A simple algorithm for real-time decomposition of first order Ambisonics signals into sound objects controlled by eye gestures

GISO GRIMM[1,2], JOANNA LUBERADZKA[1], JANA MÜLLER[1] AND VOLKER HOHMANN[1,2]

[1]Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany
e-mail: g.grimm@uni-oldenburg.de
[2]HörTech gGmbH, Marie-Curie-Str. 2, D-26129 Oldenburg, Germany

September 23rd 2016.

Abstract

Spatial filtering and decomposition of sounds into acoustic source objects is increasingly investigated for speech enhancement in hearing aids. However, with increasing performance and availability of these 'space aware' hearing aid algorithms, knowledge of the user's personal listening preferences and knowledge of the attended source becomes crucial. Here we present a prototypical algorithm which decomposes a first order Ambisonics stream into multiple acoustic objects using space and frequency dependent object probability as a de-mixing gain. Eye gestures recorded by electrooculography are used to select the desired acoustic object for re-synthesis. The methods of scene decomposition and re-synthesis as well as the eye-gesture controlled object selection are described. The aim of this paper is to demonstrate the principle functioning and applicability of a decomposition algorithm with eye gesture control. The performance of object separation was assessed in simple and more realistic acoustic environments. Although further improvement of the decomposition algorithm and an increased robustness of the gesture detection are required for practical applications, the results indicate that the proposed eye-gesture based object decomposition has the potential of providing a benefit for hearing aid users.

1 Introduction

Current hearing aid technology employs advanced signal processing techniques for speech enhancement such as noise cancellation, dereverberation and directional filtering. Nevertheless, for a hearing aid user, speech communication in many of the common acoustic situations (i.e., the classical cocktail-party) remains challenging [3].

To overcome the noise problem, Computational Auditory Scene Analysis (CASA) algorithms are investigated, which extract information on the acoustic scene similarly to a human listener to build objects, i.e., separated information streams, based on the noisy acoustic input. One of the future goals of hearing-aid technology is to use these CASA algorithms to create space-aware hearing devices, which gather detailed information about the acoustic objects in the surrounding and resynthesize or enhance the desired object depending on the hearing wish. Together with a method to detect the hearing wish, this technology could lead to an interactive hearing aid, which provides the user with the chosen content of the acoustic scene in a way that minimises the detrimental effect of a hearing impairment.

The user's hearing wish should be estimated by the hearing device in an unobtrusive way, which doesn't require any external devices. Electrooculography (EOG), which extracts the information about the gaze direction, is a possible solution that has already been used for application in audiology [12] and human-computer-interfaces [15]. This contribution explores the possibilities of using a simplified CASA algorithm in combination with EOG-based control of the hearing wish for the design of a space-aware hearing aid.

2 Methods

2.1 CASA algorithm

A simple CASA algorithm is proposed, which decomposes an acoustic environment into multiple acoustic objects. This algorithm requires horizontal first order Ambisonics (FOA) input signals in the B-format in Furse-Malham normalisation, i.e., the w, x and y channels; the signal coherence and direction is analysed in time and frequency. This method is related to DirAC [13] and harpex [4]. However, the aim of those methods is the spatial up-sampling for reproduction of B-format content on playback systems with higher spatial resolution, whereas the proposed method aims at decomposing the acoustic environment into a small number of acoustic objects, for object-based audio coding. The discrete time-domain signals are transformed into the frequency domain using a short-time Fourier transform (STFT) with an overlap-add method [2]. The block size was 1024 samples at a sampling rate f_s of 44.1 kHz. The window length was set to 2048 samples, i.e., 50% overlap of blocks, and the FFT length L was 4096 samples, resulting in a frequency resolution of 10.8 Hz per frequency bin. The short-time input spectra are denoted with $W(k,\nu)$, $X(k,\nu)$ and $Y(k,\nu)$, respectively, with the time index k and the frequency index ν. The signal coherence is estimated by the absolute value of the estimation value of the normalised product of X with the complex conjugate $\overline{Y}$:

$$c(k,\nu) = \left| \left\langle \frac{X(k,\nu)\overline{Y(k,\nu)}}{\left|X(k,\nu)\overline{Y(k,\nu)}\right|} \right\rangle_{\tau} \right| \tag{1}$$

Temporal averaging is achieved here with a first-order IIR lowpass filter with time constant τ. The time constant τ was set to 40 ms, corresponding to a cut-off frequency of 4 Hz.

Only if a signal component is coherent, i.e., c is close to one, the direction of arrival estimation $\varphi = \angle\{X + jY\}$ is valid (plane wave decomposition). This approach is motivated by the finding that only components with a high coherence are relevant for perceptual localisation [11, 7]. However, opposite to the approach of [11], not a sparse selection of glimpses is used for characterisation, but the coherence c is taken as a continuous weighting function.

In a next step, the intensity of the coherence-weighted sound pressure, $I_{coh}(k,\nu) = |c(k,\nu)W(k,\nu)|^2$, is calculated. It is accumulated in a two-dimensional sliding histogram $I_{hist}(k,\varphi,b)$, where φ is the sampled azimuth, sampled in 24 bins, and b is the frequency band index of a logarithmic frequency scale $b_s(\nu) = 4\log_2(\frac{\nu}{125L/f_s})$ with four bands per octave in the frequency range from 125 Hz to 4 kHz; $b(\nu) = \text{floor}(b_s(\nu))$. For each frequency bin ν, the intensity was added to the intensity histogram at the frequency band $b(\nu)$ and at the linearly sampled estimated direction of arrival $\varphi(\nu)$. The forgetting time constant (first-order IIR lowpass per azimuth and frequency bin) of the histogram was set to 500 periods of the band centre frequency, limited to a maximum of 1 second for frequencies below 500 Hz. Since all the introduced measures depend on time, we simplify the notation by omitting the time index k.

With the non-linear frequency scale of the intensity histogram, sounds with pink-noise frequency characteristics result in an equal distribution of intensity across frequency bands. Bandlimited stimuli, such as speech or noise, can be roughly approximated by a Gaussian distribution across frequency bands. As a model function for a single sound object, the product of a raised cosine function as a function of azimuth φ with a Gaussian as a function of frequency bands b was used:

$$I_{mod,n}(\varphi,b) = A_n \left(\frac{1+\cos(\varphi-\varphi_n)}{2}\right)^{w_n} e^{-\frac{(b-b_n)^2}{2s_n^2}}, \tag{2}$$

where n is the object number, A_n is an object intensity coefficient, φ_n the object direction of arrival, w_n the source width, b_n the spectral centre, and s_n the spectral extension. These object parameters A_n, φ_n, w_n, b_n and s_n were estimated by minimising the squared difference between the modelled intensity and the intensity histogram, summed across all discrete directions and frequency bands:

$$e = \sum_{\varphi}\sum_{b}\left(I_{hist}(\varphi,b) - \sum_{n=1}^{N} I_{mod,n}(\varphi,b)\right)^2 \tag{3}$$

A gradient-search method with one iteration in each processing cycle was used. The number of modelled objects N was fixed, and set to 3 in this study.

For decomposition into the estimated object source signals, for each object n the FOA signal is sampled in the object direction φ_n using a $\max_{rE}$ decoder [6], corresponding to a hyper-cardioid microphone steered towards the estimated direction of arrival φ_n. Additionally, for each object, a de-mixing gain is computed in each frequency bin:

$$G_n(\nu) = \frac{I_{mod,n}(\varphi_n, b_s(\nu))}{\sum_{l=1}^{N} I_{mod,l}(\varphi_l, b_s(\nu))}, \tag{4}$$

where φ_n is the direction of the tested object n. Finally, the resulting output signal in the STFT-

representation is:

$$S_n(\nu) = (W(\nu) + \cos(\varphi_n)X(\nu) + \sin(\varphi_n)Y(\nu))G_n(\nu) \quad (5)$$

The $\max_{rE}$ weighting $\frac{1}{\sqrt{2}}$ is already part of the Furse-Malham normalisation and thus not explicitly given in Eq. 5. The time domain signal at the output of the CASA algorithm for object n is the re-synthesised STFT representation S_n.

Effectively, this algorithm consists of three steerable hyper-cardioid microphones combined with frequency weighting based on the model function.

2.2 EOG control

The human eye acts as an electrical dipole. The potential difference between the positively charged cornea and the negatively charged retina is in the order of 0.5 to 1 mV. In case of horizontal eye movements, the orientation of the dipole is changed, and the potential difference can be measured with a pair of simple electrodes at the left and right side of the head. The voltage is proportional to the gaze direction up to an angle of approximately 30 degrees [16].

For measurement of the EOG potential, a mobile measurement amplifier with a high input impedance and an amplification by a factor of 1000 was developed. The amplified signal was digitised with 10 bits word length and a sample rate of 50 Hz using an Arduino Nano board, and transferred to the signal processing computer as a Bluetooth serial stream. The device was battery driven to achieve a full electrical separation from any other electrical device.

The CASA object was selected via eye gestures. The EOG signal was high-pass filtered with a cut off frequency of 0.025 Hz to compensate for baseline drift. Baseline drift may be caused by interfering background, electrode polarisation or electrode contact [10] and appears as a low frequency signal change, which is not correlated with actual eye movements [5]. The eye gesture was detected each time the EOG potential exceeded a threshold of $\pm 350\mu$V for more than 500 ms. Eye gestures to the left switched to the next output stream on the left of the current output stream, and eye gestures to the right selected the next output stream on the right of the currently selected object.

2.3 Stimuli

All stimuli used in this study were virtual acoustic environments (VAEs), generated with a toolbox for acoustic scene creation and rendering (TASCAR) [9]. The signals were rendered to a virtual FOA microphone. VAEs from [8] were used, including early reflections, late reverberation, source and listener movement. The VAEs included an anechoic reference condition with a speech signal from the front and three interferers from 90, 180 and 270 degrees (SNR: -4.8 dB), and seven more realistic virtual acoustic environments, simulating a street (SNR: -4.7 dB), a supermarket (SNR: -4.1 dB), a conversation in a cafeteria (SNR: -2.7 dB), public announcements in a train station (SNR: 1.9 dB), a dialog in a kitchen (SNR: -1.7 dB), a monologue in a forest (SNR: -0.2 dB), and a panel discussion (SNR: 37.7 dB). Clearly localised sound sources, e.g., the target speech signal, were simulated based on anechoic recordings. Diffuse sounds, e.g., distant traffic or babble noise, were added as FOA recordings. Room acoustics were calculated using a geometric image source model for early reflections, and a feedback delay network [1, 14] for late reverberation. All environments contained a target speech signal and multiple spatially distributed interferers. The direction of the target speech signal differed across acoustic environments: In the reference condition, the 'street' and the 'kitchen' environment it was coming from the front. In the supermarket the target was in front of the virtual listener, however, the listener moved the head, so the environment was rotated according to the head movement. In the 'train station' the target was played back via a virtual announcement system above the listener, while the listener was walking on a platform. In the 'cafeteria' and 'panel discussion' the target was distributed across multiple alternating speakers in the frontal hemisphere. In the 'nature' environment the target was from 45 degrees to the right of the listener.

For the EOG controlled prototype, also the output signals of the CASA algorithm were synthesised in the direction of the estimated direction of arrival of the sound objects. The selected sound sources were played back to the listener via a 16 channel circular loudspeaker setup, using horizontal 7th order Ambisonics.

3 Experimental Results

For validation of the source separation performance, the SNR of each estimated CASA object was calculated and compared to a static front-facing directional microphone. If at least one CASA object has a higher SNR than the directional microphone, this algorithm can provide a benefit for the user if the optimal CASA object can be selected.

The SNR is shown in Figure 1. The input SNR

of the VAEs is shown as horizontal lines. The SNR of the three CASA objects is shown by vertical grey lines, with the digits 1 to 3. The length of the line denotes the SNR benefit (line pointing upwards) or detriment (line pointing downwards). The vertical dark grey line shows the SNR of the static directional microphone. Those conditions in which at least one object performed better than the directional microphone are indicated with a star.

The EOG control was tested only in a preliminary pilot experiment, to prove the general concept. Test subjects reported that the option of selecting streams increases subjective speech intelligibility in situations with concurrent talkers. They also reported that the identification of eye gestures needs further improvement, to avoid false alarm in case of head motion.

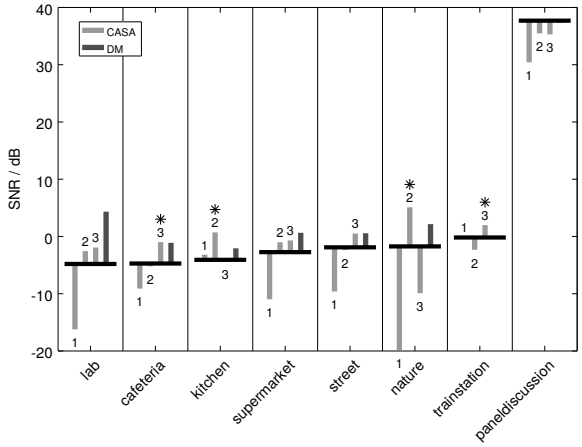

Figure 1: SNR analysis of the environments (horizontal lines) and the algorithm performance (vertical grey lines) as well as the performance of a static directional microphone (dark grey lines). A star indicates those conditions where at least one CASA object has a higher SNR than the directional microphone.

4 Conclusion

In the current paper, a CASA algorithm for a decomposition of an acoustic scene into objects and their re-synthesis was introduced. This algorithm was developed to demonstrate the general concept of object-based scene decomposition and re-mixing with eye control as a hearing aid algorithm. An analysis of the SNR indicates that the algorithm could be beneficial in some conditions, where at least one CASA object showed a higher SNR than a static directional microphone, as commonly used in hearing aids.

Eye gestures recorded by electrooculography can serve for selection of the desired acoustic stream. Typical hearing aids do not provide Ambisonics signals, however, by combining the directional microphones of the left and right hearing aid a first-order Ambisonics signal can be generated at least in the lower frequency range, which might be sufficient for hearing aid users. Implementing an improved version of the algorithm on a hearing device could be beneficial for a hearing aid user.

Acknowledgements

Work funded by DFG FOR1732 “Individualized Hearing Acoustics”.

References

[1] Fons Adriaensen. Rev1 quickguide. http://kokkinizita.linuxaudio.org/linuxaudio/zita-rev1-doc/quickguide.html, 2015. Accessed August 4, 2015.

[2] Jont B. Allen. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, 25(3):235–238, June 1977.

[3] Ruth A. Bentler. Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, 16(7):473–484, 2005.

[4] Svein Berge and Natasha Barrett. High angular resolution planewave expansion. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics May*, 2010.

[5] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. Wearable eog goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments*, 1(2):157–171, 2009.

[6] Jérôme Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scnes sonores complexes dans un contexte multimdia*. PhD thesis, Université Pierre et Marie Curie (Paris VI), Paris, 2001.

[7] Mathias Dietz, Stephan D. Ewert, and Volker Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592 – 605, 2011.

[8] Giso Grimm, Birger Kollmeier, and Volker Hohmann. Spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluation. *Journal of the American Academy of Audiology*, 2016. in press.

[9] Giso Grimm, Joanna Luberadzka, Tobias Herzke, and Volker Hohmann. Toolbox for acoustic scene creation and rendering (tascar): Render methods and research applications. In Frank Neumann, editor, *Proceedings of the Linux Audio Conference*, Mainz, Germany, 2015. Johannes-Gutenberg Universität Mainz.

[10] Jason Jianjun Gu, Max Meng, Albert Cook, and M Gary Faulkner. A study of natural eye movement detection and ocular implant movement control using processed eog signals. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 2, pages 1555–1560. IEEE, 2001.

[11] Angela Josupeit, Norbert Kopčo, and Volker Hohmann. Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. *J. Acoust. Soc. Am.*, 139(5):2911–2923, 2016.

[12] Jana A. Müller, Dorothea Wendt, Birger Kollmeier, and Thomas Brand. Comparing eye tracking with electrooculography for measuring individual sentence comprehension duration. *PLOS ONE*, 2016. submitted.

[13] Ville Pulkki. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc*, 55(6):503–516, 2007.

[14] Torben Wendt, Steven van der Par, and Stephan D. Ewert. Perceptual and room acoustical evaluation of a computational efficient binaural room impulse response simulation method. In *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, Berlin, 2014.

[15] Shang-Lin Wu, Lun-De Liao, Shao-Wei Lu, Wei-Ling Jiang, Shi-An Chen, and Chin-Teng Lin. Controlling a human–computer interface system with a novel classification method that uses electrooculography signals. *IEEE transactions on Biomedical Engineering*, 60(8):2133–2141, 2013.

[16] Laurence R Young and David Sheena. Survey of eye movement recording methods. *Behavior research methods & instrumentation*, 7(5):397–429, 1975.