FEELING TO LISTEN: SEPARATING SIMULTANEOUS IMPACT SOUNDS USING MULTI-MODAL LEARNING

A PREPRINT

Jui-Hsien Wang
Stanford University
jw969@stanford.edu

March 19, 2019

ABSTRACT

We propose a new deep learning model to separate simultaneous impact sounds. Impact sounds like tapping on a wine glass in real life are almost always associated to at least two vibrating objects (in this case your fingernails and the glass), and thus the sound we hear are the result of two vibrating objects combined. Inspired by the recent advances in speech separation, we train a deep neural network to output complex ratio masks in order to separate out the two impact sounds. The input to the network is a (mixed) sound waveform and vibrometer data on one object. We train the network using a novel synthetic dataset that is constructed ground-up using a special physics-based acoustic simulator. The mean square error of the predicted masks reach up to 75% accuracy using the current approach.

- Title: Feeling to Listen: Separating Simultaneous Impact Sounds using Multi-modal Learning
- Category: Others (Learned physical understanding)
- Members: Jui-Hsien Wang (jw969)

1 Introduction

Impact sounds are caused by vibrations induced when two objects come into contacts. It is an important, and sometimes the only, cue that can be used to reason about the world we are in. For example, gently tapping on the table with our fingernails gives us a sense of the material composition, be that of wood, plastic, or metal. In a dark room, that might be the only way to guess the material of the table. Unfortunately, by definition, the impact sounds are always a superposition of (at least) two different vibrating objects and thus two different sounds. In the above example, one from the table and one from the fingernail. If instead of a fingernail, we use a plastic pen, then it might very well affect our ability to reason about the table's material.

In this project, we show that a deep neural network can learn learn a general representation to separate the two coupled impact sounds reasonably well. Note that the problem is very challenging since the two objects can have overlapping frequencies in which they vibrate at, acoustic transfer can affect how loud each components sound (e.g., one object might produce faint sounds on some configurations but loud sounds on others), and depending on the impact position only a part of the vibrational modes will be active. Moreover, there is a non-uniqueness problem caused by symmetry – if only the input mixed sound is observed, there are at least two solutions to the problem. Therefore, in order to help disambiguate the problem and break the symmetry, we place an accelerometer on one of the objects that detect surface acceleration. This is reasonable as every phone now has an accelerometer that can measure the vibrations. This is also inline with the general trend of multi-modal learning, where several *views* of the data can result in effective, self-supervised training.

In summary, the problem we are interested in is: *Given a mixed sound of a two-object impact event, and surface vibrational data on one of the objects, we want to separate the mixed sound into two sounds, one for each impacted object.* An example illustration of the problem is given in Figure 1.

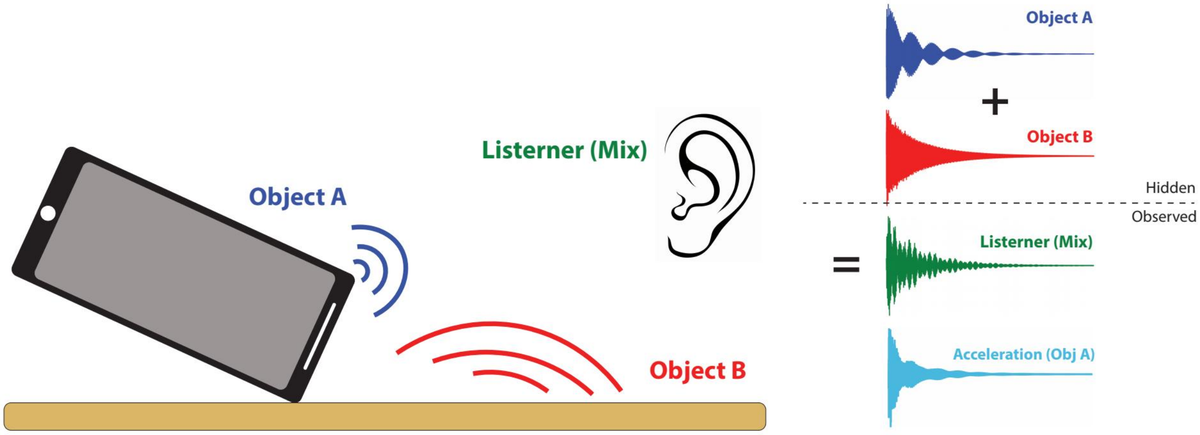

Figure 1: From the observed mixed sound and surface acceleration on one object, we want to infer the hidden states of the separated sounds for each object. The sound is mixed due to acoustic radiation and can contain, e.g., closely packed frequencies.

The network architecture is introduced in §4; the network is trained on a novel, synthetic dataset we constructed ground-up, explained in §3; finally, the results and analysis are shown in §5.

2 Related work

Information in the real-world often comes in several different modalities. For example, videos often contain frames of images and synchronized audio, captioned images contain text describing the image, and modern robots and autonomous vehicles have integrated sensors to obtain vision, LIDAR, sound, pressure, vibration, and other sensor information. These are examples of the many possible *views* of the same data of some events. Multi-modal learning is a learning paradigm that aims at using multiple modalities of the data in order to provide self-supervision learning or to increase the performance of the learned model.

Sound is being actively explored as a viable modalityin conjuction with vision in order to aid material and shape classification [1, 2, 3], to provide self-supervision for image classification [4], and to localize visual events assoicated with sound sources [5, 6]. Large audio-only datasets (e.g., DEMAND [7]) or audio-visual datasets (e.g., AudioSet [8]) exist, with either synthetic sounds, or recordings, or a mix of both [1]. However, for our task we need clean, isolated impact sounds for *each* impacting object in order to create meaningful labels, and there is currently no dataset that contains this.

Influenced by a recent success in multi-modal learning for speech separation by Ephrat et al. [9], we first construct a large dataset of impact sounds by superimposing isolated, synthetic, single-object impact sounds, and then use a deep neural network to learn a general representation that can separate the impact sounds.

3 Dataset and Features

We introduce a new, large-scale synthetic dataset comprising of clean impact sound clips with no interfering background noises. In our dataset, there are 33 hours of 1-second long clips, and in each clip there are two impacting objects randomly chosen out of 4 precomputed object models (wooden bowl, ceramic plate, wine glass, and plastic bunny); for each clip there is also a time-series surface acceleration data associated with it for one of the objects. Representative examples of the data are shown in Figure 2.

3.1 Dataset creation pipeline

Inspired by the dataset created for speech separation [9], we superimpose two synthesized, single-object impact sounds into a single mixed sound using an accurate physical simulation engine. This circumvents the tedious and difficult process of isolating impact sounds manually in real-world recordings and allows us to test the ideas quickly.

The single-object sound is synthesized using accurate finite-element analysis [10] of the objects followed by a technique called precomputed acoustic transfer [11] in order to obtain correct amplitude scalings due to acoustic wave propagation.

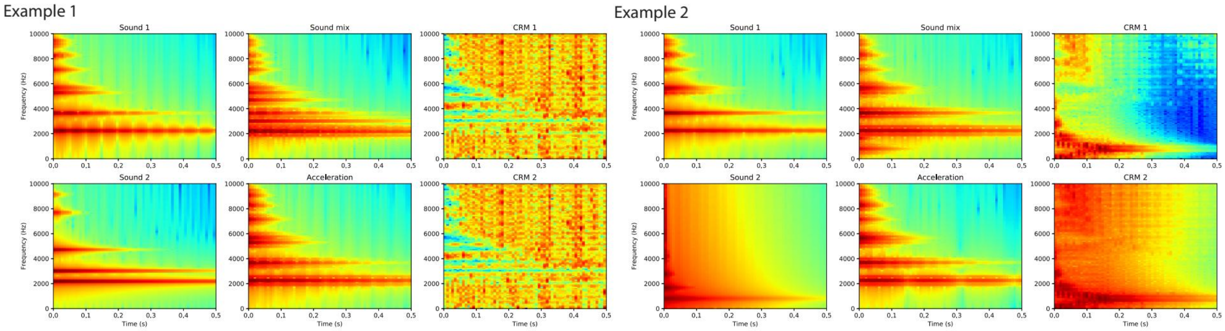

Figure 2: Some examples in our dataset. Left column: ground-truth separated sounds; Middle column: input to the neural network, the mixed sound and the surface acceleration data; Right column: complex ratio masks that, if applied (complex multiplication) to the mix sound can recover the left column. This is also the ground-truth data used to calculate the loss.

Precomputed acoustic transfer works by precomputing the solution of the Helmholtz equation to take the intricate near-field wave effects into account. This technique is widely used in computer graphics in generating realistic sound clips for near-rigid objects like the ones we have, and is used in previous work for generating synthetic sounds for training neural networks [12, 13]. Pairs of the single-object sounds are then combined to form the mixed sound.

3.2 Input-output pairs and complex ratio masks

Instead of directly inputing the time-series data into the network, we first convert both the audio and vibrational data into spectrogram via Short-Time Fourier Transform (STFT). We use a 512 hanning window with a hop size of 160. Since the sampling rate is 48 kHz, this results in the frequency resolution of $48000/512 = 93.75$ Hz for the FFT bins.

Instead of directly outputing the time-series waveform data, we predict two complex ratio masks (CRM) [9], which are complex-multiplicative masks that can be applied directly onto the input spectrogram to predict the output spectrograms. The final waveforms are computed by taking the inverse Short-Time Fourier Transform (ISTFT) on the predicted spectrograms. This process is illustrated in Figure 3.

4 Neural Network Model

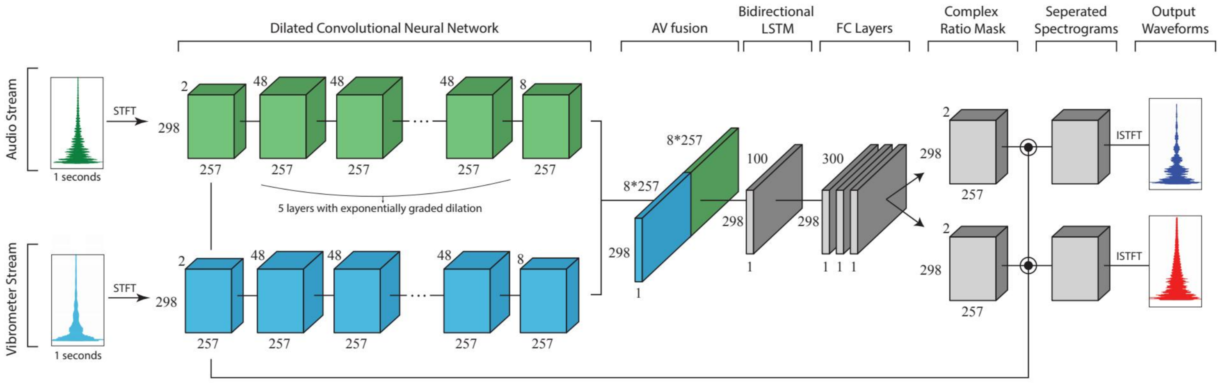

Figure 3: The neural network architecture used in the work consists of two subnetworks, sound and vibro networks, followed by a fusion network.

The neural network architecture used in this work is summarized in Figure 3. It consists of two modular networks, one for the mix sound and the other for the vibration data. The input STFT features pass through the designated subnetworks with 6 layers of convolutional neural networks (CNN). The result is then concatenated in the fusion layer before being passed into the bidirectional Long short-term memory (LSTM) and finally the fully-connected layers (FC) to produce the predicted CRMs. The dilation rate for the CNN layers are designed to increase exponentially in order to get a large receptive field. Our network, especially the CNN layers, are shallower compared to the one in [9] despite other key differences; this is because we have less data compared with the similar previous work. There are in total 4.2 million trainable parameters.

4.1 Implementation Details

The model is implemented in Keras. The loss function is the mean squared error (MSE) on the CRMs. Batch normalization is used for all layers and adam optimizer is used. The learning rate is 0.001 for the first 10 epochs, 0.0002 for the next 10 epochs, and 0.0001 for the rest 30 epochs (we trained for 50 epochs in total). The details of the dilated CNN layers are shown in Table 1. Training loss converges well and is shown in Figure 4. Training is done on Google Cloud Compute engine using 4 Tesla K80s, and was run for 8.5 days. We have experimented with larger and deeper neural networks, and with two different learning rate schedulers, all producing similar results. The validation loss also falls roughly in the same range as training loss, indicating we did not overfit the dataset. The dataset split is 90/5/5.

Table 1: Dilated convolutional layers comprising our sound and vibro subnetworks.

	conv1	conv2	conv3	conv4	conv5	conv6
Number Filters	48	48	48	48	48	8
Filter Size	1×7	7×1	5×5	5×5	5×5	1×1
Dilation	1×1	1×1	1×1	2×1	4×1	1×1

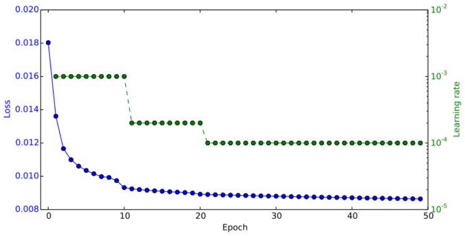

Figure 4: Convergence of the training loss.

5 Results and Future Work

Using a relatively small dataset, the results are rather satisfying especially given how challenging the problem is. Some of the representative results are shown in Figure 5. We now summarize some of the findings and provide an analysis:

- On the test set, the average loss is 0.009 (Pascal), which is similar to the training loss shown in Figure 4.
- Adding the acceleration data indeed breaks the symmetry and we have not seen any example in the test set where the solution is incorrectly swapped between object A and object B.
- For certain cases, there are some errors (e.g., the predicted B spectrogram in Figure 5 has an incorrect frequency component at about 500 Hz). We think that this is because the FFT bins have only frequency resolution of 93 Hz in our setup, so certain frequencies are not properly masked.
- We pleasantly found that predictions on object B are more accurate for many cases. This is important for many applications. For example, in robotics, a robot might be actively probing an object to understand the material composition. In that case we need accurate prediction for the sound made by object B only, since object A is the robot which we has prior knowledge about.
- We explored different architectures and hyperparameters; however, the results do not change much. We think that it would be very interesting to see how does the model do with a more diversified dataset (say 10,000 different objects). It will also be interesting to see if the results can be generalized to unseen objects or real-world scenarios where the frequency content is unknown to the network.
- We suspect a better normalization strategy for the acceleration data will result in better results, since the physical units for the acceleration data and sound are different, they could result in orders of magnitudes difference. Other forms of loss functions should also be investigated.

- Lastly, the model is very slow to train, taking weeks of compute on multiple high-end GPUs. It will be interesting to see if we can use a more efficient architecture without deteriorating the results.

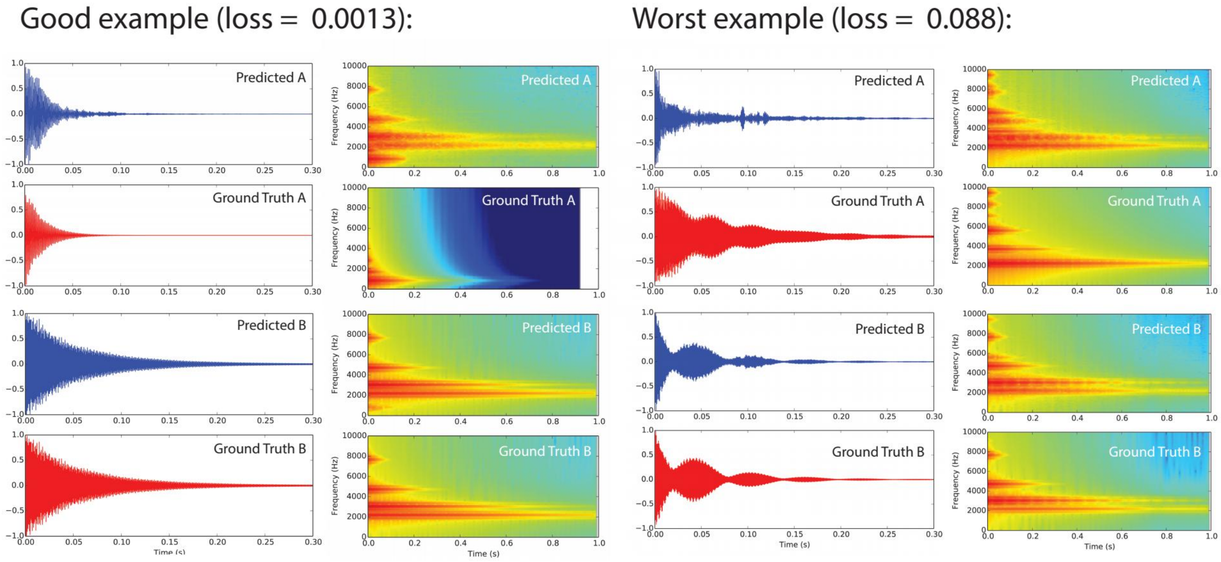

Figure 5: Representative results in the test set. We show a good example where loss is well below the average test loss and the worst example with the highest loss. Both the temporal waveform envelope and the general spectral shape can be recovered using our network. Note that we generally care more about accuracies in the separated object B sound, which comes from the object we are "investigating" (with, for example, our cell phone). There are certain errors in predicted A but we got the general timbre right. Odd columns: waveforms for both sounds; Even columns: spectrograms (all colorbars are identical).

References

[1] Auston Sterling, Justin Wilson, Sam Lowe, and Ming C. Lin. Isnn: Impact sound neural network for audio-visual object classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[2] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. Generative modeling of audible shapes for object perception. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman. Shape and material from sound. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1278–1288, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. *CoRR*, abs/1705.08168, 2017.

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. *CoRR*, abs/1712.06651, 2017.

[6] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95 vol. 1, June 2005.

[7] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. *Proceedings of Meetings on Acoustics*, 19(1):035081, 2013.

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[9] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, abs/1804.03619, 2018.

[10] Ahmed A Shabana. *Theory of Vibration: An Introduction*. Springer Science & Business Media, 2012.

[11] Doug L. James, Jernej Barbic, and Dinesh K. Pai. Precomputed Acoustic Transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics*, 25(3):987–995, July 2006.

[12] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman. Shape and material from sound. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1278–1288, 2017.

[13] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. Generative modeling of audible shapes for object perception. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.