An efficient code for environmental sound classification (L)

Raman Arora

*Department of Electrical and Computer Engineering and Auditory Behavioral Research Laboratory,
University of Wisconsin, Madison, Wisconsin 53706*

Robert A. Lutfi

*Department of Communicative Disorders and Auditory Behavioral Research Laboratory,
University of Wisconsin, Madison, Wisconsin 53706*

(Received 19 January 2009; revised 29 April 2009; accepted 30 April 2009)

An efficient code for classifying environmental sounds is described that exploits a recent significant advance in signal processing known as compressed sensing (CS) [cf. Donoho, D. (2006). *IEEE Trans. Inf. Theory* **52**, 1289–1306]. CS involves a novel approach to sampling in which the salient information in signals is recovered from the projection onto a small set of random basis functions. The advantage of the random basis over traditional Fourier or wavelet representations is that it allows accurate classification at low target-to-interference ratios based on few samples and little or no prior information about signals.

© 2009 Acoustical Society of America. [DOI: 10.1121/1.3139982]

PACS number(s): 43.66.Ba [DOS]

Pages: 7–10

I. INTRODUCTION

Much work has been devoted in recent years to the goal of developing an automated sound recognition system that can accurately and efficiently classify a wide variety of common environmental sounds according to their generating source. The effort is driven, in part, by the desire to understand through computation modeling our own ability to reconstruct from sound an accurate perception of everyday objects and events present in our natural environment (Bregman, 1990; Ellis, 1996; Lutfi, 2008). It is also motivated by the potential for significant benefit arising from the practical application. Some of the most promising applications are now being pursued in the areas of remote surveillance (Cristani *et al.*, 2004, 2007; Cowling and Sitte, 2003; Wang *et al.*, 2008) and multimedia database search (Wang *et al.*, 2006).

The development of an automated environmental sound recognition system poses a number of significant challenges, but first among these is the problem of how best to encode the salient information in signals. The information rates associated with environmental signals are exceedingly high, and not all information in signals will be diagnostic regarding the identity of their source. Preserving a large amount of information in the initial encoding of signals is costly because it requires increased rates of sampling and subsequent computation. At the outset, then, one must decide what information in signals should be preserved and what information can be discarded without significant loss in classification accuracy. The traditional approach to this problem has been to represent signals partially in some basis that shares a common structure with the signals to be encoded, typically a Fourier or wavelet basis. Signals are sampled densely in this basis and all information other than that provided by a small number (M) of significant coefficients is discarded for further processing. This approach preserves the salient information

in signals, but is clearly wasteful in that it requires a good deal of upfront processing of information that ultimately will not be used.

A second and equally important issue pertaining to encoding is the interference produced by ambient noise in the environment. Target signals rarely occur in isolation; rather, they are usually accompanied by other unwanted environmental sounds that compete with the targets. In many applications the target signal-to-noise ratio is also low (e.g., military surveillance applications involving the detection of intentionally concealed targets). In such cases much of the information in targets may be lost in Fourier or wavelet basis since the maximum-valued coefficients are likely to be determined predominantly by the noise. Prior knowledge of the interference can help in these cases to isolate the target from the interference, but such knowledge may be incomplete or entirely lacking in many applications.

The present paper offers a new approach to the classification of environmental sounds that deals effectively with the high information rates associated with potential signals and the interference produced by extraneous sound sources. It does so by exploiting a recent significant advance in the efficient encoding of signals known as compressed sensing (cf. Donoho, 2006). Compressed sensing (CS) is an emergent technology that has found increasing application in the areas of broadband signal monitoring and image reconstruction. It involves a novel approach to sampling in which the salient information in signals is recovered from the projection onto a small set of random basis functions. In the present paper the advantage of CS-based classification over traditional Fourier-based classification is demonstrated by applying both to a representative case of an environmental sound classification task.

II. COMPRESSED SENSING

To appreciate the power of CS and the insight on which it is based, it is instructive to revisit the logic underlying the

traditional approach to signal encoding. This logic says that we should project signals onto a basis that mirrors as closely as possible the underlying structure of the signals. In so doing, we can expect that a small number of basis functions, those having the largest coefficients, will capture most of the salient information in signals. The logic is correct, but it provides no recipe on how to identify the salient functions without first projecting signals onto the complete basis; a rather wasteful exercise. CS avoids this problem by taking an exactly opposite approach to encoding. Instead of matching the basis to the inherent structure of signals, it requires that we project onto a basis that is void of any structure, one for which the basis functions share no feature in common with signals. The one basis having this property for all signals is the random basis, which has noise waveforms as basis functions. The logic may seem counterintuitive; but, consider that, unlike a Fourier or wavelet basis, each function of the random basis is virtually ensured to have some measurable correlation with the signal, positive or negative. This is a direct consequence of the noise basis functions being broadband. The insight of CS is that only a small number of these correlations are required to recover the signal without error—we need not project onto the entire basis. All that is required is that the signals be sparse in some basis; that they be well-represented in some basis by a small number of non-zero coefficients.¹ Sparsity, in fact, is a feature of environmental sounds that underlies the success of popular compression schemes such as mp3. Consider that a vast number of the sounds we encounter in our everyday environment are emitted by objects having relatively simple geometries; bells, bars, plates, and membranes with different supports. Because of their simple geometries these objects have relatively few prominent modes of vibration, and so most of the salient information in their emitted sounds is captured by a few nonzero Fourier coefficients. Other environmental sounds, not sparse in the frequency domain, are nonetheless represented by a small number of nonzero values in time. The sound of footsteps, birds chirping, hands clapping, and many machine sounds are but to name a few. Indeed, with the exception of continuous broadband noise, it is difficult to imagine a natural sound that cannot practically be considered sparse in either frequency or time. In what follows, then, we use the sparse property of environmental sounds to advantage in their classification.

III. METHOD

The 50 environmental sounds used in the simulation (25 targets and 25 interferers) were taken from the high-quality sound effects CDs “Hollywood Leading Edge Sound FX, The General.” They are listed in Table I. These sounds were selected because they have been shown to be easily identified by human listeners in a previous psychophysical study (see Gygi *et al.*, 2004); hence, they were deemed, in this sense, to be “common” environmental sounds. The sounds were also selected so that they would span a broad range of different sound categories (e.g., machine, human, weather, and animal sounds). The original sound recordings ranged in duration from 0.5 to 3.6 s. For the simulation they were normalized in

TABLE I. Waveforms used in the simulation as named in the Hollywood Leading Edge Sound FX CDs.

Targets	Interferers
89_BellChrch.wav	89_CarStart.wav
89_GlassSmash.wav	89_HelicPassby.wav
89_IceDropIntoGlass.wav	89_SprtBowlingStrike.wav
BABYCRY.wav	BASKBALL.wav
BIRD.wav	BUBLES.wav
CATX.wav	CHICOUGH.wav
CLAPSA.wav	CLOCK.wav
COWA.wav	CRASHA.wav
CRICKETS.wav	CYMBALA.wav
DOGX2.wav	DOOROC.wav
DRUMS.wav	FOOTSTP.wav
GUN.wav	HAMMERNG.wav
HARP.wav	HORSERUN.wav
HORSEWNA.wav	LAUGH.wav
MATCH.wav	PEELOUT.wav
PHONE.wav	PINGPA.wav
POURWATR.wav	PlanePropPassby.wav
ROOSTER.wav	SHEEP.wav
SHOVEL.wav	SIREN.wav
SNEEZE.wav	SPLASH.wav
STAPLER.wav	ScissorsHrCut.wav
TENNIS.wav	THUNDER.wav
TOILET.wav	TRAIN.wav
TYPEWRI.wav	WistlePlc.wav
WtrRain.wav	ZIPPERA.wav

duration to 3.6 s by zero padding where necessary. They were also equated in total rms. Finally, because of the significant amount of computation required for the simulation, the sounds were down-sampled from 44.1 to 4 kHz. Each waveform then was effectively low-pass filtered at 2000 Hz and contained a total of $N=3.6 \times 4000/s=14\,400$ samples.

The simulation was conducted as follows: (1) Select at random M Gaussian noise waveforms each of length N to construct an $M \times N$ matrix $\mathbf{R}$ as the random basis to be used in all conditions. (2) Compute and store the $M \times 1$ vector $\mathbf{a}_i$ of random basis coefficients for each of the $i=1 \cdots 25$ targets,

$$\mathbf{a}_i = \mathbf{R}\mathbf{x}_i, \quad (1)$$

where the $N \times 1$ vector $\mathbf{x}_i$ is the i th discrete target waveform. (3) Compute the $M \times 1$ vector $\mathbf{b}_j$ of basis coefficients for each of the $j=1 \cdots 625$ possible combinations of target and interferer,

$$\mathbf{b}_j = \mathbf{R}(\mathbf{x}_i + \alpha \mathbf{y}_k), \quad (2)$$

where the $N \times 1$ vector $\mathbf{y}_k$ is the k th discrete interferer waveform and α determines the target-to-interference ratio. (4) For each target+interferer combination given by j identify the target by the index $\hat{i}$, where $\mathbf{a}_{\hat{i}}$ yields the largest inner-product with $\mathbf{b}_j$,

$$\hat{i} = \arg \max_{i=1 \cdots 25} \langle \mathbf{a}_i, \mathbf{b}_j \rangle. \quad (3)$$

(5) Repeat steps 1–4 for values of M ranging from 1 to 256 and at different target-to-interference ratios ranging from -20 to 20 dB. (6) Repeat steps 1–5 replacing the $M \times N$

FIG. 1. Percent correct identification performance for CS (circles), MAXF (triangles), and RANDF (squares) is plotted as a function of the target-to-interference ratio for different values of M indicated at the top of each panel.

random basis $\mathbf{R}$ with the $N \times N$ Fourier basis and choosing for $\mathbf{a}_i$ and $\mathbf{b}_j$ the M maximum Fourier coefficients in each case (MAXF classifier). (7) Repeat steps 1–4 replacing the $M \times N$ random basis $\mathbf{R}$ with the incomplete $M \times N$ Fourier basis, using the same M randomly-selected rows of the complete Fourier basis in all conditions (RANDF classifier).

IV. RESULTS

Figure 1 shows the performance of the three classifiers as a function of the target-to-interference ratio with the number of coefficients M as parameter. Note first that the poorest performance overall is achieved by RANDF. Only for $M > 32$ and a target-to-interference ratio greater than 0 does RANDF begin to match the performance of CS. This shows that random sampling alone cannot account for the high performance levels achieved by CS. Compared to MAXF the performance of CS is everywhere poorer for $M < 8$, but somewhere between $M=8$ and 16 the performance of CS begins to better that of MAXF as M is increased. The reversal is particularly evident at the lower target-to-interference ratios where the performance of CS continues to improve and is near perfect by $M=128$. The failure of MAXF performance to improve similarly over this range is due to the interference, which at the lower target-to-interference ratios determines the significant Fourier coefficients. This point merely underscores one of the fundamental differences between CS and MAXF described in the Introduction: MAXF can only improve its performance at low target-to-interference ratios by using prior knowledge of the interference to separate target and interference. CS makes no attempt at separation and so requires no such knowledge. The projection onto the random basis ensures that any portion of the target not obscured by the interference will contribute to the classification. Finally, we note that CS is overall considerably more efficient than MAXF. CS performs as well or better than MAXF at all target-to-interference ratios with as few as 16 projections, compared to $N=3.6 \times 4000/s = 14\,400$ projections for MAXF.

V. DISCUSSION

We have shown that, at target-to-interference ratios as low as -20 dB, CS achieves near perfect classification of an arbitrary set of environmental sounds with only 128 projections, and equal or better classification performance than the M maximum Fourier coefficients with as few as 16 projections. The results, of course, are for a highly restricted case in which there is only one exemplar of each target class, and in which targets and distracters can reasonably be considered sparse. A much stronger test of the algorithm would require its application to more realistic tasks in which there are multiple exemplars of each target class, including non-sparse targets and distracters. Notwithstanding, the results encourage speculation as to how CS might be recruited in the effort to model human sound source classification. Traditional approaches based on the extraction of structured features and programed schema for separating sound sources require high information rates and much prior knowledge of signals (Ellis, 1996; Martin, 1999). Yet, they still fall pitifully short of the human capacity for classifying sounds in everyday listening. The present results suggest that the problem may be more manageable than these efforts imply. The information rates, perhaps, need not be as high if compression can occur simultaneously with sampling, and the amount of prior signal knowledge, perhaps, need not be as great if classification can occur frequently without the need to separate sound sources. It remains to be seen whether the attractive features of CS could be incorporated into a computational model that would eventually approach the remarkable performance of the human classifier in more realistic listening situations than considered here. The present results are, at least, encouraging.

ACKNOWLEDGMENT

This research was supported by NIDCD Grant No. 5R01DC006875-04.

¹The reconstruction in this case involves minimization of the L_1 norm, which virtually ensures errorless reconstruction. For a very readable introduction to the theory of compressed sensing the reader is referred to the review article by Baraniuk (2007).

- Baraniuk, R. (2007). "Compressive sensing," IEEE Signal Process. Mag. **24**, 118–121.
- Bregman, A. S. (1990). *Auditory Scene Analysis* (MIT, Cambridge, MA).
- Cowling, M., and Sitte, R. (2003). "Time-frequency environmental sound recognition for autonomous robot surveillance," Proceedings of AMiRE, Brisbane, Australia.
- Cristani, M., Bicego, M., and Murino, V. (2004). "On-line adaptive background modelling for audio surveillance," in Proceedings of the International Conference on Pattern Recognition (ICPR 2004), pp. 399–402.
- Cristani, M., Bicego, M., and Murino, V. (2007). "Audio-visual event recognition in surveillance video sequences," IEEE Trans. Multimedia **9**, 257–267.
- Donoho, D. (2006). "Compressed sensing," IEEE Trans. Inf. Theory **52**, 1289–1306.
- Ellis, D. P. W. (1996). "Prediction-driven computational auditory scene analysis," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
- Gygi, B., Kidd, G. R., and Watson, C. S. (2004). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am. **115**, 1252–1265.

- Lutfi, R. A. (2008). "Human sound source identification," *Auditory Perception of Sound Sources* (Springer, New York), pp. 13–42.
- Martin, K. D. (1999). "Sound-Source Recognition: A Theory and Computational Model," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
- Wang, J. C., Lee, H. P., Wang, J. F., and Lin, C. B. (2008). "Robust environmental sound recognition for home automation," *IEEE Trans. Autom. Sci. Eng.* **5**, 25–31.
- Wang, J. C., Wang, J. F., He, K. W., and Hsu, C. S. (2006). "Environmental sound classification using hybrid SVM/KNN and MPEG-7 audio low-level descriptor," *International Joint Conference on Neural Networks*, Vancouver, BC, Canada.