Audio Retrieval By Voice Imitation

Samah Khawaled , Mohamad Khateeb, Hadas Benisty
Signal and Image Processing Laboratory (SIPL) ,
Andrew and Erna Viterbi Faculty of Electrical Engineering,
Technion – Israel Institute of Technology
{ssamahkh, sm7emad}@campus.technion.ac.il, hadas.benisty@gmail.com

***Abstract*—Existing sound retrieval systems are mostly based on a textual query. Using text to describe a sound signal is not intuitive and is often inaccurate due to subjective impression of the user; different people may use different words to describe the same sound which makes theses system complex to design and unintuitive to use. Vocal imitation, however, is the most natural human way to describe a sound. In this paper we consider a newly rising approach for sound retrieval based on vocal imitations, where the user records himself imitating the desired sound, and the system retrieves a ranked list of the most similar sounds in the dataset. In this work we represent sound signals using histograms, obtained with respect to a Gaussian Mixture Model (GMM), representing the spectral domain. This recently proposed approach was successfully applied for word representation in a keyword spotting task. Having a fixed length representation for vocal imitation signals allows us to train a robust classifier using support vector machine (SVM). Given a test imitation signal, we apply the classifier and use the output score to rank the retrieved signals, based on a majority vote. Our simulation results show that the proposed system yields a more accurate ranking compared with other existing solutions.**

Keywords – Audio Retrieval, Classification, GMM, SVM.

I. INTRODUCTION

With the rapid evolution of technology, there has been a growing interest in voice retrieval systems. These systems have become useful in many audio applications, specifically those involving human interactions. Standard systems based on textual quarries are not effective since they often require a detailed (and sometimes subjective) description of the voice content which makes the process complex, unintuitive and time consuming. In addition, due to the subjective nature of describing voice signals by text, the system may retrieve irrelevant output signals [1]. A Query-by-Example (QBE) sound retrieval systems [2] overcomes these difficulties of text based systems. Given an input signal, the system searches for the most similar signal in the library database. QBE techniques are frequently used in real applications thanks to their efficiency and effectiveness, Microsoft, for example, implemented the "Visual Query by Example" system that improved the usability of searching images [3].

In this paper we propose an audio retrieval by voice imitation system based on the QBE approach. Given a test imitation signal, the main task is to retrieve a ranked list of the top-k sounds that are most similar to the input signal. The advantage of this system is that it allows the user to describe the desired audio sound in the most intuitive manner – by imitation – rather than describing it using text. Moreover, communicating between humans with different languages mostly depends on mimicking by vocals or hand movements. Therefore, when a human asked to describe a specific sound concept, the most trivial response will be imitating it.

Despite the advantages of the retrieval by vocal imitation mentioned above, designing the system still presents challenges such as extracting appropriate features from the received input imitations: the imitation sound input could be noisy, imitation of the same sound could differ from human to human because of the different intonations and accents, languages and gender. Therefore, a major challenge for such system is to effectively represent the imitated signals and to capture similarities and dissimilarities, such that the correct sound would be retrieved while irrelevant differences (such as noise, accent, gender etc.) would be discarded.

In [8], an audio retrieval by vocal imitation system was proposed. The proposed system adapts the automatic feature learning approach [11], based on Neural Networks (NN), to fit the training data in an unsupervised way. In this system, a time-frequency representation, the Constant-Q Transform (CQT) [12] is used as an input to the NN, instead of using the audio waveforms. Training of the NN involves a heavy computational load and requires fast processors and large memory resources, as well as a large data set for training.

In this paper, we propose a more light-weighted system, not requiring heavy computations. Inspired by a recently proposed approach for Keyword Spotting (KWS) [6], we represent the imitation signals using histograms. The histograms are extracted with respect to a Gaussian Mixture Model (GMM), trained to model the spectral domain. This approach proved to be efficient, in terms of accuracy in detecting the keyword, even when the training set is small and or under noisy conditions. As a first stage in training the proposed system, we evaluate the GMM parameters using Mel Frequency Cepstral Coefficients (MFCC's) [4] as features. Given the MFCC's related to an imitation signal, a histogram is extracted based on the posterior probabilities with respect to the Gaussian components. We feed the histograms along with their labels (the actual sound that was imitated) to a Support Vector Machine (SVM) classifier using the standard LIBSVM package [7]. Given a test imitation signal we apply the SVM

978-1-5386-6378-3/18/$31.00 ©2018 IEEE

classifier onto its histogram representation to produce a probabilistic classification and obtain the output ranking.

The remainder of this paper is organized as follows: In Section 0, we provide a short description of the approach proposed in [8], which is our main benchmark. Section III, addresses the general setup of our sound retrieval system by vocal imitation and the structure of the dataset we used. Section IV presents the proposed retrieval system and its implementation. Section V addresses the evaluation measures used to evaluate the system performance. Finally, in section VI we present our experimental results, evaluating our system performance and comparing it to the benchmark system proposed in [8].

II. Related work

Yichi et al. [8] proposed a supervised system based on automatic feature learning via neural network (NN). The features were learned using a Stacked Auto Encoder (SAE) with two hidden layers. The input of the SAE was a 6-octave CQT – a time-frequency representation of the imitation signals. The learnt features were used to train an SVM multi-class classifier. Training of the SAE requires heavy computational resources, both in terms of processing and in terms of memory. These issues were taken into consideration in our work; our GMM based system demands far less complexity and very small memory space. As aforementioned, we consider the supervised system proposed in [8] as a benchmark, to allow performance comparison. In [8], the authors compares their system performance with the baseline system, proposed in [13]. This system differs from the SAE-based system only in the feature extraction, where the authors used the Timbre Toolbox [14] for extraction 472-d feature vector.

III. Problem Setup

To train and evaluate our proposed system we used the VocalSketch Data Set v1.0.4 [5], which includes four categories of different types of real life sounds: Acoustic instruments, commercial synthesizers, every-day and single-synthesizer. For each category there are two types of imitations, the first, noted "excluded" consists recordings of users asked to imitate various sounds without hearing it before. The second type, noted "included", consists recordings where users listened the sounds and then imitated them. Yichi et al. [8] used this data set to evaluate their system as well as their benchmark system. In order to compare their system performance with another system that used the same database. To allow a head to head comparison, we evaluated the performance of our proposed approach using the very same experimental setup.

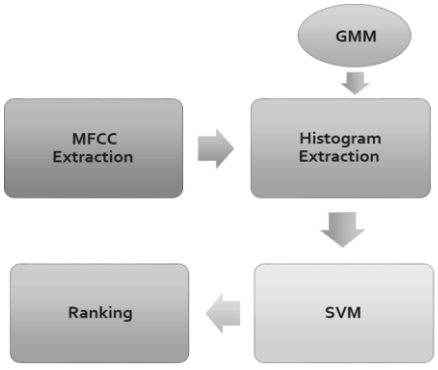

Fig. 1: Block diagram of our proposed system in testing mode, where the test imitation is the input signal

IV. Proposed System

In this section we describe in detail the building blocks of our proposed approach for vocal imitation retrieval, as presented in Fig. 1.

MFCC extraction: We extract 39 MFCC to obtain a spectral-temporal representation of the imitation signals. We use a 25msec frame-size and a 10msec frameshift, taking delta and double-delta features and 3 additional energy coefficients as described in [10], ending up with 39 features, for each frame.

GMM Training: GMM parameters, i.e., weights, mean vectors and covariance matrices are estimated using the Expectation Maximization algorithm [9]. We use MFCC features extracted from the "excluded" database to train the GMM, which provides a sufficiently large dataset for this purpose, without involving data samples for testing and cross-validation of our classifier.

Histogram extraction: We perform the following procedure (also explained in [6]) to extract a histogram representation for each imitation signal. Firstly, we obtain 39 MFCC features $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_w}\right)$, where $t = 1, \ldots, T_w$, is the frame index. Secondly, we evaluate the posterior probabilities, $z_t(m)$ of each MFCC vector, $\mathbf{x}_i$, with respect to the GMM:

$$z_t(m) = P(m \mid x_t) \tag{1}$$

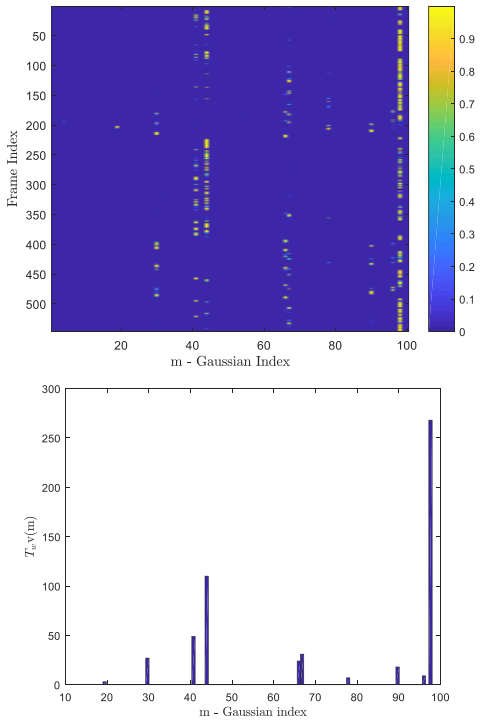

Fig. 2 Histogram extraction: top - posterior probability matrix evaluated using eqn. (1); bottom - extracted histogram using eqn. (2-3).

where $m = 1,\ldots,M$ is the Gaussian component (we used $M = 100$). For each time frame we find the dominant Gaussian component which has maximal probability by setting it to 1 and the rest to be zero, to obtain the following indicator:

$$u_t(m) = \begin{cases} 1 & m = \text{argmax } z_t(m) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Thirdly, we evaluate the histogram representation by averaging the posterior probabilities through time. Formally, the histogram representation of the imitation signal, noted by **v** is evaluated by

$$v_m = \frac{1}{T_w}\sum_{t=1}^{T_w} u_t(m) \quad (3)$$

where v_m is an element in the vector v. Altogether, the histogram representation, as evaluated by eqn. (1)-(3), counts the fraction of times a certain Gaussian component led to the highest probability, as visually demonstrated in Fig. 2.

SVM Classification: As described above, the overall task is to rank the relevance of the given imitation signal to each of the sound classes in the data set. We use the sound classes of each imitation given in the data set as labels, and train a multi-class SVM classifier using the standard LIBSVM package [7]. We use a Radial Basis Function (RBF) kernel, where parameters are set by a 5-fold cross-validation process. We used an all-pairs configuration, meaning that we train a set of binary classifiers, one per each pair-wise combination of sound classes. The output score of these binary classifiers is used to evaluate the output ranked list, as described in the next section.

Ranking: Let N be the number of the sound classes. The amount of possible pairs of classes is $N(N-1)/2$. Therefore, given a test histogram, the output of the all-pairs classifier is a $N(N-1)/2$ vector of scores, each indicating the probability that the histogram relates to one of the examined sound class, or to the other. We evaluate the majority vote by counting the amount of 'battles' each class won, i.e., it lead to the higher score. We sort the sound classes according to their votes from high to low which finally comprises the final output of our proposed system.

V. EVALUATION MEASURES

We use two standard measures to evaluate the performance of our proposed system.

1. Classification accuracy: defined as the percentage of correctly classified imitations among all imitations in the test set. We calculated the accuracy for each one of the four categories of the dataset. The classification accuracy measures performance of the system, considering only the first element of the ranked list.

2. Mean Reciprocal Rank (MRR): note $rank_i$ as the ranking of the sound class of the given test imitation and Q as the overall size of the testing set. The MRR is defined as [1],

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

A successful retrieval system would rank the correct sound class as one of the first items in the list, so the *MRR* value would be closer to 1.

VI. RESULTS

In this section we present the overall performance of our system, evaluated on the VocalSketch dataset. As mentioned above, we used the same data set as [8] to allow a head to head comparison.

TABLE1 presents the performance (both accuracy and MRR measures) for the baseline system [13], the proposed supervised system in [8], and our proposed system, for the four categories. Our proposed solution outperforms the baseline system and the SAE based system proposed in [8] in two categories (marked in boldface) and comes in the second place in the other two categories.

Moreover, it is worth mentioning that we performed the same process (histograms calculation, classification and ranking) for the same imitation examples but with another type of extracted features – CQT. This experiment didn't lead to a good performance compared with the presented results. The classification results of the CQT, accuracy and MRR, are not as high as the proposed, baseline and SAE based system results.

TABLE1 RESULTS AND COMPARISON WITH EXISTING SYSTEMS PEFORMANCE

	Baseline system		**System from [8]**		**Our system**	
Category	***Accuracy (Baseline)***	***MRR (Baseline)***	***Accuracy*** [8]	***MRR*** [8]	***Accuracy (Proposed)***	***MRR (Proposed)***
Acoustic Instruments	27.00%	0.5114	35.5%	**0.5437**	**41.94%**	0.515
Commercial Synthesizer	**29.00%**	**0.4547**	23.50%	0.3881	27.21%	0.3431
Everyday	26.33%	0.4168	27.50%	0.4197	**39.7%**	**0.47**
Single Synthesizer	30.50%	0.4832	43.00%	**0.5822**	**52%**	**0.581**

VII. CONCLUSION

In this paper, we proposed an alternative method which adapts the approach of sound retrieval by vocal imitations. The proposed retrieval system by voice imitation system is a supervised system that views the retrieval task as a classification problem, where SVM is used for this task. The SVM classifier is trained on the histogram calculated from the imitations signals. Histograms are extracted using a pre-trained GMM that models the MFCC vectors extracted from the sounds.

Our GMM-based system is simpler and more easily trained in terms of computational complexity and memory requirements than the SAE based system. Still, according to our simulations, it came in best in two of the four categories, and second (almost comparable) in the other two.

ACKNOWLEDGMENT

The authors would like to thank the Signal and Image Processing Lab (SIPL) staff for their support. We also want to thank them for choosing this project as distinguished project.

REFERENCES

[1] Zhang, Y., & Duan, Z. (2016). IMISOUND: An unsupervised system for sound query by vocal imitation. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016–May, 2269–2273. https://doi.org/10.1109/ICASSP.2016.7472081

[2] M.M.Zloof, “Query-by-Example:A Data Base Language,” IBM Systems J., vol. 16, no. 4, pp. 324–343 (1977 Dec.). http://dx.doi.org/10.1147/sj.164.0324

[3] Zha, Z.-J., Yang, L., Mei, T., Wang, M., & Wang, Z. (2009). Visual query suggestion. Proceedings of the Seventeen ACM International Conference on Multimedia - MM ’09, 15. https://doi.org/10.1145/1631272.1631278.

[4] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” Acoustics, Speech and Signal Processing, IEEE Transaction on, vol. 28, no. 4, pp. 357–366 (1980 Aug.). http://dx.doi.org/10.1109/TASSP.1980.1163420

[5] M. Cartwright and B. Pardo, “VocalSketch: Vocally Imitating Audio Concepts,” Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, South Korea, 2015), pp. 43–46. http://dx.doi.org/10.1145/2702123.2702387

[6] Benisty, H., Katz, I., Crammer, K., & Malah, D. (2018). Discriminative Keyword Spotting for limited-data applications. Speech Communication, 99(February), 1–11. https://doi.org/10.1016/j.specom.2018.02.003

[7] C. C. C. a. C. J. Lin, "“LIBSVM: A Library for Support Vector Machines",ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, pp. 1–27 (2011 Apr.).

[8] Zhang, Y., Duan, Z., & Member, A. E. S. (2016). Supervised and Unsupervised Sound Retrieval, 64(7), 1–11.

[9] Figueiredo, M.A.T.; Jain, A.K. (March 2002). "Unsupervised Learning of Finite Mixture Models". IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (3): 381–396. doi:10.1109/34.990138.

[10] S. Furui (1986), "Speaker-independent isolated word recognition based on emphasized spectral dynamics".

[11] G. E. Hinton, S. Osindero, and Y. W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554 (2006 Jul.). http://dx.doi.org/10.1162/neco.2006.18.7.1527

[12] C. Schorkhuber and A. Klapuri, “Constant-Q Transform Toolbox for Music Processing,” Proc. the 7th Sound.

[13] D. S. Blancas and J. Janer, “Sound Retrieval from Voice Imitation Queries in Collaborative Databases,” presented at the AES 53rd International Conference: Sematic Audio (2014 Jan.), conference paper P2-8.

[14] G. Peeters, B. L. Giordano, P. Susini et al., “The Timbre Toolbox: Extracting Audio Descriptors from MusicalSignals,” *J Acous. Soc. Amer.*, vol. 130, no. 5, pp. 2902–2916 (2011 Nov.). http://dx.doi.org/10.1121/1.3642604