

FIRST PRINCIPLES FOR SOUND EVENT RECOGNITION FOR SOUNDSCAPE RESEARCH

JD Krijnders INCAS³, Assen, the Netherlands
TC Andringa Auditory Cognition Group, Artificial Intelligence, Faculty of Mathematics and Sciences, University of Groningen, Groningen, the Netherlands

1 INTRODUCTION

Soundscape research has focussed on describing the way people perceive the soundscape and how the soundscape influences human behavior. Soundscapes can influence people both in a positive way¹, such as in parks and churches and a negative way², such as in cities with traffic and construction work. Models of this influence exist mainly for the negative case³, but these do not include the essential step of sound event recognition. However sound event recognition is an essential basis for these models^{3,4}, both to generate a high-level summary of the soundscape as well as identify the individual sources. In this paper we will discuss the demands for a sound event recognition system that operates in everyday environments.

1.1 Existing sound recognition techniques

Compared to other sound recognition tasks, sound event recognition(ESR) poses a number of different demands. In many pattern recognition tasks it is not unreasonable to demand that the input consists of a single source class. For example, in speech recognition the input can be assumed to be speech in a known language that is produced by a cooperative speaker and has minimal, or known and stable, transmission effects. Therefore it is effective to use features, like Mel-frequency cepstral coefficients (MFCC's, which were developed for speech recognition with Hidden Markov Models(HMM)). However, in the case of environmental sound recognition there is no such thing as a "cooperative source". In fact, the pattern of sources has to be estimated from the signal, since no restrictions on the signals content, other than that it stems from a superposition of physical sources, can be assumed safely. Additionally, all sound sources are influenced by varying, and typically unknown degrees of transmission effects; especially in the form of reflections that add delayed copies of the source signal to the input.

Another important difference is that both the input and the output of a speech recognition system are ordered developments (albeit in very different domains). This constrains and therefore facilitates decoding and is essential for the automatic speech recognizers (ASR) design. In contrast, in ESR the individual sound sources may be either uncorrelated or subject to complex within-class and between-class correlations. Moreover, different sound sources can be defined on quite different temporal scales. The result is a varying superposition of sources instead of a sequence of events with a temporal ordering that is reliable enough to guide decoding. Finally, sounds from more distant or more diffuse sources tend to merge and form a changing diffuse background that might be quite different from the sources that constitute it. Unlike in ASR this background is informative and needs to be described as well.

1.2 Human sound recognition

Humans have little trouble to solve this targets-in-noise problem (for the specific case of speech generally known as the cocktail-party effect^{5,6,7}). Computer implementations on the other hand have problems when speech is mixed with low levels non-stationary background sounds^{8,9,10}. One of the problems of current sound recognition systems is that the MFCC or similar features work best in clean conditions. Concurrent sources influence all coefficients in unpredictable ways. As long as the other sources can be treated as a small perturbation of the target signal this may work, but in many cases this cannot be guaranteed.

The unpredictability of the perturbations makes it hard to separate concurrent sources from the target sound and to properly recognize the target¹⁰.

1.3 ASA and Auditory Object

Human performance seems to benefit from the combination of signal-driven processing and top-down knowledge¹¹. To mimic such an approach both processes should share common representations and have the possibility to reason about multiple interpretation hypotheses. These hypotheses could relate to the auditory objects that appear in modern cognitive research¹¹. What an auditory object is is still unclear^{12,13}, but it should represent information from a single source. The representations for ESR should help to select time-frequency regions that are likely to stem from a single source, which is to be contrasted to approaches where the complete scene is identified as a whole^{14,15,16}. However, these methods can activate top-down knowledge to disambiguate the sound sources found with the proposed methods¹⁷.

The main difference between human perception and modern ASR performance might be related to a difference in signal representation. As a number of decades of research on Auditory Scene Analysis (ASA) have shown¹⁸, humans are able to track the development of (patterns of) signal components such as tones and pulses. The important characteristic of these patterns is that it allows the auditory system to form interpretation hypotheses that are very likely to stem from a single source¹⁸. Unlike the name of the research domain suggest, ASA has not often been aimed directly at real auditory scenes. Instead ASA has shown which rules govern the grouping of evidence into perceptual streams by focussing on basic patterns of tones, noises, and clicks. This streaming behaviour has been modeled^{19,20}, and these implementations have solved some problems in sound source separation and robustness to noise. However, these implementations typically involve re-synthesizing audio from a selection, called a mask, and using this as input for a standard ASR system. This extra step requires a hard (yes/no) decision on what to include in the re-synthesized sound before the signal is actually recognized and positively identified as target. This strict early-state selection may be suboptimal because it is error-prone. However this step is necessary, because the features from computational auditory scene analysis (CASA) do not have the desirable properties of MFCC's required by modern ASR systems²¹.

A related approach, called missing data theory^{22,23}, accounts for the fact that some regions of the time-frequency plane might be more important than other regions and ought to be weighted differently in the decoding process. Both approaches involve the estimation of a mask of which low-level signal properties indicate that it is more likely to represent target than non-target.

1.4 The local signal-to-noise ratio

As noted by Haykin²⁴, the cocktail-party phenomenon is, fifty years after Cherry's seminal work, still an enigma and the answer to the cocktail-party phenomenon requires deep understanding of many fundamental issues that are deemed to be of theoretical and technical importance." The cocktail-party phenomenon can be described as the ability to detect and recognize target sounds that are mixed with and partially masked by similar sounds. So even when the target does not stand out in terms of energy or spectral content, i.e. it is non-salient, it can be detected and recognized. Natural pattern detection and recognition can also rely on more subtle cues than saliency. As summarized in ²⁴ human auditory scene analysis relies the estimation of coherent units of single-source evidence, time-frequency elements or signal components, in combination with a number of principles to group these signal components:

- proximity, which characterizes the distances between the auditory features with respect to their onsets, pitch, and intensity (loudness);
- similarity, which usually depends on the properties of a sound signal, such as timbre;
- continuity, which features the smoothly varying spectra of a time-varying sound source;
- closure, which completes fragmentary features that have a good Gestalt -the completion can be understood as an auditory compensation for masking-, and;
- common fate, which groups together activities (onset, glides, or vibrato) that are synchronous.

Signal component estimation can be related to a sixty year old result by Fletcher²⁵, reviewed by Allen²⁶ who has determined the local signal-to-noise ratio (SNR) as a necessary and sufficient

indicator of the reliability of acoustic evidence: time-frequency regions with a negative local SNR (in dB) did not contribute to recognition performance, but a positive local SNR improved phoneme recognition. A local SNR exceeding 30 dB did not lead to further improvement. The combination of signal component estimation and the notion that all regions with a positive local SNR should be able to contribute to the probability of a correct recognition result, forms the basis of this work. In situations where the signal-driven evidence is reliable it will lead to the activation of interpretation hypotheses for possible groupings. When the signal-driven evidence is less reliable, grouping hypotheses based on context, task demands, or prior knowledge can use the less reliable evidence to decide on the best grouping of signal components. The resulting pattern of grouped signal components provides information about the development of the source that produced it.

Furthermore, when the context provides a matching pitch contour, the system can, ideally, switch from an orienting mode, in which the signal drives processing, to a checking-mode in which knowledge-driven hypotheses search for matching evidence. In the orienting mode the signal must be sufficiently unambiguous to drive processing. In the checking mode, knowledge and expectations counteract signal ambiguity, which allows the system to function in more (adverse) situations. We suggest that the interplay between orienting and checking modes of pattern recognition are an important, and hitherto, neglected ingredient of the cocktail party phenomenon.

2 REQUIRED PROPERTIES

Based on the observations in the first part of this paper, we formulate the required properties of a representation that allow sound event recognition. Note that this list focusses on signal that are well localized in time or frequency. For broadband signals a similar list of properties will be subject of future work.

The representation should:

1. be based on local properties of the time-frequency plane,
2. allow the creation of hypothesis of energy stemming from a single source,
3. be sensitivity to noise in a predictable and local SNR dependent manner,
4. be level independent (i.e. only local SNR dependent),
5. preserve continuity: a smooth (continuous and continuous first derivative) development through time and frequency should lead to a single signal component with a similar development,
6. be frequency independent,

The level and frequency independence prevent unnecessary biases. The relation with the local SNR will allow the estimation of the salience and guarantee that the measures work in the circumstances they are supposed to work in. Continuity preservation is important if the representation is to serve as a basis for object formation²⁷.

3 CONCLUSIONS

We have argued that the current features and recognition strategies for sound recognition are unsuitable for the recognition of sound events in soundscapes. Suitable representation should allow the recognition system to separate different source. Therefore the representations should only be based on local properties in the time-frequency plane. To estimate the saliency of a time-frequency point, estimating the local signal-to-noise will be a useful property as local signal-to-noise ratio is the only measure that influences human performance in speech recognition tasks. In adverse conditions the representation should allow for multiple interpretation hypothesis and for influence of knowledge-driven expectancies.

4 BIBLIOGRAPHY

1. Irvine et al. Green space, soundscape and urban sustainability: an interdisciplinary, empirical study. *Local Environment* (2009) vol. 14 (2) pp. 155-172
2. Truax, *Acoustic Communication*, Greenwood Publishing Group, ISBN 9781567505368, 2nd edition, 2001
3. De Coensel and Botteldooren. Modeling auditory attention focusing in multisource environments. *Proceedings of Acoustics'08* (2008)
4. COST Action 0804, *Soundscape of European Cities and Landscapes*, European Cooperation in the field of Scientific and Technical Research, december 2008
5. Bronkhorst. The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acta Acustica United with Acustica* (2000) vol. 86 (1) pp. 117-128
6. Cherry and Taylor. Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America* (1954) vol. 26 (4) pp. 554
7. Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America* (1953) vol. 25 (5) pp. 975
8. Gong. Speech recognition in noisy environments: A survey. *Speech Communication* (1995) vol. 16 (3) pp. 261-291
9. Lippmann. Speech recognition by machines and humans. *Speech Communication* (1997) vol. 22 (1) pp. 1-15
10. O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition* (2008) vol. 41 (10) pp. 2965-2979
11. Shinn-Cunningham. Object-based auditory and visual attention. *TRENDS in cognitive sciences* (2008) vol. 12 (5) pp. 182-186
12. Carlyon et al. Effects of Attention and Unilateral Neglect on Auditory Stream Segregation. *Journal of Experimental Psychology: Human Perception and Performance* (2001) vol. 27 (1) pp. 115-127
13. Griffiths and Warren. What is an auditory object?. *Nature Reviews Neuroscience* (2004) vol. 5 pp. 887-892
14. Aucouturier et al. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes *The Journal of the Acoustical Society of America* (2007) vol. 122 (2) pp. 881-891
15. Chu et al. Environmental Sound Recognition With Time-Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing* (2009) vol. 17 (6) pp. 1142-1158
16. Eronen et al. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* (2006) vol. 14 (1) pp. 321-329
17. Krijnders et al. Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters* (2010)
18. Bregman. *Auditory Scene Analysis*. Book (1990)
19. Wang and Brown. *Computational Auditory Scene Analysis*. Book (2006)
20. Ellis. *Prediction-driven computational auditory scene analysis*. PhD Thesis (1996)
21. Hermansky. Should recognizers have ears?. *Speech Communication* (1998) vol. 25 (1-3) pp. 3-27
22. Cooke. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America* (2006)
23. Cooke et al. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication* (2001) vol. 34 (3) pp. 267-285
24. Haykin and Chen. The Cocktail Party Problem. *Neural Computation* (2005) vol. 17 pp. 1875-1902
25. Fletcher. The Perception of Speech and Its Relation to Telephony. *The Journal of the Acoustical Society of America* (1950) vol. 22 (2) pp. 89-151
26. Allen. How do humans process and recognize speech?. *IEEE Transactions on Audio, Speech and Language Processing* (1994) vol. 2 (4) pp. 657-577
27. Niessen et al. Disambiguating sounds through context. *International Journal on Semantic Computing* (2009) vol. 2 (3) pp. 1-15