Modeling the Integration of Audio-Visual Distance Information

Ľuboš Hládek[1*], Aaron R Seitz[2], Norbert Kopčo[1]

[1] Institute of Computer Science, P. J. Šafárik University in Košice, Slovakia

[2] Department of Psychology, University of California, Riverside, USA

Introduction

Auditory and visual stimuli can fuse into a single percept even if each of them is presented from a different location. We recently examined this illusion, called the ventriloquism effect (VE), in the distance dimension (Hládek et al., 2013; Hládek et al., 2014). Typically, the fused object's perceived location is near the visual signal location, even if the subject is instructed to localize the sound

Previously we observed that VE in the distance dimension was influenced by the reference distance and direction of the AV disparity despite the fixed value of the AV disparity (Hladek et al., 2013; Hládek et al., 2014). (Figure 1, Figure 2).

The underlying model likely involves statistically optimal integration of the inner representations of the audio and visual information (Alais, Burr, 2004) formulated as Maximum likelihood estimation (MLE) integration (Ernst, Banks, 2002). However, auditory (also visual) distance perception is known to be highly compressed (Zahorik, 2005) which could be related to the prior information which also enters the integration process.

Körding et al., (2007) showed that the VE in the horizontal plane does not always follow the predictions of the optimal integration especially at the high magnitudes of the AV disparity because then the integration of the stimulus starts to fall apart. Their experiments suggested that the transition from the full integration to full segregation is not necessarily linear process (linear combination of the two distributions) but is influenced by the process in which the subjects are trying to infer the cause of the event (whether the sound and light were produced by the same event or produced by two separate events).

In the VE experiments in distance (Hladek et al., 2013; Hládek et al., 2014) the disparity was fixed on logarithmic scale, however, the AV disparity was increasing with distance on the linear scale which may have caused that the subjects perceived two stimuli nonunified, i.e., segregated.

We examined whether cue combination models are sufficient to explain the observed data or whether casual inference needs to be considered.

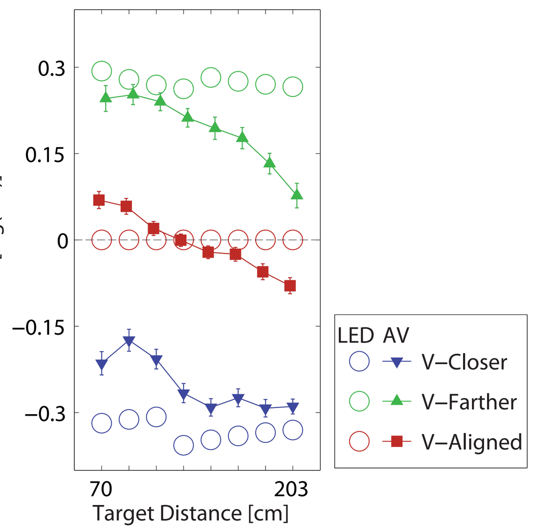


Figure 1 Localization bias of AV stimuli as a function of the target distance. Responses are compressed and biased in the direction of visual component (circles). Circles show the amount of the actual AV disparity.

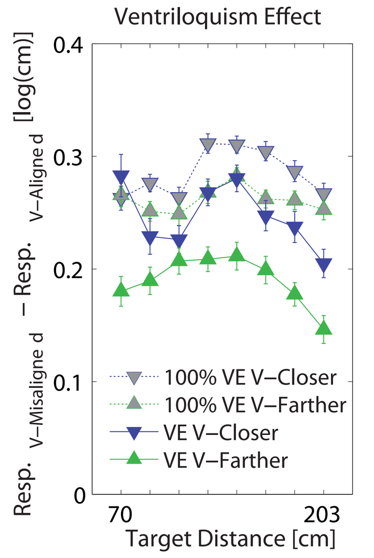


Figure 2 Magnitude of the VE was defined as the amount of bias due to the AV disparity (re. V-Aligned). Gray lines show the theoretical positions extrapolated from the V-Aligned data.

What is the underlying mechanism of the AV integration in distance dimension?

Methods

Experiment

- the experiment was run in a small semi-reverberant room (T_{60} =408 ms)
- the subjects (n=114) were seated in front of 8 linearly spaced loudspeakers with LED lights above the array positioned in the median plane of the subject (LEDs served for presentation and collection of responses)
- the task was to localize sounds (300 ms noise bursts) in distance dimension presented along with the flashes of the LED lights with the fixed AV disparity from one pseudo randomly chosen loudspeaker. The lights were always approximately 30% closer or farther.
- subjects performed 2 experimental sessions (V-Closer+V-Farther),(V-Closer+V-Aligned), or (V-Farther+V-Aligned) with counterbalanced order

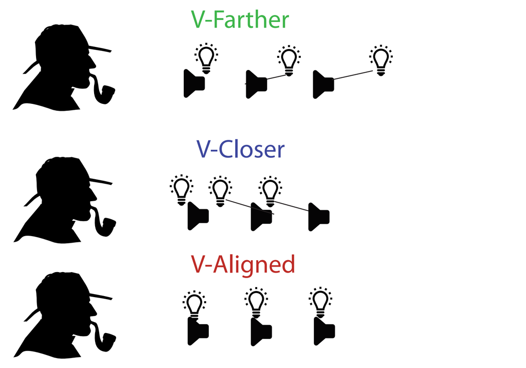
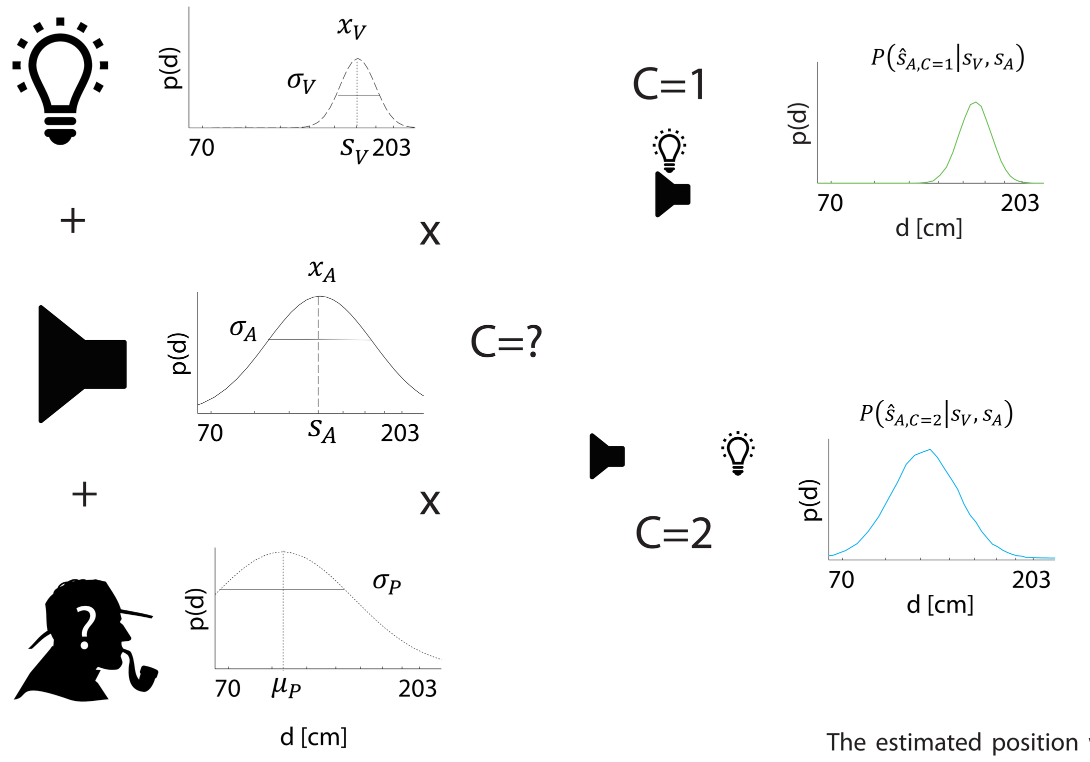


Figure 3 Schematics of the relative positions of the subject, target sounds, and LED lights in all three conditions. (Upper) V-Farther (middle) V-Closer (bottom) V-Aligned (zero AV disparity).

Model Structure



$$\hat{s}_{A,C=1} = \frac{\frac{x_V}{\sigma_V^2} + \frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}}$$

$$\hat{s}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}}$$

Figure 4 The model structure. Inputs of the model are $x_V = N(s_V,\sigma_V)$ and $x_A = N(s_A,\sigma_A)$ the noisy inner representations (measurements) of the visual and auditory components. C=? means that the underlying cause is not known to the observer. If they have common cause (integrated) (C=1) they are optimally combined together with prior distribution. If the sound is perceived with a independent cause (C=2) only auditory and prior functions are optimally combined.

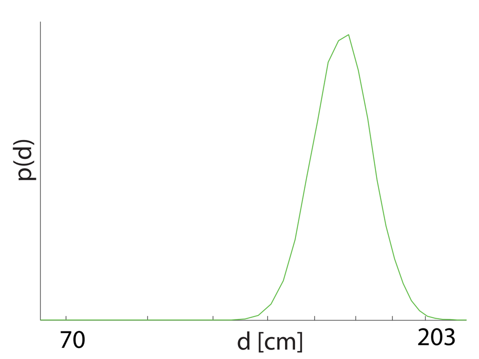
The estimated position was computed as an the average of the two possible causes.

$$\hat{s}_A = P(C = 1|x_V, x_A)\hat{s}_{A,C=1} + (1 - P(C = 1|x_V, x_A))\hat{s}_{A,C=2}$$

Model histograms were obtained by Monte Carlo simulations. The parameters of the model histograms were fit to subject histograms (multinomial distribution) by maximizing the likelihood function.

Models

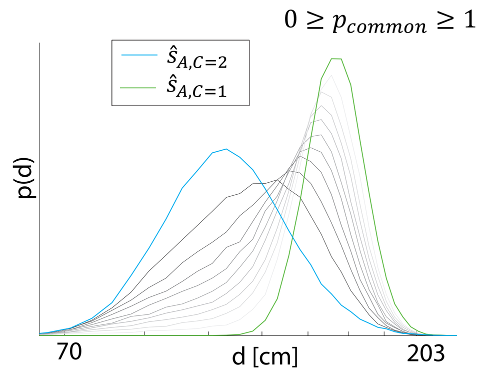
MLE



$$P(C = 1|x_V, x_A) = 1$$

Figure 5 MLE is the forced fusion model. d = Target distance.

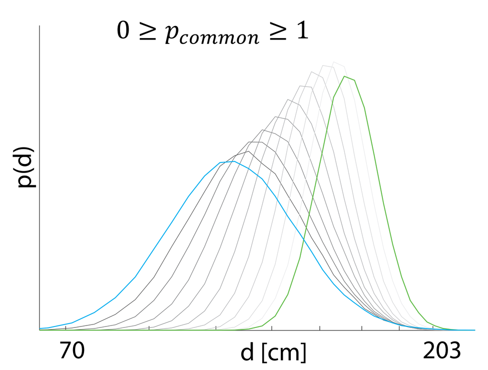
Causal Inference



$$P(C = 1|x_V, x_A) = \frac{P(x_A, x_V|C = 1)p_{common}}{P(x_A, x_V|C = 1)p_{common} + P(x_A, x_V|C = 2)(1 - p_{common})}$$

Figure 6 The causal structure depends on the actual measurements and the noise of the underlying distributions. $P(x_A,x_V|C=1)$ is a Gaussian distribution of the integrated percept and $P(x_A,x_V|C=2)$ distribution of the segregated percept. p_{common} is parameter of the distrubition.

Linear Combination



$$P(C = 1|x_V, x_A) = p_{common}$$

Figure 7 Alternative model to Causal Inference. The resulting distribution is only a linear combination of the two extreme cases. p_{common} is parameter of the distribition.

Preliminary Analysis

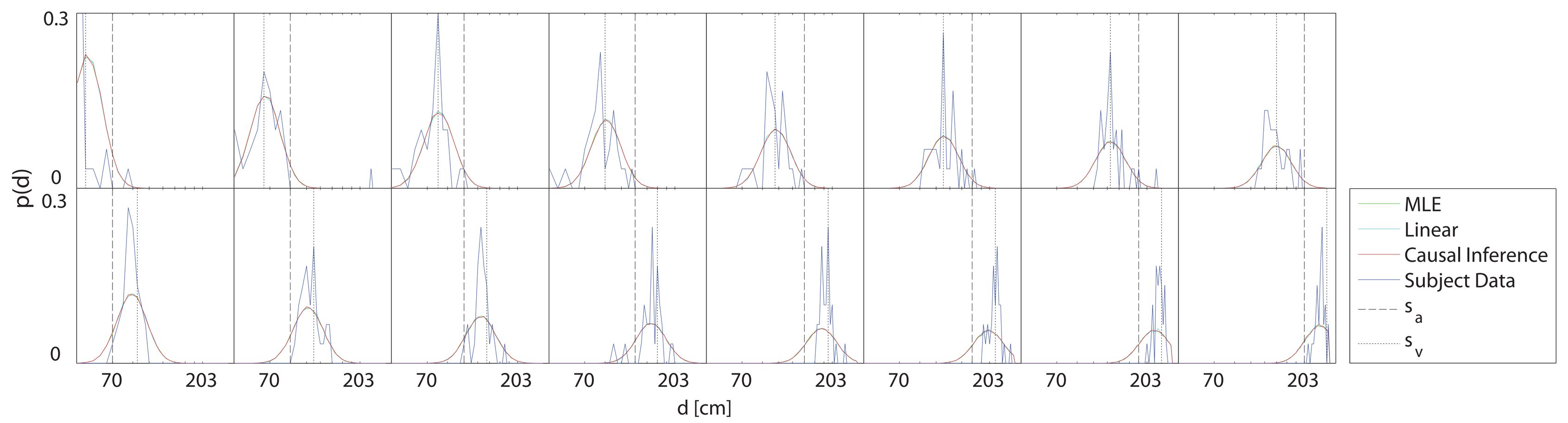


Figure 8 An example of the fitting procedure for one subject in V-Closer (upper row) and V-Farther (bottom row) conditions. Data show response histograms (blue) and model fits for each target location (columns) and two conditions of AV disparity (rows) on log scale. For this particual subject all three models had identical best fit therefore three model lines are on the top of each other.

	MLE	CAUSAL	LINEAR
σ_A	0.57 ± 0.02	0.57 ± 0.02	0.57 ± 0.02
σ_V	0.25 ± 0.01	0.24 ± 0.01	0.25 ± 0.01
σ_P	1.31 ± 0.04	1.33 ± 0.04	1.33 ± 0.04
μ_P	3.87 ± 0.02	3.87 ± 0.02	3.87 ± 0.02
p_{common}		0.96 ± 0.00	0.96 ± 0.00
ΔBIC		-9.29 ± 5.4	1.54 ± 1.3

Table 1 Summary of the estimated mean parameter values and across subject SEM for three models. BIC is a measure that can compare likelihood of the models with various number of parameters. Negative ΔBIC means that a given model performs better but none of the models seems to outperform MLE model given the magnitudes of the standard error.

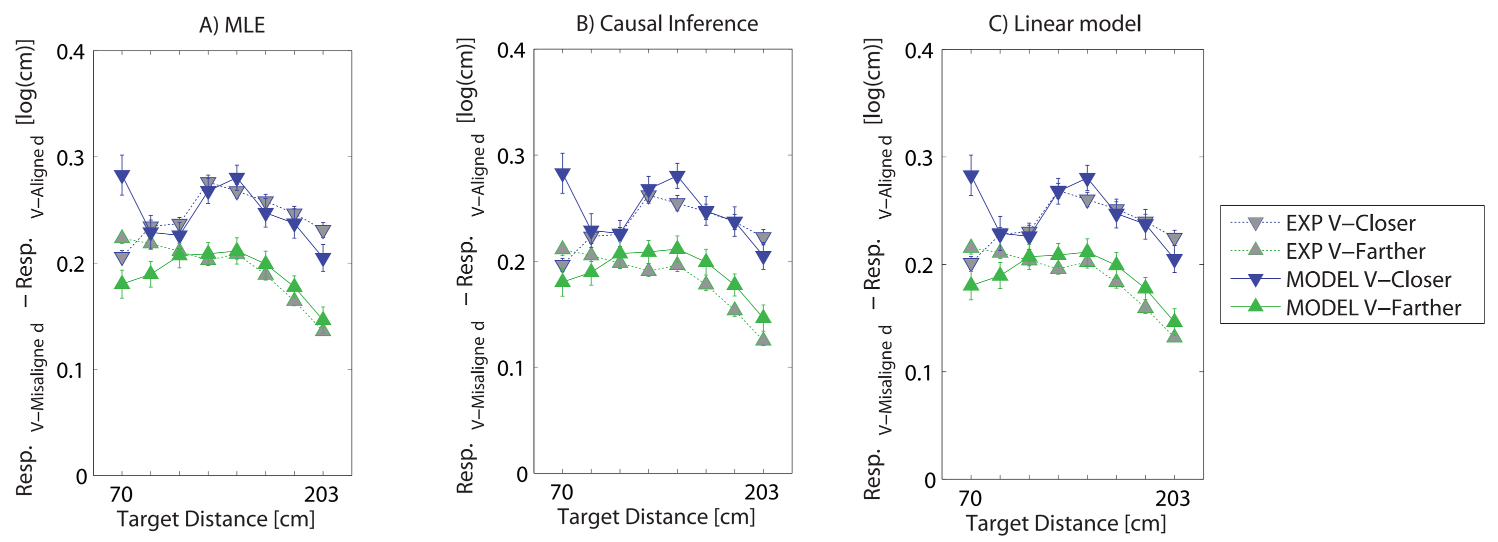


Figure 9 (A-C) Three models produce almost equal results because forced fusion always dominate the fit. The small deviation of Causal Inference model is within the error bars. The deviation of the blue point at 70cm only relates to the fact that the V component of the AV stimulus was placed really close to the listener away from the loudspeaker array.

Data of each subject were fit separately in a similar fashion as shown on Figure 8. Four or five parameters were computed from 480 data points. Model data on Figure 9 were computed from histograms with the best fit. The estimated responses were computed as a center of gravity of the histogram.

Current data suggest that the MLE model is sufficient in explaining current data.

Total variance of the MLE model can be expressed as:

$$\frac{1}{\sigma_{TOTAL}^2} = \frac{1}{\sigma_V^2} + \frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}$$

which gives the approximate value $\sigma_{TOTAL} = 0.22$ which is close to the σ computed from the subject histograms (~0.20).

Discussion

The current results are 'work in progress'.

The aim of the current modeling is to explain observed data of the VE experiments in the distance dimension and to identify possible underlying mechanisms.

Current preliminary analysis suggest that the subject response pattern can be explained by the MLE model with least number of parameters in which the A and V components were always perceptually integrated and additional parameter is not necessary. The reason could be the magnitude of the AV disparity relative to standard deviations in the current experiment.

However, no validation was made on the estimates of the internal variables σ_A and σ_V. And, the model was not tested with A-only or V-only stimuli, or with AV stimuli with varying disparity. Thus, the range of data used was not sufficient to critically test individual models.

The important factor seems to be that the model involved a prior function. Possibly, subjects create strong expectations about the scene which can account for the compression of the distance perception.

The auditory distance perception has been traditionally modeled by the power law function (Zahorik, 2005) which predicts only response biases. Here, we evaluate a Bayesian approach that can describe the data in terms of both biases and standard deviations.

Acknowledgements

* current address: MRC / CSO Institute of Hearing Research - Scottish Section, Glasgow, UK (lubos.hladek@ihr.gla.ac.uk)

This work was supported by APVV-0452-1, EU H2020-MSCA-RISE-2015 grant no. 69122. TECHNICOM (ITMS: 26220220182) EU R&D OP funded by ERDF.

References:

Alais, D., and Burr, D. (2004). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol., 14, 257–62. doi:10.1016/j.cub.2004.01.029

Ernst, M. O., and Banks, M. S. (2002). "Humans integrate visual and haptic information in a statistically optimal fashion," Nature, 415, 429–33.

Hládek Ľ, Le Dantec C, Seitz A, Kopčo N (2014) "Visual calibration of auditory distance perception" Presented at the 37th MidWinter meeting of the Association for Research in Otolaryngology, San Diego, CA.

Hládek Ľ, Le Dantec C, Kopčo N, Seitz A (2013). "Ventriloquism effect and aftereffect in the distance dimension," In: Proceedings of Meetings on Acoustics, Volume 19, pp. 050042, June 2013; Presented at the ICA 2013, 165th Meeting of the Acoustical Society of America, 52nd Meeting of the Canadian Acoustical Assocation and the 21st International Congress on Acoustics, Montreal, Quebec, Canada, 2-7 June 2013

Körding, K. P., Beierholm, U. R., Ma, W. J., Quartz, S. R., Tenenbaum, J. B., and Shams, L. (2007). "Causal inference in multisensory perception," PLoS One, 2, e943. doi:10.1371/journal.pone.0000943

Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). "Auditory distance perception in humans: A summary of past and present research," Acta Acust. united with Acust., 91, 409–420.