IMINET: CONVOLUTIONAL SEMI-SIAMESE NETWORKS FOR SOUND SEARCH BY VOCAL IMITATION

Yichi Zhang and Zhiyao Duan

University of Rochester, Dept. of Electrical and Computer Engineering, Rochester, NY 14623, USA
yichi.zhang@rochester.edu, zhiyao.duan@rochester.edu

ABSTRACT

Searching sounds by text labels is often difficult, as text labels cannot always provide sufficient information for the sound content. Previously we proposed an unsupervised system called IMISOUND for sound search by vocal imitation. In this paper, we further propose a Convolutional Semi-Siamese Network (CSN) called IMINET. IMINET uses two towers of Convolutional Neural Networks (CNN) to extract features from vocal imitations and sound recordings, respectively. It then adopts a fully connected network to predict the similarity between vocal imitations and sound recordings. We propose three different configurations of the CSN by choosing different weight sharing strategies between the two towers. We also propose late fusion of the retrieval results of IMINET's different configurations and those of IMISOUND as a baseline. Experiments show significant improvements of the retrieval performance from the IMISOUND baseline to the fusion of IMINET's different configurations, and to different fusions between IMINET and the IMISOUND baseline.

Index Terms— Vocal imitation, information retrieval, convolutional Siamese network, metric learning

1. INTRODUCTION

Vocal imitation is a common human behavior that uses vocal organs to mimic sounds. It is an effective way to convey ideas that are difficult or insufficient to describe with languages in human communication. These difficulties may be due to the language barrier between the communicating parties, or due to the fact that certain sounds do not have a definite semantic meaning such as computer-synthesized sound effects. In many scenarios, vocal imitation also augments language descriptions and makes the concepts being conveyed more vivid. For example, audiences of the National Public Radios Car Talk show [1] call in to describe symptoms of their vehicles by imitating the sounds caused by mechanical or electrical failures to seek advice. These imitations make the conversations more effective and fun.

Designing computer systems that can recognize vocal imitation for sound search [2, 3] extends human-computer interaction and has broad applications in multimedia retrieval, music production, security and surveillance, and biomonitoring. Current large-scale sound libraries such as freesound.org are indexed by text labels. These text labels, however, often do not convey enough details of the sound, and even if they do, memorizing them is difficult. Vocal-imitation-based search allows users to search sounds based on details not described by text labels. This is especially useful for large-scale libraries where many different sounds share the same text labels and search for target sounds in long recordings.

There are two main challenges in designing vocal-imitation-based sound search systems: feature representation and matching algorithms. Feature representations of vocal imitations and real sounds need to be robust to different aspects (e.g., pitch, timbre, rhythm) that humans emphasize in different imitations for different sounds. They also need to consider differences between imitations and real sounds due to the physical constraints of the human vocal system. The matching algorithm needs to work with the feature representations to discern target sounds from irrelevant ones for a given query. In [4, 5], we proposed an unsupervised system for sound search by vocal imitation called IMISOUND. We employed a Stacked Auto-Encoder (SAE) to learn feature representations from training vocal imitations and applied the same representation for both imitation queries and sound candidates. We then calculated their similarity through Kullback-Leibler (K-L) divergence [6], Dynamic Time Warping (DTW) [7], and the cosine distance. The feature representation and matching algorithm in IMISOUND, however, were designed separately.

In this paper, we propose another neural network model called IMINET for sound search by vocal imitation that jointly optimizes feature learning and the matching algorithm. As shown in Figure 1, IMINET is a Convolutional Semi-Siamese Network (CSN) that contains 1) two Convolutional Neural Network (CNN) towers for feature learning for vocal imitations (query) and sound recordings (candidate) respectively; and 2) a Fully Connected Network (FCN) for feature learning that classifies the query-candidate feature concatenations into positive and negative pairs. Different weight sharing strategies between the two convolutional towers are proposed, resulting in different versions of IMINET. Through this joint optimization, feature representations learned by the convolutional layers are better tuned for the FCN's metric learning. Experiments show that different versions of IMINET achieve comparable or higher sound search performance than IMISOUND. When retrieval results from these versions are fused, IMINET clearly outperforms IMISOUND. By fusing outputs of IMINET and IMISOUND, the retrieval performance is further boosted significantly.

2. RELATED WORK

Query by vocal imitation falls into the task of Query by Example (QEB) [8]. QEB has been applied to sound related applications like query by humming [9], query by beat boxing [10], cover song recognition [11], and spoken document retrieval [12]. However, little work has been reported on sound search by vocal imitation.

Roma and Serra [13] designed a system that allows users to search sounds on freesound.org by recording audio with a micro-

This work is funded by the National Science Foundation grant No. 1617107. We acknowledge NVIDIA's GPU donation for this research.

978-1-5386-1631-4/17/$31.00 ©2017 IEEE

phone, but no formal evaluation was reported. Blancas et al. [2] built a supervised system using hand-crafted features by the Timbre Toolbox [14] and an SVM classifier. A vocal imitation query was classified to a class and sounds in that class were retrieved. The major limitation of supervised systems, however, is that they cannot retrieve sounds that do not have training imitations. Helén and Virtanen [15] designed a query by example system for generic audio. Hand-crafted frame-level features were extracted from both query and sound samples and the query-sample pairwise similarity was measured by probability distribution of the features. In our previous work [3], we first proposed a supervised system using a Stacked Auto-Encoder for automatic feature learning and an SVM for imitation classification. Considering the limitation of supervised systems, we then proposed an unsupervised system called IMISOUND [4]. The SAE was adopted for feature extraction and various distances were adopted for query-sample similarity measure.

In another aspect, Siamese network was first proposed by Bromley et al. [16] for signature verification. Since then, it has been successfully applied to many image/video tasks such as face verification [17] and image recognition [18]. More recently, Bertinetto et al. [19] proposed a fully-convolutional Siamese network for object tracking in videos. Han et al. [20] proposed the MatchNet for patch-level image matching, a two tower structure with convolutional layers for feature extraction and fully connected layers for metric learning. However, little work has been reported in the audio domain. Chen and Salman [21] proposed a regularized Siamese deep network to extract speaker-specific information from MFCCs.

3. THE IMINET MODEL

The IMINET model is a Convolutional Semi-Siamese Network (CSN). The overall structure is shown in Figure 1. The two towers of convolutional layers receive a vocal imitation query and a sound candidate as their input, respectively, and learns feature representations. These features are then concatenated and fed into a Fully Connected Network (FCN) for metric learning. The final output of the IMINET is a probability indicating whether the two inputs are of the same concept (positive pair) or not (negative pair). The IMINET network structure is built using Keras v2.0.3 [22].

3.1. Preprocessing

Both the vocal imitations and sound recordings are first downsampled to 16 kHz. A 6-octave (50-3200 Hz) Constant-Q Transform (CQT) is then employed to calculate their spectrograms using the MATLAB CQT toolbox [23]. The CQT uses 12 bins in each octave and a hop size of 26.25 ms. Considering the fixed size input for convolution in the two towers, imitation and recording CQT spectrograms are truncated to 129 frames. Spectrograms shorter than 129 frames are zero-padded. Therefore, the CQT spectrograms have a dimensionality of 72*129 (frequency bins * time frames). The reasons to use CQT instead of linear-frequency spectrograms are two fold: 1) the log-frequency scale in CQT better corresponds to human auditory perception; 2) the representation is more compact for the ease of network training.

3.2. Convolutional layers for feature learning

Each tower of the Siamese network is a Convolutional Neural Network (CNN) with 4 convolutional layers. The parameters are shown on the upper right side in Figure 1. Both towers receive a 72*129

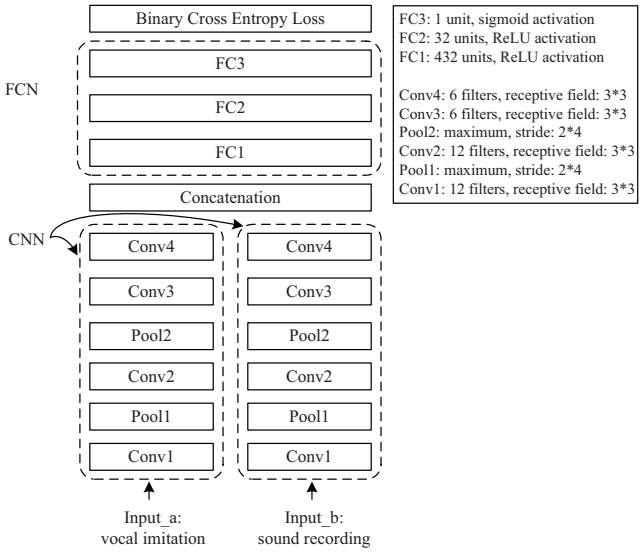

Figure 1: The proposed IMINET structure.

sized CQT spectrogram as input. Both Conv1 and Conv2 have 12 filters with a receptive field of 3*3, and followed by a Rectified Linear Unit (ReLU) activation function. They are then each followed by a 2*4 (both shape and stride) max-pooling layer with 2 in frequency and 4 in time. For Conv3 and Conv4, each has 6 filters with ReLU activations, but no pooling layer follows.

Besides sharing the same architecture, Siamese networks usually tie the parameters of the two towers, i.e., the two inputs pass through the exactly same networks for feature learning. This is suitable when the two inputs share many traits in common, i.e., image matching [20]. In our work, however, vocal imitations lie in a much more restricted sound space than general sound recordings, due to the physical constraints of the human vocal system. Conceptually, vocal imitations and sound recordings should pass through two different feature learning networks. Therefore, in IMINET we propose three configurations when designing the convolutional towers:

(1) Tied: The two towers share the exactly same weights and biases in all layers.

(2) Untied: The two towers do not share weights and biases at all, although their structures are the same. This allows the two towers to be tuned for their input domains independently.

(3) Partially tied: The weights and biases in the two towers are not shared for Conv1 and Conv2 layers, but are shared for Conv3 and Conv4 layers. The rationale behind this design is that layers close to the input should be tuned to adapt to the input's unique characteristics and extract surface-level features that are closely related to the specific input domain, while deeper layers should behave like "grandmother cells" [24] that extract more complex and highly conceptual features [25] that are shared across input domains.

As in both untied and partially tied configurations, the symmetry between the two towers are less strict, we call such structures semi-Siamese networks.

3.3. Fully connected layers for metric learning

After the features from the two towers are extracted, they are flattened and concatenated. Then they are fed into a 3-layer Fully Con-

nected Network (FCN). There are 432 and 32 neurons in FC1 and FC2, respectively, where the ReLU activation function is adopted. To avoid overfitting, we use 20% dropout on both FC1 and FC2. FC3 has only one neuron which uses the sigmoid activation function to squash the output value between 0 and 1. This value is viewed as the probability indicating whether the imitation-recording pair is a positive pair (i.e., correct match).

The FCN can be viewed as a metric learning network that learns the similarity between vocal imitations and sound recordings from positive and negative training pairs. Compared to common distance/similarity measures such as Euclidean distance or cosine distance, this metric/similarity is learned together with the feature representations of the vocal imitations and sound recordings, likely leading to a better retrieval performance.

3.4. Training

Training the network requires positive and negative pairs of vocal imitations and sound recordings. We combine vocal imitations with the original sound recordings that they imitate into positive pairs, and with other sound recordings as negative pairs. We adopt Adaptive Moment Estimation (Adam) optimization algorithm [26] to minimize the loss function of binary cross-entropy between the probability output and the ground-truth label, where 1 and 0 denotes positive and negative pairs, respectively. The learning rate is 0.001, β_1 and β_2 are 0.9 and 0.999, respectively, ϵ is 1e-8. The batch size is 128. Early stopping based on validation loss with patience of 5 epochs is employed for training termination.

3.5. Sound retrieval

Once IMINET is trained, it can be used to search sounds for a vocal imitation query. To do so, we pair the imitation query with each sound candidate in the library and use IMINET to calculate its likelihood of being a positive pair. Let the likelihood for the i-th sound candidate be $L_{csn}(i)$. We then rank all sound candidates by their likelihood from high to low and return the top ones to the user.

It is noted that although IMINET is trained in a supervised way, its use for sound retrieval is totally unsupervised. In other words, IMINET needs not to be trained on imitation-sound pairs of a certain sound concept for it to be used to retrieve that sound concept. This is similar to unsupervised sound retrieval systems that use predefined distance/similarity measures [4, 5, 15].

3.6. Late fusion

Inspired by ensemble learning [27], we consider to fuse the retrieval results of the three configurations of IMINET by multiplying their outputs (i.e., likelihood values):

$$L_{fusion}(i) = L_{tied}(i) * L_{untied}(i) * L_{partial}(i), \quad (1)$$

where $L_{tied}(i)$, $L_{untied}(i)$, and $L_{partial}(i)$ are the pairing likelihood between the query and the i-th sound candidate, by tied, untied, and partially tied models, respectively.

We also consider to fuse the retrieval results of IMINET with those of IMISOUND [5]. As described before, IMISOUND uses a two-hidden-layer Stacked Auto-Encoder (SAE) [28] to extract features for a vocal imitation and a sound candidate. It then calculates the cosine distance between their feature representations. To fuse this result with that of IMINET, we convert the distance to a likelihood through a softmax function:

$$L_{sae}(i) = \frac{e^{-D(i)}}{\sum_{n=1}^{N} e^{-D(n)}}, \quad (2)$$

where $D(i)$ is the cosine distance between the vocal imitation and the i-th sound candidate; N is the total number of sound candidates in the library. Then the fusion between IMINET and IMISOUND can be done by multiplying their likelihood values:

$$L_{fusion}(i) = L_{csn}(i) * L_{sae}(i). \quad (3)$$

IMINET and IMISOUND have different structures and training objectives. In particular, IMINET learns feature representations in a supervised way with the goal of helping discriminate positive and negative pairs, while IMISOUND learns features in an unsupervised way which aims at a good reconstruction of the input. In addition, IMINET learns the similarity between vocal imitations and sound recordings from training data, while IMISOUND uses a pre-defined distance measure. Therefore, it is expected that they perform differently on the same imitation-sound pair and fusing their results may improve the retrieval performance.

4. EXPERIMENTS

4.1. Dataset

We adopt VocalSketch Data Set v1.0.4 [29] in our experiments. This dataset contains 120 sounds with distinct concepts and 10 vocal imitations of each recording obtained from different Amazon Mechanical Tuckers. The sounds and imitations are 3-second long on average. The sounds fall into 4 categories, namely Acoustic Instruments (AI), Commercial Synthesizers (CS), Everyday (ED), and Single Synthesizer (SS). The number of sounds in these categories is 40, 40, 120, and 40, respectively. We choose half of the sounds of each category (i.e., 20, 20, 60, and 20 from AI, CS, ED, SS, respectively) and all of their imitations to compose a dataset to train and validate the IMINET. We use the other half sounds and their imitations to test the IMINET. Therefore, training and testing materials do not share any sounds nor imitations.

For the 120 sounds used for training and validation, we choose 7 imitations of each sound to form 120 * 7 = 840 positive pairs and 840 negative pairs to train the IMINET. Positive pairs are pairs of an imitation and its target sound. Negative pairs are created by randomly pairing an imitation with an irrelevant sound. We use the rest 3 imitations of each sound to compose 120 * 3 = 360 positive pairs and 360 negative pairs to validate the IMINET.

We evaluate IMINET sound search performance within each category of the remaining 120 sounds and their imitations. By taking one vocal imitation from the AI category for example, it is paired with all the 20 candidate sound recordings to form 20 pairing test samples. In total there are 10 * 20 = 200 pairing test samples in the AI category. To get statistically reliable results, we train the IMINET 10 times with different initializations and evaluate their sound search performance.

4.2. Evaluation measure

Same as our previous work, we employ Mean Reciprocal Rank (MRR) [30] to evaluate the search performance in each category:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}, \quad (4)$$

Table 1: MRR (mean ± std) comparisons of various IMINET configurations with the baseline system.

	Configuration	Acoustic Instr.	Comm. Synthesizers	Everyday	Single Synthesizer
Baseline	IMISOUND	0.425 ± 0.015	0.311 ± 0.007	0.126 ± 0.003	0.373 ± 0.012
Proposed	Untied	0.377 ± 0.019	0.318 ± 0.020	0.154 ± 0.014	0.325 ± 0.020
Proposed	Partial	0.384 ± 0.027	0.304 ± 0.015	0.154 ± 0.015	0.340 ± 0.031
Proposed	Tied	0.401 ± 0.028	0.327 ± 0.019	0.158 ± 0.012	0.380 ± 0.018
Proposed	Untied + Partial + Tied	0.438 ± 0.015	0.343 ± 0.020	0.175 ± 0.012	0.382 ± 0.013
Proposed	Untied + IMISOUND	0.470 ± 0.025	0.356 ± 0.011	0.168 ± 0.010	0.402 ± 0.022
Proposed	Partial + IMISOUND	0.496 ± 0.018	0.346 ± 0.025	0.173 ± 0.014	0.417 ± 0.025
Proposed	Tied + IMISOUND	0.504 ± 0.014	0.355 ± 0.016	0.171 ± 0.009	0.452 ± 0.020
Proposed	Untied + Partial + Tied + IMISOUND	0.520 ± 0.020	0.371 ± 0.013	0.188 ± 0.007	0.447 ± 0.012

where $rank_i$ is the rank of the target sound among all sounds in the same category for the i-th vocal imitation query; Q is the number of imitations in each category. MRR ranges from 0 to 1 with a higher value indicating a better sound retrieval performance. For example, an MRR of 0.5 suggests that, on average, the target sound is ranked the second among all sounds in the category, while an MRR of 0.25 suggests that, on average, the target sound ranks the fourth. We report the average MRR and standard deviation across 10 runs of the system. We compare with our previous IMISOUND system [5].

4.3. Experimental results

Table 1 shows performance comparisons of different configurations of the proposed IMINET, different fusing strategies, and the IMISOUND baseline [5]. Several interesting observations can be made. First, from untied to partially tied to tied configurations of IMINET, the MRR increasing trend is observed in all categories. We also observe the similar trend when these configurations are fused with IMISOUND. This is unexpected, as we thought that partially tied or untied configuration could better account for the differences between vocal imitations and sound recordings and result in better retrieval performance. A possible explanation could be that the number of parameters is reduced in the tied configuration, which makes the network easier to train considering the small amount of training data. This may suggest that data scarcity is a bottleneck hindering the potential exploitation of more complicated models.

Second, the best performing IMINET configuration, tied, outperforms IMISOUND on two categories (Commercial Synthesizers and Everyday), underperforms on the Acoustic Instrument category, and achieves comparable performance on the Single Synthesizer category. An unpaired t-test shows that the MRR improvement is statistically significant for Commercial Synthesizers (p = 1.17e-2) and Everyday (p = 6.15e-6) at the significance level of 0.05.

Third, by fusing different configurations of IMINET, the MRR is better than each configuration itself. The MRR improvements for all categories except Signal Synthesizer are statistically significant (Acoustic Instruments p = 3.08e-2, Commercial Synthesizers p = 2.70e-4, and Everyday p = 8.30e-8), at the significance level of 0.05. This is because under different weight constraints, each configuration tends to learn its unique features. We believe that these features are complementary to some extent, explaining why the fused model outperforms every single configuration.

Fourth, by late fusion of the IMISOUND with each IMINET configuration, the MRR is boosted significantly. The lowest MRR after late fusion comes from Untied + IMISOUND combination, but it still significantly outperforms both IMISOUND and Untied, at the significance level of 0.05. In the Acoustic Instruments category, IMISOUND gets an MRR of 0.425. This means that on average, the target sound is ranked between the 2nd and the 3rd. By combining IMISOUND with Untied CSN, the MRR is increased by 10.6%. For the rest 3 categories, the MRR is also increased by 14.5%, 33.3%, and 7.8%, respectively. This observation verifies our hypothesis earlier, that thanks to the intrinsic differences between IMISOUND and IMINET on both network structure and training objective, their retrieval results are likely to be complimentary and fusing them can improve the performance significantly.

Finally, we can achieve the highest MRR in general by fusing all configurations of IMINET together as well as IMISOUND. In the Acoustic Instruments category, the MRR is as high as 0.520. The Everyday sound category has the lowest MRR of 0.188, but still improves significantly from IMISOUND's MRR of 0.126. It suggests that the target sound is ranked between the 5th and 6th on average, among the 60 sound candidates in that category. By conducting the unpaired t-test, we found salient MRR improvement comparing with IMISOUND: Acoustic Instruments p = 8.41e-10, Commercial Synthesizers p = 3.72e-9, Everyday p = 1.67e-12, and Single Synthesizer p = 2.51e-11, at the significant level of 0.001.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a Convolutional Semi-Siamese Network (CSN) called IMINET to search sounds by vocal imitation from a sound library in an unsupervised manner. It uses two towers of Convolutional Neural Networks (CNN) to extract features from vocal imitations and sound recordings, respectively, and then uses a Fully Connected Network (FCN) to predict the similarity between the imitation and the sound. We proposed three different configurations of the CSN by choosing different weight sharing strategies between the two towers. We also proposed late fusion of the retrieval results of IMINET's different configurations and those of a baseline system named IMISOUND. Experiments show significant improvements of the retrieval performance from the IMISOUND baseline to the fusion of different configurations of IMINET, and to different fusions between IMINET and the IMISOUND baseline. For future work, we would like to implement data augmentation of the CSN model to alleviate the data scarcity problem. We also would like to combine Recurrent Neural Networks (RNN) with Siamese Networks together to model the temporal evolution of both vocal imitations and sound recordings. Finally, we would like to conduct subjective studies for our system.

6. REFERENCES

[1] http://www.cartalk.com [Accessed 07/27/2017].

[2] D. S. Blancas and J. Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *Proc. Audio Engineering Society 53rd International Conference on Semantic Audio*, 2014, pp. 1–6.

[3] Y. Zhang and Z. Duan, "Retrieving sounds by vocal imitation recognition," in *Proc. Machine Learning for Signal Processing (MLSP), 2015 IEEE International Workshop on*, 2015, pp. 1–6.

[4] ——, "IMISOUND: an unsupervised system for sound query by vocal imitation," in *Proc. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 2269–2273.

[5] ——, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of the Audio Engineering Society*, vol. 64, no. 7/8, pp. 533–543, 2016.

[6] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[7] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transaction on*, vol. 26, no. 1, pp. 43–49, 1978.

[8] M. M. Zloof, "Query-by-example: A data base language," *IBM Systems Journal*, vol. 16, no. 4, pp. 324–343, 1977.

[9] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," in *Proc. the 3rd ACM International Conference on Multimedia*, 1995, pp. 231–236.

[10] A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beating-boxing: Music retrieval for the DJ," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2004, pp. 170–177.

[11] T. Bertin-Mahieux and D. P. Ellis, "Large-scale cover song recognition using hashed chroma landmarks," in *Proc. Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, 2011, pp. 117–120.

[12] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "A lattice-based approach to query-by-example spoken document retrieval," in *Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 363–370.

[13] G. Roma and X. Serra, "Querying freesound with a microphone," in *Proc. the 1st Web Audio Conference (WAC)*, 2015.

[14] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signal," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[15] M. Helén and T. Virtanen, "Audio query by example using similarity measures between probability density functions of features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 179303, 2009.

[16] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 1994, pp. 737–744.

[17] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, 2005, pp. 539–546.

[18] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. the 32nd International Conference on Machine Learning (ICML)*, 2015.

[19] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 850–865.

[20] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proc. Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 3279–3286.

[21] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized siamese deep network," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 298–306.

[22] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[23] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. the 7th Sound and Music Computing Conference*, 2010.

[24] C. G. Gross, "Genealogy of the 'grandmother cell'," *The Neuroscientist*, vol. 8, no. 5, pp. 512–518, 2002.

[25] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. the 26th annual international conference on machine learning*, 2009, pp. 609–616.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] D. W. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *The Journal of Artificial Intelligence Research*, pp. 169–198, 1999.

[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[29] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 43–46.

[30] D. R. Radev, H. Qi, H. Wu, and W. Fan, "Evaluating web-based question answering systems," *Ann Arbor*, vol. 1001, p. 48109, 2002.