# A Shadowing Experiment with Natural and Synthetic Stimuli

Iona Gessinger[1], Eran Raveh[1,2], Johannah O'Mahony[1], Ingmar Steiner[1–3], Bernd Möbius[1,2]

[1] *Computational Linguistics & Phonetics* [2] *Multimodal Computing & Interaction, Saarland University* [3] *DFKI*

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

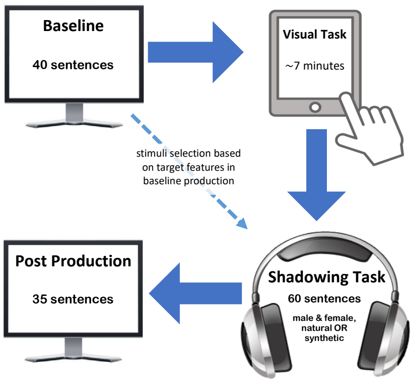
→ Do human speakers also converge to synthesized speech?

## 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.:

| sentence | target feature |
|---|---|
| Die Best**ä**tigung ist für Tanja. | [ɛː] vs. [eː] |
| Ich bin süch**tig** nach Schokolade. | [ɪç] vs. [ɪk] |
| Wir begleit**en** dich zur Taufe. | [ən] vs. [n̩] |

The experimental procedure consists of four tasks:


Baseline
40 sentences
Visual Task
~7 minutes
stimuli selection based on target features in baseline production
Post Production
35 sentences
Shadowing Task
60 sentences
male & female, natural OR synthetic

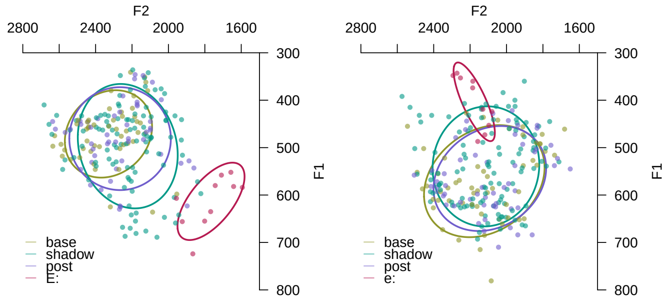
In the shadowing task, the participants are presented with productions of two model speakers that contain the opposite target feature realization of that observed in the participants' baseline productions.

[1] J. Pardo. On phonetic convergence during conversational interaction. *JASA*, 119(4), 2006.

[2] N. Lewandowski. *Talent in nonnative phonetic convergence*. PhD thesis, Universität Stuttgart, 2012.

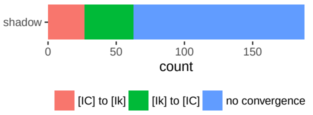
[3] M. Babel et al. Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, 5(1), 2014.

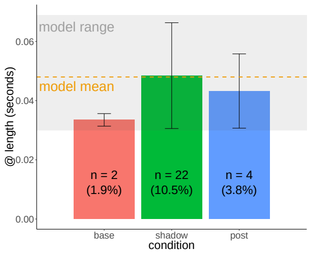
## 3 Results for Natural Stimuli



Target feature [ɛː] vs. [eː] produced by participants with baseline preference [eː] (left figure; n = 11) and [ɛː] (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 [ɛː] (left figure) and [eː] (right figure) (**red**). The ellipses visualize the confidence level of the estimated true mean (here: ±1 standard deviation).

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





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

There were 8.6% more occurrences of [ə] 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.

UNIVERSITÄT DES SAARLANDES

Computational Linguistics & Phonetics

M2CI CLUSTER OF EXCELLENCE

DFKI