Modeling of Speech Localization in a Multitalker Mixture using “Glimpsing” Models of Binaural Processing

Peter Toh1, Angela Josupeti1, Norbert Kopco1, Volker Hohmann2

1Charles University Prague, Czech Republic; 2Medical Physics Section, Department of Medical Physics and Acoustics and Centre of Excellence “HearingLab”, Osnabrück University, Germany; 3Tatárik University Kosice, Slovakia.

INTRODUCTION

A recent study measured the human ability to locate a speech target masked by a mixture of four talkers in a room (Kopco et al., 2010, fig.1). 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.

In this study, we investigate the performance of two binaural auditory “glimpsing” models in experiments comparable to Kopco et al. (2010).

1. A model (Faller and Merimaa, 2004) based on short-term cross-correlation

2. A model (Dietz et al., 2011) based on interaural phase difference (IPD) extraction

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 (Cooke, 2006).

RESULTS

Model predictions were compared to three measures of human performance: the increase of RMS errors due to maskers, the biases due to maskers, and the misses/false alarm rates.

Lowering glimpsing threshold

Performance of glimpsing models was examined by different settings of glimpsing thresholds (g). Default thresholds for both models were 0.56. We observed that lowering thresholds results get better. Figures show default (g = 0.98) and the best (g = 0) parameterizations for each model.

Lowering binaural processor time constant

Lower time constant means shorter integration window for processing binaural cues and more emphasis on actual states of cues. We have observed improved performance in localization with lowering time constant. In fact, to match humans’ performance we had to set time constant to very low values (ca. 0.1 times fundamental period).

On-masker and off-masker target locations

Psychophysical data shows that human localization performance is better when target and maskers are spatially separated. Both glimpsing models (in every parameterization) show opposite behavior: the farther the target from the masker, the worse localization performance is.

Target location estimation

A way to evaluate information in selected glimpses to make final target location estimation have a big importance. Although simple methods like mean TDI could have low overall RMS errors, methods based on suppressing non-selected glimpses are closer to experimental data.

High frequency bands

The models can perform very well based only on high frequency channels despite high TDI ambiguity. The best results were achieved when both low and high frequency channels were processed together.

Miss rates and false alarm rates

The subjects in the experiments were asked to indicate when they did not hear the target. Inability to detect targets in normal trials is measured by miss rate while inability to indicate no presence in catch trials by false alarm rate. Default versions of the models have miss rates quantitatively comparable to humans. However, we can observe the same on-masker/offset-masker disparity as in the RMS data. With lower glimpsing thresholds miss rates dropped to zero. False alarm rates were unrealistically high in all cases.

CONCLUSIONS

The tested binaural models were able to capture several characteristics of human performance. The main responses for lateral target locations were mostly biased; the RMS errors were smaller for central target locations, and the overall performance varied with TMR. 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.

Contrary to the expected behavior of glimpsing models, high thresholds do not yield better results. It seems that higher numbers of selected samples are beneficial for further processing in more complex scenarios, where information tends to be noisy so more samples are needed for better estimation. TDI/phase ambiguity can be another reason for having more samples being an advantage.

It seems that on-masker/off-masker disparity is a direct consequence of the way the glimpsing model works. Target glimpses can contain also masker information that causes shifts in estimation – this shift is bigger when maskers are more distant (humans don’t make this kind of errors).

The differences between the model predictions and human performance might be due to differences in interaural level difference processing, across-channel feature integration, the assumed method of combination of target and masker glimpses and due to the usage of B4.

Kopco et al. (2010) also examined the effect of providing a priori information about masker locations on performance. The current models need to be extended to be more use of this kind of information.

REFERENCES


Dietz et al. (2011). “Auditory model based direction estimation to prominent sources from binaural signals” Speech Communication 53(5)


MATERIALS & METHODS

The following framework was used to test the two models:

A: multitalker signal generation

To create sound inputs close to those in the experiment, the stimuli were simulated by convoking speech talkers from the experiment with binaural room impulse responses recorded in a reverberant space similar to the experimental room.

B: auditory peripheral preprocessing

Standard auditory preprocessing mimicking gammatone filterbank was used. Specific parametrization was adopted from (Dietz et al., 2011).

C: binaural processing

Models use different methods to extract binaural cues – while Faller and Merimaa based their binaural processing on computing IC in range of time shifts, Dietz computes interaural phase differences (IPD) directly. Both models require defined essential-decaying time window upon which they do computations.

D: selection of glimpses

Both models claims to have an ability to select spectro-temporal bins dominated by energy from only one source (“glimpses”). Binaural cues contained in selected time should point to location of that source.

Faller and Merimaa use glimpsing based on interaural coherence (IC) – high short-term correlation between the left and the right signal indicates that one source is dominant in particular time-frequency window.

Dietz used interaural vector strength (IVS) of interaural phase differences to determine glimpses – i.e. dominance of one source is indicated by low fluctuation of IPD values.

E: target-related information selection

As the main focus of this work is to explore the best possible performance of the glimpsing models, we used a simple bias mask (BM, Cooke, 2006) to relate individual glimpses to target source and masker sources. BM was computed as a ratio between target energy and masker energy above a certain threshold (default set to 0 dB).

F: target location estimation

We examined several methods for integration of information from glimpses across time and frequency bands. They relied on ITD information. Methods of localization estimation varied from simple (mean/median of ITD of selected samples, maximum of histogram) to more complicated which used information from both selected and non-selected glimpses (combining Gaussian kernel-based pdf functions for each group of glimpses).

ACKNOWLEDGEMENTS

Work supported by EU NNI PIH4-2014-073779 245743, VEGA-1/0492/12, the DFS SF/PRR 31 “The Active Auditory System”, NII/PIH4-2014-073779, and the PGS program “Hearing.”

Fig.3. RMS errors for control cases (only target-presented)

Fig.4. RMS errors for various masker locations and masker patterns (TMR = 0)

Fig.5. Increase of RMS error due to the presence of the masker. Predictions of the Dietz and Faller and Merimaa models with optimized parameterizations and the Kopco et al. (2010) data. At a gain source with right tympanic and left data were integer flip/flop operation

Fig.6. Miss rates across target positions and masker patterns (TMR = 0)

Fig.7. Overall and on-masker/off-masker locations mean RMS errors

Fig.8. RMS error re control (only target-presented)

Fig.9. RMS error re control (only target-presented)

Fig.10. RMS error re control (only target-presented)

Fig.11. RMS error re control (only target-presented)