2019 27th European Signal Processing Conference (EUSIPCO)

Investigation of Reverse Mode Loudspeaker Performance in Urban Sound Classification

Gyorgy Kalmar
University of Szeged
Szeged, Hungary
kalmargy@inf.u-szeged.hu

***Abstract*—The loudspeaker is a transducer that converts electrical signals to sound. However, it is well-known that in reverse mode, it can convert sound to an electrical signal. In this paper, the reverse mode behavior is investigated through the analysis of its influence on urban sound classification accuracy by comparing the results of deep learning based classifiers. As no audio datasets recorded by loudspeakers are available, a popular traditional dataset was used and transformed into forms as they would have been recorded by reverse mode speakers. These transformations simulated the loudspeakers' electrical responses to acoustical excitation signals based on their reverse mode transfer functions, which were derived from equivalent mechanical circuits. The details of this reverse mode modeling are also included. The transformed datasets were used during the trainings of the classifiers, and the effects of different speaker parameters and noise levels were examined and compared. The results showed that smaller, full-range speakers performed better than bigger woofers. The types of well-classified events revealed that loud, impulsive events could be classified more accurately.**

***Index Terms*—loudspeaker, reverse mode, sound classification**

I. Introduction

Loudspeakers are transducers that convert electrical signals to sounds. However, in *reverse mode*, they can convert acoustical signals to electrical signals as microphones do. This property is well-known and one could easily find descriptions about how to use speakers as microphones [1], [2]. The current paper investigates the reverse mode behaviour of speakers by analyzing their performance in an audio classification task.

In academic literature, the only recent related work was published by Guri et al. in [3]. They developed a proof-of-concept malware, named *Speake(a)r* that allowed using headphones as microphones by port retasking, which could be realized silently on any computer containing Realtek audio CODEC chips. The software altered the functionality of output ports and turned them into input ports. Thus, a headphone connected to the output line became an input device. They also carried out experiments and recorded normal conversations from close ranges with acceptable quality, therefore showed that eavesdropping and cyber attacks would be feasible.

Based on their results, the question arises: Would it be possible to use loudspeakers (not headphones) for "good"? For example, in security applications, where suspicious acoustical event detection is important while protecting privacy. In this paper, moving-coil loudspeakers are investigated as input devices for automated urban sound classification. Only passive speakers - without built-in amplifiers - are examined, as the voltage generated in reverse mode cannot reach the driving cable in active speakers.

The loudspeaker based audio recording is sub-optimal compared to the microphone based solutions. However, if a deployed system already contains multiple, spatially separated but connected passive speakers that are driven by the same source, the whole area could be protected by monitoring the driving cable with only one device. This setup can be found in schools, stations, hospitals, etc., where the speakers broadcast announcements. The other advantage comes from the structure of the loudspeakers. As they are designed to radiate, not to record sound, their sensitivity is low in reverse mode. The signal-to-noise ratio is therefore reduced, disabling their usability in potential spying attacks.

Contributions: The introduction of the reverse mode modeling of loudspeakers, which has not been presented before; and the utilization of this model to simulate the speakers' responses to various acoustical inputs. The effect of the reverse mode on the audio classification accuracy is examined by using a labeled urban sound dataset and convolutional neural network based classifiers.

II. Speaker models

A moving-coil speaker contains a suspended coil, which is placed in a gap between permanent magnets. When alternating current flows through the wire, force is being induced that moves the coil and the attached diaphragm, the cone, back and forth. That rapid movement of the cone generates pressure waves in the air.

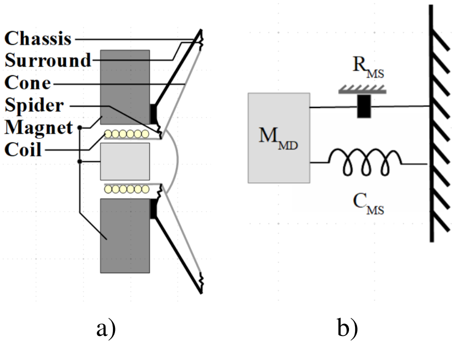

Fig. 1. Cross-section of a moving-coil loudspeaker (a), and its equivalent mechanical mass-sprint-damper model (b).

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

The cross-section of a loudspeaker can be observed in Fig. 1.a. The explained structure can be approximated by a mechanical mass-spring-damper system presented in Fig. 1.b, where M_{MD} is the mass of the moving parts, C_{MS} is the compliance of the suspension and R_{MS} is responsible for the additional losses.

A. Equivalent Circuit of the Moving-coil Loudspeaker

In this section, the loudspeaker's electrical equivalent circuit and the steps required to obtain it are summarized. These results are well-known in the literature [4]–[7], however, the outline is presented here as it will help the reader to understand the concept of reverse mode modeling.

The loudspeaker is driven in the electrical domain that induces mechanical force, which moves the cone, producing waves in the acoustical domain. This complex system can be modeled with an equivalent circuit presented in Fig. 2.a by using electrical impedance, mechanical *mobility* and acoustical impedance [4], [7].

In Fig. 2.a, a voltage generator with neglected output impedance models the driving source of the loudspeaker. The coil resistance, R_e is represented by a resistor while its inductance is negligible at the relevant sound frequencies.

It is known that the force on the coil is given by the product of the flux density in the gap, B (T), the length of the coil wire l (m) and the current (A). Therefore the constant that connects the electrical and mechanical domain is the Bl product. This parameter is usually given in the loudspeakers' datasheets. The force generation and impedance transformation can be modeled by a virtual transformer with turn ratio of *Bl*:1. The impedance connected to the secondary of this transformer is reflected back to the primary side by this Bl ratio.

In the mechanical domain, the mechanical mobility (the inverse of the mechanical impedance) is used to model the mass M_{MD}, the compliance C_{MS}, and the losses G_{MS} of the mass-spring-damper system.

The air load has two main parts, the reactive (C_A) and the resistive (R_A) parts. The radiated sound energy (dissipated on R_A) is proportional to the square of the diaphragm area S. This load modeled by applying the mechanical velocity to the primary of another virtual transformer with turn ratio of S, as shown in Fig. 2.a. On the secondary side, the voltage is proportional to the volume velocity U and the current is proportional to the sound pressure P.

The first transformer can be eliminated by bringing the mechanical load to the primary side with impedance inversion and conversion. The second transformer can be eliminated in the same way, but it does not invert the impedance. After these eliminations, the equivalent circuit will be transformed into the electrical domain, where all the components are replaced by equivalent electrical ones. This simplified model can be observed in Fig. 2.b. The detailed description can be found in [4], [7].

The parameter values of the model can be calculated from the well-known Thiele-Small parameters [8]. These electromechanical parameters define the low-frequency behavior of a speaker unit and usually are used for enclosure design. They are measured and published by the speaker manufacturers. The complete list of the Thiele-Small parameters and the relations with the values required by the model in Fig. 2.b can be found in [8], [9].

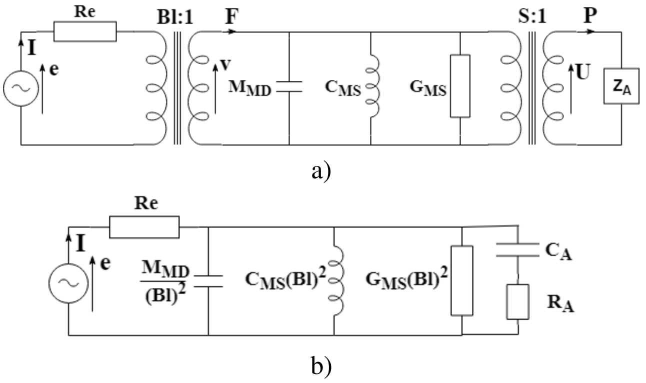

Fig. 2. Equivalent circuits of a moving-coil speaker [7]. In (a), virtual transformers are used to represent the mechanical force generation and acoustical energy radiation. In (b), these transformers are eliminated by impedance conversion, and the loads are transformed into the electrical domain.

B. Equivalent Circuit of the Reverse Mode

When an external pressure signal is applied on the surface of the diaphragm, it starts vibrating, and the attached coil oscillates in the magnetic field. According to Faraday's law of inductance, voltage is generated in the coil. This can be modeled in the same way as it was presented in Fig. 2.a, but the driving source is moved into the acoustical domain. As the measurement of the reverse mode voltage is required, the electrical voltage generator is replaced by a resistor, which represents the input impedance R_O of a voltage meter or operational amplifier. The resulting circuit can be examined in Fig. 3.a. The acoustical domain resistance and reactance are eliminated as the driving pressure directly acts on the surface of the diaphragm. The mechanical domain model remains the same, but the excitation comes from the other direction.

To simulate a speaker's response to a given acoustical signal, the model is converted into the mechanical domain by eliminating the two virtual transformers. The resistance of the coil, R_e is negligible compared to the input impedance of the voltage meter ($R_e << R_O$), therefore it is omitted.

The acoustical driving source is a volume velocity signal. The relation between this signal and the mechanical velocity of the diaphragm is determined by the surface area of the cone, the transmission coefficient of the cone material and the shape of the cone. The exact description is complex, but an approximation can be made, as the U volume velocity can be brought to the other side of the virtual transformer by using the turn ratio. This step is simplified here as the original intensities of the acoustical input signals are unknown, thus the exact simulation is impossible.

In Fig. 3.a, mechanical mobility was used to represent the mechanical components. This part of the circuit can be replaced by its dual mechanical *impedance* type analogy. The

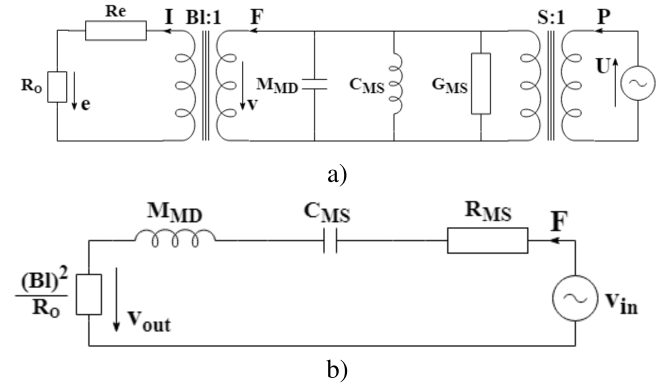

Fig. 3. Equivalent circuits of the reverse mode. In (a), virtual transformers are used to represent the force and electrical signal generation steps. The driving source is moved to the acoustical domain. The electrical output signal can be measured on R_O. In (b), the simplified mechanical equivalent circuit is presented after the elimination of the virtual transformers. The elements are transformed into the mechanical domain by using mechanical impedance.

dual components then form an equivalent single loop circuit with elements connected in series. This circuit is presented in Fig. 3.b. The output signal is the velocity drop across the impedance converted resistor R_O.

From the simplified equivalent circuit, the reverse mode transfer function of a loudspeaker can be derived as:

$$H(s) = \frac{R_O C_{MS} \cdot \boldsymbol{s}}{M_{MD} C_{MS} \cdot \boldsymbol{s}^2 + (R_O + R_{MS}) C_{MS} \cdot \boldsymbol{s} + 1}.$$

The parameter values, as in the previous section, can be calculated from the available Thiele-Small parameters. The band-pass filter nature of the derived transfer function can be observed in Fig. 5, where three different speakers' simulated reverse mode transfer functions are presented. (Details are explained in Section III.) These are similar to the measured electrical impedance curves usually included in the speakers' datasheets. This similarity is reasonable as the impedance seen by the electrical driving unit is also dominated by the mechanical properties of the speakers. These transfer functions will be used to simulate the speakers' responses to various v_{in} acoustical excitation signals.

III. Methods

The paper investigates the behavior of loudspeakers in reverse mode and examines the classification performances in cases when they are used as microphones to record sounds. A labeled urban sound dataset was used during the tests and - by using the derived reverse mode transfer functions - it was transformed into forms, as they would have been recorded by loudspeakers. These transformations were carried out with different speakers and with different noise levels. The classification algorithm is based on a deep learning approach. This section describes the applied methods, and the results are discussed in Section IV.

A. Reverse mode simulation

It is possible to record sounds with speakers, however, no dataset is available to evaluate their effect on the recorded signals. Therefore, a dataset recorded by microphones was used and the responses of speakers to these inputs were simulated by using their reverse mode transfer functions $H(s)$.

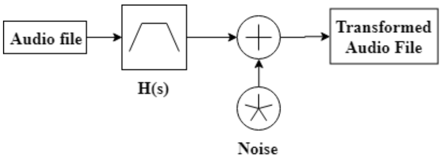

Fig. 4. Block diagram of the reverse mode simulation. The input files came from an urban sound dataset, and the responses to these acoustical signals were simulated by using the speakers' transfer functions. The final outputs were produced by adding noise to the response signals.

The original sound pressure levels (SPLs) of the input recordings and the amplification factors, microphone types, etc. were unknown, thus no exact simulation was possible. To simulate the effect of different input intensities, noises with different power levels - no noise, -10dB, -30dB, -40dB - were added to the transformed signals. High noise power lowers the SNR, thus lower input sound intensities were simulated. White Gaussian noise was applied with zero mean and the variance was set according to the required noise power level. The block diagram of the simulation step is presented in Fig. 4.

B. Data

A dataset called UrbanSound8k was used [10]. This dataset contains 8732 labeled sound excerpts ($\leq 4s$) of urban sounds from 10 classes: *air conditioner, car horn, children playing, dog bark, drilling, enginge idling, gunshot, jackhammer, siren, and street music*. The dataset is organized into 10 folds.

The audio files had various lengths, sample rates, bit-depths and number of channels. To unify these parameters, all of the recordings were converted to a single channel, 16-bit format with a sample rate of 22.05 kHz. First, each signal went through the reverse mode simulation and then they were split into overlapping segments with lengths of 0.95 s. The same preprocessing was carried out in [11], [12]. Each segment of a signal received the original signal's label. These labeled, transformed segments were used during the classification phase.

C. Urban Sound Classification

Modern classification algorithms rely on deep learning and neural networks. The best results were achieved by using these methods in the field of sound classification too [13]. Convolutional neural networks (CNN) are commonly used, which process their input data (mainly images, or multidimensional data) by successive convolution steps, and the kernels of these filters are formed during the training phase [14].

In this paper, a CNN was used to perform the classification task. The structure of the network is similar to the one published in [15]. The input format was changed; log-scaled mel-spectrograms were used as it was presented in [11], [12]. The neural network structure and the training process were implemented in Keras [16].

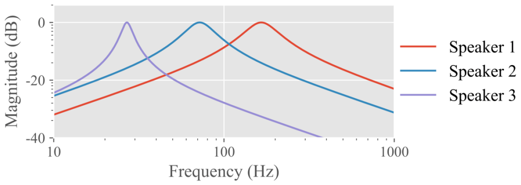

Fig. 5. Reverse mode transfer functions of three different loudspeakers. The band-pass filter nature of the curves can be observed. Amplifications were applied to set the magnitudes of the transfer functions to unity at the resonance frequencies.

Three different speakers were tested with 5 cm (Speaker 1, full-range [17]), 10 cm (Speaker 2, full-range [18]) and 20 cm (Speaker 3, woofer [19]) cone diameters. Based on their published Thiele-Small parameters, their reverse mode transfer functions were calculated and are presented in Fig. 5. The whole audio dataset was transformed four times per speaker with the four noise power levels described earlier. That resulted in 12 transformed datasets.

The CNN was trained on the 12 transformed datasets separately until convergence (early stopping was employed based on validation loss). From the datasets, 8 folds were used for training, 1 for validation and 1 for testing. To simplify the training method, instead of the required 10-fold cross-validation scheme, only one training was carried out on each dataset with the same folds selected for training, validation and testing ([1,8], 9, 10 respectively). Thus the results are comparable and do not require the 10× repetition of the training process on each dataset. The baseline accuracy was determined similarly without the reverse mode simulation step. The optimal learning rate parameter of the training procedure was determined on the baseline system first, and then the same value was used during all the other trainings. Further optimization could be made in each separate case, which might increase the accuracy, however, these results are beyond the interest of the current paper.

IV. Results

Loudspeakers are sub-optimal microphones, therefore it is foreseeable that their performance affects the classification accuracy negatively. Table I summarizes the resulting accuracies (true positives/all samples) in all the tested cases. The accuracy on the original dataset was 70%, which seems close to the state-of-the-art [12], however, no exact comparison can be made, because the 10-fold cross validation scheme was violated in this work. Still, the results are comparable within the table, since all the trainings and testings were carried out on the same, but differently transformed datasets.

It can be examined in Table I that the speakers' performances are at least with 9% worse than the baseline accuracy. This is mainly originated from the nature of the reverse mode frequency response. It is noticeable that the higher the diameter and lower the resonance frequency, the lower the classification accuracy becomes. Speaker 1 achieved the best performance, which can be explained by its highest resonance frequency, thus relevant frequencies are less attenuated.

TABLE I

Accuracies of the trained classifiers. In the columns the noise levels and in the rows the baseline setup and the different speakers are presented.

	No noise added	Noise: -40dB	Noise: -30dB	Noise: -10dB
Original	70%	-	-	-
Speaker 1	61%	60%	60%	50%
Speaker 2	56%	54%	55%	42%
Speaker 3	56%	58%	53%	32%

The noise level also had an impact, however, it was not necessarily negative in reasonable ranges (-40dB, -30dB). This is not surprising; adding noise to signals is a commonly used data augmentation method [15]. At -10dB noise power level, most of the input signals fell in the range of the noise, therefore only loud events were classified correctly.

To investigate the details of the accuracy drops caused by the reverse mode speakers, in Fig. 6 the difference between two confusion matrices are presented: *(baseline matrix - Speaker 1, no noise matrix)*. In the resulting matrix, the positive numbers on the diagonal represent the number of miss-classified events compared to the baseline classifier. Interestingly, the loud, impulsive events like *gunshots*, *dog bark*, and *car horn* were well-classified with the reverse mode speaker based classifier as well. Most of the error came from the less intense events with periodic nature like *drilling*, *air conditioner*, and *engine idling*. With the high frequencies attenuated, these periodical events produced similar spectrograms, therefore misled the classifier.

To better illustrate the nature of well-classified sounds, Table II presents the top 3 best performing classes in each test cases.

As it can be observed in Table II, events with high SPL and transient nature were more distinguishable by reverse mode speakers. For example, Speaker 1, a full-range speaker, preserved enough information to enable high gunshot detection accuracy. In all the cases, the *music* events were usually classified correctly, mainly because of their variant nature

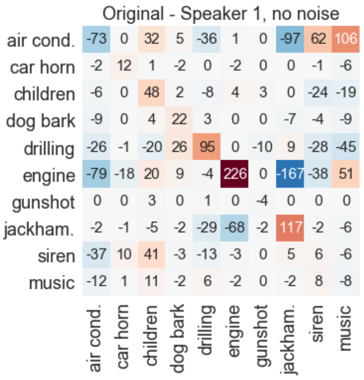

Fig. 6. The results of the baseline classifier trained on the original dataset are compared to the results of a classifier trained on a transformed dataset by illustrating the difference between their confusion matrices.

TABLE II
BEST PERFORMING CLASSES IN THE EXAMINED TEST CASES.

	No noise added	**Noise: -40dB**	**Noise: -30dB**	**Noise: -10dB**

Orig.	jackh.(93%)
	gunsh.(85%)
	car h.(81%)
		-	-	-

Sp. 1	gunsh.(91%)
	music(82%)
	jackh.(73%)
		gunsh.(89%)
		music(79%)
		dog b.(71%)
			gunsh.(89%)
			music(82%)
			car h.(74%)
				jackh.(73%)
				music(67%)
				car h.(63%)

Sp. 2	music(73%)
	jackh.(69%)
	drill(68%)
		music(78%)
		gunsh.(72%)
		drill(68%)
			jackh.(78%)
			music(72%)
			dog b.(70%)
				jackh.(66%)
				music(60%)
				dog b.(59%)

Sp. 3	jackh.(74%)
	music(73%)
	dog b.(63%)
		music(74%)
		jackh.(73%)
		dog b.(66%)
			jackh.(74%)
			music(68%)
			dog b.(64%)
				engi.(59%)
				music(41%)
				air c.(38%)

both in the frequency and in the time domains. However, most of the falsely categorized samples ended up in this class as well. As speakers with bigger cone diameters tend to attenuate higher frequencies more, the impulsive events lost their transient nature, therefore got miss-classified more times in these situations.

A. Discussion

From the classification results, it can be concluded that reverse mode speakers could be used for event detection. However, the type and nature of these events are limited. For example, reliable speech recognition could hardly be achieved because of the low sound pressure levels. At the same time, loud, impulsive events like gunshots, explosions, screaming, etc. could be detected with sufficient accuracy. The other limiting factor is the speakers' type. Woofers and subwoofers have lower resonance frequencies, therefore attenuate the relevant part of the spectrum more. This restriction is less critical, as typically full-range speakers are used in everyday applications.

V. SUMMARY AND FUTURE WORK

In this paper, the reverse mode (microphone mode) of loudspeakers was investigated in terms of urban sound classification performance. A reverse mode equivalent circuit and the corresponding transform function were derived from a well-known electrical equivalent circuit. These transfer functions were employed to simulate the speakers' responses to acoustical excitation signals. The model parameters can be calculated easily from manufacturer datasheets. Three different speakers were modeled and their effect on the classification accuracy were examined and compared. A labeled urban sound dataset and its transformed versions served as the input of the classification, which was carried out by a state-of-the-art neural network based classifier. The sound pressure level uncertainty of the original dataset was handled by adding noises with different power levels to the data.

The results suggested that loudspeakers could be used for event detection, however only loud, impulsive events like gunshots, explosions, screaming, etc. could be detected accurately by full-range speakers.

Based on the results, the research will continue in two directions in the future. First, investigating the possibility of designing 'smart speakers', which could detect events in their inactive state, while preserving privacy. The other direction is to examine the event detection capabilities during active state – when the speaker is being actively driven by electrical signals.

VI. ACKNOWLEDGEMENTS

The research was supported by the EFOP-3.6.3-VEKOP-16-2017-0002 and by the Ministry of Human Capacities, Hungary grant 20391-3/2018/FEKUSTRAT.

REFERENCES

[1] D. Medairos, "How to turn a speaker into a microphone." https://www.techwalla.com/articles/how-to-turn-a-speaker-into-a-microphone. [Online; accessed 27-February-2019].

[2] NPS Physics, "Speaker as a microphone - Why can a speaker work as a microphone?." https://www.youtube.com/watch?v=xpBt4AgkGXE. [Online; accessed 27-February-2019].

[3] M. Guri, Y. Solewicz, A. Daidakulov, and Y. Elovici, "SPEAKE(a)R: Turn Speakers to Microphones for Fun and Profit," in *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, (Vancouver, BC), USENIX Association, 2017.

[4] F. V. Hunt, *Electroacoustics: The Analysis of Transduction, and Its Historical Background (Harvard Monographs in Applied Science)*. Cambridge, MA, USA: Harvard University press, 1954.

[5] J. Borwick, *Loudspeaker and headphone handbook*. CRC Press, 2012.

[6] F. A. Everest, *Master handbook of acoustics*. ASA, 2001.

[7] L. L. Beranek and T. Mellow, *Acoustics: sound fields and transducers*. Academic Press, 2012.

[8] R. H. Small, "Direct radiator loudspeaker system analysis," *Journal of the Audio Engineering Society*, vol. 20, no. 5, pp. 383–395, 1972.

[9] Wikibooks, "Engineering acoustics/transducers - loudspeaker — wikibooks, the free textbook project." https://en.wikibooks.org/w/index.php?title=Engineering_Acoustics/Transducers_-_Loudspeaker&oldid=3232758, 2017. [Online; accessed 27-February-2019].

[10] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, (Orlando, FL, USA), pp. 1041–1044, Nov. 2014.

[11] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Sep. 2015.

[12] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental sound classification with dilated convolutions," *Applied Acoustics*, vol. 148, pp. 123 – 132, 2019.

[13] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11 – 26, 2017.

[14] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354 – 377, 2018.

[15] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, pp. 279–283, March 2017.

[16] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[17] "Dayton Audio DMA58-4, full-range loudspeaker." "http://www.loudspeakerdatabase.com/Dayton/DMA58#4%CE%A9". [Online; accessed 27-February-2019].

[18] "Dayton Audio DMA105-8, full-range loudspeaker." "http://www.loudspeakerdatabase.com/Dayton/DMA105#8%CE%A9". [Online; accessed 27-February-2019].

[19] "Dayton Audio E220CF-8, woofer loudspeaker." "http://www.loudspeakerdatabase.com/Dayton/E220CF-8". [Online; accessed 27-February-2019].