, a LeakyReLU activation, and a 2D max pooling layer with pooling size (2, 1) and stride (2, 1). Followed by two bidirectional LSTM (BiLSTM) layers with 256 units and 0.5 dropout. Lastly, it moves through a fully connected layer. The posteriors, i.e. $\hat{z}_{\text{OMR}}$, are then used to create a word graph representation using Chen and Goodman's modified Kneser-Ney smoothing [13].

Baseline DNN_{A2S}

The baseline DNN_{A2S} moves the preprocessed input first through a 2D convolutional layer with 8 filters, kernel size (10, 2), an stride 1, followed by batch normalization, a LeakyReLU activation, and a 2D max pooling layer with pooling size (2, 2) and stride (2, 2). Then, it moves through a 2D convolutional layer with 8 filters and kernel size (8, 5), batch normalization, a LeakyReLU activation, and a 2D max pooling layer with pooling size (2, 1) and stride (2, 1). Then for the baseline, it moves further through two BiLSTM layers with 256 units and 0.5 dropout, followed by a fully connected layer. The posteriors, i.e. $\hat{z}_{\text{A2S}}$, are then used to create a word graph representation using Chen and Goodman's modified Kneser-Ney smoothing [13].

Fusion methods

The word graphs for both tasks are fused together using the Kaldi toolkit [43] and the SRILM toolkit [58] to create the late multimodal fusion using the four methods present in the baseline [4]: Minimum Bayes Risk (MBR), confusion networks (CN), Smith-Waterman (SW) local alignment, and lightly-supervised (both lightly-supervised OMR and lightly-supervised A2S) [16].

Evaluation metric

In order to gain insight into the effect of the individual DNN_{OMR} and DNN_{A2S} on the final fusion, the partitions are divided into high (H), medium (M), and low (L) Symbol Error Rate (SER) scores of the respective DNN, which is the performance metric used in the baseline. This creates nine unique scenarios where a high SER refers to an approximately 30% SER, a medium SER refers to an approximately 20% SER, and a low SER refers to an approximately 10% SER. This is also where the performance gap influences the scenarios, since the scenario partitions are constrained by the less performing DNN, which in this case is DNN_{A2S}.

The SER metric is expressed in mathematical terms in Equation 1, where $\mathcal{S}$ denotes a set of test data from either the OMR or A2S task, ED denotes the string edit distance [33], and $\hat{z}_i$ and z_i represent the estimated and ground-truth sequences, respectively.

$$SER(\%) = \frac{\sum_{i=1}^{|\mathcal{S}|} ED(\hat{z}_i, z_i)}{\sum_{i=1}^{|\mathcal{S}|} |z_i|} \tag{1}$$

Lastly, both DNNs are trained for 150 epochs using a CTC loss with an ADAM optimizer [30].

Overview of GradSPSA

The complete architecture of the GradSPSA is shown in Figure 1b, and consists of four components: both DNNs, word graph creation, and late multimodal fusion. Compared to the baseline [4], DNN_{OMR} is kept the same, while DNN_{A2S} has been adapted due to the observed performance gap between the

two DNNs in the preceding work of the baseline [3], where DNN_{A2S} performs worse than DNN_{OMR}. Furthermore, GradSPSA uses a 60% subset of the cleansed Camera-PrIMuS dataset as shown in Table 1. While it is not a common approach in other studies to use a subset of the dataset, this decision has been made due to the time constraint of this work.

GradSPSA's DNN_{A2S}

While the baseline's DNN_{A2S} only has two convolutional blocks, in GradSPSA there is an additional third convolutional block, which consists of a 2D convolutional layer with 8 filters and kernel size 3, followed by batch normalization, a LeakyReLU activation, and a 2D max pooling layer with pooling size (1, 2) and stride (1, 2). This convolutional block is inspired by the baseline's DNN_{A2S} [49] and its base model described in [55], which uses pooling size (1, 2) due to its benefit of recognizing narrow shapes. Then, it is followed by a fully connected layer. Here, the third convolutional block is used to downsample the features to make the memory usage of the multi-head attention (MHA) unit in the interaction block more manageable. Furthermore, the output of the fourth OMR convolutional block, i.e. the fourth OMR 2D max pooling layer output, is followed by a fully connected layer. The output of the fully connected layer for both tasks are fed into a MHA layer with 2 heads, 64 query and key dimension, and a dropout of 0.1. This is followed by Add & Norm block, a feed forward network, and another Add & Norm block. This interaction block is based on [64] and is used to jointly learn from different representation subspaces [61]. Specifically the first BiLSTM layer is replaced instead of the second due to the attention mechanism used in the interaction block enabling later BiLSTM layers, in this case only the second A2S BiLSTM layer, to receive feature sequences related to the output [60]. Note that positional encoding, which is common for the transformer architecture, is not necessary for GradSPSA's DNN_{A2S} due to the presence of convolutional and recurrent layers in the proposed model [31].

Starting from the third convolutional block, the layer after which the features are not further down sampled [27, 21], dense connections are used in GradSPSA's DNN_{A2S} to lessen the performance gap between the two tasks by increasing variation in the input of subsequent layers and improving efficiency [25].

Following the first dense connection, is a BiLSTM layer with 256 units and 0.5 dropout. After this layer, another dense connection is made before it moves through a fully connected layer.

Gradual soft parameter sharing

A component unique to the GradSPSA's DNN_{A2S} is the gradual soft parameter sharing loss added to the A2S CTC loss from the baseline architecture. Here, soft parameter sharing is used due to OMR and A2S both being derived from the Plaine and Easie Code (PAEC) file present in the dataset [9]. This means in essence that OMR and A2S are related but only loosely due to having different representations.

Inspired by [40], the proposed GradSPSA uses unidirectional gradual soft parameter sharing to lessen the performance gap, i.e. improve A2S while protecting OMR from adverse effects, and lessen the model's dependency on a large amount of data. The inspiration [40] transitions gradually from a hard parameter sharing in the encoder to a soft parameter sharing in the decoder in a bidirectional manner. Since the inspiration uses an encoder-decoder model [40], this method needs to be adapted for the encoder-only architecture of the proposed model [4]. Furthermore, the method also needs to be adapted for unidirectional instead of bidirectional gradual soft parameter sharing. The hard parameter sharing needs to be removed since the tasks are only loosely related. Lastly, the layers suitable for soft parameter sharing need to be selected since not all of them are suitable due to their representation.

Adapting the method for unidirectional gradual soft parameter sharing means that the gradual soft parameter sharing loss is only added to A2S, which in turn means that DNN_{OMR} needs to be trained separate from and before DNN_{A2S}. This also makes it unnecessary to freeze DNN_{OMR}'s parameters when these are used in DNN_{A2S}, even though OMR normally would need to be protected from the performance gap [64].

Loss function

The original gradually soft parameter sharing loss as described in [40] is expressed in mathematical terms in Equation 2, where N denotes the number of layers and θ_n^T denotes the decoder's parameters for the n-th layer and task T with $(\text{OMR}, \text{A2S}) \in T$, and $||\cdot||^2$ denotes the squared Frobenius norm as used in [40], which penalizes the distance between the tasks' parameters.

$$\mathcal{L}_{GS_{dec}} = \gamma \sum_{n=1}^{N-1} (e^{\frac{N-n}{N}} - 1)||\theta_n^{\text{OMR}} - \theta_n^{\text{A2S}}||^2 \tag{2}$$

The translation of the encoder-decoder-based loss to an encoder-based loss is characterized by reflecting the slope over the x-axis, adding the constant $e - 1$ to keep the values in the same range as the original loss function, and by starting n at index 2 in order to let the first layer be entirely task-specific due to the feature representation of the first layer being still too low-level for soft parameter sharing to have a positive influence [40]. This entire step is expressed in mathematical terms in Equation 3.

$$\mathcal{L}_{GS_{enc}} = \gamma \sum_{n=2}^{N-1} (-e^{\frac{N-n}{N}} + e)||\theta_n^{\text{OMR}} - \theta_n^{\text{A2S}}||^2 \tag{3}$$

To remove the hard parameter sharing from the loss function, the variable S replaces the variable N and denotes the number of shared layers. Since S is the index for a set of layers, s will start at index 1 and $S - 1$ will be replaced with S. Furthermore, θ_{s_T} denotes the s-th layer's parameters for task T for the encoder. In Equation 4, this is expressed in mathematical terms.

$$\mathcal{L}_{GS_{soft,enc}} = \gamma \sum_{s=1}^{S} (-e^{\frac{S-s}{S}} + e)||\theta_{s_{\text{OMR}}} - \theta_{s_{\text{A2S}}}||^2 \tag{4}$$

The layers in S are based on two factors: feature representation level and number of layers in the DNN_{OMR} and DNN_{A2S}. Due to the feature representation being still too low level for soft parameter sharing to have a positive influence, the first layer for both tasks are excluded from S [40]. To keep the last layer for both tasks entirely task-specific, the last layer for both tasks are also excluded from S [20].

Furthermore, the number of remaining layers for both tasks are not the same. This means that one extra layer in DNN_{OMR} is also excluded. Specifically the second convolutional OMR layer is excluded, since all other layers have their own counterpart taking into account that the deeper layers are preferred since these have a higher feature representation.

On top of this, the fourth convolutional OMR layer and third convolutional A2S layer are also excluded since they are already included in the interaction block's parameters. This results in the parameters from OMR layers Conv 3, BiLSTM 1, and BiLSTM 2 being in $\theta_{S_{\text{OMR}}}$, and the parameters from A2S layers Conv 2, interaction block, and BiLSTM being in $\theta_{S_{\text{A2S}}}$.

This results in the gradually soft parameter sharing loss described in Equation 4 being added to the original A2S CTC loss as shown in Equation 5, while the OMR CTC loss remains the same as in the multimodal late fusion baseline [4].

$$\mathcal{L}_{\text{A2S}} = \mathcal{L}_{\text{A2S}} + \mathcal{L}_{GS_{soft,enc}} \tag{5}$$

Furthermore, GradSPSA uses Witten-Bell smoothing [63] during the word graph creation stage due to the small size of the dataset used in the experiments in combination with the performance gap between the two DNNs, which Chen and Goodman's modified Kneser-Ney smoothing [13] is unable to handle.

4 Experiments

The baseline for all experiments in this work is the state-of-the-art multimodal late fusion model as depicted in Figure 1a. Both the baseline and GradSPSA are trained with 150 epochs using the Adam optimizer [30] with a learning rate of 0.001 and a batch size of 16 for DNN_{OMR} [4]. During the training of DNN_{A2S}, used in the baseline, a batch size of 4 was used [4]. To accommodate the increased memory requirements of MTL for GradSPSA's DNN_{A2S} a batch size of 2 was used due to using two tasks' features at once.

The original baseline method in [4] selects the test fold based on the best validation fold. This is a non-standard cross validation method which could bias the results upwards. Therefore, in our experiments both the original non-standard cross validation method and a standard k-fold cross validation method are used for the baseline and GradSPSA.

The initial values for hyperparameter γ and α are chosen based on [40] and [4], respectively, where γ equals to $1e$-07, and α equals to 0.5. Furthermore, the learning rate is chosen based on [4] and equals to 0.001.

Due to running into smoothing problems during the grid search in the word graph creation stage, we changed the baseline's smoothing method from Chen and Goodman's modified Kneser-Ney smoothing [13] to Witten-Bell smoothing [63, 59], and moved on to the experiments on a 60% subset of the dataset as shown in Table 1 due to the time constraint of this work. Here, the 60% subset of the dataset, obtained through simple random sampling, is only used for GradSPSA, while the full dataset is used for the baseline [4].

5 Results

The initial experiment shown in Table 2 used a total of 150 epochs for both DNNs and shows that 80 epochs is sufficient for the proposed DNN_{A2S}. However, due to the time constraint of this work, the experiments are a continuation of the grid search on 60% of the dataset, which in essence means that 150 epochs are used for both DNNs.

Fold	OMR	A2S
Fold 0	130	62
Fold 1	133	72
Fold 2	95	31
Fold 3	126	66
Fold 4	150	55

Table 2: The initial experiment used a total of 150 epochs. Here, it shows that 80 epochs is sufficient for A2S, while OMR could benefit from more than 150 epochs.

Fusion method as described in [4]	Scenarios								
	$1_{H,H}$	$2_{H,M}$	$3_{H,L}$	$4_{M,H}$	$5_{M,M}$	$6_{M,L}$	$7_{L,H}$	$8_{L,M}$	$9_{L,L}$
Late multimodal fusion model [4] as in the experiments by [4]									
Minimum Bayes Risk	23.3	16.8	7.2	17.4	14.4	7.1	8.0	7.2	5.8
Lightly-supervised A2S	28.5	28.4	28.0	20.7	19.3	19.6	10.9	10.3	11.4
Lightly-supervised OMR	25.8	17.1	9.1	25.4	16.8	8.9	24.9	16.5	9.1
Confusion network	24.6	15.9	6.3	18.2	15.0	5.8	8.1	7.3	5.5
Smith-Waterman	20.8	13.8	6.1	18.2	11.1	4.5	14.5	8.1	3.0
Late multimodal fusion model [4] as in our experiments									
Minimum Bayes Risk	28.3	29.4	46.3	21.6	21.0	27.5	13.0	12.5	21.6
Lightly-supervised A2S	28.0	32.0	-	-	21.1	28.3	15.6	13.6	20.0
Lightly-supervised OMR	23.4	27.6	-	-	19.5	23.2	18.8	15.8	19.3
Confusion network	29.6	32.5	49.3	25.3	22.0	30.3	18.0	15.2	22.4
Smith-Waterman	24.7	25.4	44.0	17.9	16.8	25.3	10.3	8.9	17.7
GradSPSA (ours; cross validation as implemented by [4])									
Minimum Bayes Risk	**2.7**	22.9	**8.5**	**5.4**	**9.7**	48.9	**3.8**	**5.6**	50.6
Lightly-supervised A2S	19.9	32.2	78.4	21.0	34.0	75.7	20.4	33.5	78.5
Lightly-supervised OMR	10.7	97.3	18.0	13.6	21.6	**42.4**	11.9	18.4	**35.8**
Confusion network	18.3	30.3	34.1	21.2	24.6	55.9	19.9	22.1	50.6
Smith-Waterman	6.0	**21.6**	64.8	7.7	14.6	75.7	6.7	12.3	68.7
GradSPSA (ours; k-fold cross validation)									
Minimum Bayes Risk	**4.2**	**13.9**	**6.4**	**8.1**	**7.3**	48.2	**5.9**	**4.3**	35.0
Lightly-supervised A2S	76.5	25.3	72.0	31.0	26.8	73.3	30.5	26.3	72.5
Lightly-supervised OMR	49.9	37.8	16.4	11.3	16.9	**44.8**	9.8	14.4	**33.9**
Confusion network	15.0	24.1	28.9	18.2	21.6	54.6	16.7	19.7	46.7
Smith-Waterman	22.9	14.3	50.2	24.4	10.4	63.8	23.5	8.5	55.7

Table 3: The nine combinations of high (H), medium (M), and low (L) Symbol Error Rate (SER) for the tasks, i.e. the scenarios. Here, the multimodal late fusion baseline uses the full dataset with a modified Kneser-Ney smoothing [13] and GradSPSA uses 60% of the dataset with Witten-Bell smoothing [63]. Here, underlined values beat the baseline as in our experiments and boldface values are the best SER scores for their respective validation method.

Table 3 shows the results of the experiments described in Section 4. Here, a scenario is denoted

by $N_{\text{OMR performance level,A2S performance level}}$, where N denotes the scenario and the subscripts denote the performance level of high, medium, or low SER score for task OMR and A2S, respectively. Furthermore, scenario 3 and 4 in Table 3 do not have results for the lightly-supervised method for the baseline due to the modified Kneser-Ney smoothing being unable to work with small datasets. On top of this, scenarios 3, 6, and 9 in the baseline only have the modified Kneser-Ney smoothing applied to order 2 instead of order 2, 3, 4, and 5 due to the small amount of data available in these scenarios. Furthermore, fold 3 does not have suitable sample pairs for all scenarios except 1, 4, and 7 as shown in Table 4, which means the results for scenario 2, 3, 5, 6, 8, and 9 are averaged across four folds instead of five.

Fusion method performance

In the results in Table 3, we observed that Minimum Bayes Risk (MBR) outperforms all other fusion methods in almost all scenarios for the GradSPSA k-fold cross validation results. The only scenarios for these results where MBR does not perform as the best method is scenario 6 and 9. In these scenarios lightly-supervised A2S performs the best with MBR a close second. This might be due to these scenarios containing a low amount of test data as shown in Table 4.

Cross validation schemas

Furthermore, we observe in Table 3 that the best SER for the validation scheme from [4] performs better than the best SER for the k-fold cross validation scheme for scenarios 1, 4, 6, and 7. Notably, all these scenarios, except for scenario 6, have the most amount of data available across all scenarios. Scenario 6 is the odd one out, but outperforms the k-fold cross validation results only for the best SER, i.e. lightly-supervised OMR, while the other fusion methods do not. On top of this, in Table 3 the k-fold cross validation results for all other scenarios, i.e. scenarios 2, 3, 5, 8, and 9, outperform the validation scheme for all fusion methods. This suggests that the bias in the cross validation scheme from [4] has an even bigger bias when the amount of available data is smaller.

Baseline performance

However, we also observed in Table 3 that the baseline in our experiments has a higher SER for most scenarios and fusion methods than the baseline experiments performed by [4]. This difference in SER could influence the results of GradSPSA by biasing the GradSPSA results downwards.

Scenario performances

In Table 4 the A2S M and L levels have a direct impact on the number of test samples.

Furthermore, we compare all scenarios in Table 3 to their performance level. Scenario 1 outperforms the baseline in most of the fusion methods with the exception of the lightly-supervised methods for the k-fold cross validation. Considering the performance level of scenario 1, which is approximately 30% SER for both tasks, most fusion methods perform much better, i.e. at a medium, low, or even extremely low performance level, than the individual tasks.

Scenario 2 performs better than expected for MBR and Smith-Waterman for the k-fold cross validation scheme. This is shown by these fusion methods performing at a low performance level, while both tasks have at least a medium performance level.

Scenario 3 performs well for MBR, which performs at a low performance level, while all other fusion methods perform at a medium, high, or extremely high level. This might be due to the low amount of data available in this scenario as shown in Table 4. Furthermore, both lightly-supervised methods perform as expected, since the high OMR SER only worsens A2S, which results in a high SER for lightly-supervised A2S, and the low A2S SER improves the high OMR SER for lightly-supervised OMR.

	Scenarios		
	$1_{H,H}$ $4_{M,H}$ $7_{L,H}$	$2_{H,M}$ $5_{M,M}$ $8_{L,M}$	$3_{H,L}$ $6_{M,L}$ $9_{L,L}$
Fold 0	2674	1249	68
Fold 1	2674	1016	27
Fold 2	2674	1992	145
Fold 3	2674	0	0
Fold 4	2674	1644	71
Total	13370	5901	311

Table 4: The number of test sample pairs in a fold and the total test sample pairs per scenario $N_{\text{OMR performance level,A2S performance level}}$ for GradSPSA, where N denotes the scenario, the subscripts denote the performance level of high, medium, or low SER score for task OMR and A2S respectively, and the fold partitions are a subset of the original fold partitions in [4]. Here, A2S impacts the number of test sample pairs due to A2S performing at a high SER level, while the thresholds for medium and low SER performance level are relatively low. Furthermore, the scenarios constrained by these thresholds have an uneven distribution of the sample pairs across the folds due to the heterogeneous nature of the dataset.

Scenario 4 outperforms the baseline unexpectedly for MBR and lightly-supervised OMR for both the validation and cross validation scheme, and Smith-Waterman for only the validation scheme, because these fusion methods perform at a low performance level. As expected, most other fusion methods perform at a medium performance level.

Scenario 5 performs mostly as expected at the medium performance level, which is the same performance level as the individual tasks, with the exception of the MBR and Smith-Waterman fusion methods, which perform at a low performance level.

Scenario 6 performs much worse than expected, i.e. even worse than the high SER score, but considering the lack of test sample pairs as shown in Table 4, this is to be expected.

Scenario 7 performs mostly as expected, since most fusion methods perform at a low performance level. However, lightly-supervised A2S performs at the expected high performance level since OMR has a higher performance level than A2S, which worsens lightly-supervised A2S.

Scenario 8 performs well and as expected. MBR has a very low SER even for a low performance level.

Scenario 9 performs much worse than expected, but considering the lack of test sample pairs in this scenario as shown in Table 4, this is to be expected.

In Table 4, the scenarios are shown based on the sample pairs distribution per scenario. This shows that DNN_{A2S} constraints the number of sample pairs present in the partitions. Furthermore, A2S shows a clear preference for fold 2 for the medium SER, i.e. scenarios 2, 5, and 8, while showing a clear preference for fold 4 for all other A2S SER as shown in Table 5. Fold 2 is also the fold with the most sample pairs across the folds, which could suggest that the model performs best on more data.

	Scenarios	
	$1_{H,H}$	
	$3_{H,L}$	
	$4_{M,H}$	
	$6_{M,L}$	$2_{H,M}$
	$7_{L,H}$	$5_{M,M}$
	$9_{L,L}$	$8_{L,M}$
Fold	4	2

Table 5: The selected test fold for the original cross validation scheme in [4] for each scenario $N_{\text{OMR performance level,A2S performance level}}$, where N denotes the scenario and the subscripts denote the performance level of high, medium, or low SER score for task OMR and A2S, respectively. Here, the scenarios with a medium Symbol Error Rate (SER) for A2S have a clear preference for fold 2, while all other scenarios prefer fold 4, where the folds correspond to the folds in Table 4.

Noteworthy is that scenario 1 and 5 for several fusion methods outperform their individual tasks, even though both tasks' performance level is the same. In scenario 9 the two tasks also have the same low performance level, but it performs way worse than expected. However, this could be explained by the low amount of data available as shown in Table 4.

Furthermore, the scenarios where the OMR SER is lower than the A2S SER perform better for lightly-supervised OMR than for lightly-supervised A2S, which is to be expected since a high A2S worsens a low OMR SER.

For the scenarios where the OMR SER is higher than the A2S SER, i.e. scenarios 2, 3, and 6, both scenario 2 and 3 perform better than the individual tasks for MBR. On the other hand, scenario 6 performs worse overall but mostly likely this is due to the low amount of test data as shown in Table 4.

Lastly, the baseline uses the full cleansed dataset for the training, validation and test partitions, while the GradSPSA method uses only 60% of the cleansed dataset for these partitions.

6 Discussion

GradSPSA is evaluated on nine unique scenarios in order to observe what the influence of the individual tasks are on their fusion. Specifically, we find the scenarios where OMR SER is smaller than A2S SER and in the order of biggest performance gap to lowest performance gap, i.e. scenario 7, 8, and 4, the most important. Following in importance are the scenarios where OMR SER is equal to A2S SER, i.e. scenario 1, 5, 9. We find the least important scenarios are the scenarios where OMR SER is bigger than A2S SER. Furthermore, scenarios 3, 6, and 9 should be considered with care due to the small amount of test samples as shown in Table 4.

We observe based on the results in Section 5 that GradSPSA outperforms the baseline and the individual tasks for all scenarios using the Minimum Bayes Risk (MBR) fusion method when there are enough sample pairs for that scenario. Here, scenario 6 and 9 demonstrate that 311 sample pairs spread across four folds is not enough to get reliable results even with MBR. Furthermore, these scenarios also show along with Table 4 that while the performance gap between OMR and A2S has lessened, it is still present. On average GradSPSA outperforms the baseline with an average improvement of 37.2% in SER and up to an improvement of 86.2% in SER for the third scenario.

Furthermore, the results also suggest that the scarcity of AMT and A2S datasets does not have to be a limitation, since GradSPSA performs well with a relatively small amount of data. However, this would need to be verified by evaluating GradSPSA, which is trained on 36% of the cleansed dataset, on the test partition of the full dataset. Only if this is indeed verified, could we claim that GradSPSA lessens the demand on the scarce annotated AMT and A2S datasets.

Other fusion methods such as Smith-Waterman (SW) local alignment and confusion networks (CN) are also able to outperform the baseline, but are in most scenarios not able to beat the individual tasks.

Furthermore, we observed that the original cross validation scheme as used in the baseline [4] is not consistent with a regular k-fold cross validation scheme as additionally used in this work. Notable is that the scenarios where the original cross validation scheme used in [4] beats the k-fold cross validation scheme belong to the group of scenarios with the most amount of data available, i.e. the scenarios with a high A2S SER. The only exception is scenario 6, where only the best performing fusion method from the validation scheme beats the k-fold cross validation scheme, while all other fusion methods do not. All other scenarios outperform the validation scheme for all fusion methods. Therefore, scenario 6 is more in line with these scenarios. This seems to indicate a bias in the original cross validation scheme when more data is available.

7 Conclusion

This work proposes GradSPSA, a DNN with unidirectional gradual soft parameter sharing for multimodal music transcription. Our experimental results show that Minimum Bayes Risk (MBR) is the best performing fusion method and outperforms the state-of-the-art with an average improvement of 37.2% in Symbol Error Rate (SER). This improvement is up to 86.2% for the third scenario.

Furthermore, the performance gap between the two tasks in GradSPSA has lessened compared to the state-of-the-art model, but is also still present.

On top of this, late fusion enables GradSPSA to be partially optimized. While this is not the same as retraining a DNN, it does allow for some optimization through the hyperparameter α.

In short, we observed that the proposed GradSPSA could potentially alleviate the demands on scarce dataset, i.e. the dataset does not have to be as big as for the unimodal task to achieve the same performance.

Future work

A limitation to this work is the change in smoothing method from modified Kneser-Ney [13] to Witten-Bell smoothing [63] due to the modified Kneser-Ney smoothing's inability to work well on a very small vocabulary. We expect this to not be very limiting since modified Kneser-Ney consistently outperforms Witten-Bell smoothing on the same architecture in [13]. Therefore, Witten-Bell smoothing can only produce worse results than modified Kneser-Ney smoothing, which mitigates the risk of biasing the results upwards. A future work could include comparing the results for the GradSPSA to the baseline model when both are using Witten-Bell smoothing.

Furthermore, while the GradSPSA experiments improve upon the state-of-the-art, these experiments will only serve as initial experiments due to the necessity of verifying our findings on the full test set, i.e. 4,457 sample pairs, in order to rule out any biases due to only using a 60% subset of the test set, i.e. 2,674 sample pairs. Only after this verification can a claim be made about whether GradSPSA lessens the demand on the scarce annotated AMT and A2S datasets. A future work

could be the verification of our findings on the full test set.

Another limitation is the usage of simple random sampling for a dataset that we falsely assumed to be homogeneous. In a future work, using a sampling method suitable for creating a subset of a heterogeneous dataset in combination with using either a better performing base A2S architecture or using higher thresholds for the creation of the scenarios could result in a better comparison with existing methods.

Another limitation is that the Camera-PrIMuS dataset has been cleansed, which is not common practice in other studies. This could bias the results upwards because the cleansing process removes sample pairs that contain long rests, i.e. sample pairs that are harder to align. In a future work, our findings can be verified on the non-cleansed Camera-PrIMuS dataset.

Furthermore, the hyperparameters are not fully optimized due to the time constraint of this work. This is not expected to be a big limitation, since hyperparameter optimization can only improve the results.

An additional limitation is that there has not been conducted an ablation study to gain insight into the effect of dense connections, the interaction block, the third A2S convolutional block, and unidirectional gradual soft parameter sharing individually on the GradSPSA due to the time constraint of this work. This can limit our understanding of the individual components and their interaction on the results. In a future work it would be interesting to gain insight into this.

Lastly, other possible research avenues include using MTL on a different architecture, training GradSPSA with a bigger batch size, and training GradSPSA on a live music, i.e. non-synthesized, dataset. It would also be interesting to penalize wrong predictions based on the predictions' closeness to the ground truth, i.e. predictions that are not far off of the ground truth are penalized less than predictions that are far off from the ground truth. This is something Smith-Waterman local alignment does not take into account.

References

[1] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727, 2018.

[2] M. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] M. Alfaro-Contreras, J. Valero-Mas, J. Iñesta, and J. Calvo-Zaragoza. Multimodal strategies for image and audio music transcription: A comparative study. In *Pattern Recognition, Computer Vision, and Image Processing. International Conference on Pattern Recognition 2022 International Workshops and Challenges*, pages 64–77, Cham, 2023. Springer Nature Switzerland.

[4] M. Alfaro-Contreras, J. Valero-Mas, J. Iñesta, and J. Calvo-Zaragoza. Late multimodal fusion for image and audio music transcription. *Expert Systems With Applications*, 216:119491, 2023.

[5] A. Baró, P. Riba, and A. Fornés. Musigraph: Optical music recognition through object detection and graph neural network. In *Frontiers in Handwriting Recognition*, pages 171–184, Cham, 2022. Springer International Publishing.

[6] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

[7] J. Calvo-Zaragoza, J. Martinez-Sevilla, C. Penarrubia, and A. Ríos-Vila. Optical music recognition: Recent advances, current challenges, and future directions. In *Document Analysis and Recognition - ICDAR 2023 Workshops*, pages 94–104, Cham, 2023. Springer Nature Switzerland.

[8] J. Calvo-Zaragoza, A. Pertusa, A. Gallego, J. Iñesta, L. Micó, J. Oncina, C. Perez-Sancho, P. Ponce de León, and D. Rizo. Multiscore project: Multimodal transcription of music scores. In *Proceedings of the 14th Machine Learning and Music Workshop*, page 3, 2021.

[9] J. Calvo-Zaragoza and D. Rizo. Camera-PrIMuS: Neural end-to-end optical music recognition on realistic monophonic scores. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 248–255, 2018.

[10] J. Calvo-Zaragoza and D. Rizo. End-to-end neural optical music recognition of monophonic scores. *Applied Sciences*, 8(4), 2018.

[11] J. Calvo-Zaragoza and D. Rizo. Semtantic to MIDI converter. https://grfia.dlsi.ua.es/primus/, 2018.

[12] S. Chen, Q. Jin, J. Zhao, and S. Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26, 2017.

[13] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[14] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 49:167–192, 2017.

[15] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1643–1654, 2010.

[16] J. Fainberg, O. Klejch, S. Renals, and P. Bell. Lattice-Based Lightly-Supervised Acoustic Model Training. In *Proceedings Interspeech 2019*, pages 1596–1600, 2019.

[17] C. Fuente, J. Valero-Mas, F. Castellanos, and J. Calvo-Zaragoza. Multimodal image and audio music transcription. *International Journal of Multimedia Information Retrieval*, 11:77–84, 2022.

[18] A. Galan-Cuenca, J. Valero-Mas, J. Martinez-Sevilla, A. Hidalgo-Centeno, A. Pertusa, and J. Calvo-Zaragoza. MUSCAT: A multimodal music collection for automatic transcription of real recordings and image scores. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 583–591, New York, NY, USA, 2024. Association for Computing Machinery.

[19] Y. Gong, A. Liu, A. Rouditchenko, and J. Glass. UAVM: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, 2022.

[20] H. Guo, R. Pasunuru, and M. Bansal. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[21] J. Han and P. Zeng. Residual BiLSTM based hybrid model for short-term load forecasting in buildings. *Journal of Building Engineering*, page 111593, 12 2024.

[22] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck. Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 50–57. ISMIR, Sept. 2018.

[23] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[24] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[25] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[26] F. Jamshidi, G. Pike, A. Das, and R. Chapman. Machine learning techniques in automatic music transcription: A systematic survey. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, 2024.

[27] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2025. Online manuscript released August 24, 2025.

[28] E. Kerinec, A. Søgaard, and C. Braud. When does deep multi-task learning work for loosely related document classification tasks? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–8. Association for Computational Linguistics, 2018.

[29] J. Kim and J. Bello. Adversarial learning for improved onsets and frames music transcription. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 670–677. ISMIR, Nov. 2019.

[30] D. Kingma and J. Ba. ADAM: A method for Stochastic Optimization. In *3rd International Conference for Learning Representations*, 2015.

[31] D. Kodati and R. Tene. Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers. *Neural Computing and Applications*, 37(5):3077–3110, 2024.

[32] S. Lee, J. Hwang, M. Song, and H. Park. A method based on dual cross-modal attention and parameter sharing for polyphonic sound event localization and detection. *Applied Sciences*, 12(10), 2022.

[33] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[34] J. Liang. Harmonizing minds and machines: survey on transformative power of machine learning in music. *Frontiers in Neurorobotics*, 17, 2023.

[35] L. Liu and E. Benetos. From audio to music notation. *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, pages 693–714, 2021.

[36] L. Liu, V. Morfi, and E. Benetos. Joint multi-pitch detection and score transcription for polyphonic piano music. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 281–285. IEEE, 2021.

[37] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7613–7617. IEEE, 2021.

[38] E. Manilow, P. Seetharaman, and B. Pardo. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 771–775, 2020.

[39] T. Moebert, C. J., and M. Weseloh. Fluidsynth. https://www.fluidsynth.org/.

[40] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, 2021.

[41] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6548–6552. IEEE, 2018.

[42] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop*, pages 513–520. IEEE, 2018.

[43] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[44] A. Ríos-Vila, E. Fuentes-Martinez, and J. Calvo-Zaragoza. Towards sheet music information retrieval: A unified approach using multitask transformers. In *6th International Workshop on Reading Music Systems*, page 7, 2024.

[45] A. Ríos-Vila, J. Iñesta, and J. Calvo-Zaragoza. On the use of transformers for end-to-end optical music recognition. In *Pattern Recognition and Image Analysis*, pages 470–481, Cham, 2022. Springer International Publishing.

[46] A. Ríos-Vila, D. Rizo, J. Iñesta, and J. Calvo-Zaragoza. End-to-end optical music recognition for pianoform sheet music. *International Journal on Document Analysis and Recognition*, 26(3):347–362, 2023.

[47] M. Román, A. Pertusa, and J. Calvo-Zaragoza. An end-to-end framework for audio-to-score music transcription on monophonic excerpts. In *Proceedings of 19th International Society for Music Information Retrieval Conference*, pages 34–41, 2018.

[48] M. Román, A. Pertusa, and J. Calvo-Zaragoza. A holistic approach to polyphonic music transcription with neural networks. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 731–737, Delft, The Netherlands, 2019.

[49] M. Román, A. Pertusa, and J. Calvo-Zaragoza. Data representations for audio-to-score monophonic music transcription. *Expert Systems with Applications*, 162:113769, 2020.

[50] S. Ruder. An overview of multi-task learning in deep neural networks. https://arxiv.org/pdf/1706.05098, 2017.

[51] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard. Sluice networks: Learning what to share between loosely related tasks. https://arxiv.org/pdf/1705.08142v1, 2017.

[52] C. Schörkhuber and A. Klapuri. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64. SMC, 2010.

[53] E. Shatri and G. Fazekas. DoReMi: First glance at a universal omr dataset. In *Proceedings of the 3rd International Workshop on Reading Music Systems*, 2021.

[54] S. Shatri and G. Fazekas. Optical music recognition: State of the art and major challenges. In *Proceedings of the International Conference on Technologies for Music Notation and Representation - TENOR'20/21*, pages 175–184, Hamburg, Germany, 2020. Hamburg University for Music and Theater.

[55] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[56] K. Shibata, E. Nakamura, and K. Yoshii. Non-local musical statistics as guides for audio-to-score piano transcription. *Information Sciences*, 566:262–280, 2021.

[57] A. Sigalov. Soundfont file. https://huggingface.co/spaces/asigalov61/Melody-Harmonizer-Transformer/blob/main/SGM-v2.01-YamahaGrand-Guit-Bass-v2.7.sf2, 2024.

[58] A. Stolcke. SRILM-an extensible language modeling toolkit. In *International Conference on Spoken Language*, volume 2, pages 901–904, 2002.

[59] A. Stolcke, D. Yuret, and N. Madnani. SRILM-FAQ. http://www.speech.sri.com/projects/srilm/manpages/srilm-faq.7.html.

[60] G. Tong, Y. Li, H. Gao, H. Chen, H. Wang, and X. Yang. MA-CRNN: a multi-scale attention CRNN for Chinese text line recognition in natural scenes. *International Journal on Document Analysis and Recognition*, 23(2):103–114, 11 2019.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[62] H. Wang, X. Jin, Y. Du, N. Zhang, and H. Hao. Adaptive hard parameter sharing method based on multi-task deep learning. *Mathematics*, 11(22):4639, 2023.

[63] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 2002.

[64] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T. Liu. Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2015–2024, 2022.

[65] J. Yu, Y. Dai, X. Liu, J. Huang, Y. Shen, K. Zhang, R. Zhou, E. Adhikarla, W. Ye, Y. Liu, et al. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. 2024.

[66] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for video Technology*, 28(10):3030–3043, 2017.

[67] Y. Zhang, J. Xiao, S. Yang, and S. Shan. Multi-task learning for audio-visual active speaker detection. *The ActivityNet Large-Scale Activity Recognition Challenge*, 4:2, 2019.