Parsing the Blooming Buzzing Confusion: Identifying Natural Auditory Scenes

Work supported by the National Institute on Aging and the Veterans Affairs Medical Research.

Brian Gygi, East Bay Institute for Research and Education, Martinez, CA

ABSTRACT

A corpus of 115 naturally-occurring auditory scenes was collected and analyzed, both for acoustic features and semantic properties of the scenes. Relationships were found between several acoustic features and semantic attributes such as the type of scene, Urban/Rural and the presence of music. Identification of a subset of 33 of these scenes was quite good, mean p(c) = 0.71. However, none of the acoustic variables strongly predicted the identification results. This preliminary work indicates some of difficulties involved in studying auditory scenes.

INTRODUCTION

While some research has shown the ability of listeners to identify environmental sounds presented in isolation (e.g., Gygi *et al.*, 2004, Shafiro 2004), little work has investigated how people attend to numerous sounds contained in an auditory 'scene'. Investigations of the statistics of classes of soundscapes have shown regular features (Boersma,1997; De Coensel, *et al.*, 2003) such as a 1/f dependency for a number of descriptors of the soundscapes. This work is the beginning of a program to study the relation between the perception of individual sounds, perception of auditory scenes and the acoustic factors present in both instances.

The Corpus of Scenes

115 naturally-occurring auditory scenes were collected from various sources, such as field recording professionals, sound effects CDs or on-line databases. The files were all 44100 kHz 16-bit, and with three exceptions were all stereo files. They represented 31 different canonical scene types, e.g.:

Street scene Office Indoor Sporting Event Forest Library Bar Market Hospital

Durations ranged from 10 – 423 s with a mean of 80.6 s. Scenes were selected to be familiar, representative and to contain multiple sources if possible. They were normalized for level. For most scene types there were at least two examples, the exceptions being the Household, Library, and Forest Fire scenes.

Qualitative Analysis

The number of difference sources in each scene was tabulated (mean source/scene = 4.92). In addition, the scenes were coded for Inside/Outside, Urban/Rural, and for the presence of voices, music and machinery.

Quantitative analysis

In the manner of Voss & Clarke (1978) and De Coensel (2003), Instantaneous Power and Instantaneous Pitch of the scenes were calculated and the spectra fitted to polynomials. The LT power spectra were similarly fitted. The slopes were all calculated in log-log units.

Acoustic variables included from Gygi *et al.* (2004) were No. of Bursts, Burst/Total Duration as well as moments of the LT Spectrum, Spectral Velocity, Pitch Salience and the autocorrelation matrix Figure 1 at right shows a spectrogram of the Library scene plotted along the Instantaneous Pitch, Instantaneous Power and Bursts (and the events causing the bursts).

In an effort to determine if the acoustics reflect the number and type of sources, one-way ANOVA were performed on the Quantitative Variables vs. Qualitative Variables

ANOVA RESULTS

Variables significant for Scene Type*

Variable	*Max*	*Min*
No. of Bursts	Kitchen	Forest Fire, Rain
LT Spectral Slope	City Park	Forest Fire
LT Spectral SD:	Forest	Farm, Rain
Mean Spec. Vel.:	Hospital	Rain

Variables significant for Urban/Rural

Variable	*Mean Urban*	*Mean Rural*
No. of Bursts	10.82	22.62
LT Spectral SD	0.50	0.53
Inst. Power Slope	-1.47	-1.70

Variables significant for Presence of Machinery

Variable	*w/o Machinery*	*w/Machinery*
LT Spectral SD	0.50	0.53
Inst. Power Slope	-1.60	-1.75

Variables Significant for Presence of Music

Variable	*w/o Music*	*w/Music*
Inst. Pitch Slope	-0.77	-0.89
No. of Bursts	14.39	6.22
LT Spec. Vel. SD *(f)*	435.85	338.40

Variables Significant for Presence of Speech

Variable	*w/o Speech*	*w/Speech*
Mean Pitch Salience	0.201	0.392
Max. Pitch Salience	0.625	0.761

PREDICTING THE NUMBER OF SOURCES

Variables significantly corr. with # of Sources

Variable	*No. of Sources*
Duration	0.396
Inst. Pitch Slope	-0.418
No. of Bursts	0.302
Bursts/Tot. Dur.	0.301

Voss & Clarke (1978) found a 1/f relation for the spectral density of Instantaneous Pitch and Instantaneous Power in both speech and music, and De Coensel (2003) found a similar relationship in rural soundscapes. The distribution of slopes for this corpus is somewhat different, perhaps due the shorter duration of the sound clips in this study:

	Mean	*SD.*
Inst. Pitch	-0.802	0.237
Inst. Power	-1.673	0.298
LT Spectrum	-1.314	0.713

**Household and Library were omitted from this analysis, each having only one example, but Forest Fire was retained since it was perfectly identified and so thought to be representative of the scene type.*

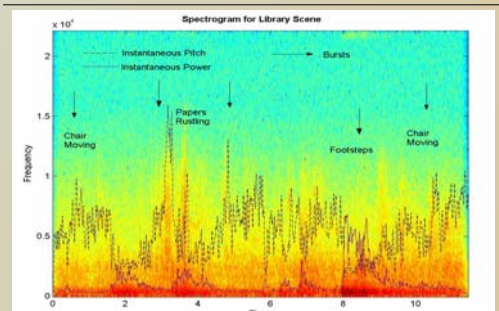

Figure 1. The spectrogram of the Library scene is displayed, plotted along with the Instantaneous Power (purple line) and Instantaneous Pitch (black line) of the scene. The starting points of bursts are also noted, along with the events causing the bursts.

IDENTIFICATION STUDY

Thirty-three of the scenes were selected for an identification study, on the basis of clarity and representativeness. The scenes were edited down to an average duration of 10.42 s and presented binaurally to seven NH listeners at 80 db SPL. Listeners were first asked to make an immediate identification. Then they could play back the scene and were asked to note as many different sounds sources as they could identify. Two judges rated the answers for correctness on a fractional scale and tabulated the number of sources correctly identified (*r* of the judges' ratings = 0.96).

Scene	*p(c)*	*No. of Sources*	*Scene*	*p(c)*	*No.of Sources*
Forest fire	1.00	2.71	Rural Night	0.79	3.71
Train station	0.93	3.14	Casino	0.71	3.29
Playground	0.93	3.57	Xmas Tree	0.71	3.14
Farm	0.86	2.43	Kitchen	0.71	1.71
Bowling Alley	0.86	3.57	Auto Repair	0.71	3.29
Beach	0.86	3.29	Bar	0.64	2.29
Fireworks	0.86	3.00	Construction	0.64	2.57
Forest	0.86	2.57	Hockey game	0.64	3.43
Grocery	0.86	3.86	Horse race	0.64	2.29
House Foyer	0.86	4.57	Market	0.57	3.86
Hospital	0.86	3.43	Factory	0.50	2.50
Street protest	0.86	3.71	ATM	0.43	2.57
Auto race	0.86	2.43	Rainy street	0.36	1.71
Restaurant	0.86	2.43	Laundromat	0.36	1.57
Tennis match	0.86	2.71	Office	0.14	2.86
Video arcade	0.86	2.71	Library	0.07	2.43
Crosswalk	0.79	2.71	***Mean***	***0.71***	***2.91***

PREDICTING THE IDENTIFICATION OF SCENES

Sig. Correlation with p(c): Instantaneous Power Slope, r =0.42, No. of Correct Sources, r = 0.40

Sig. Correlation with No. of Correct Sources: LT Spectrum *SD,* r = 0.39, Instantaneous Power Slope, r = 0.36.

No ANOVA with the qualitative variables was significant, although the p(c) by Urban/Rural approached significance, p= 0.08

CONCLUSION

Listeners can reliably identify a number of familiar auditory scenes, as well as a number of the individual sound sources in those streams. The acoustics of the scenes can provide information about semantics, such as whether the scene is urban, the presence of machinery or music, and give some clues as to the number of different sources. However, whether that information is actually used by listeners in identifying the scene is unclear. It may be that for most scenes only it may be necessary to identify only one or two key sources and the remaining information is redundant.

A problem with the corpus used is that the scenes are too identifiable, since nearly all of them were identified > 0.50. Less identifiable scenes that forced listeners to attend to more details might reveal more of what is informative in a scene.

References

Boersma, H. F. (1997). "Characterization of the natural ambient sound environment: Measurements in open agricultural grassland." J. Acoust. Soc. Am. 101(4) 2104–2110.

De Coensel, B. Botteldooren, D., & De Muer , T. (2003). "1/f Noise in Rural and Urban Soundscapes." Acta Acust. Un. Acust., 89, 287-295 .

Gygi, B., Kidd, G.R. & Watson, C.S. (2004). "Spectral-temporal factors in the identification of environmental sounds." J. Acoust. Soc. Am., 155 (3), 1252–1265.

Shafiro, V. (2004). "Perceiving the sources of environmental sounds with a varying number of spectral channels." Unpublished doctoral dissertation, CUNY, New York, NY.

Voss, R. F. & Clarke, J. (1978). "1/f noise in music: Music from 1/f noise." J. Acoust. Soc. Am. , 63(1), 258-263.