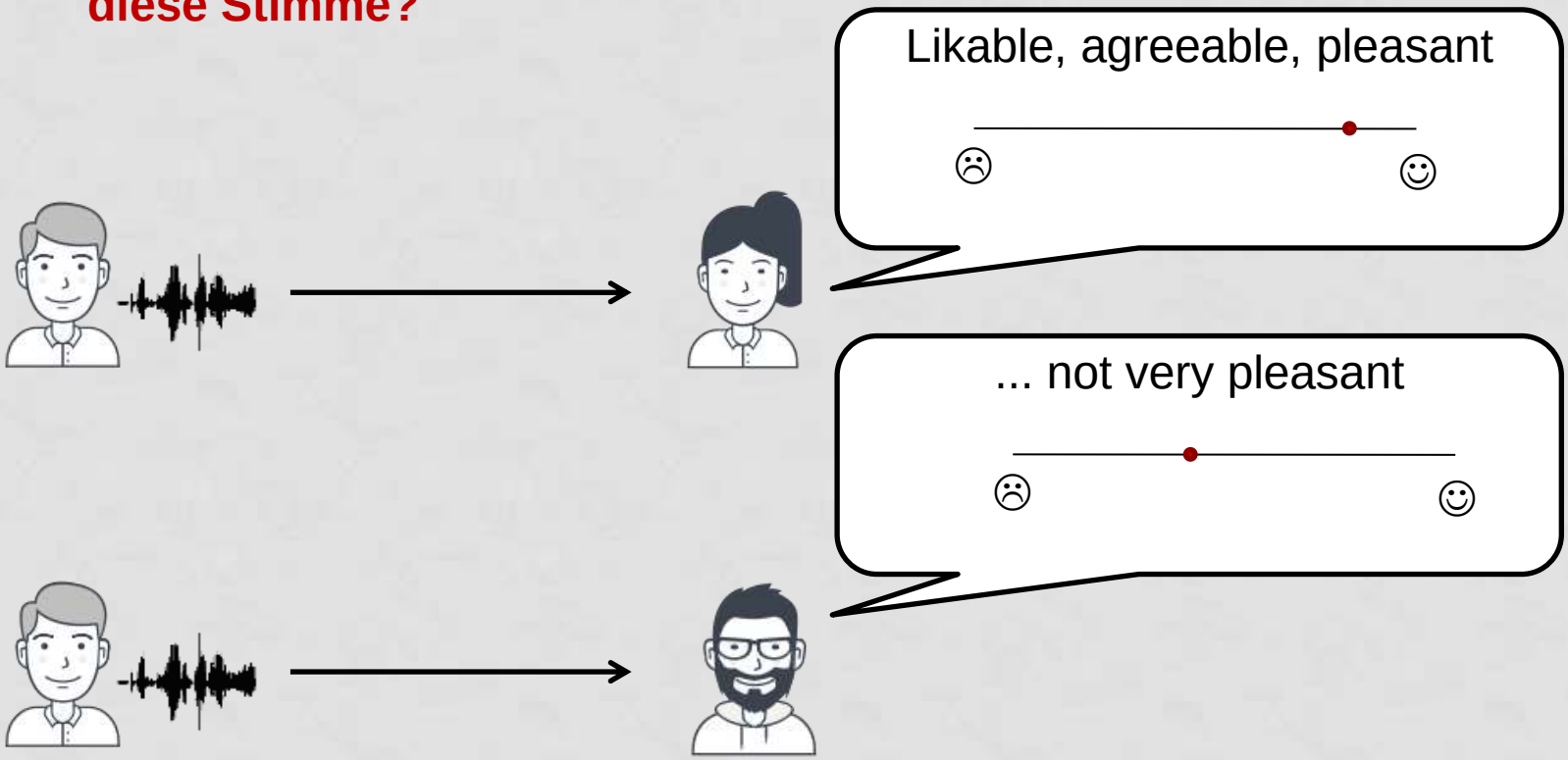
A Paired-Comparison Listening Test for Collecting Voice Likability Scores

Laura Fernández Gallardo

Quality and Usability Lab, Technische Universität Berlin, Germany

ITG Conference on Speech Communication 2016. October 5–7, Paderborn, Germany

Wie sympathisch finden Sie diese Stimme?



Welche Stimme finden Sie sympathischer?

Outline

- Motivation
- Speech stimuli and listening test
- Preference choice analyses
  - Paired-comparison test: Consistency checks
  - Direct scaling test: Consistency checks
  - Scaling listeners' preference
- Contrasting the test approaches
- Conclusions

Motivation

- Likert scale: disagreement between raters

- We propose: paired-comparison listening test
- Apply the Bradley-Terry-Luce (BTL) model to derive a ratio scale of preference

- Goals:
  - 1. To ascertain whether the BTL model can be fit to listeners' paired comparisons of voice likability
  - 2. To contrast the paired-comparison test with the direct scaling test in [3]

[3] L. Fernández Gallardo and B. Weiss, "Speech Likability and Personality-based Social Relations: A Round-Robin Analysis over Communication Channels," in Interspeech, 2016.

Outline

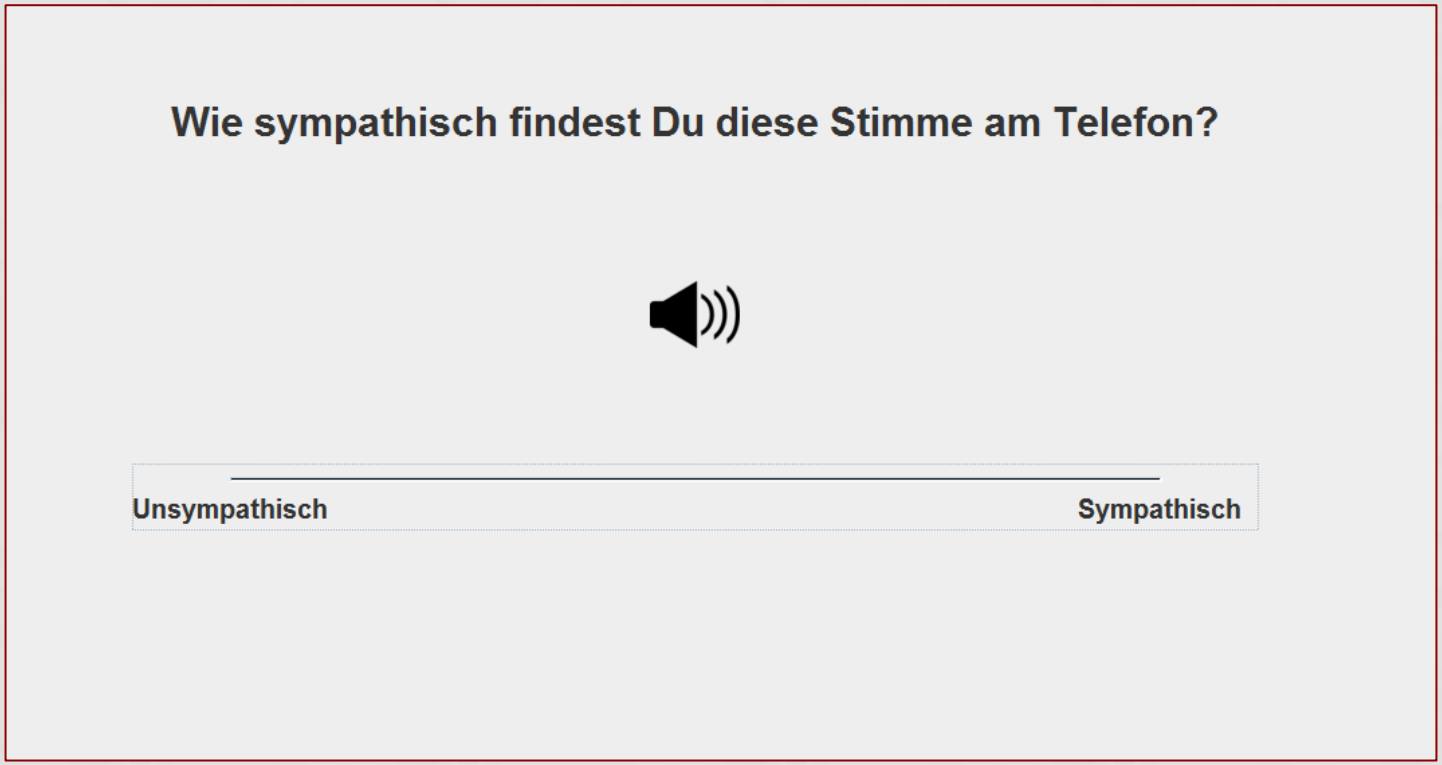
- Motivation
- Speech stimuli and listening test
- Preference choice analyses
  - Paired-comparison test: Consistency checks
  - Direct scaling test: Consistency checks
  - Scaling listeners' preference
- Contrasting the test approaches
- Conclusions

Speech stimuli

[22] L. Fernández Gallardo, "Recording a High-Quality German Speech Database for the Study of Speaker Personality and Likability," accepted in 12. Tagung Phonetik und Phonologie im deutschprachigen Raum, 2016.

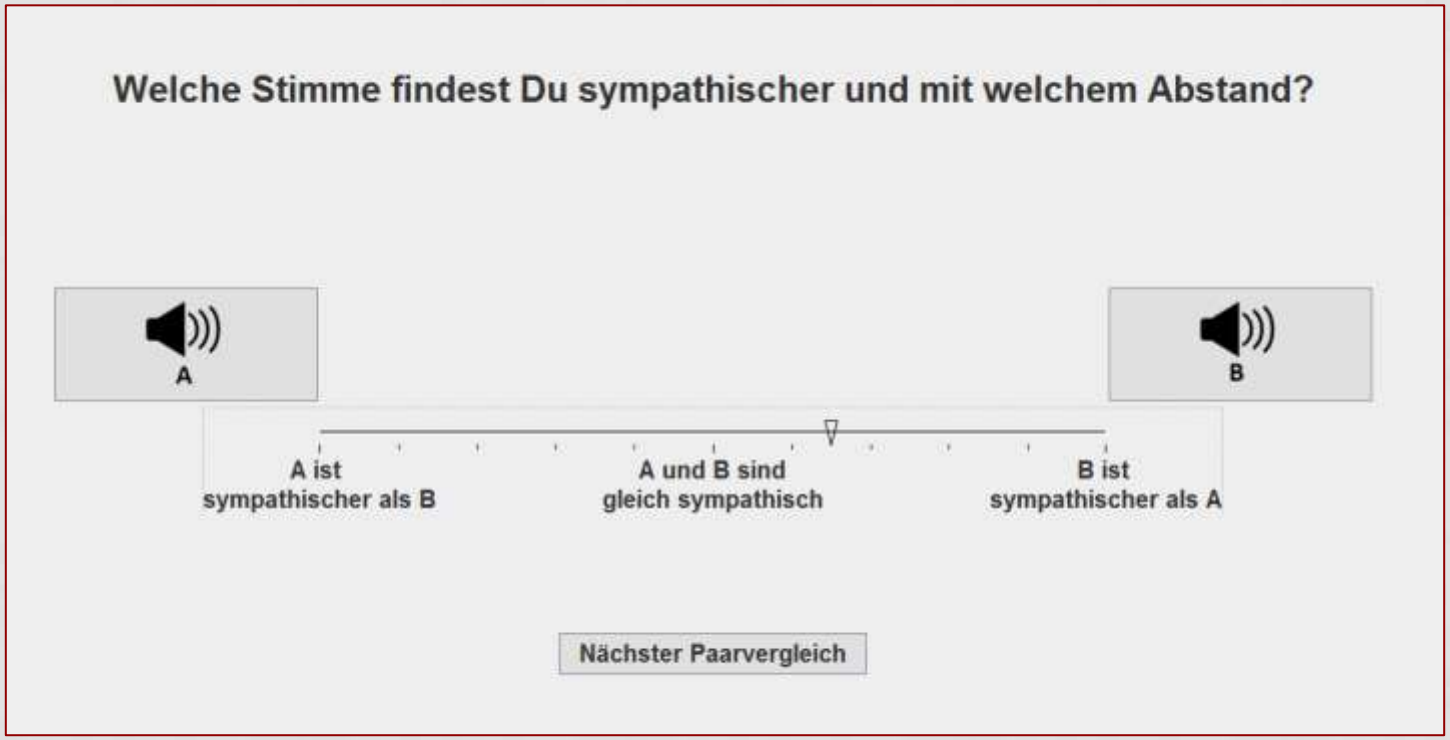
Listening test

- [3] L. Fernández Gallardo and B. Weiss, "Speech Likability and Personality-based Social Relations: A Round-Robin Analysis over Communication Channels," in Interspeech, 2016.



Listening test

- Paired-comparison listening test



Listening test

- Direct scaling test [3]
  - Stimuli: "Ich würde auf die SMS gern verzichten und meine Frei-Minuten dafür erhöhen“
  - 15 stimuli from male speakers rated by 15 female listeners
- Paired-comparison listening test
  - $\binom{15}{2} = 105$ pairs presented in the test, randomly
  - Unchanged:
    - male speakers
    - female listeners (13 out of 15)
    - speech material (wideband)
    - headphones
    - test room

Outline

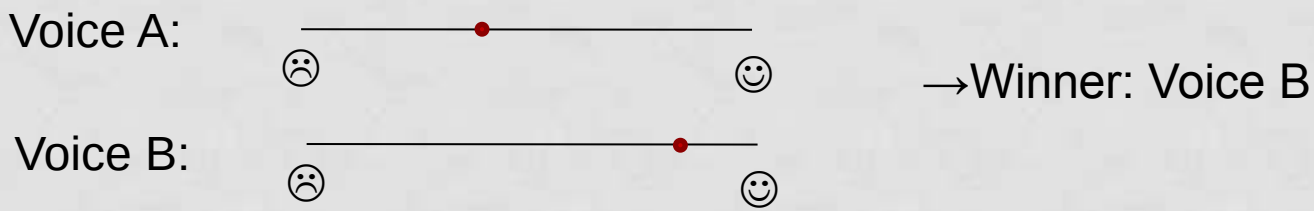
- Motivation
- Speech stimuli and listening test
- Preference choice analyses
  - Paired-comparison test: Consistency checks
  - Direct scaling test: Consistency checks
  - Scaling listeners' preference
- Contrasting the test approaches
- Conclusions

Consistency checks

- Consistency of responses for each rater individually
  - average Kendall's coefficient of consistence $\zeta = 0.78$ (range: 0.63–0.93)
  - all ζ > chance value ($p < .001$) → participants did not make their choices at random
- Agreement amongst raters
  - Kendall's coefficient of agreement u
  - u = 0.19 (range: -0.08–1)
- Checking stochastic transitivity properties
  - If x > y and y > z, a transitivity is violated if z > x
  - The transitivity violations found can be attributed to randomness (no systematic) →The BTL model can be fit

Consistency checks

- Simulated paired-comparison data from the direct scaling test



- All participants were consistent
- Low agreement amongst raters
- No systematic transitivity violations

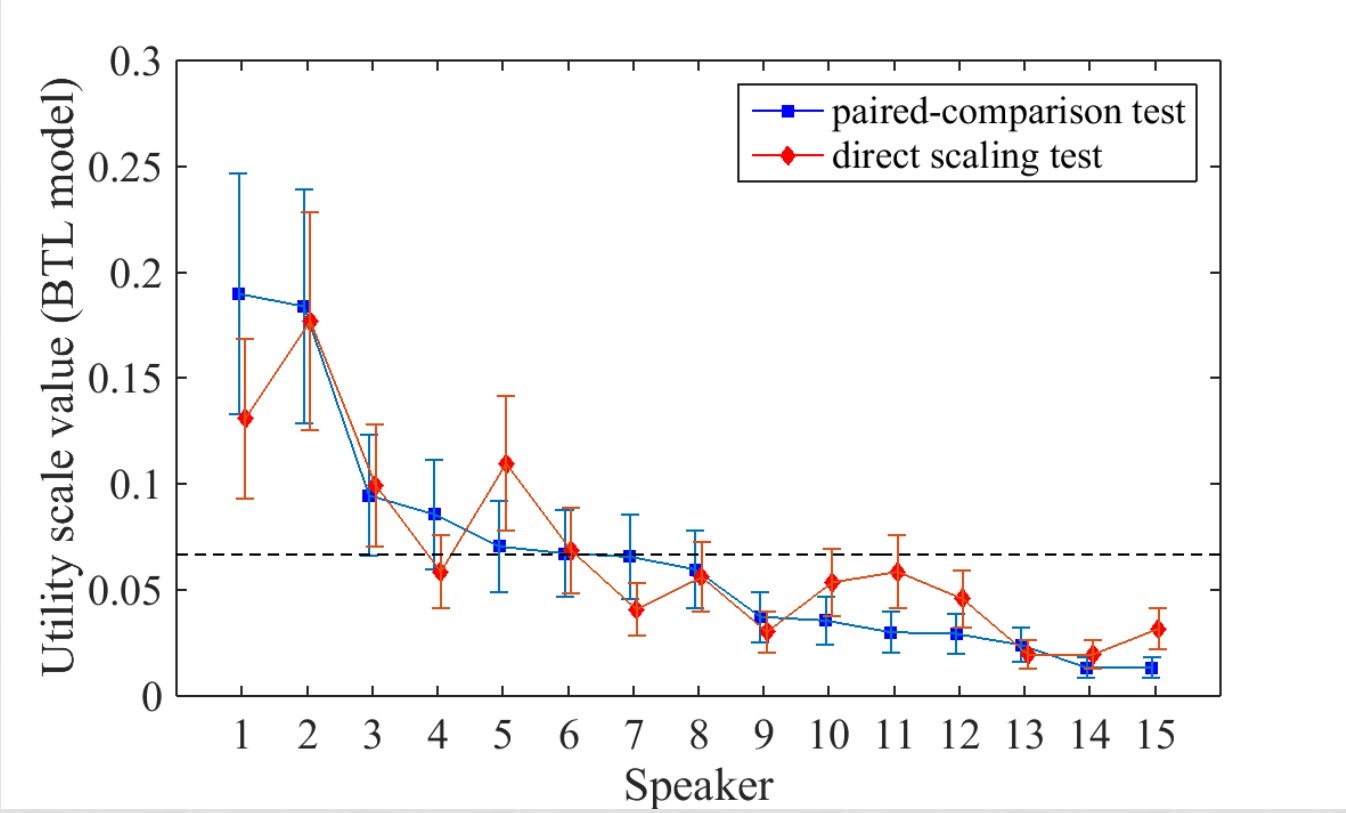
Consistency checks

- Transitivity violations and Kendall's coefficient of agreement

Test	Paired-comparison	Direct scaling
Weak stochastic transitivity	10	8
Moderate stochastic transitivity	28	18
Strong stochastic transitivity	134	127
Kendall’s u	0.19	0.11

Scaling listeners' preference

- Ratio scale preferences estimated by the BTL model



Outline

- Motivation
- Speech stimuli and listening test
- Preference choice analyses
  - Paired-comparison test: Consistency checks
  - Direct scaling test: Consistency checks
  - Scaling listeners' preference
- Contrasting the test approaches
- Conclusions

Contrasting the test approaches

- Strong and significant correlations between the score series
  - pair-comparison *u*
  - simulated pair-comparison *u‘*
  - mean likability [3]

	u	*u‘*	[3]
u			
u‘	$R^2 = 0.90$		
[3]	$R^2 = 0.81$	$R^2 = 0.91$	

Contrasting the test approaches

- Direct scaling and the paired-comparison tests provide very similar likability scores
- Higher agreement between raters and greater discriminability have been found for the paired-comparison test with respect to the direct scaling test
- The number of pairs in the test $\binom{N}{2}$ grows quadratically (Q(N2)) with the number of voices N to be scaled

Outline

- Motivation
- Speech stimuli and listening test
- Preference choice analyses
  - Paired-comparison test: Consistency checks
  - Direct scaling test: Consistency checks
  - Scaling listeners' preference
- Contrasting the test approaches
- Conclusions

Conclusions

- A paired-comparison listening test
  - for collecting subjective voice likability ratings
  - contrasted to the direct scaling test in [3]
- The BTL probabilistic choice model could be successfully applied and ratio scale preference measures were derived
  - paired-comparison constitutes a reliable method
  - enables simple comparative judgments
- Contrasting paired-comparison and direct scaling tests
  - results highly correlated
  - paired-comparison leads to a somewhat higher agreement between raters and greater discriminability
  - considerable number of pairs in the paired-comparison test $\binom{N}{2}$!
  - direct scaling tests may be therefore preferred despite the detriment to raters' agreement

Thank you for your attention!

Laura Fernández Gallardo, Ph.D.
laura.fernandezgallardo@tu-berlin.de

Questions?

Laura Fernández Gallardo, Ph.D.
laura.fernandezgallardo@tu-berlin.de