SIMILARITY MEASURES FOR VOCAL-BASED DRUM SAMPLE RETRIEVAL USING DEEP CONVOLUTIONAL AUTO-ENCODERS

Adib Mehrabi, Keunwoo Choi, Simon Dixon, Mark Sandler

Queen Mary University of London, London, UK
Centre for Digital Music, EECS
E1 4FZ, London, UK
{a.mehrabi, keunwoo.choi}@qmul.ac.uk

ABSTRACT

The expressive nature of the voice provides a powerful medium for communicating sonic ideas, motivating recent research on methods for query by vocalisation. Meanwhile, deep learning methods have demonstrated state-of-the-art results for matching vocal imitations to imitated sounds, yet little is known about how well learned features represent the perceptual similarity between vocalisations and queried sounds. In this paper, we address this question using similarity ratings between vocal imitations and imitated drum sounds. We use a linear mixed effect regression model to show how features learned by convolutional auto-encoders (CAEs) perform as predictors for perceptual similarity between sounds. Our experiments show CAEs outperform three baseline feature sets (MFCCs, temporal features and spectrogram-based representations) at predicting the subjective similarity ratings. We also investigate how the size and shape of the encoded layer affects the predictive power of the learned features. The results show preservation of temporal information is more important than spectral resolution for this application.

Index Terms— vocalisation, audio similarity, convolutional neural networks, auto-encoders

1. INTRODUCTION AND RELATED WORK

Searching for audio samples is a core part of the electronic music making process, yet is a time consuming task, and a key area for future technological development [1]. This task typically involves browsing lists of badly labelled files, relying on filenames such as 'big_kick' or 'hi-hat22'. Such methods for browsing sound libraries limit the users' ability to efficiently find the sounds they are looking for. Meanwhile, the voice provides an attractive medium for effectively communicating sonic ideas [2, 3], as it can be used to express timbral, tonal and dynamic temporal variations [4]. Moreover, previous research demonstrates that musicians are able to accurately vocalise important acoustic features of musical sounds [5, 6].

Query by vocalisation (QBV) is the process of searching for sounds based on vocalised examples of the desired sound. Typically, QBV systems extract audio features from a vocalisation, which can then be compared to the features of sounds in a sample library (to return class labels or a ranked list of sounds). Initial approaches to QBV used heuristic based features [7, 8]. Morphological features describing the high-level temporal evolution of sounds have also been applied to QBV [9], however drum sounds generally have similar high-level temporal morphology (i.e. rise-fall), so these types of features are less applicable here.

Recent work has shown that learned features using stacked auto-encoders (SAEs) outperform heuristic descriptors such as MFCCs (Mel-frequency cepstral coefficients) in 2 common scenarios: supervised learning, using features to train a classifier [10]; and unsupervised search using distance in a feature vector space [11, 12]. SAEs utilise a deep learning structure where multiple layers learn an efficient representation to encode the input. Furthermore, in [13] the authors present a QBV system based on convolutional neural networks (CNNs) implemented in a semi-Siamese network structure. In this case the convolutional layers are trained to learn feature representations from constant Q spectrograms (CQT) of vocal imitations and the imitated sounds. The CNN is followed by fully connected layers to match input vocalisations to audio samples, requiring each sample in a sound library to be compared to a vocal query. The system shows promising results for matching vocal imitations to the imitated sounds, however in the general case QBV systems require efficient, deployable querying. Using this method a single query on a dataset with N data samples requires N forward-pass computations of the network, which is significantly demanding, for example compared to nearest neighbour search in a feature vector space.

Whilst both SAE and CNN approaches show promising performance in terms of retrieving an *imitated* sound from a set of audio samples, none of the above mentioned QBV methods consider the *perceptual similarity* between the query and retrieved sounds. Central to the evaluation of these approaches is the assumption that the target sound is indeed the sound that was imitated, and the task is to match the imitations and imitated sounds accordingly. However, we consider a use case in which the query is not necessarily an imitation of a sound in the database, and investigate which feature representations correlate well with the perceptual similarity between an imitation and a set of audio samples.

In this paper, we address this by evaluating the performance of both heuristic and learned features for QBV of drum sounds. An overview of our approach is illustrated in Fig. 1. We present a set of convolutional auto-encoders (CAEs) trained on a dataset of $\sim$ 33k audio samples and $\sim$ 6k vocalisations. These are used to extract features from 420 vocal imitations of 30 drum sounds. The feature sets are evaluated using perceptual similarity ratings between the vocal imitations and the imitated drum sounds, provided by a group of 51 listeners. We include 4 sets of features: (1) MFCCs; (2) temporal descriptors; (3) a spectrogram based representation from [14], which the authors show to correlate strongly with perceptual similarity between drum sounds; (4) encoded representations from the CAEs. We compare 11 CAEs, which differ in both the size of the encoded feature tensor and the shape of the encoded layer in the temporal and spectral dimensions.

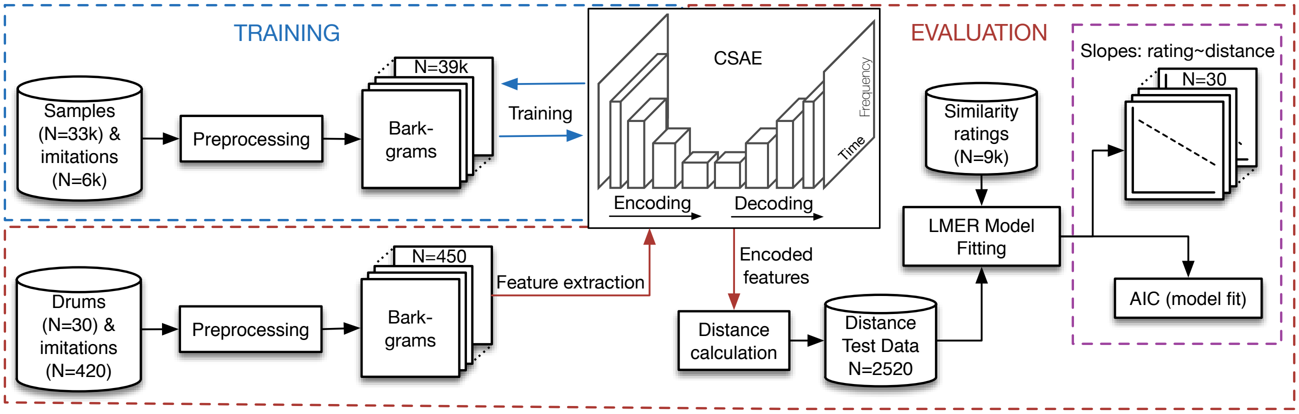

Fig. 1: Overview of the complete work flow. All audio (training and test data) is preprocessed to create 128x128 barkgram representations. The trained CAE is used to extract features from the test data. Euclidean *distance* between each imitation and its imitated sound is then computed, and fitted with the *rating* data to an LMER model. Performance of the 14 feature sets (3 baselines and 11 CAE networks) is measured by 1) AIC for model fit, and 2) the proportion of imitated sounds that have a significantly negative slopes for $rating \sim distance$.

2. PROBLEM DEFINITION

The task is to establish which audio features best correlate with perceptual similarity between real drum sounds (the *imitated sounds*) and vocal imitations of drum sounds (the *imitations*). Specifically, we are interested in i) how heuristic descriptors perform compared to learned features using CAEs, and ii) the importance of temporal vs spectral dimensions and the size of the encoded tensors from the CAEs. We limit the problem to a set of 30 drum sounds: 6 from each of 5 classes (kick, snare, cymbal, hi-hat, tom-tom), and consider only the similarity between the imitations and within-class sounds (i.e. similarity between the imitation of a kick drum and the 6 kick drum sounds).

3. EXPERIMENTS

3.1. Baseline Methods

We use 3 baseline methods: one spectral, one temporal and one that combines both aspects. The first, and main baseline of interest (PK08) is a simple spectrogram-based model of similarity from [14]. This has been shown to provide feature spaces that correlate highly with listeners' perceptual similarity ratings between within-class drum sounds, and we are interested in how well it transfers to our application. In summary, similarity between two sounds is measured as the Euclidean distance between their vectorised barkgrams, constructed from a spectrogram with the following parameters: 93ms time window; 87.5% overlap; Bark scale (72 bins); loudness in dB and scaled using Terhardt's model for the outer and middle ear [15]. The barkgrams are time-aligned, and where two sounds are not of the same length the shorter barkgram is zero padded to the length of the longer one. This essentially treats the flattened barkgram as a feature vector, encapsulating both spectral and temporal aspects of the sounds.

For the second method (MFCC) we calculate the first 13 MFCCs for each sound (excluding MFCC 0) with first and second order derivatives, using a 93ms time window and 87.5% overlap. The mean of each MFCC and its derivatives are calculated for each sound, yielding 39 features. The third method (TEMP) is a set of 5 temporal features: log attack time (LAT); temporal centroid (TC); LAT/TC ratio; temporal crest factor (TCF) and duration. We calculate LAT and TC as per the definition in [14]. TCF is calculated over the entire time domain signal, and is the maximum value divided by the root mean squared.

3.2. CAE Networks

3.2.1. *Model Architecture*

The basic architecture is a CAE with four 2D convolution layers in its encoder/decoder. Each convolutional layer is followed by batch normalisation and (ReLU) activation layers. To avoid checker board artefacts caused by deconvolution layers [16] we apply upsampling prior to each decoding convolutional layer. As such, each decoding deconvolution layer is an upsampling layer followed by a 2D convolution layer with $(1, 1)$ stride.

We vary the kernel size of the first and last layers while using fixed $(10, 10)$ kernels for the other convolution layers. The encoding layers have [8, 16, 24, 32] kernels (layers 1-4 respectively) which is mirrored in the decoder, i.e., [32, 24, 16, 8]. Finally, a single-channel convolution layer is used as an output layer.

The kernel size is varied in order to compare the shape of the encoded representation (i.e. square, wide, tall) and how this interacts with the shape of the kernels over layers. We present the results from 11 variants of the above model, as per Table 1. The activation of the last layer of the encoder is flattened into a 1D vector which is used as the feature vector.

3.2.2. *Training Data and Pre-processing*

The network is designed to learn a wide range of vocal and percussion related sounds including *i)* short, percussive/non-percussive and pitched/unpitched sounds, and *ii)* a wide range of vocalisations. The training dataset is made up of 24,294 percussion sounds, 4,884 sound effects and 4,523 single note instrument samples. In addition, we included 4,429 vocal imitations of instruments, synthesisers and everyday sounds from [17], and 1,387 vocal imitations of 72 short synthesised sounds from [6]. This results in a dataset of $\sim$ 39k sounds, of which $\sim$ 6k are vocal imitations.

For each sound in the training set we compute the barkgrams from spectrograms with a 93 ms time window and 87.5% overlap, using 128 Bark bins. As with the PK08 baseline, the magnitudes are modified via decibel scaling and Terhardt's ear model curves [15]. To achieve a fixed size representation for all sounds, we either zero-pad or truncate the barkgrams to 128 frames ($\approx$ 1.5 seconds).

3.2.3. *Training Procedure*

The models are implemented using Keras [18] and Tensorflow [19]. Training and validation sets are 70%:30% split from the training data (Section 3.2.2). As the training dataset contains 5.5 times more audio samples than vocal imitations, and we are equally interested in learning both sound types, we specify a 50%/50% split of audio samples/vocal imitations for each batch (128 data samples). The models were all fitted using the Adaptive Moment estimation (ADAM) optimiser [20] with a learning rate of 0.001, and mean squared error loss function. We use the early-stopping scheme for no improvement in validation loss after 10 epochs. The best (i.e. lowest validation loss) model for each parameter setting is selected for the analysis.

4. EVALUATION

4.1. Test data

The 30 drum sounds were taken from the fxpansion[1] *BFD3 Core* and *8BitKit* sample libraries, which include a range of acoustic and electronic drum samples. Vocal imitations of each sound were recorded by 14 musicians (>5 years experience), giving 420 imitations. The recordings took place in an acoustically treated room at the Centre for Digital Music, Queen Mary University of London[2].

Perceptual similarity ratings between the imitations and each of the within-class drum sounds were collected from 63 listeners via a web based listening test, using a format based on the MUSHRA protocol for subjective assessment of audio quality [21]. Each listener was presented with 30 tests. For each test the listener was presented with a (randomly selected) vocal imitation and the 6 within-class drum sounds (one being the imitated sound). The listener then rated the similarity between the imitation and each drum sound (giving 6 similarity ratings per test), on a continuous scale from 'less similar' to 'more similar'. Although the MUSHRA standard is typically used to compare a set of sounds to a reference, as we do here, it also provides an inherent ranking of and pairwise comparison between the test sounds [22].

Of the 30 test pages, 28 were unique and 2 were random duplicates. These were included for post-screening of the listeners, as recommended in the MUSHRA standard [21]. Listener reliability was assessed using the Spearman rank correlation between the two duplicate test pages for each listener. We considered reliable listeners as those who were able to replicate their responses for at least one of the duplicates with $\rho >= 0.5$, i.e. large positive correlation [23]. There were 51 reliable listeners, for whom ρ = 0.63/0.04 (mean/standard error), giving 9,126 responses from 1521 tests (excluding duplicates). We then computed Kendall's coefficient of concordance, W [24] on the ranked responses for each imitation. The mean/standard error of W = 0.61/0.01, indicating moderate to strong agreement amongst the reliable listeners [25].

Analysis of the ratings indicated that listeners were able to correctly identify the imitated sound with above chance accuracy (37% of cases, chance = 16%), and the imitated sound was rated first or second most similar to the imitation in 60% of tests. This indicates that although the imitations were often rated as being most similar to the imitated sounds, there are a considerable number of cases (up to 40%) where 2 of the 6 within-class sounds were rated more similar to the imitation than the imitated sound. This highlights the potential importance of perceptual similarity measures for tasks such as QBV, depending on whether the task is to identify and return an *imitated* sound, or to return the *most similar* sound. The 9126 similarity ratings are used as as a ground truth from which to measure the performance of each of the feature sets.

4.2. Linear mixed effect regression modelling

For a given feature set, distance is measured between each of the 420 imitations and the 6 within-class sounds, giving 2520 distance values. We use Euclidean distance in keeping with the baseline method PK08. Linear mixed effect regression (LMER) models are then fitted for predicting the continuous ratings from the continuous predictor variable. LMER is well suited to this task given that all listeners did not provide ratings for all imitations but only a randomly-selected set of 28 imitations (giving an unbalanced dataset). In addition, it allows us to include the dependencies between ratings for each listener and imitated sound.

Maximum likelihood parameters for the models are estimated using the lme4 package in R [26]. The general model is fitted with rating y_{ijk} as the dependent variable for each rating i, random intercepts for each listener k, and fixed effects of distance x_{ij} and imitated sound j, with an interaction term between distance and imitated sound. The model is given by:

$$y_{ijk} = \alpha_j + \beta_{1j} x_{ij} + \gamma_k + \epsilon_{ijk} \qquad (1)$$

where β_{1j} is the slope of the rating over distance for a given instance of j, and γ_k is the random intercept for a given listener k.

We note that model analysis showed heteroskedasticity in the residuals. Parameter estimates were therefore compared to those from robust models [27], and no major differences were found. As such the non-robust models were used for the analysis.

Wald 95% confidence intervals (CIs) were then calculated for each interaction ($\beta_{1j}x_{ij}$). For imitated sounds where the upper CI for $\beta_{1j}x_{ij} < 0$, we can infer the slope is significantly below 0 ($\alpha < 0.05$). This indicates that the feature set is a good predictor for the imitated sound in question.

The performance of each feature set is evaluated using two metrics: The percentage of imitated sounds for which $\beta_{1j}x_{ij}$ is significantly below 0 (accuracy); and Akaike's information criterion (AIC), which gives a measure of model fit (note: lower AIC = better model fit). An ideal feature set would have a significantly negative $\beta_{1j}x_{ij}$ (perfect predictor = -1.0) for all 30 imitated sounds, and be a good fit to the rating data given the model in Eq. 1.

5. RESULTS AND DISCUSSION

The results are given in Table 1. The encoded features from all CAEs outperform the baseline feature sets. The best performing feature set (11) shows slopes for $rating \sim distance$ that are significantly less than 0 ($\alpha < 0.05$) for 83.3% (25/30) of the imitated sounds, and has the lowest AIC. This shows the feature set is generally a good predictor of perceptual similarity between the vocal imitations and imitated sounds tested here, and has the best fitting LMER model.

Interestingly, preservation of the temporal resolution appears more important than spectral resolution for our task. For feature sets

[1] https://www.fxpansion.com

[2] http://www.eecs.qmul.ac.uk/facilities/view/control-room

Type	Feat. set	L1/8 kernel	Strides of conv./upsampling layers				Encoded layer (L4)		Results	
			L1/8	L2/7	L3/6	L4/5	Shape (×32)	Size	AIC	Acc.
CAE (Square)	1	(5, 5)	(2, 2)	(2, 2)	(2, 2)	(2, 2)	(8, 8)	2048	**1820**	**73.3**
	2	(5, 5)	(2, 2)	(2, 2)	(2, 2)	(4, 4)	(4, 4)	512	1925	66.7
	3	(5, 5)	(2, 2)	(2, 2)	(4, 4)	(4, 4)	(2, 2)	128	1958	66.7
CAE (Tall)	4	(5, 3)	(2, 2)	(2, 2)	(2, 2)	(2, 4)	(8, 4)	1024	**1609**	**73.3**
	5	(5, 3)	(2, 2)	(2, 2)	(2, 4)	(2, 4)	(8, 2)	512	1647	70.0
	6	(5, 3)	(2, 2)	(2, 4)	(2, 4)	(2, 4)	(8, 1)	256	2361	63.3
	7	(5, 3)	(2, 2)	(2, 4)	(2, 4)	(4, 4)	(4, 1)	128	2523	56.7
CAE (Wide)	8	(3, 5)	(2, 2)	(2, 2)	(2, 2)	(4, 2)	(4, 8)	1024	1921	66.7
	9	(3, 5)	(2, 2)	(2, 2)	(4, 2)	(4, 2)	(2, 8)	512	1866	73.3
	10	(3, 5)	(2, 2)	(4, 2)	(4, 2)	(4, 2)	(1, 8)	256	1395	83.3
	11	(3, 5)	(2, 2)	(4, 2)	(4, 2)	(4, 4)	(1, 4)	128	**1298**	**83.3**
PK08	12	–	–	–	–	–	– –	–	**2388**	**53.3**
TEMP	13	–	–	–	–	–	– –	–	2692	40.0
MFCC	14	–	–	–	–	–	– –	–	3162	30.0

Table 1: Details of the CAE variants and results for all feature sets (1–14). The CAEs vary only in the kernel shape of layers 1 and 8, and the shape of the encoded representations (determined by strides): with 3 square, 4 tall (in frequency) and 4 wide (in time). Results are given in terms of i) the LMER model fit (AIC), and ii) the percentage of imitated drum sounds for which the $rating \sim distance$ slope is significantly less than 0 ($\alpha < 0.05$). Note: lower AIC = better model fit.

wide in time and narrow in frequency (8–11) performance improves as the size of the encoded layer decreases. This indicates there is a great deal of redundancy in the spectral information: encoded spectral dimension shapes greater than 1 have an adverse effect on performance.

We only use similarity ratings of sounds from the same class (i.e. kick, snare etc.), and there will naturally be high spectral similarity within each class. As such, differences in temporal envelopes may be more salient than spectral variation, providing the decisive cues used by listeners when giving the ratings. This hypothesis is supported by comparing the square and tall CAEs. However there is also some redundancy here, as can be seen comparing feature sets 10 and 11. As a post-hoc analysis we tested variants of CAE 11 using smaller encoded kernel shapes: $(1, 2)$ and $(1, 1)$, and found that below $(1, 4)$ performance started to decrease. This effect can also be seen in models 4–7, where performance decreases as width is reduced from 4 to 1.

Regarding the baseline feature sets the 5 temporal features (TEMP) outperform the MFCCs. However, the PK08 feature set outperforms TEMP, indicating the small set of 5 temporal features may not be sufficient to capture the temporal variation required to predict similarity. Performance of the MFCCs is surprisingly low given their popularity for many music and speech related tasks.

Further analysis of the LMER model for the best performing feature set (11), shows the individual slopes for each drum sound (Fig. 2). Here there is considerable variation between the imitated sounds. In particular, we note that the 5 sounds for which the upper CI crosses 0 (3 kicks and 2 toms) are all pitched. They are not the only pitched sounds in the dataset (indeed, all the toms are pitched). This suggests that reducing the size of the encoded spectral shape 1 may work best over all the drum sounds used here, however the predictions for some pitched sounds suffer as a result.

Finally, we note the slopes, although generally below 0, do not approach -1. Listener rating data is inherently noisy, and the concordance amongst listeners varies across the sounds. As such, there will clearly be a glass ceiling for performance, and a perfect model fit would not be useful for a real world application of the LMER model. Indeed, a perfect model fit is not desirable if one is interested in generalisability of the fitted LMER model.

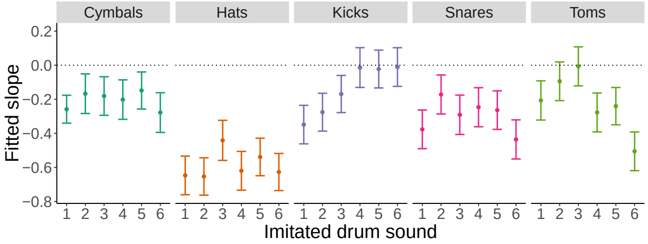

Fig. 2: Slope estimates (with 95% CIs) for the LMER model fitted on the performing feature set (11). A negative slope indicates a decrease in perceptual similarity with an increase in distance, i.e. sounds for which the feature set performs well.

6. CONCLUSIONS AND FUTURE WORK

In this paper we apply convolutional auto-encoders (CAEs) to the problem of query by vocalisation (QBV) for drum sound retrieval. We present a novel evaluation using perceptual similarity ratings between vocal imitations and the imitated drum sounds, providing insight into how learned features perform at predicting these ratings. Specifically, we compare CAEs that differ in both the size and shape of the encoded layer, in terms of the spectral and temporal dimensions.

Our experiments show CAEs outperform 3 baseline feature sets (MFCCs, temporal features and spectrogram based representations) by a considerable margin. Furthermore, we show that reducing the size of the encoded layer by reducing the spectral height increases the predictive power of the learned features, yet reducing the temporal width has the opposite effect. This finding is partly unexpected given that drum sounds generally have a similar overall temporal envelope (attack/transient followed by a decay), however understandable given that we compare only within-class sounds (i.e. kick, snare etc.), which are also likely to share similar spectral distributions. For future work we would like to investigate more fine-grained morphological features to represent the temporal evolution that appears to be so important here. In addition we would like to investigate the transferability of the best performing fitted LMER model to other

QBV tasks, to determine how a model fitted on one set of sounds and similarity ratings performs given a larger sound library, as might be used in a typical music production environment.

7. REFERENCES

[1] Kristina Andersen and Florian Grote, "Giantsteps: Semi-structured conversations with musicians," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Seoul, Korea, 2015, pp. 2295–2300.

[2] Guillaume Lemaitre, Arnaud Dessein, Patrick Susini, and Karine Aura, "Vocal imitations and the identification of sound events," *Ecological Psychology*, vol. 23, no. 4, pp. 267–307, 2011.

[3] Guillaume Lemaitre and Davide Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[4] Johan Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, Illinois, USA, 1989.

[5] Guillaume Lemaitre, Ali Jabbari, Olivier Houix, Nicolas Misdariis, and Patrick Susini, "Vocal imitations of basic auditory features," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2268–2268, 2016.

[6] Adib Mehrabi, Simon Dixon, and Mark B Sandler, "Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid," *The Journal of the Acoustical Society of America*, vol. 141, no. 2, pp. 783–796, 2017.

[7] David Sanchez Blancas and Jordi Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *Proceedings of the 53rd Audio Engineering Society Conference*, London, England, 2014, pp. 2–8.

[8] Gerard Roma and Xavier Serra, "Querying freesound with a microphone," in *Proceedings of the First Web Audio Conference*, Paris, France, 2015.

[9] Enrico Marchetto and Geoffroy Peeters, "A set of audio features for the morphological description of vocal imitations," in *Proc. of the 18th International Conference on Digital Audio Effects (DAFx)*, Trondheim, Norway, 2015.

[10] Yichi Zhang and Zhiyao Duan, "Retrieving sounds by vocal imitation recognition," in *Proceedings on the 25th IEEE International Workshop on Machine Learning for Signal Processing*, Boston, USA, 2015, pp. 1–6.

[11] Yichi Zhang and Zhiyao Duan, "Imisound: An unsupervised system for sound query by vocal imitation," in *ICASSP*, 2016.

[12] Yichi Zhang and Zhiyao Duan, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 533, 2016.

[13] Yichi Zhang and Zhiyao Duan, "Iminet: Convolutional semi-siamese networks for sound search by vocal imitation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2017, pp. 304–308.

[14] Elias Pampalk, Perfecto Herrera, and Masataka Goto, "Computational models of similarity for drum samples," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 408–423, 2008.

[15] Ernst Terhardt, "Calculating virtual pitch," *Hearing research*, vol. 1, no. 2, pp. 155–182, 1979.

[16] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.

[17] Mark Cartwright and Bryan Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Korea, 2015.

[18] François Chollet et al., "Keras," 2015.

[19] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[20] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] International Telecommunication Union, "ITU 1534-1: Method for the subjective assessment of intermediate quality level of coding systems," Tech. Rep., International Telecommunication Union, 2003.

[22] Thomas Sporer, Judith Liebetrau, and Sebastian Schneider, "Statistics of mushra revisited," in *Proceedings of the 127th Audio Engineering Society Convention*, New York, USA, 2009, pp. 323–331.

[23] Jacob Cohen, *Statistical power analysis for the behavioral sciences*, Erlbaum, NJ, USA, 2nd edition edition, 1988.

[24] Maurice G Kendall and B Babington Smith, "The problem of m rankings," *The Annals of Mathematical Statistics*, vol. 10, no. 3, pp. 275–287, 1939.

[25] Roy C Schmidt, "Managing delphi surveys using nonparametric statistical techniques," *Decision Sciences*, vol. 28, no. 3, pp. 763–774, 1997.

[26] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[27] Manuel Koller, "robustlmm: An r package for robust estimation of linear mixed-effects models," *Journal of statistical software*, vol. 75, no. 6, 2016.