Pooling Robust Shift-Invariant Sparse Representations of Acoustic Signals

Po-Sen Huang[†], Jianchao Yang[†], Mark Hasegawa-Johnson[†], Feng Liang[‡], Thomas S. Huang[†]

[†]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA
[‡]Department of Statistics, University of Illinois at Urbana-Champaign, USA
{huang146, jyang29, jhasegaw, liangf, t-huang1}@illinois.edu

Abstract

In recent years, designing the coding and pooling structures in layered networks has been shown to be a useful method for learning high-level feature representations for visual data. Yet, such learning structures have not been extensively studied for audio signals. In this paper, we investigate the different pooling strategies based on the sparse coding scheme and propose a temporal pyramid pooling method to extract discriminative and shift-invariant feature representations. We demonstrate the superiority of our new feature representation over traditional features on the acoustic event classification task.
Index Terms: sparse coding, pooling, acoustic event classification

1. Introduction

Non-stationary, harmonic quasi-periodicities, and time-relative structures provide useful cues for different types of audio processing applications, e.g., speech recognition, audio coding, audio localization, footstep tracking, and music genre identification [1, 2, 3, 4]. The importance of these acoustic cues has long been recognized, but reliable and efficient extraction of such information is difficult. The difficulty is largely due to the fact that most signal analysis/representation approaches are block-based, i.e., the signal is processed in a series of discrete blocks. Hence, non-stationary periodicities and transients in the signal can be temporally smeared out across blocks. Furthermore, block-based representations are sensitive to temporal shifts, which could create ambiguity for representing a signal. In the case of audio and speech signal recognition, for example, common feature representations, Perceptual Linear Prediction (PLP) and Mel-frequency cepstral coefficients (MFCCs) , are extracted based on temporal shifting blocks. Although standard windowing techniques can alleviate these effects, it is desirable to have signal representations which are inherently insensitive to signal shifts.

A desirable representation should also capture the underlying time-frequency structures in signals with good coding efficiency, so that the structures can be easily extracted for further processing. To efficiently capture and represent signal structures, sparse coding methods have been studied [5, 6, 7]. Sparse coding provides a class of algorithms for finding succinct representations of stimuli by assuming a signal can be represented by a small number of basis functions taken from an overcomplete dictionary. Grosse et al. [5] proposed a shift-invariant sparse coding model to efficiently encode a time-series input signal with the basis functions in all possible shifts and demonstrated its usefulness in audio classification tasks. Smith and Lewicki proposed a sparse and shift-invariant signal representation [1], where a signal is encoded with a set of time-shiftable gammatone kernel functions, approximating the cochlear filters, with associated magnitudes and temporal positions. The representation is called spikegram because the nonzero entries in the sparse basis expansion can be interpreted as neural spikes: action potentials on the auditory nerve. This spikegram can encode the audio signal sparsely and efficiently in both frequency and time.

In image analysis, effective feature learning structures have been proposed, which contain two basic layers: a "coding" layer for efficiently encoding the signals, and a "pooling" layer to extract compact and translation-invariant representations from the coding responses for discriminant analysis or further processing. Some recent works based on pooling over sparse codes have led to state-of-the-art performance on several visual recognition benchmark datasets [8, 6, 7]. Among different spatial pooling methods, *max pooling* [6] has shown great potential for many existing coding schemes. However, the problem of effectively extracting discriminant information from sparse representations has not been extensively studied for audio signals. For audio signals, both biological and psychoacoustic evidence suggest that humans have an intensity pooling mechanism within critical bands, and loudness pooling (cube root of intensity) across bands (e.g., [9]). In this paper, we investigate different pooling strategies for sparse coding schemes and propose a temporal pyramid pooling method for extracting discriminative shift-invariant representations.

The organization of this paper is as follows. Section 2 introduces the feature extraction scheme for audio sig-

This research was supported by U.S. ARL and ARO under grant number W911NF-09-1-0383.

nals. Section 3 presents the experiments and results. Finally, section 4 concludes the paper with discussions on future works.

2. Sparse Time-Relative Coding

In this section, we introduce the model we use for sparsely encoding the audio signals, which was originally proposed by Smith and Lewicki [1, 10] and Manzagol et al. [2].

2.1. Model for Signal Representation

Smith and Lewicki proposed an efficient time-relative coding structure using spikes for audio signals [1, 10]. In this model, a signal $x(t)$ is represented by a set of gammatone kernel functions, $\phi_1, \cdots, \phi_M$, which can be positioned independently and arbitrarily along the time axis. The model can be expressed in a convolutional form:

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} x_i^m \phi_m(t - \tau_i^m) + \epsilon(t), \qquad (1)$$

where x_i^m and τ_i^m are the coefficient of the i^{th} instance of the m^{th} gammatone kernel ϕ_m and its temporal position, respectively. The notation n_m indicates the number of instances for ϕ_m, which does not have to be the same for different kernels. Moreover, the kernels are not restricted in length or form. Here, $\epsilon(t)$ represents the additive noise.

Motivated by both natural sound statistics and biology studies [1], the gammatone filters are given by

$$\phi_m(t) = t^{(l-1)} e^{-2\pi b t} \cos(2\pi f_m t), t > 0, \qquad (2)$$

where l is the filter order, t is time, b is the filter's bandwidth, and f_m is the center frequency of the filter and is distributed on equivalent rectangular band (ERB) scales between 65 Hz and 14k Hz.The gammatone functions are known to approximate the cochlear filters and can be modeled with Slaney's auditory toolbox [11]. The number of gammatone kernels determines the spectral and temporal representation precision. In this paper, we use 64 normalized gammatone kernels ($l = 4$ in Eq. (2)) with frequencies ranging from 20 Hz to the Nyquist frequency, suggested by previous works [1, 2] to achieve a trade-off between representation precision and computational complexity.

2.2. Encoding Algorithm

To encode a signal with the 64 gammatone kernels, *matching pursuit* is employed to achieve a tradeoff between reconstruction error and computational complexity [1]. The algorithm iteratively approximates an input signal $\mathbf{x}$ with successive orthogonal projections onto the chosen basis functions [12], so that $\mathbf{x}$ is a linear combination of few elementary "atoms" from the set of gammatone kernels $D = \{\phi_m\}$. It works iteratively as follows: First, the signal is cross-correlated with all the basis atoms. Second, the best fitting projection is selected and its corresponding atom, scaling, and placement are stored. Finally, the projection is subtracted from the signal and the procedure continues over the residual.

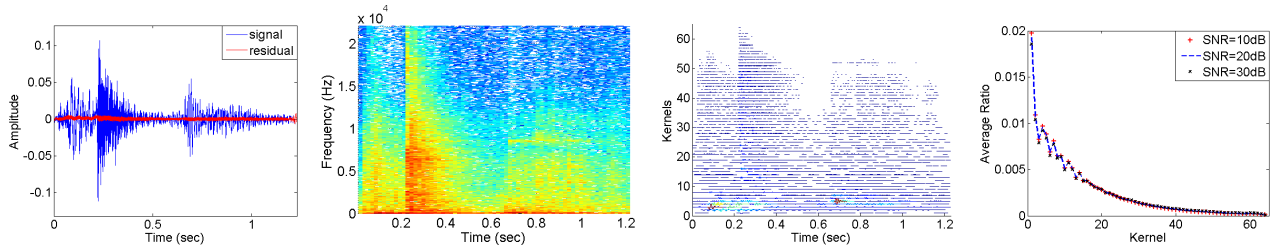
After encoding with matching pursuit, the signal is decomposed into a set of sparse spike codes, called a *spikegram*. The spike pattern usually represents the firing of the action potentials at auditory nerves. One important characteristic of the coding is that kernels are placed at time locations precisely. This characteristic allows precise localization of signals, for example, onset of a music signal [2]. Another property is that spikes are used on a per-need basis. The spikegram is adaptive in the number of spikes for a given encoding ratio. This characteristic contrasts with the uniform distribution of encoding resources in the spectrogram. Figure 1 shows an example of the representation comparison between a spectrogram and a spikegram for a segment of door slamming sound. As shown, spikegram is much sparser than the traditional spectrogram. Suggested in many previous works, representing the signal in the most parsimonious form is advantageous for discriminant analysis, which is also verified in this work.

2.3. Temporal Pyramid Pooling

Acoustic signals represent a temporal phenomena, for which capturing the time varying characteristics is important for recognition. Similar to the translation-invariant spatial pooling for visual data, temporal pooling over audio signals can achieve shift-invariant representations. Let X and Y denote the spikegrams of two acoustic signals. In analogy to the spatial pyramid matching work [8], we can construct a temporal pyramid by dividing the spikegram according to a sequence of regular grids at resolution $0, ..., L$, such that the grid at level ℓ has 2^ℓ regular cells along the temporal dimension of the spikegram. Denote $X_n^\ell = \{x_i^m : b_n^\ell \leq \tau_i^m \leq e_n^\ell\}$, where x_i^m is as defined in Eq. (1), b_n^ℓ and e_n^ℓ are the start and end times of the n^{th} cell on level ℓ. Y_n^ℓ and y_i^m are defined similarly for signal Y. Then we can define an additive kernel between the two sub-spikegrams given a chosen pooling function f:

$$\mathcal{I}(X_n^\ell, Y_n^\ell) = \sum_{m=1}^{M} \mathcal{I}(f(X_n^\ell(m)), f(Y_n^\ell(m))), \qquad (3)$$

where $X_n^\ell(m)$ denotes the set of responses of the m^{th} gammatone filter in the sub-spikegram of X_n^ℓ, $\mathcal{I}$ is the additive kernel (e.g., intersection kernel $\mathcal{I}(x, y) = \min(x, y)$), and the pooling function computes a statistical number from the set of responses $X_n^\ell(m)$. Different pooling functions can be defined, e.g., average pooling computes the average response of $X_n^\ell(m)$. Putting all pieces together, we get the following pyramid temporal



(a) Signal and Residual (b) Spectrogram (c) Spikegram (d) Sparsity

Figure 1: An example of a door slamming event representation: (a) signal and matching pursuit residual in the time domain, (b) spectrogram in the time-frequency domain, (c) spikegram in the time-frequency domain, and (d) average ratio of non-zero coefficients at each gammatone kernel for different SNR cases.

pooling kernel:

$$\kappa(X,Y) = \sum_{\ell=0}^{L} \sum_{n=1}^{2^\ell} w_{\ell n} \mathcal{I}(X_n^\ell, Y_n^\ell), \qquad (4)$$

where the set of weights $\{w_{\ell n}\}$ balance the contributions of different sub-spikegrams. For simplicity, we set equal weights for all subsequent experiments. As the feature pooling is performed in multiple temporal scales, the defined kernel can achieve different levels of shift-invariance. Compared with the deep belief networks [13], our feature learning only consists of one coding layer and one to two pooling layers. While extending the current framework into multiple layers or a deep learning structure is plausible in the future, we mainly investigate the pooling properties over the spikegram in this paper.

3. Experiments

3.1. Dataset and Setup

To evaluate the performance of the proposed pyramid temporal pooling kernel over the sparse spikegram, especially for non-stationary and transient natural signals, we work on the acoustic event dataset collected by Universitat Politecnica de Catalunya [14]. The database contains recordings of eleven target acoustic events (AEs) in a meeting room environment with six T-shaped 4-microphone clusters (Fs=44,100 Hz). The eleven acoustic events recorded include: Knock door/table, Door slam, Steps, Chair moving, Spoon/cup jingle, Paper work–listing and warping, Key jingle, Keyboard typing, Phone ringing/Music, Applause, and Cough. There are approximately 90 instances per event class for the whole dataset, which is divided into six sessions (S01-S06). We use S01-S04 for training and testing, and S05 and S06 as the development set. Among sessions S01-S04, we choose three for training and one for testing each time. All reported results are averaged using this four-fold cross validation.

To make the task more realistic, we add different levels of Gaussian white noise to the recorded clean audio. For baseline comparison, we compare our pooling techniques with traditional Perceptual Linear Prediction (PLP) features, which are extracted from 30 ms Hamming windows with a temporal step of 20 ms. Support vector machines with the intersection kernel [8] is used as the classifier for our feature representations.

3.2. Evaluated Pooling Methods

In order to test the pooling methods, we design and evaluate different pooling strategies for the spikegram coefficients. Besides the well-known pooling methods, including *max pooling*, *average pooling*, *energy pooling*, and *magnitude pooling*, we also introduce two more pooling methods inspired by the Mel-frequency cepstral coefficients (MFCC) feature and the Perceptual Linear Prediction (PLP) feature that are commonly used in speech/music recognition.

In MFCC, after the Mel-frequency filter bank, the action of taking log of the power at each Mel frequency is similar to a *log pooling* for each gammatone kernel. On the other hand, in the PLP feature, the action of taking intensity-loudness power law compression is similar to a *cuberoot pooling* for gammatone kernel responses. In Table 1, we list all the individual pooling methods we tested. Note that *log pooling*, *energy pooling*, *magnitude pooling*, and *cuberoot pooling* are equivalent in terms of the sufficient statistics, but they perform differently for a specifically chosen classifier. Manzagol et al. proposed the average-pooling-like feature from the spikegram [2], which achieves the same accuracy as Mel-frequency cepstral coefficients (MFCC) for the music genre recognition task. In our experiments, we use their algorithm as a baseline for comparing different pooling methods.

3.3. Experimental Results

Table 2 shows the performance comparisons for different pooling methods on the acoustic dataset [14] with SVM, compared with the traditional PLP feature (without pooling) with Hidden Markov Model [15], and segmental PLP feature [16] (without pooling) and average pooling methods [2] with intersection kernel SVM. In each cell of the pooling results, the upper and lower row show the result

Table 2: Classification accuracy with different features under different SNRs. The PLP†, PLP‡, and average§ correspond to the features in [15], [16] and [2] respectively. In each cell of the pooling results with SVM, the upper and lower row correspond to one ($L = 0$) and two ($L = 1$) layers of temporal pyramid pooling from kernel SVM, respectively.

classifier	HMM	SVM						
features	PLP†	PLP‡	average§	magnitude	energy	cuberoot	log	max
10dB	28.05±4.40	33.86±4.09	38.64±1.90	42.51±2.34 42.90±2.03	39.15±2.16 38.30±2.60	$\mathbf{43.40 \pm 2.67}$ 39.67±2.12	41.35±2.05 37.06±1.45	32.20±4.55 33.67±3.62
20dB	51.54±5.21	55.82±4.36	54.47±4.75	$\mathbf{67.93 \pm 5.57}$ 67.25±5.19	66.63± 4.23 64.53±3.84	67.55±4.96 66.31±5.85	64.16±2.64 61.38±2.93	54.30±6.11 57.30±5.50
30dB	77.45±6.96	73.30±4.42	79.04±3.87	83.99±3.23 $\mathbf{85.08 \pm 3.03}$	83.01±3.45 82.57±3.24	84.10±2.60 84.21±3.63	77.82±3.69 77.09±3.19	77.50±3.55 80.49±3.98

Table 1: Feature pooling methods for the spikegram. x_i^m is the i^{th} response coefficient for the m^{th} gammatone kernel in Equation (1).

Pooling Method	Equation
average	$f(X_n^l(m)) = \sum_i \lvert x_i^m \rvert$
max	$f(X_n^l(m)) = \max_i(\lvert x_i^m \rvert)$
log	$f(X_n^l(m)) = \log(\sum_i \lvert x_i^m \rvert^2)$
energy	$f(X_n^l(m)) = \sum_i \lvert x_i^m \rvert^2$
magnitude	$f(X_n^l(m)) = \sqrt{\sum_i \lvert x_i^m \rvert^2}$
cuberoot	$f(X_n^l(m)) = (\sum_i \lvert x_i^m \rvert^2)^{0.33}$

with one ($L = 0$) and two ($L = 1$) layers of pyramid pooling with kernel SVM (intersection kernel), respectively. From the results, we have several interesting observations:

1. Although *magnitude pooling*, *log pooling*, *energy pooling*, and *cuberoot pooling* are scalar transformations of energy pooling, they perform differently due to the limited discriminant power of the chosen classifier. They all perform better than *average pooling*.
2. *Max pooling* does not work well compared with other methods, although it performs well with visual sparse codes [6].
3. The pooling methods outperform the PLP feature uniformly under all SNRs by a remarkable margin (8% to 12% absolute error reduction and 14% to 34% relative), indicating that our pooling feature representation is discriminative and robust to noise.
4. Non-linear amplitude compression, i.e., *magnitude pooling*, *cuberoot pooling* and *log pooling* achieve better performance than *energy pooling*. The results match the findings of non-linear relationship between the intensity of sound and its perceived loudness in psychoacoustic studies [9].

4. Conclusion

In this paper, we proposed different pooling schemes over sparse spikegram coefficients of acoustic signals to extract discriminative and shift-invariant feature representations. Compared with PLP, the new pooled shift-invariant feature achieves error rate reduction of 8% to 12% absolute (14% to 34% relative) on the acoustic event classification task across a range of SNRs. Although we only perform our experiments on one real-world acoustic event dataset, we hope our analysis and observations will inspire more research on designing efficient and effective coding and pooling structures that can be extended to multiple layers in the deep learning framework both for audio and visual signals.

5. References

[1] E. Smith and M. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computing*, vol. 17, no. 1, pp. 19–45, 2005.

[2] P. A. Manzagol, T. Bertin-mahieux, and D. Eck, "On the use of sparse time-relative auditory codes for music," in *ISMIR 2008*, 2008.

[3] P.-S. Huang, M. Hasegawa-Johnson, and T. Damarla, "Exemplar selection methods to distinguish human from animal footsteps," in *HLVD and Light Vehicle Detection Workshop*, 2011.

[4] P.-S. Huang, T. Damarla, and M. Hasegawa-Johnson, "Multi-sensory features for personnel detection at border crossings," in *Fusion*, July 2011, pp. 1–8.

[5] R. Grosse, R. Raina, H. Kwong, and A. Ng, "Shift-invariant sparse coding for audio classification," in *UAI*, 2007.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.

[7] J. Yang, K. Yu, and T. S. Huang, "Supervised translation-invariant sparse coding," in *CVPR*, 2010, pp. 3517–3524.

[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[9] H. Fletcher, "A space-time pattern theory of hearing," *Journal of the Acoustical Society of America*, vol. 1, no. 3, pp. 311–343, 1930.

[10] E. Smith and M. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 800–805, 2006.

[11] M. Slaney, "Auditory toolbox," 1994.

[12] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[13] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computing*, pp. 1527–1554, 2006.

[14] C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J. R. Casas, "Audiovisual event detection towards scene understanding," in *CVPR*, 2009, pp. 81–88.

[15] P.-S. Huang, X. Zhuang, and M. A. Hasegawa-Johnson, "Improving acoustic event detection using generalizable visual features and multi-modality modeling," in *ICASSP*, 2011.

[16] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *ICASSP*, vol. 2, 1999, pp. 585 – 588.