IMISOUND: AN UNSUPERVISED SYSTEM FOR SOUND QUERY BY VOCAL IMITATION

Yichi Zhang, Student Member, IEEE, Zhiyao Duan, Member, IEEE

Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA

ABSTRACT

Vocal imitation is widely used in human interactions. In this paper, we propose a novel human-computer interaction system called IMISOUND that listens to a vocal imitation and retrieves similar sounds from a sound library. This system allows users to search sounds even if they do not remember their semantic labels or the sounds do not have these labels (e.g., synthesized sound effects). IMISOUND employs a Stacked Auto-Encoder (SAE) to extract features from both the vocal imitation (query) and sounds in the library (candidates). The SAE is pre-trained using training vocal imitations of sounds not in the library to automatically learn more suitable feature representations than human-engineered features such as MFCC's. It then measures the similarity between the query and each sound candidate, using the K-L divergence and Dynamic Time Warping distance between their feature representations, and finally retrieves the closest sounds. IMISOUND is an unsupervised system in the sense that no training is performed for the target sound, nonetheless, experiments show that it achieves comparable performance to a previously proposed supervised system which requires pre-training on sounds to be retrieved. Experiments also show that IMISOUND significantly outperforms an unsupervised MFCC-based baseline system, validating the advantage of the SAE feature representation.

Index Terms—Vocal imitation, information retrieval, automatic feature learning, stacked auto-encoder

1. INTRODUCTION

Vocal imitation is of great importance in human interactions. We use it to convey concepts of sounds that are difficult to describe in language [1]. This may be because the communicating parties speak different languages, or because the language description of the sound lacks the desired vividness (e.g., a "Christmas tree" dog barking sound), or even because the sound does not have a clear association to a language description (e.g., many computer-synthesized sounds). Commonly used vocal imitations are further abstracted into onomatopoeia and perpetuated into languages [2].

Computer systems that are able to recognize vocal imitations will enable novel human-computer interactions. For example, current sound libraries are indexed by text labels. To search for a sound, a user has to remember the sound's text labels such as name, keyword, and the way the sound was produced, etc. This process can be very tedious when the sound library is large or when the sound does not have a clear association to text labels. A system that supports sound search through vocal imitation can make the search process more efficient and effective.

However, there are several challenges to design such a sound retrieval system. One of the most difficult problems is finding appropriate feature representations for vocal imitations. People tend to imitate different aspects for different sounds. For instance, to imitate a "du-du (car horn)" sound, constant pitch and the time gap between the two horns are likely to be emphasized, while for a "cat meowing" imitation, timbral evolution might be paid more attention. Even for the same sound concept, different people may imitate differently in terms of pitch, timbre, rhythm, loudness, and their temporal evolution. On the other hand, due to the physical constraints of human vocal folds and tract, the variety of vocal imitations are confined within a small subspace relative to the entire sound space.

In our previous work [3], we proposed a supervised vocal imitation recognition system for sound retrieval. We employed a Stacked Auto-Encoder (SAE) [4] to automatically learn features from training vocal imitations. We found that these learned features are more suitable to represent vocal imitations than those commonly-used, hand-crafted Mel-Frequency Cepstral Coefficient (MFCC) features [5]. However, this system requires training imitations for the sound concept to be recognized and retrieved. It cannot retrieve sounds that do not have training imitations, which limits its use cases.

In this paper, we propose a novel unsupervised system called IMISOUND for sound retrieval by vocal imitation. We adopt the SAE-based automatic feature learning module proposed in [3] to extract features for both the vocal imitation (query) and sounds in the library (candidates). We calculate the similarity between the query and each candidate using the Kullback-Leibler (K-L) divergence [6] and the Dynamic Time Wrapping (DTW) distance [7] between their feature representations. The most similar candidates are returned as the retrieved sounds for the query. Compared to [3], IMISOUND does not require any training imitations on sounds to be retrieved, nevertheless, experiments show that it achieves a comparable retrieval performance. In addition, IMISOUND significantly outperforms a baseline system that uses MFCC features for similarity calculation. This again validates the advantage of using SAE-learned feature

representations for vocal imitations over hand-crafted features.

2. RELATED WORK

Retrieving sounds by vocal imitation is essentially one instance of Query by Example (QbE) [8]. Thanks to its intuitive interaction, QbE has been proposed for different tasks in sound-related applications, such as query-by-humming [9][10], spoken document retrieval [11][12], etc.

Up to date little work has been done regarding sound retrieval by vocal imitation. Roma and Serra [13] proposed a system that allows the user to query sounds on Freesound.org, but no formal evaluation was reported. Blanca el al. [14] built a supervised system using temporal and spectral features and an SVM classifier. It hence cannot retrieve sounds that do not have training imitations. In addition, the hand-crafted features may be difficult to represent the complex acoustic aspects of vocal imitations of a large variety of sounds. Helén and Virtanen [15] designed an audio query system by measuring feature similarities but it again extracts hand-crafted features. We previously proposed a supervised system based on automatic feature learning and an SVM classifier [3]. The automatically learned features have shown to outperform the MFCC features, however, the supervised nature again prevents its usage in retrieving sounds that do not have training imitations.

3. THE IMISOUND SYSTEM

Figure 1 shows the structure of the IMISOUND system. For the first two modules, we use the same design described in our previous paper [3]. Given a vocal imitation query, we first convert it into a constant-Q spectrogram and segment it into short overlapping patches. We then use a pre-trained Stacked Auto-Encoder (SAE) with two hidden layers to extract features in each patch. To achieve unsupervised sound retrieval, we propose to adopt the K-L divergence and DTW distance to calculate the feature distance between the imitation query and each sound candidate. Finally, we rank candidates according to their distances and return closest ones as the retrieved sounds.

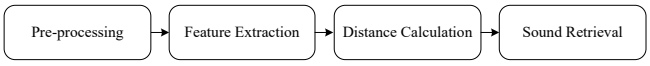

Figure 1. The proposed IMISOUND system.

(1) *Pre-processing*: Taking a vocal imitation query, we first downsample it to 16 kHz. Then a 6-octave Constant-Q transform [16] (12 elements per octave and hop size of 26.25 ms) is employed to convert the waveform into a logarithmic-frequency spectrogram for the accordance with human hearing perception and dimensionality reduction. Then the spectrogram is segmented into overlapping patches. We set the length of each patch to 525 ms and the hop size to 262.5 ms. This length covers the smallest phonic unit carrying semantic meanings [17]. Then the patches are converted into vectors with 1,440 dimensions for further processing.

(2) *Automatic Feature Learning*: In our previous work [3], we demonstrated that features extracted by SAE significantly outperforms hand-crafted features like MFCC's in a supervised vocal imitation recognition setting. Here for the unsupervised setting, we employ the same two-hidden-layer SAE (with non-tied weights) to extract features for both the vocal imitation query and the sound candidates in the library. The input layer has 1,440 neurons, each for one dimension of the vector. The first and second hidden layer contains 500 and 100 neurons, respectively. After passing through the SAE, each patch is finally represented by a 100-d vector. To train the SAE, we use vocal imitations of half of the sound concepts in the VocalSketch Data Set v1.0.4 [18]. These imitations and sound concepts are not used in evaluating the retrieval performance of the system.

(3) *Distance Calculation*: After the SAE feature extraction, each imitation query and sound candidate is represented by a sequence of 100-d vectors. The length of the sequence varies depending on the length of the file. We then measure the distance between the query and each sound candidate.

If the imitation query and the sound candidate are similar, their feature sequences tend to resemble each other in terms of both the probability distribution and temporal evolution. Therefore we define the distance between the two feature sequences using both their K-L divergence (modeling feature distributions) and their DTW distance (modeling temporal evolution).

For K-L divergence, we ignore the time information and view each feature sequence as a bag of feature vectors. Given the relatively large dimensionality (100) and small number of vectors (about 15) of each sequence, we further assume independence between different dimensions, and calculate the symmetric K-L divergence in the *i*-th dimension as

$$D_{K\text{-}L_{(i)}}(P \| Q) = \frac{1}{2}(D_{kl}(P \| Q) + D_{kl}(Q \| P)) = \frac{1}{2}(\sum_j P(j) \ln \frac{P(j)}{Q(j)} + \sum_j Q(j) \ln \frac{Q(j)}{P(j)}) \quad , \quad (1)$$

where P and Q represents the distribution along one dimension of the query and the sound candidate respectively, and j indexes the histogram bins within the i-th dimension.

Figure 2 illustrates the K-L divergence calculation between features from a vocal imitation and a sound candidate. Each imitation and sound candidate is represented by a sequence of 100-d feature vectors. The length of a sequence is the number of patches in each file. For each dimension across all patches within one file, features obey a certain 1-d probability distribution (e.g., distribution P_1 in *Imitation*, which can be approximated by its histogram). K-L divergence between the imitation and the sound candidate is then calculated in each dimension (e.g., $K\text{-}L_1$ calculated by P_1 and Q_1), and all dimensions are then summed together to obtain the overall symmetric K-L divergence $D_{K\text{-}L}$.

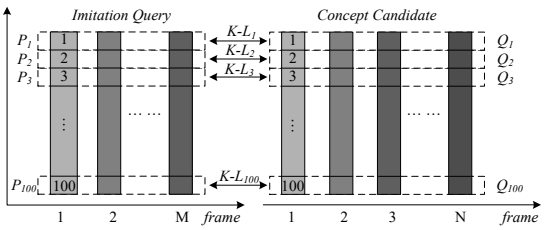

Figure 2. K-L divergence calculation.

K-L divergence is good at modeling mismatches in dynamic ranges between the imitation query and the sound candidate along each dimension, however, it loses temporal information, which can be very important in describing the similarity between sounds. In addition, the independence assumption misses the covariance between different dimensions. Therefore, we complementarily calculate Dynamic Time Warping (DTW) distance between the imitation query and the sound candidate, considering how the 100 dimensions evolve collectively over time. It better models timbre and pitch evolution between vocal imitations and sound candidates. To perform DTW on the two feature sequences, we use cosine distance for the local cost measure [19], which ignores the absolute energy difference. We align the first vectors and the last vectors of the two sequences, and find the warping path that gives the lowest overall cost. The cost is the DTW distance we want, denoted by D_{DTW}.

We combine $D_{K\text{-}L}$ and D_{DTW} in an L-1 space, i.e., summing them as the final distance. To make sure they are of the same scale, we normalize them by their maximal values before the summation. The final distance is thus calculated as

$$D = \frac{D_{K\text{-}L}}{\max(D_{K\text{-}L})} + \frac{D_{DTW}}{\max(D_{DTW})}. \quad (2)$$

Figure 3 shows an example of the distance calculation between a vocal imitation query for "marimba hit with a rubber mallet" and 20 sound candidates within the category of acoustic instruments. Most pitched sounds are of the same pitch. We see that the target sound "marimba hit with a rubber mallet" is indeed the closest to the origin (the vocal imitation) in this 2-d space. After listening to the sound candidates, we find some interesting aspects. The closest candidates (e.g., "thaigong", "vibraphone (sustained)", "piano", "woodblock", and "violin"), including the target sound, are all percussive sounds except "vibraphone (bowed)". Their K-L divergences are smaller than other candidates. We argue that this is because percussive sounds have a wider dynamic range than non-percussive sounds in each dimension, and this is captured by the K-L divergence. In addition, the several furthest candidates (e.g., "triangle", "orchestra bells", and "wind gong") have very different frequency distributions in the CQT spectrogram from the vocal imitation, even though they are also percussive. Therefore, their 100-d feature vectors obtained by passing the spectrogram through the SAE are very different from those of the imitation as well. This makes both their K-L divergences and the DTW distances large.

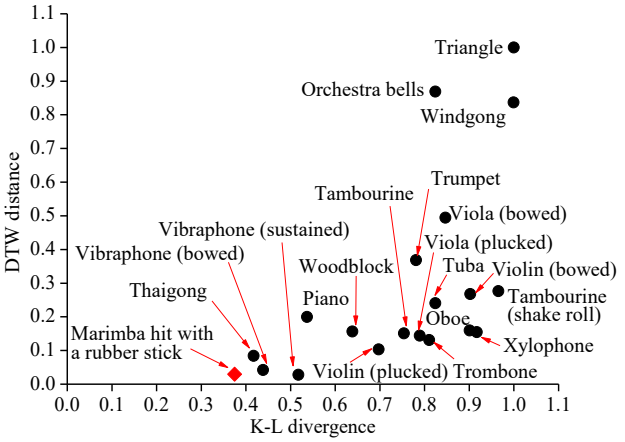

Figure 3. Distance calculation between a vocal imitation and sound candidates of different acoustic instruments.

(4) *Sound retrieval*: Distances between the imitation query and all sound candidates are then ranked, and candidates with the shortest distances are returned to the user.

4. EVALUATIONS

4.1 VocalSketch Data Set

We adopt the VocalSketch Data Set v1.0.4 for experimental evaluation [18]. The dataset includes sound recordings and their vocal imitations in 4 categories: Acoustic Instruments (AI), Commercial Synthesizers (CS), Everyday (ED), and Single Synthesizer (SS), which contains 40, 40, 120, and 40 recordings, respectively. Each recording has 10 vocal imitations from different people.

We use all imitations of half of the recordings in each category, namely 20, 20, 60, and 20 for AI, CS, ED, and SS respectively, to train the Stacked Auto-Encoder (SAE), and then use the other half of the recordings and their imitations to evaluate the retrieval performance. Sound recordings are retrieved within each category instead of across all categories, because in practice people usually know what category they should be searching in for a sound concept in their mind.

4.2 Evaluation Measures

We use *Mean Reciprocal Rank (MRR)* [20] to evaluate the retrieval performance of IMISOUND:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}, \quad (3)$$

where $rank_i$ is the rank of the target sound candidate in the distance ranking list of the i-th vocal imitation query. Q is the total number of imitation queries in the experiment. MRR ranges from 0 to 1 with a higher value for a better retrieval performance. For example, an MRR value of 0.33 would suggest that on average the target sound ranks about the 3rd in the retrieved list of all sounds.

Table 1. MRR comparisons between IMISOUND and baseline systems.

Category (No. of candidates)	SAE-based features			MFCC features			Supervised system [3]
	K-L & DTW	K-L	DTW	K-L & DTW	K-L	DTW	
Acoustic Instruments (20)	**0.392**	0.384	0.371	0.280	0.228	0.248	0.388
Comm. Synthesizers (20)	**0.318**	0.283	0.284	0.212	0.236	0.187	0.326
Everyday (60)	**0.109**	0.103	0.090	0.105	0.090	0.094	0.226
Single Synthesizer (20)	**0.377**	0.361	0.332	0.231	0.230	0.217	0.395

4.3 Baseline Methods

We compare our method to several baseline systems. The first is our previously proposed supervised system [3]. It uses the same SAE to extract features from the vocal imitation query and sound candidates, but trains an SVM classifier from training imitations of the target sound to recognize the sound concept of the imitation query. This is to validate the unsupervised design of IMISOUND. The second system is created by simply replacing the combination of DTW and K-L distances in IMISOUND with either only DTW distance or K-L distance, to evaluate the complementary nature of the two distances. The third system is to replace the SAE feature extraction module of IMISOUND with MFCC feature calculations. A 39-d MFCC feature vector is used, including the original 13 MFCC coefficients, their 13 first-order derivatives, and 13 second-order derivatives, which is commonly applied in audio and speech processing tasks. Different distances are also compared in this MFCC-feature-based system. This baseline is to validate the advantage of automatic feature learning over hand-crafted features for our task.

4.4 Experimental Results

Table 1 shows performance comparisons between the proposed system and various baseline methods. We describe several interesting observations in the following.

First, the MRR values of IMISOUND (first column) are comparable to our previously proposed supervised system in three out of the four categories. In the Acoustic Instruments category, IMISOUND even slightly outperforms the supervised system, achieving 0.392 MRR. This means that on average, the target sound is ranked around the 3rd among the 20 recordings in that category. For the Everyday category, there is a big gap between IMISOUND and the supervised baseline. This may be due to the larger amount and diversity of sounds in this category. Nevertheless, the 0.109 MRR value suggests that the target sound is ranked between the 9th and 10th among the 60 recordings in the category. It is noted that the MRR measure is very conservative in describing the system's performance in practice, since a user does not necessarily know precisely which sound he/she wants to retrieve. Sounds that are similar enough to the query should be all of some interest.

Second, IMISOUND is the best in all compared unsupervised systems (the first 6 columns). The highest MRR values are in bold within the unsupervised settings in the table. This comparison has two aspects: (1) For all three kinds of distance measure, systems with automatic feature learning significantly outperform systems using hand-crafted MFCC features in all categories except Everyday. This confirms the finding in our previous paper [3] that automatically learned features are more suitable to represent vocal imitations, and extends it to the unsupervised retrieval setting. For the Everyday category, the SAE-based features achieve comparable results with the MFCC features, and both are significantly below the supervised performance. This suggests that the advantage of automatic feature learning cannot be shown in this challenging category. (2) For both SAE and MFCC based systems, the MRR's obtained by combining of K-L divergence and DTW distance is better than those using either K-L divergence or DTW distance individually. This is because K-L divergence only measures the distribution difference of features, while DTW distance compares the difference of temporal evolution.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an unsupervised query-by-vocal-imitation system IMISOUND to retrieve sounds from a sound library. A two-hidden-layer Stacked Auto-Encoder (SAE) is adopted to extract features from the vocal imitation and sound candidates. Then feature similarity is calculated by the K-L divergence and DTW distance between the vocal imitation and each sound candidate. Experiments show that IMISOUND achieves a comparable retrieval performance to a previously proposed supervised system. Experiments also show that the SAE-based features outperforms MFCC features, validating automatic feature learning in the representation of vocal imitations in the unsupervised setting. For future work, we would like to conduct human subject studies to evaluate the system's performance in large sound libraries. We also would like to adopt more advanced deep neural networks such as Recurrent Neural Networks (RNN) to model the temporal evolution of vocal imitations.

6. REFERENCES

[1] Guillaume Lemaitre and Davide Rocchesso, “On the effectiveness of vocal imitations and verbal descriptions of sounds,” *The journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862-873, 2014.

[2] Shiva Sundaram and Shrikanth Narayanan, "Classification of Sound Clips by Two Schemes: Using Onomatopoeia and Semantic Labels." in *Proc*. *IEEE International Conference on Multimedia and Expo*, pp. 1341-1344, 2008.

[3] Yichi Zhang and Zhiyao Duan, “Retrieving sounds by vocal imitation recognition,” *in Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, 2015.

[4] Geoffrey E. Hinton and Ruslan R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504-507, 2006.

[5] Steven B. Davis and Paul Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[6] Solomon Kullback and Richard A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, pp. 79-86, 1951.

[7] Stan Salvador and Philip Chan, “FastDTW: Towards accurate dynamic time warping in linear time and space,” *in Proc. KDD workshop on mining temporal and sequential data*, 2004.

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

[9] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith, “Query by humming: musical information retrieval in an audio database,” *in Proc. 3rd ACM International Conference on Multimedia (MULTIMEDIA)*, New York, pp. 231-236, 1995.

[10] Qiang Wang, Zhiyuan Guo, Gang Liu, Chungguang Li, and Jun Guo. “Local Alignment for query by humming,” *in Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3711-3715, 2013.

[11] Shih-Hsiang Lin, Ea-Ee Jan, and Berlin Chen. “Handling verbose queries for spoken document retrieval,” *in Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5552-5555, 2011.

[12] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng. “A lattice-based approach to query-by-example spoken document retrieval,” *in Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, New York, pp. 363-370, 2008.

[13] Gerard Roma and Xavier Serra, “Querying Freesound with a microphone,” *in Proc. 1st Web Audio Conference (WAC)*, 2015.

[14] David S. Blancas and Jordi Janer, “Sound Retrieval from Voice Imitation Queries in Collaborative Databases,” in *Proc. AES 53rd International Conference on Semantic Audio*, pp.1-6, 2014.

[15] Marko Helén and Tuomas Virtanen. “Audio query by example using similarity measures between probability density functions of features,” *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1-12, 2010.

[16] Christian Schörkhuber and Anssi Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound and Music Computing Conference*, pp. 3-64, 2010.

[17] Emily M. Mugler, James L. Patton, Robert D. Flint, Zachary A. Wright, Stephan U. Schuele, Joshua Rosenow, Jerry J. Shih, Dean J. Krusienski, and Marc W. Slutzky, “Direct classification of all American English phonemes using signals from functional speech motor cortex,” *Journal of Neural Engineering*, vol. 11, no. 3, pp. 1-8, 2014.

[18] Mark Cartwright and Bryan Pardo, “VocalSketch: Vocally imitating audio concepts,” *in Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, 2015.

[19] Robert J. Turetsky and Daniel P.W. Ellis, “Ground-truth transcriptions of real music from force-aligned MIDI syntheses,” *in Proc. 4th International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, pp. 135-141, 2003.

[20] Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan, “Evalutating web-based question answering systems,” *in Proc. 3rd International Conference on Language Resources and Evaluation (LREC)*, 48109, 2002.