Semantically-Informed Deep Neural Networks For Sound Recognition

Michele Esposito, Giancarlo Valente, Yenisel Plasencia-Calaña, Michel Dumontier, Bruno L Giordano, Elia Formisano

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095606

SEMANTICALLY-INFORMED DEEP NEURAL NETWORKS FOR SOUND RECOGNITION

Michele Esposito[1], Giancarlo Valente[1], Yenisel Plasencia-Calaña[2], Michel Dumontier[3], Bruno L. Giordano[4], Elia Formisano[1,2]

[1] Department of Cognitive Neuroscience - Maastricht University, Maastricht, The Netherlands
[2] BISS institute - Maastricht University, Maastricht, The Netherlands
[3]Institute of Data Science - Maastricht University, Maastricht, The Netherlands
[4] Institut des Neurosciences de La Timone - Université Aix-Marseille, Marseille, France

ABSTRACT

Deep neural networks (DNNs) for sound recognition learn to categorize a barking sound as a "dog" and a meowing sound as a "cat" but do not exploit information inherent to the semantic relations between classes (e.g., both are animal vocalisations). Cognitive neuroscience research, however, suggests that human listeners automatically exploit higher-level semantic information on the sources besides acoustic information. Inspired by this notion, we introduce here a DNN that learns to recognize sounds and simultaneously learns the semantic relation between the sources (semDNN). Comparison of semDNN with a homologous network trained with categorical labels (catDNN) revealed that semDNN produces semantically more accurate labelling than catDNN in sound recognition tasks and that semDNN-embeddings preserve higher-level semantic relations between sound sources. Importantly, through a model-based analysis of human dissimilarity ratings of natural sounds, we show that semDNN approximates the behaviour of human listeners better than catDNN and several other DNN and NLP comparison models.

Index Terms— natural sound recognition, deep neural networks, auditory semantics, semantic embeddings, acoustic-to-semantic transformation

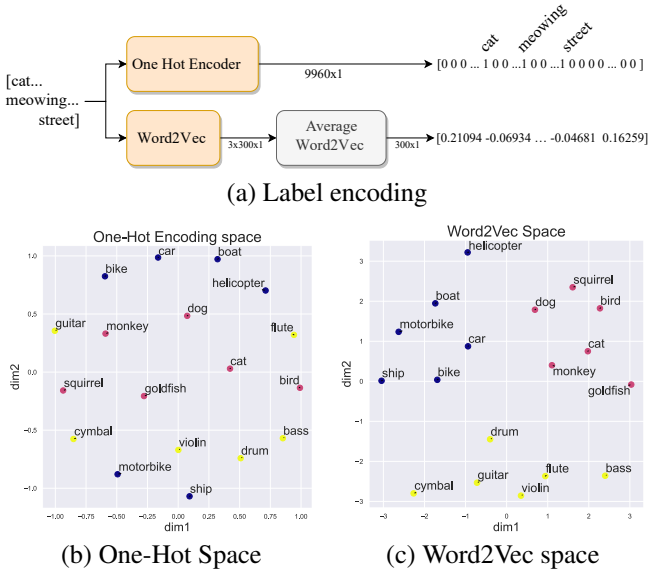
1. INTRODUCTION

Human sound recognition involves the transformation of acoustic waveforms into meaningful representations of the sound-producing source or event. Whereas this ability is automatic and effortless in humans, engineering artificial systems that reproduce human recognition performance has proven challenging. In machine learning (ML), sound recognition has been typically formulated as a classification problem, where sounds are assigned to predefined classes based on the analysis of various features extracted from the input acoustic signal. Different ML approaches have been proposed, showing promising results in several applications [1]. Recently, deep neural networks (DNNs) have been shown to outperform other conventional ML algorithms. Mimicking similar research on visual object recognition [2], sound-to-event DNNs have been used for sound classification tasks [3], [4], [5]. Trained on a large-scale dataset of human-labelled sounds (Audioset, [6]), Google's VGGish and Yamnet have provided remarkable performances. These networks receive spectrogram representations as input and can classify sounds in up to 527 and 521 classes, for VGGish and Yamnet, respectively. Although a taxonomic organization of labels has been proposed (Audioset, [6]), in most cases the information on the (hierarchical) relation between labels is not used explicitly to train the networks (but see Jimenez et. al [7]). Typically, labels are encoded as binary categorical variables, using one-hot or multi-hot (in case of multiple simultaneous labels) encoding; Fig. 1 (a,b).

Interestingly, recent cognitive neuroscience research has shown that sound-to-event DNNs (including VGGish and Yamnet) provide a good approximation of human listeners' behaviour in several real-world auditory perception tasks [8, 9]. Giordano et al. [9] considered behavioural data consisting of perceived sound (dis)similarities, estimated with a hierarchical sorting task [10] and examined to what extent sound-to-event DNNs, and other acoustic, auditory perception and semantic (natural language processing, NLP) models could explain these behavioural data. Results not only showed that sound-to-event DNNs outperformed all other models in predicting human sound dissimilarity judgements but also that NLP models, namely *word2vec* [11] and GloVe [12], predicted variance of behavioural data that could not be accounted for by sound-to-event DNNs trained using categorical labels. These findings suggest that, when listening to (and comparing) real-world sounds, human listeners automatically exploit higher-level semantic information on the sources besides acoustic information.

Inspired by these results, we sought to develop DNNs that - mimicking human perception [13] - learn to recognize

Fundings/support: Data Science Research Infrastructure (DSRI; Maastricht University); Dutch Research Council (NWO 406.20.GO.030 to EF); French National Research Agency (ANR-21-CE37-0027-01 to BLG).



(a) Label encoding

(b) One-Hot Space

(c) Word2Vec space

Fig. 1: **Categorical vs Semantic label encoding** (a) Label-encoding step for the two approaches. (b,c) Example visualization of one-hot vs word2vec spaces

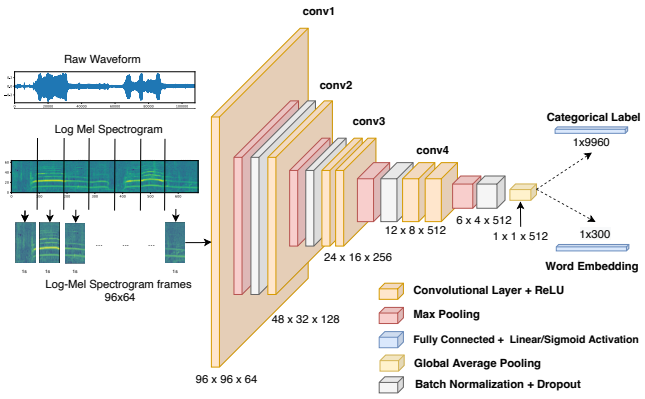


Fig. 2: **Pre-processing and catDNN/semDNN's architecture**. Note how the architectures differ only at the last dense layer.

sounds and simultaneously learn the semantic relation between the sources. To this aim, we formulated sound recognition as a deep-learning regression problem of mapping spectrograms onto a continuous, multidimensional space, quantitatively capturing the semantic relations between sound sources and events. In the present study, we obtained this multidimensional space from the *word2vec* embeddings of linguistic sound descriptions, Fig. 1 (a,c). To evaluate the effect of semantics on sound recognition DNNs, we trained two networks (semDNN and catDNN) with identical architecture (except for the output layer, see below) using, in one case, 300-dimensional *word2vec* embeddings of the linguistic sound description (semDNN; [11]), and categorical, one-hot encoded single words in the other case (catDNN).

To avoid biases in the comparison, our DNNs were trained from scratch, as all available pre-trained networks have been trained using categorical coding. Furthermore, for training, we curated a dataset of 388,211 sounds (2,584 hours), covering a broad range of real-world sounds (Super Hard Drive Combo [14]), characterized by the rich semantic description that we derived from the database metadata (see below).

Based on [9], we expected that semDNNs would better approximate human behaviour in auditory cognitive tasks than catDNNs, as the *word2vec* embeddings preserve the semantic relation between sound sources, which are instead lost with one-hot encoding, Fig. 1 (b,c).

Similar to our approach, other recent studies proposed to combine sound-to-event DNNs with language embeddings. Xie et al. [15] combined audio feature embeddings from VGGish and semantic class label embeddings from *word2vec* at the output stage using a bilinear model. However, the audio network was not explicitly trained to learn semantic embeddings in this case. Recently, Elizalde et al. [16] employed contrastive learning to combine sentence embeddings (BERT, [17]) with an audio encoder into a joint multimodal space. Our present work, however, focuses on evaluating the effects of semantic representation type (continuous vs categorical) on sound-to-event DNNs and assessing networks' ability to predict human perceptions.

2. METHODS AND MATERIAL

2.1. Network Architecture

We developed two different networks: semDNN and catDNN (Fig. 2). Both networks resemble VGGish (four main convolutional blocks; 64, 128, 256, and 512 filters), including the sound preprocessing and log-mel spectrogram input to the network. Compared to VGGish, we added a dropout layer (rate = 0.2; [18]) and a batch normalization layer [19] after each downsampling operation, and after the fully connected layers. We also applied global average pooling after the last convolutional block to summarize the feature maps into a fixed-length vector. The two networks differed only at the output layer, where semDNN has a 300-units (N dimensions of semantic embedding = 300) layer with linear activation, and catDNN has a 9,960-units (N dictionary words = 9,960) dense layer with sigmoid activation. We additionally trained a convolutional autoencoder (CAE) to assess the network behaviour with acoustic inputs only (no category/semantic label task; same architecture as in Fig. 2 for the encoder and reverted architecture for the decoder; see below for additional control networks). The loss function was adapted to the network task. We used binary cross entropy for catDNN (multi-classification task), an angular distance for semDNN (regression task), and mean square error for the CAE.

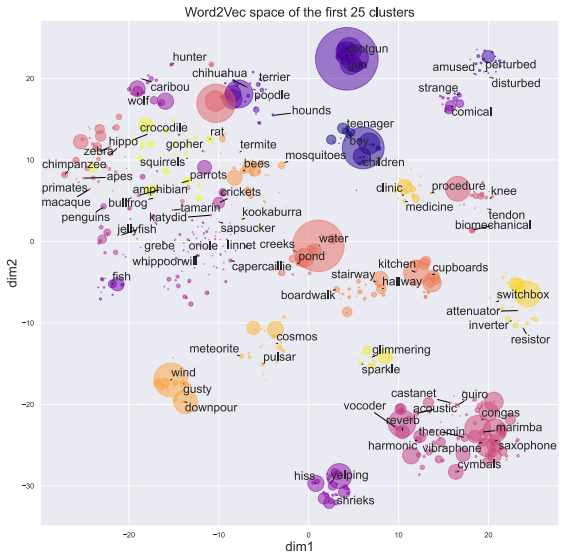


Fig. 3: **t-SNE visualization of 25 clusters of the semantic space**. Circle size indicates the number of occurrences of each word.

2.2. Semantically balanced dataset for training, validation, internal and external evaluation

Networks were trained using sounds and labels from SuperHard Drive Combo [14], a collection of 388,211 variable-length sounds (2,584 hrs) covering a wide range of sound sources and events. We then used a natural language processing pipeline to extract a dictionary of 9,960 sound-descriptive words from the database metadata (median nb. words/sound= 3, range = 1–15).

The distribution of the word descriptors in the database was unbalanced (Fig. 3). Thus, we created a *semantically balanced* dataset based on a hierarchical clustering analysis of the *word2vec* representations of the dictionary (input = pairwise cosine distance; ward-linkage). We considered a clustering solution with $K = 300$ semantic word clusters (Fig. 3). The resulting balanced dataset included 273,940 sounds (training set = 90% = 246,546 sounds; 1,366,848 frames; validation set 5%; internal evaluation = 5%).

Without further training, the networks were validated on four publicly-available external datasets: FSD50k [20], ESC-50 [21], Urban Sound 8K [22] and MSOS [23].

2.3. Comparative networks evaluation

We compared semDNN and catDNN relative to two prediction-accuracy metrics, one requiring the conversion of semDNN *word2vec* embedding predictions onto word predictions (Ranking score), and the other requiring the computation of *word2vec* embeddings of catDNN word predictions (average maximum cosine similarity, see below). To obtain single-word predictions for semDNN, we projected the predicted semantic embeddings, potentially reflecting a mixture of words, onto the single-word embeddings in the dictionary using non-negative least squares (NNLS) regression [24].

Ranking Score We first sorted the NNLS β-values (semDNN) and the output probabilities (catDNN) in ascending order to compare prediction accuracy. The *ranking score* was defined as:

$$m = 1 - \frac{rank - 1}{N - 1} \quad (1)$$

where N is the dictionary length and $rank$ is the position of β/probability corresponding to the "true" label. When labels included multiple words, we averaged the ranking score obtained for every single word.

Average maximum cosine similarity score (AMCSS) For each sound, we computed the cosine similarity between the $topN$ ($5 \leq topN \leq 15$) predicted words and a word in the true label and then considered its maximum value. This operation was repeated for each word in the label. The *AMCSS* was obtained as the average of these values (e.g. if the true label is 3 words long, 3 max values are obtained and then averaged).

2.4. Prediction of human behavioural data

We evaluated to what extent layer-by-layer embeddings of semDNN and catDNN, and of several control networks, including the CAE, and of additional control models, approximated perceived dissimilarity judgements obtained with humans ([10], Exp. 2, hierarchical sorting task; N sounds = 80). We adopted a cross-validated representational similarity analysis (RSA, [25, 9]), implying the comparison of behavioural data and model representations in the distance domain (model distance = cosine between-sound distances). We considered as additional comparison models. First, two NLP embeddings (word2vec and GloVe [11, 12]) to compare our audio-based learning of semantic relations in semDNN with text-based learning. Second, we considered three pre-published categorical sound-to-event DNNs (Yamnet, VGGish and Kell [4, 6, 8]), and three variants of the semDNN network (semDNN$_{\text{unbal}}$, trained with a randomly selected semantically unbalanced dataset, semDNN$_{\text{GloVe}}$, trained to learn GloVe embeddings, and semDNN$_{\text{notrain}}$, the random initialization of an untrained semDNN network).

3. RESULTS

3.1. Internal and external network-prediction accuracy

Fig. 4 shows the ranking score and AMCSS for semDNN vs catDNN, averaged over all evaluation sounds, for the internal (SuperHardDrive) and the four external datasets. In all cases, the ranking score was higher for semDNN, which

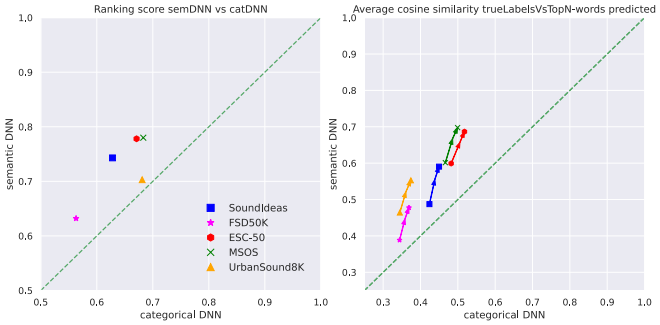


Fig. 4: **Networks comparison**, Sounds-averaged ranking scores (left) and AMCSS (right) for semDNN and catDNN. Arrows indicate increasing top-N from 5 to 15.

indicates that training using semantic embeddings produces more accurate predictions than categorical labels. AMCSS was also higher for semDNN in all cases, indicating that Top-N ranked words for semDNN predictions are more semantically related to the true labels than catDNN. This advantage increases when a larger number of words is considered (Fig. 4, arrows).

A relevant hypothesis was that semDNN embeddings would predict higher-order semantic relations between sounds better than catDNN. We tested the MSOS dataset (see 2.2), for which sounds are organized in five macro-classes (effects, human, music, nature, urban). For both semDNN and catDNN, we computed the pairwise cosine distance between sound embeddings in the last intermediate layer (Fig. 2, arrow). Fig. 5 shows that the semDNN embedding (middle panel) reflects better the macro-class organization (left panel) than the catDNN embedding (right panel; correlation with True categorical model = 0.330 and 0.193 for semDNN and catDNN, respectively).

3.2. Comparisons on human behavioural data

Fig. 6 shows the results of the RSA of human dissimilarity ratings of natural sounds for SemDNN, catDNN and other

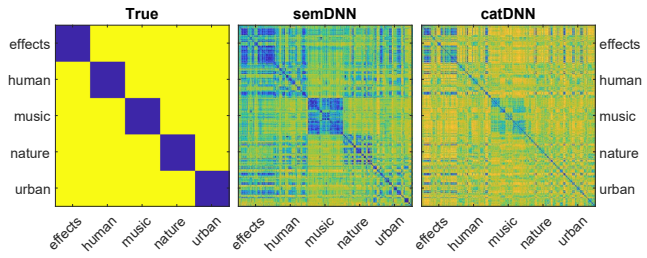


Fig. 5: Dissimilarity matrices obtained from the embeddings at the last common layer for semDNN (centre) and catDNN (right). Matrix on the left reflects the true macro-classes (min/max distance = blue/yellow).

compared models. SemDNN (R^2_{cv}=0.371) outperformed catDNN (R^2_{cv} = 0.199) and other DNNs trained using categorical labels (VGGish, R^2_{cv} = 0.226; Yamnet, R^2_{cv} = 0.302; Kell, R^2_{cv} = 0.179. Also, embeddings from SemDNN explained human behavioural data better than embeddings derived by applying NLP models (word2vec (R^2_{cv} = 0.211; GloVe, R^2_{cv} = 0.156) to sound descriptors. These results confirm our hypothesis that a network combining acoustic and semantic information approximates human behaviour in auditory cognitive tasks better than models considering acoustic (categorical sound-to-event DNNs) or semantic (NLP) information alone. Finally, SemDNN outperformed several other DNNs with the same architecture that we trained in different configurations to evaluate the effects of individual factors: CAE (convolutional autoencoder, R^2_{cv} = 0.090), $semDNN_{notrain}$ (random initialization for untrained semDNN, R^2_{cv} = 0.034), $semDNN_{unbal}$ (semDNN trained on unbalanced dataset, R^2_{cv} = 0.265), and $semDNN_{GloVe}$ (semDNN trained to learn GloVe embeddings, R^2_{cv} = 0.191).

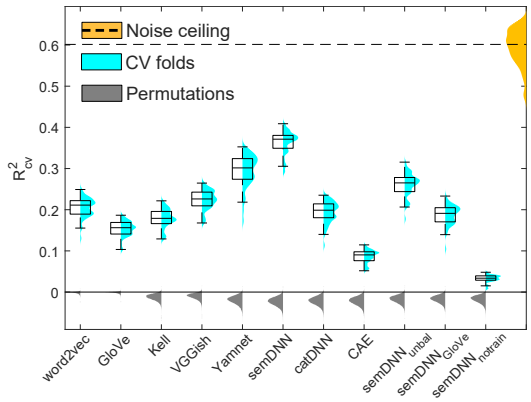


Fig. 6: **Behaviour prediction.** SemDNN outperforms all other models at predicting perceived sound dissimilarity.

4. CONCLUSIONS

We systematically investigated the effects of training DNNs for sound recognition with continuous semantic embeddings (*word2vec*) vs categorical labels (one-hot encoding). We showed that training with continuous embeddings is beneficial, as it produces semantically more accurate labelling of sounds. Importantly, using human behavioural data, we showed that DNNs trained with continuous semantic embeddings approximate human behaviour better than categorical DNNs. Here, we considered *word2vec* embeddings of the linguistic sound descriptions to retain information on the (linguistic) semantic relations between the sound sources. In the future, the same approach could be extended to different types of semantic embeddings, for example, derived from natural sound ontologies [26].

5. REFERENCES

[1] A. Bansal and N. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent Systems with Applications*, p. 200115, 2022.

[2] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.

[3] J. J. Huang and J. J. A. Leanos, "Aclnet: efficient end-to-end audio classification cnn," *arXiv preprint arXiv:1811.06669*, 2018.

[4] S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, et al., "CNN architectures for large-scale audio classification," in *Proc. ICASSP 2017*, 2017, pp. 131–135.

[5] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM*, vol. 28, 2020.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, 2017.

[7] A. Jimenez, B. Elizalde, and B. Raj, "Sound event classification using ontology-based neural networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2018, vol. 9.

[8] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, 2018.

[9] B. L. Giordano, M. Esposito, G. Valente, and E. Formisano, "Intermediate acoustic-to-semantic representations link behavioural and neural responses to natural sounds," *Nature Neuroscience (In Press)*, 2023.

[10] B. L. Giordano, J. McDonnell, and S. McAdams, "Hearing living symbols and nonliving icons: Category-specificities in the cognitive processing of environmental sounds," *Brain Cogn*, vol. 73, pp. 7–19, 2010.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint: 1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 (EMNLP)*.

[13] L. M. Heller, B. Elizalde, B. Raj, and S. Deshmukh, "Synergy between human and machine approaches to sound/scene recognition and processing: An overview of icassp special session," 2023.

[14] "SuperHardDriveCombo," https://www.sound-ideas.com/Product/28/Super-Hard-Drive-Combo, [Online].

[15] H. Xie, O. Räsänen, and T. Virtanen, "Zero-shot audio classification with factored linear and nonlinear acoustic-semantic projections," 2020.

[16] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*.

[20] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM*, vol. 30, pp. 829–852, 2022.

[21] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[22] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," *Proceedings of the 2014 ACM*, , no. 3, pp. 1041–1044, 2014.

[23] "MSOS-dataset," https://cvssp.org/projects/makingsenseofsounds/site/challenge/, 2013, [Online].

[24] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, SIAM, 1995.

[25] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis – connecting the branches of systems neuroscience," *Front Syst Neurosci*, vol. 2, 2008.

[26] B. L. Giordano, R. de Miranda Azevedo, Y. Plasencia-Calaña, E. Formisano, and M. Dumontier, "What do we mean with sound semantics, exactly? a survey of taxonomies and ontologies of everyday sounds," *Frontiers in Psychology*, vol. 13, 2022.