Going Deep with Segmentation of Field Recordings

Matija Marolt
University of Ljubljana
matija.marolt@fri.uni-lj.si

ABSTRACT

In the paper, we explore the performance of deep residual convolutional networks for labelling ethnomusicological field recordings. Field recordings are integral documents of folk music performances captured in the field, and typically contain performances, intertwined with interviews and commentaries. As these are live recordings, captured in non-ideal conditions, they usually contain significant background noise. Labelling of field recordings is a typical step in segmentation of these recordings, where short sound excerpts are classified into one of a set of predefined classes. In the paper, we explore classification into four classes: speech, solo singing, choir singing (more than one voice) and instrumental performances. We describe the dataset gathered for the task and the labelling tools developed for gathering the reference annotations. We compare different input representations and convolutional network architectures based on residual modules for labelling short audio segments and compare them to the more standard feature based approaches, where an improvement in classification accuracy of over 5% was obtained.

1. INTRODUCTION

Field recordings are documents of folk song and music performances taken "in the field", usually in environments familiar to musicians. They aim to preserve entire recording sessions and the context in which they were recorded, and are thus a mix of performances and speech, which often consists of interviews with musicians. Since recordings are taken in everyday environments, they are often very noisy due to background noise (e.g. people talking, doors closing etc.), poor recording equipment or the recording environment itself. Segmentation of field recordings is one of the first tasks that ethnomusicologists perform when studying the recorded materials, as they separate the contents into different units, such as speech or individual performances. It is also a prerequisite for computational analysis of field recordings.

In the audio processing and music information retrieval research fields, automatic segmentation of recordings is a well-studied task. It is important for segmentation of broadcast news and radio broadcasts, where recordings are usually separated into speech and music units, as well as in other domains such as for removal of non-speech parts in speech recognition systems. Most approaches either first label short segments of the recording into a set of classes (e.g. speech, music) and then find segment boundaries (Lie, Stan, & Hong-Jiang, 2001; Williams & Ellis, 1999), or first find the segment boundaries and later apply classification into classes (Panagiotakis & Tziritas, 2005; Tzanetakis & Cook, 1999). Pikrakis et al. (2008) used a three step approach: first they identified regions in the signal which are very likely to contain speech or music with a region growing algorithm. Then, they segmented the remaining regions with a maximum likelihood model and finally, a boundary correction algorithm was applied to improve the found boundaries. Marolt (2009) also used a three step procedure where signal fragments were first labelled into five classes, then candidate boundaries established and finally the actual boundaries estimated with a maximum-likelihood criterion.

More recently, within the Mirex 2015 Music/Speech Classification and Detection task ("Mirex 2015 Results," 2015), 9 authors submitted their algorithms for classifying recordings into either speech or music, and for finding segment boundaries in a set recordings, which also included a number of field recordings. The algorithms were very successful for the first task, reaching 99.7% accuracy (Lidy, 2015), which might indicate that the task of music/speech classification is *solved*, however it is more likely that the evaluation dataset was too basic and did not include enough challenging cases for the algorithms. This is already obvious if we observe results of the same approaches for the second task, where frame-based F1 measure of the best system dropped to 89.4% (Marolt, 2009), while the F1 score of finding segment boundaries was only at 40.3%.

In the past years, deep learning had become the prevalent approach for classification problems in image and audio domains. It is therefore not surprising that it was also applied to segmentation of audio recordings. The aforementioned best music/speech classifier at Mirex 2015 by Lidy (2015) was based on convolutional neural networks. Similarly, Kruspe et al. (2017) use deep networks to discriminate between speech and music sections in broadcast signals and reports over 99% F1 measure for speech and 91% for music discrimination. Authors from Google (Hershey et al., 2017) compared a number of deep architectures for large-scale audio classification on tagged audio from the YouTube-100M dataset, as well as on a large scale dataset of labelled sound clips from YouTube videos – Audio Set (Gemmeke et al., 2017).

In this paper, we explore deep neural networks for labelling ethnomusicological field recordings. Unlike broadcast recordings, field recordings are more challenging to label and segment due to their noisy nature. In contrast to most speech/music discriminators, we aim to separate between four rather than two classes: speech, solo singing, choir singing (more than 1 voice) and instrumental recordings. We chose the four classes as they are very representative for a number of field recordings from different regions that we analyzed. Also, in contrast to most previous work, we do not aim to segment (clean) broadcast recordings, but field recordings, which may be of varying quality, as already described previously. We describe the architecture used for classification, the dataset used and our first results.

2. DATASET

Exploration of field recordings revealed four major classes of recordings that appear in a variety of cultures: solo singing, choir (more than one voice) singing, instrumental performances, and speech. Our goal was therefore to classify field recordings into the four classes, and not to limit ourselves to just speech and music. To train deep learning classifiers, large datasets are needed - the larger the better as recent deep learning experiences show. Apart from the Audio Set (Gemmeke et al., 2017), which is an excellent large-scale audio classification dataset, there are few suitable datasets available for the task. In the presented work, we decided not to begin with the Audio Set, as its categories are not ideal for our purpose; for example, there is no solo singing category, examples labeled with singing are mostly accompanied by music, while musical genres are mostly oriented towards popular music genres (pop, rock etc.).

We therefore gathered short excerpts from a variety of recordings from ethnomusicological (and related) archives that put their collections online in recent years. The sources include: the British Library world & traditional music collection[1], Alan Lomax recordings[2], sound archives of the CRNS[3] and a number of recordings from the Slovenian sound archive Ethnomuse and the Norwegian national library, which are not available online, but were made available to us by ethnomusicologists with the respective institutions. These field recordings were augmented by the well-known GTZAN music/speech collection and the Mirex 2015 music/speech detection public dataset.

Altogether 7,000 5 second long excerpts were extracted from these sources. To manually label them into the four target classes, we enhanced the web-based audio annotator tool (Cartwright et al., 2017), so that it can be controlled exclusively by the keyboard. This enabled fast multi-user annotation of audio excerpts into the four categories, augmented by three additional categories of "voice over instrumental", "noise" and "not clear". The latter was to be applied when the audio clip was either too noisy to be recognized or contained too many short fragments of different types of materials, so that it was difficult to select a single label. The annotator's goal was namely, to select a single label for the five second clip, where clips were randomly chosen from the set of unlabeled clips for each participating annotator. The user interface was kept very similar to the original audio annotator and is shown in Figure 1.

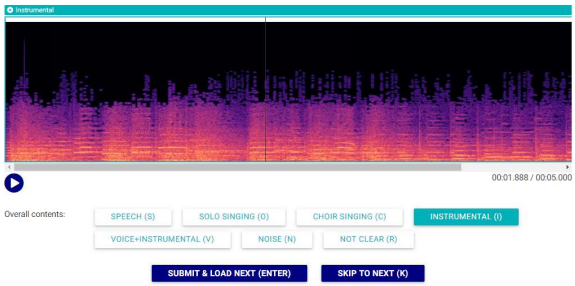

Figure 1. The annotation interface.

3. EXPERIMENT

Our goal was to evaluate the performance of deep networks for the classification task at hand. All the audio excerpts were first downsampled to 22050 Hz, mixed to a single channel and normalized.

We compared several input representations for the task: 46 ms FFT frames (252 bins between 50 and 7000 Hz) and 64 channel mel-scale spectrograms (50-8000 Hz) extracted from FFT frames of 23ms, 46 ms, and 92ms. We log-scaled all representations (adding 1e-5 before applying the logarithm) and used 1 or 2 second long feature blocks with 50% overlap as network inputs. Stacking of different resolution frames (23ms, 46ms, 96ms) was also tested.

We chose convolutional deep networks as our main classification tool and focused specifically on residual networks (K. He, Zhang, Ren, & Sun, 2015), which previously demonstrated good performance for a variety of image, as well as audio-based tasks. The main feature of residual networks are their shortcut connections that implement identity mappings and enable convolutional blocks to learn residuals between the underlying mapping of features and the input.

The overall network architecture is shown in Figure 2. The input layer is first processed by m nxn convolutions, optionally enhanced with m nxn dilated convolutions with rate 2, to expand the receptive field of filters. A max pooling layer was added to reduce the size of feature maps, followed by p resnet v2 blocks (Kaiming He, Zhang, Ren,

[1] https://sounds.bl.uk//World-and-traditional-music
[2] http://research.culturalequity.org/home-audio.jsp
[3] http://archives.crem-cnrs.fr/

& Sun, 2016), where the size of feature maps is halved (in each dimension) within each block and the number of filters doubled. The batch normalized output of resnet blocks is gathered by 1x1 convolutions into a 2D feature map. The map is finally processed by a small fully connected layer with four outputs, where the softmax activation yields final class probabilities. We tested different values for the described parameters, which we outline in section 4. Batch normalization, as well as l2 regularization were used for regularizing the network, to avoid overfitting. To introduce non-linearity, we compare the performance of standard ReLU activation functions with exponential linear units ELU (Clevert, Unterthiner, & Hochreiter, 2015).

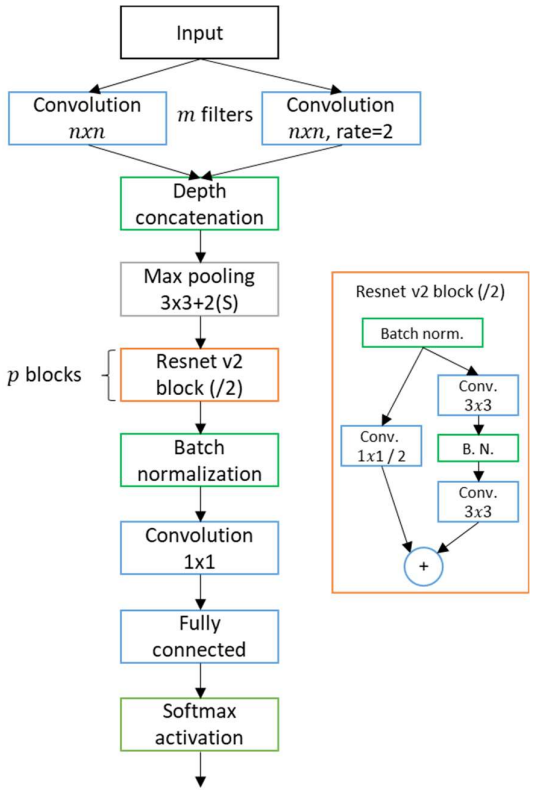

Figure 2. The network architecture.

Three-fold cross validation was used to assess the performance of each network, where 2/3 of the dataset was used for training, the remaining 1/3 for testing, and the procedure repeated three times. The networks were trained with minibatches of 128 examples. For each audio example, the block of input features was drawn from a random location within the audio, so that for each epoch, the feature blocks used to train the network differed in their location within training files. Such *time translation* diversifies the limited training data available and improves performance, as was also demonstrated elsewhere (Jansen et al., 2017). For testing, the entire test files were used.

Stochastic gradient descent was used for training over 500 epochs, and the learning rate set to decay from 0.1 by 0.75 each 500 steps. The experiments were implemented in Tensorflow.

4. RESULTS

4.1 Input representation

A comparison of different input representations is shown in Table 1. We report average accuracy over all classes over the three cross-validation splits in the last column. The same network architecture (described in 4.2) was used for all comparisons. We compare two different input representations: mel compressed spectrograms vs. FFT, two different block sizes (1.1 vs. 2.2 second long blocks of input features), three different window and two different step sizes for FFT calculation.

We see a significant difference only in the choice of block sizes: features covering 1.1 seconds of audio give around 2% lower accuracy as 2.2 second blocks, indicating that it is beneficial for the network to have more context in order to distinguish between the categories. Indeed, even when listening to, for example speech vs. solo singing, in many cases one second of audio cannot not reveal the correct category. This is even truer for field recordings, which are typically amateur performances, many times by older people, include strong dialects etc. There are no significant differences between different window and step sizes in 2.2 second blocks. Stacking of different window sizes also does not improve the performance significantly. We therefore decided to use 2.2 second blocks of 64 channel mel spectrograms calculated from FFT frames of 46 ms with 23ms step size (network input size 96x64) in our further experiments.

feature	*block (s)*	*step (ms)*	*window (ms)*	*input size*	*accuracy*
mel	1.1	12	12	96x64	0.861
			23	96x64	0.868
			46	96x64	0.868
			12,23,46	96x64x3	0.875
	2.2	12	12	192x64	0.882
			23	192x64	0.887
			46	192x64	0.887
			12,23,46	192x64x3	0.891
		23	23	96x64	0.886
			46	**96x64**	**0.890**
			92	96x64	0.891
			12,23,46	96x64x3	0.895
fft	2.2	23	46	96x252	0.892

Table 1. A comparison of different input representations.

4.2 Network architectures

The overall network architecture was described in section 3. We tested the influence of the following parameters on network performance: the number of filters in the first convolutional layer (2, 4, 6, 8), the sizes of these filters (4, 6, 8, with or without stacked dilated convolutions of the same size), the number of resnet blocks (3, 4, 5) and the activation function (ReLU vs ELU). Table 2 lists the key results.

The networks are not very sensitive to the size of input filters. When the number of layer one filters m increases

up to 6 filters, performance improves, while higher numbers do not have a large effect. Adding an additional set of dilated filters (rate=2) helps, although this also increases the number of network parameters. The optimal number of resnet blocks was determined to be 4, an additional block does not add much to accuracy, but increases the number of network parameters substantially. The ELU activation function seems to improve training (consistently higher accuracy by approx. 1%) over ReLU.

activation	*dilated*	*nxn L1 size*	*m L1 filters*	*p resnet blocks*	*accuracy*
elu	yes	4	6	5	0.893
				4	**0.890**
				3	0.882
			2	4	0.874
			4	4	0.881
	no	4	6	4	0.883
relu	yes	4	6	4	0.882

Table 2. Comparison of network architectures.

Based on the evaluation, our final network architecture uses ELU activations, 6 4x4 convolutions stacked with 6 dilated 4x4 convolutions (rate=2) on the first layer, followed by 4 resnet blocks. The final fully connected layer is small (24x4) and has no hidden layer, but directly maps into the four outputs. The entire network is not very deep, as we have a limited amount of training data, and contains 172,936 trainable parameters.

4.3 Comparison to other approaches

To put the obtained results into perspective, Table 3 lists the performance of three other approaches on the same dataset (also using 3-fold cross-validation):

- a *standard* deep convolutional network with two 3x3 convolutions (one with stride 2) in place of each resnet block (no shortcut links), trained on the same mel-spectrogram input data representation;
- a multilayer perceptron with one hidden layer of 16 neurons trained on VGGish (Hershey et al., 2017) features extracted from the data. VGGish are audio classification features extracted from a VGG-like deep model trained on a large YouTube dataset and made available by Google. Input to the MLP consisted of two consecutive 128-dimensional VGGish vectors, each summarizing 1 second of audio;
- a simple logistic regression model trained on handcrafted features, as described in (Marolt, 2009).

model	*number of parameters*	*accuracy*
proposed resnet	172,936	**0.890**
standard deep	166,556	0.862
MLP on VGGish	4,180	0.881
logistic regression	51	0.837

Table 3. A comparison to other approaches.

The proposed model outperforms the others. It has the highest number of trainable parameters, however care has been taken to avoid overfitting by including batch normalization and l2 regularization during training, as well as using 1/3 of the dataset for testing at each run, so it is safe to assume that its performance is realistic for a wide variety of materials. VGGish features come close second.

An analysis of errors showed many *logical* mistakes, which can be attributed to several factors. First, some of the recordings are very noisy and even a human listener can have some difficulty to discern the contents. Such recordings were often mistakenly classified as instrumentals, as the noise was considered part of the performance.

The confusion matrix in Table 4 shows that many mistakes are made between *neighboring* classes: solo singing is misclassified as choir singing or speech, choir mostly as solo, instrumentals as choir or speech as solo. Some confusions may be due to the particularity of the contents, e.g. some short excerpts of dialectal speech may sound very much like singing. Some mistakes are not really mistakes – an excerpt may be correctly classified, and wrongly labelled. Namely each five second audio clip in our dataset is only labelled with a single class, even though parts of it may contain another class. An example is a choir recording, where some parts are sung solo and then evolve into choirs. As the network only classifies short 2 second excerpts, it may correctly label the solo part as solo, however the entire example is labelled as choir, so this is considered a misclassification. Choir parts sung in unison are another case that is difficult to classify – they are labelled as choir singing in our dataset, but may sound very similar to solo singing.

The final trained network is integrated into the publicly available SeFiRe tool for segmentation of field recordings[1].

		predicted			
		solo	choir	instr.	speech
true	solo	0.87	0.07	0.01	0.05
	choir	0.06	0.89	0.02	0.03
	instr.	0.02	0.04	0.92	0.02
	speech	0.06	0.01	0.01	0.92

Table 4. The confusion matrix.

5. CONCLUSION

In the paper, we demonstrated the performance of a medium sized deep convolutional network applied to classification of field recordings into four classes. We also provide a comparison of different input representations and network architectures for the task. The database used and the final trained model will be made available to the community.

In our future work, we will aim to enhance the dataset with additional sources of field recordings. We will also

[1] http://lgm.fri.uni-lj.si/portfolio-view/sefire/

make use of the Audio Set, currently the largest annotated audio classification dataset, to enlarge our training data. Our second goal is to increase the number of target categories into typical instrument categories and introduce non-exclusive categories (e.g. singing over instrumental), which are currently labeled as instrumentals.

6. REFERENCES

Cartwright, M., Seals, A., Salamon, J., Williams, A., Mikloska, S., MacConnell, D., . . . Nov, O. (2017). Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact., 1*(CSCW), 1-21. doi:10.1145/3134664

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR, abs/1511.07289*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., . . . Ritter, M. (2017). *Audio Set: An ontology and human-labeled dataset for audio events*, New Orleans, LA.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv e-prints*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. *CoRR, abs/1603.05027*.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., . . . Wilson, K. (2017). CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jansen, A., Plakal, M., Pandya, R., Ellis, D. P. W., Hershey, S., Liu, J., . . . Saurous, R. A. (2017). Unsupervised Learning of Semantic Audio Representations. *CoRR, abs/1711.02209*.

Kruspe, A., Zapf, D., & Lukashevich, H. (2017). *Automatic speech/music discrimination for broadcast signals*.

Lidy, T. (2015). *Spectral Convolutional Neural Network for Music Classification*. Paper presented at the Mirex 2015, Malaga, Spain.

Lie, L., Stan, Z. L., & Hong-Jiang, Z. (2001). *Content-based audio segmentation using support vector machines*. Paper presented at the IEEE International Conference on Multimedia and Expo.

Marolt, M. (2009). *Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings.* Paper presented at the ISMIR, 10th International Society for Music Information Retrieval Conference, Kobe, Japan.

Mirex 2015 Results. (2015). Retrieved from http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection

Panagiotakis, C., & Tziritas, G. (2005). A speech/music discriminator based on RMS and zero-crossings. *Multimedia, IEEE Transactions on, 7*(1), 155-166.

Pikrakis, A., Giannakopoulos, T., & Theodoridis, S. (2008). A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks. *Multimedia, IEEE Transactions on, 10*(5), 846-857.

Tzanetakis, G., & Cook, P. (1999). *Multifeature audio segmentation for browsing and annotation.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on.

Williams , G., & Ellis, D. P. W. (1999). *Speech/music Discrimination Based On Posterior Probability Features.* Paper presented at the Eurospeech'99, Budapest, Hungary.