Verification and Characterization of Users by Their Voices

Laura Fernández Gallardo, Ph.D.

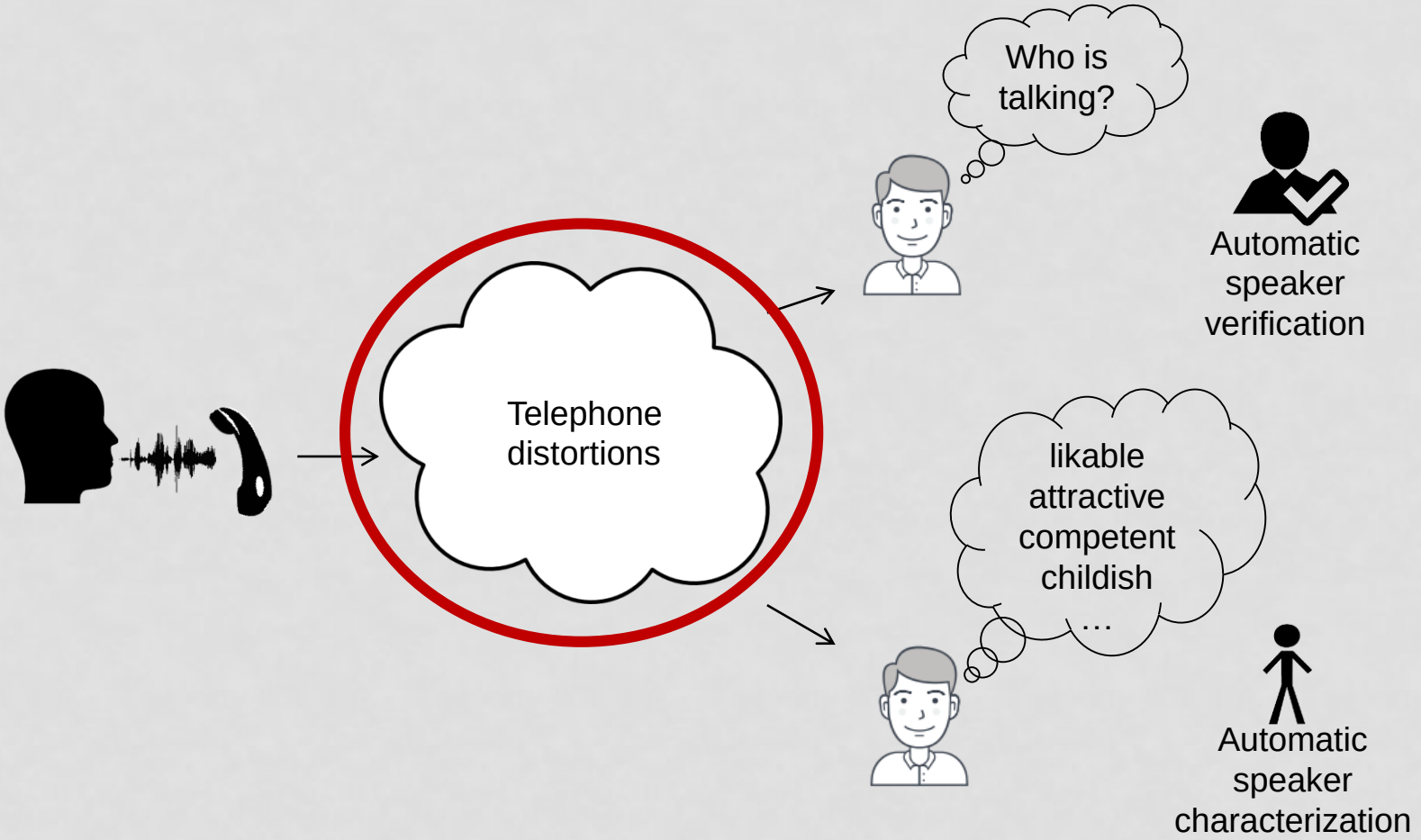
Quality and Usability Lab, Technische Universität Berlin, Germany

Who I am

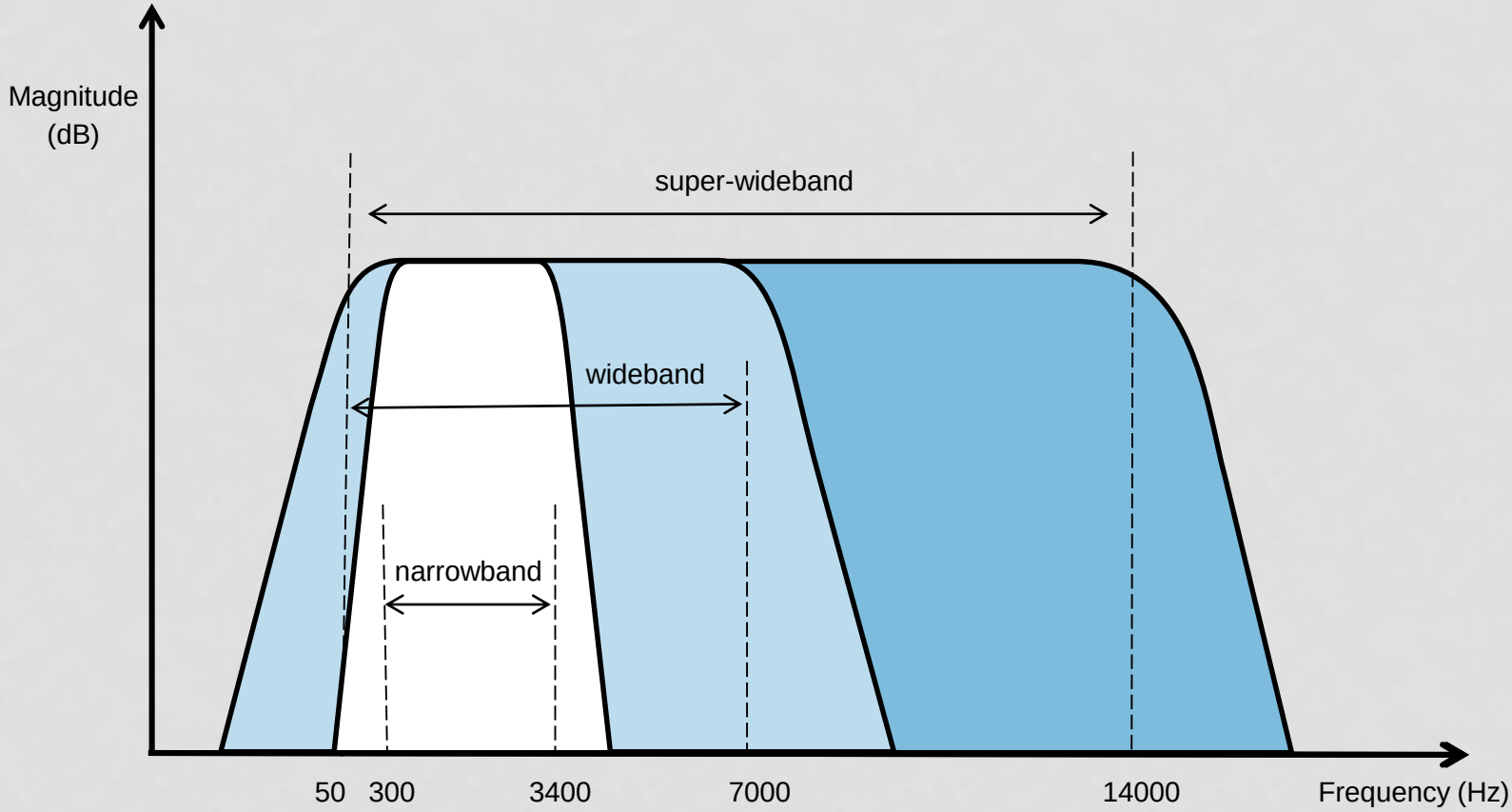
Laura Fernández Gallardo

- 2005 – 2011: MSc. Telecommunications Engineering (University of Granada, Spain)
- 2012 – 2015: PhD in Computer Science (University of Canberra, Australia)
- 2015 – 2018: Postdoc Researcher (Technische Universität Berlin, Germany)

Outline

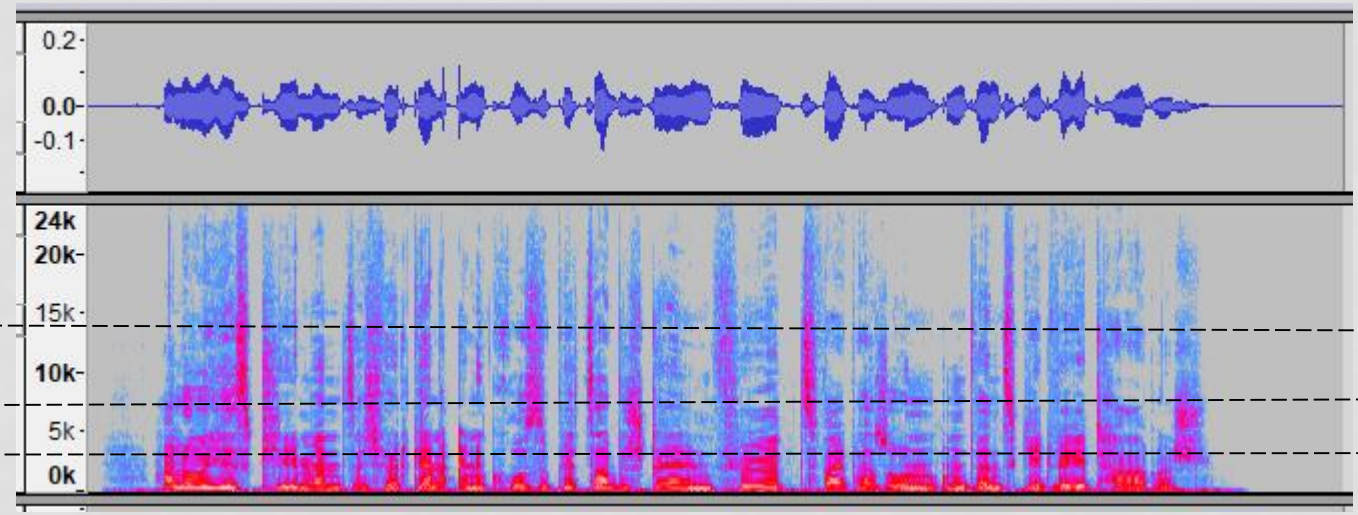

Who is talking?
Automatic speaker verification
Telephone distortions
likable
attractive
competent
childish
…
Automatic speaker characterization

Telephone distortions



Telephone distortions

Clean

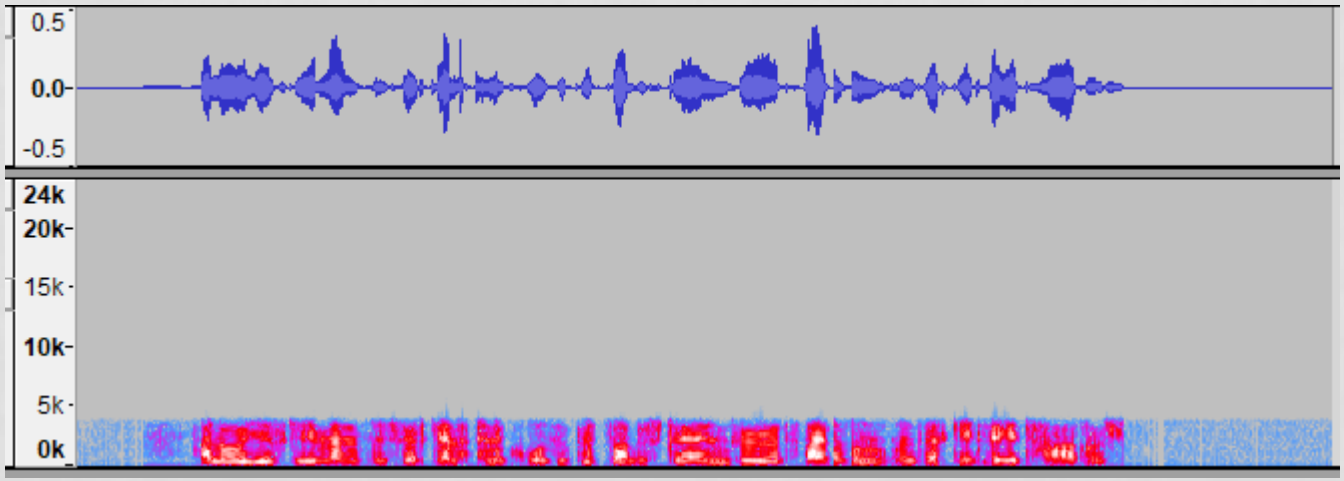


SWB

WB

NB

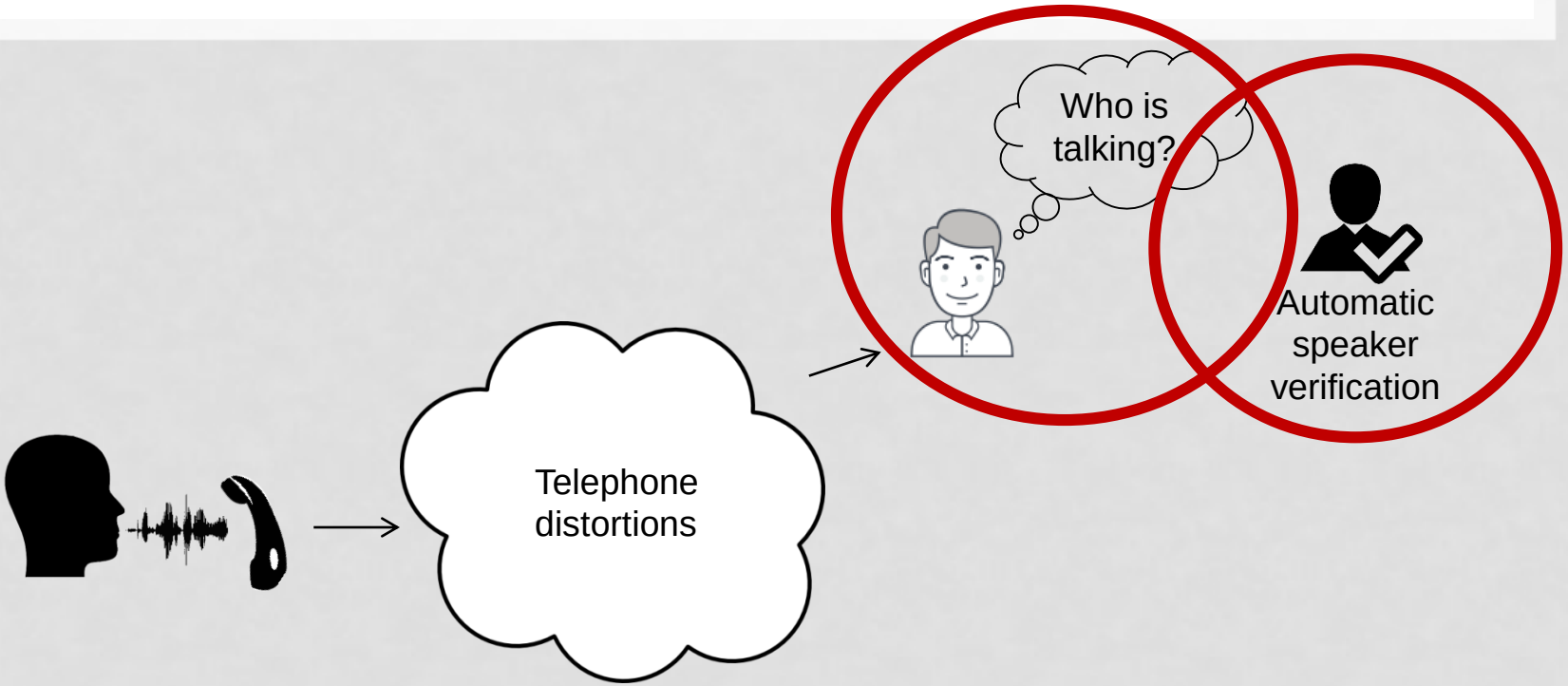
Narrowband,
AMR-NB@4,75,
packet loss 10%,
jitter 10ms



Outline

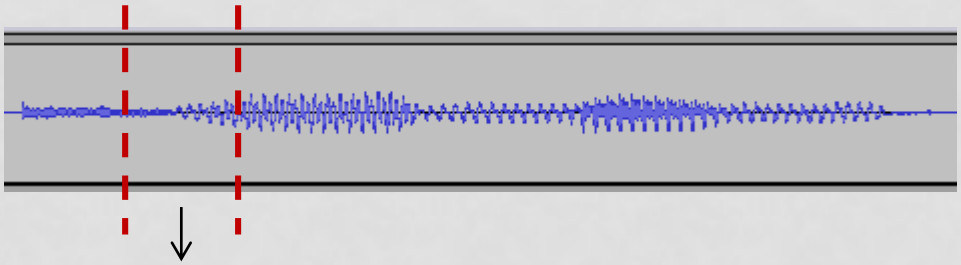
My Phd

Outline

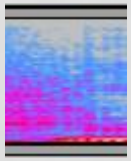

Who is talking?
Automatic speaker verification
Telephone distortions

Extracting speech features

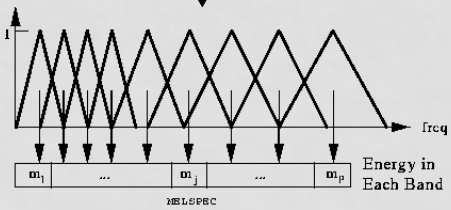
framing



Fourier Transform



Mel filterbank



Speech features: MFCCs
(Mel Frequency Cepstral Coefficients)

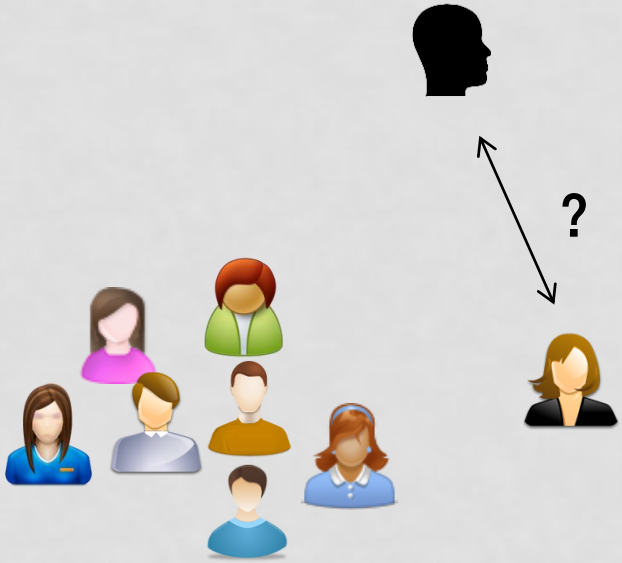
Automatic speaker verification (ASV)

Automatic speaker verification (ASV):

- GMM-UBM
- i-vectors

Test utterance:

Identity claim:

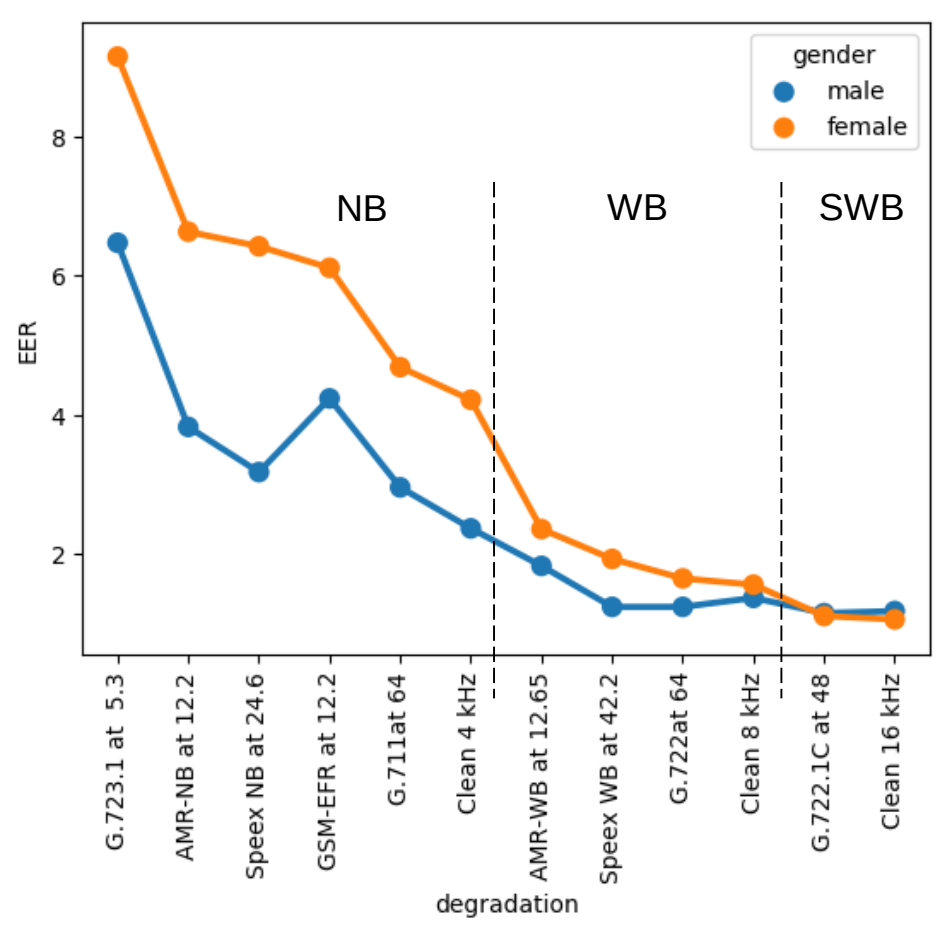


Enrolled users

Automatic speaker verification (ASV)

Too few speech databases sufficient bandwidth (> 16 kHz)

ASV evaluation: GMM-UBM



My PhD: contributions

- ✓ Significant improvement from NB to WB, yet not from WB to SWB
- ✓ Effects of handset, codecs, and packet loss
- ✓ Speaker-discriminative fricative sounds in WB

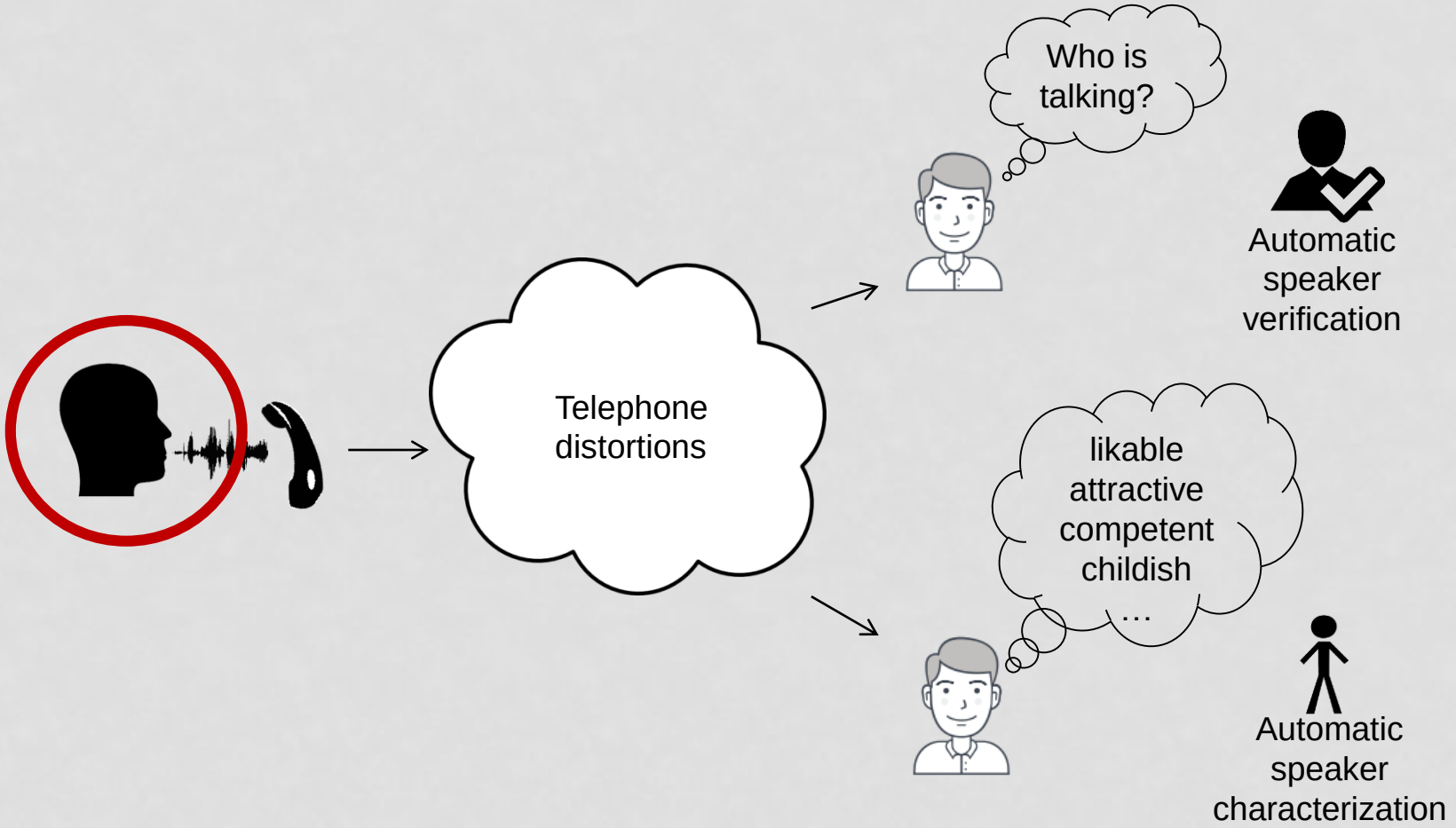
Automatic speaker verification

- ✓ Significant improvement from NB to WB, and from WB to SWB, especially for females
- ✓ Effects of channel impairments and of channel mismatch
- ✓ Effects of codecs on speaker-specific spectral regions
- ✓ Unvoiced fricatives effective at > 4 kHz

Outline

My Postdoc

Outline


Who is talking?
Automatic speaker verification
Telephone distortions
likable
attractive
competent
childish
…
Automatic speaker characterization

Nautilus Speaker Characterization (NSC) Corpus

300 speakers
(126 m, 174 f)
Native German

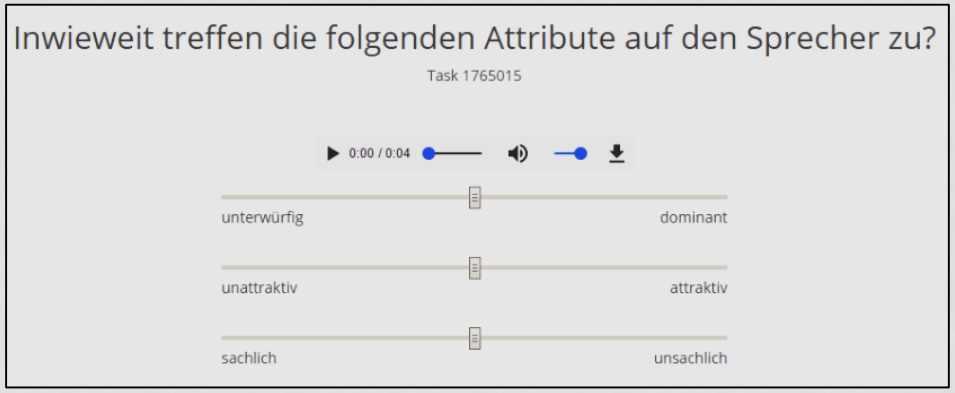
acoustically-isolated room

AKG C 414B-XLS
microphone

Scripted, semi-spontaneous and spontaneous conversational speech
Fs = 48 kHz

Released for non-commercial research and teaching purposes only

Speaker characteristics

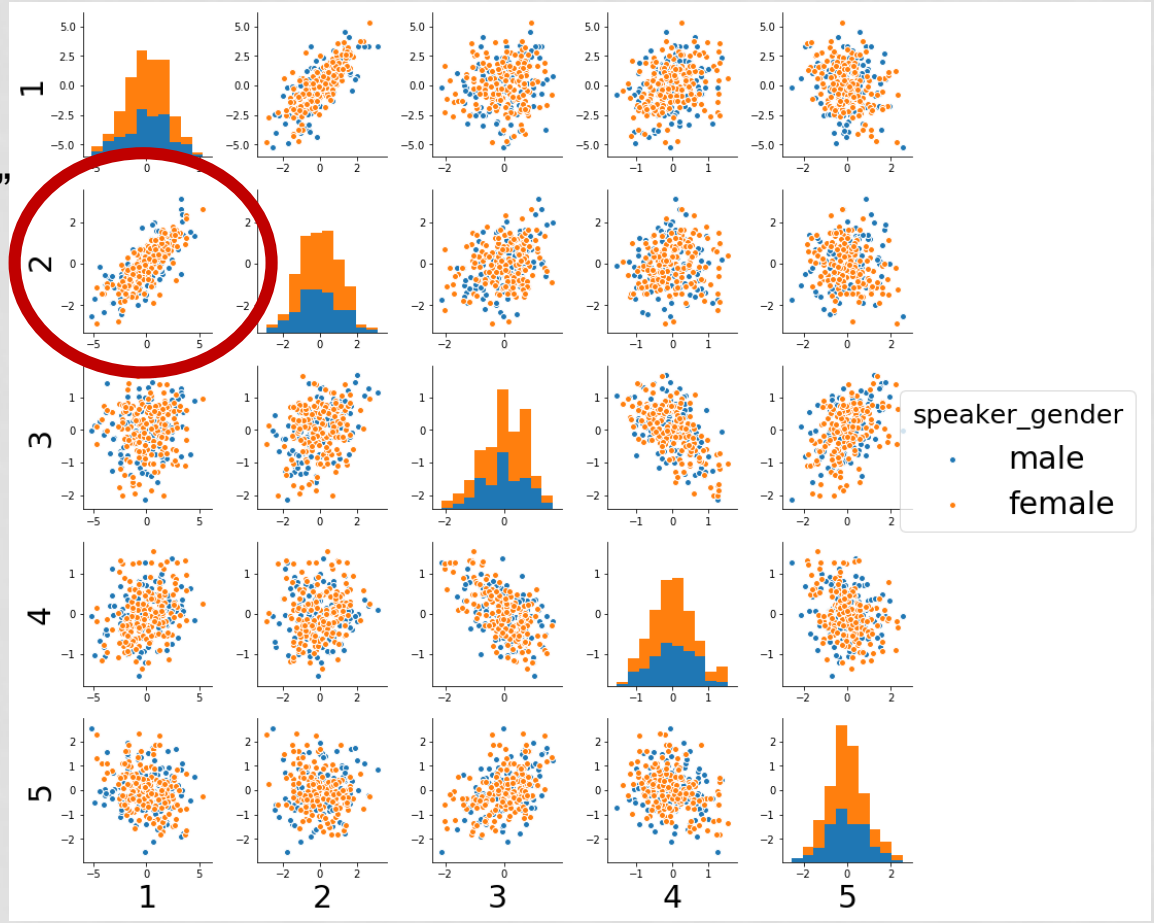


item	Antonyms (German)	
1	sympathisch	unsympathisch
2	unsicher	sicher
3	unattraktiv	attraktiv
4	verständnisvoll	verständnislos
5	entschieden	unentschieden
6	aufdringlich	unaufdringlich
7	nah	distanziert
8	interessiert	gelangweilt
9	emotionslos	emotional
10	genervt	nicht genervt
11	passiv	aktiv
12	unangenehm	angenehm
13	charaktervoll	charakterlos
14	reserviert	gesellig
15	nervös	entspannt
16	distanziert	mitfühlend
17	unterwürfig	dominant
18	affektiert	unaffektiert
19	gefühlskalt	herzlich
20	jung	alt
21	sachlich	unsachlich
22	aufgeregt	ruhig
23	kompetent	inkompetent
24	schön	hässlich
25	unfreundlich	freundlich
26	weiblich	männlich
27	provokativ	gehorsam
28	engagiert	gleichgültig
29	langweilig	interessant
30	folgsam	zynisch
31	unaufgesetzt	aufgesetzt
32	dumm	intelligent
33	erwachsen	kindlich
34	frech	bescheiden

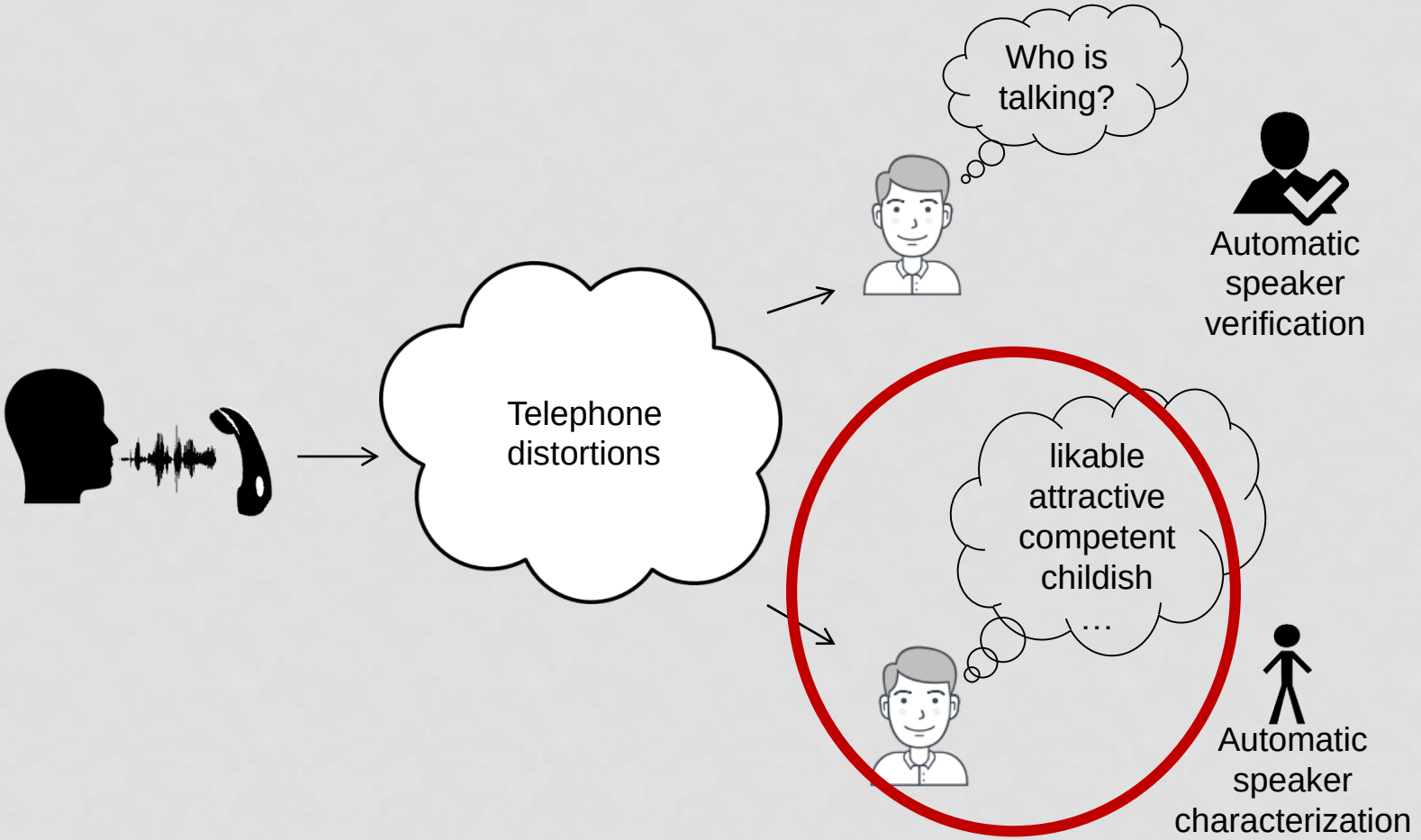
The space of perceptual traits

“WAAT”

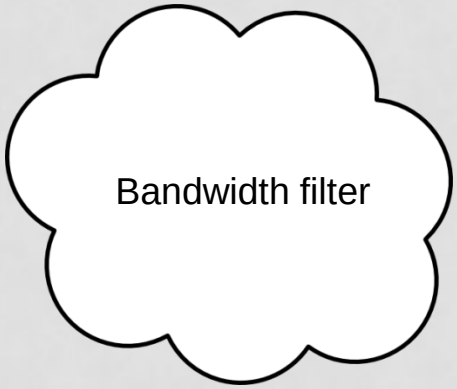
1: warmth
2: attractiveness
3: confidence
4: compliance
5: maturity



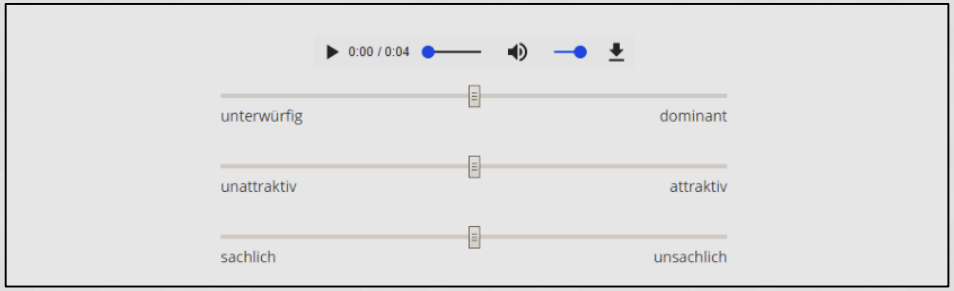
Outline


Who is talking?
Automatic speaker verification
Telephone distortions
likable
attractive
competent
childish
…
Automatic speaker characterization

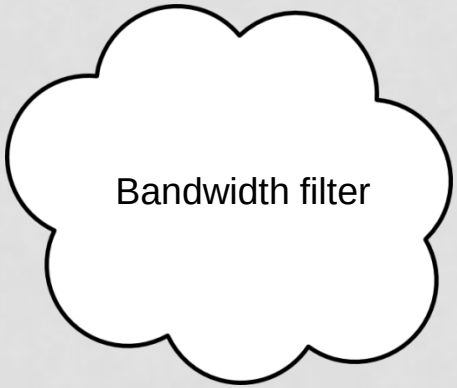
MOS - Speaker characteristics



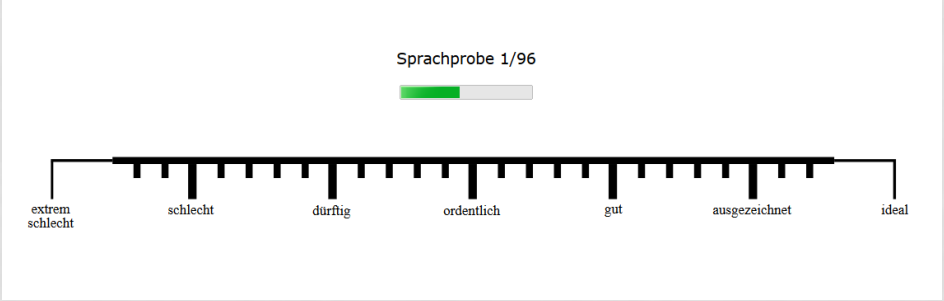
„Inwieweit treffen die folgenden Attribute auf den Sprecher zu?“

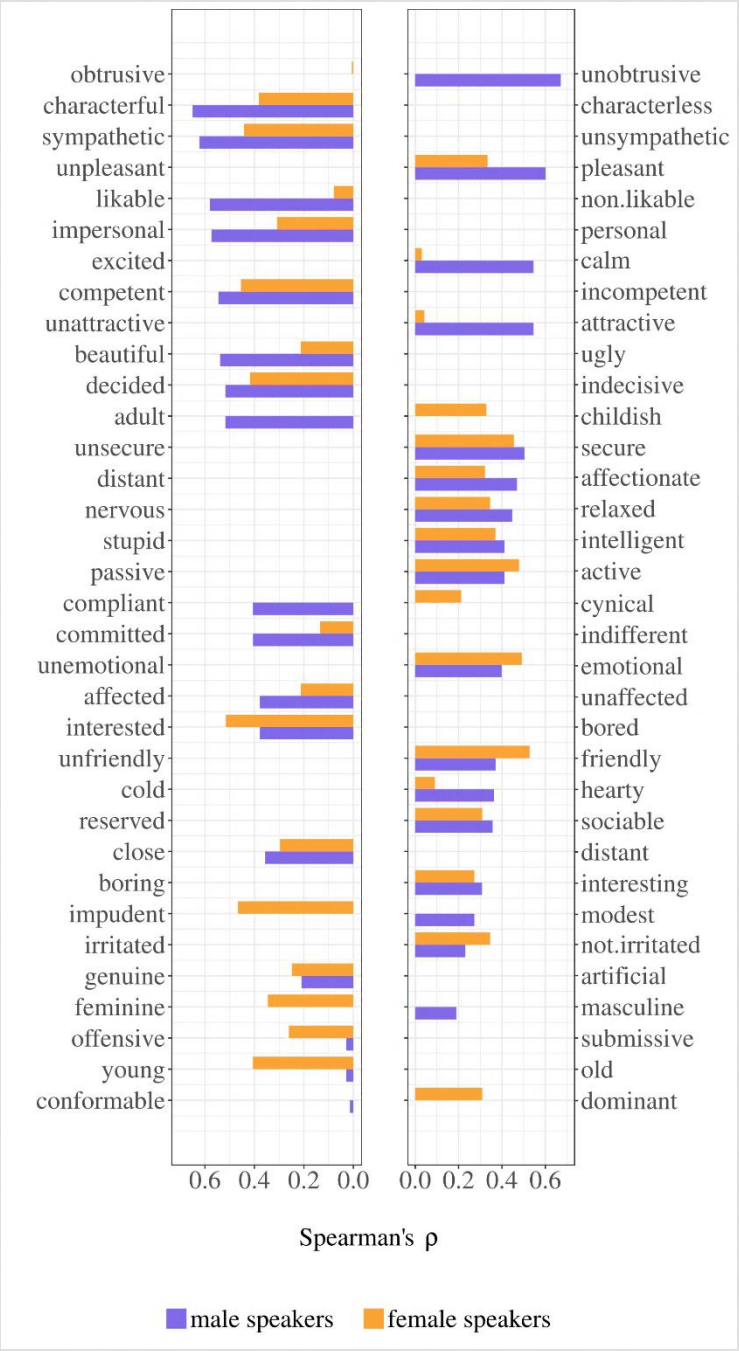


MOS - Speaker characteristics



„Bitte beurteilen Sie die nachfolgenden Sprachproben nach ihrer Gesamtqualität“





With higher speech quality (MOS), speakers characteristics highlighted:

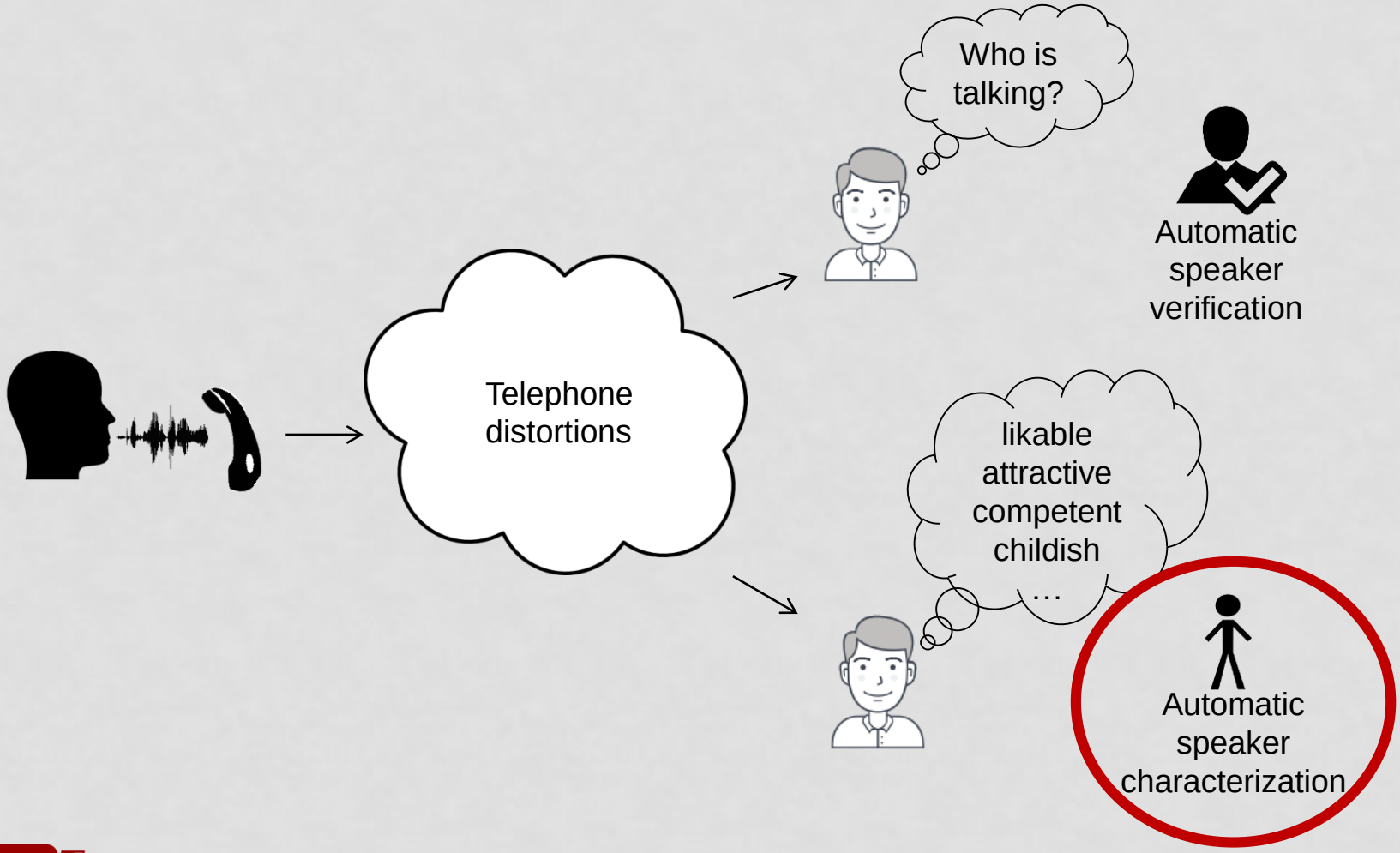
- unobtrusive, characterful, sympathetic, likable
- pleasant, attractive, beautiful
- impersonal, competent, calm, adult, decided, secure

- friendly, sympathetic, characterful
- interested, emotional, active, young
- impudent, competent, secure, decided, intelligent

Outline

My Postdoc:
ongoing work

Outline


Who is talking?
Automatic speaker verification
Telephone distortions
likable
attractive
competent
childish
…
Automatic speaker characterization

Machine learning (tools)

openSMILE:)
by audEERING™

eGeMAPS:

- Frequency
- Energy
- Spectral
- Temporal

(88)

caret
mlr

Machine learning (labels)

300 x 34 item ratings

1: non likable
2: secure
3: attractive
4: unsympathetic
5: indecisive
6: unobtrusive
7: distant
8: bored
9: emotional
10: not irritated
11: active
12: pleasant
(…)

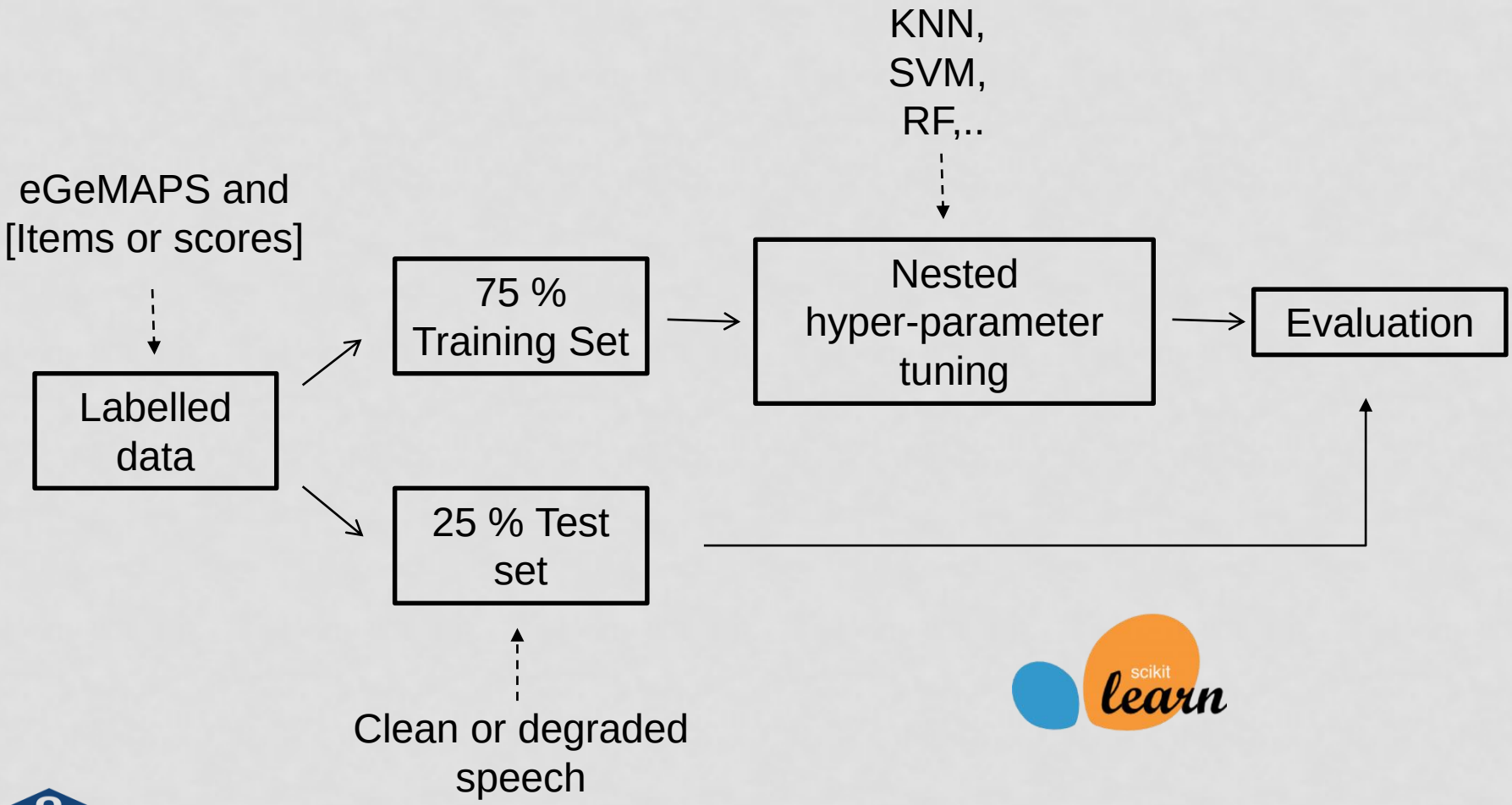
300 x 5 trait scores

1: warmth
2: attractiveness
3: confidence
4: compliance
5: maturity

Evaluation metrics:

- Average per-class accuracy (cls)
- RMSE (reg)

Machine learning (pipeline)

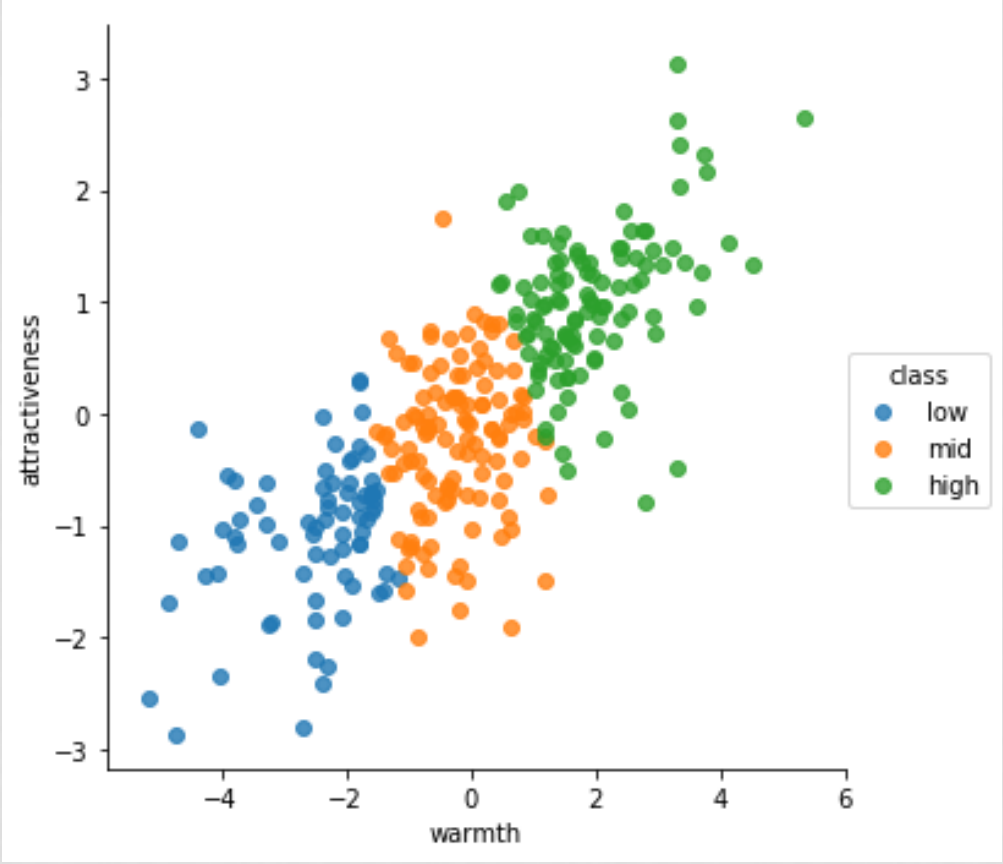


Machine learning (pipeline)

Nested hyperparameter tuning

- Split train data into A (80%) and B (20%) sets
- For each model:
  - RandomizedSearchCV on hyperparameters and on SelectKBest (Cross-validation)
  - Evaluate performance on B
- Choose best model based on performance on B
- Train best model with all train data

Binary classification

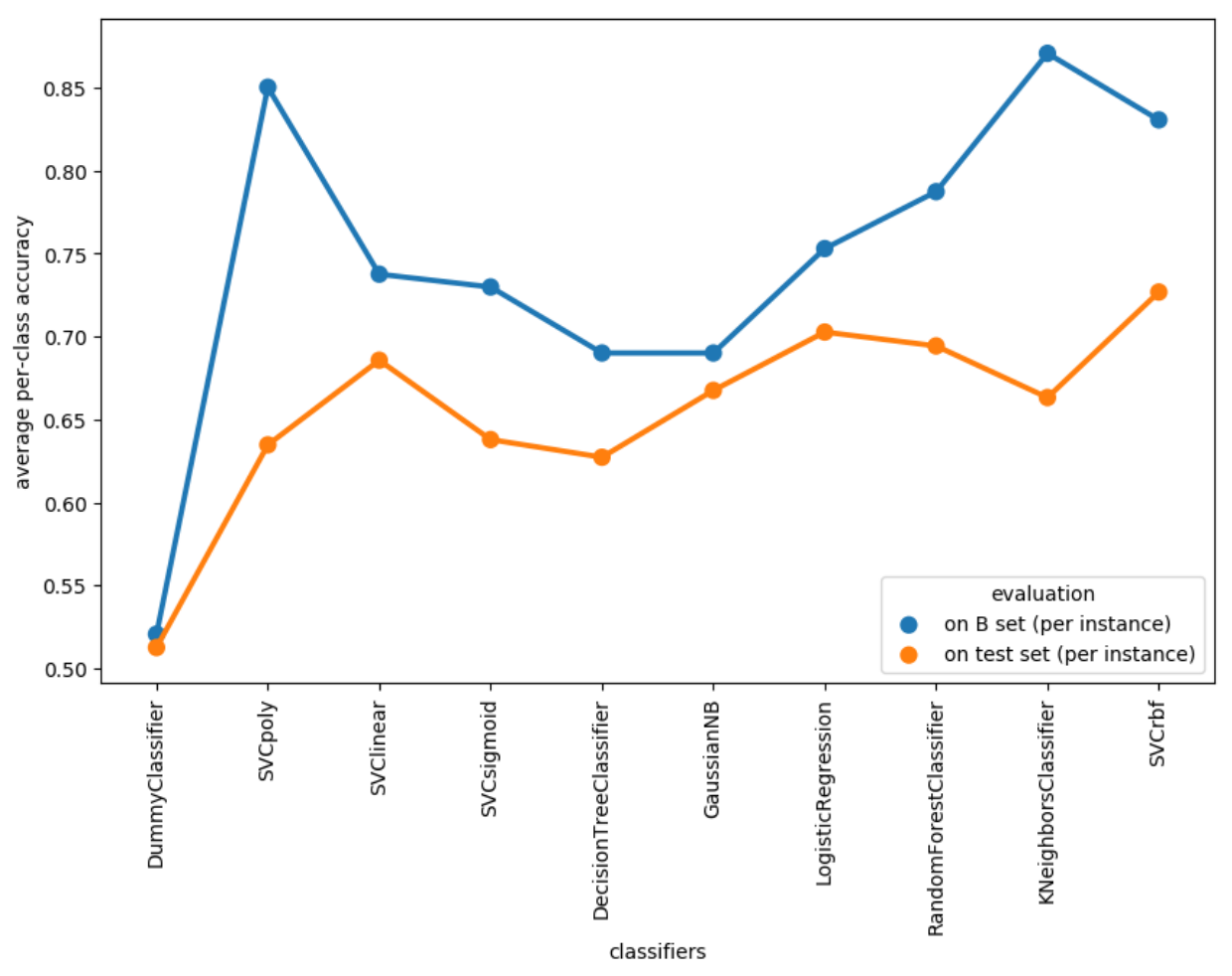


- WAAT
- K-means: 3 classes
- Binary: ”low” / “high”

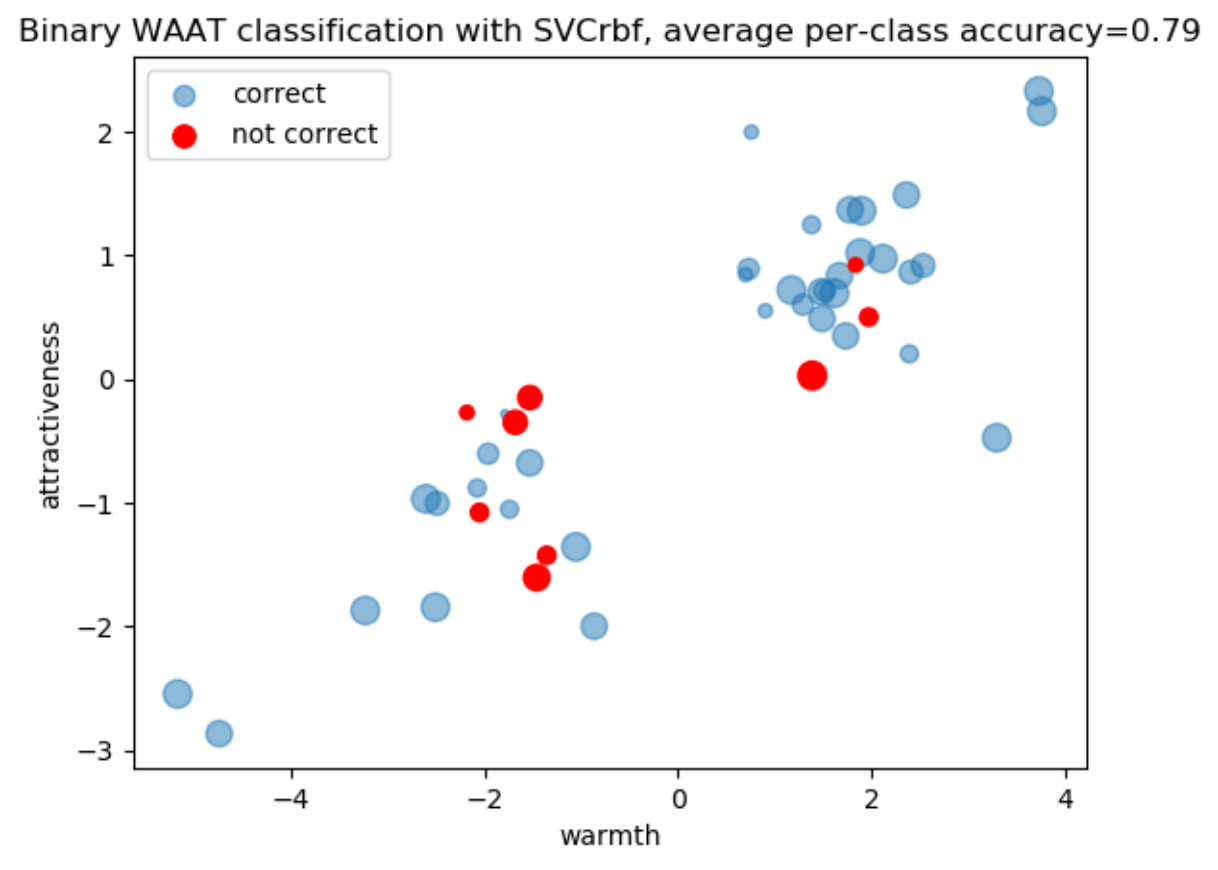
Binary classification

- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- Random Forest
- Support Vector Machines

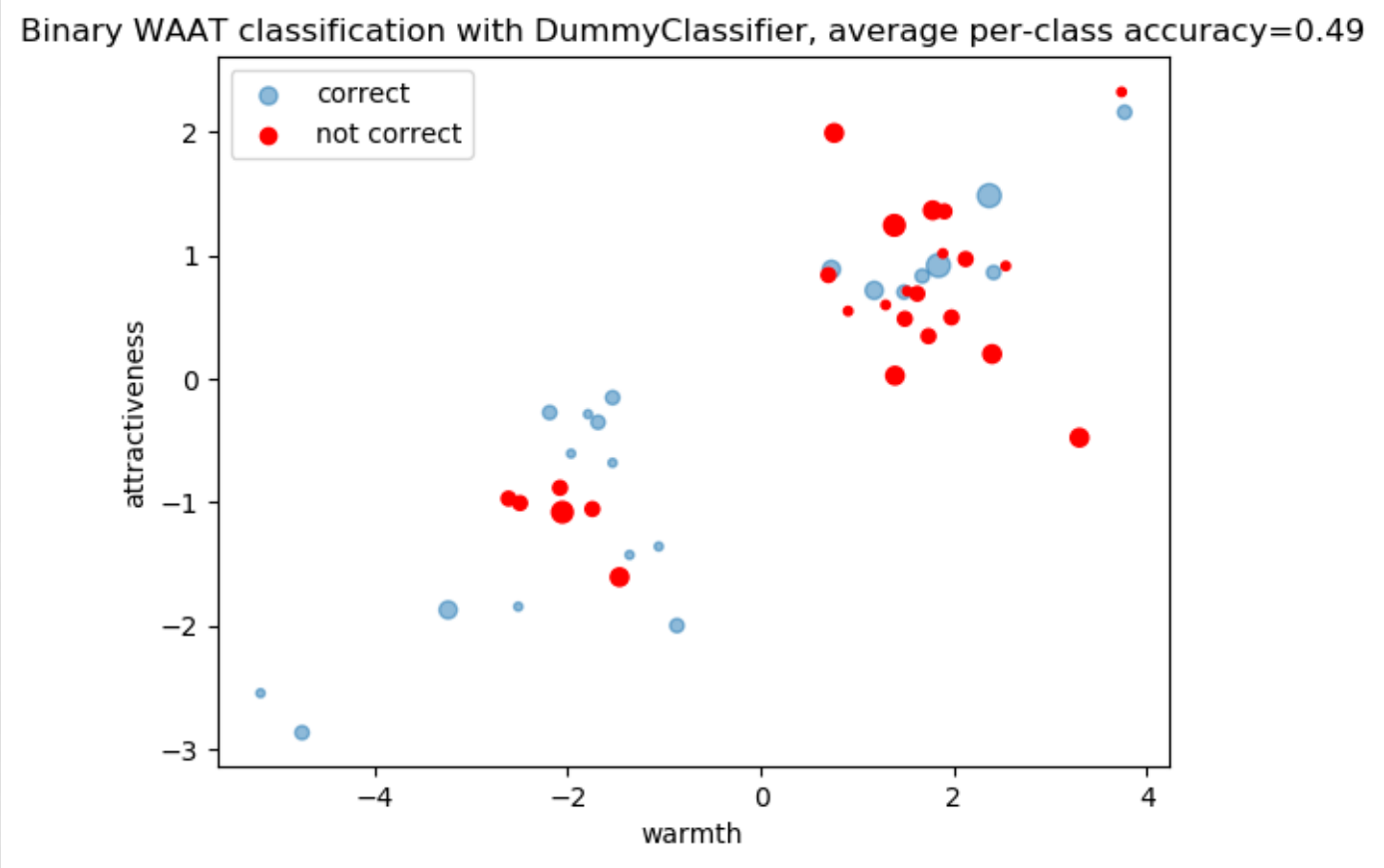
Binary classification



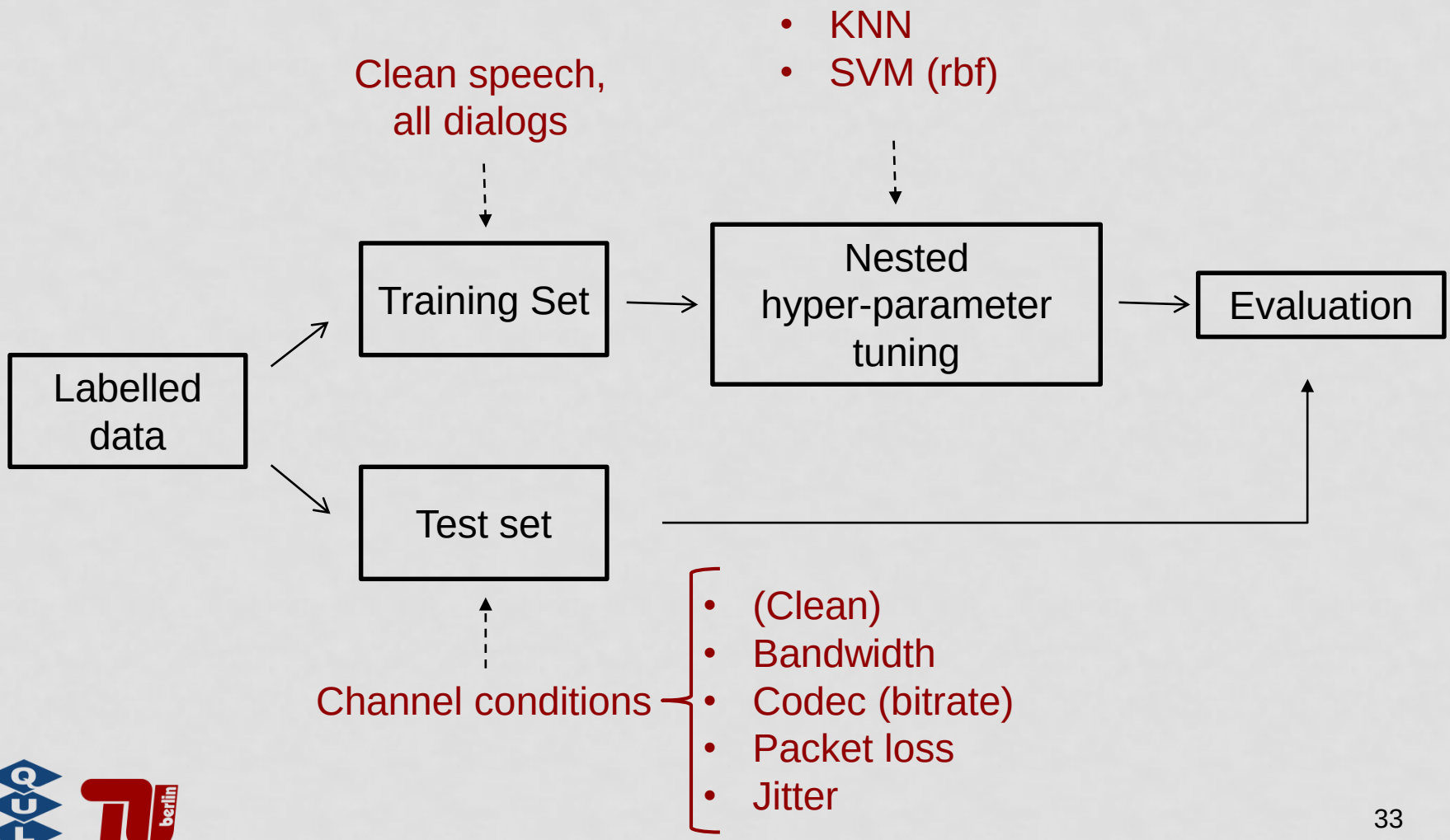
Binary classification



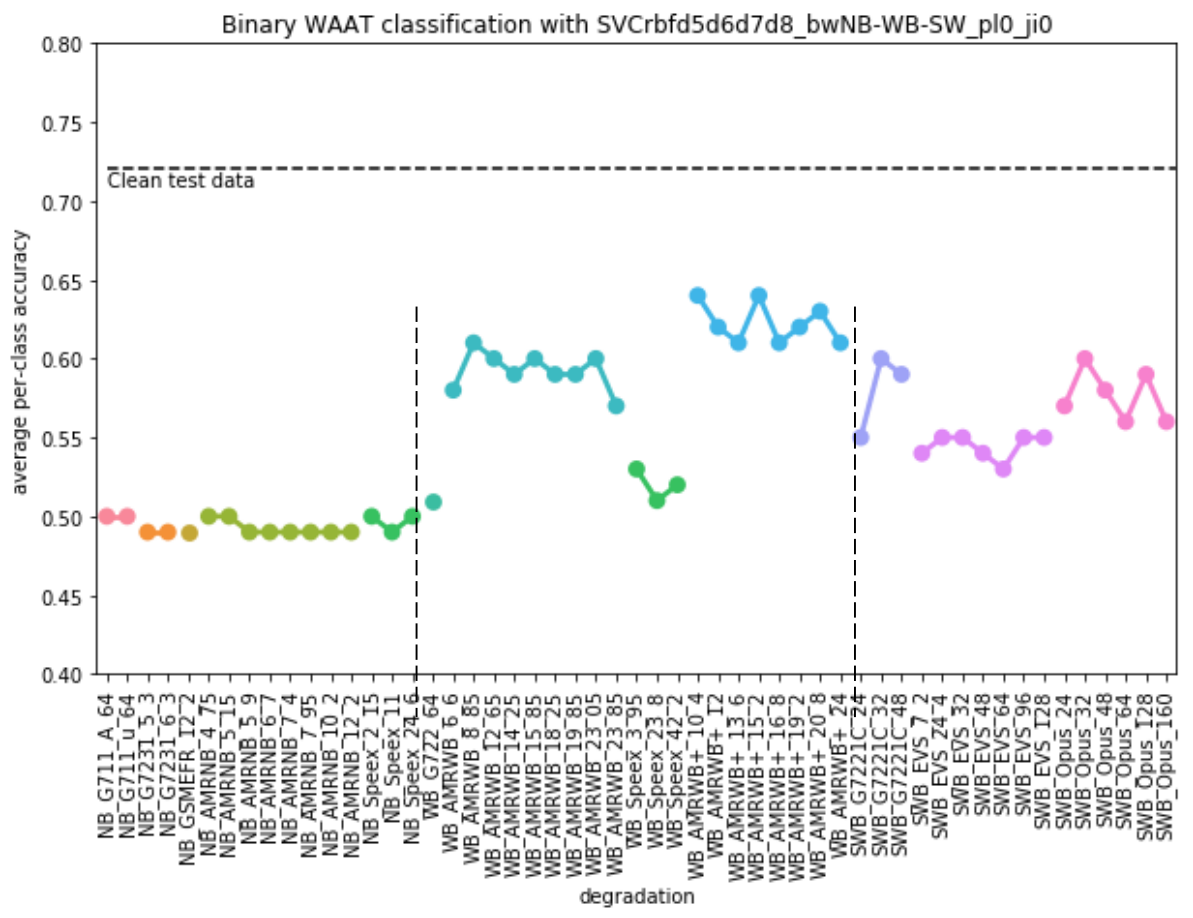
Binary classification



Effects of transmission channels



Effects of transmission channels



Ongoing / future work

- Improving the performance
  - Feature engineering
  - Other techniques for feature selection
  - “similarity” as feature
- Re-evaluate channel effects

My Postdoc: contributions

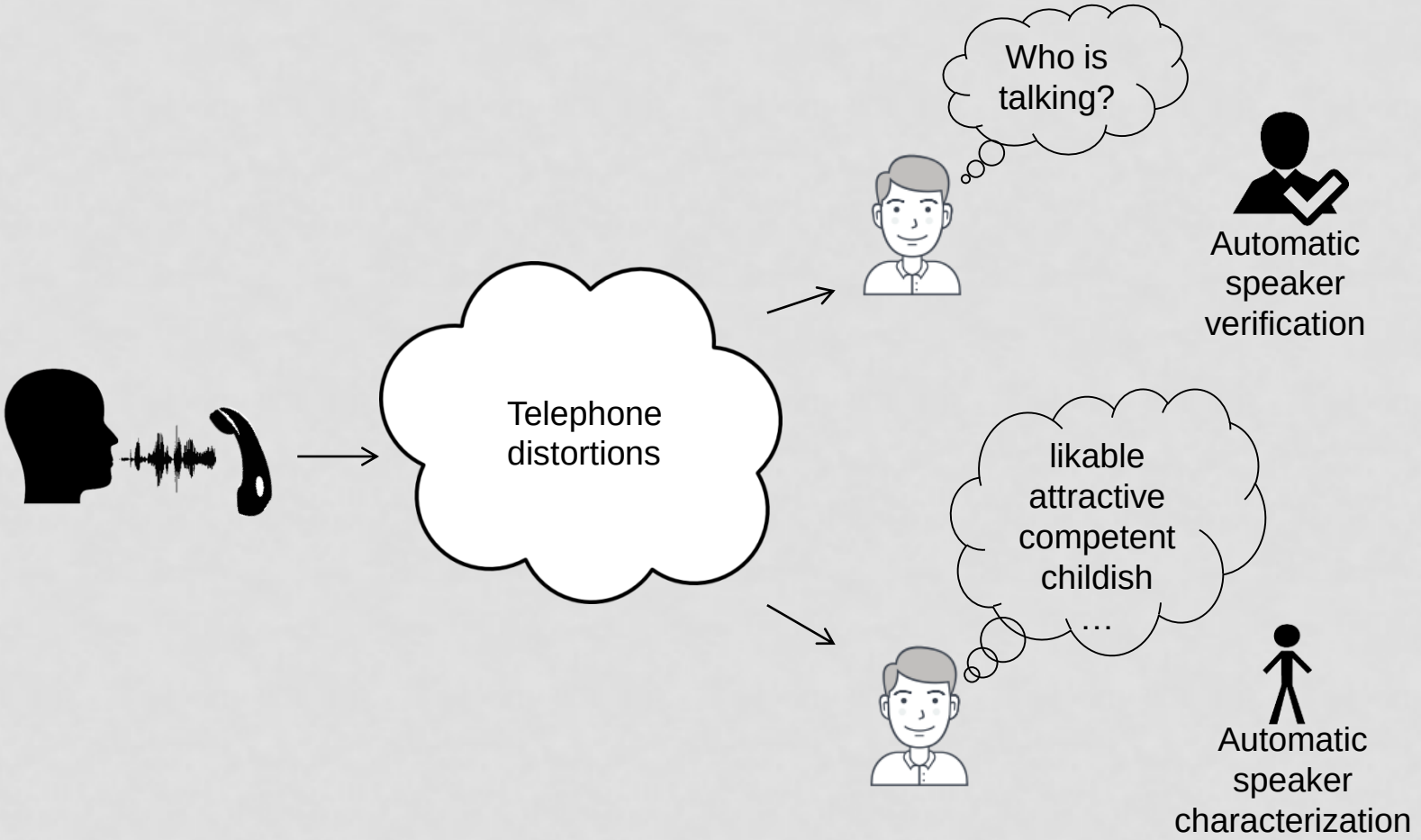
- ✓ New labelled speech database (much needed)

- ✓ Main traits of speakers' characteristics
- ✓ Speech quality ↔ speaker characteristics
- ✓ Subjective voice descriptions ↔ speaker characteristics

Automatic speaker characterization

- ✓ Important features for speaker characterization
- ✓ Open-source pipeline for classification and regression
- ✓ Effects of degraded test speech on performance

Outline


Who is talking?
Automatic speaker verification
Telephone distortions
likable
attractive
competent
childish
…
Automatic speaker characterization

Thank you for your attention!

Laura Fernández Gallardo
laufergall@gmail.com
http://www.qu.tu-berlin.de/?id=lfernandez

 laufergall

 laura-fernandez-gallardo

(Backup slides)

GMM-UBM

- UBM
  - is a GMM from a speaker population
  - typically 1024 or 2048 mixtures
  - each mixture defines the pdf of a cluster of n-dimensional MFCCs
  - fit by the EM algorithm

- Speaker-specific GMM
  - adapting a well-trained UBM with the target speaker data

$$LLR = log\left(p(X|\Theta_S)\right) - log\left(p(X|\Theta_{UBM})\right)$$

i-vectors

$$M = m + \boldsymbol{T}w$$

Speaker- and channel-dependent supervector

UBM mean supervector

Total-variability matrix

i-vector

MSR Identity Toolbox: A MATLAB Toolbox for Speaker Recognition Research, Microsoft

$$score(w_{target}, w_{test}) = \frac{\langle w_{target} | w_{test} \rangle}{\|w_{target}\| \, \|w_{test}\|}$$

ASV evaluations: i-vectors

- Conditions
  - Bandwidths: NB and WB. (No data for SWB)
  - Codecs
- Speech data
  - Development (fitting UBM): 670 speakers
  - Training / test (a): 112 speakers from TIMIT
  - Training / test (b): 991 speakers from NIST SRE 2010
- Features: MFCCs + delta + delta-delta
- UBMs: 1024 Gaussian mixtures
- 400 total factor (i-vector dimensionality)

ASV evaluations: i-vectors

Relative EER reduction NB to WB:
32% to 57% (clean)
30% to 64% (transmitted)

	TIMIT_test EER (%)	NIST SRE10 c1 EER (%)
Clean 4kHz	3.41	8.48
G.711	**4.29**	**10.10**
AMR-NB	5.01	10.99
Clean 8kHz	1.46	5.76
G.722	**1.80**	7.07
AMR-WB	2.52	7.07

Extracting speech features

VOICEBOX: Speech Processing Toolbox for MATLAB, by Mike Brookes, Imperial College, London

openSMILE:)
by audEERING™

Speech features

For each speaker…

openSMILE:)
by audEERING™

88 eGeMAPS features

Box Cox transformed
centered
scaled
collinear features removed

x 2 genders
x 5 dimensions

linear correlation

"Feature importance"

Speech features

- Higher warmth:
  1. Higher F0 range
- Higher attractiveness:
  1. Higher F0 range
  2. Higher std of F0
  3. Lower median F0

- Higher confidence:
  1. Lower median F0
  2. Higher F0 range
- Higher compliance:
  1. Lower std length voiced segments
  2. Lower std of F1 frequency
  3. Lower std of falling slope for loudness
  4. Higher F1 and F2 bandwidth
- Higher maturity:
  1. Lower median F0
  2. Higher std of F3 frequency
  3. Higher std of F3 bandwidth

Most warm+attractive

Least warm+attractive

Speech features

- Higher warmth:
  1. Higher F1 frequency
  2. Higher F0 range
  3. Higher std of F2 bandwidth
  4. Higher std of spectral flux

- Higher attractiveness:
  1. Higher F1 frequency
  2. Higher std of F2 bandwidth
  3. Lower std of F1 frequency
  4. Higher F0 range
  5. Lower spectral slope 0-500Hz

- Higher compliance:
  1. Lower std of F1 frequency
  2. Higher F1 frequency
  3. Lower loudness range

- Higher confidence:
  1. Higher std of falling slope for loudness
  2. Higher F0 range

- Higher maturity:
  1. Lower median F0
  2. Higher mean mfcc4
  3. Lower F1 frequency
  4. Higher mean mfcc2

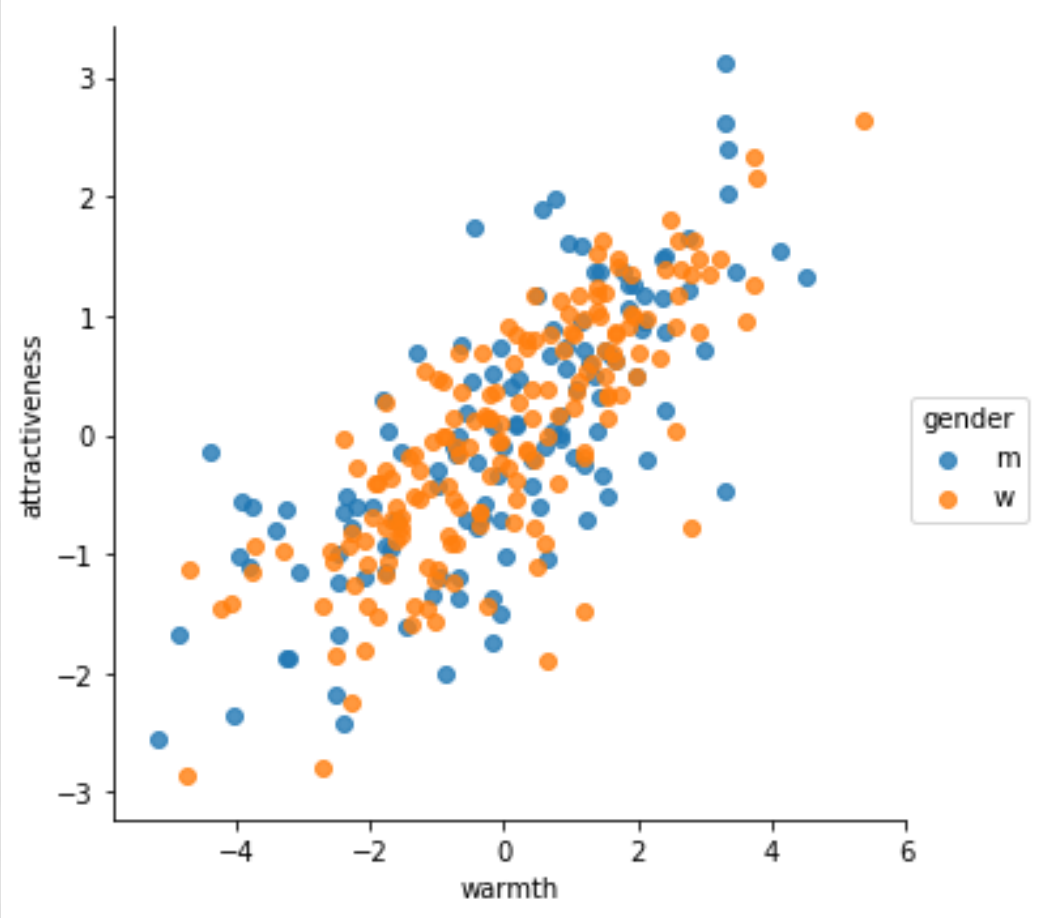
Speaker characteristics

- Factor analysis for male and for female speakers separately

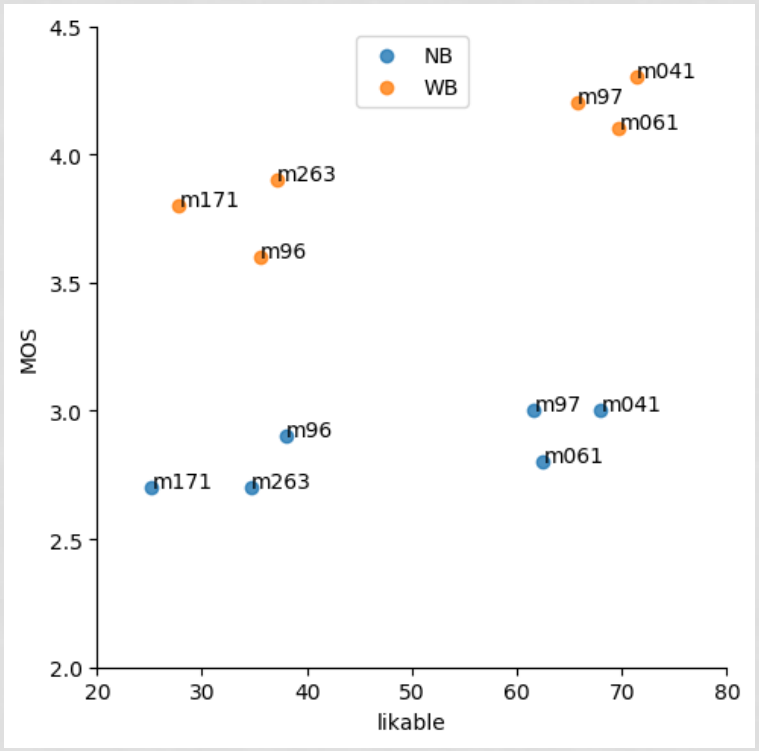
	Warmth	Attract.	Confid.	Compl.	Matur.
herzlich	0.85				
mitfuehlend	0.84				
distanziert	-0.76				
freundlich	0.59				
verstaendnislos	-0.58				
unsympatisch	-0.52				
nicht.genervt	0.51				
attraktiv		0.85			
haesslich		-0.79			
angenehm		0.58			
interessant		0.48			
sicher			1.00		
unentschieden			-0.6		
gehorsam				0.87	
zynisch				-0.71	
alt					0.82
kindlich					-0.73

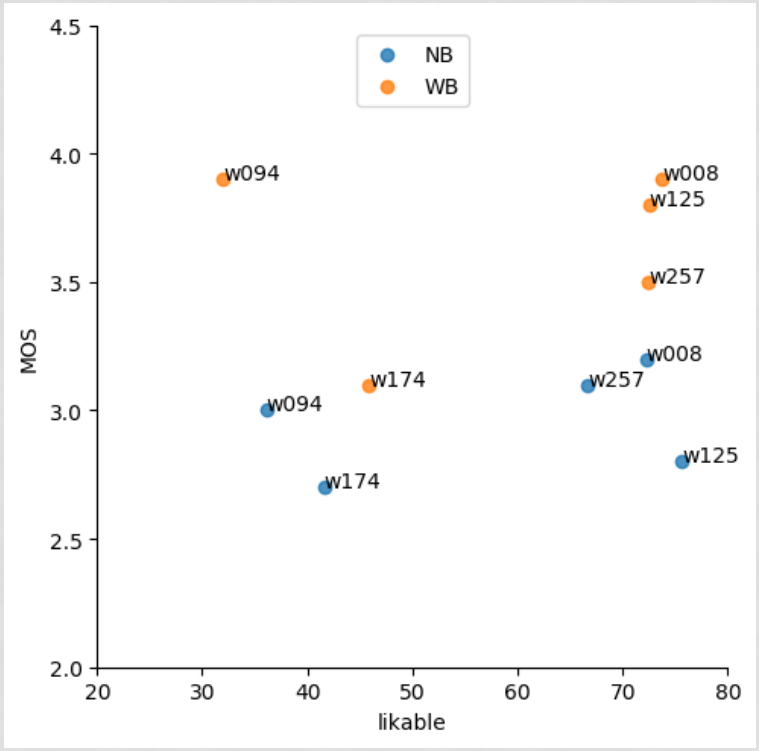
	Warmth	Attract.	Compl.	Confid.	Matur.
herzlich	0.84				
mitfuehlend	0.84				
distanziert	-0.78				
freundlich	0.56				
verstaendnislos	-0.49				
nicht.genervt	0.49				
unsympatisch	-0.45				
attraktiv		0.83			
haesslich		-0.81			
angenehm		0.59			
gehorsam			0.80		
zynisch			-0.72		
sicher				0.82	
unentschieden				-0.81	
kindlich					-0.81
alt					0.68

The “WAAT” space

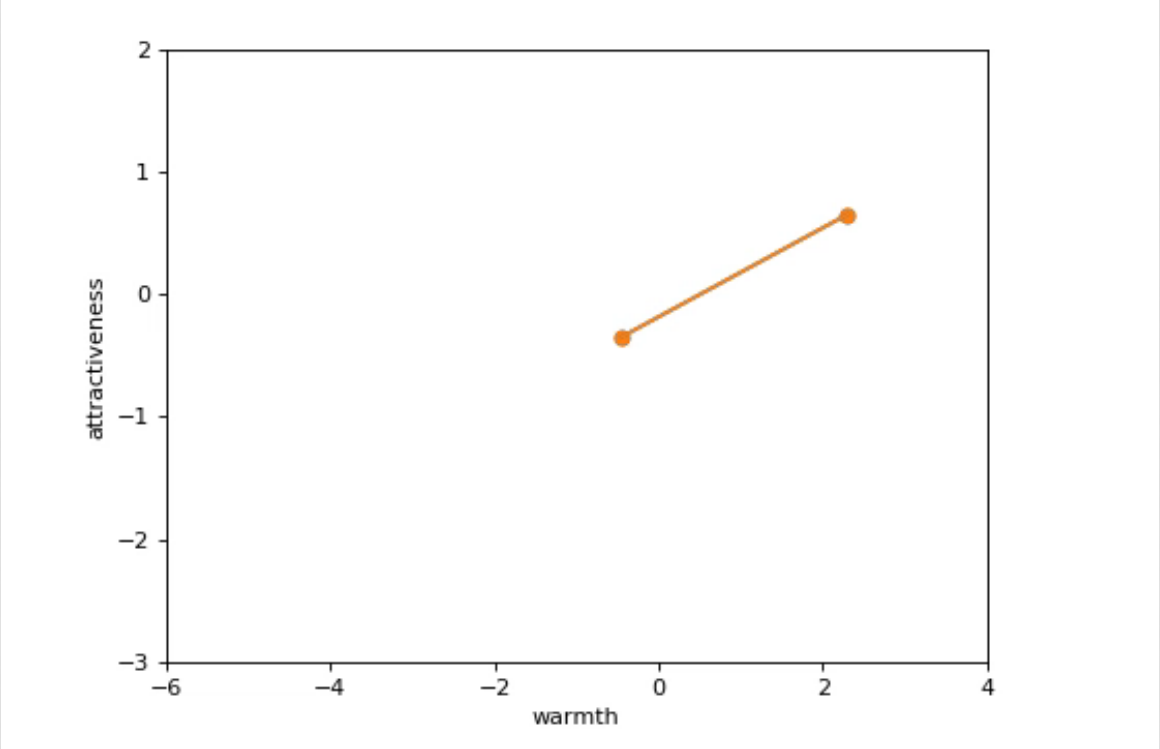


MOS - Speaker characteristics





Intuition of error in regression (WAAT)



Multioutput regression

Random Forest:

Overall RMSE = 1.07

Baseline:

Overall RMSE = 1.18