ACOUSTIC STOPWORDS FOR UNSTRUCTURED AUDIO INFORMATION RETRIEVAL

Samuel Kim, Shiva Sundaram†, Panayiotis Georgiou, Shrikanth Narayanan

Signal Anlaysis and Interpretation Lab. (SAIL)
University of Southern California, Los Angeles, USA.

†Deutsche Telekom Laboratories
Quality and Usability Lab, TU-Berlin, Berlin, Germany.

kimsamue@usc.edu

ABSTRACT

The notion of *acoustic stopwords* is proposed to improve the performance of generic unstructured audio information retrieval systems. The rationale behind this is based on the assumption that not all portions of a generic audio signal contribute toward deriving specific descriptive categories (semantic words and onomatopoeias in this work). Detecting the non-salient regions in the audio can hence lead to more robust mapping of signal to categorical descriptions. Using the latent perceptual indexing (LPI) based framework, we propose to remove the proposed acoustic stopwords from the extracted acoustic features, which may include little information on descriptive categories. The acoustic stopwords are selected based on data-driven frequency-related measurements such as *document frequency* (DF) and *inverse document frequency* (IDF). The empirical results with BBC sound effect library show that removing acoustic stopwords based on the IDF measurement improves the audio classification performance especially for onomatopoeic labels.

Index Terms — unstructured audio, audio information retrieval, acoustic stopwords

1. INTRODUCTION

Detecting salient regions in processing multimedia data is often critical in various aspects, such as highlight extraction [1] and feature frame selection [2]. Particularly, in this work, our focus is on extracting salient regions in generic unstructured audio signals with respect to the descriptive categories in an audio information retrieval framework.

One possible approach is to extract prominent segments that attract users' attention. Kalinli *et al.* proposed an auditory attention model to extract gist features from audio signals [3]. They used 2-D Gabor filters motivated by biological observations and presented promising results both in speech recognition [3] and unstructured audio scene classification [4].

In this work, we use a data-driven method which neither require prior knowledge nor elaborate modeling of the human auditory system. We process the given data in a unsupervised way for extracting salient regions for the purpose of audio information retrieval. Here we use the latent perceptual indexing (LPI) based audio information retrieval system proposed in our previous work [5]. The LPI method was inspired by latent semantic analysis (LSA) which was originally devised for text processing applications [6]. By drawing analogies between text and sound, we have proposed various notions such as *acoustic words* and *latent acoustic topics* to facilitate text-like audio signal processing. The empirical results with BBC sound library has shown promising results in classifying semantic labels and onomatopoeias.

Here, we continue our work toward selecting salient regions through the notion of acoustic stopwords, signal portions which may include little or no information with respect to descriptive categories of audio signals. Drawing parallels between text processing and audio signal processing, the notion of acoustic stopwords is inspired by the idea of stopwords in text processing applications [7]. In text applications, stopwords such as articles, conjunctions, and prepositions are often necessary to be removed prior to actual processing since they are considered as non-predictive and non-discriminating words. Likewise in this work, we derive a list of acoustic stopwords and subsequently remove these acoustic stopwords from the processed audio signal.

In the next section, we provide a review of the unstructured audio information retrieval system that utilizes the LPI framework extended by the proposed stopword removal process. The detailed description of the proposed stopword removal process is given in Section 3 followed by experimental results and discussion.

2. UNSTRUCTURED AUDIO INFORMATION RETRIEVAL SYSTEM

We use the LPI-SVM based audio information retrieval system proposed in [5]. It utilizes LPI for feature extraction and support vector machine (SVM) for classification. Fig. 1 shows the basic diagram of the proposed framework (the shaded block represents the new stopword removal process), and the following subsections describe individual steps in detail.

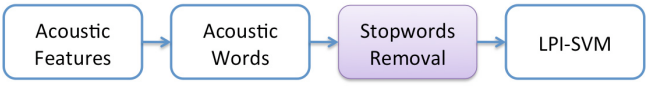


Figure 1: Diagram of the LPI-SVM based unstructured audio information retrieval system along with the proposed stopword removal process.

2.1 Acoustic features

Using frame-based analysis, we calculate mel frequency cepstral coefficients (MFCC) to represent the acoustic properties of the audio signal. MFCCs provide spectral information considering human auditory properties, and have been widely used in many sound related applications, such as speech recognition and audio classification tasks [8]. In this work, we used 20 ms hamming windows with 50% overlap to extract 12-dimensional feature vectors.

2.2 Acoustic words

With a given set of acoustic features, we derived an acoustic dictionary of codewords using the well-known *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) algorithm [9]. Similar ideas to create acoustic words can be also found in [5, 10, 11]. The rationale is to cluster audio segments which have similar acoustic characteristics and to represent them as discrete entities like words in a text document. Once the dictionary is built, the extracted acoustic feature vectors from sound clips can be mapped to acoustic words by choosing the closest word in the dictionary.

2.3 Stopword removal

In text processing applications, various methods have been proposed to remove stopwords [7]. In addition to the standard stopwords, groups of words that have either low or high document frequency are usually considered as stopwords in text mining applications since both high frequency and low frequency groups are considered to carry linguistic content and so that they facilitate the meaning of the text.

By drawing an analogy between text words and acoustic words, we also extend that analogy to acoustic stopwords. We hypothesize that this will preserve more salient regions of the audio signals. Detailed description of choosing acoustic stopwords is provided in Section 3.

2.4 LPI-SVM

After removing the stopwords from the extracted acoustic words, we generate a *word-clip co-occurrence matrix* which describes a histogram of acoustic words in individual audio clips. The word-clip co-occurrence matrix is an $N \times M$ matrix whose element f_{ij} is the frequency of acoustic word w_i in document d_j, where N is the number of words in the dictionary and M is the number of audio clips. Each column is normalized to sum to one so that f_{ij} denotes the probability of word w_i in document d_j which is also known as *term frequency*.

The word-clip co-occurrence can be decomposed into three matrices using singular value decomposition (SVD), i.e.,

$$\mathbf{F} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T \tag{1}$$

where $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ represent a matrix of word vectors, a diagonal matrix with singular values and a matrix with description vectors. This procedure allows LPI to capture the association between a set of descriptions and words in a semantic space. A reduced rank approximation can be obtained by retaining only R greatest singular values, i.e.,

$$\mathbf{F} \simeq \mathbf{U}_{N\times R} \cdot \mathbf{S}_{R\times R} \cdot {\mathbf{V}_{M\times R}}^T \tag{2}$$

where $R < min(N,M)$. The value of R is determined experimentally.

In this work, we use the matrix with description vectors, $\mathbf{V}_{M\times R}$, to represent audio clips. Each audio clip, therefore, will be represented with a single R-dimensional feature vector. The feature vectors will be fed into the SVM framework for training or classification.

3. ACOUSTIC STOPWORDS SELECTION METHOD

Let W be an acoustic dictionary that includes N acoustic words, i.e., $W = \{w_1, w_2, \cdots, w_N\}$ and $\Phi = \{\phi_1, \phi_2, \cdots, \phi_N\}$ be the corresponding measurement for individual words. In this work, we choose document frequency (DF) and inverse document frequency (IDF) as the measurements. Although there are many other methods to evaluate the saliency of words such as entropy and mutual information [6, 7], we only consider these frequency-related measurements in this work.

- **Document Frequency (DF)**: The DF of a word is the number of documents which include the corresponding word, i.e.,
$$DF(w_i) = |\{d : w_i \in d\}| \,. \tag{3}$$
It reflects the contribution of the corresponding word in a database and widely used in text processing applications such as text clustering and text classification.
- **Inverse Document Frequency (IDF)**: We also use the IDF which is a variant of the DF measurement.
$$IDF(w_i) = \log \frac{D - DF(w_i) + 0.5}{DF(w_i) + 0.5}. \tag{4}$$
Constant values are added in both numerator and denominator to avoid division-by-zero or logarithm-of-zero. The values are inversely proportional to the DF values.

Fig. 2 shows the IDF and DF values of individual acoustic words with the given database. The acoustic word indices are sorted in ascending order of IDF values (in other words, descending order of DF values).

Once individual words are evaluated, we sort the acoustic words with respect to the term scores in ascending order. Then, we select a list of stopwords using a threshold. In this

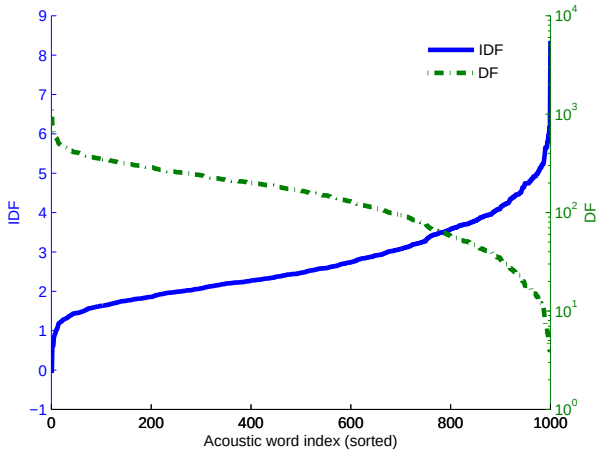


Figure 2: IDF (solid line) and DF (dashed line) values of individual acoustic words. The indices are sorted in ascending order of IDF values (hence, descending order of DF values).

work, we set the threshold based on the number of stopwords rather than specific threshold value. The set of stopwords $\overline{W}$, therefore, can be written as follows.

$$\overline{W} = \{w_i : I(\phi_i) \leq \lfloor \gamma \cdot N \rfloor\}, \quad 0 \leq \gamma \leq 1 \tag{5}$$

where γ represents the desired number of stopwords among the total number of acoustic words in terms of percentage and $I(\cdot)$ represents the index of the sorted list of word measurements. $\lfloor \gamma \cdot N \rfloor$ denotes a maximum integer that does not exceed $\gamma \cdot N$.

We then eliminate the generated acoustic stopwords from the audio clips. In other words, we build a new dictionary that includes only non-stopwords, i.e.,

$$\widehat{W} = \{w_i : w_i \in W \text{ and } w_i \notin \overline{W}\} \, . \tag{6}$$

This will reduce the size of word-clip co-occurrence matrix to $(N - \lfloor \gamma \cdot N \rfloor) \times M$ in the LPI-SVM procedure.

In case of using DF as measurement, low-DF acoustic words will be removed from the vocabulary being considered as stopwords. In case of using IDF, on the other hand, low-IDF (i.e., high-IDF) acoustic words will be removed. Using both measurements enables us to investigate the effects of high and low frequency acoustic words toward the information retrieval task performance.

4. EXPERIMENTAL SETUP

4.1 Database

We have collected 2,140 audio clips from the BBC Sound Effects Library [12] and labeled each file with onomatopoeic labels, semantic labels, and short descriptions. The semantic labels and short descriptions are provided with the database. The semantic labels are given as one of predetermined 21 different categories. They include *transportation*, *military*, *ambience*, *human*, and so on. For deriving the onomatopoeic words, we performed subjective annotation to label individual audio clips. We asked subjects to label the corresponding audio clip choosing from among 22 onomatopoeic descriptions. The audio clips are originally recorded with 44.1kHz (stereo) sampling rate and downsampled to 16 kHz (mono) for acoustic feature extraction.

4.2 Evaluation

We set the initial acoustic dictionary size to 1,000 and the size of stopwords was experimentally controlled by γ (we put a constraint $\gamma \leq 0.2$ for convenience). To evaluate the performance of the proposed framework, we use the *F-measure* which is widely used for evaluating information retrieval systems. The metric considers both *precision* and *recall* and can be written as

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{7}$$

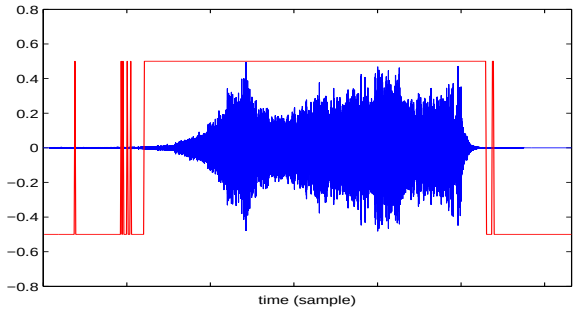
and is evaluated through 10-fold cross validation for individual descriptive categories, i.e., onomatopoeias and semantic words.

5. RESULTS AND DISCUSSION

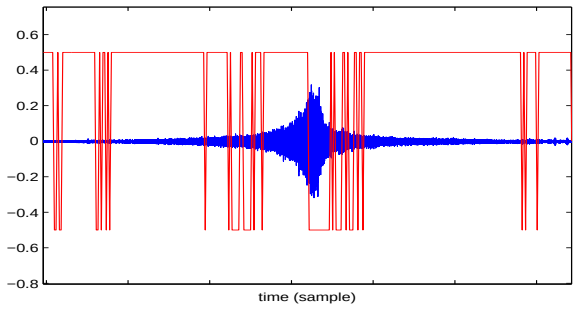
Fig. 3 illustrates examples of output of stopword removal process ($\gamma = 0.2$, low-IDF acoustic words are considered as acoustic stopwords). As shown in the figure, some segments are removed by being considered as stopwords while some segments are retained so that they can be used for audio classification. In Fig. 3(a), it seems that the retained segments represent sound active region in terms of amplitude (although it includes some false alarms). However, note that selecting acoustic stopwords depends on not energy-related measurements but only frequency-related measurements (either DF or IDF). If we consider the example shown in Fig. 3(b), it is clear that the acoustic stopwords are not chosen by energy-related measurements.

Fig. 4 shows the classification results of audio clips in terms of F-measure for both onomatopoeic words (Fig. 4(a)) and semantic labels (Fig. 4(b)) according to the number and the method of choosing acoustic stopwords. Solid lines and dashed lines denote the performance for using IDF and DF measurements respectively, while the x-axis is the γ value which represents portion of acoustic stopwords extracted among the original acoustic words.

In case of classifying onomatopoeic labels, from Fig. 4(a), it is easily observed that the performance is improved as the size of acoustic stopwords increases with the IDF measurement. This supports our hypothesis in previous work [10] that salient region detection method is necessary especially for deriving onomatopoeic labels for audio. The human subjects might describe what they hear based on specific local sound contents rather than global sound contents when they are asked to annotate with onomatopoeias. However, no significant improvement can be seen using the DF metric for stopword identification. This shows that the

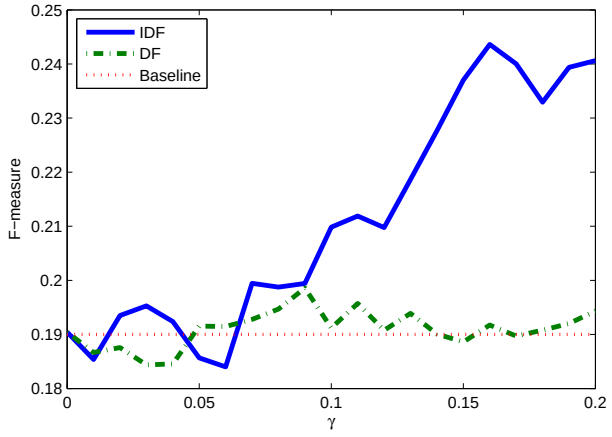


(a)

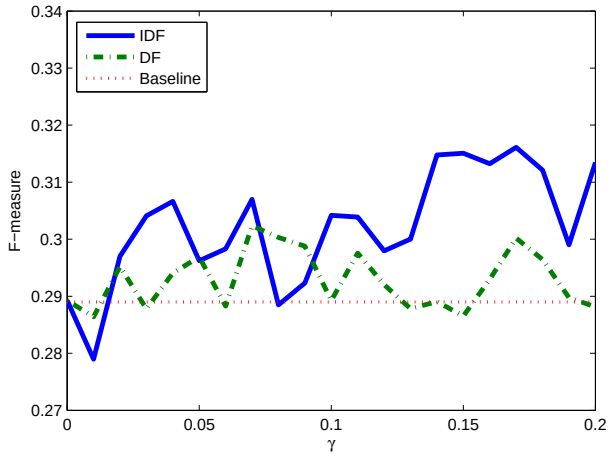


(b)

Figure 3: Examples of stopword removal process results ($\gamma = 0.2$, low-IDF acoustic words are considered as acoustic stopwords).



(a) Onomatopoeic words



(b) Semantic labels

Figure 4: Classification results of audio clips in terms of F-measure for (a) onomatopoeic words and (b) semantic labels according to the threshold γ which determines the number of stopwords.

acoustic words that commonly exist across sound clips can be considered as acoustic stopwords and also that removing these words can improve information retrieval accuracy. As shown in Fig. 4(b), in classifying semantic labels, the improvement by removing acoustic stopwords is not as significant as classifying onomatopoeic labels although the absolute performance in terms of F-measure is higher than the baseline.

6. CONCLUDING REMARKS

In this paper, we proposed a method to remove non-salient regions in audio signals with respect to descriptive categories, i.e., semantic words and onomatopoeias. We investigated data-driven frequency-related measurements such as IDF and DF to generate a list of acoustic words that potentially include acoustic stopwords. Empirical results within the LPI-SVM framework show that removing the stopwords that are generated by the IDF measurement can improve the performance in classifying the onomatopoeias.

In the future, we will apply alternative measurements, such as entropy and mutual information, in addition to frequency-based measurements. We will also investigate bio-inspired salient detection algorithms in this framework such as in [4].

REFERENCES

[1] C.-Y. Chao, H.-C. Shih, and C.-L. Huang, "Semantics-based highlight extraction of soccer program using DBN," vol. 2, March 2005, pp. 1057 – 1060.

[2] C. Jung, M. Kim, and H. G. Kang, "Selecting feature frames for automatic speaker recognition using mutual information," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[3] O. Kalinli and S. S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, Jul. 2009.

[4] O. Kalinli, S. Sundaram, and S. S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Rio de Janeiro, Brazil, Oct. 2009.

[5] S. Sundaram and S. Narayanan, "Classification of sound clips by two schemes: using onomatopoeia and semantic labels," in *IEEE International Conference of Multimedia and Expo*, 2008.

[6] J. R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–80, 2005.

[7] M. Makrehchi and M. S. Kamel, "Automatic extraction of domain-specific stopwords from labeled documents," in *Advances in Information Retrieval*, vol. 4956 of Lecture Notes in Computer Science, 2008, pp. 222–233.

[8] L. R. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[9] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[10] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009.

[11] G. Chechik, E. Ie, M. Rehn, S. Bengio, and R. F. Lyon, "Large-scale content-based audio retrieval from text queries," in *ACM International Conference on Multimedia Information Retrieval (MIR)*, 2008.

[12] The BBC sound effects library - original series. [Online]. Available: http://www.sound-ideas.com