A Shadowing Experiment with Natural and Synthetic Stimuli
Iona Gessinger¹, Eran Raveh¹-², Johannah O’Mahony¹, Ingmar Steiner¹-³, Bernd Möbius¹,²
¹Computational Linguistics & Phonetics, ²Multimodal Computing & Interaction, Saarland University, ³DFKI
lastname@coli.uni-saarland.de

1 Phonetic Convergence
- Defined as an increase in segmental and suprasegmental similarities between two speakers [1]
- Found in both conversational and non-conversational human-human interaction [2, 3]
- Received little to no attention so far in the field of human-computer interaction
  → Do human speakers also converge to synthesized speech?

2 Shadowing Experiment
We examine three segmental features that show variation across native speakers of German. The target features are embedded in short German sentences, e.g.:

<table>
<thead>
<tr>
<th>sentence</th>
<th>target feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die Bestätigung ist für Tanja.</td>
<td>[e:] vs. [e:]</td>
</tr>
<tr>
<td>Ich bin sicher, nach Schokolade.</td>
<td>[k] vs. [k]</td>
</tr>
<tr>
<td>Wir begleiten dich zur Taufe.</td>
<td>[a:] vs. [a:]</td>
</tr>
</tbody>
</table>

The experimental procedure consists of four tasks:

- Baseline: 40 sentences
- Visual Task: 7 minutes
- Post Production: 35 sentences
- Shadowing Task: 35 sentences

In the shadowing task, the participants are presented with productions of two model speakers that contain the opposite target feature realization of that observed in the participants’ baseline productions.

3 Results for Natural Stimuli
Target feature [e:] vs. [e:] produced by participants with baseline preference [e:] (left figure; n = 11) and [e:] (right figure; n = 10) in the production tasks base, shadow and post, as well as the human models (n = 2) they heard in the shadowing task, producing [e:] (left figure) and [e:] (right figure) (red). The ellipses visualize the confidence level of the estimated true mean (here: ±1 standard deviation).

Target feature [e:] vs. [k] produced by all participants (n = 21) in the shadowing task: in 34% of the cases convergence was observed.

Target feature [a:] vs. [n]: Occurrences of [a] in all participants (n = 21) and their respective mean length per condition. Only segments of at least 30ms were considered as [a]. The length of [a] produced by the model speakers ranged from 30ms to 69ms (mean = 48ms).

There were 8.6% more occurrences of [a] in the shadowing condition than in the baseline condition.

4 Conclusion & Future Work
Convergence was observed in all three target phenomena in the natural condition. The degree of convergence varied across the participants. These results will be used as the baseline for the synthetic condition.

Segment durations and pitch contours of the natural stimuli will be used to generate the synthetic stimuli with MaryTTS. This will control for potential differences in information structure between the two conditions.