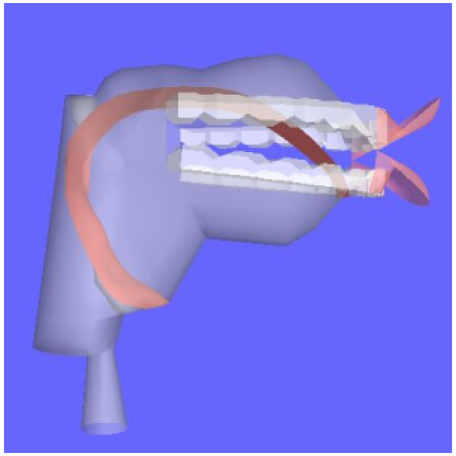
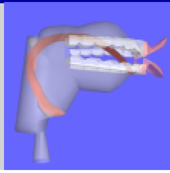
Machine vs. Human: A Cross-Discipline Study on Synthetic Speaker Age Recognition



Eva Lasarcyk, Michael Feld, Christian Müller

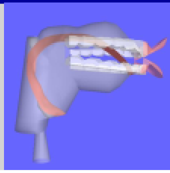
FEAST
May 6, 2009
Saarbrücken

WHAT are we trying to achieve?!
Idea of a cross-discipline study on vocal age



- Imagine you are talking on the phone to someone you don‘t know. Without seeing the person you can make some reasonable assumptions about e.g. their age. But you can never be sure that the young lady you think you‘re talking to is in reality an elderly woman with a “young“ voice impression.
- How well does age recognition work over the phone anyway? (Limited bandwidth; already a tough task)
- **Exploratory** nature of study with *synthetic* voices. (Limited experience, since we are very experienced only in the *natural* world; makes it an even tougher task.)
- Plus: Comparing the "human ear" with an age classifier.

Motivation of This Talk



- Show an example of collaboration between speech sciences (aka phonetics) and speech technology
- Present an explorative model of synthetic vocal aging
- Compare human listeners and an automatic age classifier
- Discuss what we can learn from this approach in order to improve the age classification system

Our Goals/Research Questions

1. Can age cues that are derived from the literature be implemented into synthetic voices in order to let human listeners recognize the age class *Young, Adult, Senior*?

2. What is the relative importance of individual cues for human perception of speaker age?

3. Would a speaker age recognition system, which is solely trained on natural voices, produce meaningful results when presented with the same synthetic voices?
    1. Are the voices natural enough to “fool” the system?
    2. Does the system (with its statistical model based on short-term cepstral features) in fact catch up some of the theoretically motivated age cues?

Problem

- A person’s voice changes due to
  - **Aging**
  - Emotional conditions
  - Pathological conditions
  - …

- Knowledge applicable for
  - Security
  - Medical applications
  - Speech technology
  - …
  - Scientific curiosity

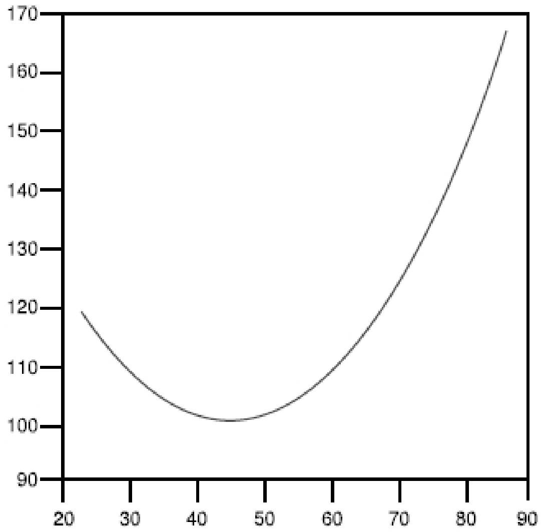
Physiological Changes

- Vocal tract lengthening
- Reduction in pulmonary function
- Ossification of laryngeal cartilages
- Increased vocal fold stiffness
- Reduced vocal fold closure

- Habits? Sicknesses?

Acoustic Changes

- Mean F0
  - Raised in old males
- Increased F0 variability
- Lower formant frequencies
- Greater noise
- Slower speaking rate



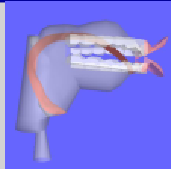
Müller 2005/Linville 2001

Findings in the literature are sometimes contradictory

Outline

- Anatomy of Vocal Aging
- Modeling of synthetically aged voices
- Evaluation "Systems": Listeners and age classifier
- Results
- Conclusions and Discussion (work in progress)

Modeling with VocalTractLab (Birkholz 2006, Birkholz&Kröger 2006)

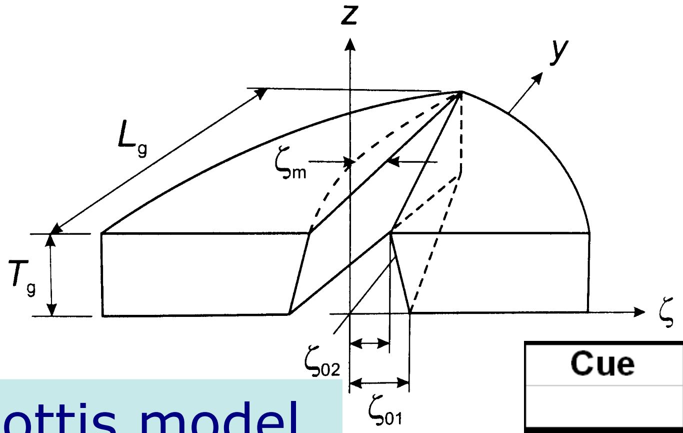


F0 = F0base+sin(A1*2pi*JF)+sin(A2*2pi*JF2)+...

3 age classes: Young (15-24), Adult (25-54), Senior (55-80)

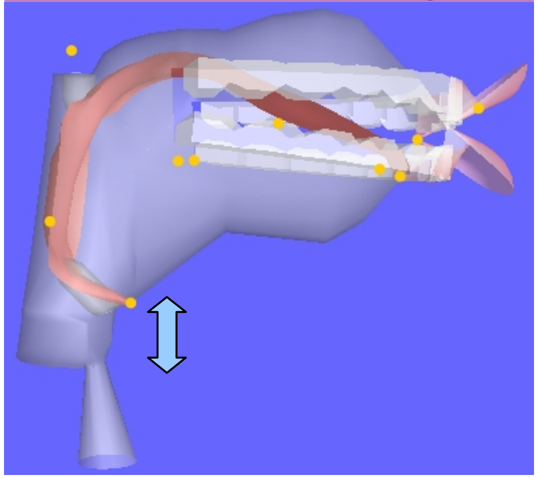
12 "voices" per age class

Contents: aI-aU, aU-OI



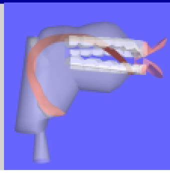
Glottis model

Vocal tract shape



Cue					Class	
				Young	Adult	Senior
F0 [Hz]		High		130	100	150
		Medium		120	107,5	127,5
		Low		110	115	105
VQ	noise	breathy	Displacement	0,4	0,6	0,8
			Arytenoid Area	-1	1	2
			Vertical Phase Difference	0,1	0,2	0,3
		modal	Displacement	0,5	0,7	0,9
			Arytenoid Area	0	1,5	2,5
			Vertical Phase Difference	0,2	0,3	0,4
	irregularities		Shimmer	30/40	40/50	50/60
		regular	Jitter Amplitudes	0,3	0,5	1,3
		irregular	Jitter Amplitudes	0,4	0,6	1,6
Larynx [VTL:HY]				high	medium	low

Evaluation "Systems" I

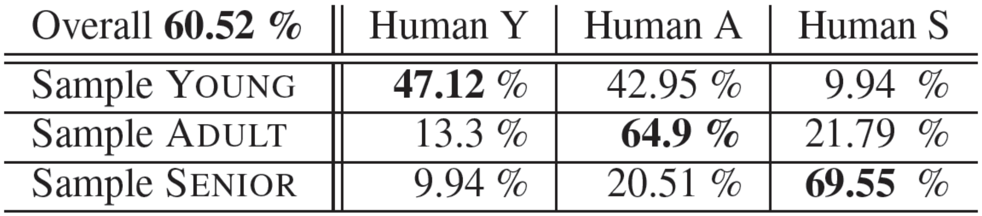


- Forced-choice classification task
  - Web-based listening test with warm-up procedure
  - 12 voices in 3 age classes, with two wordings
  - 2 presentations of each stimulus (144 total)
  - Possibility to provide feedback at end of test
- 26 Listeners (1 Young, 20 Adult, 5 Senior)
  - More or less naive to synthetic voices
  - Thanks to the ones of you who participated!

Ex. 1 Ex. 2 Ex. 3

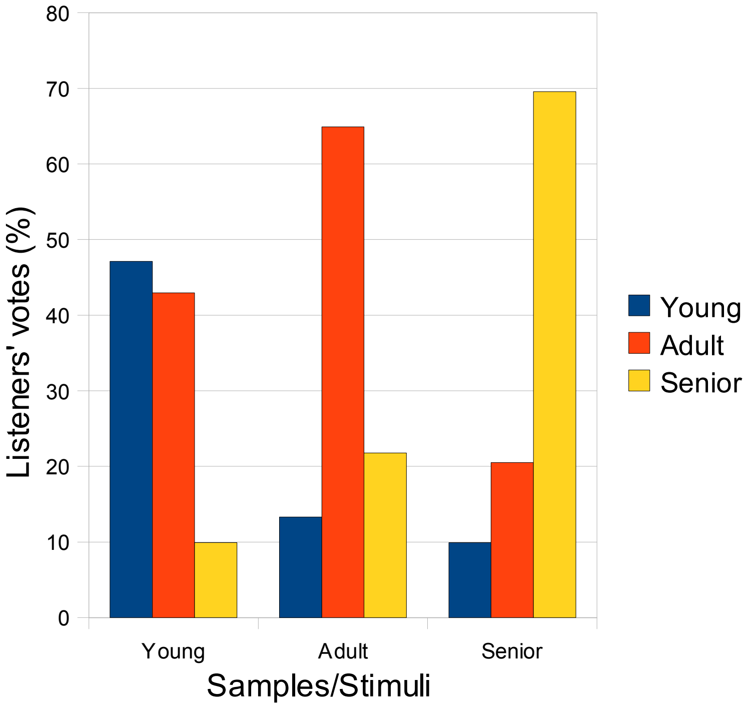
Results I: Listeners' Classification Accuracy

- Confusion matrix (3744 votes)



Overall **60.52 %**	Human Y	Human A	Human S
Sample YOUNG	**47.12** %	42.95 %	9.94 %
Sample ADULT	13.3 %	**64.9 %**	21.79 %
Sample SENIOR	9.94 %	20.51 %	**69.55** %

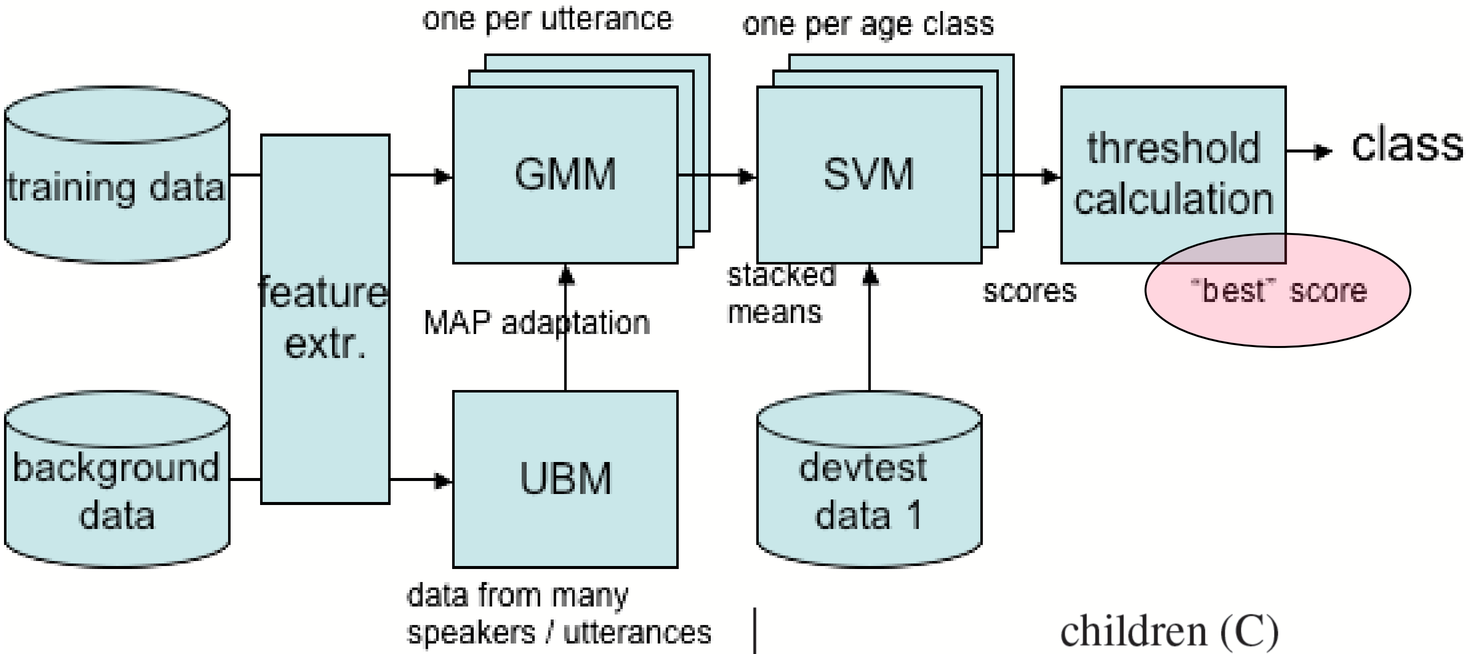
Young: High Young F0, Senior: Low F0



- Verbal feedback of participants
  - Human/synthetic/mechanical
  - Tuning in, discrimination/identification
  - Jittery = old, young/adult hard
  - Fitness
- Consistency of answers (Which were the "hard" voices?)

Evaluation "Systems" II

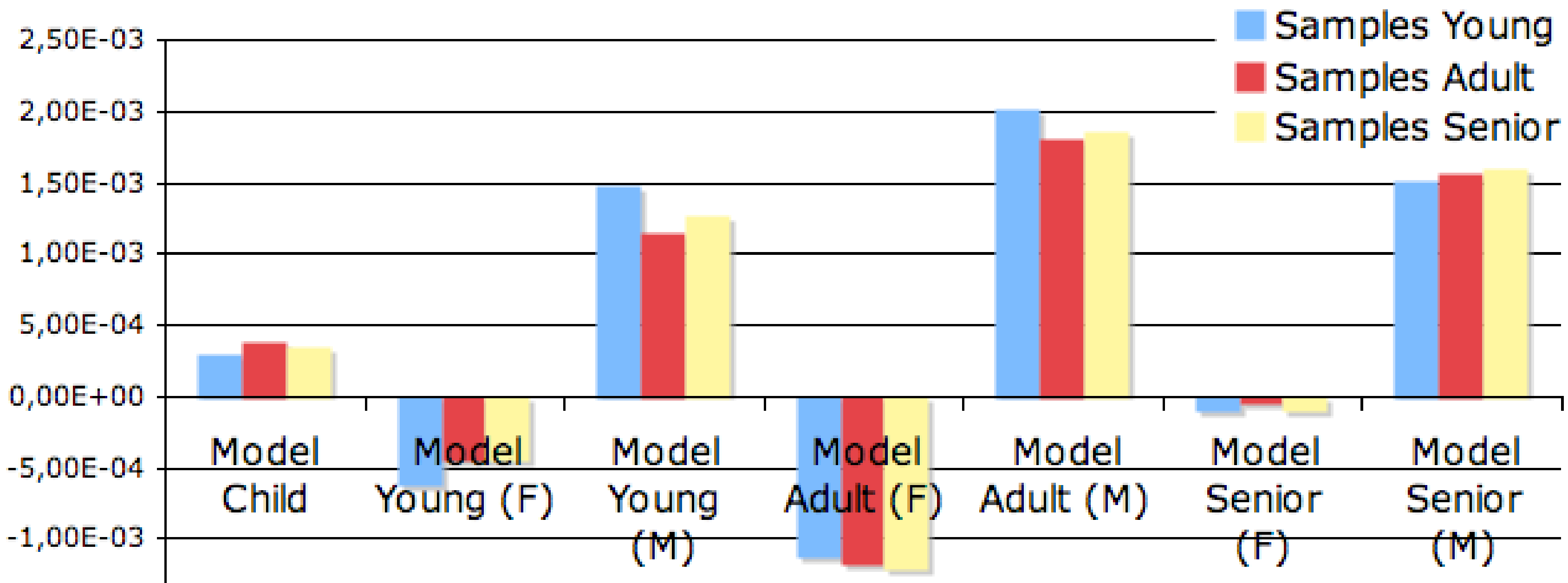
- Age classification system
  - Trained on conversational telephone speech
  - Not tuned for test data (synthetic)



children (C)		<15
young females (YF)	young males (YM)	15 – 24
adult females (AF)	adult males (AM)	25 – 54
senior females (SF)	senior males (SM)	55 – 80

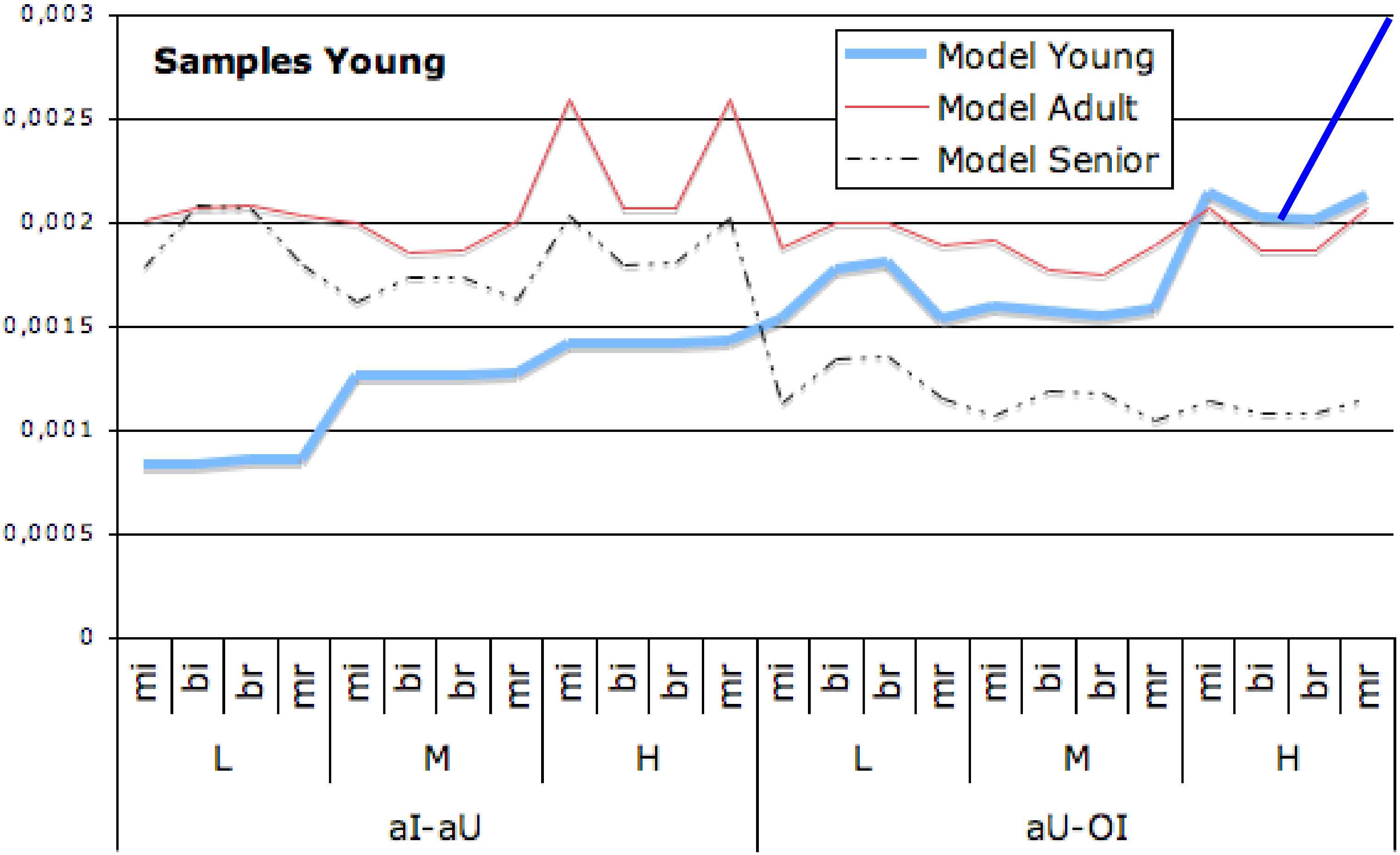
Results II: Age classifier

- Mean scores per age model
  - Reasonable output in general ("male" models)



Results II:
Scores of "male models" for YOUNG samples

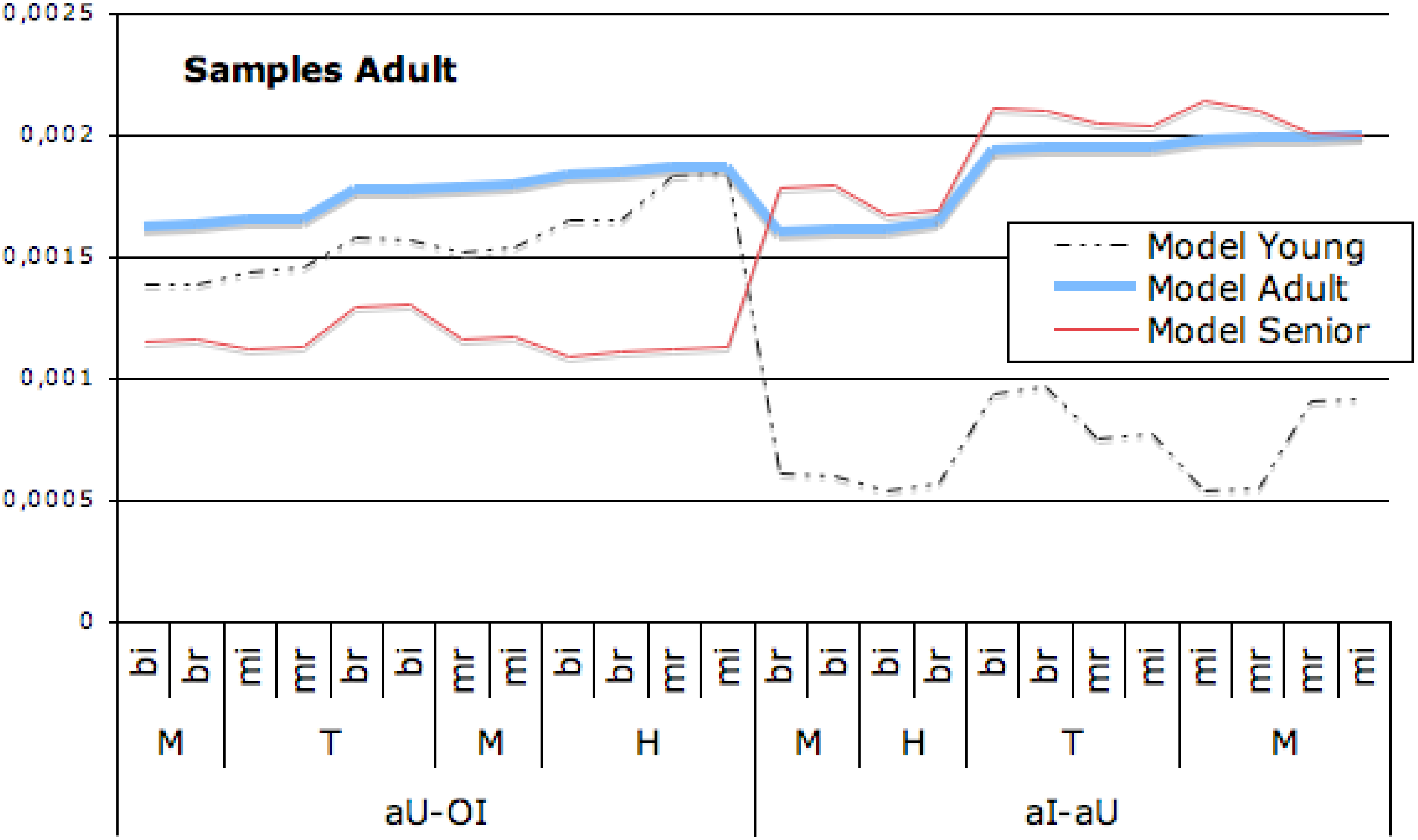
- As a function of synthetic age cues: Clear effect
- Only if target model scores highest = Correct classification



Curve of target model

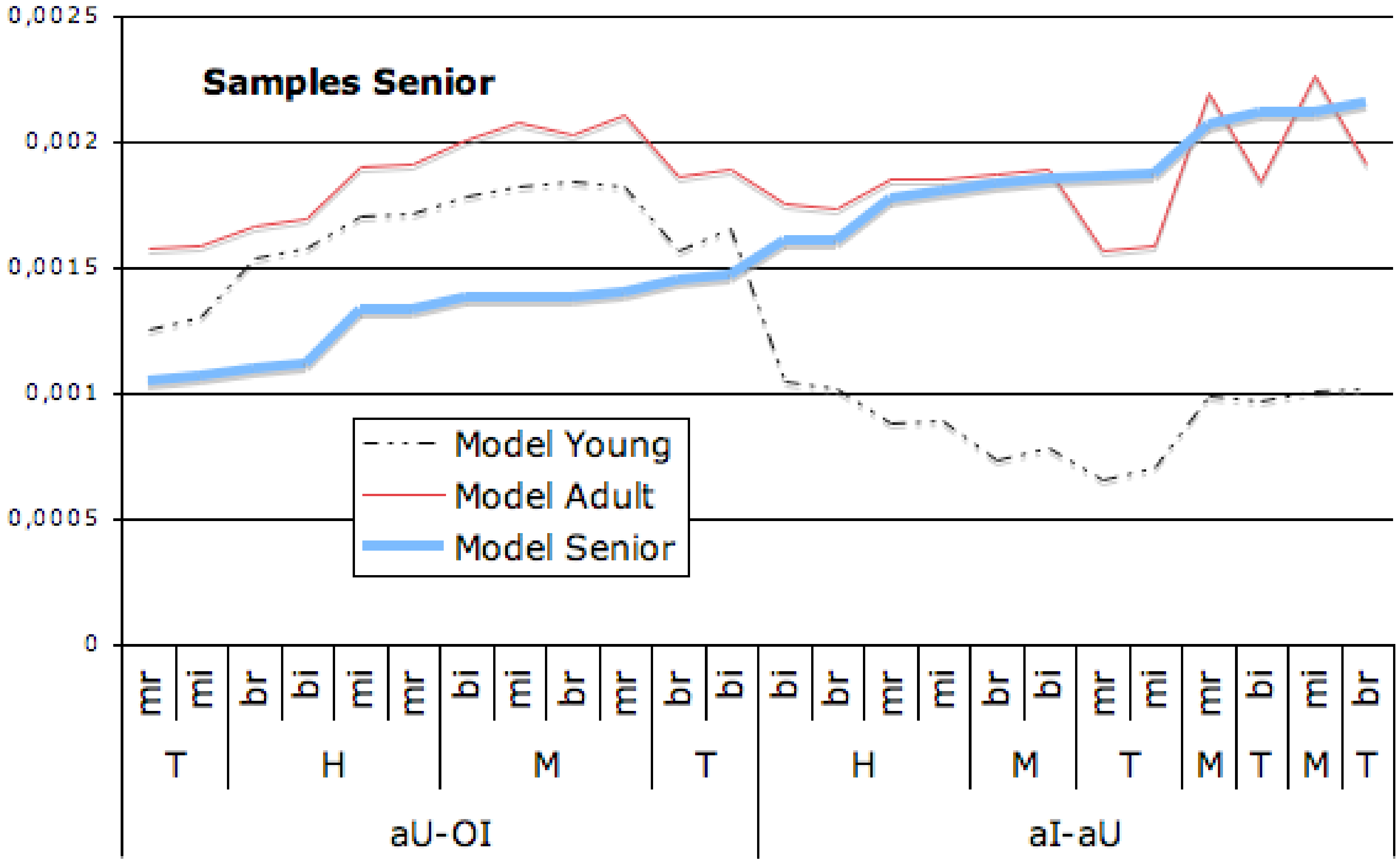
Results II:
Scores of "male models" for ADULT samples

- Content largest effect, other cues not so clearly sorted
- Within-class variance higher than for YOUNG in training (?)



Results II:
Scores of "male models" for SENIOR samples

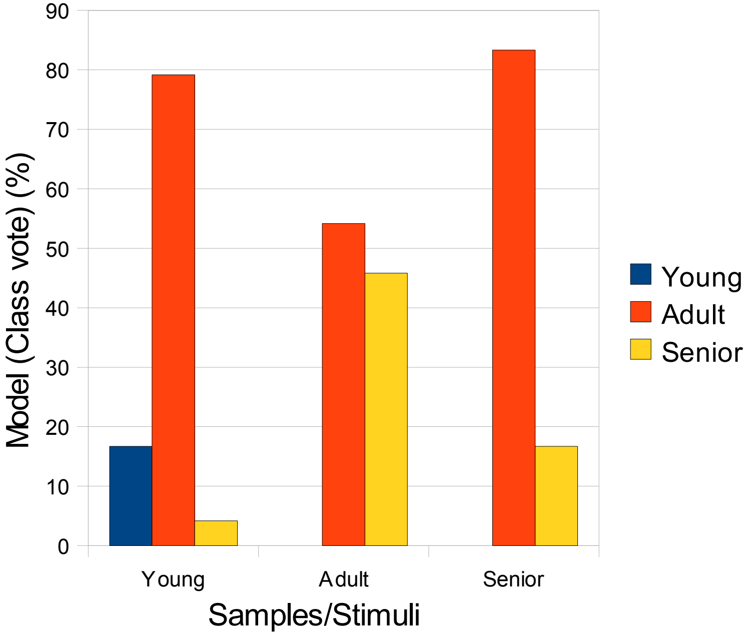
- Similar picture as with ADULT samples



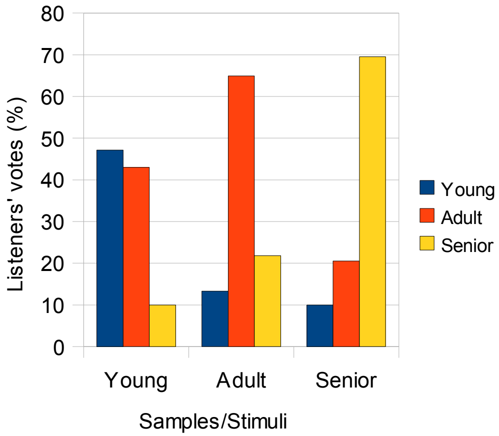
Results II: Age Classifier Accuracy

- Confusion matrix

Overall **29.17 %**	Model Y	Model A	Model S
Samples Young	**16.67 %**	79.17 %	4.17 %
Samples Adult	0 %	**54.17 %**	45.83 %
Samples Senior	0 %	83.33 %	**16.67 %**



- ADULT wins often
- Jittery = Old?



Our Goals/Research Questions Revisited

1. Can age cues that are derived from the literature be implemented into synthetic voices in order to let human listeners recognize the age class *Young, Adult, Senior*?

2. What is the relative importance of individual cues for human perception of speaker age?

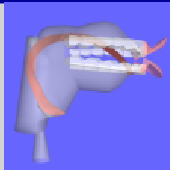
3. Would a speaker age recognition system, which is solely trained on natural voices, produce meaningful results when presented with the same synthetic voices?
    1. Are the voices natural enough to "fool" the system? (Meaningful scores)
    2. Does the system (with its statistical model based on short-term cepstral features) in fact catch up some of the theoretically motivated age cues?

Conclusions and Discussion

- Limits of the stimuli set due to design reasons
- Indications of quality of the age model (consistency)
- General topic of synthetic "world" and naive listeners
- Ways to improve the age classifier? (Control conditions)

- Successful collaboration between speech sciences and speech technology

References



P. Birkholz. 3D-Artikulatorische Sprachsynthese. Dissertation, published by Logos (Berlin), 2006.

P. Birkholz and B.J. Kröger, “Vocal tract model adaptation using magnetic resonance imaging,” in Proc. 7th ISSP, Ubatuba, 2006, pp. 493–500.

S.E. Linville, Vocal Aging, Singular, 2001.

C. Müller, Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Twolayered Context-Sensitive Speaker Classification on the Example of Age and Gender], Ph.D. thesis, Computer Science Institute, University of the Saarland, Germany, 2005.