Enhancing sound sources by use of binaural spatial cues

Johannes Nix, Volker Hohmann

AG Medizinische Physik
Universität Oldenburg, Germany
jnix@medi.physik.uni-oldenburg.de

Abstract

A highly desirable possible application of simulation of auditory scene analysis is the development of hearing aids capable of suppressing disturbing sound sources in complex noise environments. It is well-known that in such situations spatial cues contribute a small, but often decisive hint for auditory scene analysis. Concerning noise suppression, they have the interesting advantage that they do not require prior knowledge of the sound source spectrum, and that the directions of the sound sources change relatively slowly.

Part of these spatial cues are interaural phase and level differences. Using a statistical approach, it is possible to use these parameters from a mixture of sound sources to estimate sound source direction. It has been shown that this works well even in moderate to low signal to noise ratios with an accuracy comparable to what human subjects achieve[10, 8].

Using known head-related transfer functions, estimates of the sound source direction can be directly used to enhance one signal from a two-microphone recording of two signals. However, if more than two sound sources are involved, additional information is needed to estimate the contributions of the different sources. This information can be gained by ”binding” or integrating temporal information from the signal envelope with information on sound source direction. Therefore correlations of direction information and spectro-temporal features are examined. Results show clear correlations between the envelope of the sources and the directional information.

1. Introduction

In difficult listening situations, humans exploit spatial characteristics of sound signals by mechanisms of binaural hearing. Part of the signal features that result from the directional filtering of head and pinna are the interaural phase and level differences.

Interaural phase and level differences show characteristic patterns for each direction. In quiet environments, these patterns have been used for the classification of sound source direction by neural networks; however in environments with high levels of noise, this is a relatively difficult task [7, 4, 3]. The reason for this is that the identification of directions depends partly on the fine structure of phase and level differences which is disturbed strongly in noise. Phase and level differences of the noise field interact with the parameters of the sound sources, leading to some characteristics, e.g. the mean level difference of an isotropic noise field will be zero, and thus the resulting level differences from any sound source will be ‘pulled’ towards zero on average. Therefore, for a reliable estimation of sound source direction, this influence has to be taken into account.

Once the sound source directions are known, there exist two possible strategies for enhancing a specific sound source. If only two sources are present in one instant of time, both sources can be reconstructed by inverse filtering of the microphone signals using the known head related transfer functions (HRTF) for each detected direction. If more than two sound sources are present, this information is insufficient for an exact estimation of the original sources, and additional information is needed. However, the sound source direction can be used to reveal part of the dependencies between the source spectra and the observed two-microphone signal. Furthermore, this paper specifically investigates to what extent a statistical localization model can be used to gain information about the envelope of the sound sources.

2. Sound localization based on statistics of binaural parameters

Let $H_{\alpha,\theta,r}(f)$ be the head-related transfer function (HRTF) for the direction with azimuth α and elevation θ for the right ear, and $H_{\alpha,\theta,l}(f)$ the HRTF for the left ear, both assumed to be non-zero. Then

$$I_{(\alpha,\theta)}(f) = \frac{H_{(\alpha,\theta),r}(f)}{H_{(\alpha,\theta),l}(f)}$$

is the interaural transfer function for a sound signal from this direction. In the following we assume all spectra to be averaged over discrete frequency bands of half the equivalent rectangular bandwidth (ERB) of critical bands. Assuming $X(f)$ being the short time critical bands spectrum of a sound signal with non-zero power densities, interaural phase differences $\Delta\phi(f)$ and level differences $\Delta L(f)$ of $X(f)$ filtered by the left and right HRTF are the phase and level of $I_{(\alpha,\theta)}(f)$. Superposition of signals from different directions will result in fluctuating phase and level differences, thus $\Delta\phi(f)$ and $\Delta L(f)$ become random variables. We assume one directional sound source from a certain direction superposed with a perturbing noise field with a certain signal-to-noise ratio (SNR). Given the statistics of these binaural parameters, the sound source direction can be estimated by a Bayesian classification, as will be shown.

By grouping phase and level differences for each instant in time in a feature vector

$$\vec{x} = (\Delta\phi(1), \Delta\phi(2), \Delta\phi(3), \ldots \Delta L(1), \Delta L(2), \Delta L(3), \ldots)$$

a multidimensional probability density function $P(\vec{x})$ can be defined. Due to the large dimension, the distribution of this probability density function (PDF) cannot be estimated from observations, but the PDF of the components x_f of $\vec{x}$ can be estimated by observing histograms of the phase and level differences. These histograms represent the estimated marginal

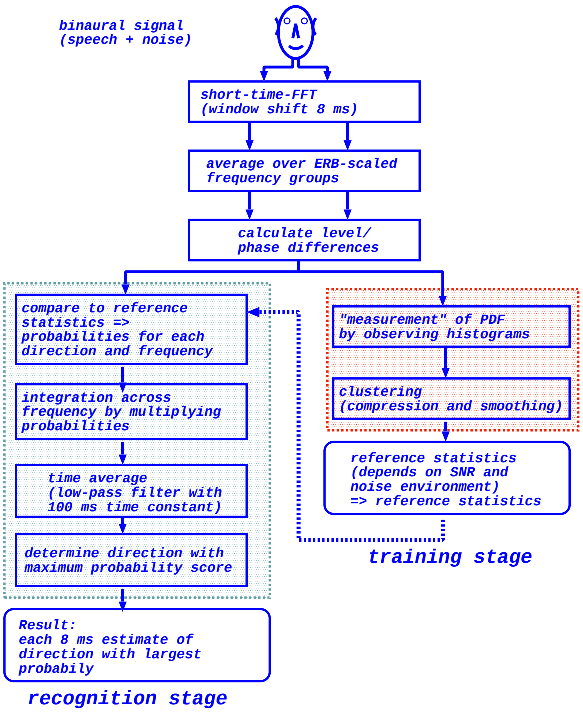

Figure 1: Block diagram Bayes-cluster-algorithm

distributions $\tilde{P}(x_f)$. From these, the estimate $\tilde{P}(\vec{x})$ can be approximated as the product of the marginal distributions by assuming independent random processes:

$$\tilde{P}(\vec{x}) = \prod_f \tilde{P}(x_f)$$

The estimated PDF $\tilde{P}(\vec{x})$ depends on azimuth and elevation of the target sound source. Now, given that the estimates $\tilde{P}_{\alpha,\theta}(\vec{x})$ for all directions $(\alpha, \theta))$ are known, and a feature vector from some source in the same noise field is observed, the probability for the presence of the direction $(\acute{\alpha}, \acute{\theta})$ can be calculated using Bayes' formula:

$$P((\acute{\alpha},\acute{\theta})) = \frac{\tilde{P}_{(\acute{\alpha},\acute{\theta})}(\vec{x})}{\sum_{(\alpha,\theta)} \tilde{P}_{(\alpha,\theta)}(\vec{x})}$$

The resulting probabilities for each direction are smoothed by a first-order low-pass filter with 200 ms time constant. The result of this smoothing will be referred to as "likelihood score" for this direction. Because the likelihood score is the probability for the presence of a source in this direction, its value is between 0 and 1. The direction $(\hat{\alpha}, \hat{\theta})$ with the highest likelihood score can be used as estimate of the direction of the most active sound source.

The a-priori parameters $\tilde{P}_{(\alpha,\theta)}(\vec{x})$ can be viewed as a type of "learned" references, the learning step being similar to the training of neural networks. A diagram of the complete algorithm is given in figure 1. The described scheme was implemented on a digital signal processor system described described by Wittkop et. al. [12]. In this implementation, the time signal was sampled at 25000 kHz. The time signal was subsequently analyzed using a 512 point FFT, yielding one short-time spectrum every 8 ms. A maximum frequency resolution of 0.5

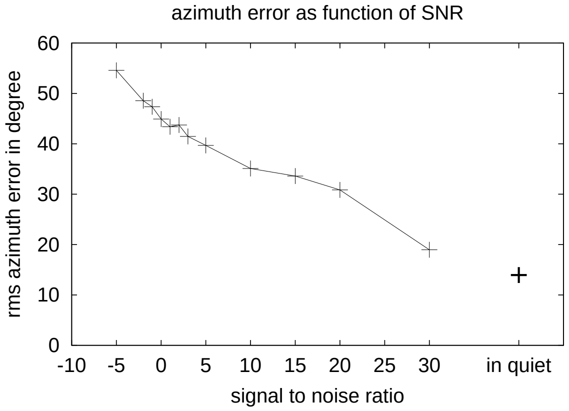

Figure 2: RMS errror of azimuth estimate as function of SNR

ERB was used by grouping the frequency axis in 43 bands between 50 Hz and 8kHz. To minimize front/back errors and to allow for estimation of the elevation along the "cones of confusion", which show largely symmetrical physical cues, a high spatial resolution is necessary, resulting in a large number of histograms. In order to reduce the amount of data, the resulting histograms were compressed and smoothed by a hierarchical WARD cluster analysis [6]. This compression allows for a reduction of histogram data by a factor of 70 and reduces the computational load by the factor 6.

3. Localization performance in noise

The Bayesian classification localization algorithm was tested in a number of different target/noise conditions. In this tests, the reference condition used to estimate the a-priori PDF were derived from recordings of a mixture of four speakers from 430 directions with each recording having a duration of 20 sec. As the noise, a 25 sec recording from a busy university cafeteria was used. All sound samples were captured with ITE hearing aids worn by the same human subject and recorded on DAT tape.

The signals used to test the localization performance were recordings of one male speaker from 96 directions. Each recording had 25 sec signal length, so that each data point in the resulting plots is derived from 120,000 direction estimates.

Figure 2 shows the RMS average of the azimuth error as a function of the SNR. This error measure increases from 11.7°at an SNR of 30 dB to35°at an SNR of 0 dB. Figure 3 shows the percentage of front/back confusions as a function of the SNR. It indicates that the percentage of front/back confusions increases from 24 % at high SNR to about 40 % at an SNR of -5 dB. Both results agree well with the performance range of human subjects for directional white noise as noise environment and a click train as target sound source, as reported by [5].

4. Generalization ability

Because the a-priori reference probabilities depend on the type of the noise field, the Bayesian classification will perform less than optimum when the references are applied to a different type of sound field. However, extensive tests had the result that the localization performance depends mostly on the SNR of the test situation, and that as long as the noise in the "learning" situation as well as the noise in the test situation is com-

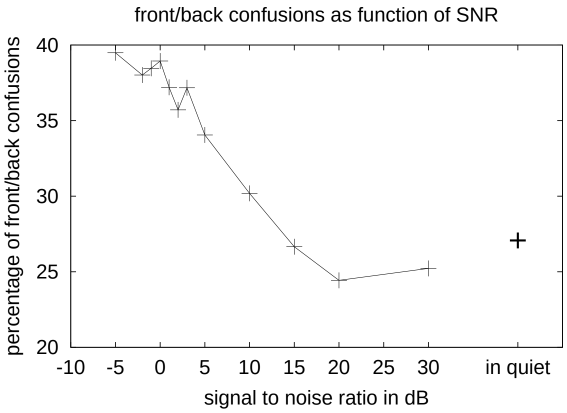

Figure 3: Percentage of front/back confusions as function of SNR

trained noise	C	S	M	T1	T2	A
tested noise						
Cafeteria	34.0	36.7	37.7	37.0	**43.7**	**47.1**
Station concourse	35.2	36.8	37.8	37.7	**42.0**	**49.8**
Metal workshop	36.3	38.4	37.8	36.7	**42.1**	**45.5**
Traffic 1	35.6	37.5	37.5	36.1	**43.5**	**46.5**
Traffic 2	**40.9**	**41.4**	**44.0**	**49.2**	34.0	**44.6**
Automobile (inside)	**33.3**	**35.0**	**35.3**	31.1	**35.3**	30.1

Table 1: Percentage of front/back confusions for different noise environments in training and test phase, SNR always 5 dB. Percentages which are higher than 5 % above the minimum of the row, are printed in boldface.

posed of some type of speech, performance is degraded only slightly [10]. An example for this result is given in table 1, which contains a matrix of noise conditions containing speech and other noise sources, for 36 training / test condition pairs. Numbers, in which the percentage of front back confusions is 3 % or higher above the minimum for this reference, are printed in boldface.

5. Localization of concurrent speakers

The localization algorithm yields two distinct maxima for the likelihood scores for two interfering speakers, even if the level of the speakers differ by 15 dB. For example, in a test series with one fixed speaker from five directions, and one speaker from 96 directions, the more intense speaker reached always a percentage of ‘most likely direction’ between 30 % and 50 %, and the percentage for the less dominant speaker is typically about 10 % to 20 % . Excluding front/back confusions and elevation errors, likelihood scores for non-active directions have been always less than 2 %. This property is due to the fact that by the spectral and temporal integration of the probability values, high likelihood scores for existing sources are reached as long as the sources do not overlap completely in the time and in the frequency domain.

6. Demixing two sound sources

Assuming $X_{r,f}$ is the short time spectrum of the right ear signal for the frequency band f, and $X_{l,f}$ the corresponding spectrum of the left ear signal, and $S_{f,n}$ are the n sound sources, then the filtering of the sound sources by the known HRTF $H_{f,r,n}$ and $H_{f,l,n}$ can be described as a linear operation:

$$\begin{pmatrix} X_{r,f} \\ X_{l,f} \end{pmatrix} = \begin{pmatrix} H_{f,r,1} & H_{f,r,2} & \dots & H_{f,r,N} \\ H_{f,l,1} & H_{f,l,2} & \dots & H_{f,l,N} \end{pmatrix} \begin{pmatrix} S_{f,1} \\ S_{f,2} \\ S_{f,3} \\ \vdots \\ S_{f,N} \end{pmatrix} \tag{1}$$

which can be written

$$\vec{x}_f = \mathbf{H_f}\vec{s}_f \tag{2}$$

This, if $\vec{X}_f$ and the sound source directions, and therefore $H_{f,l/r,n}$ are known, and N equals 2, the original sound sources can be demixed by matrix inversion of $\mathbf{H_f}$. Because the sound source direction changes only slowly in time, the directions can be estimated by the long-term average of the first two maxima of the likelihood score. In this case, demixing the sources is equivalent to a two-microphone adaptive beamformer, the time constant for the adaption being in the order of 1 s.

7. Correlations of envelopes and likelihood score

If $N > 2$, the inverse of the mixing matrix $\mathbf{H_f}$ is not well-defined. However, using additional restrictions, a solution with minimum error can be defined.

Potentially useful restrictions can be derived from statistical properties of speech. First, it is known that the spectral power densities of the short-time spectra are correlated in adjacent frequency bands [2, 1]. Second, the likelihood score of direction information is correlated to the envelope of the corresponding sound source. Both properties can be used to reduce the possible solutions for the inverse of $\mathbf{H_f}$.

The figures 4 and 5 show the envelopes of a mixture of two speakers at 0 dB on the top, the envelopes of one speaker (target or jammer) in the mid, and the logarithmized likelihood scores for the direction of each speaker at the bottom. The figures reveal that likelihood scores are correlated with the envelopes. The correlations values are 0.46 for the signal selected in figure 4, and 0.12 for the second signal. Smoothing by a low-pass filter increases the correlation.

8. Conclusions

- From binaural recordings, direction information can be extracted by Bayesian classification of interaural phase and level differences.
- The localization algorithm works well at moderate to low SNR. The performance of sound localization is comparable to human subjects.
- If trained and tested noise environment do not match, but both trained and tested environment contain speech noise, the localization performance degrades only slightly.
- For a two-source situation, demixing of the sound sources is possible by inversion of a matrix composed of the HRTF. Depending on the accuracy of the sampled HRTF, additional adaptation steps may be necessary.
- For more than two sound sources, the direction information is not sufficient to unmix the sources, and additional constraints are necessary.

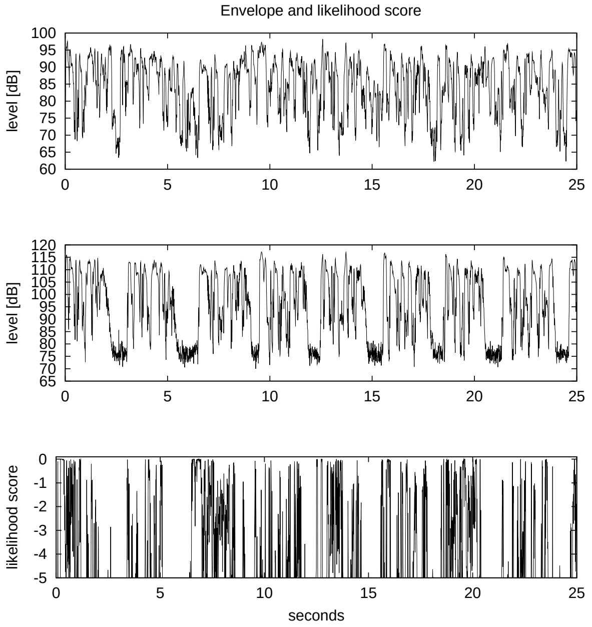

Figure 4: Envelopes of mixed signal (top) and target signal at -90° azimuth (mid), and log likelihood score for the target speaker in a two-speaker situation (bottom).

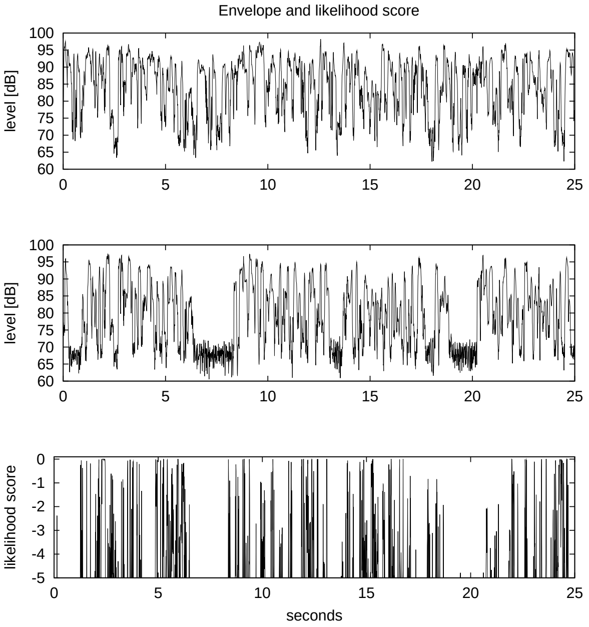

Figure 5: Envelopes of mixed signal (top) and noise signal at 155° azimuth (mid), and log likelihood score for the jammer in a two-speaker situation (bottom).

- The likelihood score from the Bayesian classification contains information concerning the envelopes of the component signals. This could be used e.g. for envelope filtering of the components signals.

9. References

[1] Jörn Anemüller. *Across-frequency processing in convolutive blind source separation.* PhD thesis, University of Oldenburg, Germany, July 2001.

[2] Jörn Anemüller and Birger Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In P. Pajunen and J. Karhunen, editors, *Proceedings of the second international workshop on independent component analysis and blind signal separation*, pages 215–220, 2000. http://www.physik.uni-oldenburg.de/Docs/medi/members/ane/pub.

[3] M. S. Datum, F. Palmieri, and A. Moiseff. An artificial neural-network for sound localization using binaural cues. *J. Acoust. Soc. Am.*, 100(1):372–383, June 1996.

[4] R. O. Duda. Elevation dependence of the interaural transfer function. In Robert H. Gilkey and Timothy R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 49–75. Lawrence Erlbaum Assoc., Mahwah, New Jersey, 1997.

[5] Michael D. Good and Robert H. Gilkey. Sound localization in noise: The effect of signal-to-noise ratio. *J. Acoust. Soc. Am.*, 99:1108–1117, February 1996.

[6] Bruno Kopp. Hierarchical classification III: Average-linkage, median, centroid, WARD, flexible strategy. *Biometrical J.*, 20(7/8):703–711, 1978.

[7] C. Neti, E. D. Young, and M. H. Schneider. Neural network models of sound localization based on directional filtering by the pinna. *J. Acoust. Soc. Am.*, 92(6):3140–3156, 1992.

[8] Johannes Nix and Volker Hohmann. Statistics of interaural parameters in real sound fields employing one directional sound source and its application to sound source localization. *J. Acoust. Soc. Am.*, –. in preparation, http://www.physik.uni-oldenburg.de/Docs/medi/members/jnix/.

[9] Johannes Nix and Volker Hohmann. Lokalisation im Störgeräusch auf der Basis der Statistik binauraler Parameter. In Albert Sill, editor, *Fortschritte der Akustik-DAGA'98*, pages 474–475, Oldenburg, 1998. DEGA (Deutsche Gesellschaft für Akustik e. V.), DEGA.

[10] Johannes Nix and Volker Hohmann. Statistics of binaural parameters and localization in noise. In Torsten Dau, Volker Hohmann, and Birger Kollmeier, editors, *Psychophysics, Physiology and Models of Hearing*, pages 263–266, Singapore, 1999. World Scientific Publishing Co.

[11] Johannes Nix and Volker Hohmann. Robuste Lokalisation im Störgeräusch auf der Basis statistischer Referenzen. In Albert Sill, editor, *Fortschritte der Akustik- DAGA'2000*, Oldenburg, 2000. DEGA (Deutsche Gesellschaft für Akustik e. V.).

[12] Thomas Wittkop, Stephan Albani, Volker Hohmann, Jürgen Peissig, William S. Woods, and Birger Kollmeier. Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. *Acustica united with acta acustica*, 83(4):684–699, 1997.