Assessing and modeling apparent source width perception

Johannes Käsbach, Manuel Hahmann, Tobias May and Torsten Dau

Centre for Applied Hearing Research, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, Email: johk@elektro.dtu.dk

Introduction

Spatial perception is a primary function of the human auditory system. It is essential in decoding the auditory scene surrounding a listener. Each sound source in such a scene has a certain location and distance with respect to the listener. This spatial separation helps the listener in distinguishing concurrent sources from each other, e.g. a target speaker from interfering noise sources. The perceived horizontal extent of sound sources is typically described by the apparent source width (ASW). A reduced sensitivity to ASW as e.g. found in hearing-impaired listeners [6] may have consequences on the ability of spatially separating sound sources. Therefore, it is important to understand the contributing cues to ASW perception. According to literature, three binaural cues are mainly contributing to ASW: The interaural time differences (ITDs) and the interaural level differences (ILDs), that are as well important for determining the location of a sound source in the horizontal plane, and the interaural coherence (IC). Due to reflections in rooms and from the head and torso of the listener all three cues fluctuate over time. With increasing amount of room reflections, the IC decreases and larger variations in ITDs and ILDs occur, leading to an increased ASW. The psychophysical relation between these three binaural cues and ASW can be exploited by binaural auditory models.

Traditional models of ASW are used to evaluate the quality of concert halls by analyzing the interaural cross-correlation (IACC) function [1]. Based on the IACC, the interaural coherence (IC) is extracted as the absolute maximum value normalized by the root-mean-square (RMS) value of the left and right ear-signal. Hereby, an inverse relation between IC and ASW exists. Okano et al. (1995) [9] proposed a frequency-specific weighting of the IC, termed $\mathrm{IACC_{E3}}$ that averages the IC in three octave bands 0.5, 1 and 2 kHz. The $\mathrm{IACC_{E3}}$ is calculated for the first 80 ms of the binaural impulse recordings (BRIRs) since early reflections are known to contribute mostly to ASW [4]. Zotter et al. (2013) [13] observed a high correlation of $r = 0.97$ between the $\mathrm{IACC_{E3}}$ and perceptual data obtained in a stereo loudspeaker measurement setup.

Similar ideas as suggested by Okano were implemented in a complex binaural auditory model by van Dorp Schuitman et al. (2013) [10] which splits the input signal in a direct and a reverberant stream. From the direct stream the model extracts ITDs up to 2 kHz as the time-lag at the maximum IC and estimates the ASW by averaging their standard deviation. In contrast to the traditional IC-based measures, this model is applied on binaural recordings. The model showed higher correlations with perceptual data compared to the $\mathrm{IACC_{E3}}$. Note that both studies, Okano et al. and van Dorp Schuitman et al. further considered the monaural sound pressure level (SPL) as an additional cue for ASW.

Blauert and Lindemann (1986) [3] suggested that both, ITD and ILD fluctuations, contribute to ASW. They combined the standard deviation of both cues with equal weights and reported a higher correlation with perceptual data ($r = 0.75$) as opposed to an IC-based model ($r = 0.61$). Later Mason et al. (2005) [7] developed an ASW model that combined both ITDs and ILDs according to the duplex theory by using ITDs at low frequencies and ILDs at high frequencies [8] plus including loudness information.

Obviously, several models of ASW have been suggested in literature, each validated on different perceptual datasets. The present study investigates generalizability of the introduced models by evaluating their performance across datasets. Hereby, it will be addressed whether (i) a correlation-based approach, i.e. either extracting IC or ITD as suggested by Okano et al. and van Dorp Schuitman et al., respectively, is sufficient for the estimation of ASW, (ii) the in literature suggested frequency region of up to 2 kHz is optimal in such approach or if high-frequency ICs or ITDs are contributing to ASW as well and (iii) a model combining ITDs and ILDs as suggested by Blauert and Lindemann and Mason et al. is feasible. The models have been validated on two experimental datasets presented in Käsbach et al. 2014 and 2015 [5],[6].

Summary of the perceptual studies

Two perceptual studies were conducted to measure ASW ([5] and [6]), in the following referred to as Experiment A and B, respectively. Distinct sensations of ASW were generated by using stereo loudspeaker setups. In such a setup the listener perceives a phantom sound image in the center of the two loudspeakers. The ASW was measured as a function of the physical source width (PSW) which was controlled by two experiment-specific settings, the loudspeaker layout and applied signal processing. In the measurement procedure listeners indicated the perceived ASW on a degree scale as illustrated in Figure 1. Note that in Experiment B, listeners could indicate the left and right most boundary of the sound source separately, whereas in Experiment A, the response had to be given symmetrically. In the present study only 3 source signals per experiment will be used.

In Experiment A, the stereo setup at an angle of ± 30 degrees was used indicated by the red dashed rectangles in Figure 1. Five distinct PSW values, denoted by PSW #1 to PSW #5, were generated by varying the coherence between the two loudspeaker channels accordingly

to $IC_{LS} = 1, 0.8, 0.6, 0.3$ and 0. The source signal was either Gaussian white noise, band-pass filtered with a bandwidth of 2 octaves at a center frequency of 0.25 kHz or high-pass (HP) filtered at 8 kHz. The stimuli had a duration of 4 s and were presented at 70 dB SPL.
In Experiment B, the PSW was controlled by varying the angle between the stereo speakers. In addition, a source widening algorithm was applied as described in Zotter et al. (2013) [13]. Specifically, a line-array of 3 stereo loudspeaker pairs (Type Dynaudio BM6) plus an additional loudspeaker in the center of the array was used as indicated by the grey rectangles in Figure 1. In total, five distinct PSW values were generated. The source signals were pink noise, male speech and a guitar sample. The stimuli had a duration of 6 s and were presented at 70 dB SPL.
In Figure 2, the perceived ASW as a function of PSW averaged across listeners is shown for Experiment A (left panel) and Experiment B (right panel) . The error bars represent the standard deviation across listeners. It can be seen that ASW increases with increasing PSW. In Experiment A (left panel), the different signal types (represented by the different symbols and linestyles) show similar results with a tendency that the bandpass-filtered signal at 250 Hz and the white noise signal were perceived with larger ASW than the HP filtered signal at 8 kHz. In a statistical analysis with a mixed model the factor PSW showed a similar effect size ($F = 113.6, p < 0.001$) compared to the factor source signal ($F = 97.2, p < 0.001$). In Experiment B (right panel), it can be seen that ASW increases as well with PSW in a similar manner to Experiment A. Small differences are present between the source signals, such that the noise source is perceived generally with larger ASW compared to speech and guitar. In a statistical analysis with a mixed model the factor PSW showed a dominating effect size ($F = 114.8, p < 0.001$) compared to the non-significant factor source signal ($F = 3.9, p = 0.06$).

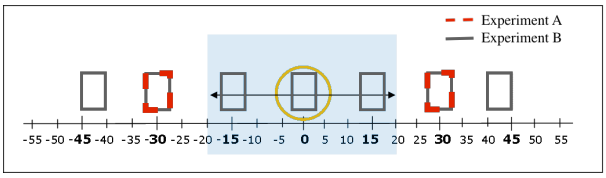

Figure 1: Sketch of the experimental set-up. The loudspeaker pairs generate a phantom source at 0 degree. Listeners were asked to indicate the ASW in degree, for both boundaries of the source image.

The ASW model

Binaural recordings were obtained with a head and torso simulator (HATS) that was placed at the listener's position. The functional model consisted of various processing stages, including gammatone filtering, inner hair-cell transduction (IHC) and absolute threshold of hearing (ATH). Given the binaural signal, the model extracted ITDs, ILDs and IC, in order to predict ASW.

Front-end

The auditory processing was based on the auditory-front-end (AFE) developed by the TWO!EARS consortium [12]. The binaural signals were first analyzed by a gammatone filterbank to represent the frequency selectivity of the basilar membrane. The 39 filters were set to a bandwidth of 1 equivalent rectangular bandwidth (ERB) in the frequency range between 80 to 20000 Hz. In the second stage, IHC transduction was simulated, i.e. the loss of phase locking to the stimulus' fine structure at high frequencies. The IHC processing was performed according to Bernstein et al. (1999) [2], suggesting a cut-off frequency of 425 Hz and also simulating basilar membrane compression. In a following stage, the activity in each frequency band was estimated. The signals had been calibrated to a root-mean-square (RMS) value corresponding to the 70 dB SPL of the experimental stimuli. Frequency bands below the ATH as defined in Terhardt (1979) [11] were excluded from further processing. In the last stage, ITDs, ILDs and ICs were calculated per time-frequency units. The signals of both ears were analysed in short time hanning windows of 20 ms duration with an overlap of 50 % which resulted in a time-frequency representation of each ear signal. The IC and ITD were extracted from the normalised interaural cross-correlation function per time-frame. The IC was equal to the maximal coherence and the ITD corresponded to the time-lag at this value. Time-lags were limited to a range of ± 1.1 ms. The ILDs were defined as the energy difference in dB between the two ear signals.

Back-end

In the model's back-end the ASW estimation was based on the variability of the binaural cues that increases due to increasing room reflections which is leading to a larger ASW accordingly. The variability of binaural cues was estimated by percentiles, capturing the width of the cue's statistical distribution, where the 20 % percentile corresponded to the left most boundary and the 80 % percentile to the right most boundary of the sound source. The first back-end, termed DUPLEX, combined ITDs and ILDs according to the duplex theory [8] motivated by Blauert and Lindemann and Mason et al.. The time-frequency representations of ITDs and ILDs were used to estimate their fluctuations in each frequency band, separately. In order to combine ITDs and ILDs, the corresponding binaural cues were normalized to the overall maximal value observed for the presented stimuli in the chosen percentiles, i.e. 1.1 ms for ITDs and 12 dB SPL for ILDs. ITDs and ILDs were combined across frequency in accordance with the duplex theory using ITDs up to 1.5 kHz and ILDs above this frequency limit. The final ASW prediction was then obtained by calculating the mean value across all frequency channels. In a second back-end, termed ITD_{low}, only the ITD-percentiles were analysed with a upper frequency limit of 2 kHz according to van Dorp Schuitman. The third back-end used the IC for the ASW prediction, termed IC_{E3}, resembling a short-term analysis of $IACC_{E3}$. In total 16 gammatone

filters of the freont-end were selected corresponding to the frequency range between 0.35 to 2.83 kHz defined by the octave wide filters in $IACC_{E3}$ at 0.5, 1 and 2 kHz. The frame-based values of IC were averaged with equal weights across all frames and frequency channels. As a reference model dealt the $IACC_{E3}$ according to Okano et al. in a long-term analysis.

Calibration of the model

A calibration stage was required in order to map the output of the model to ASW in degrees. A linear fitting approach was chosen here that allowed for a sensitivity parameter a and an offset b such that the calibrated model output was $y_{cal} = ay + b$, where y represents the uncalibrated model output. The calibration was performed using two data points measured with the white noise stimulus of Experiment A for PSW #1 and PSW #5.

Modeling results and discussion

The individual model performance was accessed by calculating Pearson's correlation coefficient r^2 and the RMS-error between the calibrated model outputs and all experimental data, i.e. for Experiment A and B together including all source signals and conditions. The corresponding values are displayed in Table 1. In general, all models provide a very high correlation with the perceptual data (ranging from $r^2 = 0.8$ to $r^2 = 0.97$). This is due to the fact that PSW is the dominating factor compared to the source stimulus which is captured correctly by all models. Analysing r^2 across source signals for single PSW values did not show any differences between the models' performance.

Table 1: Model performances in terms of correlation coefficient r^2, r and the RMS-error.

Model	r^2	r	RMS-error [°]
$IACC_{E3}$	0.97	0.98	3.87
IC_{E3}	0.91	0.95	10.5
IC_{full}	0.86	0.93	15.87
ITD_{low}	0.90	0.95	10.7
ITD_{full}	0.80	0.89	16.93
DUPLEX	0.89	0.94	11.25

In Figure 3, the outputs of the four tested models, $IACC_{E3}$, IC_{E3}, ITD_{low} and DUPLEX are presented for Expermient A (left panels) and for Experiment B (right panels). Note that the first two models are inversive proportional to ASW and are therefore shown as $1 - IACC_{E3}$ and $1 - IC_{E3}$, respectively. Further, both models produced a single output value and are therefore plotted with a symmetric ASW. It can be seen that all models are able to predict the general trend of the data, i.e. that the perceived ASW increases with PSW. Differences occur mainly in the slopes of the predicted boundaries of ASW and between source signals. The $IACC_{E3}$ is shown on the top panels of Figure 3. The model achieves the highest correlation of

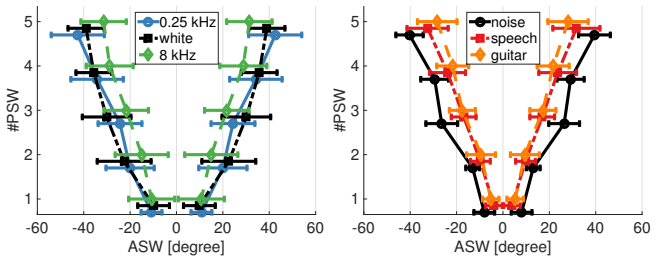

Figure 2: Perceptual results of ASW for Experiment A (left panel) and Experiment B (right panel) in degrees. ASW is shown as a function of the physical source width (PSW), denoted by PSW #1 (narrow) to #5 (wide). Plotted are the mean and standard deviation. The different symbols and linestyles represent the different source signals.

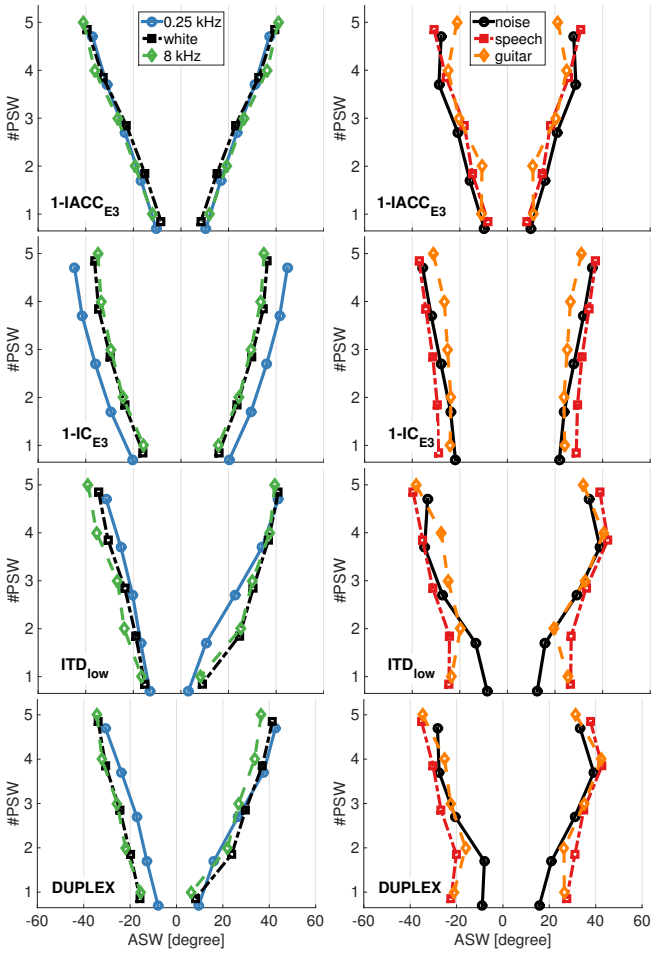

Figure 3: Modeling results of ASW for Experiment A (left panels) and Experiment B (right panels) in degrees. From top to bottom: $1-IACC_{E3}$, $1-IC_{E3}$, ITD_{low} and DUPLEX. ASW is shown as a function of the physical source width (PSW), denoted by PSW #1 (narrow) to #5 (wide). The different symbols and linestyles represent the different source signals.

the considered models with $r^2 = 0.97$ ($r = 0.98$ which is identical to findings in [13]) due to the fact that it captures correctly the dynamic range in ASW, i.e. the difference between smallest and largest ASW, for both Experiments. However, the model does not capture the increase in ASW for PSW #5 in Experiment B (right top panel) and does only reveal minor differences in the source signals. Considering the model denoted by IC_{E3},

the performance decreases to $r^2 = 0.91$. This indicates that a short-term analysis of IC and a higher frequency resolution (16 gammatone filters as opposed to 3 octave-wide filters in $IACC_{E3}$) are not required to predict the perceptual data considered in this study. Its output is shown in the second row of Figure 3 and it becomes prominent that this model has a reduced sensitivity, i.e. a more shallow slope of the boundaries. It captures minor source signal differences which however are contradicting the data. The ITD_{low} model is shown in the third row of Figure 3. Note that the model output is more asymmetric due to the fact that the boundaries are estimated separately by the corresponding percentiles. Its performance is with $r^2 = 0.90$ similar to the IC_{E3} model. Since both models, IC_{E3} and ITD_{low} are based on the IACC this result is reasonable. The output of the ITD model deviates though from the IC_{E3} model. Prediction errors are here found due to the models asymmetric output (potentially caused by asymmtric HATS positioning) and an overestimation in case of the speech and guitar source signal in Experiment B (right panel in the third row) for low PSW values. In Table 1, the performance of both models is shown as well when including the entire bandwidth for the analsyis, i.e. including envelope information. The models are denoted as IC_{full} and ITD_{full} and show with $r^2 = 0.86$ and $r^2 = 0.80$, respectively, a decreased performance compared to their low frequency estimates. This suggests that high frequency components in IACC-based measures do not provide useful information for ASW. The DUPLEX model, shown on the bottom of Figure 3 provides an identical output and performance as the ITD_{low} model. Therefore, adding ILDs in the analysis did not provide a further benefit. An analysis of the ILD percentiles showed that the used stimuli provided a small dynamic range of ILD fluctuations, i.e. the difference for PSW #1 and PSW #5, was less than 1 dB. This is small considering that the average fluctuations of ILDs were around 4 to 5 dB and maxima occurred at 12 dB. Even though the usage of ILDs in the analysis cannot be justified for the considered stationary stimuli (even for the speech and music signal the variations across PSW were small), ILDs might provide a larger dynamic range in real rooms and hence might become more relevant for the ASW estimation.

Summary and conclusions

In this study, two experiments have been presented where the ASW has been measured as a function of the PSW. The stimuli were also analysed by four binaural functional models to predict ASW. Hereby, a model that combines ITDs and ILDs according to the duplex theory has been developed (DUPLEX) and compared to other existing approaches in literature, i.e. $IACC_{E3}$, IC_{E3}, and ITD_{low}. Comparing model performances by means of r^2 it can be concluded that (i) models based on the interaural cross-correlation function, i.e. either extracting IC or ITD, produce equivalent results for the estimation of ASW. Hereby, the best performance was obtained by a long-term analysis of the binaural signals with $IACC_{E3}$. (ii) The correlation-based models operate thereby in their optimal frequency range and adding higher frequency components, here envelope ICs or ITDs, deteriorated the ASW estimation. (iii) The DUPLEX model including also ILDs could not provide any further benefit in the ASW estimation possibly due to the stationary character of the chosen stimuli.

Acknowledgement

The Centre for Applied Hearing Research is supported by a consortium of Oticon, Widex and GNResound. The binaural models were implemented together with Manuel Hahmann and are based on the auditory front-end (AFE) of the TWO!EARS consortium [12].

References

[1] Ando, Y. (2007): Concert hall acoustics based on subjective preference theory. The Springer Handbook of Acoustics (Springer Science + Business Media, New York), pp. 351-386.

[2] Bernstein, L., R., van de Par, S. and Trahiotis, C. (1999): The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise. J. Acoust. Soc. Am. 106 (2), pp. 870-876.

[3] Blauert, J. and Lindemann, W. (1986): Auditory spaciousness: Some further psychoacoustic analyses. J. Acoust.Soc. Am. 80 (2), pp. 533-542.

[4] Bradley, J. S. (2011): Review of objective room acoustics measures and future needs. Elsevier Applied Acoustics 72, 713-720.

[5] Käsbach, J., May, T., Le Goff, N. and Dau, T. (2014): The importance of binaural cues for the perception of ASW at different sound pressure levels. DAGA, Oldenburg.

[6] Käsbach, J., Wiinberg, A., May, T., Løve Jepsen, M. and Dau, T. (2015): Apparent source width perception in normal-hearing, hearing-impaired and aided listeners. DAGA, Nürnberg.

[7] Mason, R., Brookes, T., Rumsey, F. and Neher, T. (2005): Perceptually motivated measurement of spatial sound attributes for audio-based information systems. EPSRC Project Reference: GR/R55528/01.
http://iosr.uk/projects/PMMP/index.php

[8] Macpherson, E. A. and Middlebrooks, J. C. (2002): Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited, J. Acoust. Soc. Am. 111(5), pp. 2219-2236.

[9] Okano, T., Beranek, L. L., Hidaka, T. (1995): Interaural cross-correlation, lateral fraction, and low- and high- frequency sound levels as measures of acoustical quality in concert halls. J. Acoust. Soc. Am. 98 (2), pp. 255-265.

[10] van Dorp Schuitman, J., de Vries, D., Lindau, A. (2013): Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. J. Acoust. Soc. Am. 133 (3), pp. 1572-1585.

[11] Terhardt, E. (1979): "Calculating virtual pitch. Hear. Res. Vol. 1, pp. 155-182.

[12] TWO!EARS Consortium (2013-2016): A computational framework for modelling active exploratory listening that assigns meaning to auditory scenes. EU-Project, no. 618075, Coordinator: Prof. Dr. Alexander Raake, TU Berlin.
http://twoears.aipa.tu-berlin.de

[13] Zotter, F., Frank, M. (2013): Efficient phantom source widening. Archives of Acoustics, Vol. 38, No.1, pp. 27-37.