Modeling of Speech Localization in a Multitalker Mixture Using "Glimpsing" Models of Binaural Processing

Peter Toth[1], Angela Josupeit[2], Norbet Kopco[3], Volker Hohmann[2]

[1]Charles University in Prague, Czech Republic, [2]Carl von Ossietzly Universitat Oldenburg, Germany, [3]Safarik University, Kosice, Slovakia

Background

A recent study measured the human ability to localize a speech target masked by a mixture of four talkers in a room [Kopco et al., JASA 127, 2010, 1450-7]. The presence of maskers resulted in increases in localization errors that depended on the spatial distribution of maskers, the target-to-masker energy ratio (TMR), and the listener's knowledge of the maskers' locations. The current study investigated the performance of two binaural auditory "glimpsing" models in simulated experimental conditions. The models were tested under the assumption that optimal information about the TMR in individual spectro-temporal glimpses is available, quantifying the ability of the models to encode spatial properties of complex acoustic scenes.

Methods

The framework for the modeling consisted of: 1. auditory preprocessing, 2. extraction of binaural cues, 3. identifying the "glimpses", i.e., the spectro-temporal bins dominated by energy from only one source, 4. selecting target-related glimpses based on Ideal Binary Masks, and 5. estimating the target position. Two binaural models, one based on short-term running interaural coherence [Faller and Merimaa, JASA 116, 2004, 3075-89] and one on instantaneous interaural phase difference [Dietz et al., Speech Communication 23, 2011, 592-605] were modified and implemented. The stimuli were simulated by convolving speech tokens from the experiment with binaural room impulse responses recorded in a reverberant space similar to the experimental room.

Results

The two models produced similar predictions, both slightly worse than human performance. However, many trends in the data were captured by the models. E.g., the mean responses for lateral target locations were medially biased, the RMS errors were smallest for central target locations, and the overall performance varied with TMR. However, there were also qualitative differences. E.g., the models predicted best performance near the masker locations while humans were better at localizing targets far from the maskers.

Conclusion

The tested binaural models were able to capture several characteristics of human performance. Even though each model extracts binaural information in a different way, the model predictions were comparable, suggesting that the extracted features are equivalent and integrated in similar ways. The differences between the model predictions and human performance might be due to differences in interaural level difference processing, across-channel feature integration, or the assumed method of combination of target and masker glimpses.

Funding

Work supported by EU FP7-247543, VEGA-1/0492/12, the DFG SFB/TRR 31 "The Active Auditory System", and the PhD program "Hearing".