ForumAcusticum
2 0 0 5

# Separable representations for cocktail party processing

Alain de Cheveigné
Ecole Normale Supérieure and CNRS, 29 rue d'Ulm, F-75230, Paris, France e-mail: **Alain.de.Cheveigne_at_ens.fr**,

Perceptual parsing of a complex acoustic scene requires the following ingredients: (a) a representation that is "separable" in the sense that it allows patterns to be split into parts that belong to diverse sources, (b) rules and cues to guide the partition, and (c) processes that can make sense of the partitioned information in the event that the partition was imperfect and the partitioned information is incomplete. Acoustic waveforms can usefully be represented in the *time domain*, in the *frequency domain*, in the *cepstral domain*, etc. These domains are transforms one of the other, and for tasks such as recognizing speech in quiet they may in principle be used interchangeably. In the presence of noise or competing sources, however, partitioning may be easier within one representation than within others. This paper explores the idea that the auditory system forms a set of diverse representations that are redundant in quiet, but useful in cluttered acoustic scenes. These representations are produced by a combination of spectral analysis in the cochlea and time-domain neural processing within the nervous system. Partitioning is based on cues and rules such as described by Bregman (1990). The final pattern-matching stage must be able to use the possibly incomplete information that survives the partitioning stage. Similar ideas may be applied to machine analysis of complex acoustic scenes.

## 1 Introduction

This paper is about acoustic scene analysis by man and machine. It departs slightly from current wisdom on auditory scene analysis (ASA, [2]) and computational ASA (CASA, [25]) by exploring four ideas. The first is that successful scene analysis depends on finding a *separable* representation within which the correlates of different sources can be distinguished. The familiar time-frequency plane is one example, but other representations are possible. The second idea is that this representation should be understood as a means to *suppress* interfering sources, rather than, or in addition to, enhance the target source. The third idea is that segregation is inherently imperfect: parts of the target are likely to be suppressed together with the interference, and the incomplete target thus needs to be interpolated or reconstructed. This requires internal *models*. The fourth idea is that it may be useful to assume a set of *redundant* representations with different separability properties. In quiet they are equivalent, but in noise one or the other may allow better suppression of interference. Throughout this paper we consider both natural and artificial processing.

Accounts of auditory scene analysis [2] usually describe perceptual organization in terms of *grouping* and *segregation* of components ("partials") belonging to each source. For example if one partial of a vowel starts earlier than the others, the timbre of that vowel shifts as if the partial were discounted in the calculation of the vowel's spectrum [6]. A partial that is mistuned or modulated may emerge as a perceptual entity distinct from the rest of a harmonic complex tone [16]. The easier identification of concurrent vowels with fundamental frequencies ( $F_0$s) that are different, rather than the same, is usually explained by saying that partials of either vowel are segregated because they belong to distinct harmonic series [27], etc. Implicit in such explanations is the notion that sounds are "made up" of sinusoidal partials with frequencies that follow arbitrary time courses, and with precisely delimited temporal extents, and that the auditory system has access to individual representations of each. This view is problematic, as the temporal and spectral resolution of the auditory system is known to be limited. Indeed, the well-known uncertainty relation of Gábor [15] implies that no measurement device – whether human or machine – can have perfect time *and* frequency resolution.

Segregation necessarily operates within the constraints of physiology (ASA by humans) or physically realizable operations (CASA by machines). However, these constraints allow some flexibility. Auditory processing is limited by the selectivity of initial filtering within the cochlea, but the *temporal structure* of neural signals that are carried to the brain allows this selectivity to be adjusted to some degree. For example spectral resolution can be enhanced (at the expense of temporal resolution) by cross-channel or within-channel neural subtraction (e.g. [7, 28]) or reduced (for the benefit of temporal resolution) by cross-channel neural summation [22]. CASA systems have a yet larger range of representations to choose from, and a range of time-frequency tradeoffs. These include the short-term Fourier transform, wavelets, chirps, etc. One representation may give better frequency resolution, another better time resolution, etc. Each may be of use in some particular situation.

Most systems or models choose, among these many alternatives, a single representation that offers a compromise

between conflicting requirements. Here we explore instead the idea of multiple representations. Traditionally, the focus has been on finding a representation that best represents features of the *target* pattern, or best allows them to be selected from a mixture. Here we take a different viewpoint: the representation should allow *interference* to be suppressed. Representations are judged on their capacity to allow such suppression. Different forms of interference may require different representations, and this leads to the notion of multiple representations.

## 2 Separable representations

A *separable representation* designates any representation that allows correlates of interference to be discounted. The key property is that the representation of the composite scene (target + interference) is separable into two parts, one affected by interference, and the other only by the target. Observable data are, as it were, *rotated* and *projected* such as that the interference projects to a limited part of the representation, that can be discounted. As a trivial example, interference concentrated within a limited time interval can be discounted within a time-domain representation. Narrow band interference can be suppressed in a frequency representation (spectrum). Interference that extends over a limited patch of time and frequency may be attenuated within a time-frequency representation.

Models of auditory segregation, such as the concurrent vowel identification model of Meddis and Hewitt [21], or the earlier binaural model of Lyon [19], operate by gating channels of a cochlear model according to whether they are dominated by one source or another. A similar idea is used in CASA systems, such as the early system of Weintraub [29], or later work such as [4]. These models and systems create a separable representation on the basis of cochlear filtering. Depending on the model, this consists either of a two-dimensional time-frequency map (with pixels belonging to either source), or a more complex set of autocorrelation patterns, also indexed by time and frequency. Each pixel is assigned to either source. The spectro-temporal segregation principle is ubiquitous in segregation models, from the pioneering work of Parsons [23] to recent efforts such as [26, 30].

However, there exist alternative approaches to separation. For example a harmonic interferer can be suppressed by a time-domain comb-filter [11]. There is evidence that the auditory system may use an analogous form of processing applied to neural spike trains [8]. The equalization-cancellation (EC) model of [14] performs a similar operation based on binaural interaction. Ideally, in the case of perfectly periodic (or interaurally correlated) interference, these operations suppress all evidence of the interfering source.

At this point it is worth mentioning two basic approaches to interference suppression: subtractive and multiplicative. In the subtractive approach, a representation of the interference is subtracted from the representation of the scene, leaving (hopefully) an intact representation of the target. In the multiplicative approach, correlates of the interference are suppressed by setting them to zero. The subtractive approach requires a linear representation (waveform, complex spectrum), whereas multiplicative suppression can apply to a non-linear representation (e.g. power spectrum). Multiplication by zero in essence "destroys" features contaminated by interference, while subtraction allows (in principle) the correlates of the target to emerge intact. If phase is available, then spectral subtraction may reasonably be applied to the complex spectrum (magnitude and phase). If phase is unknown, then a power spectrum is more appropriate, as power then combines additively on average. Note however that if the statistics of the sources are *sparse* within the time-frequency representation (as is often the case with sources such as speech or musical instruments), the particular scaling (magnitude, power, log) is indifferent as each pixel is usually dominated by one source only [26]. This explains the success of time-frequency plots as separable representations.

## 3 Interference cancellation

Given a separable representation, should the system focus on correlates of interference to suppress them, or on correlates of the target to enhance them? The previous section took the former for granted, but this is in contrast with ASA models and CASA systems that usually assume target enhancement. There are several reasons to favor interference cancellation. One is that there is converging evidence to suggest that the auditory system uses such a strategy, at least in presence of harmonic or binaurally correlated interference [7, 12, 13, 5, 14]. Another is that the cancellation strategy is functionally superior in the sense that (if successful) it removes dependency of segregated patterns on the interferer. In this way the system cannot mistake features of the interference as belonging to the target. See [11] for an in-depth discussion of the cancellation principle in acoustic scene analysis.

## 4 Missing features

Ideally, we would wish the representation to allow perfect suppression of interference, while at the same time preserving the target. Such is rarely the case: often some target correlates are suppressed by the cancellation mechanism. There is usually a tradeoff between the amount of target suppression and the amount of residual interfer-

ence. Arguably it should be arbitrated in favor of less interference, as leak-through from interference corrupts features in ways that are unpredictable and impossible to compensate for. In contrast, target distortion is often predictable from knowledge of the segregation process, and thus easier to handle. Solutions to target distortion depend on the context within which scene analysis is performed. If the task is pattern-matching, known target distortion can be compensated within the pattern-matching stage. If instead the goal is to output a "clean" target waveform, for example to present to a listener's ears, then distortion or missing features are a more serious problem. It is best solved by applying *models* constrained by learning, physical principles, etc, to interpolate across missing parts. Useful models may be instrument models, voice production models, grammars, etc. "Missing feature" theory is a hot topic in speech recognition [3, 18]. See [11] for a discussion in the context of scene analysis. The ability to deal with incomplete targets is a key requirement for acoustic scene analysis.

# 5 Multiple redundant representations

As mentioned earlier, interference local in time or frequency favors a temporal or spectral representation respectively. Other spectro-temporal patterns may fit different spectro-temporal analyses (short-term Fourier transform, wavelets, chirp transforms, etc.), while periodic interference fits a representation incorporating harmonic cancellation. Obviously no single representation offers every advantage, and for this reason it is worth considering to entertain *multiple* representations and models in parallel. To the degree that they are transforms one of the other, the information they carry is redundant, but one representation or another may offer a better "separable representation" depending on the interference. Multiple representations entail a cost in terms of storage and processing, but this is not a serious obstacle for the massively parallel auditory nervous system, or for modern computational systems.

## 5.1 Auditory processing

One can speculate that the auditory system is designed to elaborate multiple transforms of incoming acoustic information, and that the purpose of at least some of these redundant representations is to allow the contribution of interfering sources to be discounted [9]. The first steps are well documented: quasi-linear filtering followed by hair-cell transduction to form approximately 3000 channels, each coded by about 10-20 afferent auditory nerves. Within the cochlear nucleus and subsequent nuclei, these patterns are recoded into a very wide variety of responses [24]. Because of their complexity, their functional relevance for the processing of sound is obscure. However, it has been argued that apparently "random" non-linear transforms of input patterns can be used to implement sophisticated pattern-matching functions [20]. Furthermore, there are reasons to believe that functionally useful steps, such as redundancy reduction, will result in hard-to-interpret responses [1]. One may hypothesize that many such elementary operations are performed systematically within the auditory system, particularly at lower levels within the brainstem, and that higher level mechanisms *select* among them those that turn out to be most useful for a given task [20].

Among these elementary transforms, some may be useful for segregation. Such is obviously the case of the initial stage of cochlear filtering. The above-mentioned EC model [14] is another example: the "equalization" step of that model can be understood as a particular projection that cancels interaurally-correlated interference. The "harmonic cancellation" model of [7] performs a similar operation on the basis of harmonic structure. Simple neural operations such as coincidence may be assembled to effectively perform segregation operations such as harmonic cancellation [10]. In other words, the system performs, at each level, many "projections" of the incoming patterns in the hope that one at least will fit the structure of interference and allow it to be discounted. It is not essential to assume that all representations are calculated in parallel in every situation. The essence of the conjecture is rather that the higher levels can accommodate whatever representation is chosen to escape the interference at hand. In other words, internal "templates" exist in multiple formats. This conjecture is hard to prove or disprove experimentally, but it may be of help to make sense of the complex structures and responses revealed by studies of the auditory system.

## 5.2 Automatic speech recognition

Automatic speech recognition (ASR) usually operates on the basis of a particular signal feature representation (e.g. mel spectrum, mel-frequency cepstral coefficients or MFCC, RASTA-PLP, etc.). Much effort has been invested into the choice and optimization of this representation. However it has been found that features that have proven most effective in quiet (e.g. MFCCs) are difficult to deploy in systems that deal with interference [3, 18, 17]. An interesting alternative would be to use a set of redundant features, all trained on the same training data, comprising for example spectral and cepstral features with diverse spectro-temporal resolution properties. On clean speech, all features are expected to give roughly the same performance. For mixtures, one or the other may allow better segregation and interference suppression, and effective application of missing-feature

techniques [3, 18]. For example, the system of [17] operates on a spectro-temporal representation with resolution properties that are presumably optimal for some but not all sorts of interference. Extending the system to include *multiple* representations would presumably allow it to deal effectively with a wider range of interference.

# 6 Summary

This paper put forward several ideas. An acoustic scene analysis system (natural or artificial) can benefit by incorporating multiple redundant representations, each tailored to allow various forms of interference to be suppressed. The key to effective scene analysis is suppression of interfering sources, so that their correlates do not interfere with those of the target. Interference suppression is likely to result also in distortion of the target. The suppression parameters being known, the system can interpolate over the missing parts on the basis of internal models. It is expected that these ideas may be of practical use in machine-based systems. They may also offer a key to understanding scene analysis processes that go on within the auditory system.

# References

[1] H.B. Barlow. Possible principles underlying the transformations of sensory messages. In W.A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge Mass, 1961.

[2] A. S. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, Mass., 1990.

[3] M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *ICASSP*, volume II, pages 863–866, 1997.

[4] M. P. Cooke. *Modeling auditory processing and organisation*. PhD thesis, Sheffield, Department of Computer Science, 1991.

[5] J.F. Culling and Q. Summerfield. Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.*, 98:785–797, 1995.

[6] C.J. Darwin and R.W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *J. Acoust. Soc. Am.*, 103:1080–1084, 1998.

[7] A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acoust. Soc. Am.*, 93:3271–3290, 1993.

[8] A. de Cheveigné. Concurrent vowel identification iii: A neural model of harmonic interference cancellation. *J. Acoust. Soc. Am.*, 101:2857–2865, 1997.

[9] A. de Cheveigné. The auditory system as a separation machine. In J. Breebaart, A. J. M. Houtsma, A. Kohlrausch, V. F. Prijs, and R. Schoonhoven, editors, *Physiological and Psychophysical Bases of Auditory Function*, pages 453–460. Shaker Publishing BV, Maastricht, The Netherlands, 2001.

[10] A. de Cheveigné. Correlation network model of auditory processing. In *Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis*, Aalborg (Danmark), 2001.

[11] A. de Cheveigné. The cancellation principle in acoustic scene analysis. In P. Divenyi, editor, *Perspectives on Speech Separation*, pages 243–257. Kluwer, New York, 2004.

[12] A. de Cheveigné, S. McAdams, J. Laroche, and M. Rosenberg. Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *J. Acoust. Soc. Am.*, 97:3736–3748, 1995.

[13] A. de Cheveigné, S. McAdams, and C. Marin. Concurrent vowel identification ii: Effects of phase, harmonicity and task. *J. Acoust. Soc. Am.*, 101:2848–2856, 1997.

[14] N.I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am.*, 35:1206–1218, 1963.

[15] D. Gábor. Acoustical quanta and the theory of hearing. *Nature*, 159:591–594, 1947.

[16] W.M. Hartmann and S.L. Doty. On the pitches of the components of a complex tone. *J. Acoust. Soc. Am.*, 99:567–578, 1996.

[17] G. Hu and D.L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans on Neural Networks*, 15:1135–1150, 2004.

[18] R.P. Lippmann and B.A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *ESCA Eurospeech*, pages KN–37–40, 1997.

[19] R.F. Lyon. A computational model of binaural localization and separation. In W. Richards, editor,

*Natural computation*, pages 319–327. MIT Press, Cambridge, Mass, 1983-1988. reprinted from Proc. ICASSP 1983, 1148-1151.

[20] W. Maass, T. Natschläger, and H. Markram. Computation models for generic cortical microcircuits. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, pages 575–605. CRC-Press, 2003.

[21] R. Meddis and M.J. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Am.*, 91:233–245, 1992.

[22] A. Palmer, D. Jiang, and D.H. Marshall. Responses of ventral cochlear nucleus onset and chopper units as a function of signal bandwidth. *J. Neurophysiol.*, 75:780–794, 1996.

[23] T.W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60:911–918, 1976.

[24] W.S. Rhode and S. Greenberg. Physiology of the cochlear nuclei. In A.N. Popper and R.R. Fay, editors, *The mammalian auditory pathway: neurophysiology*, pages 94–152. Springer Verlag, New York, 1992.

[25] D.F. Rosenthal and H.G. Okuno. *Computational auditory scene analysis*. Lawrence Erlbaum, 1997.

[26] S. Roweis. One-microphone source separation. In *Advances in NIPS*, page 609–616. MIT Press, Cambridge MA, 2000.

[27] M.T.M. Scheffers. *Sifting vowels*. PhD thesis, Gröningen, 1983.

[28] S. A. Shamma. Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J. Acoust. Soc. Am.*, 78:1622–1632, 1985.

[29] M. Weintraub. *A theory and computational model of auditory monaural sound separation*. PhD thesis, Stanford, 1985.

[30] M. Wu, D. Wang, and G.J. Brown. A multipitch tracking algorithm for noisy speech. *IEEE Trans. ASSP*, 11:229–241, 2003.