Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis

Source: Speech Communication 52 (2010)

Author : Michael Pucher, Junichi Yamagishi

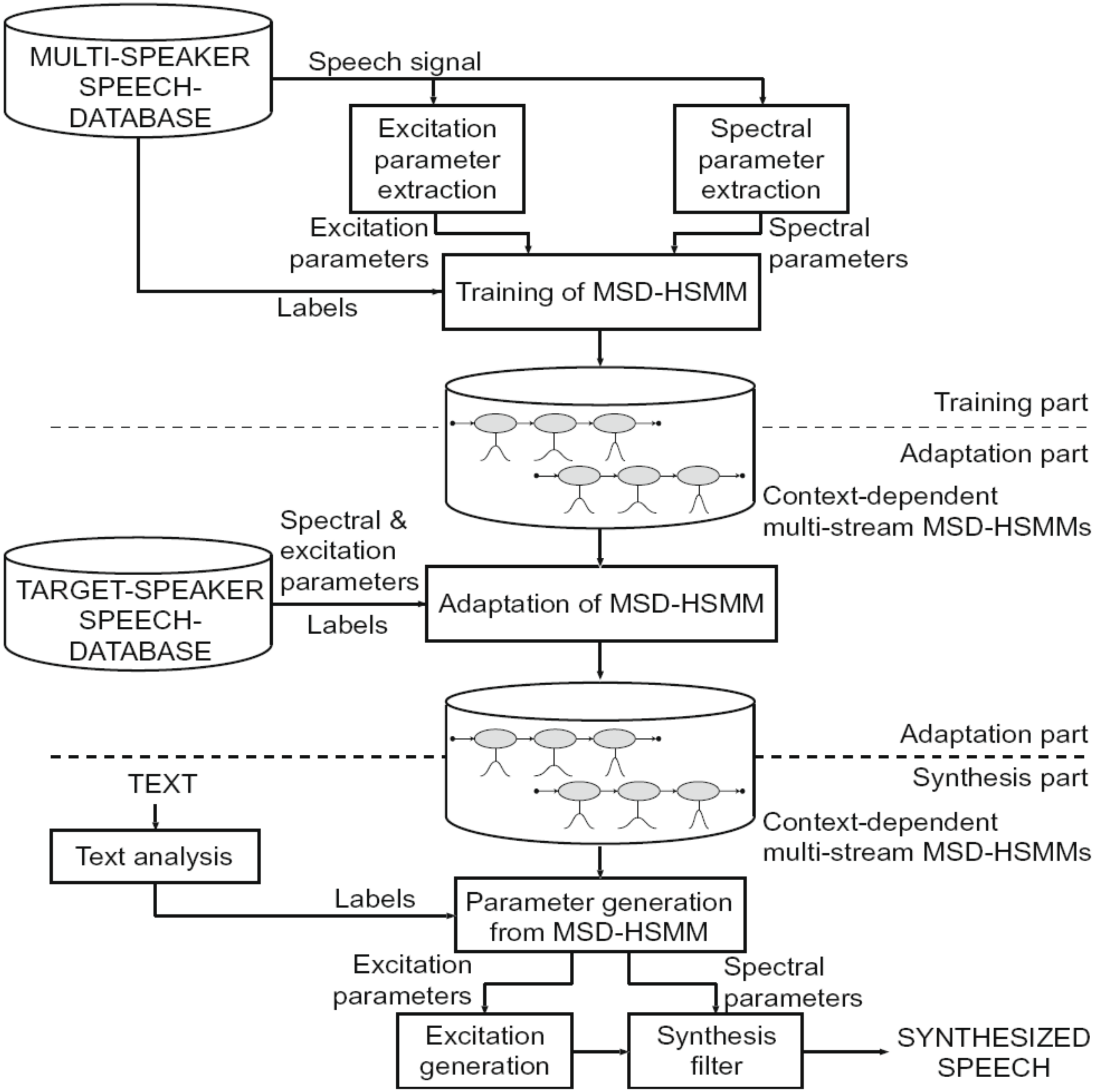
Professor : 陳嘉平

Reporter : 楊治鏞

Introduction

- An HMM-based speech synthesis framework is applied to both standard Austrian German and a Viennese dialectal variety and several training strategies for multi-dialect modeling such as dialect clustering and dialect-adaptive training are investigated.

- For bridging the gap between processing on the level of HMMs and on the linguistic level, we add phonological transformations to the HMM interpolation and apply them to dialect interpolation.


MULTI-SPEAKER SPEECH-DATABASE
Speech signal
Excitation parameter extraction
Spectral parameter extraction
Excitation parameters
Spectral parameters
Labels
Training of MSD-HSMM
Training part
Adaptation part
Context-dependent multi-stream MSD-HSMMs
Spectral & excitation parameters
TARGET-SPEAKER SPEECH-DATABASE
Labels
Adaptation of MSD-HSMM
Adaptation part
Synthesis part
Context-dependent multi-stream MSD-HSMMs
TEXT
Text analysis
Labels
Parameter generation from MSD-HSMM
Excitation parameters
Spectral parameters
Excitation generation
Synthesis filter
SYNTHESIZED SPEECH

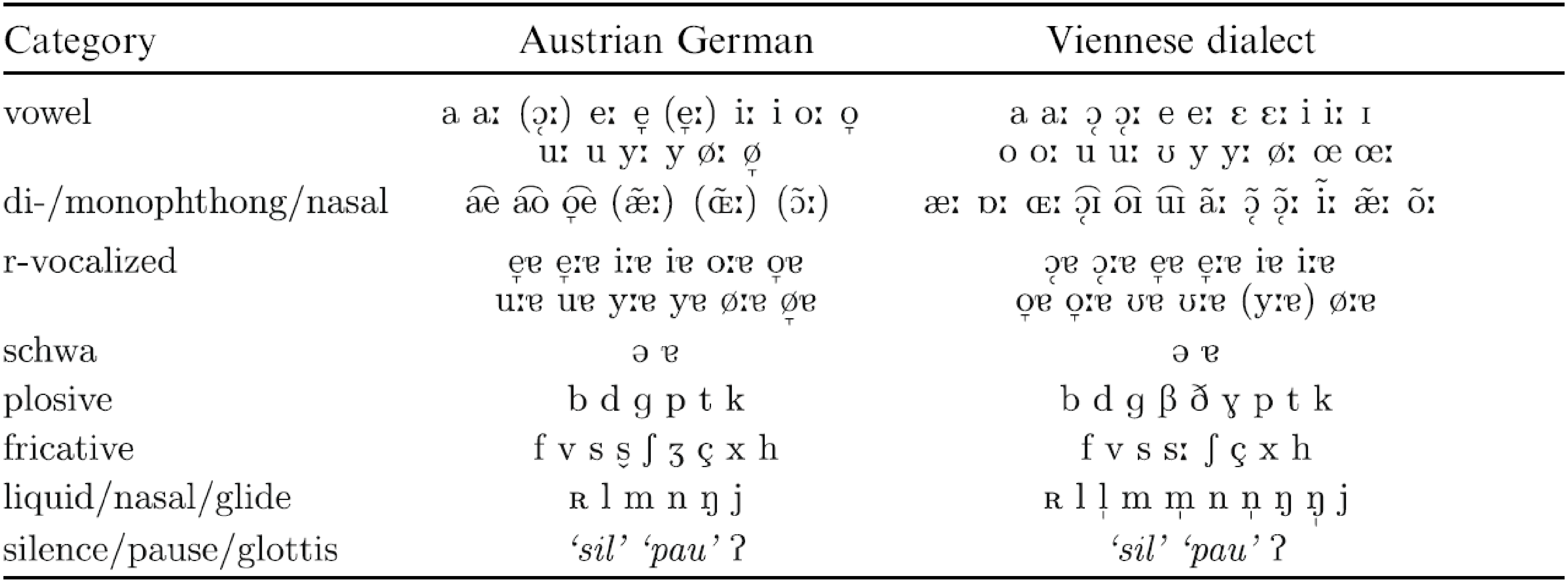
Speech database for Austrian German and Viennese dialect

- Phonesets used for standard Austrian German and Viennese dialect are shown in Table 1.

- For training and adaptation of Austrian German and Viennese dialect voices, a set of speech data comprising utterances from 6 speakers was used.

- Table 2 shows details of the speakers and number of utterances recorded for each.

Speech database for Austrian German and Viennese dialect



Category	Austrian German	Viennese dialect
vowel	a aː (ɔ̜ː) eː e̝ (e̝ː) iː i oː o̝ uː u yː y øː ø̝	a aː ɔ̜ ɔ̜ː e eː ɛ ɛː i iː ɪ o oː u uː ʊ y yː øː œ œː
di-/monophthong/nasal	a͡e a͡o o̝͡e (æ̃ː) (ɶ̃ː) (ɔ̃ː)	æː ɒː ɶː ɔ̜͡ɪ o͡ɪ u͡ɪ ãː ɔ̜̃ ɔ̜̃ː ĩː æ̃ː õː
r-vocalized	e̝ɐ e̝ːɐ iːɐ iɐ oːɐ o̝ɐ uːɐ uɐ yːɐ yɐ øːɐ ø̝ɐ	ɔ̜ɐ ɔ̜ːɐ e̝ɐ e̝ːɐ iɐ iːɐ o̝ɐ o̝ːɐ ʊɐ ʊːɐ (yːɐ) øːɐ
schwa	ə ɐ	ə ɐ
plosive	b d g p t k	b d g β ð ɣ p t k
fricative	f v s s̮ ʃ ʒ ç x h	f v s sː ʃ ç x h
liquid/nasal/glide	ʀ l m n ŋ j	ʀ l l̩ m m̩ n n̩ ŋ ŋ̩ j
silence/pause/glottis	*'sil'* *'pau'* ʔ	*'sil'* *'pau'* ʔ

Speech database for Austrian German and Viennese dialect

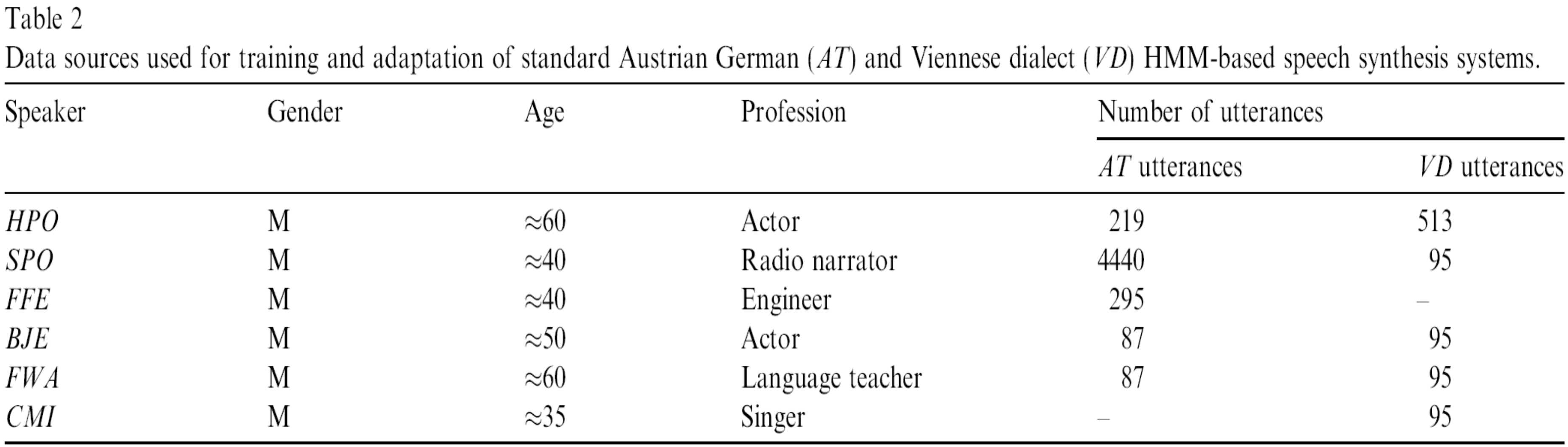


Table 2
Data sources used for training and adaptation of standard Austrian German (*AT*) and Viennese dialect (*VD*) HMM-based speech synthesis systems.

Speaker	Gender	Age	Profession	Number of utterances	
				AT utterances	*VD* utterances
HPO	M	≈60	Actor	219	513
SPO	M	≈40	Radio narrator	4440	95
FFE	M	≈40	Engineer	295	–
BJE	M	≈50	Actor	87	95
FWA	M	≈60	Language teacher	87	95
CMI	M	≈35	Singer	–	95

Speech database for Austrian German and Viennese dialect

- However since such a well-balanced database is not available yet and there are always fewer resources for non-standard varieties, we explore the best modeling for both AT and VD from the available unbalanced database.

Modeling approaches

- SD and SI refer to speaker-dependent and speaker-independent modeling.

- Likewise we can consider dialect-dependent and dialect-independent modeling.

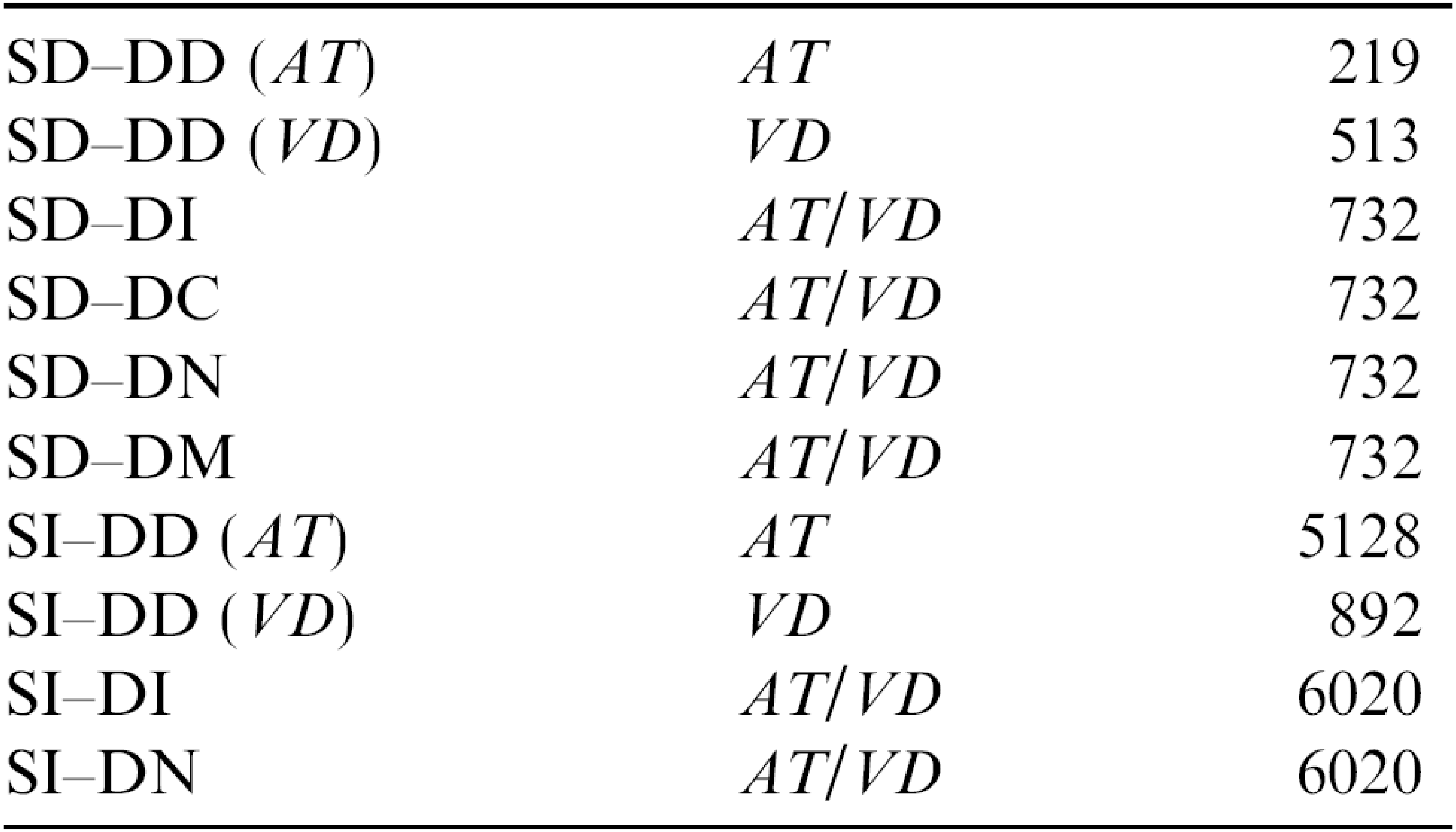
- The first is to add dialect information as a context for sub-word units and perform decision-tree-based clustering of dialects in the training of the HMMs.

Modeling approaches

- The second is to divide a set of speech data in both varieties uttered by one speaker into two subsets of speech data in different varieties uttered by two different pseudo speakers.

- DD, DI, DC and DN refer to dialect-dependent, dialect-independent, dialect clustering and dialect-adaptive training, respectively.

- DM refers to “DC plus DN”.

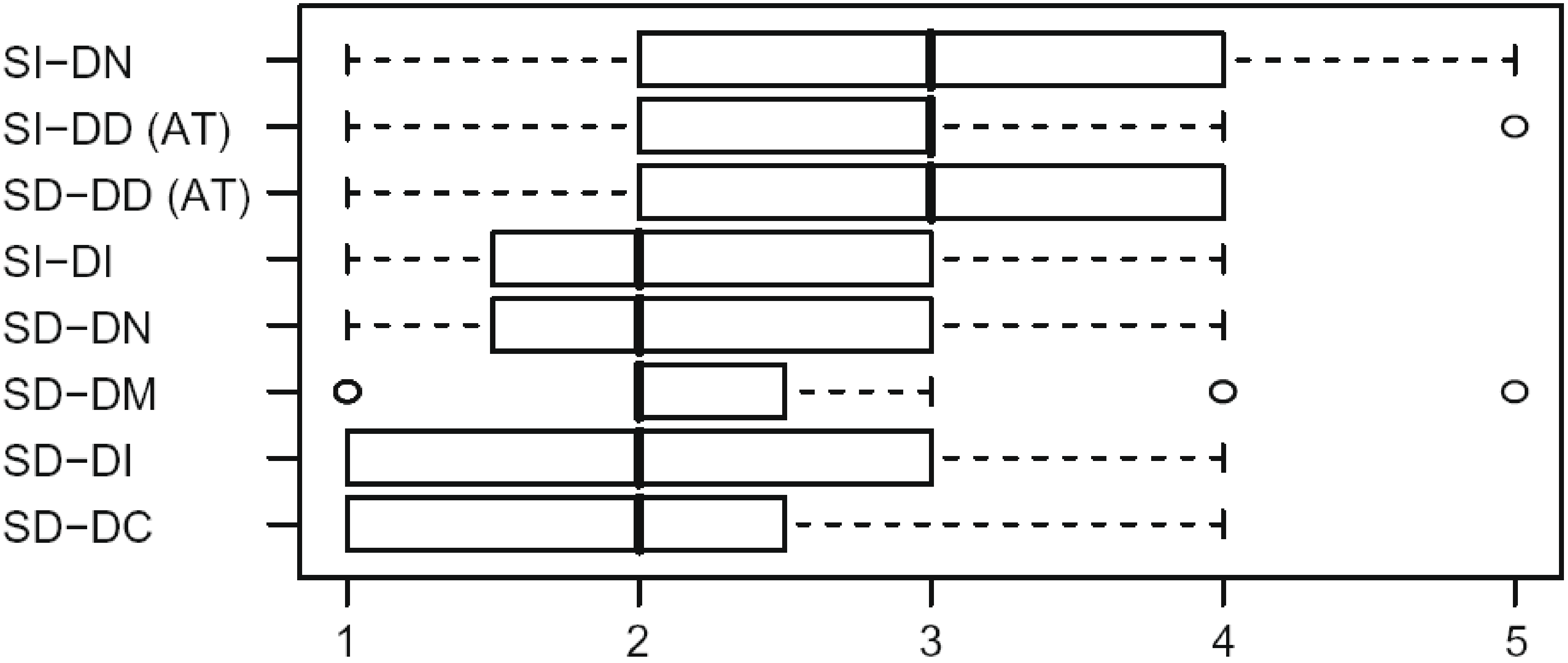


SD–DD (*AT*)	*AT*	219
SD–DD (*VD*)	*VD*	513
SD–DI	*AT*/*VD*	732
SD–DC	*AT*/*VD*	732
SD–DN	*AT*/*VD*	732
SD–DM	*AT*/*VD*	732
SI–DD (*AT*)	*AT*	5128
SI–DD (*VD*)	*VD*	892
SI–DI	*AT*/*VD*	6020
SI–DN	*AT*/*VD*	6020

Evaluation

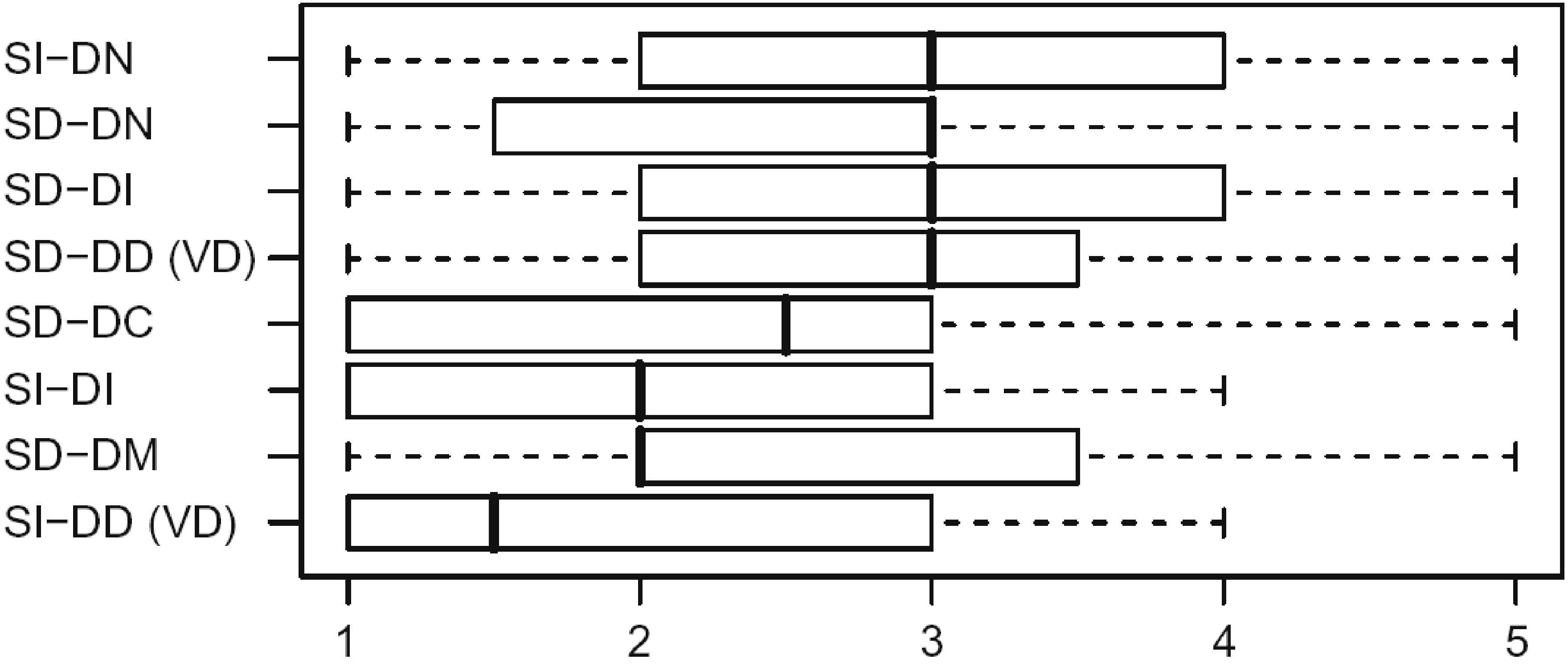
- The listening evaluation consisted of two parts: in the first part listeners were asked to judge the overall quality of synthetic speech utterances generated from several models using the different training strategies from Table 3.
- In the second part, after hearing a pair (in random order) of synthetic speech samples generated from the models, the listeners were asked which synthetic speech sample they preferred.

Fig. 3



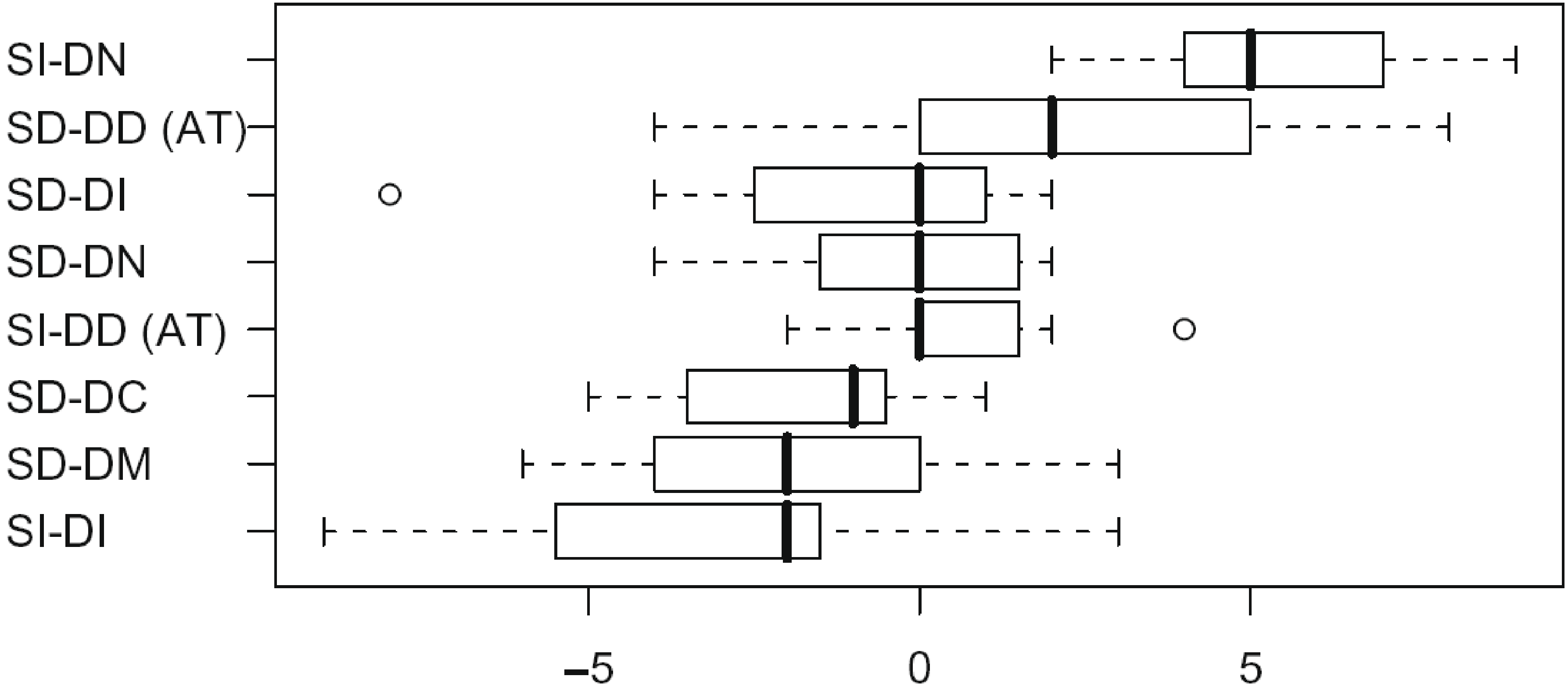
(a) Austrian German voices

Fig. 3



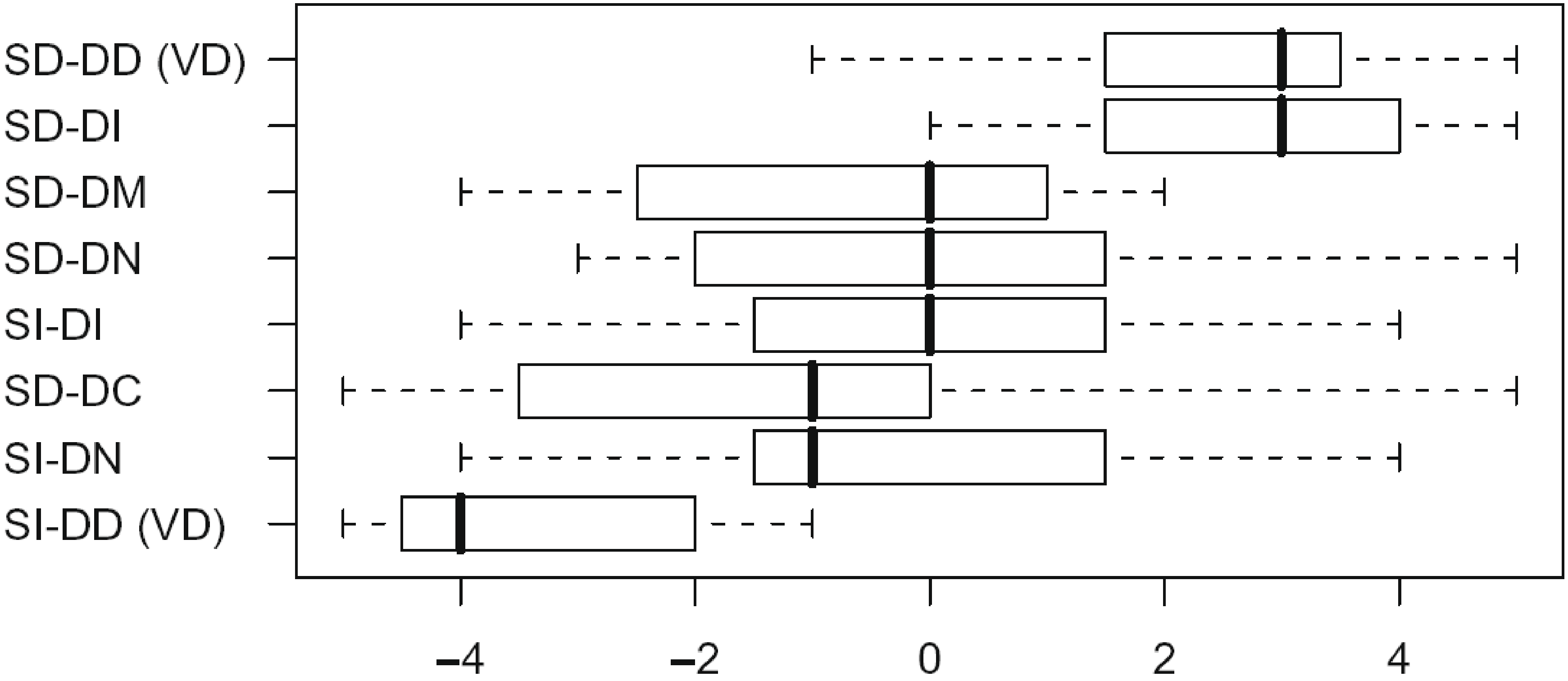
(b) Viennese dialect voices

Fig. 4



(a) Austrian German voices

Fig. 4



(b) Viennese dialect voices

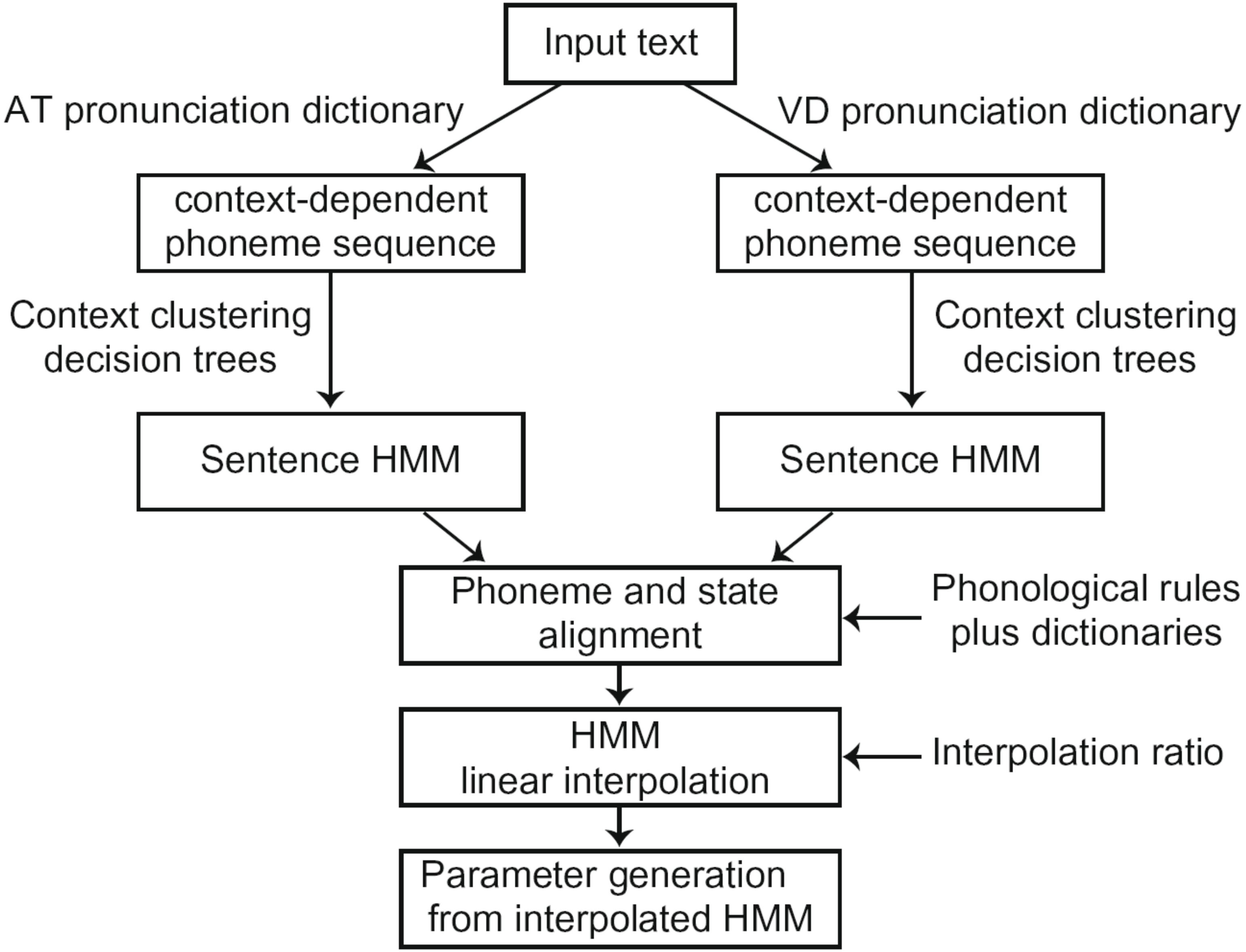

Input text
AT pronunciation dictionary
VD pronunciation dictionary
context-dependent phoneme sequence
context-dependent phoneme sequence
Context clustering decision trees
Context clustering decision trees
Sentence HMM
Sentence HMM
Phoneme and state alignment
Phonological rules plus dictionaries
HMM linear interpolation
Interpolation ratio
Parameter generation from interpolated HMM

Fig. 5. Flow of dialect interpolation.

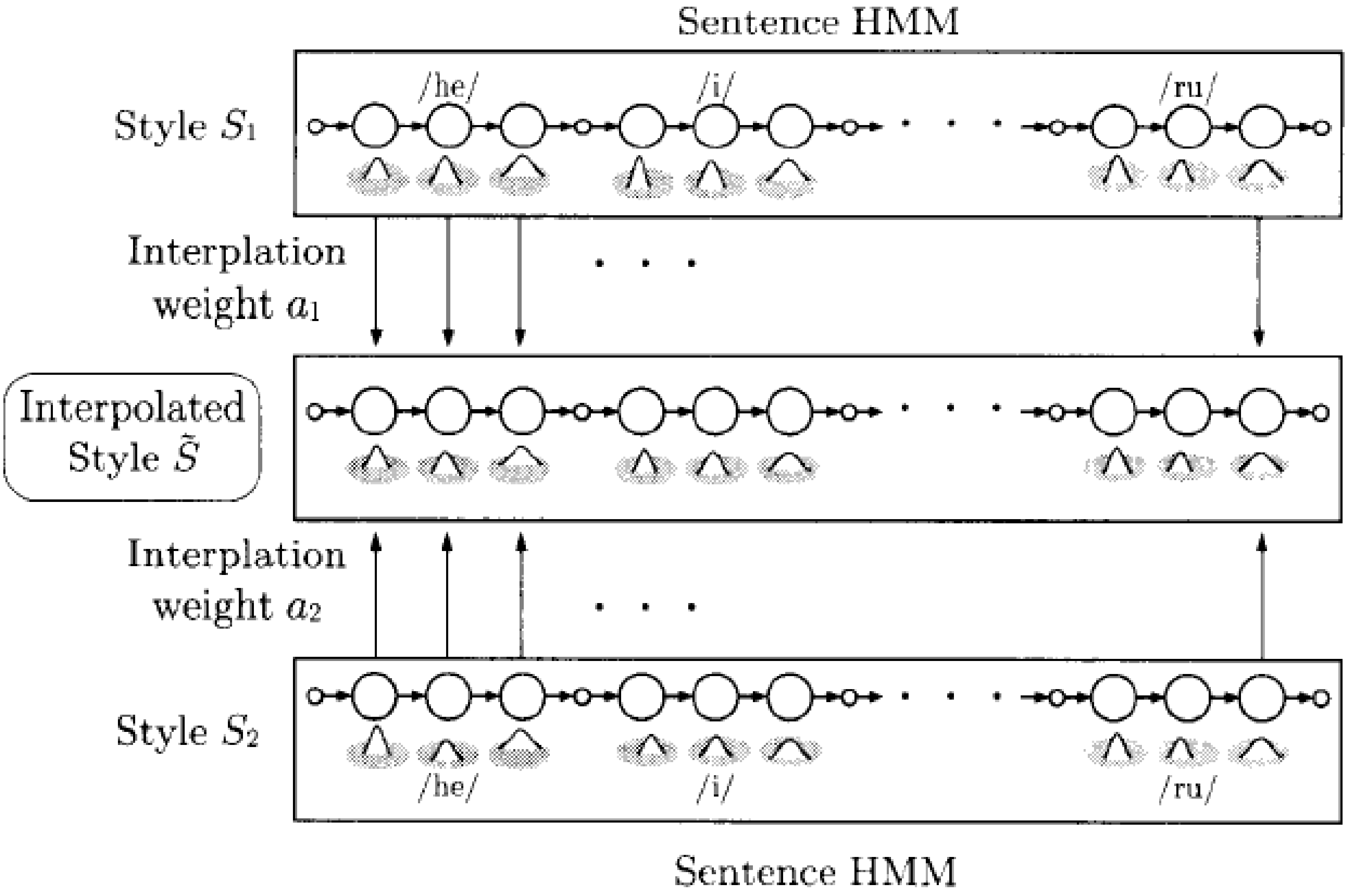

Sentence HMM
Style S_1
/he/
/i/
/ru/
Interplation weight a_1
Interpolated Style $\tilde{S}$
Interplation weight a_2
Style S_2
/he/
/i/
/ru/
Sentence HMM

Fig. 1 Example of interpolation of two style models.

HMM linear interpolation

- Let μ_i^{AT} and μ_i^{AD} be mean vectors of Gaussian pdfs for AT and VD voices, respectively, at aligned state i.
- Likewise Σ_i^{AT} and Σ_i^{VD} are their covariance matrices.

$$\hat{\boldsymbol{\mu}}_i = w\boldsymbol{\mu}_i^{AT} + (1-w)\boldsymbol{\mu}_i^{VD},$$

$$\widehat{\boldsymbol{\Sigma}}_i = w^2\boldsymbol{\Sigma}_i^{AT} + (1-w)^2\boldsymbol{\Sigma}_i^{VD},$$

- where w is an interpolation ratio between AT and VD voices.

Phonological processes between the standard variety of Austrian German and the Viennese dialect

- The phonological differences between the language varieties under consideration can be classified according to formal criteria that also have a significant impact on the way one can interpolate between the models associated with different phones or phone strings.

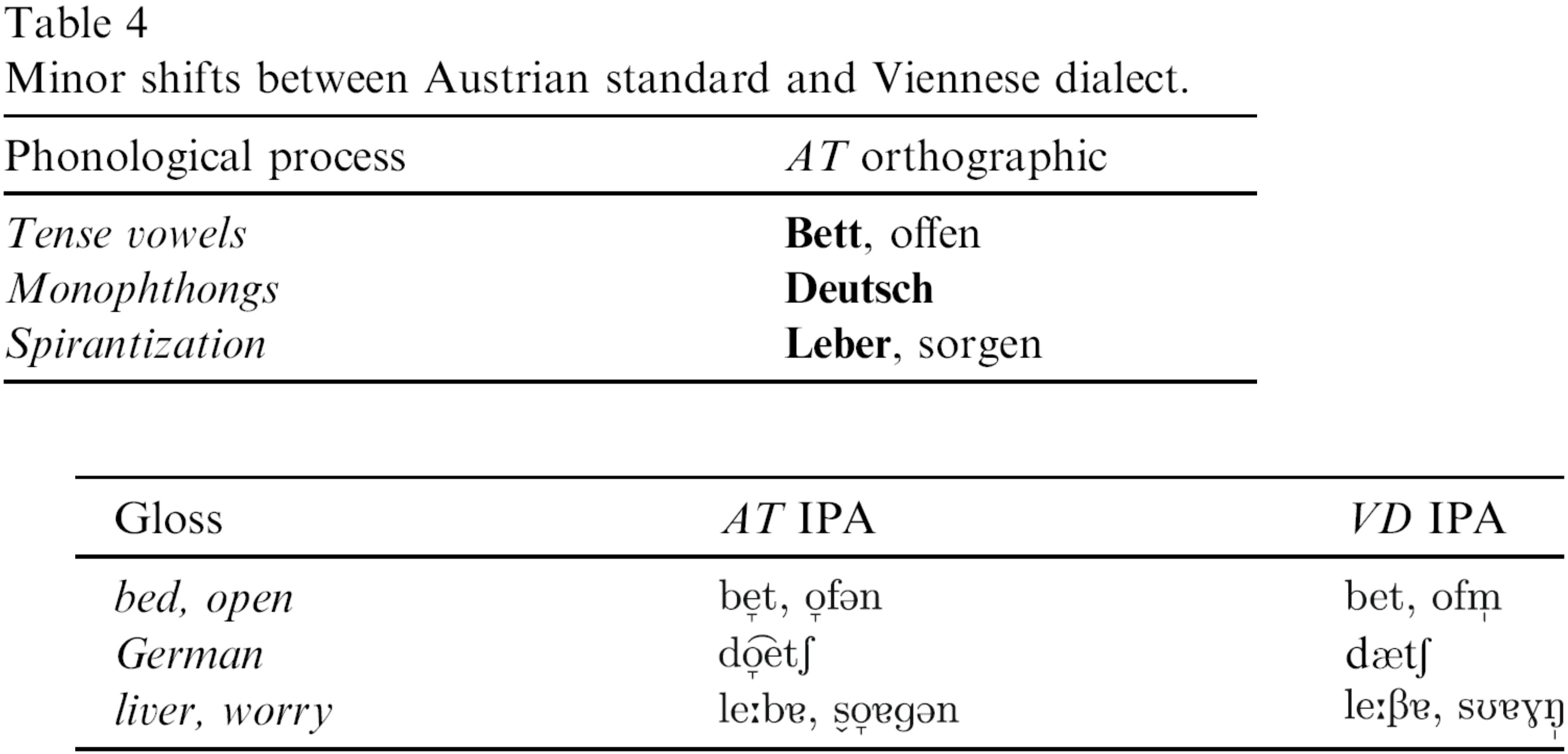


Table 4
Minor shifts between Austrian standard and Viennese dialect.

Phonological process	*AT* orthographic	Gloss	*AT* IPA	*VD* IPA
Tense vowels	**Bett**, offen	*bed, open*	be̝t, o̝fən	bet, ofm̩
Monophthongs	**Deutsch**	*German*	do̝͡etʃ	dætʃ
Spirantization	**Leber**, sorgen	*liver, worry*	leːbɐ, s̬o̝ɐgən	leːβɐ, sʊɐɣŋ̩

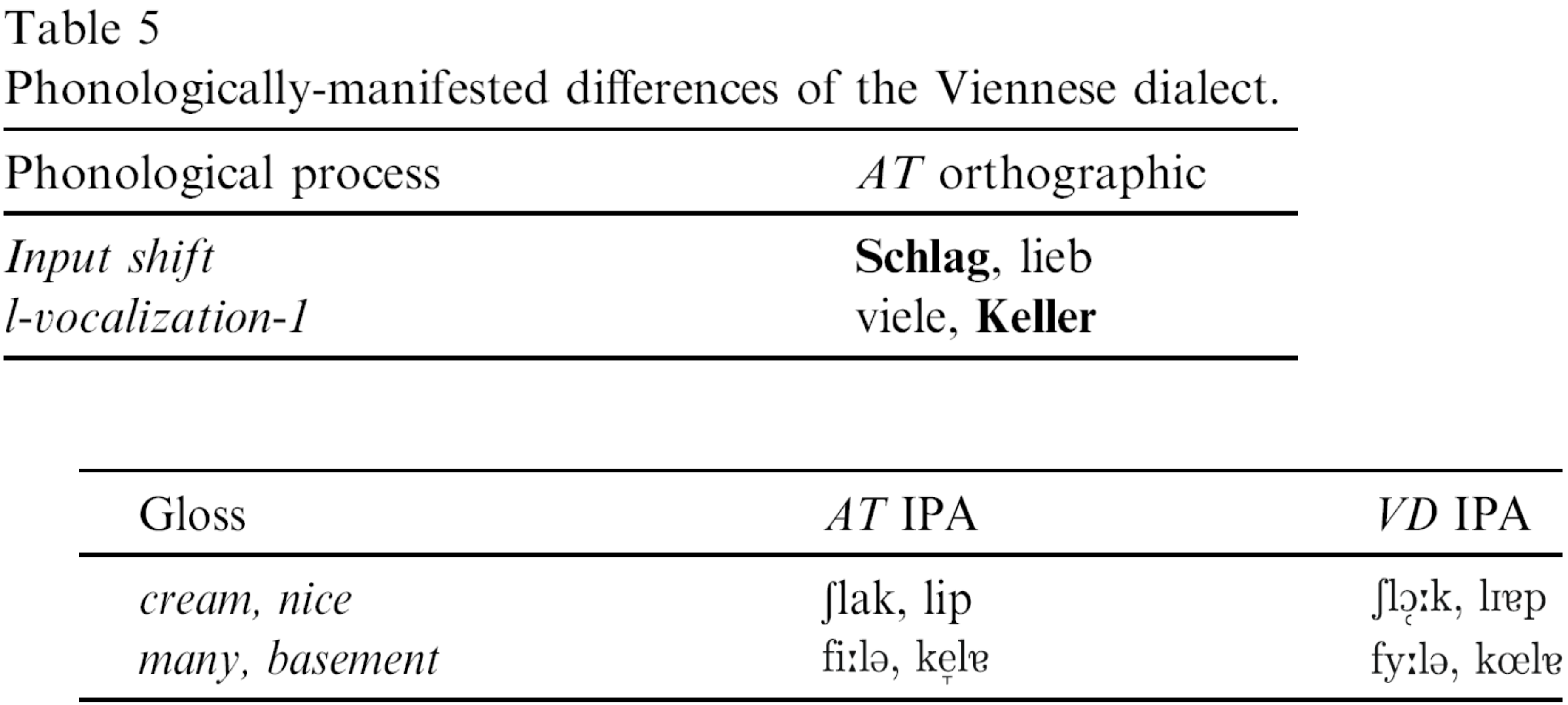


Table 5
Phonologically-manifested differences of the Viennese dialect.

Phonological process	*AT* orthographic	Gloss	*AT* IPA	*VD* IPA
Input shift	**Schlag**, lieb	*cream, nice*	ʃlak, lip	ʃlɔ̜ːk, lɪɐp
l-vocalization-1	viele, **Keller**	*many, basement*	fiːlə, ke̞lɐ	fyːlə, kœlɐ

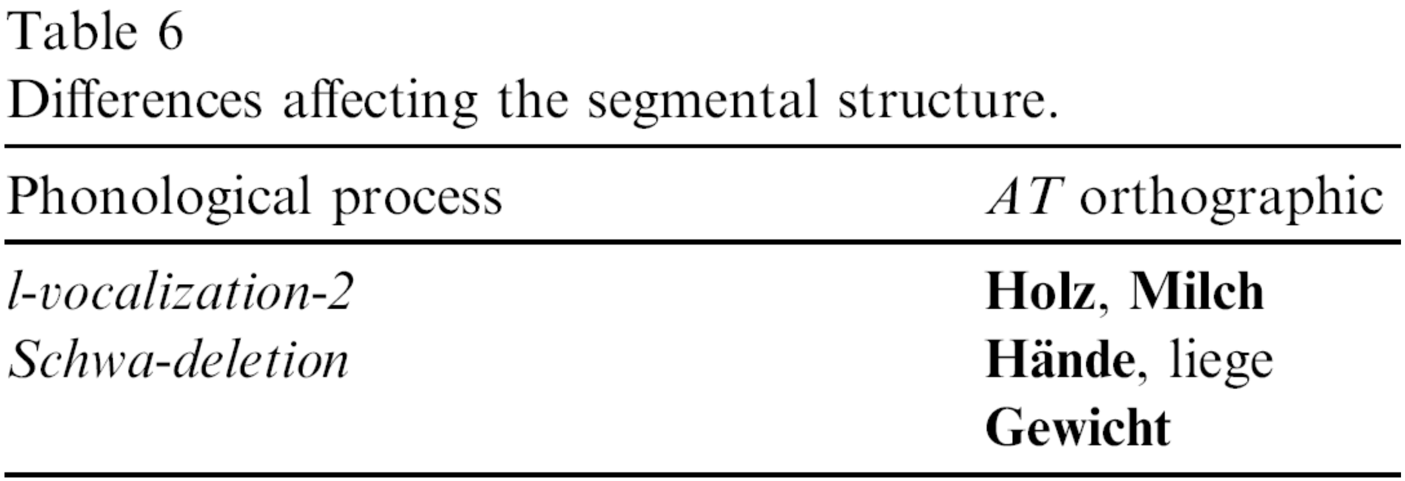
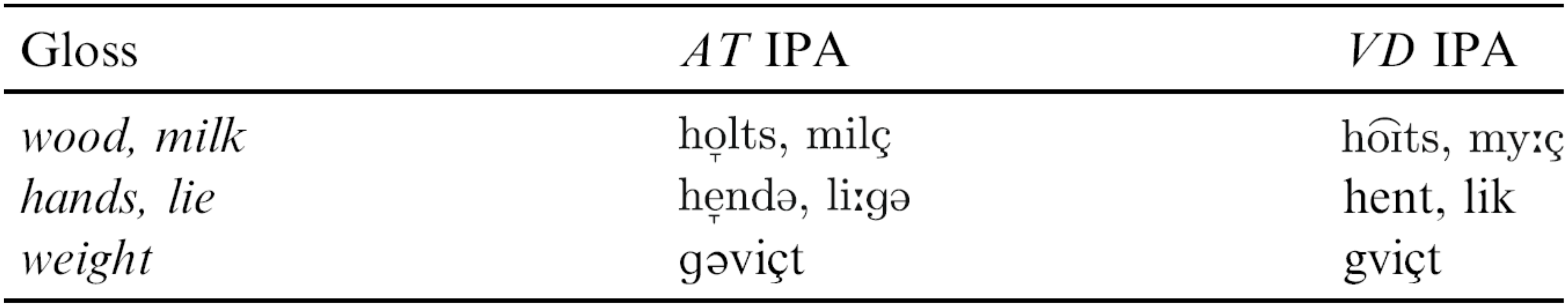


Table 6
Differences affecting the segmental structure.

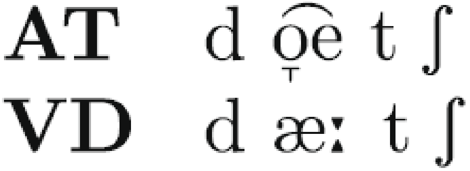
Phonological process	*AT* orthographic
l-vocalization-2	**Holz**, **Milch**
Schwa-deletion	**Hände**, liege
	Gewicht



Gloss	*AT* IPA	*VD* IPA
wood, milk	ho̞lts, milç	ho͡ɪts, myːç
hands, lie	he̞ndə, liːgə	hent, lik
weight	gəviçt	gviçt

Phonological constraints for HMM interpolation

- For the first group mentioned in the previous subsection, we can straightforwardly apply HMM interpolation since they have the same number of phones in Austrian and Viennese.



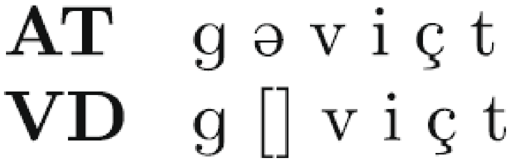
AT	d	o̞͡e	t	ʃ
VD	d	æː	t	ʃ

Phonological constraints for HMM interpolation

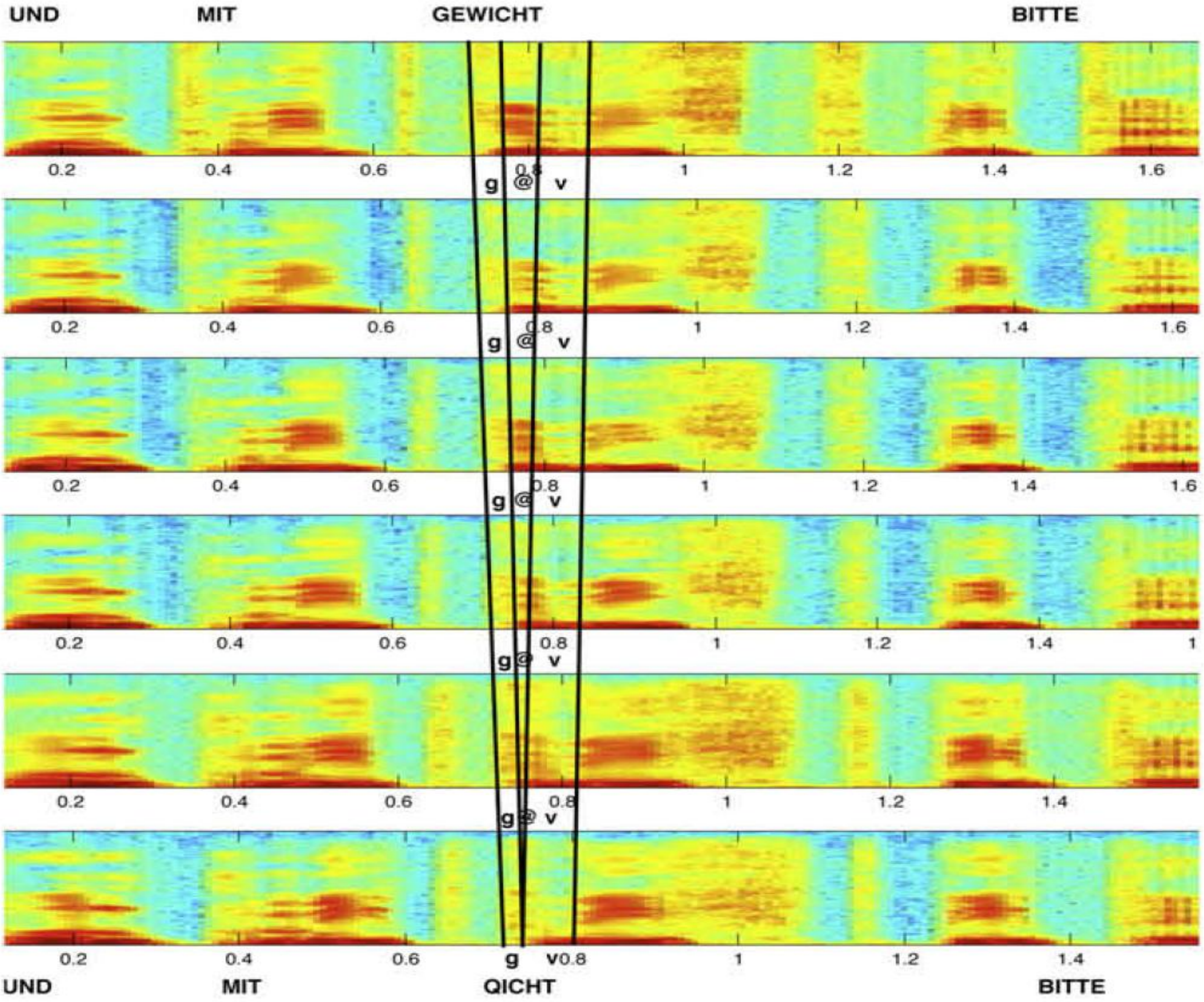
- For the second group, which does not have in-between variants, we utilize simple switching rules which disable the HMM interpolation and switch the target phone for one variety to the other variety at some intermediate point (threshold).

Phonological constraints for HMM interpolation

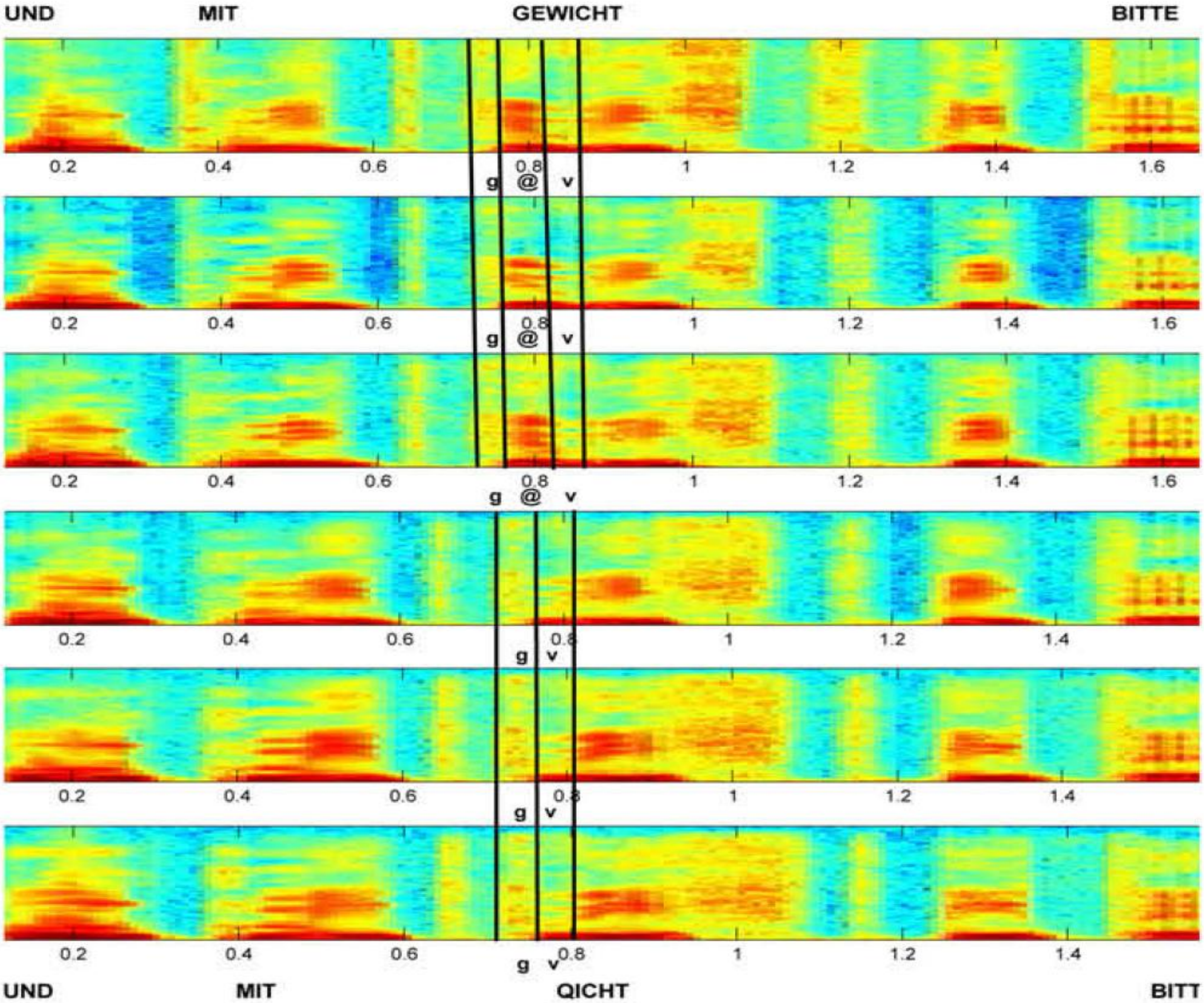
- For the third group (having words consisting of different numbers of phones in standard and dialect versions), we introduce a null phone , which simply corresponds to a phone model with zero duration.



AT	g	ə	v	i	ç	t
VD	g	[]	v	i	ç	t



(a) without switching rules

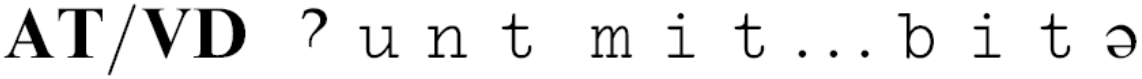


(b) with switching rules

Evaluation

- We designed a carrier sentence “Und mit . . . bitte” (And with . . . please) whose slot was filled with the words shown in bold in Tables 4–6.

- The phonetic transcription of the carrier sentence is provided in example.

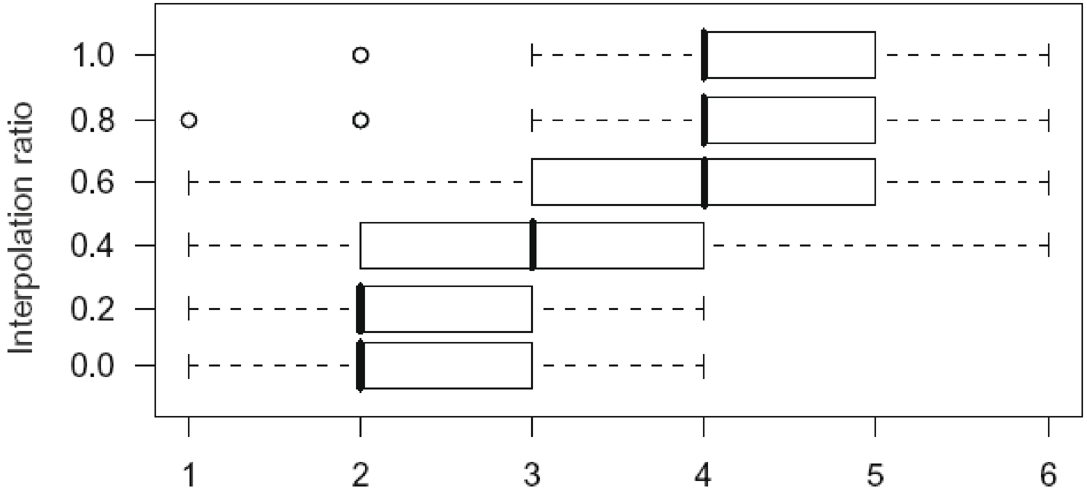


AT/VD ʔ u n t m i t . . . b i t ə

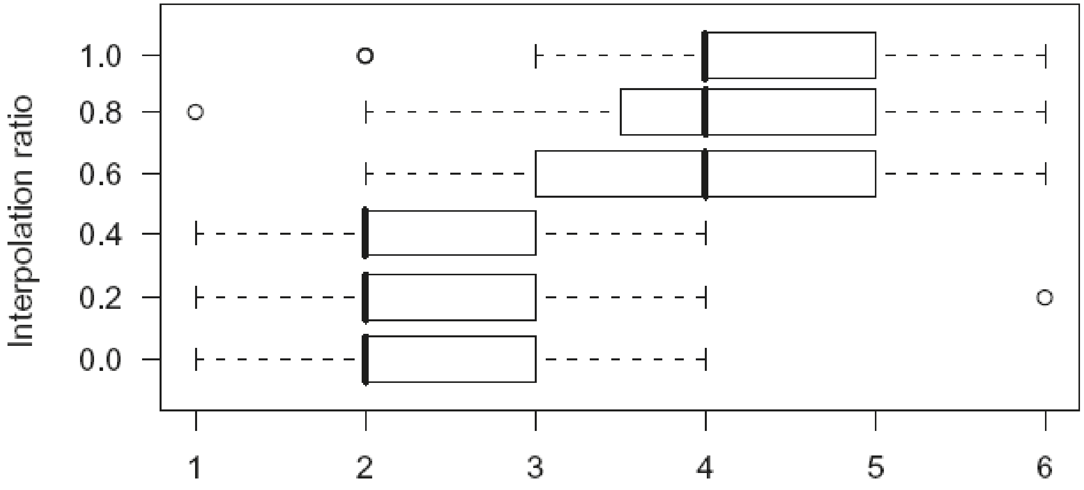
Evaluation

- For the rating, we used a scale from 1 (‘strongly Viennese’) to 6 (‘strongly standard’).

- In the figure, a ratio of 0.0 corresponds to the VD non-interpolated speech samples and 1.0 corresponds to the AT non-interpolated speech samples.



(a) HMM interpolation only



(b) HMM interpolation plus switching rules