**用CNN对音频信号进行音色分析**

**摘要**

The focus of this work is to study how to efficiently tailor Convolutional Neural Networks (CNNs) towards learning timbre representations from **log-mel magnitude spectrograms**.We first review the trends when designing CNNs architectures. Through this literature overview we discuss which are the crucial points to consider for efficiently learning timbre **representations（都被我翻译成了：表达）** using CNNs. From this discussion we propose a design strategy meant to capture the relevant time-frequency contexts for learning timbre, which permits using domain knowledge for designing architectures. In addition, one of our main goals is to design efficient CNN architectures – what reduces the risk of these models to over-fit, since CNNs’ number of parameters is minimized. Several architectures based on the design principles we propose are successfully assessed for different research tasks related to timbre: singing voice phoneme classification, musical instrument recognition and music auto-tagging.

这项工作的重点是研究如何根据**对数梅尔幅度谱图**有效地调整卷积神经网络（CNN）从而学习音色。我们先回顾了一下设计CNN结构的趋势。通过文献综述，我们讨论的是“使用CNN高效学习音色表达的关键点是什么”。从这个讨论中，我们提出了一个设计策略，旨在为了学习音色而捕捉相关的时间频率，它允许使用领域知识设计架构。此外，我们的主要目标之一是设计高效的CNN架构，它能降低了这些模型过度拟合的风险，因为CNN的参数数目被最小化了。根据我们提出的设计原则，成功地评估了几种与音色相关的研究任务：歌唱声音音素分类、乐器识别和音乐自动标注。

1. **介绍**

Our goal is to discover novel deep learning architectures that can efficiently model music signals, what is a very challenging undertaking. After showing that is possible to design efficient CNNs [1] for modeling temporal features –tempo & rhythm– we now focus on studying how to efficiently learn timbre 1 [注释][[1]](#footnote-0)representations, one of the most salient musical features.

我们的目标是发现新的深层学习架构，可以有效地建模音乐信号，这是一个非常具有挑战性的事业。在显示可以设计高效的神经网络“建模的时空特征[ 1 ]–节奏与韵律–”之后，我们现在致力于研究如何有效地学习音色表达，一个最为突出的音乐特征 。

Music audio processing techniques for timbre description can be divided in two groups: (i) bag-of-frames methods and (ii) methods based on temporal modeling. On the one hand, bag-of-frames methods have shown to be limited as they just model the statistics of the frequency content along several frames [2]. On the other hand, methods based on temporal modeling consider the temporal evolution of frame-based descriptors [3][4] – some of these methods are capable of representing **spectro-temporal** patterns, that can model the temporal evolution of timbre [4]. Then, for example, **attack-sustain-release** patterns can be jointly represented.

音色描述的音乐音频处理技术可分为两类：（一）帧包的方法和（ii）基于时间建模的方法。一方面，帧的方法已被证明是有限的，因为他们只是沿有限帧建立统计频率内容的模型[ 2 ]。另一方面，基于时序建模方法考虑的时间演化的帧描述[ 3 ] [ 4 ]–有些方法能够代表**spectro-temporal**模式，能模拟音色[ 4 ]随时间的演化。然后，例如，**攻击持续释放模式**可以共同代表。

Most previous methodologies –either based on (i) or (ii)–require a dual pipeline: first, descriptors need to be extracted using a pre-defined algorithm and parameters; and second, (temporal) models require an additional framework tied on top of the proposed descriptors – therefore, descriptors and (temporal) models are typically not jointly designed. Throughout this study, we explore modeling timbre by means of deep learning with the input set to be magnitude spectrograms. This quasi end-to-end learning approach allows minimizing the effect of the fixed pre-processing steps described above. Note that no strong assumptions over the descriptors are required since a generic perceptually-based pre-processing is used: log-mel magnitude spectrograms. Besides, deep learning can be interpreted as a temporal model (if more than one frame is input to the network) that allows learning **spectro-temporal** descriptors from spectrograms (i.e. with CNNs in first layers). In this case, learnt descriptors and temporal model are jointly optimized, what might imply an advantage when compared to previous methods.

以往的方法–基于（i）或（ii）–需要双管道：第一，描述需要使用一个预先定义的算法和参数的提取；第二，（时间）模型需要一个额外的框架，它被放在提议的描述符的顶部--因此描述符和（时间）模型通常是不联合设计的。在本研究中，我们探讨通过将输入设置为幅度图谱的深度学习的方法来建模音色。这种类似准端到端学习方法允许最小化上述固定的预处理步骤的效果。注意，没有很强的假设的描述是必需的，因为一个通用的基于感知的预处理是：对数梅尔幅度谱图。此外，深度学习可以被解释为一个时间模型（如果多于一帧输入到网络），它允许从谱图中学习**谱-时间**的描述符（即第一层CNNs）。在这种情况下，学习的描述符和时间模型是联合优化的，这或许是在暗示与以前的方法相比存在着优势。

From the different deep learning approaches, we focus on CNNs due to several reasons: (i) by taking spectrograms as input, one can interpret filter dimensions in time-frequency do-main; and (ii) CNNs can efficiently exploit invariances –such

as time and frequency invariances present in spectrograms–by sharing a reduced amount of parameters. We identified two general trends for modeling timbre using spectrogram-based CNNs: using small-rectangular filters (m << M and n << N)

[6] or using high filters (m ≤ M and n << N) 2 [注释][[2]](#footnote-1)[7][8].

从不同深度的学习方法，我们专注于CNN由于几个原因：（一）以谱图作为输入，可以及时解析滤波器维度--以频率为主；及（ii）神经网络可以有效地利用不变性–如时间和频率在谱图中的不变性–通过分享一个简化的的参数数量。我们发现利用基于CNN的谱图进行音色建模有两大趋势：使用小矩形滤波器（m<<M和n<<N）或使用高通滤波器（m≤M和n << N）CNNs的输入设置为维度为M×N的对数梅尔频谱，同时CNN滤器维度是m×n，M和m代表频点数，N和n代表时间帧数。

· Small-rectangular filters inquire the risk of limiting the representational power of the first layer since these filters are typically too small for modeling spread spectro-temporal patterns [1]. Since these filters can only represent sub-band characteristics (with a small frequency context: m <<M) for a short period of time (with a small time context: n <<N) these can only learn, for example: onsets or bass notes [9][10]. But these filters might have severe difficulties on learning cymbals’ or snare-drums’ time-frequency patterns in the first layer since such a spread context can not fit inside a small-rectangular filter. 3 [注释][[3]](#footnote-2)

小的矩形滤波器查询限制第一层的表达能力，因为这些滤波器的风险对于建立传播的频率--时间模型通常太小。由于这些滤波器只能代表子带特征（用一个小的频率背景： m<< M）在一小段时间中（一个小的时间背景：N＜＜n）这些只能学习，例如：声母或者低音。但这些滤波器可能在第一层时频模式中对学习钹或小鼓”存在很严重的困难，由于这样的传播语境不能放在一个小的矩形滤波器。

· Although high filters can fit most **spectral envelopes**, these might end up with a lot of weights to be learnt from (typically small) data – risking to over-fit and/or to fit noise. See Fig. 1 (right) for two examples of filters fitting noise as a result of having available more context than the required for modeling onsets and harmonic partials, respectively. 3

虽然高滤波器可以满足大多数**频谱包络**，但这些可能最终会有大量的权重需要从（通常是小的）数据中学习-冒着过度拟合和/或匹配噪声的风险。图1（右）两个滤波器拟合噪声的例子，由于分别有更多的上下文比建模所需的起始和谐波泛音。

Additionally, most CNN architectures use unique filter shapes in every layer [5][7][6]. However, recent works point out that using different filter shapes in each layer is an efficient way to exploit CNN`s capacity [1][11]. For example, Pons et al. [1] proposed using different musically motivated filter shapes in the first layer to efficiently model several musically relevant time-scales for learning temporal features. In Section II we propose a novel approach to this design strategy which facilitates learning musically relevant time-frequency contexts while minimizing the risk of noise-fitting and over-fitting for timbre analysis. Out of this design strategy, several CNN models are proposed. Section III assesses them for three research tasks related to timbre: singing voice phoneme classification, musical instrument recognition and music auto-tagging.

此外，大多数神经网络结构在每一层中采用独特的滤波器的形状。然而，最近的作品，指出利用各层不同的滤波器的形状是利用CNN能力的有效途径。例如，Pons等人建议在第一层使用不同的音乐动机的滤波器形状，以此为学习时间特征有效地模拟几个音乐相关的时间尺度。在第二部分中，我们提出了一种新的方法，这种设计策略，有利于学习音乐相关的时频环境，同时最大限度地减少音色分析中噪声拟合和过拟合的风险。出于这种设计，几种CNN模型得以建立。第三部分评估了与音色有关的三项研究任务：歌唱语音音素分类、乐器识别和音乐自动标注。

**II. CNNS DESIGN STRATEGY FOR TIMBRE ANALYSIS**

**二、针对音色分析的CNNs设计战略**

Timbre is considered as the “color” or the “quality” of a sound [12]. It has been found to be related to the spectral envelope shape and to the time variation of spectral content [13]. Therefore, it is reasonable to assume timbre to be a time- frequency expression and then, magnitude spectrograms are an adequate input. Although phases could be used, these are not considered – this is a common practice in the literature [5][7][6], and this investigation focuses on how to exploit the capacity of spectrograms to represent timbre. Moreover, timbre is often defined by what it is not: “a set of auditory attributes of sound events in addition to pitch, loudness, duration, and spatial position” [14]. Then, we propose ways to design CNN architectures invariant to these attributes:

音色被认为是声音的“颜色”或者“品质”。我们发现他和频谱包络以及随时间变化的频谱内容有关。因此，音色是一个时频表达然后幅度谱图是足够的输入是合理的假设。虽然相位可以使用，这些都是毋庸置疑的–这是在文献中常见的做法，这一调查的重点是如何利用谱图的能力代表音色。此外，音色通常被定义为：“音高、响度、持续时间和空间位置之外的声音事件的一组听觉属性”。然后，我们对这些属性提出了设计CNN架构不变的方式：

**Pitch invariance**. By enabling filters to convolve through the frequency domain of a mel spectrogram (a.k.a. f 0 shifting), the resulting filter and feature map can represent timbre and pitch information separately. However, if filters do not capture the whole spectral envelope encoding timbre –because these model a small frequency context–, previous discussion does not necessarily hold. Additionally, depending on the used spectrogram representation (i.e. STFT or mel) CNN filters might be capable of learning more robust pitch invariant features. Note that STFT timbre patterns are f 0 dependent. However, mel timbre patterns are more pitch invariant than STFT ones because these are based in a different (perceptual) frequency scale. Besides, a deeper representation can be pitch invariant if a max-pool layer spanning all over the vertical axis 4[注释] [[4]](#footnote-3)of the feature map (M’) is applied to it: MP(M’,·).

**音调不变。**通过使滤波器卷积通过Mel谱频率域（即F 0转移），由此产生的滤波和特征映射可以分别代表的音色和音调信息。然而，如果滤波器没有捕获整个频谱包络编码音色，（因为这些模型在一个小频率环境），以前的讨论并不一定成立了。此外，根据用图表示（即STFT或MEL）CNN滤波器可以学习更强大的音调不变特征。注意，STFT音色模式依赖F 0。然而，梅尔音色模式比STFT更能使音调不变因为这些都是建立在一个不同的（感性）频率范围。此外，如果一个最大池化层可以横跨特征映射（M）D 所有垂直轴【N~和M~一般表示任何特征映射的维数。因此，尽管筛选器映射尺寸取决于滤波器大小，但我们用相同的名称表示这些维度：N~和M~】他可以被应用于：MP（M，·），那么我们就有可能获得一个更深层的表达。

**Loudness invariance** for CNN filters can be approached by using weight decay – L2-norm regularization of filter weights. By doing so, filters are normalized to have low energy and energy is then expressed into feature maps. Loudness is a perceptual term that we assume to be correlated with energy.

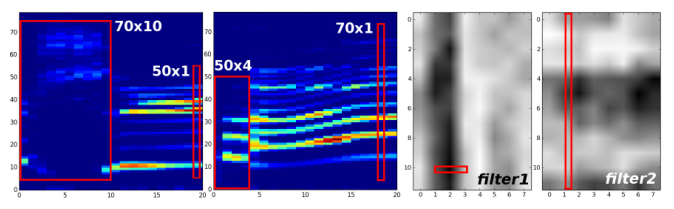
CNN滤波器的**响度不变**性可以用权衰减-滤波器范数的L2范数正则化来逼近。通过这样做，滤波器被归一化为低能量，然后能量被表示成特征映射。响度是一个感性的术语，我们假设它与能量有关。

**Duration invariance.** Firstly, m×1 filters are time invari- ant by definition since these do not capture duration. Temporal evolution is then represented in the feature maps. Secondly, sounds with determined length and temporal structure (i.e. kick drums or cymbals) can be well captured with m×n filters. These are also duration invariant because such sounds last a fixed amount of time. Note the resemblance between first layer m×1 filters with frame-based descriptors; and between first layer m×n filters with spectro-temporal descriptors.

**时间不变性。**首先，m×1滤波器通过定义来让时间不变因为这些不捕获时间。随时间的演变，然后在特征图表示。其次，确定长度和时间结构（即踢鼓和钹的声音）可以用m×n滤波器很好地捕捉。这些也都是因为这样的声音持续不变的最后一段固定的时间。注意第一层m×1滤波器和基于帧的描述之间的相似性；注意第一层m×n滤波器和频率时间的描述

**Spatial position invariance** is achieved by down-mixing (i.e. averaging all channels) whenever the dataset is not mono.From previous discussion, we identify the filter shapes of the first layer to be an important design decision – they play a crucial role for defining pitch invariant and duration invariant CNNs. For that reason, we propose to use domain knowledge for designing filter shapes. For example, by visually inspecting ***Fig. 1***(left) one can easily detect the relevant time-frequency contexts in a spectrogram: frequency ∈ [50,70] and time ∈ [1,10] – which can not be efficiently captured with several small-rectangular filters. These measurements provide an intuitive guidance towards designing efficient filter shapes for the first CNN layer.

**空间位置不变**通过下混合来实现（即平均所有通道）当数据集不是单声道的。从前面的讨论，我们确定的滤波器形状的第一层是一个重要的设计决策，他们起着至关重要的作用–定义音调不变和持续不变的CNN。出于这个原因，我们建议使用领域知识设计滤波器形状。例如，通过视觉检查图1（左）可以很容易地检测频谱相关时的背景：频率∈ [50,70]、时间∈ [1,10] –不能有效地捕捉几个小矩形滤波器。这些测量为设计CNN第一层的高效滤波器形状提供了直观的指导。



*Fig. 1. Left: two spectrograms of different sounds used for the singing voice phoneme classification experiment. Right: two trained small-rectangular filters of size 12×8. Relevant time-frequency contexts are highlighted in red.*

图1。左：用于歌唱语音音素的分类实验谱图不同的声音。右：两个训练有素的小矩形滤波器尺寸12×8。有关时背景用红色突出显示。

Finally, we discuss how to efficiently learn timbre features with CNNs. Timbre is typically expressed at different scales in spectrograms – i.e. cymbals are more spread in frequency than bass notes, or vowels typically last longer than consonants in singing voice. If a unique filter shape is used within a layer, one can inquire the risk of: (a) fitting noise because too much context is modeled and/or (b) not modeling enough context.

最后，我们将讨论如何有效地学与CNNs的音色特征。音色通常在谱图–即钹不同尺度的表达更传播频率比低音，或元音比辅音通常唱歌的声音持续时间更长。如果在一个层中使用唯一的筛选器形状，则可以查询（a）拟合噪声的风险，因为建模和/或（b）未建模足够上下文的上下文太多。

Risk (a). *Fig. 1* (right) depicts two filters that have fit noise. Observe that filter1 is repeating a noisy copy of an onset throughout the frequency axis, and filter2 is repeating a noisy copy of three harmonic partials throughout the temporal axis. Note that much more efficient representations of these musical concepts can be achieved by using different filter shapes: 1×3 and 12×1, respectively (in red). Using the adequate filter shape allows minimizing the risk to fit noise and the risk to over-fit the training set (because the CNN size is also reduced).

风险（a）。图1（右）描绘了两个滤波器，有合适的噪声。观察滤波器是重复出现嘈杂的复制整个频率轴，和滤波器是重复三次谐波泛音嘈杂的复制整个时间轴。注意，更有效的表示这些音乐的概念可以通过使用不同的滤波器的形状来实现：1×3和12×1，分别为（红色）。使用适当的滤波器形状允许最小化风险以适应噪音和过度适应训练集的风险（因为CNN的大小也减少了）。

Risk (b). The frequency context of filter2 is too small to model the whole harmonic spectral envelope, and it can only learn three harmonic partials – what is limiting the representational power of this (first) layer. A straightforward solution for this problem is to increase the frequency context of the filter. However note that if we increase it too much, such filter is more prone to fit noise. Using different filter shapes allows reaching a compromise between risk (a) and (b).

风险（b）。滤波器的频率背景太小，模型的整体谐波的频谱包络，它只能学三谐波泛音–什么是限制这一代表性的力量（一）层。这个问题的简单的解决方案是增加滤波器的频率下。但是请注意，如果我们增加太多，这样更容易适应噪声滤波器。使用不同的滤波器的形状可以达到风险之间的妥协（A）和（B）。

Using different filter shapes within the first layer seems crucial for an efficient learning with spectrogram-based CNNs. This design strategy allows to efficiently model different musically relevant time-frequency contexts. Moreover, this design strategy ties very well with the idea of using the available domain knowledge for designing filter shapes – that can intuitively guide the different filter shapes design so that spectro-temporal envelopes can be efficiently represented within a single filter. Note that another possible solution might be to combine several filters (either in the same layer or going deep) until the desired context is represented. However, several reasons exist for supporting the here proposed approach: (i) the Hebbian principle from neuroscience [15]: “cells that fire together, wire together”, and (ii) learning complete spectro-temporal patterns within a single filter allows to inspect and interpret the learnt filters in a compact way.

第一层内使用不同的滤波器的形状似乎是至关重要的一个有效的谱图基于细胞神经网络的学习。这种设计策略可以有效地模拟不同音乐背景有关时。此外，这种设计策略与利用领域知识设计滤波器的形状–，可以直观地引导不同的滤波器的形状设计使时空信封可以有效地表示一个单一的滤波器内的理念很好。注意，另一种可能的解决方案可能是合并几个滤波器（在同一层或深入），直到所需的上下文被表示。然而，支持在这里提出的方法存在着几方面的原因：（一）从神经科学[ 15 ] Hebbian原理：“细胞火一起，连接”，和（ii）在一个单一的滤波器允许检查和解释在一个紧凑的方式学会学习完整的时空模式滤波器。

Above discussion introduces the fundamentals (in bold italics) of the proposed design strategy for timbre analysis.

以上讨论了基本面（粗体斜体）对音色的分析所提出的设计策略。

**III. E XPERIMENTS**

1. **实验**

Audio is fed to the network using fixed-length log-mel spectrograms. Phases are discarded. Spectrograms are normalized: **zero mean and variance one.** Activation functions are ELUs [16]. Architectures are designed according to the proposed strategy and previous discussion – by employing: weight decay regularization, monaural signals, and different filter shapes in the first layer. Each network is trained optimizing the cross-entropy with SGD from random initialization [17]. The best model in the validation set is kept for testing.

音频使用固定长度的对数梅尔频谱图输入到网络中。**相位**被丢弃。谱图进行归一化处理：**零均值和方差**。激活操作三ELUS。架构是根据被提议的战略和提前的讨论来设计的：采用重量衰减正规化，单声道信号，并在第一层使用不同的滤波器形状。每层网络在随机初始化中通过SGD被训练而优化交叉熵。验证集中最好的模型用于测试。

In the following, we assess the validity of the proposed design strategy with 3 general tasks based on timbre modeling:

在下面，我们根据音色模型的来评估所提出的3个一般任务设计策略的有效性：

**III. A. Singing voice phoneme classification**

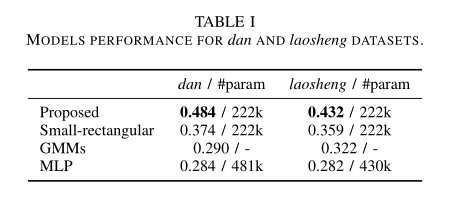
**III. A. 声音分类**

**The jingu** [注释][[5]](#footnote-4) a cappella singing audio dataset used for this study [18] is annotated with 32 phoneme classes[注释][[6]](#footnote-5) and consists of two different role-types of singing: dan (young woman) and laosheng (old man). The dan part has 42 recordings (89 minutes) and comes from 7 singers; the laosheng part has 23 arias (39 minutes) and comes from other 7 laosheng singers. Since the timbral characteristics of dan and laosheng are very different, the dataset is divided in two. Each part is then randomly split –train (60%), validation (20%) and test (20%)– for assessing the presented models for the phoneme classification task. Audio was sampled at 44.1 kHz. STFT was performed using a 25ms window (2048 samples with zero-padding) with a hop size of 10ms. This experiment assesses the feasibility of taking architectural decisions based on domain knowledge for an efficient use of the network’s capacity in small data scenarios. The goal is to do efficient deep learning by taking advantage of the design strategy we propose. This experiment is specially relevant because, in general, no large annotated music datasets are available – this dataset is an example of this fact. The proposed architecture has a single wide convolutional layer with filters of various sizes. Input is of size 80×21 – the network takes a decision for a frame given its context: ±10ms, 21 frames in total. We use 128 filters of sizes 50×1 and 70×1, 64 filters of sizes 50×5 and 70×5, and 32 filters of sizes 50×10 and 70×10 – considering the discussion in section II. A max-pool layer of 2×N 0 follows before the 32-way softmax output layer with 30% dropout[[7]](#footnote-6). MP(2,N’) was chosen to achieve time-invariant representations while keeping the frequency resolution.

用于研究的**金谷**无伴奏合唱音频数据[ 18 ]是有32个音素类和由两个不同类型唱歌角色所组成的：旦（年轻女子）和老生（老人）。旦部分有42段记录（89分钟），来自7个歌者；老生部分有23种唱腔（39分钟），来自其他7个老生歌者。因为旦和老生的音色特征有很大的不同，所以数据集被分为2个。然后，每一部分都随机地分割——训练（60%）、验证（20%）和测试（20%）——用于评估音素分类任务所提出的模型。音频采样在44.1千赫。STFT（short-time Fourier transform 短时间傅里叶变换）使用25ms窗口进行的（用0填充的2048个样本）其中的10ms大小在不断变化。这个实验评估基于领域知识，在小数据情况下的网络容量的有效利用的采用架构决定的可行性。我们的目标是利用我们提出的设计策略来进行有效的深度学习。这个实验特别相关，因为一般来说，没有大的有注释的音乐数据集可用——这个数据集就是这个事实的一个例子。所提出的架构有一个单一带有各种大小滤波器的宽卷积层。输入的大小为80×21–这个网络中的一帧由它的环境决定：±10ms，共21帧。我们使用尺寸50×1和70×1的滤波器128个，50×5和70×5的滤波器64个，50×10 and 70×10的滤波器32个----想想在Ⅱ中的讨论。在32路带有30% dropout的softmax输出层前加一2\*N~的最大池化层。MP（2，N~）被选择去实现时间不变的表达，同时保持频率分辨率。

We use overall classification accuracy as evaluation metric and results are presented in ***Table I***. As a baseline, we also train a 40-component Gaussian Mixture Models (GMMs), a fully-connected MLP with 2 hidden layers (MLP) and Choi et al.’s architecture [5], that is a 5-layer CNN with small-rectangular filters of size 3×3 (Small-rectangular). All architectures are adapted to have a similar amount of parameters so that results are comparable. GMMs features are: **13 coefficients MFCCs, their deltas and delta-deltas**. 80×21 log-mel spectrograms are used as input for the other baseline models. Implementations are available online [注释][[8]](#footnote-7)

我们用总体分类精度作为评价指标及结果，见表1。作为基准，我们也训练40组的高斯混合模型（GMM），一个带有2个隐藏层（MLP）的完全连接MLP，Choi等人的结构[ 5 ]，这个结构是一个这大小为3\*3（小矩形）的小矩形滤波器的5层CNN。所有体系结构都有类似的参数，因此结果是可比的。GMMs的特点是：**13系数MFCCs，它们的三角洲和三角洲的三角洲**。80×21对数梅尔频谱图作为其他基线模型的输入被使用。实验可在线使用。



Proposed architecture outperforms other models by a significant margin (although being a single-layer model), what denotes the potential of the proposed design strategy. Deep models based on small-rectangular filters –which are state-of-the-art in other datasets [5][6]– do not perform as well as the proposed model in these small datasets. As future work, we plan to investigate deep models that can take advantage of the richer representations learnt by the proposed model.

我们提出的模型胜于其他模型是通过一个重要的margin（盈余）（虽然是一个单层模型），他贡献了潜在的设计策略。基于小矩形滤波器的深层模型——在其他数据集（5）[ 6 ]中最先进的模型，在这些小数据集中的性能不如所提出的模型好。作为未来的工作，我们计划研究深层的模型，他们可以利用我们所提出的模型学习更丰富的表示。

**Ⅲ.B. Musical instrument recognition**

**Ⅲ.B. 乐器识别**

IRMAS [19] training split contains 6705 audio excerpts of 3 seconds length labeled with a single predominant instrument. Testing split contains 2874 audio excerpts of length 5∼20 seconds labeled with more than one predominant instrument. 11 pitched class instruments are annotated. Audios are sampled at 44.1kHz. The state-of-the-art for this dataset corresponds to a deep CNN based on small-rectangular filters (of size 3×3) by Han et al. [6]. Moreover, another baseline is provided based on a standard bag-of-frames approach + SVM classifier proposed by Bosch et al. [19]. We experiment with two architectures based on the proposed design strategy:

“IRMAS训练集”包含3秒长度的6705个音频摘录，它们被标记一个单一的主要仪器。测试集包含长度为5∼20秒的2874个音频摘录，它们被标记与一个以上的主要仪器。11个斜班乐器有注释。音频是在44.1kHz的采样。这个数据库的最先进的和一个基于han等人的小矩形滤波器的深度CNN相匹配。此外，基于Bosch等人提出的标准帧包+ SVM分类器，提供了另一个基线。[ 19 ]。我们根据提出的设计策略对两种体系结构进行了实验：

· Single-layer has a single but wide convolutional layer with filters of various sizes. The input is set to be of size 96×128. We use 128 filters of sizes 5×1 and 80×1, 64 filters of sizes 5×3 and 80×3, and 32 filters of sizes 5×5 and 80×5. We also max-pool the M’ dimension to learn pitch invariant representations: MP(M’,16). 50% dropout is applied to the 11-way softmax output layer.

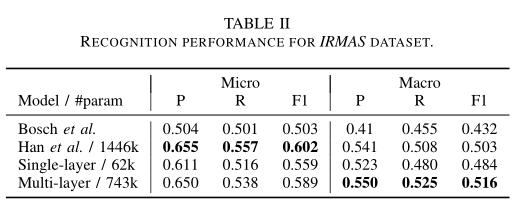
单层有一个带有不同尺寸滤波器的单一而宽的卷积层。输入设置为大小96×128。我们使用128个尺寸5×1和80×1的滤波器， 64个尺寸5×3和80×3的滤波器，32个尺寸5×5和80×5的滤波器。我们还最大池化M~维来学习音高不变表示：MP（M~，16）。50%dropout 应用于11路softmax输出层。

· Multi-layer architecture’s first layer has the same settings as single-layer but it is deepen by two convolutional layers of 128 filters of size 3×3, one fully-connected layer of size 256 and a 11-way softmax output layer. 50% dropout is applied to all the dense layers and 25% for convolutional layers. Each convolutional layer is followed by max-pooling: first wide layer - MP(12,16); deep layers - MP(2,2).

多层结构的第一层和单层有相同的设置，但它被尺寸为3×3的128滤波器的两个卷积层深化，一个尺寸256的全连接层和11路softmax输出层。50%dropout适用于所有致密层和25%的卷积层。每个卷积层后跟一个最大池化：第一个宽层MP（12,16）；深层MP（2,2）。

Implementations are available online 8 [[9]](#footnote-8). STFT is computed using 512 points FFT with a hop size of 256. Audios where down-sampled to 12kHz. Each convolutional layer is followed by batch normalization [21]. All convolutions use same padding. Therefore, the dimensions of the feature maps out of the first convolutional layer are still equivalent to the input – time and frequency. Then, the resulting feature map of the MP(12,16) layer can be interpreted as an eight-bands summary (96/12=8). This max-pool layer was designed considering: (i) is relevant to know in which band a given filter shape is mostly activated – as a proxy for knowing in which pitch range timbre is occurring; and (ii) is not so relevant to know when it is mostly activated. To obtain instrument predictions from the softmax layer we use the same strategy as Han et al. [6]: estimations for the same song are averaged and then a threshold of 0.2 is applied. In ***Table II*** we report the standard metrics for this dataset such as micro- and macro- precision, recall and f-beta score (f1). The micro- metrics are calculated globally for all testing examples while the macro-metrics are calculated label-wise and the unweighted average is reported.

实现在线可用。STFT用带有256自由变换的尺寸的512点FFT计算。音频下采样到12khz。每个卷积层之后批处理标准化[ 21 ]。所有的卷积使用相同的填充。因此，第一卷积层的特征映射的维度仍然相当于输入的时间和频率。然后，得到的MP（12,16）层的特征地图可以被解释为一个**eight-bands（频带？） summary**（96 / 12 = 8）。这个最大池层的设计考虑到：（i）与一个给定滤波器形状多半会被激活的频带有关，作为知道音高范围内音色发生的代理；（ii）与知道什么时候它大部分激活时不相关。**（与时间不想关于事件相关？）**我们使用和han等人的相同的策略，从softmax层获得乐器的预测。[ 6]: 对同一首歌曲的估计是平均值，然后定义它的阈值为0.2。表2是我们记录此数据集像微观和宏观的精度标准度量，recall和f-beta评分（F1）。微观指标被全面的计算为了所有测试事例；宏观指标被计算参数更新过程的优化目标，它不加权的平均被记录。



Multi-layer achieved similar results as the state-of-the-art with twice fewer #param. This result denotes how efficient are the proposed architectures. Moreover, note that small

filters are also used within the proposed architecture. We found these filters to be important for achieving state-of-the-art performance – although no instruments with such small time-frequency signature (such as kick drum sounds or bass notes) are present in the dataset. However if m=5 filters are substituted with m=50 filters, the performance does not drop dramatically. Finally note that single-layer still achieves remarkable results: it outperforms the standard bag-of-frames + SVM approach.

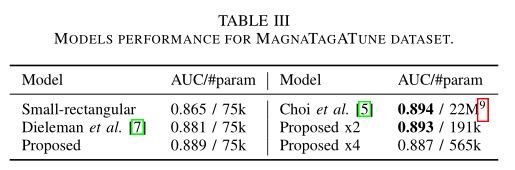
多层取得了与带两个参数的最优解类似的结果。这个结果说明了所提出的架构是如何有效的。此外，还要注意小的滤波器也需要在我们提出的结构中使用。我们发现这些滤波器对于实现最先进的性能是很重要的——尽管在数据集中没有任何具有如此小的时频特征的乐器（如鼓鼓音或低音音）存在。然而，如果m＝5个滤波器被m＝50个滤波器替换，性能不会急剧下降。最后指出单层仍然取得了显著的效果：它优于帧包标准+ SVM方法。

**Ⅲ.C. Music auto-tagging**

**Ⅲ.C. 音乐自动标注**

Automatic tagging is a multi-label classification task. We approach this problem by means of the MagnaTagATune dataset [20] – with 25.856 clips of ≈ 30 seconds sampled at 16kHz. Predicting the top-50 tags of this dataset (instruments, genres and others) has been a popular benchmark for comparing deep learning architectures [5][7]. Architectures from Choi et al. [5] and Dieleman et al. [7] are set as baselines – that are state-of-the-art examples of architectures based on small-rectangular filters and high filters, respectively. Therefore, this dataset provides a nice opportunity to explore the trade off between leaning little context with small-rectangular filters and risking to fit noise with high filters. Choi et al.’s architecture consists of a CNN of five layers where filters are of size 3×3 with an input of size 96×187. After every CNN layer, batch normalization and max-pool is applied. Dieleman et al.’s architecture has two CNN layers with filters of M×8 and M 0 ×8 size, respectively. The input is of size 128×187. After every CNN layer a max-pool layer of 1×4 is applied. Later, the penultimate layer is a fully connected layer of 100 units. An additional baseline is provided: Small-rectangular, which is an adaption of Choi et al.’s architecture to have the same input and number of parameters as Dieleman et al. All models use a 50-way sigmoidal output layer and STFT was performed using 512 points FFT with a hop size of 256.

自动标注是一种多标签分类任务。我们研究这个问题借助于magnatagatune集[ 20 ]–25.856剪辑≈30秒采样16khz。预测这个数据集的前50的标签（乐器、文体等）已经成为比较深度学习结构[ 5 ] [ 7 的]流行基准。以Choi et al. [5] 和 Dieleman et al. [7]的框架被作为基准，这俩框架分别是是基于小矩形滤波器和高通滤波器的框架实例最优。因此，这个数据集提供了一个很好的机会来探索 用小矩形滤波器学习小背景和用高通滤波器冒险匹配之间的交换。 Choi et al.的架构包括一个输入为大小96×187的3×3滤波器的5层的CNN。在每一层CNN之后，批量正常化和最大池化被启用。dieleman等人的架构有两层CNN，滤波器大小分别为：M×8 and M 0 ×8 。 输入的大小为128×187。在每一个CNN层后，有一个 1×4的最大池化被启动。稍后，倒数第二层是100个单元的完全连接层。提供一个额外的基线：小矩形，这是chio等人的一个结果。结构和dieleman等人的有相同的输入和相同的参数的数目。所有的模型都使用50路输出层，STFT采用512点256变化的FFT。

[[10]](#footnote-9)

Our experiments reproduce the same conditions as in Diele- man et al. since the proposed model adapts their architecture to the proposed design strategy – we uniquely modify the first

layer to have many musically motivated filter shapes. Other layers are kept intact. This allows to isolate our experiments from confounding factors, so that we uniquely measure the impact of increasing the representational capacity of the first layer. Inputs are set to be of size 128×187 – since input spectrograms (≈3 seconds) are shorter than the total length of the song, estimations for the same song are averaged. We consider the following frequency contexts as relevant for this dataset: m=100 and m=75 to capture different wide spectral shapes (e.g. genres timbre or guitar), and m=25 to capture shallow spectral shapes (e.g. drums). For consistency with

Dieleman et al., we consider the following temporal context: n=[1,3,5,7]. We use several filters per shape in the first layer:

我们的实验重现Diele等人的同样的条件。由于所提出的模型将其体系结构与所提议的设计策略相适应，我们独特地修改了第一层，使其具有许多音乐驱动的滤波器形状。其他层保持完整。这允许我们从混杂因素中分离实验，以便我们唯一地测量增加第一层代表能力的影响。输入设定为128×187的大小–因为输入谱图（≈3秒）小于歌曲的总长度，为同一首歌的估计平均。我们考虑以下频率背景作为相关的这个数据集：M = 100和M = 75，以捕捉不同的宽光谱形状（如流派音色或吉他），M = 25捕捉浅光谱形状（如鼓）。与dieleman等人的相同，我们考虑下面的时间背景：n=[1,3,5,7]。我们在第一层使用每个形状的多个过滤器：

· m=100: 10x 100×1, 6x 100×3, 3x 100×5 and 3x 100×7.

· m=75: 15x 75×1, 10x 75×3, 5x 75×5 and 5x 75×7.

· m=25: 15x 25×1, 10x 25×3, 5x 25×5 and 5x 25×7.

For merging the resulting feature maps, these need to be of the same dimension. We zero-pad the temporal dimension before first layer convolutions and use max-pool layers: MP(M’,4) – note that all resulting feature maps have the same dimension: 1×N 0 , and are pitch invariant. 50% dropout is applied to all dense layers. We also evaluate variants of the proposed model where the number of filters per shape in the first layer are increased according to a factor – other layers are kept intact. Implementations are available online 10 [[11]](#footnote-10)

为了合并生成的特征映射，这些需要具有相同的维度。我们在第一层卷积前在每个时间维度初始化0，并且使用最大池层：MP（m，4）–注意，所有产生的特征映射具有相同的尺寸：1×N 0，音高不变。50%dropout被用于所有致密层。我们还评估了该模型的变体，其中第一层的每一个形状的滤波器数量根据一个因素而增加，其他层保持不变。实验可在线

We use area under the ROC curve (AUC) as metric for our experiments. ***Table III*** (left column) shows the results of three different architectures with the same number of parameters. The proposed model outperforms others, denoting that architectures based on the design strategy we propose can better use the capacity of the network. Moreover, ***Table III*** (right column) shows that is beneficial to increase the representational capacity of the first layer – up to the point where we achieve equivalent results to the state-of-the-art while significantly reducing the *#param* of the model.

我们使用ROC曲线(AUC) 下的面积作为衡量我们实验的指标。表三（左栏）显示了三个相同参数数目的不同体系结构的结果。该模型优于其他模型，表示基于我们提出的设计策略的架构可以更好地利用网络的容量。此外，表三（右列）表明，有利于提高第一层的表达能力----直到我们达到和最优解相同的结果，同时大大降低了模型的#参数。

**IV. C ONCLUSIONS**

1. **结论**

Inspired by the fact that it is hard to identify the adequate combination of parameters for a deep learning model –which leads to architectures being difficult to interpret–, we decided to incorporate domain knowledge during the architectural design process. This lead us to discuss some common practices when designing CNNs for music classification – with a specific focus on how to learn timbre representations. This discussion motivated the design strategy we present for modeling timbre using spectroram-based CNNs. Several ideas were proposed to

achieve pitch, loudness, duration and spatial position invariance with CNNs. Moreover, we proposed actions to increase the efficiency of these models. The idea is to use different filter shapes in the first layer that are motivated by domain knowledge – namely, using different musically motivated filter shapes in the first layer. A part from providing theoretical discussion and background for the proposed design strategy, we also validated it empirically.

我们从深度学习模模型难以确定参数组合中获得启发（这导致了架构难以被诠释），我们决定在建模设计过程中纳入领域知识。这导致我们当设计CNN音乐分类时讨论一些通用的方法–而这类方法又专注于如何学习音色表现。这个讨论推动了我们提出了使用基于CNN的图谱建模音色的设计策略。几个思路用CNN实现音高、响度、持续时间和与的空间位置不变。此外，我们还提出了提高这些模型效率的措施。其思想是在第一层中使用不同的滤波器形状，这是由领域知识驱动的，也就是说，在第一层使用不同的音乐驱动的的滤波器形状。在为设计策略提供理论上讨论和背景这一部分中，我们也通过经验验证了它。

Several experiments in three datasets for different tasks related to timbre (singing voice phoneme classification, musical instrument recognition and music auto-tagging) provide empirical evidence that this approach is powerful and promising. In these experiments, we evaluate several architectures based on the presented design strategy – that has proven to be very effective in all cases. These results support the idea that increasing the representational capacity of the layers can by achieved by using different filter shapes. Specifically, proposed architectures used several filter shapes having the capacity of capturing timbre with high enough filters. Moreover, we found very remarkable the results of the proposed single-layer architectures. Since single-layer architectures use a reduced amount of parameters, these might be very useful in scenarios where small data and a few hardware resources are available. Furthermore, when deepen the network we were able to achieve equivalent results to the state-of-the-art – if not better. As future work we plan to relate these findings with previous research (where a similar analysis was done for designing CNNs for modeling temporal features [1]), to extend this work to non-musical tasks, and to inspect what filters are learning.

在三个不同的数据集上进行的有关音色的测试（歌唱语音音素分类、乐器识别和音乐自动标注）的实验提供了有效的证明这种方法是强大的和有前途的。在这些实验中，我们评估了几种基于所提出的设计策略的架构，这些策略在所有情况下都非常有效。这些结果支持了通过使用不同的滤波器形状来增加层的代表能力的想法。具体地说，所提出的体系结构使用了具有足够高的滤波器捕获能力的多个滤波器形状。此外，我们发现所提出的单层结构的结果非常显著。由于单层架构使用的参数数量减少，在小数据和少量硬件资源可用的情况下，这些结构可能非常有用。此外，在深化网络后，我们可能把相同的结果实现最优化，如果没有更好的。在未来的工作中，我们计划把这些发现和先前的发现（为建模时间特征做一个相似的分析，设计CNNs ）联系起来，把这项工作扩张到没有音乐的测试中，并且去检查什么滤波器正在学习。

**REFERENCES**

1. **参考文献**

[1] J. Pons and X. Serra, “Designing efficient architectures for modeling temporal features with convolutional neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2017.

[2] A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra, “Acous- ticbrainz: a community platform for gathering music information obtained from audio,” in International Society for Music Information Retrieval Conference, 2015.

[3] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77,no. 2, pp. 257–286, 1989.

[4] A. Roebel, J. Pons, M. Liuni, and M. Lagrange, “On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2015.

[5] K. Choi, G. Fazekas, and M. Sandler, “Automatic tagging using deep convolutional neural networks,” in International Society of Music Information Retrieval, 2016.

[6] Y. Han, J. Kim, and K. Lee, “Deep convolutional neural networks for predominant instrument recognition in polyphonic music,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1,pp. 208–221, 2017.

[7] S. Dieleman and B. Schrauwen, “End-to-end learning for music audio,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[8] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,”in Advances in Neural Information Processing Systems, 2009.

[9] J. Pons, T. Lidy, and X. Serra, “Experimenting with musically motivated convolutional neural networks,” in Content-Based Multimedia Indexing,2016.

[10] K. Choi, J. Kim, G. Fazekas, and M. Sandler, “Auralisation of deep convolutional neural networks: Listening to learned features,” in International Society of Music Information Retrieval, Late-Breaking/Demo Session, 2015.

[11] H. Phan, L. Hertel, M. Maass, and A. Mertins, “Robust audi event recognition with 1-max pooling convolutional neural networks,”

[12] D. L. Wessel, “Timbre space as a musical control structure,” Computer Music Journal, pp. 45–52, 1979.

[13] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams,“The timbre toolbox: Extracting audio descriptors from musical signals,” The Journal of the Acoustical Society of America, vol. 130, no. 5, pp.2902–2916, 2011.

[14] S. McAdams, “Musical timbre perception,” The psychology of music,pp. 35–67, 2013.

[15] D. O. Hebb, “The organization of behavior: A neuropsychological approach,” 1949.

[16] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv:1511.07289,2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,”arXiv:1502.01852, 2015.

[18] D. A. Black, M. Li, and M. Tian, “Automatic identification of emotional cues in chinese opera singing,” in International Conference on Music Perception and Cognition and Conference for the Asian-Pacific Society for Cognitive Sciences of Music, 2014.

[19] J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, “A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals,” in International Society for Music Information Retrieval (ISMIR), 2012.

[20] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging.” in International Society for Music Information Retrieval, 2009.

[21] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning, 2015.

1. See first paragraph of Section II for a formal definition of timbre. [↑](#footnote-ref-0)
2. CNNs input is set to be log-mel spectrograms of dimensions M×N and the CNN filter dimensions to be m×n. M and m standing for the number of frequency bins and N and n for the number of time frames. [↑](#footnote-ref-1)
3. Section II further expands this discussion with more details. [↑](#footnote-ref-2)
4. N’ and M’ denote, in general, the dimensions of any feature map.Therefore, although the filter map dimensions will be different depending on the filter size, we refer to these dimensions by the same name: N’ and M’. [↑](#footnote-ref-3)
5. Jingu” is also known as “Beijing opera” or “Peking opera”. [↑](#footnote-ref-4)
6. Annotation and more details can be found in: <https://github.com/MTG/jingjuPhonemeAnnotation> [↑](#footnote-ref-5)
7. Dropout是指在模型训练时随机让网络某些隐含层节点的权重不工作，不工作的那些节点可以暂时认为不是网络结构的一部分，但是它的权重得保留下来（只是暂时不更新而已），因为下次样本输入时它可能又得工作了 [↑](#footnote-ref-6)
8. <https://github.com/ronggong/EUSIPCO2017> [↑](#footnote-ref-7)
9. https://github.com/Veleslavia/EUSIPCO2017 [↑](#footnote-ref-8)
10. Although equivalent results can be achieved with 750k parameters. [↑](#footnote-ref-9)
11. https://github.com/jordipons/EUSIPCO2017 [↑](#footnote-ref-10)