Modeling auditory scene analysis by multidimensional statistical filtering may stimulate advances in hearing-aid signal processing

VOLKER HOHMANN

Medical Physics, University of Oldenburg, D-26111Oldenburg, Germany

‘Auditory Scene Analysis’ (ASA) denotes the ability of the human auditory system to decode information on sound sources from a superposition of sounds in an extremely robust way. ASA is closely related to the 'Cocktail-Party-Effect' (CPE), i.e., the ability of a listener to perceive speech in adverse conditions at low signal-to-noise ratios. This contribution discusses theoretical and empirical evidence suggesting that robustness of source decoding is partly achieved by exploiting redundancies that are present in the source signals. Redundancies reflect the restricted spectro-temporal dynamics of real source signals, e.g., of speech, and limit the number of possible states of a sound source. In order to exploit them, prior knowledge on the characteristics of a sound source needs to be represented in the decoder/classifier (‘expectation-driven processing’). In a proof-of-concept approach, novel multidimensional statistical filtering algorithms such as ‘particle filters’ have been shown to successfully incorporate prior knowledge on the characteristics of speech and to estimate the dynamics of a speech source from a superposition of speech sounds (Nix and Hohmann, 2007).

LIMITATIONS OF CURRENT NOISE REDUCTION STRATEGIES

The problem of compensating for the reduced ability of most hearing-impaired listeners to communicate in difficult acoustical conditions is most challenging. Although advances in noise reduction have been made using digital signal processing in the last few years, the performance of hearing-impaired listeners with hearing-aids is still worse than that of normal-hearing listeners in most listening conditions characterized by background noise and reverberation. This section briefly introduces current noise reduction strategies and identifies their limitations.

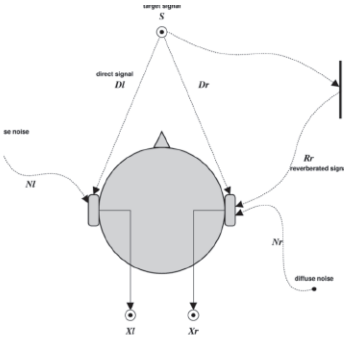

Fig. 1: Source configurations considered for noise reduction techniques in hearing-aids.

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

Figure 1 shows the signal model considered for hearing-aid applications. A directional target signal is assumed to be present, which is in many cases assumed to be at or close to the frontal direction. Reverberation is added to the target depending on the room acoustics. In addition to further directional distracting targets (not shown in Fig. 1), diffuse noise sources may be present. At each ear, one or more microphones pick up the signal (only one signal at each ear shown in Fig. 1). Subsequent processing may use a single microphone input only (monaural processing) or may combine information from several microphones positioned at one ear (small microphone arrays), or from both ears (binaural processing). Two general types of algorithms for noise reduction based on the signal model described above can be identified:

Quasi-stationary spatial filtering with multi-microphone input:

Blind Source Separation (BSS, e.g., Anemüller and Kollmeier, 2000; Parra *et al.,* 1998); Directional Microphones and Adaptive Beamforming (e.g., Elko *et al.,* 1995); Greenberg and Zurek, 1992; Kompis and Dillier, 1994; Kates and Weiss, 1996).

Time-variant envelope filtering:

Single-channel noise reduction (e.g., variants of Wiener filtering (e.g., Levitt *et al.,* 1993; Ephraim and Malah, 1985)); multi-microphone noise reduction, e.g., Bodden (1993).

'Quasi-stationary' spatial filtering means that the filter characteristics does not change with the dynamics of speech, but with the slower changes in the spatial configuration of the sound sources. In conjunction with minor filter constraints, this property guarantees an almost artifact-free output signal. Fixed or adaptive directional microphones using the output of two or three microphones mounted closely in one hearing-aid shell are therefore commonly used in current hearing-aids. They provide good-quality output and improvements in speech reception threshold in noise (SRT) of 3-6 dB. Common to all spatial filtering schemes is the fact that they cannot separate more sources than the number of microphones being used. This is because of the underlying linear signal mixture model, which is sketched for the case of two signals and two receivers/microphones in Eq. 1 (convolutive mixture, equation applies separately to each frequency bin of a Fourier analysis):

$$\begin{pmatrix} z_l \\ z_r \end{pmatrix} = \begin{pmatrix} h_{1l} & h_{2l} \\ h_{1r} & h_{2r} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \qquad \text{(Eq. 1)}$$

$$h_{ik} = g_k\left(\alpha_i, \varphi_i\right) \qquad \text{(Eq. 2)}$$

The spectral values z_l and z_r of the signals observed at the two microphones (close-distance microphones or binaural configuration) are assumed to be a linear mixture of the spectral values s_1 and s_2 of the two (uncorrelated) signals. The mixing matrix contains the coefficients of the head-related transfer functions (HRTF), which depend on the signal's direction (azimuth α and elevation φ). Spatial filtering schemes like BSS generally reconstruct the signals by estimating and inverting the mixture matrix, which is possible in case the matrix is square. For multi-talker configurations, de-mixing is not possible in general, because Eq. 1 forms an underdetermined system of equations in this case (e.g., a 2x3 mixing matrix for 2 sensors and 3 sources). The filter effect is therefore limited, and target suppression often occurs in these configurations due to filter convergence errors. Also, if the spatial filter implements a sharp spatial zero in its transfer-function, slight head movements that cannot be avoided in hearing-aid applications will lead to a strong modulation of the level of the noise signal, because it is moving on the steep skirts of the transfer function close to its zero.

Whereas spatial filters aim at canceling directional distracting sources by adapting a zero of the transfer function towards the distractor's direction, envelope filters, on the other hand, aim at reconstructing the target signal's envelope in frequency subbands. Envelope filtering therefore requires filter adaptation and variation with a much higher temporal resolution of about the syllable rate of speech. In general, these algorithms estimate the envelope of the target source in several frequency bands, which is equivalent to short-time signal-to-noise ratio (SNR) estimation. The target is generally defined by the frontal direction in case the envelope filter exploits spatial information (multi-microphone input), or is defined by different statistical properties of target speech and noise in single-microphone systems, e.g., amplitude modulated target speech signal vs. non-modulated (stationary) noise signal. The signal is then modulated with the estimated envelope in order to reconstruct the target's envelope and thereby to suppress the other sources. When defining source separation based on the envelope rather than the complete fine structure of the signals, envelope filtering is not limited to suppressing one noise source only. In multi-talker configurations including reverberation, however, SNR estimators generally show significant errors. Envelope filtering is therefore vulnerable to generate disturbing artifacts due to the limited information on the true target envelope. Envelope filtering schemes using single-microphone input (monaural noise reduction, e.g. Ephraim and Malah (1985)) are well established in current hearing-aids. They are restricted to improving perceived noisiness of a noise condition, but fail to improve speech intelligibility (see, e.g., Marzinzik and Kollmeier, 2003).

In summary, the current techniques of noise reduction have the following limitations:

Spatial filters:

Limited SNR-improvement and directivity (low-order directional microphones)

Adaptation time too long compared to changes in the spatial environment (e.g., head rotation) (BSS schemes)

Limited to the n-source n-microphone case

Envelope filters:

Limited accuracy of short-term sub-band SNR estimation

Prone to generating audible and disturbing processing artifacts, because of the errors in estimating the SNR

No increase in speech reception threshold (SRT) in noise; only reduced listening effort observed

The work presented in this paper investigates possible strategies to overcome the limitations of envelope filters. First, possible reasons for the limited success of the current methods of envelope filtering for noise reduction are discussed. Then, a generic model is proposed that may resolve these issues by incorporating *a-priori* knowledge on the statistical properties of the sound sources. Finally, the application of this model to separating multiple speech sources from a binaural signal is introduced in a proof-of-concept approach.

AUDITORY SCENE ANALYSIS AND EXPECTATION-DRIVEN PROCESSING

Three major reasons why current techniques of noise reduction fail in difficult noise conditions where the normal hearing system successfully separates the target can be identified based on the available literature:

Ambiguity of the source separation problem: More sources than microphones make the problem of separating the sources ambiguous and prevent linear solutions from being applicable (cf. Eq. 1 for the case of more sources than receivers).

Random fluctuations of signal-derived parameters: Diffuse background noise and reverberation impose statistical fluctuations on all signal-derived parameters, reducing the accuracy of parameter estimates used for filtering (e.g., SNR estimates, or interaural differences (see, e.g., Nix and Hohmann, 2006).

Missing information: Linearly superposed daily-life signals overlap significantly in the time-frequency domain. Thus, major parts of the information on the different sources is totally masked in the time-frequency representation, leading again to a limited accuracy in SNR estimates for envelope filters (notion of ‘disjoint orthogonality, see, e.g., Hohmann *et al.*, 2002).

The possible principles for solving the above-mentioned limitations commonly believed to be used by the hearing system for the analysis of the acoustical scene are as follows:

Using statistical a-priori knowledge on the statistical properties of sound sources in the environment allows for solving ambiguities and for filling out missing information by exploiting the sources’ redundancies.

Statistical combination of several noise-deteriorated parameters allows for noise-robust extraction of information on the sound sources.

An example for 2. is the noise-robust estimation of sound source direction by integration of (noisy) interaural parameters across frequency (Nix and Hohmann, 2006). As an example for 1. we might assume as a thought-experiment that the signals’ spectral values at two different frequencies f_1 and f_2 are 100% correlated (i.e., they are identical). In this case, Equation 1 contains the same values on the right-side for the two different frequencies. Thus, we end up having four equations for only three unknowns (see Eq. 3), and the ambiguity for the case of more sources than receivers is completely solved by using the sources' redundancy.

$$f_1 : \begin{pmatrix} z_{l,f_1} \\ z_{r,f_1} \end{pmatrix} = \begin{pmatrix} h_{1l,f_1} & h_{2l,f_1} & h_{3l,f_1} \\ h_{1r,f_1} & h_{2r,f_1} & h_{3r,f_1} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix}$$

$$f_2 : \begin{pmatrix} z_{l,f_2} \\ z_{r,f_2} \end{pmatrix} = \begin{pmatrix} h_{1l,f_2} & h_{2l,f_2} & h_{3l,f_2} \\ h_{1r,f_2} & h_{2r,f_2} & h_{3r,f_2} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} \qquad \text{(Eq. 3)}$$

In reality, the redundancy is not 100%, but significant correlations of spectral values across frequency have been observed, in particular in speech (Anemüller, 1999). This means that exploiting redundancies generally is not possible by putting up more equations. Instead, statistical frameworks are needed for exploiting them.

A general framework for both incorporating statistical source models describing the sources' redundancies and combining information from different signal-derived parameters in an optimal statistical way are Sequential Monte-Carlo methods (SMC, also known as 'particle filters'; see, e.g., Arulampalam *et al.*, (2002)). First applied to scene analysis problems in vision, interest in applying these schemes in Computational Auditory Scene Analysis (CASA) is rapidly increasing.

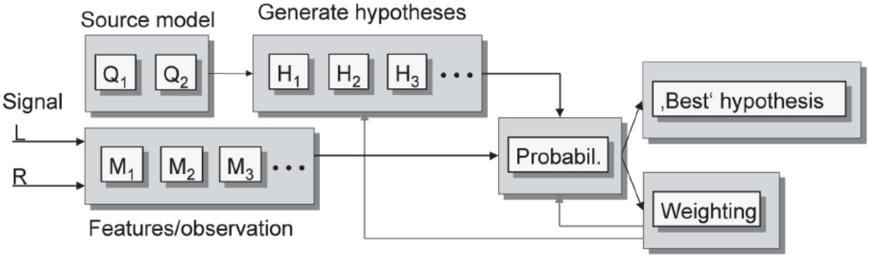

Fig. 2: Processing framework for combining the signal-driven internal representation (features/observation) with a-priori knowledge on sound sources (source model) in a probabilistic way (see text for details).

Figure 2 shows a generic block diagram of SMC processing schemes applied to binaural acoustic input. An auditory preprocessing model generates an internal representation, which forms the observation and contains masked, incomplete and ambiguous information on the sound sources. A source model (e.g. for speech) generates hypotheses of the possible states of the sources, which establishes a ‚mirror' of the ‚outer world'. Hypotheses do not necessarily reflect the properties of the sound generated by the source, but may include the parameters of a physical model of the source (e.g., articulatory parameters in case of a speech source). Statistical decoding (e.g., SMC) combines hypotheses and observation optimally and non-hierarchically. The decoding process estimates the 'best hypothesis given the observation and assigns weights (probabilities) to the hypotheses. Ambiguities are resolved in this framework because the number of possible states of the source is restricted largely by the source model. Masked and incomplete information is amended by exploiting the ability of the source model to predict probable states of the sources from the current 'best' hypothesis even if the observation fails intermittently.

Figure 3 sketches the implications of the use of a source model when separating superposed speech sources from a spectro-temporal representation. Whereas current envelope filters for noise reduction estimate the relative contribution of target and noise in a single time-frequency bin from the signal observed in that particular bin, the inclusion of the source model allows to estimate the bin from the passed frames (spectro-temporal information). This estimation is very different from just averaging, because averaging would smear out the highly dynamic spectro-temporal characteristics of speech. A source model of speech, on the other hand, takes these characteristics into account, e.g., by anticipating the probability of common onsets across frequency after a speechpause.

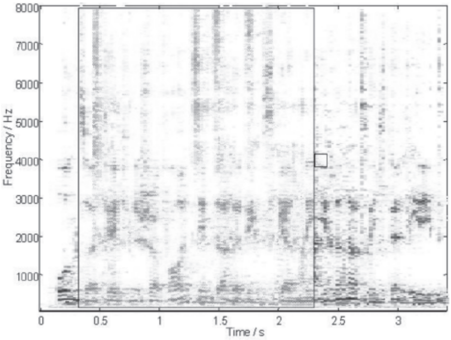

Fig. 3: Spectro-temporal representation of two superposed speech signals: Estimating speech in single time-frequency bins (small rectangle) from information gathered in passed frames (large rectangle) (see text for details).

PRINCIPLES OF PARTICLE-FILTERING

The general principles of particle filtering shall be outlined briefly in this section in order to clarify the idea of (i) the inclusion of a source model and (ii) the optimal combination of hypotheses and observation. Further information can be found in Arulampalam *et al.* (2002).

Figure 4 shows the processing steps of particle filters for a model system describing an harmonic oscillator. The system is completely described by a state space, which is two-dimensional in this case (position and velocity variables). The true temporal evolution of the system is specified by an unknown trajectory in the state space (blue circle in Fig. 4). In order to estimate it, a set of particles are distributed in the state space, which sample possible system states (bullets in Fig. 4). Starting from an initial distribution of particles, four processing steps are taken for each time step:

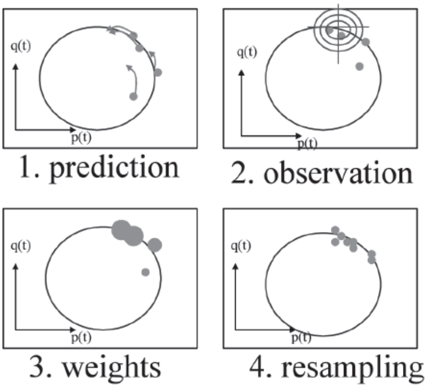

Fig. 4: Generic processing steps of particle filters. The model system is an oscillator described by a two-dimensional state space (position q and velocity p) (see text for details).

1. Predict the possible system state at the next time step separately for each particle. This prediction step is denoted by the arrows in the upper left panel and needs source statistics, i.e., statistical *a-priori* information on the possible evolution of the system. In this case, a circling motion in the state space is expected.

2. The predicted possible states are compared to an observation of the state, which might be fuzzy, incomplete and deteriorated by noise (denoted by the cross-hair in the upper left panel).

3. By assessing the congruence of observation and particles using an observation statistics, a weight is assigned to each particle, that denotes the likelihood of each particle given the observation. Weights are denoted by the different sizes of the bullets in the lower left panel.

4. Finally, particles with low weights are discarded and replaced by new particles close to the particles with high weights. The procedure continues for the next time step with step one using the set of particles generated by the last step four.

The procedure sketched above ensures that the probability density function (PDF) of the system state given the observation is sampled by the particle positions and their assigned weights. Estimates of the true trajectory can therefore be calculated by taking the weighted mean of the particle positions. Taking the processing principle to the audio domain, a-priori information on the possible temporal evolution of sound sources is used in step one, allowing for the extrapolation of their evolution in case part of the received acoustical information on these objects is ambiguous or even fully masked. Step three explicitly performs the combination of several signal-derived parameters, because the observation statistics calculates the likelihood of a hypothesis/particle given both the hypothesized state value and the values of **all** observed variables.

The next section introduces a proof-of-concept approach using the particle filtering technique for separating and localizing superposed speech sources from a binaural signal.

SEPARATING SPEECH SOURCES USING A PARTICLE FILTER APPROACH

Nix and Hohmann (2007) introduced a framework for tracking sound source directions and spectral envelopes of superposed speech signals based on the generic processing scheme sketched in Fig. 2. The approach uses a detailed statistical description of the high-dimensional spectro-temporal dynamics of speech, which establishes prior information (expectation) and was measured from a large speech database. Fig. 5 shows a sample spectrogram generated with the speech model by starting with a random speech spectrum and sampling subsequent spectra from the dynamical model using a trained Markov chain. It can be seen that the model generates speech characteristics like onsets and typical sequences of vowel- and consonant-like spectra.

The results of Nix and Hohmann (2007) show that the algorithm tracks sound source directions very precisely, separates the voice envelopes with algorithmic convergence times down to 50ms, and enhances the signal-to-noise ratio in adverse conditions, requiring very high computational effort. The approach has a high potential for improvements of efficiency and could be applied for voice separation and reduction of non-stationary noises. For further details refer to Nix and Hohmann (2007).

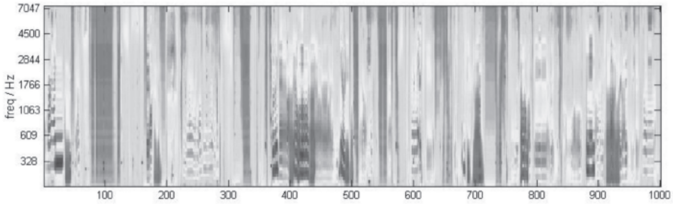

Fig. 5: Sample spectrogram generated by the dynamic speech model used by Nix and Hohmann (2007) for establishing in their particle-filter-based source separation scheme.

ACKNOWLEDGMENTS

Thanks to Birger Kollmeier for his continuous support and contribution to this work. Thanks also to the members of the Medical Physics group, in particular to Johannes Nix. Work supported by DFG (SFB-TR31).

REFERENCES

Nix, J., and Hohmann, V. (**2007**). "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering" IEEE Trans. Audio, Speech and Lang. Proc. **15**(3): 995-1008.

Anemüller, J., and Kollmeier, B. (**2000**). "Amplitude modulation decorrelation for convolutive blind source separation," In: P. Pajunen und J. Karhunen (Eds.), Proc. of the second int. workshop on independent component analysis and blind signal separation, Helsinki, pp. 215-220.

Parra, L., Spence, C., and de Vries, B. (**1998**). "Convolutive blind source separation based on multiple decorrelation," In: IEEE Neural Networks and Signal Processing Workshop, Cambridge, 1998.

Elko, G. W., and Anh-Tho, Nguyen Pong (**1995**). "A simple adaptive first-order differential microphone," IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA.

Greenberg, J. E., and Zurek, P. M. (**1992**). "Evaluation of an adaptive beamforming method for hearing aids," J. Acoust. Soc. Am. **91**(3):1662-76.

Kompis, M., and Dillier, N. (**1994**). "Noise reduction for hearing aids: combining directional microphones with an adaptive beamformer," J. Acoust. Soc. Am. **96**(3):1910-13.

Kates, J. M., and Weiss, M. R. (**1996**), "A comparison of hearing-aid array processing techniques," J. Acoust. Soc. Am. **99**(5):3138-48.

Levitt, H., Bakke, M, Kates, J.M. (**1993**). "Signal processing for hearing impairment, "Scand. Audiol. Suppl. **38**.

Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Sig. Proc. ASSP **33**(2): 443-445.

Bodden, M. (**1993**). "Modeling human sound-source localization and the cocktail-party-effect," Acta Acustica **1**:43-56.

Marzinzik, M., and Kollmeier, B. (**2003**). "Predicting the Subjective Quality of Noise Reduction Algorithms for Hearing Aids," Acta acustica/Acustica, **89**, p. 521-529.

Nix, J., and Hohmann, V. (**2006**). "Sound source localization in real sound fields based on empirical statistics of interaural parameters," J. Acoust. Soc. Am. 119(1): 463-479.

Hohmann, V., Nix, J., Grimm, G., Wittkop, T. (**2002**). "Cocktail party processing based on interaural parameters," Forum Acusticum, Sevilla, SEA, ISBN 84-87985-06-8.

Anemüller, J. (**1999**). "Correlated modulation: a criterion for blind source separation," In Joint Meeting "Berlin 99" integrated 25th German Acoustics DAGA Conference, Berlin, p. 4.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (**2002**). "A tutorial on particle filters for online nonlinear/non-Gaussian bayesian tracking," IEEE Transactions on Signal Processing, **50**, 174–188.