LEARNING AUDITORY MODELS OF MACHINE VOICES

Kelly Dobson and Brian Whitman

MIT Media Lab
Massachusetts Institute of Technology
Cambridge MA 02139 USA
monster, bwhitman@media.mit.edu

Daniel P.W. Ellis

LabROSA, Electrical Engineering
Columbia University
New York NY 10025 USA
dpwe@ee.columbia.edu

ABSTRACT

Vocal imitation is often found useful in Machine Therapy sessions as it creates an emphatic relational bridge between human and machine. The feedback of the machine directly responding to the person's imitation can strengthen the trust of this connection. However, vocal imitation of machines often bear little resemblance to the target due to physiological limitations. In practice, we need a way to detect human vocalization of machine sounds that can generalize to new machines. In this study we learn the relationship between vocal imitation of machine sounds and the target sounds to create a predictive model of vocalization of otherwise humanly impossible sounds. After training on a small set of machines and their imitations, we predict the correct target of a new set of imitations with high accuracy. The model outperforms distance metrics between human and machine sounds on the same task and takes into account auditory perception and constraints in vocal expression.

1. INTRODUCTION

Machines and humans can not make the same sounds, and we need a way to detect what machine a human is vocally imitating. We are particularly driven by our work in Machine Therapy in which humans try to vocally imitate machines, but this task also informs problems such as query-by-vocalization for sound effects and music signals, and affect detection. For example, a birdsong database could be driven by a vocal query front end, but the problem of mapping from human imitation of animal sounds to the targets remains unsolved.

Our need is for a generalized model that computes symmetric similarity between machine and human sounds. Given a new machine, we would like to detect a human imitation without training. And given a human imitation of a machine, we would like to distinguish the specific machine they are imitating without training. In this paper we learn a model that links machine-generated sounds and human imitations and generalizes to new machines and new human vocal imitations.

Our process is described in Figure 1. We could directly compute similarity between the features (shown as the faint dotted line.) This approach has worked for us previously as in Blendie, a Machine Therapy apparatus that allowed people to control the motor speed of a kitchen blender by imitating its sound. Blendie used simple similarity through a statistical model based on distance to a ground truth blender machine sound. [1] We want to move from this approach to a more universal approach by designing a generalized auditory model inclusive of all machine sounds and their projections into human vocal space and vice versa. In this paper, we learn this model (the black box in Figure 1) by regressing the feature dimensions of the machine sound against the feature dimensions of the human imitation. Then, new machine sounds can be projected into 'human space,' with better accuracy in the similarity task.

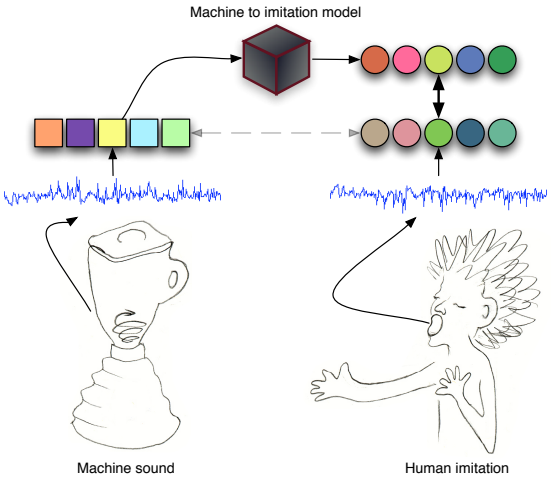

Figure 1: Our goal: a model that translates machine sounds to a predicted human imitation so that we can correctly predict the triggering of a machine and vary its parameters.

2. BACKGROUND

Machines share some expressive elements with people and harbor meaning and emotional material that is affective but often unrecognized by people. Machine Therapy, an alternate to traditional therapy akin to art therapy and music therapy, utilizes the sounds of machines as relational elements for people to vocally connect with. These unconventional vocal expressions can facilitate access to a human's own personal sounds and linked psychological states. In traditional psychoanalytic situations, non-verbal vocalizations by people are recognized as often very important and meaningful experiences.

Work in general sound analysis and categorization concentrates on human entered labels or similarity and clustering among databases of sounds [2]. Similarly, work in music instrument detection [3] attempts to detect an instrument sound in a mixed or

monophonic set of audio signals. Related work investigates speech vs. music detection [4] using a wide range of auditory-influenced features. In speech analysis, work has been done on laughter detection [5] and emotion detection using prosodic features [6].

Our work in acoustic machine analysis is somewhat informed by work in machine condition monitoring, where signal processing systems attempt to detect possible faults in machine processes. In [7], higher order spectra is used to detect degradation of machine health. In [8] a genetic algorithm is used to select vibration features to inform a neural network trained to detect faults in rotating machinery.

3. ANALYSIS AND PREDICTION OF MACHINE VOICES

Our problem is to estimate the quality of a particular person's imitation of a machine, taking into account the innate limitations of the sound 'gamut' that that person, or people in general, can produce, and also the specific acoustic dimensions or attributes that the person is aiming to reproduce. Ideally, we would like to find the appropriate representational space that captures all the significant aspects of variation in both machine voice and human imitation, and to learn the optimal mappings between human and machine sounds – mappings which will likely vary between subjects, but which will also share a common core.

3.1. Auditory Features

Machines have a wide variety of possible sounds, and as our task is to create a model that generalizes well to new machines, we need to create a representation that captures perceptually dominant characteristics of the sound without relying on the specifics of any particular machine. We observed that the qualities of pitch, roughness, energy, and transients were often imitated by the subjects. In this study we first tried to work with modulation cepstra [9], which largely suppresses pitch information, and our predictions did not fare well.

Instead we chose to focus on five auditory features for each short-time window (w samples, usually $w = 2048$ for a sampling rate of 44,100 Hz) of the input audio. We chose a fundamental frequency (f_0) estimation [10] [11], aperiodicity (the residual from the f_0 estimation) [12], a power estimation, spectral centroid and 'modulation spectral centroid,' the centroid of the Fourier transform of the magnitude envelope along time of each frequency channel in our initial short-time Fourier transform (STFT), $\mathcal{F}_n(\|X(k,n)\|)$, where $X(k,n)$ is the STFT for frequency band k and time window n, and $\mathcal{F}_n$ indicates the Fourier transform taken along the n (time) dimension. Our intuition for detecting modulation in the spectral bands is that often the machines have high roughness content in the upper modulation bands, while human imitations can only to try to approximate these sounds in lower bands.

An example set of features extracted from a target sound (a blender) and a vocal imitation is seen in Figure 2. None of our features encapsulates long-term time scale above our window length w; we consider an analysis of this time scale and transients to be part of future work. For the scope of our features, we treat each frame independently of the next, and there is no explicit correspondence between the time pattern of the machine sound and imitation. Other features to be considered in future work include voiced/unvoiced detection, noise detection, and bandwidth.

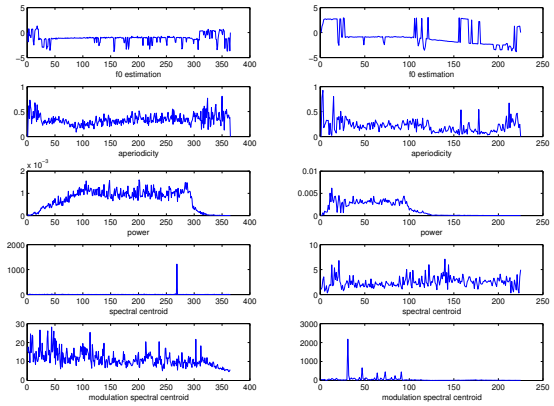

Figure 2: Features extracted from a blender sound at left, with the human imitation at right. Five perceptual features: f_0 estimation, residual (aperiodicity), power estimation, spectral centroid and modulation spectral centroid. Time segemnts on the horizontal axis, feature output on the vertical axis.

In the experiments section we perform a feature selection task that attempts to discover the best performing individual features from the set of five. Different machine types respond better to different types of auditory modeling, and to make a general model we choose the best performing feature combination.

Before learning or evaluation, all the features (in the training subsets) had their mean removed. All features were scaled to $\{-1..1\}$ before the regression model was learned. The mean and scaling constants were saved as part of the model to perform the same transform on test data.

3.2. Projection Learning Techniques

As we can see in Figure 2, the imitation seems to have some correlation to the target sound, but with different scale and dynamics. We need to *learn* this mapping of machine sound to human imitation so that for new machine sounds we can compute the similarity in a projected 'human space.'

To learn the projection between machine sound and human imitation, we used a multi-dimensional regression learning technique with support vector machines (SVM) [13]. As the usual configuration of a regression problem is to have a set of ℓ d-dimensional frames linked to a set of single-dimension scalar ys, we instead use a multi-class approach. For each input frame of machine sound, we chose a random frame from its human imitation space, and after collecting the entire set of ℓ frames of the machine sound, we train d SVM regression models, one for each dimension of the human response. Thus, each model predicts a single dimension of the human imitation space by a regression among all d dimensions of the machine sound. We note that this approach considers covariance among variables in machine space but not of the human features. In our model, covariance of machine features can influence a single human feature projection, but features that co-vary in machine space do not directly influence covariance in the human projection. This is a limitation to be addressed in future work.

The SVM regression task expects two parameters: a σ for

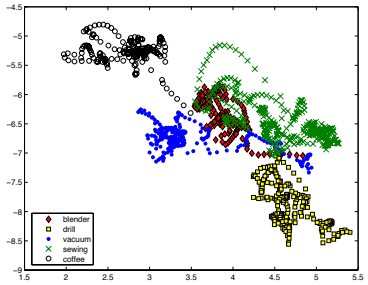

Figure 3: Five machine sounds projected into a two dimensional space.

the kernel parameter, which represents data embedded in a kernel space described through the gaussian

$$e^{-\frac{(|x_1-x_2|)^2}{\sigma^2}} \quad (1)$$

where x_1 and x_2 are frames in ℓ, and the regularization parameter C which is often considered a 'generalization knob,' affecting the tradeoff between accuracy in classification and generalization to new data.

In our work we trained a single set of regression models (one for each dimension in feature space) instead of separate models for each machine. If we trained one machine's model at a time, the regression problem could overfit to only predict one type of machine. By forcing the regression to consider all types of machines and their corresponding imitations, we hope to create a model that works well for new types of machines.

4. EXPERIMENTS AND RESULTS

Sounds of five machines and seven subjects were recorded for this study. The machines included an *Osterizer Blendor* kitchen blender, a *Dewalt* portable power drill, a *Hoover Commercial Windtunnel* stand up vacuum cleaner, a *Acorto 2000s* espresso maker and a *Singer 20* industrial sewing machine. The subjects included three females and four males with different backgrounds and various primary languages. A PZM microphone was placed between the person imitating a machine and the machine they were working with. The microphone was always within 1 m of the sound source being recorded.

A 'machine ground truth' was first recorded for each of the five machines, consisting of the machine sound alone. The human participants were introduced to the machines and then left alone with the machines in a soundproof studio with the instructions to record their imitation samples when comfortable. Each participant spent a short time alone with all of the machines, three to ten minutes, before recording their imitation of each machine onto the minidisc setup provided. Each of the imitations were between 2 s and 10 s in duration.

4.1. Similarity using Only Ground Truth

We first compute Euclidean distance between randomly chosen frames of each of the machine sounds in order to set an upper limit on classification. By comparing randomly-selected disjoint subsets of each machine, we obtain the natural confusion of the target sounds. In Figure 3 we see the five machines' audio projected through our features into two dimensions via principal components analysis [14]. While the five machines are clearly separable there is some overlap between similar sounding types of machines.

	blender	drill	vacuum	sewing	coffee
blender	**0.78**	0.07	0.05	0.09	0.01
drill	0.03	**0.93**	0.01	0.01	0.02
vacuum	0.02	0.02	**0.81**	0.14	0.02
sewing	0.06	0.03	0.13	**0.76**	0.02
coffee	0.03	0.06	0.18	0.07	**0.66**

Table 1: Confusion of machines' ground truth. For each row of ground truth, the columns indicate our prediction results for each machine. This sets an upper bound on prediction performance. Mean accuracy of classifiers = 0.77.

The results for the ground truth task in Table 1 is obviously high with a mean accuracy of 0.77 (the mean of the diagonal of the confusion matrix.) We note that the coffee machine is the least self-similar while the drill is the most self-similar.

4.2. Performance Without Model

	blender	drill	vacuum	sewing	coffee
blender	0.08	0.06	0.04	0.37	**0.44**
drill	0.02	0.09	0.14	0.07	**0.69**
vacuum	0.02	0.08	0.08	0.14	**0.68**
sewing	0.08	0.12	0.03	0.38	**0.38**
coffee	0.06	0.06	0.17	0.26	**0.46**

Table 2: Confusion of prediction of human imitations (rows) against machine ground truth (columns) without using a learned auditory model. Without the model, all imitations are classified as closest to the "coffee" ground truth machine sound, for a total prediction accuracy of 20%. Mean accuracy of classifers = 0.22.

We then show the result without the auditory model. In this task, we show the accuracy of machine prediction given the entire set of human imitations. To compute the prediction, we arrange the Euclidean distance between the set of imitations $\times$ the set of machine sounds, after both are run through our feature space calculations. For each frame of human imitation, we find the frame in machine space with the minimum distance and treat it as a vote for that class. The columns in Table 2 indicate probabilities of each machine prediction given the row imitation sound.

We see that all imitations are classified as the coffee machine, for a total accuracy of 20% for a 1-in-5 machine detection task. The overall mean accuracy of the classifiers is 0.22, very close to the random baseline of 0.20. There is a close call for the sewing machine classification, but otherwise most of the frames were incorrectly labeled. We attribute this poor performance to the lack of a perceptual model of human imitation– since the coffee machine was shown to be the least self-similar in Table 1, it follows that it had the widest variety in feature space. Therefore, without a model, much of the imitation sounds have a higher probability of matching with the coffee machine.

4.3. Performance With Model

We then learn the model as described in Section 3.2. To be able to evaluate different machines' performance through the model, we computed a round-robin evaluation, leaving one machine out each time for a total of five models. After the d regression models were learned for each of the five machines (using a C of 1000 and a σ of 0.5,) we computed our features on the machine audio and put them through its left-out model (i.e. the model trained on the data excluding both that particular machine's sound and its imitations) to compute a projection in human imitation space for the machine sound. We then computed the similarity classification as above, but instead of computing similarity of human imitation to machine sound, we computed the similarity between human imitation and machine sound projected into human imitation space.

	blender	drill	vacuum	sewing	coffee
blender	**0.44**	0.11	0.15	0.25	0.05
drill	0.27	0.03	**0.62**	0.07	0.01
vacuum	0.22	0.11	**0.46**	0.13	0.09
sewing	0.24	0.09	**0.36**	0.21	0.10
coffee	0.18	0.13	0.14	0.17	**0.37**

Table 3: Confusion of prediction of human imitations (rows) against machine ground truth (columns) projected through our learned auditory model with the highest probability for each imitation in bold. This machine prediction task scored 60% overall. Mean accuracy of classifiers = 0.30.

The results for this task are in Table 3. We see that our overall accuracy in the 1-in-5 prediction task is now at 60% over the 20% we achieved without using the model. We also see that our mean accuracy is now 0.30, compared to 0.22 for no model and 0.2 for the baseline. The missed machines include the drill, which had the highest self similarity in Table 1, and the sewing machine. We explain the poor performance of the drill due to poor generalization in our model: since the drill has high self-similarity and low similarity to any of the other machines, our model (trained on only the other machines in the round robin) did not account for its unique sound.

4.4. Evaluating Different Features

Due to the expressive range of each of the machines, we attempted to determine which of the auditory features were more valuable for each machines' individual classification task. Just as we computed a leave-one-out evaluation along the machine axis for evaluation in prediction, we here evaluate feature performance by formulating the vector of the $(2^d) - 1$ permutations. For each feature permutation, we compute the similarity evaluation as above and search the result space for the best performing overall model and also the best performing classifer for each individual machine.

machine	best features	performance
blender	aperiodicity	0.79
drill	spectral centroid, modulation centroid	0.24
vacuum	power	0.70
sewing	f_0	0.40
coffee	modulation centroid	0.68

The best overall feature space for the 1-in-5 task through model was found to be a combination of all features but the modulation spectral centroid. For the task without the model, the best performing features for the similarity classification were a combination of f_0, power, and modulation centroid.

5. CONCLUSIONS AND FUTURE WORK

We show in this paper that it is possible to project a machine sound to a human vocal space applicable for classification. Our results are illuminating but we note there is a large amount of future work to fully understand the problem and increase our accuracy. We hope to integrate long-scale time-aware features as well as a time-aware learning scheme such as hidden Markov models or time kernels for SVMs [15]. We also want to perform studies with more machines and more subjects, as well as learn a parameter mapping to automatically control the functions of the machines (speed, torque, etc.) along with the detection.

6. REFERENCES

[1] K. Dobson, "Blendie." in *Conference on Designing Interactive Systems*, 2004, p. 309.

[2] M. Slaney, "Semantic-audio retrieval," in *Proc. 2002 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002.

[3] K. D. Martin, "Sound-source recognition: A theory and computational model," Ph.D. dissertation, MIT Media Lab, 1999.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 1331–1334.

[5] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *Proc. NIST Meeting Recognition Workshop*, March 2004.

[6] T. Polzin and A. Waibel, "Detecting emotions in speech," 1998.

[7] N. Arthur and J. Penman, "Induction machine condition monitoring with higher order spectra," *IEEE Transactions on Industrial Electronics*, vol. 47, no. 5, October 2000.

[8] L. Jack and A. Nandi, "Genetic algorithms for feature selection in machine condition monitoring with vibration signals," *IEE Proc-Vis Image Signal Processing*, vol. 147, no. 3, June 2000.

[9] B. Whitman and D. Ellis, "Automatic record reviews," in *Proceedings of the 2004 International Symposium on Music Information Retrieval*, 2004.

[10] A. de Cheveigé, "Cancellation model of pitch perception," *J. Acous. Soc. Am.*, no. 103, pp. 1261–1271, 1998.

[11] M. Goto, "A predominant-f0 estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models," in *Proc. ICASSP-2001*, 2001.

[12] P. R. Cook, "Music, cognition and computerized sound," pp. 195–208, 1999.

[13] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[14] C. V. L. G.H. Golub, *Matrix Computations*. Johns Hopkins University Press, 1993.

[15] S. Rüping and K. Morik, "Support vector machines and learning about time."