

Articulatory Synthesis of Vocalized /r/ Allophones in German

Simon Stone, *Member, IEEE*, Yingming Gao, and Peter Birkholz, *Member, IEEE*

Abstract—Articulatory synthesis relies on precise, parametric vocal tract shapes to generate natural-sounding speech. In German, a particular challenge is the accurate synthesis of the vocalic /r/ allophones following vowels or in syllable coda position. Using established phonetic conventions, no satisfying results could be achieved so far, implying a possible shortcoming of these existing conventions. This study therefore analyzed a large number of natural recordings of the sounds in question from a single speaker to find the optimal number of target vocal tract shapes. Applying clustering techniques, the manifold of vocalic /r/ allophones could be reduced to two prototypical [e] variants previously undescribed in the literature. As shown by a listening experiment, which of these two sounds was preferred in which context depended not only on the respective context vowel’s tenseness, but also on its openness and acoustic distance to other context vowels ending in the same [e] variant. This indicates that the two different allophones might serve as a contrastive cue to help differentiate between otherwise similar context vowels.

Index Terms—German phonology, R allophones in German, articulatory synthesis

I. INTRODUCTION

ARTICULATORY synthesis is the computational creation of speech by simulating articulatory, phonatory, and control processes during speech production. Due to the possibility of fine-grained control of virtually all articulatory parameters, it is an invaluable tool in education and research and potentially superior to any other synthesis system [1]. An articulatory synthesis pipeline comprises several stages of the speech production process and thus requires accurate models for the vocal folds [2]–[4], the vocal tract [5]–[7], an aerodynamic and acoustic simulation [8]–[11], and an articulatory control model [12]–[15]. While much work has been published regarding each of these individual fields, only a handful of systems have tackled the daunting task of compiling an entire pipeline to actually perform the entire synthesis (e.g., the CASY system [16], the APEX system [17], the ArtiSynth project [18], and SAPWindows [19]) with varying success and persistence. The most complete, feature-rich, and continuously developed articulatory speech synthesizer, however, is arguably the VocalTractLab¹ (see Figure 1). The achievable quality of the synthesized speech has continually increased

S. Stone, Y. Gao, and P. Birkholz are with the Institute of Acoustics and Speech Communication, Technische Universität Dresden. This paper has supplementary downloadable material available at <http://www.vocaltractlab.de/index.php?page=birkholz-supplements>, provided by the author. The material includes all analyzed recordings, some illustrative synthesis examples using the current conventions for vocalic /r/ allophones, and the stimuli used in the listening test. Contact simon.stone@tu-dresden.de for further questions about this work.

¹www.vocaltractlab.de

over the last decade (see <https://www.vocaltractlab.de/index.php?page=vocaltractlab-examples> for some recent examples), but is still not quite as high as the synthesis quality of state-of-the-art unit-selection [20] or neural end-to-end [21] systems. VocalTractLab conducts an aero-acoustic simulation [9], [22], [23] using a vocal tract model [6], [24] excited by a model of the vocal folds [4], [25]. The vocal tract model is a three-dimensional representation of the supraglottal airways and articulators and is defined by a set of 23 physiologically motivated control parameters (e.g., tongue tip position, velic opening, and so on; see [6] for a thorough presentation of the parameters and their impact on the model). A particular articulatory configuration or *vocal tract shape* can be defined by specifying values for each of these control parameters. To articulate a connected utterance, a sequence of target shapes corresponding to the constituent sounds in the utterance is defined and the articulatory trajectory is finally modeled using the Target Approximation Model [12], [26]. The timing of this trajectory (determined by the duration and transition time of each segment) is very important for the naturalness and intelligibility of articulatory synthesis. Beyond that, properly configured target shapes are also critically important and may greatly increase the overall quality. These are unfortunately non-trivial to generate without some sort of reference.

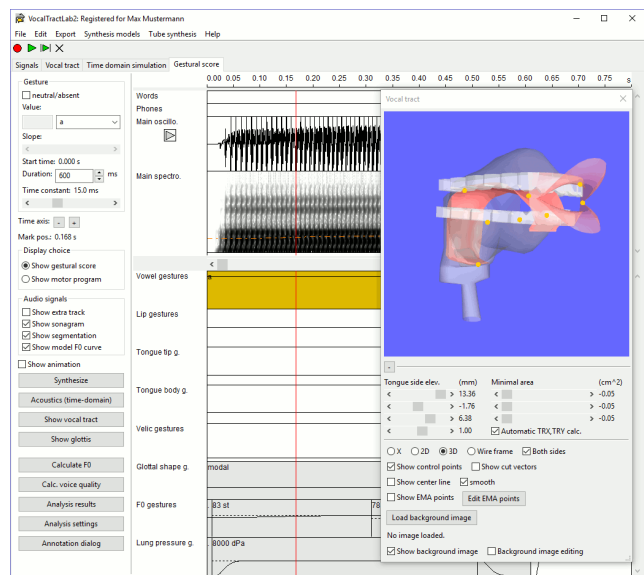


Fig. 1: The VocalTractLab graphical user interface. The vocal tract model can be seen in the foreground and the interface for the dynamic articulatory control (the gestural score) in the back.

supplemental materials). These shortcomings could of course be rooted in suboptimal modeling of the speech production processes in the articulatory synthesizer. Nevertheless, the disagreement between phoneticians on this subject motivated a thorough investigation of these /r/ allophones to identify the minimal set of vocal tract shapes necessary to synthesize the entire variety of allophones. Instead of limiting the analysis to perceptual descriptions, we conducted objective formant measurements and applied clustering techniques to identify similarly realized sounds in a large set of example utterances spanning all contexts of interest. Using the identified, prototypical sounds, we synthesized short words containing the sound constellations described above. We then conducted a pair-wise comparison of the different allophones in a subjective listening test. The results showed that the manifold of allophonic realizations could be reduced to two distinct vocalic sounds, i.e., vocal tract shapes, without significant loss in acceptance by the listeners. One of these sounds corresponds to the canonical /v/, while the other sound (we called /v_{mid}/) appears to be undescribed for German in the literature so far and is located between /v/ and /ə/ on the formant map. Which of these two sounds was preferred in which context depended not only on the context vowel’s tenseness, but also on its openness and acoustic distance to other context vowels ending in the same [v] variant, indicating that the different allophones might serve as a contrastive cue to differentiate between otherwise similar context vowels (e.g., /e:/ vs. /i:/ or /o:/ vs. /u:/).

Since this study is only based on a single speaker, its findings cannot be generalized to a larger population. However, validating the identified prototypes is only possible if the analyzed speaker is also available as a model speaker in the used articulatory synthesizer *VocalTractLab*, which currently only applies to the speaker used in this study.

II. DATA-DRIVEN INVESTIGATION OF THE GERMAN /r/ ALLOPHONES

For ease of reference, from here on we will not distinguish between the /r/ allophones [v] and [ʁ] when referring to the secondary diphthongs and even the monophthong [v] and summarize all allophonic instances as [Vv], similar to the convention by Kohler (see subsection I-A). This is not supposed to identify them as the same phone, but is merely a shorthand notation that allows a more compact presentation of our analyses.

A. Speech material

To analyze the varieties of the German /r/ allophones, we recorded a set of words spoken by a male native German speaker (age 57, born in North-East Germany, living in Berlin for more than 50 years). The words were selected so as to include 20 examples of each /r/ sound in every possible German vowel context, i.e., /a: e: i: o: u: ε: ø: y: a ε I ɔ ʊ œ ʏ/ and in word-final position (see Table III for the complete list). It was also ensured that each [Vv] appeared in different positions within the words and with different following context sounds (e.g., labial and dental consonants). That way these heterogeneous influences might, in a sense, “average out”

in the subsequent analysis to identify a prototype that only depends on the vowel context, unbiased by other factors (see below). Each word was embedded in the carrier phrase *Ich habe <word> bestellt.* - [ɪç 'ha:bə <word> bə'ʃtɛlt] (English: *I have ordered <word>*). The carrier phrase was chosen so that the word of interest would be preceded by the “neutral” vocal tract configuration of [ə] and followed by the stop [b], which facilitated the segmentation process because the end of the word could be easily spotted in the audio waveform by its closure phase. The order of the words was randomized within each group with each group consisting of samples of one [Vv] combination. The set was recorded twice on two consecutive days, with one recording session containing an entire set and consisting of approximately 30 min of audio. The recording was made in a sound-proofed recording studio using a Sennheiser MKH20P48 microphone connected to a Behringer Ultragain Mic 2000 microphone amplifier at a sampling rate of 48 kHz using 16 bit quantization and the audio was saved as one single mono PCM WAV file per session. Before further processing, the two long WAV files were each manually segmented to extract the two times 300 words of interest, which were saved as two times 300 short WAV files. The segmented audio recordings can be found in the supplemental materials.

B. Formant measurements

After the segmentation of the audio recordings (see above), we obtained two sets of 300 words containing 15 different [Vv] secondary diphthongs (i.e., 20 different words for each context). The first set of words was then perceptually evaluated by two German native speakers with a background in ear phonetics. If a realization from the first set was judged as acceptable, it was added to the final data set. If it was judged as flawed, the realization from the second set was considered. If it was acceptable, it was added to the final data set. If that realization was also flawed, the entire word had to be discarded. After this initial screening process, eleven words had to be discarded, i.e., 289 words remained in the data set. This preselection process is summarized in Figure 3.

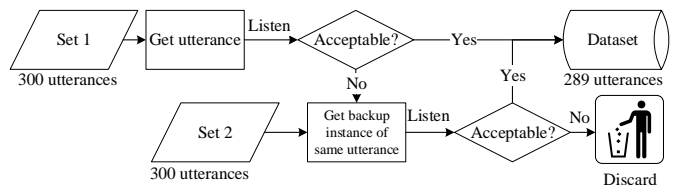


Fig. 3: Procedure for the preselection of utterances from the two recorded sets.

For each of these instances, the first three formants F_1 , F_2 , and F_3 were determined for both the context vowel sound and the [v] or [ʁ] sound independently. In the case of the words containing only the [v] sound (e.g., *Butter* - [bʊtɐ], Engl. *butter*), it was treated as a “pseudo-diphthong” and the formants were taken from stationary segments in the beginning (considered the “context vowel” in the remainder of the study) and end of the sound (considered the actual /r/ allophone).

The formants were measured using Praat 6.0.43 [35] and similar to the strategy outlined in [36]: starting at 5, the number of formants assumed in the LPC-based algorithm was increased in 0.5 increments up to 7 and the fit was visually inspected using the spectrogram. If the trajectories were sufficiently smooth and well-aligned with the high-intensity ridges in the spectrogram during the respective segment of interest, the average formant frequencies (in Hertz) F_1 , F_2 , and F_3 were determined over a range of approximately 30 ms (see Figure 4 for a segmentation example).

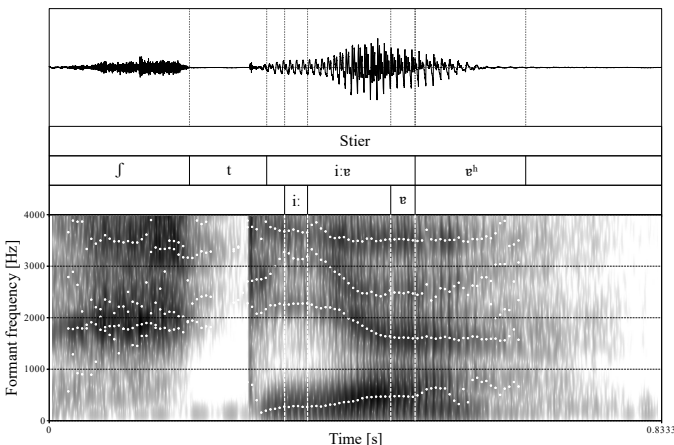


Fig. 4: Example segmentation of a word (here: *Stier* - [ʃti:ɐ], Engl. *bull*). The number of formants in the formant tracking was optimized for both parts of the secondary diphthongs independently (here: six for both). The values for the first three formants F_1 , F_2 , and F_3 were determined by averaging over the marked segments of approximately 30 ms.

C. Data processing and selection

After a careful first pass through the entire set, we obtained two sets of formants for each of the 289 words (one set for the context vowels, and one set for the [ɐ]). Before any further processing, we transformed the formant frequencies to the Bark scale so that distance calculations in the formant space more accurately reflected perceptual distances. We used the following formula² for the mapping, taken from [37]:

$$f_{\text{Bark}} = \frac{26.81 f_{\text{Hz}}}{1960 + f_{\text{Hz}}} - 0.53$$

if: $f_{\text{Bark}} < 2 \rightarrow f_{\text{Bark}} := f_{\text{Bark}} + 0.15 \cdot (2 - f_{\text{Bark}})$
if: $f_{\text{Bark}} > 20.1 \rightarrow f_{\text{Bark}} := f_{\text{Bark}} + 0.22 \cdot (f_{\text{Bark}} - 20.1)$

Even after the initial screening, there were still two major potential error sources for these semi-automatically acquired data: (i) pronunciation variants by the speaker that deviated from the canonical transcription, and (ii) formant measurement errors. Type (i) errors could result in correctly measured formants labeled with the wrong sound, while type (ii) errors (which can easily happen in LPC-based formant measurements [38]) would cause wrong formant values for correctly

labeled sounds. Some of the type (i) errors were already identified by the subjective screening of the recordings but there was no objective assurance that all of them had been spotted perceptually. Before further processing, we therefore checked all formant values for consistency within each group, i.e., within each subset of formant triples representing the same [Vɐ] sound. Assuming that the differences between the formant values within each group should generally be low, large deviations from each group’s median were considered an indicator for possibly faulty formants (potentially type (i) or type (ii) error or both). To find these outliers, we applied the outlier detection algorithm implemented in the MATLAB function `isoutlier` to both the formant triples from the context vowels and the /r/ allophones in each group separately³. The recordings that these outlying formant values were taken from were then re-analyzed by a second labeler who was not aware of the previous measurement results or the formant values of other samples within that group (blind double-check). Using the formant values from this second analysis, the outlier detection was repeated until no more outliers were found. While most outliers could indeed be attributed to measurement errors and were eliminated after re-checking the measurements, a number of data points had to be given special consideration. Some of the words were on closer inspection pronounced in a non-canonical fashion and accordingly relabeled (see Table III). Sixty remaining outliers could not be corrected with either of these two strategies and were excluded from further analysis. In many cases, the coarticulatory effect of nasal sounds directly before or after the secondary diphthongs and/or the position in the final syllable (and therefore the increased breathiness in the recording) may have negatively impacted the formant measurement. However, in some cases (e.g., *unberührt* - [ˈʊnbəʁyːɐt], Engl. *untouched*) the reasons remain unclear, but may have to do with the stress of the syllable or the preceding /r/ sound (the r-coloring). The excluded outliers are grayed out in Table III. In the processed final set of 229 words, the number of words containing each /r/ allophone was unevenly distributed, i.e., the data set was an imbalanced sample after preprocessing. However, as will be discussed in subsection II-D, the further analyses were performed on a per-group basis, meaning that no implicitly weighted calculations were made throughout the study.

D. Unsupervised clustering

In subsection II-C, we already carefully preprocessed the data to identify mismatched labels based on up to two listeners’ opinions. However, this subjective judgment may have missed more subtle pronunciations variants that walk the ill-defined line between two sound categories. Therefore, we followed up the preprocessing with a more data-driven approach to identify the actually realized context vowel of every allophone instance in three steps: (1) cluster the formant data of the context vowels, (2) relabel each context vowel using the automatically found cluster associations, (3) calculate the

²The implementation is identical to the one included in the Audio Toolbox function `hz2bark` in MATLAB 2019a and up.

³Using default values, this marked all formant tuples as outliers that were more than three scaled median absolute deviations from the median of the respective group (see documentation of `isoutlier` in MATLAB R2018b).

mean formant triple in each group of allophones using the cluster associations from step (2).

The clustering was done using the unsupervised naive k -means algorithm [39, pp. 424]. In this particular study, the final centroids could be regarded as a set of “typical” formant frequencies for a particular phone. The distance between a particular realization and any given centroid can be interpreted as a measure of perceptual similarity to that phone (because the clustering was done in Bark space).

After initializing the algorithm with the number of analyzed context vowels ($k = 15$), the cluster centroids for [e:] and [ɛ:] converged to roughly the same point in the formant space, which was in line with the observations made in [40, p. 175] that speakers from Berlin and the general north-east of Germany tend to replace the vowel [ɛ:] by the higher vowel [e:]. Therefore, realizations from both contexts were merged into a single group labeled as [e:] and k was set to 14 to avoid a redundant cluster. After the unsupervised clustering into 14 groups, all realizations in the same cluster were then relabeled using the most frequently occurring original label within the cluster. As an example: The formant triple measured in the vowel with the original label [ʊ] in the utterance *Wurst* - [vʊʁst] (Engl. *sausage*) was closer to all the realizations originally labeled as [o] than to all the other realizations originally labeled as [ʊ]. Hence, the label of this realization was changed from the original (supervised) label [ʊ] to the (unsupervised) cluster label [o] to more accurately reflect the similarities in the data. Both the original labels and the final cluster associations are shown in Figure 5 and all utterances that were relabeled are marked in Table III by an asterisk.

After the vowel clustering, these updated labels were used to regroup the formant triples measured in the [v] phones of the (pseudo-) diphthongs by their context vowel. Finally, the mean formant triple was calculated for each group (given in Table I).

Allophone	F_1		F_2		F_3	
	[Hz]	[Bark]	[Hz]	[Bark]	[Hz]	[Bark]
[e]	754	6.9	1472	11.0	2569	14.7
[aɐ]	735	6.8	1224	9.8	2388	14.2
[eɐ]	668	6.3	1556	11.3	2385	14.2
[iɐ]	522	5.1	1603	11.5	2443	14.3
[oɐ]	572	5.5	1106	9.1	2296	13.9
[uɐ]	494	4.9	1309	10.2	2347	14.1
[ɛɐ]	668	6.3	1556	11.3	2385	14.2
[øɐ]	645	6.1	1437	10.8	2245	13.8
[yɐ]	469	4.6	1466	10.9	2177	13.6
[ɪɐ]	505	5.0	1245	9.9	2394	14.2
[ɛʊ]	717	6.7	1491	11.1	2356	14.1
[ɔʊ]	550	5.3	877	7.8	2431	14.3
[øʊ]	427	4.3	835	7.5	2388	14.2
[yʊ]	473	4.7	1160	9.4	2289	13.9
[œʊ]	622	5.9	1183	9.6	2187	13.6
[e _{mid}]	670	6.3	1338	10.4	2341	14.1
[e _{low}]	494	4.9	1346	10.4	2325	14.0

TABLE I: Average formant frequencies (centroid) of all realizations of [v] in the same context and of the two identified [v] variants.

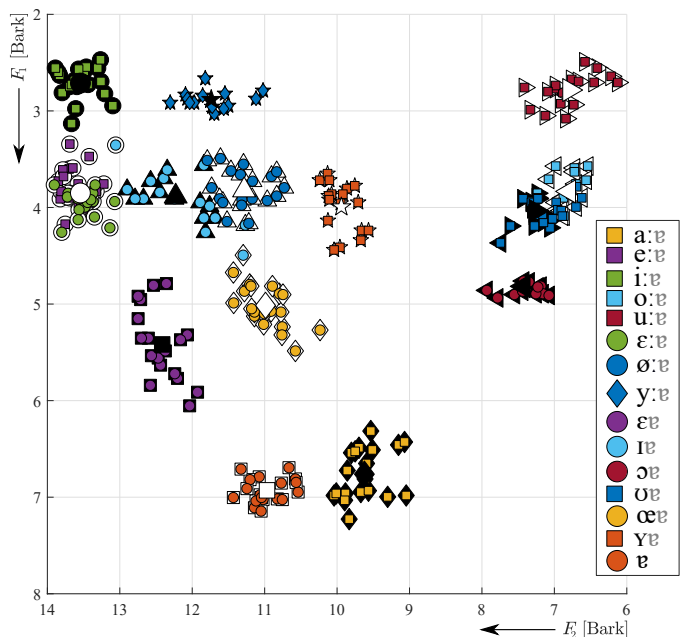


Fig. 5: Clustering result for the context vowels. The clusters were found unsupervised using a k -means algorithm and the Euclidean distance metric, with the starting centroids chosen as the mean vector of each labeled group (counting [e:ɐ] and [ɛ:ɐ] as the same group, i.e., $k = 14$). F_3 is not shown for ease of reference but was used in the clustering. Supervised sound labels are marked by color and shape while the unsupervised cluster associations are shown by the underlying shape.

E. Shape optimization

The analysis of the formant frequencies of the [v] realizations from the various vowel contexts at first seemed to imply that every [Vɐ] diphthong is produced with a different second vowel, because (almost) every context group had different values for F_1 , F_2 , and F_3 . However, this apparent manifold of phones could have been deceptive: There is always a certain degree of natural variation in fluent articulation of speech, formant values are notoriously difficult to measure with great precision [36], and the groups were mostly imbalanced after the rigorous data selection. We therefore adopted a two-fold strategy: (1) Reduce the number of /r/ allophones by finding clusters in the average allophone realizations and using their centroids as “prototype allophones”. (2) Find the optimal number of phones necessary to create natural sounding /r/ allophones in any given context. To that end, we needed to create VocalTractLab-compatible vocal tract shapes corresponding to every [Vɐ] sound. Since the recorded speaker in this study was also the speaker originally used to create the vocal tract shapes included in the VocalTractLab and a “context-free” [v] shape was already available, the context-dependent shapes were created by modifying this initial shape to obtain the desired configuration. To steer this modification, target vocal tract resonances had to be specified. Since formant frequencies F_i extracted with an LPC-based algorithm (as implemented in Praat) and vocal tract resonance frequencies f_{R_i} are not necessarily the same [41], we applied a correction

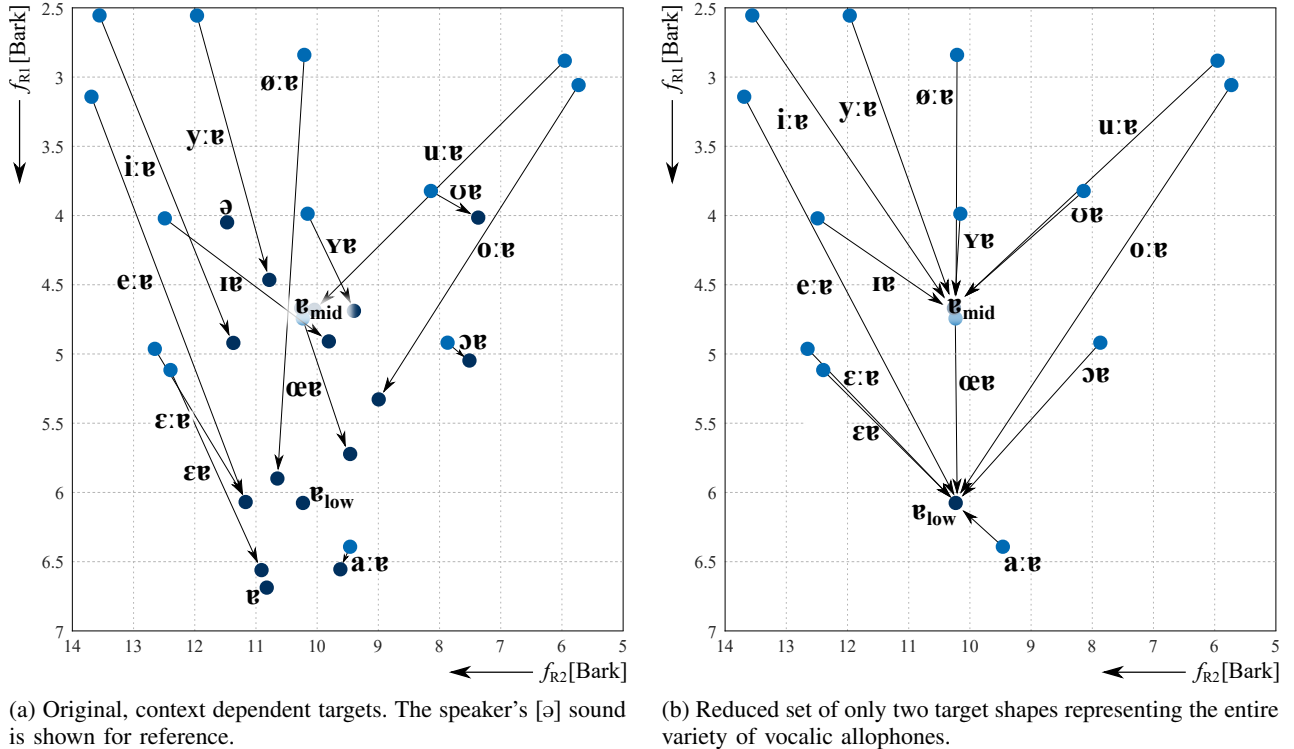


Fig. 6: Vocal tract resonances of the vowels and /r/ allophones.

factor to each formant triple, which led to the resonance map and diphthong vectors shown in Figure 6a.

Using these [v] resonance frequencies as targets, a greedy algorithm optimized the original shape according to the procedure laid out in [6]. For the tense context vowels, the initial shape was the default `VocalTractLab` [v], but for the lax context vowels the corresponding context vowel shape was used, because the acoustic error between the context vowel shape and the target resonances was generally smaller than between the [v] shape and the target resonances. In addition to these context-dependent shapes, two more shapes were created: one shape $[v_{mid}]$ representing the centroid of the cluster of [v] sounds with an f_{R1} of less than 5.5 Bark (consisting of the [v] sounds from [i:r], [r], [y:r], [y:r], and [u:r]), and one shape $[v_{low}]$ representing the centroid of the cluster of [v] sounds with an f_{R1} of more than 5.5 Bark (containing the [v] sounds in [v], [a:r], [e:r], [e:r], [e:r], [o:r], [o:r], and [o:r]) (see Figure 6a). The respective formant frequencies are given in Table I. The shapes were found by once again starting from an initial shape and then optimizing vocal tract parameters to minimize the acoustic differences between the target and actual resonances. The initial shape used in the optimization was the original `VocalTractLab` [v] shape for the $[v_{low}]$, and the [o] shape included in the `VocalTractLab` for the $[v_{mid}]$. Both initial shapes were chosen because their initial resonances were the closest to the respective target resonances. Finally, we ended up with a set of 17 new vocal tract shapes: one for each context-dependent (pseudo-)diphthong and the two prototypes $[v_{low}]$ and $[v_{mid}]$ (the latter two shapes are shown in Figure 7).

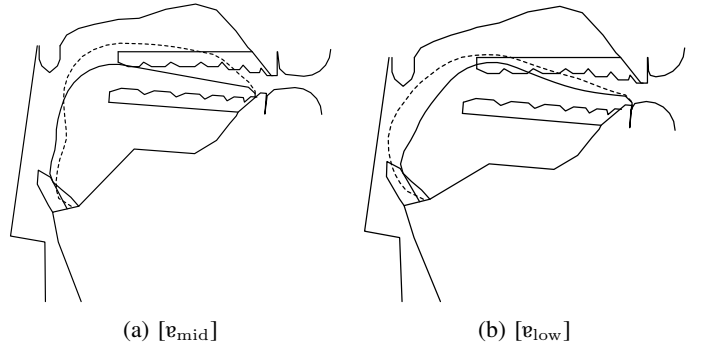


Fig. 7: 2D sagittal views of the vocal tract shapes obtained for the new prototype sounds. The dashed lines are the contours of the tongue sides.

III. LISTENING EXPERIMENT

Two questions remained at this point: Is the reduction of the 15 context-dependent [v] sounds to the two prototypes $[v_{low}]$ and $[v_{mid}]$ permissible without large losses in perceived naturalness and if so, which prototype should be chosen in which context? To answer these questions, we asked German native listeners to state their preference in a listening experiment.

A. Generation of the stimuli

For each of the 15 context variants, we chose one carrier word to contain the [v] sound from the vocabulary used for the audio recordings (see second column in Table II). Each carrier word was then synthesized using `VocalTractLab` in four versions that only differed in the choice of the [v] shape: one

used the context-dependent $[V\vartheta]$, one the monophthong $[\vartheta]$, one the $[\vartheta_{\text{low}}]$, and one the $[\vartheta_{\text{mid}}]$. In case of the syllabic $[\vartheta]$ in *Ober* - $[\text{o:b}\vartheta]$ (Engl. *waiter*), only three versions were synthesized. The set of stimuli therefore consisted of 15 (number of contexts) times 4 (number of $[\vartheta]$ variants) minus 1 words for a total of 59 stimuli. To generate the gestural scores for the synthesis, we used the recorded audio as a reference and adjusted the durations of the gestures to match the timing of the natural recording as well as possible (a paradigm called copy synthesis). The time constants were all set to 15 ms to ensure constant conditions across all shapes and words. The synthesized audio was peak-normalized and padded with 200 ms of silence at both the beginning and end. All synthesized audio stimuli are part of the supplemental materials.

B. The experimental paradigm

The experiment consisted of an exhaustive pair-wise comparison of all variants of all words. In each trial, a subject would hear, e.g., a version of the word *Fahrt* - $[\text{fa:r}\vartheta]$ (Engl. *drive*) synthesized with $[\vartheta]$ and a version of the same word synthesized with $[\vartheta_{\text{low}}]$ (separated by 500 ms of silence) and choose the one they considered to sound more natural. The playback of the stimuli could be repeated up to five times and do-overs were allowed. The order of the stimuli was randomized for every subject both within and across all trials. In total, the number of trials was therefore 14 (number of words with four variants) times 12 (number of pair-wise combinations of the four variants) plus 6 (number of combinations of the three variants of *Ober* - $[\text{o:b}\vartheta]$) = 174 trials. The experiment was completed by 26 subjects (16 male, 10 female, age 19 to 39, median age 22.5, all native German speakers with no reported hearing impairments), who gave informed consent and received a suitable monetary compensation for their time. It was conducted in a soundproofed recording studio using a Xiberia V10 USB stereo semi-open headset, and Praat version 6.0.43 on a Windows 8.1 desktop PC.

C. Results

The aim of the experiment was to rank the allophone shapes for each context using the subjects' answers. To that end, we counted for every word how often each subject preferred each of the four possible shapes in the pair-wise comparisons (which could happen at most six times because each shape was only part of six pairs for each word). This led to 15 (number of words) tables with four (number of shapes) columns and 26 (number of subjects) rows. We then conducted an exact two-sided paired Wilcoxon signed-rank test between each pair of columns to obtain a final preference ranking. The results of the experiment are summarized in Table II.

In absolute numbers, the context-dependent $[V\vartheta]$ vocal tract shapes were the highest ranked variant for only four of the fifteen contexts and for only one of these contexts ($[\text{Y}\vartheta]$) was the difference to the next highest-ranked shape significant. In other words: In 14 of the 15 cases, the context-dependent allophone shapes could be replaced by one of the two new prototypes $[\vartheta_{\text{low}}]$ or $[\vartheta_{\text{mid}}]$ without a significant loss

in preference. For $[\text{œ}\vartheta]$ and $[\text{u}\vartheta]$, one of the new prototypes was even significantly higher ranked than the original context-dependent shape, despite the quite large acoustic difference between them (see Figure 6). This is likely due to the fact that the original speaker hardly even articulated the $/r/$ sound in those combinations, as is evident from the very short transitions for $[\text{œ}\vartheta]$ and $[\text{u}\vartheta]$ in the resonance map in Figure 6a. It appears that the listeners preferred a more distinguishable secondary diphthong over the quasi-monophthong realized by the original speaker. Another notable observation is that in the word *Ober* - $[\text{o:b}\vartheta]$, where the original shape would have been the shape $[\vartheta]$, the subjects still preferred the shape $[\vartheta_{\text{low}}]$. The reason may be the regional accent of the original speaker, who produced the $[\vartheta]$ a little more open than what might be considered Standard German.

IV. CONCLUSIONS

The results from the listening experiment allow the conclusion that the variety of vocalized allophones of $/r/$ can be entirely described by only two vocal tract target shapes: $[\vartheta_{\text{mid}}]$ and $[\vartheta_{\text{low}}]$. The context that best fits each of these two shapes can also be derived from the rankings, which leads to the diphthong vectors given in Figure 6b. The emerging pattern was not predicted by any of the three conventions mentioned in subsection I-A, but could be explained by a mix of all three with one addendum: A fricative allophone is not necessary for these contexts (as was correctly described by Kohler) but a distinction between contexts still needs to be made (as was done by the Duden and the Deutsche Aussprachewörterbuch). However, in contrast to both of these conventions, the distinction should not be based on the tenseness of the context vowel, but on the vowel height *and* a contrast criterion: As can be seen in Figure 6b, the diphthongs ending in the $[\vartheta_{\text{low}}]$ configuration mostly start at low vowels, while the ones ending in $[\vartheta_{\text{mid}}]$ mostly start at mid to high vowels. The exceptions from this are $[\text{œ}\vartheta]$, $[\text{e}\vartheta]$, and $[\text{o}\vartheta]$, which are all nearly high or high vowels but still end on $[\vartheta_{\text{low}}]$. A likely explanation, reminiscent of Lindblom's Dispersion Theory [42], is that this perceptual contrast is helpful to discriminate between the otherwise very similar pairs $[\text{œ}\vartheta]$ versus $[\text{y}\vartheta]$, $[\text{e}\vartheta]$ versus $[\text{i}\vartheta]$, and $[\text{o}\vartheta]$ versus $[\text{u}\vartheta]$.

V. SUMMARY AND OUTLOOK

We conducted a single-speaker study of the German $/r/$ allophones in syllable coda position to find the minimum number of underlying vocal tract shapes representing these sounds that are required to synthesize convincing, natural sounding speech. To that end, we recorded and analyzed 300 utterances containing 20 examples of each of the 15 possible contexts. The main goal of the analysis was to find patterns in the formant data of the 300 allophonic realizations in a data-driven way with as little manual intervention as possible. Using unsupervised techniques, we identified two prototypes $[\vartheta_{\text{low}}]$ and $[\vartheta_{\text{mid}}]$, that are sufficient to synthesize natural sounding vocalic $/r/$ allophones across all investigated contexts, as confirmed by a listening experiment. These two prototypes are not free-variant, but complementary allophones since they

Allophone	Carrier word		Frequency of preference				Ranking (*significant at $p < 0.01$)
			[e]	[e _{mid}]	[e _{low}]	[V _e]	
[a _r e]	Fahrt	[fa:rət]	77	52	84	99	[e _{mid}] < [e] < [e _{low}] < [V _e]
[e _r e]	er	[e:rə]	79	22	111	100	[e _{mid}] < * [e] < [V _e] < [e _{low}]
[i _r e]	wir	[vi:rə]	33	92	74	113	[e] < * [e _{low}] < [e _{mid}] < [V _e]
[o _r e]	Ohr	[o:rə]	47	51	96	118	[e] < [e _{mid}] < * [e _{low}] < [V _e]
[u _r e]	Uhr	[u:rə]	31	119	65	97	[e] < * [e _{low}] < [V _e] < [e _{mid}]
[ɛ _r a]	Air	[ɛ:rə]	76	25	113	98	[e _{mid}] < * [e] < [V _e] < [e _{low}]
[ø _r e]	Stör	[ʃtø:rə]	25	97	94	96	[e] < * [e _{low}] < [V _e] < [e _{mid}]
[y _r e]	Tür	[ty:rə]	19	130	50	113	[e] < * [e _{low}] < * [V _e] < [e _{mid}]
[ɐ]	Ober	[ˈo:bɐ]	52	8	96	–	[e _{mid}] < * [e] < * [e _{low}]
[ɛɐ]	Verb	[vɛ:p]	106	12	101	93	[e _{mid}] < * [V _e] < [e _{low}] < [e]
[rɐ]	Hirsch	[hɪrʃ]	22	116	58	116	[e] < * [e _{low}] < * [V _e] = [e _{mid}]
[ɔɐ]	Ort	[ɔ:t]	76	35	124	77	[e _{mid}] < * [e] < [V _e] < * [e _{low}]
[ʊɐ]	Gurt	[gʊ:t]	20	132	69	91	[e] < * [e _{low}] < [V _e] < * [e _{mid}]
[œa]	Jörg	[jœɐg]	51	42	110	109	[e _{mid}] < [e] < * [V _e] < [e _{low}]
[ʏɐ]	Fürst	[fʏrʃt]	20	98	66	128	[e] < * [e _{low}] < * [e _{mid}] < * [V _e]

TABLE II: Results from the listening experiment. The numbers are the absolute frequency of preference summed up across all subjects (higher is better). Each ranking was tested for significance using an exact two-sided paired Wilcoxon signed-rank test (the relational operator “<” here means “less preferred”).

cannot be used interchangeably but are context-dependently preferred, as confirmed by the listening experiment. The discovered pattern of diphthong vectors using these prototypical sounds reconciles and expands the existing conventions for the production of vocalic /r/ allophones. Future work should investigate if the identified pattern can also be observed in larger speaker populations, since all analyses in this work were based on the reference speaker of the articulatory synthesizer VocalTractLab. To that end, more speakers’ vocal tract geometries need to be acquired and added to VocalTractLab.

ACKNOWLEDGMENT

This work was funded by the German Federal Ministry for Economic Affairs and Energy, reference number ZF4443004BZ8. We thank our speaker for the recordings, Klaus Willmes for his advice regarding the evaluation of the listening experiment, and Ursula Hirschfeld for her valuable feedback during the preparation of this manuscript. We also thank the anonymous reviewers for their comments and constructive criticism.

REFERENCES

- [1] C. H. Shadle and R. I. Damper, “Prospects for articulatory synthesis: A position paper,” in *Fourth ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW-4)*, Perthshire, Scotland, 2001.
- [2] B. Cranen and J. Schroeter, “Physiologically motivated modelling of the voice source in articulatory analysis/synthesis,” *Speech Communication*, vol. 19, pp. 1–19, 1996.
- [3] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, “A review of lumped-element models of voiced speech,” *Speech Communication*, vol. 55, no. 5, pp. 667–690, 2013.
- [4] P. Birkholz, S. Drechsel, and S. Stone, “Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis,” in *Proc. of the Interspeech*, Graz, Austria, 2019, pp. 3765–3769. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2410>
- [5] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savari-aux, “Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images,” *Journal of Phonetics*, vol. 30, pp. 533–553, 2002.
- [6] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.
- [7] M. Arnela, S. Dabbaghchian, O. Guasch, and O. Engwall, “MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2173–2182, 2019.
- [8] S. Maeda, “A digital simulation method of the vocal-tract system,” *Speech communication*, vol. 1, no. 3-4, pp. 199–229, 1982.
- [9] P. Birkholz, D. Jackèl, and B. J. Kröger, “Simulation of losses due to turbulence in the time-varying vocal system,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.
- [10] K. van den Doel and U. M. Ascher, “Real-time numerical solution of Webster’s equation on a nonuniform grid,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1163–1172, 2008.
- [11] B. Elie and Y. Laprie, “Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink,” *Speech Communication*, vol. 82, pp. 85–96, 2016.
- [12] P. Birkholz, “Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets,” in *Proc. of the Eurospeech*, Antwerp, Belgium, 2007, pp. 2865–2868.
- [13] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, “Model-based reproduction of articulatory trajectories for consonant-vowel sequences,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2011.

- [14] B. H. Story and K. Bunton, "A model of speech production based on the acoustic relativity of the vocal tract," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2522–2528, 2019.
- [15] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan, "A modular architecture for articulatory synthesis from gestural specification," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4458–4471, 2019.
- [16] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C. Browman, "CASYS and extensions to the task-dynamic model," in *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data*, Autrans, France, 1996, pp. 125–128.
- [17] J. Stark, B. Lindblom, and J. Sundberg, "APEX an articulatory synthesis model for experimental and computational studies of speech production," *TMH-QPSR*, vol. 2, no. 1996, pp. 45–48, 1996.
- [18] S. Fels, F. Vogt, K. Van Den Doel, J. Lloyd, I. Stavness, and E. Vatikiotis-Bateson, "Artisynth: A biomechanical simulation platform for the vocal tract and upper airway," in *Proc. of the 7th International Seminar on Speech Production*, Ubatuba, Brazil, 2006.
- [19] A. J. Teixeira, R. Martinez, L. N. Silva, L. M. Jesus, J. C. Príncipe, and F. A. Vaz, "Simulation of human speech production applied to the study and synthesis of European Portuguese," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1435–1448, 2005.
- [20] L.-J. Liu, C. Ding, Y. Jiang, M. Zhou, and S. Wei, "The IFLYTEK system for Blizzard Challenge 2017," in *Blizzard Challenge Workshop*, 2017.
- [21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [22] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Proc. of the Interspeech*, Jeju, Korea, 2004, pp. 1125–1128.
- [23] J. A. Marwitz, S. Stone, and P. Birkholz, "Optimierung der Numerik eines linearen Gleichungssystems für die Simulation des Schallfeldes im Vokaltrakt," in *Studenten zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, A. Berton, U. Haiber, and W. Minker, Eds. Ulm, Germany: TUDpress, Dresden, 2018, pp. 359–366.
- [24] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. I. Toulouse, France: IEEE, 2006, pp. 873–876.
- [25] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds," in *First International Workshop on Performative Speech and Singing Synthesis (p3s 2011)*, Vancouver, BC, Canada, 2011.
- [26] —, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.
- [27] T. Ellbogen, F. Schiel, and A. Steffen, "The BITS speech synthesis corpus for German," in *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004.
- [28] T. Siebs, *Grundzüge der Bühnenaussprache*. Verlag von A. Ahn, 1900.
- [29] C. Ulbrich and H. Ulbrich, "Realisations and alternations in German /r/-realisation," in *Proc. of the Interspeech*, Antwerp, Belgium, 2007, pp. 2733–2736.
- [30] R. Wiese, "The unity and variation of (German) /r/," *Etudes & Travaux: 'r-atics*, no. 4, pp. 11–26, 2001.
- [31] A. P. Simpson, "Accounting for the phonetics of German r without processes," *ZAS Papers in Linguistics*, vol. 11, pp. 91–104, 1998.
- [32] K. Kohler, "German," in *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, International Phonetic Association, Ed. Cambridge, UK: Cambridge University Press, 1999, pp. 86–89.
- [33] S. Kleiner, R. Knöbl, and M. Mangold, *Duden - Das Aussprachewörterbuch*, 7th ed., ser. Duden - Deutsche Sprache in 12 Bänden, Bibliographisches Institut GmbH, Ed. Berlin, Germany: Dudenverlag, 2015, vol. 6.
- [34] E. Krech, E. Stock, U. Hirschfeld, and L. Anders, *Deutsches Aussprachewörterbuch*. Berlin, Germany: Walter De Gruyter GmbH, 2009.
- [35] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," *Glott international*, vol. 5, pp. 341–345, 01 2001.
- [36] T. Kathiresan, D. Maurer, H. Suter, and V. Dellwo, "Enhancing the objectivity of interactive formant estimation: Introducing euclidean distance measure and numerical conditions for numbers and frequency ranges of formants," in *Studenten zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, J. Trouvain, I. Steiner, and B. Möbius, Eds. Saarbrücken, Germany: TUDpress, Dresden, 2017, pp. 130–137.
- [37] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 97–100, 1990.
- [38] C. H. Shadle, H. Nam, and D. Whalen, "Comparing measurement errors for formants in synthetic and natural vowels," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 713–727, 2016.
- [39] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.
- [40] P. von Polenz, *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart*, 2nd ed., ser. De-Gruyter-Studienbuch. Berlin, Germany: De Gruyter, 2000, vol. 1.
- [41] P. Birkholz, F. Gabriel, S. Kürbis, and M. Echternach, "How the peak glottal area affects linear predictive coding-based formant estimates of vowels," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 223–232, 2019.
- [42] J. Liljencrants and B. Lindblom, "Numerical simulation of vowel quality systems: The role of perceptual contrast," *Language*, vol. 48, no. 4, pp. 839–862, 1972.

APPENDIX
WORD LIST OF THE RECORDED DATA SET

/a:ɐ/		/e:ɐ/	
Aargau	[ˈa:ɐ,ɡaʊ]	Abkehr	[ˈapke:ɐ]
Acrylharz	[aˈkʁy:l,ha:ɐts]	allererst	[ˈalɐ,ʔe:ɐst]
Agrarminister	[aˈɡʁa:ɐmi,nɪstɐ]	Beerdigung	[bɐˈʔe:ɐdɪɡʊŋ]
Altar	[al,tɑ:ɐ]	Begehr	[bɐˈɡe:ɐ]
alveolar	[alveoˈla:ɐ]	Beschwerde	[bɐˈʃve:ɐdɐ]
Archivar	[ˌaʁçiˈva:ɐ]	der	[de:ɐ]
Arten	[ˈa:ɐtɪn]	Ehrfurcht	[ˈe:ɐfʊɾtʃt]
atomar	[atoˈma:ɐ]	er	[e:ɐ]
autark	[aʊˈta:ɐk]	Erde	[ˈe:ɐdɐ]
bar	[ba:ɐ]	erstens	[ˈe:ɐstɪns]
Barkeeper	[ˈba:ɐki:pɐ]	Erzgebirge*	[ˈɛʁtsgə,bɪʁɡə]
Bart	[ba:ɐt]	Gerhard	[ˈɡe:ɐɦaʁt]
Basar	[baˈza:ɐ]	Gewehr	[ɡəˈve:ɐ]
behaart	[bɐˈha:ɐt]	Herd	[he:ɐt]
Darbietung	[ˈda:ɐbi:tʊŋ]	quer	[kvɛ:ɐ]
Fahrsignal	[ˈfa:ɐzi,ɡna:l]	sehr	[ze:ɐ]
Fahrt	[fa:ɐt]	verehrt	[fɛʁˈʔe:ɐt]
formbar	[ˈfɔɾmba:ɐ]	Verkehr	[fɛʁˈke:ɐ]
Jahr	[ja:ɐ]	werden	[ˈve:ɐdn̩]
molar	[moˈla:ɐ]	Wert	[ˈve:ɐt]

/ɛ:ɐ/		/ɛʁ/	
Air	[ɛ:ɐ]	Absperrhahn	[ˈapʃpɛʁ,ɦa:n]
Aktionär	[aktsjoˈnɛ:ɐ]	absterben	[ˈap,ʃtɛʁbɪn]
Ärzte*	[ˈɛʁtstɐ]	abwärts	[ˈapvɛʁts]
autoritär	[aʊtoʁiˈtɛ:ɐ]	Allergie	[ˌalɛʁˈɡi:]
Bär	[bɛ:ɐ]	anlernen	[ˈan,lɛʁnɐn]
bewährt	[bɐˈvɛ:ɐt]	Ärger	[ˈɛʁɡɐ]
Chair	[tʃɛ:ɐ]	aufmerksam	[ˈaʊf,mɛʁkza:m]
elitär	[ˌeliˈtɛ:ɐ]	ausmerzen	[ˈaʊs,mɛʁtsn̩]
erklärt	[ɛʁˌkle:ɐt]	Beherrscher	[bɐˈhɛʁʃɐ]
Fährboot	[ˈfɛ:ɐbɔ:t]	Berg	[bɛʁk]
Fairness	[ˈfɛ:ɐnɛs]	Berlin	[bɛʁˈli:n]
Gärprozess	[ˈɡɛ:ɐpʁɔ,tʃɛs]	Bern	[bɛʁn]
Gebärde	[ɡɐˈbɛ:ɐdɐ]	Bertram	[ˈbɛʁtsʁam]
gefährlich	[ɡɐˈfɛ:ɐlɪç]	derb	[dɛʁp]
geklärt	[ɡɐˌkle:ɐt]	Ferse	[ˈfɛʁzɐ]
Härchen	[ˈhɛ:ɐçɐn]	fertig	[ˈfɛʁtɪç]
imaginär	[imaɟiˈnɛ:ɐ]	Handwerker	[ˈɦant,vɛʁkɐ]
Kläranlage	[ˈkle:ɐˈʔanla:ɡɐ]	Verb	[vɛʁp]
Legionär	[leɟjoˈnɛ:ɐ]	verwandt	[fɛʁˈvant]
Märchen	[ˈmɛ:ɐçɐn]	Werkstatt	[ˈvɛʁk,ʃtat]

/i:ɐ/		/ɪʁ/	
basiert	[baˈzi:ɐt]	Birne*	[bø:ɐne]
Bier	[bi:ɐ]	Dirne	[ˈdɪrɐnɐ]
Bierchen	[ˈbi:ɐçɐn]	Firma	[ˈfɪɾma]
Haustier	[ˈɦaʊs,ti:ɐ]	Hirn	[ɦɪɾn]
hierbei	[hi:ɐˈbaɪ]	Hirnnerv	[ˈɦɪɾn,nɛʁf]
Klavier	[klaˈvi:ɐ]	Hirsch	[ɦɪʁç]
Kurier	[kuˈri:ɐ]	Hirse	[ˈɦɪʁzɐ]
mir	[mi:ɐ]	Irrtum	[ˈɪɾtu:m]

Continued on next page

TABLE III – continued from previous page

Neugier	[ˈnɔ̃ɡi:ʁə]	Kirche	[ˈkɪrçə]
Papier	[paˈpi:ʁə]	Kirsche	[ˈkɪʁʃə]
Stier	[sti:ʁə]	Pfirsich	[ˈpfɪʁzɪç]
Tier	[ti:ʁə]	Quirl	[kvɪʁl]
Untier	[ˈʊnˌti:ʁə]	Schirm	[ʃɪʁm]
Vier	[fi:ʁə]	Stirn	[ʃtɪʁn]
Viereck	[ˈfi:ʁəˌʔɛk]	Vierzig	[vɪʁt͡sɪç]
viertes	[ˈfi:ʁətə]	wirbeln	[ˈvɪʁəbəl̩n]
Visier	[viˈzi:ʁə]	Wirkstoff*	[ˈvɪʁəkˌʃtɔf]
Weißbier	[ˈvaɪ̯sbɪ:ʁə]	Wirt	[vɪʁt]
wir	[vi:ʁə]	Zirkel	[ˈtsɪʁkəl]
zierlich	[ˈtsi:ʁəlɪç]	Zirkus	[ˈtsɪʁkʊs]
/o:ʁ/		/œ/	
Amor	[ˈa:mɔ:ʁə]	Ahorn	[ˈa:hɔ:ʁən]
Adaptor	[aˈdaptɔ:ʁə]	Akkordeon	[aˈkɔ:ʁədɔ:ʁən]
Bohrloch	[ˈbɔ:ʁəlɔx]	Formel	[ˈfɔ:ʁəm]
Chor	[kɔ:ʁə]	Fürsorge	[ˈfʏʁzɔ:ʁə]
Chlor	[klo:ʁə]	Horn	[hɔ:ʁən]
davor	[daˈfo:ʁə]	Organisation	[ˌɔ:ʁganizaˈt͡sjɔ:n]
Dekor	[deˈko:ʁə]	Ort	[ɔ:ʁt]
Depressor	[deˈpʁɛsɔ:ʁə]	Orgel	[ˈɔ:ʁgəl]
Diffusor	[dɪˈfu:zɔ:ʁə]	Skorbut	[skɔ:ʁbu:t]
empor	[ɛmˈpɔ:ʁə]	sofort	[zoˈfɔ:ʁt]
erfror	[ɛʁˈfʁɔ:ʁə]	Sorge	[ˈzɔ:ʁə]
Gregor	[ˈgʁɛ:ɡɔ:ʁə]	Sorgfalt	[ˈzɔ:ʁkfalt]
Humor	[huˈmɔ:ʁə]	Sporn	[ʃpɔ:ʁən]
Junior	[ˈju:njɔ:ʁə]	Sport	[ʃpɔ:ʁt]
Labor	[laˈbo:ʁə]	Steuerbord	[ˈʃtɔ:ʁəˌbɔ:ʁt]
Motor	[ˈmo:tɔ:ʁə]	Storch	[ʃtɔ:ʁç]
Ohr	[o:ʁə]	Support	[suˈpɔ:ʁt]
Rohr	[ʁɔ:ʁə]	Torf	[tɔ:ʁf]
Vorschrift	[ˈfo:ʁʃʁɪft]	Torso	[ˈtɔ:ʁzɔ]
Vorsprung	[ˈfo:ʁʃpʁʊŋ]	Worte	[ˈvɔ:ʁtə]
/ø:ʁ/		/œʁ/	
Akteur	[akˈtø:ʁə]	Beförderung	[bəˈfœʁədɔ:ʁʊŋ]
Behörde	[bəˈhø:ʁədə]	Björn	[bjœʁən]
Charmeur	[ʃaʁˈmø:ʁə]	Börse	[ˈbœʁzə]
Chauffeur	[ʃoˈfø:ʁə]	Dörfchen	[ˈdœʁfçən]
empört	[ɛmˈpø:ʁət]	Dörte	[ˈdœʁtə]
Förde	[fœ:ʁədə]	Erörterung	[ɛʁˈœʁətɔ:ʁʊŋ]
Frisör	[fʁiˈzø:ʁə]	Förderer	[ˈfœʁədɔ:ʁə]
Gehör	[gəˈhø:ʁə]	gehört	[geˈhœʁnt]
gehört	[gəˈhø:ʁət]	gekörnt	[geˈkœʁnt]
Hörfunk	[ˈhø:ʁfʊŋk]	Hörnchen	[ˈhœʁnçən]
Ingenieur	[ɪŋʒeˈniø:ʁə]	Jörg	[jœʁk]
Jongleur	[ʒɔŋˈljø:ʁə]	Körbchen	[ˈkœʁpçən]
Kommandeur	[kɔmanˈdø:ʁə]	Körper	[ˈkœʁpə]
Likör	[liˈkø:ʁə]	Mörser	[ˈmœʁzə]
Masseur	[maˈsø:ʁə]	nördlich	[ˈnœʁtlɪç]
Redakteur	[ʁɛdakˈtø:ʁə]	Örtchen	[ˈœʁtçən]
Stör	[ʃtø:ʁə]	Pförtchen	[ˈpfœʁtçən]
Öhrchen	[ˈø:ʁçən]	Schnörkel	[ˈʃnœʁkəl]
ungestört	[ˈʊŋgəˌʃtø:ʁət]	Surfbrett	[ˈsœʁfˌbrɛt]

Continued on next page

TABLE III – continued from previous page

Zubehör	['tsu:bə,hø:ə]	unförmig	['ʊn,fœʁmɪç]
/u:ə/		/ʊə/	
Abfuhr	['ap,fu:ə]	absurd*	[apsɔət]
Busspur	['bʊs,ʃpu:ə]	Burg	[bʊək]
Figur	['fi'gu:ə]	Burgunder	[bʊə'gʊndə]
Flur	[flu:ə]	Bursche	['bʊʃə]
Frisur	['fʁi'zu:ə]	Diskurs	[dis'kʊʁs]
Haarkur	['ha:ə,ku:ə]	durch	[dʊʁç]
Klausur	['klaʊzu:ə]	Endspurt*	['ɛnt,ʃpɔ:ət]
Kultur	['kʊlt'tu:ə]	Entwurf*	['ɛnt'vɔ:ʁf]
Kur	[ku:ə]	Frankfurt	['fʁaŋkfuət]
Mixtur	['miks'tu:ə]	Furcht	[fʊʁçt]
Natur	[na'tu:ə]	Gurgel	['gʊʁ,ɡl]
Parkuhr	['pɑʁk'ʔu:ə]	Gurke	['gʊʁkə]
Radtour	['ʁat'tu:ə]	Gurt	[ɡʊʁt]
Schnur	[ʃnu:ə]	hurtig*	['ho:ʁtɪç]
Spur	[ʃpu:ə]	Kurbel*	['kʊ:ʁbl]
stur	[stu:ə]	Kurse*	['kʊ:ʁzə]
Uhr	[u:ə]	Murks	[mʊʁks]
Uhrzeit	['u:ʁ,tʁaɪt]	Turm	[tʊʁm]
Urwald	['u:ʁ,vɑlt]	Wurst*	[vɔ:ʁst]
zur	[tsu:ə]	Wurzel*	['vɔ:ʁzəl]
/y:ə/		/ʏə/	
dafür	[da'fy:ə]	Bedürfnis	[be'dyʁfnɪs]
Figürchen	['fi'gy:ʁçən]	Bürde	['byʁdə]
Führer	['fy:ʁə]	Bürger	['byʁgə]
anführen	['an,ʃy:ʁən]	dürfen	['dyʁfn]
ausführen	['aʊs,ʃy:ʁən]	ehrwürdig	['e:ʁ,vyʁdɪç]
führt	[fy:ʁət]	Erstürmung	['ɛʁ'ʃtʏʁmʊŋ]
Fürsorge	['fy:ʁ,zʊʁgə]	fürchten	['fyʁçtn]
Gebühr	[gə'by:ə]	Fürst	[fyʁst]
Geschwür	[gə'ʃvy:ə]	gebürtig	[gə,bɪʁtɪç]
Gespür	[gə'ʃpy:ə]	Gewürz	[gə'vyʁts]
Kür	[ky:ə]	Gürtel	['gyʁtl]
natürlich	[na'ty:ʁlɪç]	Hürde	['hyʁdə]
Rührkuchen	['ʁy:ʁ,ku:χn]	Jürgen	['jyʁgən]
Schnürschuhe	['ʃny:ʁ,ʃu:ə]	Mürbeteig	['myʁəbə,taiç]
Schürhaken	['ʃy:ʁ,ha:kən]	Nürnberg	['nyʁn,bɛək]
Spürhund	['ʃpy:ʁ,hʊnt]	schlürfen	['ʃlyʁfn]
Tür	[ty:ə]	Schürze	['ʃyʁtsə]
Türchen	['ty:ʁçən]	stürzen	['ʃtʏʁtsn]
unberührt	['ʊnbə,ʁy:ʁt]	Würde	['vyʁdə]
Willkür	['vɪlky:ə]	Würfel	['vyʁfl]
/ɐ/			
Acker	['akɐ]	Bremser	['brɛ:mzɐ]
Ader	['adɐ]	Butter	['bʊtɐ]
Ascher	['aʃɐ]	Camper	['kɛmpɐ]
Bagger	['bagɐ]	Dampfer	['dampfɐ]
Banner	['banɐ]	Dichter	['dɪçtɐ]
Bastler	['bastlɐ]	Eber	['e:bɐ]
Bettler	['bɛtlɐ]	Eimer	['aɪmɐ]

Continued on next page

TABLE III – continued from previous page

Biber	[ˈbi:bɐ]	Farmer	[ˈfɑɪmɐ]
Bohrer	[ˈbo:ʁɐ]	Feier	[ˈfaɪ̯ɐ]
Bomber	[ˈbɔmbɐ]	Ober	[ˈo:bɐ]

TABLE III: Word list of the recorded data set. The transcription convention used here follows Kohler [32] for simplicity's sake and should not be considered phonetically precise in a narrow sense. The words marked with an asterisk were pronounced in a variant way (as transcribed) and relabeled accordingly. The grayed-out words were excluded from the set because of large deviations of their formants from the respective group's mean (see subsection II-C).
