Phonetic analysis of dialect/standard transitions synthesized by model-based interpolation

MICHAEL PUCHER[1]
SYLVIA MOOSMÜLLER (†)[1]

Abstract. Transitions between regional standard varieties of Austrian German and dialect varieties can be synthesized by means of an interpolation function based on Hidden Markov Models that allows for the generation of intermediate varieties. Hidden Markov Models of language varieties can be automatically interpolated on a sub-phonemic state level to generate speech of intermediate varieties. The interactions between regional standard varieties of Austrian German and dialects can be represented as phonological processes or input-switch-rules. Phonological processes are gradual and phonetically motivated; input-switch-rules show a different historical development for each variety and have no synchronic phonetic relation. In this contribution, we analyse a representative sample of such synthesized dialect/standard interactions for four speakers of the Austrian dialect of Innervillgraten and the transitions to regional Standard Austrian German. We show that the synthesizer produces input-switch-rules and phonological processes at the formant level by using a linear interpolation at the Mel-cepstral feature level and explain why this happens. A statistical analysis of formant differences is provided that clearly differentiates between input-switch-rules and phonological processes. This result supports the two-competence▪model, which assumes that speakers of Austrian German hold both a competence in Standard Austrian German and in a specific dialect.

INTRODUCTION

This paper deals with the interpolation of synthesized language varieties, or, more specifically, with the interpolation between the synthesis of a standard language and the synthesis of a dialect and, on that basis, elaborates the specific relationship between these two language varieties by discussing the outputs of the interpolation steps. In interpolating between the two language varieties, it turned out, to our surprise, that the interpolation algorithm produced different states of qualitatively different variables and thus corroborated the two-competence model developed by Dressler and colleagues (Dressler and Wodak, 1982; Dressler et al., 1989). The two-competence model, in the interaction of two language varieties, differentiates between alternations of phonological variables that

[1] Acoustics Research Institute, Austrian Academy of Sciences

lack a phonetic relationship and alternations that hold such a relation by showing a phonetic motivation and intermediate steps in the processing from one variable to the other.

The former are dubbed input-switch-rules; the latter refer to phonological processes. Most interestingly, in the interpolation between alternations that lack a phonetic motivation, jumps occur in the interpolation states, whereas interpolating phonological processes gives rise to continuous intermediate states. Before providing details about the two-competence model, which is an excellent tool for analysing (socio)phonological variation, since it includes a phonological explanation for the application of (socio)phonological variables, we will provide an overview of speech synthesis technology.

Using flexible state-of-the-art statistical parametric speech synthesis technology based on Hidden Markov Models (HMMs) we synthesized the above-mentioned alternations and processes through the use of model interpolation as already shown in Toman et al. (2015) and Pucher et al. (2010b). Dialect interpolation can be performed at a phonemic (Pucher et al., 2010b) or sub-phonemic state level (Toman et al., 2015). Interpolation of speaker models has been applied for speaker identity (Yoshimura et al., 1997), emotional speech (Tachibana et al., 2005), speaking rate (Pucher et al., 2010a), dialect (Pucher et al., 2010b; Toman et al., 2015), and accent (Astrinaki et al., 2013).

In the context of speaker identity we interpolate between two synthetic voices of different speakers, speaker 1 and speaker 2. Interpolation then allows for a gradual transition from speaker 1 to speaker 2. With emotional speech, one can interpolate between different emotional states of a specific speaker to realize a gradual transition. For dialect or accent interpolation, we also use data from one speaker in standard and dialect/accent and the interpolation occurs between standard and dialect/accent. In addition, adaptive approaches have received much attention in speech synthesis (Tamura et al., 1998, 2001; Yamagishi et al., 2004; Isogai et al., 2005; Yamagishi et al., 2006; Yamagishi and Kobayashi, 2007; King et al., 2008; Yamagishi et al., 2009) mainly due to the rise of statistical parametric speech synthesis (Zen et al., 2004). Adaptive modelling has been applied to the speaker (Yamagishi and Kobayashi, 2007), emotion (Qin et al., 2006), accent (Wester and Karhila, 2011; Karhila and Wester, 2011), dialect (Pucher et al., 2010b), type of articulation (Picart et al., 2014), and dysarthric speech (Veaux et al., 2012). The flexibility of HMM-based synthesis also allows for the integration of articulatory features (Ling et al., 2009) and the control of the acoustic model by articulatory features

(Ling et al., 2008), formant features (Lei et al., 2011), or visual features (Hollenstein et al., 2013).

This shows the wide range of applications of HMM-based speech synthesis. The interpolation algorithm used in this paper was first reported in Toman et al. (2015), where we also performed listening tests with interpolated samples for several dialects to showcase the method's possibility to generate intermediate language varieties in an unsupervised way. In that paper, we did not further analyse the interpolated samples and no statistical analysis of formant changes during interpolation was reported. We also used three dialects with one speaker each, which did not allow for an analysis of speaker similarities within a dialect.

In the current paper, which focuses on one dialect, we aim to show how the interpolation algorithm is able to automatically generate input-switch-rules as well as phonological processes, and how this supports the two-competence model.

THE TWO-COMPETENCE MODEL AND INTERPOLATION

The results of our work entail two important theoretical implications. From the interpolation perspective, the results show not only how linear interpolation deals with 'holes' containing no information in the case of an input-switch-rule, but also, that linear interpolation can deal with such 'either – or' forms. This shows the flexibility of the unsupervised interpolation algorithm that can deal with complex phenomena through a simple linear interpolation on the model level. With this, we can model input-switch-rules and phonological processes using the same unsupervised interpolation method, which shows the flexibility of the HMM-based synthesis paradigm.

From a sociolinguistic perspective, we were able to corroborate the two-competence model by demonstrating that (socio)phonological variation, the dialect-standard-interaction in our case, is not necessarily linear. On the one hand, linear interpolation reacts to qualitative differences of phonological variables and produces a jump for e.g., an /ʊ/ ↔ /i/ alternation, which shows no intermediate steps in real speech behaviour and on the other hand, generates continuous transitions for a change from [ɔ] → [ɔ ʊ], analogous to real speech behaviour. Since linear interpolation produces different outputs for phonetically similar inputs, this shows that there is a qualitative difference between a standard-dialect input-switch-rule, e.g., /ɑ/ ↔ /ɔ/ and an alternation which changes a vowel in a specific phonetic context, such as, e.g., [ɐ] → [ɔ]. We chose this example, because

it shows that the output of the input-switch-rule /ɑ/ ↔ /ɔ/ and the output of the phonological process [ɐ] → [ɔ] differ. The former contains jumps and the latter has continuous steps, although both alternations involve similar vowel qualities.

We propose a computational model for a speaker with competence in a standard and a dialect variety, that is trained on speaker data from both varieties and uses interpolation for phonological processes and input-switch-rules. We show that this model is able to simulate phonological processes and input-switch-rules and thus can support the two-competence model in linguistics.

The possibility of a computational model using only data from two speaker varieties is a necessary condition for the two-competence model and thus supports it. This means that if it were not possible to develop such a computational model then the two-competence model would be false. Our computational model is however not equivalent to the two-competence model, i.e. the truth of the two-competence model does not follow from the existence of such a computational model. The fact that such equivalences are hard to establish is a general problem of the comparison between computational models and linguistically motivated models or theories. What we show in this paper is that our computational model fulfils certain adequacy conditions for being a computational model of a two-competence speaker.

1. The model is adequate since it is only based on data from two varieties (i.e. competences) per speaker,
2. and it produces input-switch-rules and phonological processes on the formant level.

Previously we have already shown that the model produces perceptually sensible results in terms of dialect authenticity (Toman et al., 2015), which is also part of its adequacy.

Furthermore the process of dialect levelling can also be analysed within our model. When we interpolate between a standard and a dialect and increase the weight of the standard variety, the input-switch-rule will be applied, which will result in the standard pronunciation. For the phonological process a gradual transition between standard and dialect will take place, gearing the interpolated variety towards the standard. The resulting interpolated variety will exhibit some dialect features from the phonological processes while avoiding the marked dialect features (Auer, 2017) that are realized as input-switch-rules.

ANALYSIS OF INTERPOLATION ON THE SPECTRAL LEVEL

In this section we analyse the interpolation on the spectral envelope level, for spectra that are generated from the interpolated Mel-cepstral (MCEP) features, which is an intermediate step in the synthesis process using our HMM-based synthesis system. MCEP features allow for a low-dimensional representation of the spectral information and also have an auditory weighting that accounts for spectral perception by humans. We will restrict this analysis to one input-switch-rule of a female speaker and one phonological process of a male speaker. In Section VI, a phonetic analysis on the formant level is performed for all speakers and samples from Table 2. The spectral analysis is done at the level of individual processes within a word; the interpolation itself is, however, performed on the level of whole utterances because the speech synthesis is done on the utterance level. Through the automatic alignment of utterances with Dynamic Time Warping, words and phones are aligned automatically. We are analysing features from HMM states where the respective phones are mapped onto each other. Since the used HMMs do not allow for state skipping, the interpolated durations will generate a positive number of feature frames n for each state ($n \geq 1$).

INPUT-SWITCH-RULE (SPEAKER C)

Figure 1 shows the spectra generated from the MCEP features over time for the Regional Standard Austrian German (RSAG) to Innervillgraten (IVG) interpolation from /ʊ/ to /iː/ in unser ‘our’ ([ʊnsɐ] ↔ [iːnsɔ]). This example was synthesized with a female voice. In this case all five /ʊ/-states are mapped onto the five /iː/-states, which is, however, not necessarily so for all outcomes of the interpolation algorithm, since the algorithm can also deal with sequences having a different number of phones and thereby a different number of states.

The interpolation with our algorithm is without any direction, which means that interpolation from standard to dialect is the same as interpolation from dialect to standard. From the interpolation point of view, a process like r-vocalization, for example, can be described as the realization of a vowel with subsequent vocalization of the trill, or the undoing of r-vocalization with the production of a trill. As can be seen in Figure 1, the number of generated frames differs between the interpolated versions, which is a result of the duration interpolation on the state level with 15 frames for interpolation rate 0.0 (RSAG) and 23 frames for 1.0 with

5ms frame length (IVG dialect). Concerning the differences in phone durations, it should be kept in mind that the state durations are generated from a contextually clustered statistical model, i.e. are an average over multiple phones in a similar context. In Figure 1 we can see the switching process with the raising of F2 from around 1000 Hz to above 2000 Hz. We use this type of visualization instead of a spectrogram since we can directly see the spectral envelope that results from the synthesis system before the waveform is generated, instead of the Discrete Fourier Transform spectrum of a windowed speech waveform over time.

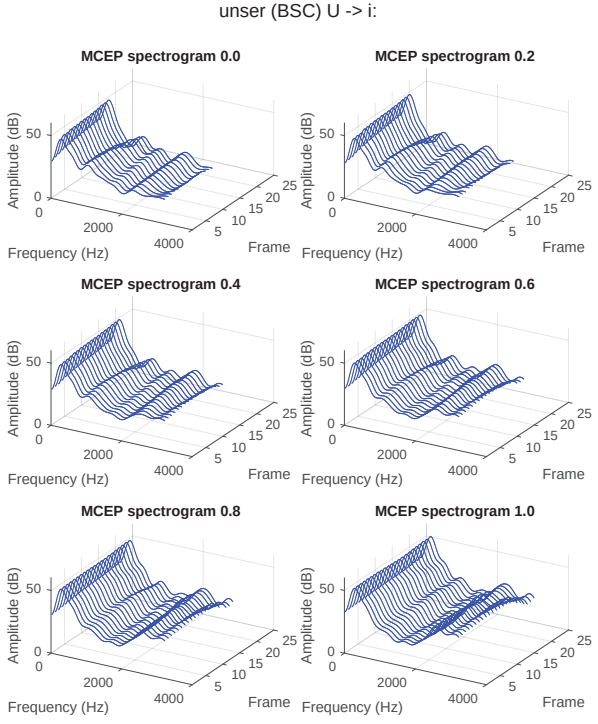

Figure 1: Interpolated MCEP spectrogram for RSAG to IVG interpolation of the input-switch-rule from /ʊ/ to / i/ in the word unser ‘our’.

PHONOLOGICAL PROCESS (SPEAKER A)

The following example ([vɔxə] ↔ [vɔʊxɛ]) was synthesized with the male Speaker A voice. Figure 2 shows again the spectra generated from the MCEP features over time for the RSAG to Innervillgraten dialect (IVG) interpolation from /ɔ/ to /ɔʊ/ in Woche. For this case, too, all five /ɔ/-states from RSAG are mapped onto the five /ɔʊ/-states from IVG. As can be seen in Figure 2, the number of generated frames differs between the interpolated versions for this case as well, with the dialect phone being longer than the standard one with 18 frames for interpolation rate 0.0 (RSAG) and 25 frames for 1.0 (IVG dialect). The reason for this difference lies in the diphthongization process from RSAG to IVG. In this phonological process, a diphthong has to be produced in the dialect. This is done by a stepwise raise of F2 in about the first third of the diphthong as shown in Figure 2.

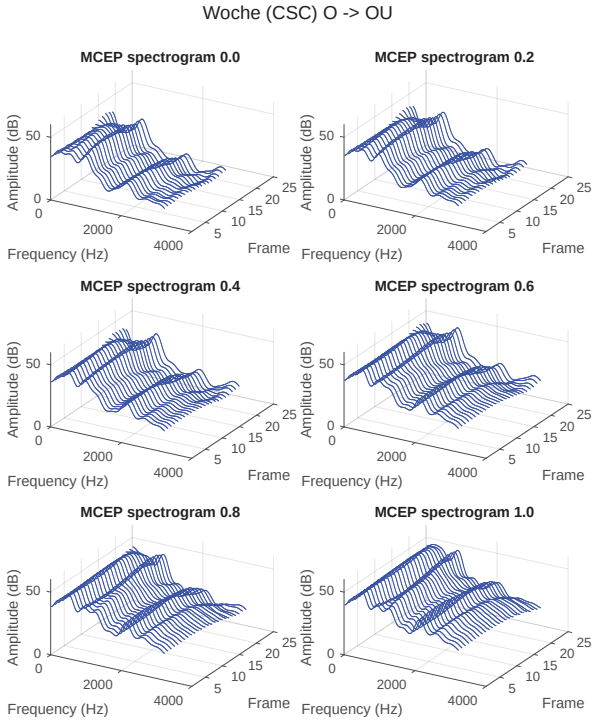

Figure 2: Interpolated MCEP spectrogram for RSAG to IVG interpolation of the phonological process /ɔ/ to /ɔʊ/ in the word Woche ‘week’

EXPLANATION OF SWITCHING AND PHONOLOGICAL PROCESS BEHAVIOR

The different behaviour for input-switch-rules and phonological processes can be explained by the behaviour of the interpolation of cepstral parameters. If we re-synthesize the spectrum from interpolated cepstral parameters as defined in Equation 1,

$$h(n) = IDFT\left(exp\left(DFT\left(\lambda c_1(n) + (1-\lambda)c_2(n)\right)\right)\right) \quad (1)$$

with λ being the interpolation control parameter and $c_1(n)$ and $c_2(n)$ being two cepstra, the spectra behave differently depending on the distance between the peaks. DFT and IDFT denote the Discrete Fourier Transform and the Inverse Discrete Fourier Transform respectively (Oppenheim and Schafer, 1999).

Figure 3 shows spectra μ generated from interpolated cepstral parameters. The first two rows show interpolations with spectra having a peak at

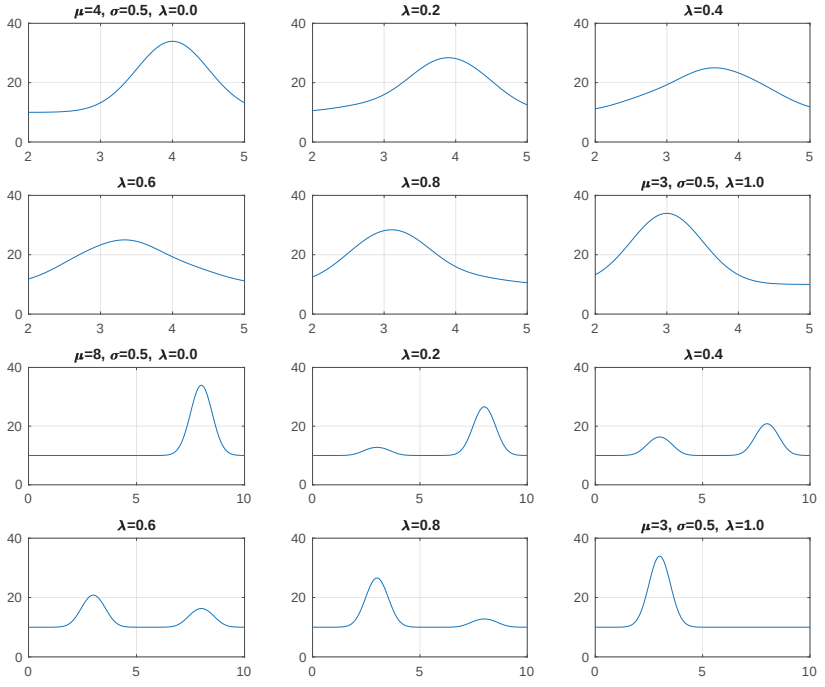

Figure 3: Re-synthesis of interpolated cepstrum parameters for spectrum with μ= 4 and μ= 3 (first and second rows) and μ= 8 and μ= 3 (third and fourth rows).

4 ($\lambda=0$) and at 3 ($\lambda=1$). In this case, the interpolation generates a gradual change from one peak to the other. Rows three and four show two spectra with peaks that are further apart. In this case the spectra generated from interpolated cepstral parameters generate the switching behaviour. While one peak is lowered, the other increases with the switching taking place between 0.4 and 0.6 where the height of the peaks change order. This behaviour can explain the switching and gradual transitions on the model interpolation, i.e. the speech production on the acoustic level. The gradual transition versus switching behaviour depends on the distance between the two peaks as well as on the variance of the peaks.

PHONETIC ANALYSIS OF INTERPOLATED SPEECH SAMPLES

In this section, we will present the phonetic analysis of two input-switch-rules and of two phonological processes from the RSAG input to the dialect output. The formant analysis in this section was based on formants extracted from the synthesized speech samples with the formant tracker from STx (Noll et al., 2007), which uses linear prediction coefficient (LPC)-based features. While the analysis in the previous section was based directly on the spectral envelope that was generated by the synthesizer, the following section uses formants extracted from the synthesis results.

Input-switch-rule 1

/ʊ/ ↔ /i/: In order to describe the interpolation steps of the input-switch-rule /ʊ/ ↔ /i/, we chose the word *unser* ‘our’ as an example. The input-switch-rule /ʊ/ ↔ /i/ involves a dramatic change in especially F2, which demands a raise of approximately 1000 Hz. F3 is raised for the vowel /i/ while for F1, no changes are expected.

The dramatic jumps in F2 are clearly visible in Figure 4. F2 of the female speakers C and D is low (below 1200 Hz) in steps 0.0, 0.2, and 0.4, and suddenly raised to > 2000 Hz in steps 0.6, 0.8, and 1.0. As already explained in Figure 3, the amplitude of one peak is lowered, while the amplitude of the other is raised in cases where peaks are far apart. This lowering and raising of the amplitude is very clearly visible in the interpolation of speakers A and B, but also holds for the female speakers, see Figure 4. For all speakers, two formant candidates are visible in step 0.4. In the first part of the vowel, the amplitude of the formant exceeding 2000 Hz is higher, whilst in the second part, the higher amplitude is visible in

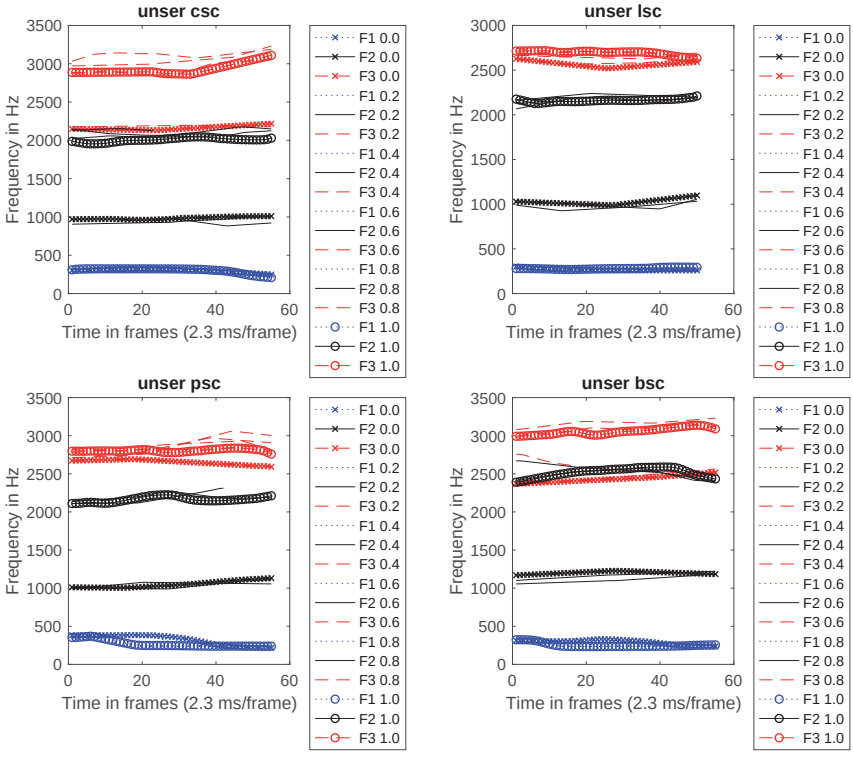

Figure 4: Formants F1-F3 for RSAG to IVG interpolation of the input-switch-rule from /ʊ/ to /i/ in the word unser ‘our’.

the formant below 1000 Hz. In both cases, the auditory impression of step 0.4 has more of an [i]-quality than a [ʊ]-quality.

During the interpolation of the samples of speakers A and C, a jump is also visible in F3. For speaker C this jump occurs from step 0.4 to 0.6; for speaker A from step 0.2 to 0.4. The mean values of F3 exceed 2800 Hz in the case of speaker A, and 3000 Hz in the case of C. This indicates a pre-palatal constriction location for the vowel [i], which is manifested by a high F3 approaching F4 (Moosmüller et al., 2015). The interpolation of the other two speakers, B and D, shows a continuous raising of F3, which is, however, below 3000 Hz for the female speaker D, and below 2700 Hz for the male speaker B. The auditory evaluation of the [i]-quality, which was performed by the authors as expert listeners, is more pronounced in the dialect sample (step 1.0) of A and C. The auditory quality was evaluated by the authors; no perception test was carried out.

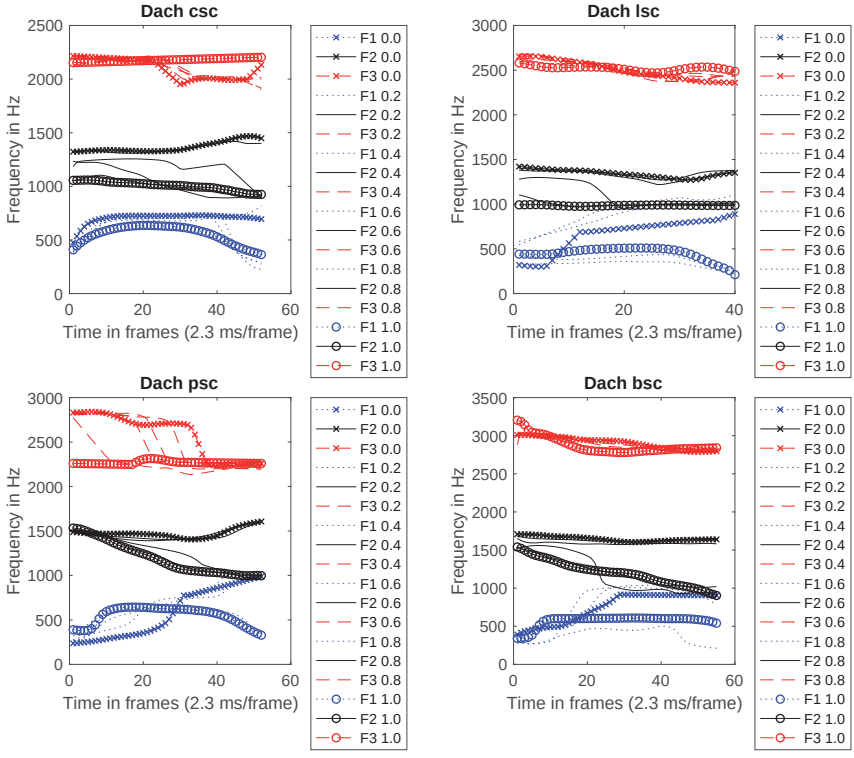

Figure 5: Formants F1-F3 for RSAG to IVG interpolation of the input-switch-rule from /ɑ/ ↔ /ɔ/ in the word Dach ‘roof”.

Input-switch-rule 2

/ɑ/ ↔ /ɔ/: As concerns the input-switch-rule /ɑ/ ↔ /ɔ/ shown in Figure 5, the differences are not as dramatic as in the previously discussed input-switch-rule /ʊ/ ↔ /i/. Nonetheless, the jumps are obvious as well. /ɑ/ ↔ /ɔ/ demands a lowering of both F1 and F2. The jump in the interpolation is most obvious in speaker B, whose F1 exceeds 600 Hz in steps 0.0, 0.2, and 0.4, while F1 is below 500 Hz in steps 0.6, 0.8, and 1.0. The same holds for F2: steps 0.0, 0.2, and 0.4 exceed 1300 Hz, whilst 0.8 and 1.0 are below 1000 Hz. Step 0.6 shows a rather diphthongal trajectory, starting with a high F2 which is lowered in the second half of the diphthong. In the samples of B, a jump is observed between steps 0.4 and 0.6. Contrary to B, the interpolation of speaker A produces only small changes in F1 and is responsible for the auditory impression of an /ɑ/-quality in all his samples. Similar to speaker B, F2 of speaker A’s steps 0.0 and steps

0.2 exceed 1300 Hz, while steps 0.6, 0.8., and 1.0 meet around 1000 Hz. Step 0.6 produces a diphthongal trajectory; the first part of the diphthong preserves a high F2, while in the second part, F2 is lowered.

The interpolation of both female speakers C and D produces the slight diphthongization for the dialectal output of the input-switch-rule /ɑ/ ↔ /ɔ/, described in Section III. For both speakers, a lowering of F1 is observable in steps 0.6, 0.8, and 1.0. This lowering comprises at least the final two-thirds of the trajectory for C, while in the samples of D, only the second half of the trajectory is affected. Nonetheless, the jump from step 0.4 to 0.6 is clearly visible for both speakers. In the trajectory of F2, the diphthongal quality is obvious for steps 0.6, 0.8 and 1.0 in the samples of both speakers, rendering [ɑɔ] as output, while steps 0.0, 0.2, and 0.4 exhibit the monophthongal quality of RSAG.

Statistical analysis of input-switch-rules

Figure 6 shows the differences in the first two formants for RSAG to IVG interpolation of the input-switch-rule /ʊ/ ↔ /i/ in the word unser 'our'. The left column shows differences in F1 for the two male speakers (A, B), the two female speakers (C, D) and for all the speakers taken together. The right column shows differences in F2.

Formant Differences (FDIFF) between interpolation steps were computed as

$$FDIFF = F_{\alpha_i} - F_{\alpha_{i+1}} \quad (2)$$

for F1 and F2 where α = (0.0,0.2,0.4,0.6,0.8,1.0) is the sequence of interpolation parameters and $F\alpha_i$ is the formant trajectory for interpolation step i. FDIFF is then a trajectory (sequence) of formant differences that will be positive if the formant is lowered from step i to step i + 1 and will be negative if the formant is raised. Figure 6 shows the boxplot for the different FDIFF trajectories with the median, the 25th and 75th percentiles and the whiskers extending to the extreme values. Outliers are shown as red crosses.

Again, a clear jump in F2 differences at F20.4 − F20.6 is visible for speakers A, B, C, and D; see Figure 6. For A and B, a larger variance in the differences is observable, but the median values are still very low. This shows that the formant trajectory of F2 is raised abruptly between 0.4 and 0.6. For all the speakers, we can also see this switching behaviour at F20.4 − F20.6. The differences between F20.4 − F20.6 and all other F2

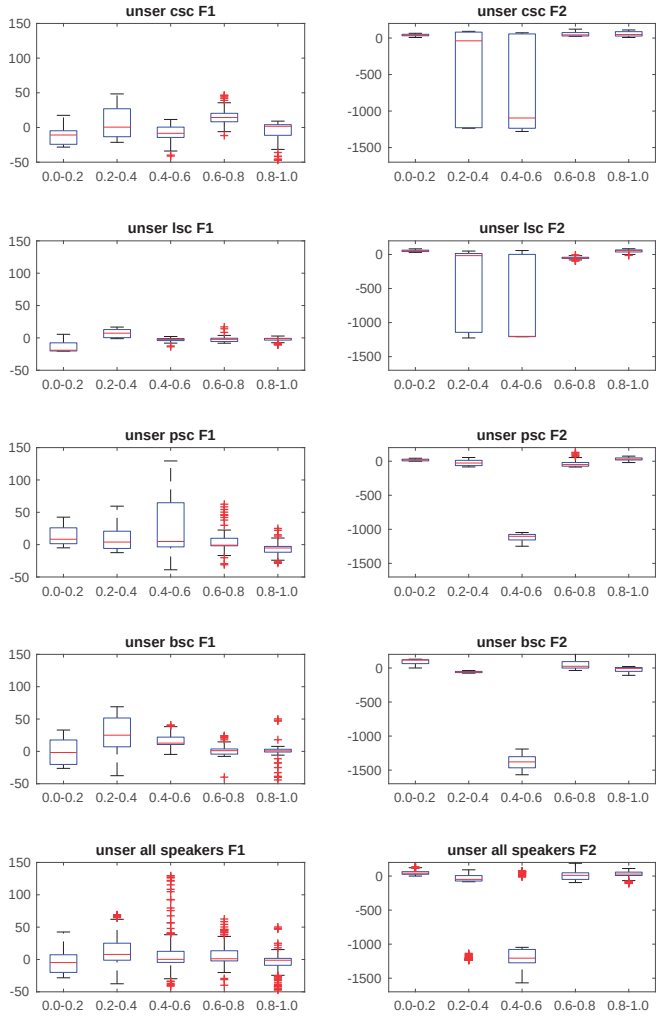

Figure 6: Differences in F1 and F2 for RSAG to IVG interpolation of the input-switch rule /ʊ/ ↔ /i/ in the word unser ‘our’.

differences are statistically significant ($p < 0.001$) according to a Wilcoxon rank sum test for equal medians, which clearly shows the input-switch from [ʊ] to [iː]. Since F1 is similar for both [ʊ] and [iː], differences in F1 do not show such abrupt changes but rather a more gradual transition behaviour for C and D and almost no changes for the male speakers A and B. Figure 7 shows the differences in the first two formants for RSAG to IVG interpolation of the input-switch-rule /ɑ/ ↔ /ɔ/ in the word Dach ‘roof’.

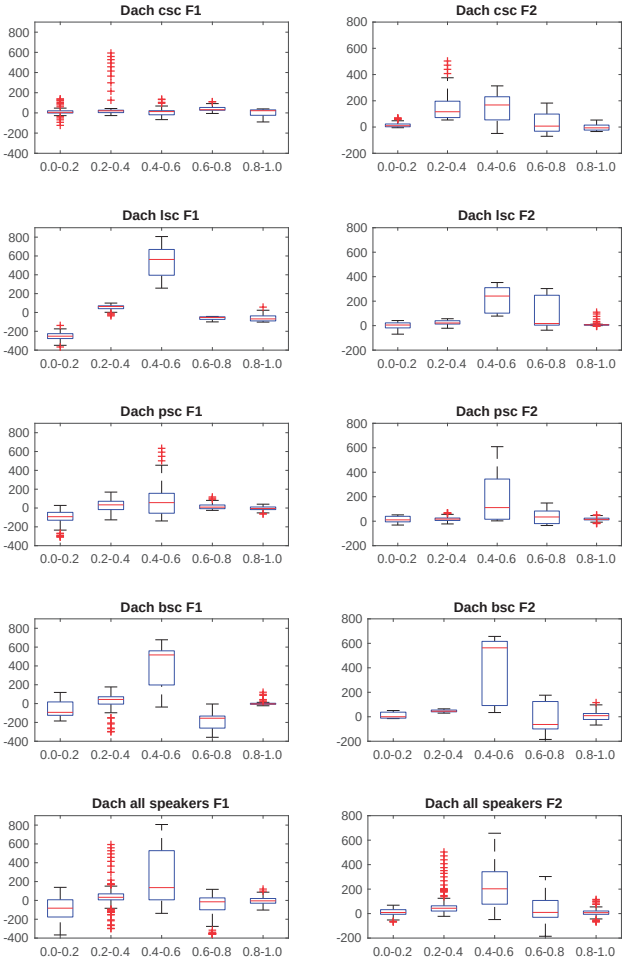

Figure 7: Differences in F1 and F2 for RSAG to IVG interpolation of the input-switch-rule /ɑ/ ↔ /ɔ/ in the word Dach ‘roof’.

A clear jump in F2 differences at F20.4 − F20.6 is visible for speakers B, C, and D; see Figure 7. This shows that the formant trajectory of F2 is lowered abruptly between 0.4 and 0.6. For all speakers we can also see this switching behaviour at F20.4 − F20.6. The differences between F20.4 − F20.6 and all other F2 differences are statistically significant ($p < 0.001$) according to a Wilcoxon rank sum test for equal medians, which clearly shows the input-switch-rule from [a] to [ɔ]. For F1 we can see a clear switch at F10.4 − F10.6 for B and C, which is slightly weaker but also present for A and D. Pooling all speakers together, the switch at F10.4 − F10.6 is significantly different ($p < 0.001$) from all other F1 differences.

Phonological process 1

[ɐ] → [ɔ]: This process affects unstressed sequences of <-er>. Preceding a bilabial▪consonant or a rounded vowel, the vowel resulting from r-vocalization changes to [ɔ]. In the same way as in the input-switch-rule /ɑ/ ↔ /ɔ/, F1 and F2 need to be lowered in order to produce the desired output [ɔ]. However, contrary to the input-switch-rule, which introduced pronounced jumps, the process shows a continuous change in formant frequencies, as becomes obvious from Figure 8 for all speakers. In the samples of B, C, and D, the output of step 0.0 is [ɐ] with a rather low F1. For this reason, F1 is continuously raised for the output of [ɔ]. On first sight, F1 of C and D seems to contain a jump, but a closer look clearly reveals intermediate formant traces in both cases. Statistical analysis reveals a significant difference only from F10.6 − F10.8; the differences between all other steps are not significant, see Section VI.F. The output of step 0.0 of A, on the other hand, is [ɑ]; therefore, F1 is continuously lowered to render [ɔ] in step 1.0. F2 shows a continuous lowering and F3 either shows no changes or a continuous lowering (C).

This example is of particular interest for our current study, since it vividly shows the difference between an input-switch-rule and a phonological process. Although the same phones are involved, we observe jumps in the case of the input-switch-rule and a continuous change in the case of the phonological process.

Phonological process 2

[ɔ] → [ɔʊ]: Again, as becomes apparent from Figure 9, a continuous change in especially F2 is visible for A, C, and D. During the interpolation of these speakers, substantial parts of F2 are continuously raised in order to arrive at the diphthongal output in 1.0. If at all, F3 is lowered (especially so in D). F1 experiences some changes too and is especially raised in the onset and lowered in the offset of the diphthong.

Statistical analysis of phonological processes

Figure 10 shows the differences in the first two formants for RSAG to IVG interpolation of the phonological process [ɐ] → [ɔ] in the word *unser* ‘our’. Figure 10 shows that there is a continuous change of F1 for speakers A, B, and C. Speaker C shows a small increase in F1 of around 200 Hz at F10.4 − F10.6. Concerning all speakers, F10.4 − F10.6 differs

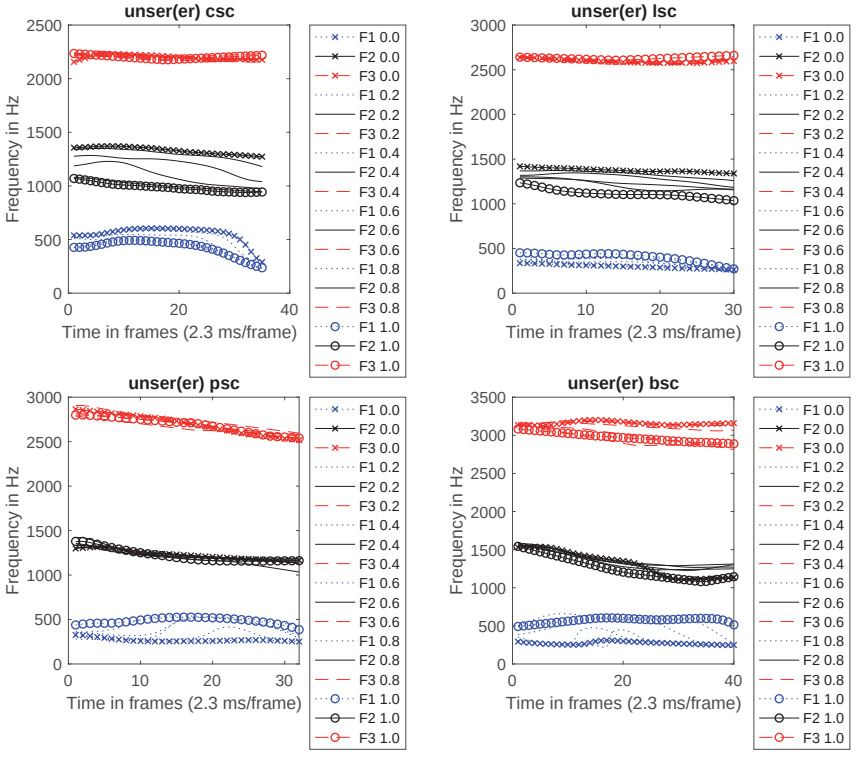

Figure 8: Formants F1-F3 for RSAG to IVG interpolation of the phonological process [ɐ] → [ɔ] in the word unser ‘our’.

significantly only from F10.6 − F10.8 ($p < 0.001$), but not from the other conditions. These differences are significant but much smaller than in the case of the input-switch-rules. In the same way, a continuous change in F2 is visible. In this case, F10.4 − F10.6 is not significantly different from any other condition for all speakers according to a Wilcoxon rank sum test for equal medians. These results for F1 and F2 show that a phonological process is involved with continuous changes from [ɐ] to [ɔ]. Figure 11 shows the differences in the first two formants for RSAG to IVG interpolation of the phonological process from /ɔ/ to /ɔʊ/ in the word Woche ‘week’. Figure 11 shows that there is a continuous change of F1 for speakers A, B, C, and D. Taking all the speakers together, F10.4 − F10.6 is significantly different from all other conditions ($p < 0.001$). These differences are, however, rather small, which is also obvious in the continuous change of F1. A similar picture emerges for F2, showing continuous changes with higher variance between all speakers. Here the condition F10.4 − F10.6 is only significantly different from the F10.2 − F10.4

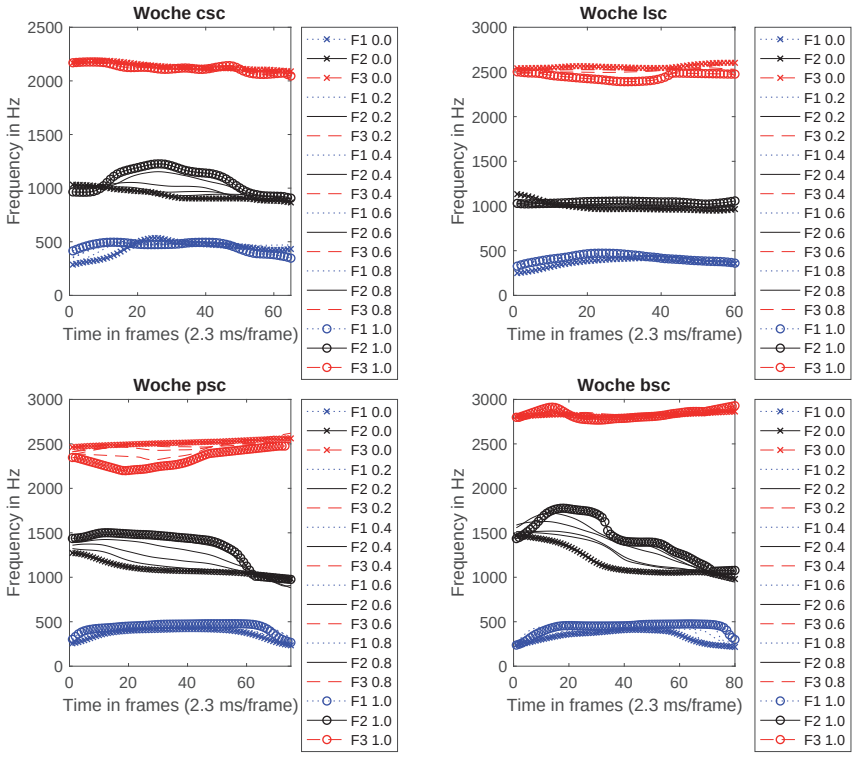

Figure 9: Formants F1-F3 for RSAG to IVG interpolation of the phonological process /ɔ/ to /ɔʊ/ in the word Woche 'week'

condition ($p < 0.001$). Overall, these results show the expected changes of a phonological process.

CONCLUSION

In this paper, we have shown how interpolation methods with state-of-the-art speech synthesis technology can be applied for the analysis of dialect variation. We analysed a representative[2] sample of dialect/standard interactions of four speakers for the Austrian dialect from Innervillgraten (IVG) and the transition to Regional Standard Austrian German (RSAG). The examples comprised input-switch-rules and phonological processes.

[2] While four speakers is certainly a small sample, our focus on two different processes (input-switch, phonological process) that appear in four words spoken by these speakers allowed us to draw a conclusion about the behaviour of the different types of processes.

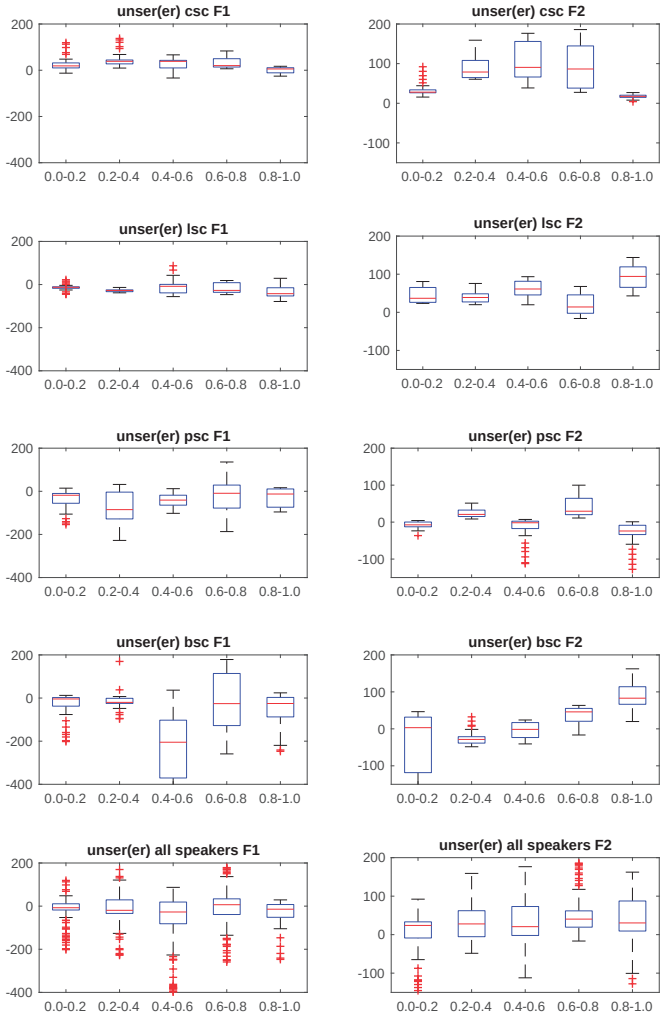

Figure 10: Differences in F1 and F2 for RSAG to IVG interpolation of the phonological process /ɐ/ to /ɔ/ in the word unser ‘our’.

The analysis was focused on the interaction between the spectral and formant level and the level of Mel-cepstral features that are used by the interpolation algorithm.

We showed that input-switch rules produce the expected non-linear behaviour at the spectral and formant level by using a linear interpolation at the Mel-cepstral feature level. A statistical analysis of formant changes within the interpolation steps shows a clear difference between input-

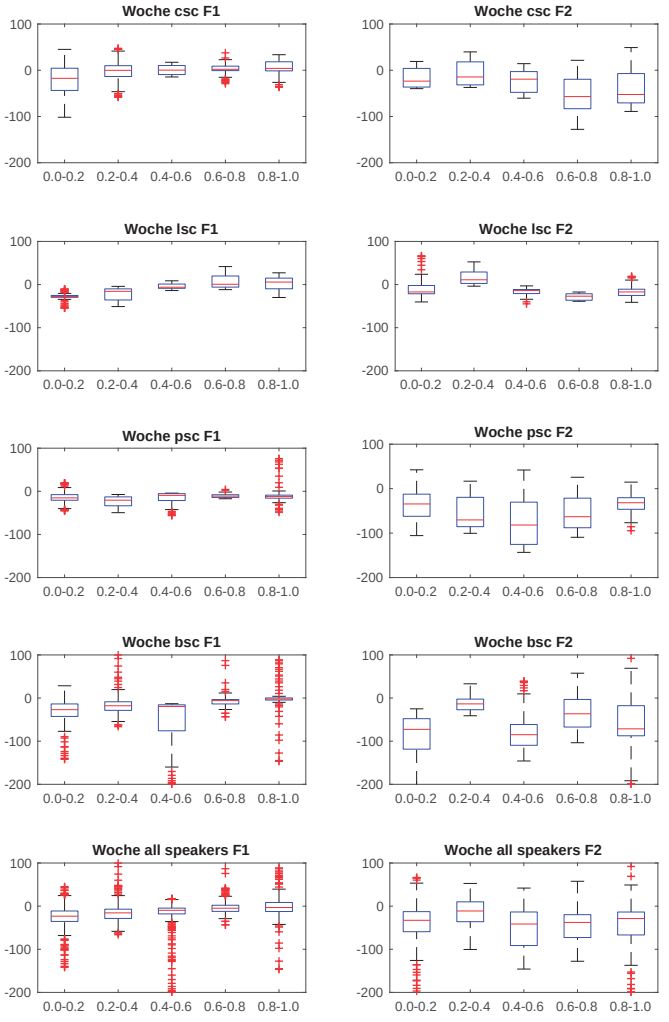

Figure 11: Differences in F1 and F2 for RSAG to IVG interpolation of the input-switch-rule /ɔ/ to /ɔʊ/ in the word Woche ‘week’.

switch rules and phonological processes. Thus, we could show by means of speech synthesis that there is a qualitative difference between input-switch-rules and phonological processes. While input-switch-rules have no intermediate steps, phonological processes are characterized precisely by the presence of intermediate steps. Consequently, in the first case, interpolation produces jumps, while in the second case, smooth transitions are generated. Therefore, we propose that the interaction between dialects

and standard varieties should be described by a two-competence model which captures the qualitative difference of phonological variables and thus provides a method for the analysis of language variation and change. Due to limited space, we provided only a few examples of the different behaviour of input-switch-rules and phonological processes in dialect levelling and sound change. For further examples, the reader is referred to Moosmüller (1991), Moosmüller and Scheutz (2013), or Soukup (2009). To generalize our results, we will extend our analysis to other Austrian dialects as well as non-German dialects in the future.

ABBREVIATIONS

AMTV	Acoustic modelling and transformation of varieties for speech synthesis
A,B	male speakers
C,D	female speakers
DFT	Discrete Fourier Transform
DiÖ	Deutsch in Österreich
F0	first formant
F2	second formant
F3	third formant
FDIFF	Formant Differences between interpolation steps
HMM	Hidden Markov Model
Hz	Hertz
IDFT	Inverse Discrete Fourier Transform
IVG	Innervillgraten
LPC	Linear Prediction Coefficient
MCEP	Mel-cepstral
RSAG	Regional Standard Austrian German
STx	Speech Tools eXtended

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF) project AMTV – Acoustic modelling and transformation of varieties for speech synthesis (P23821-N23) and project DiÖ – Deutsch in Österreich (I2539-G23).

REFERENCES

Astrinaki, M., Yamagishi, J., King, S., D'Alessandro, N., and Dutoit, T. (2013). 'Reactive accent interpolation through an interactive map application', in Proc. of the 14th Conference of the International Speech Communication Association (INTERSPEECH 2013), edited by F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, 1877–1878 (ISCA, Lyon, France).

Dressler, W., Moosmüller, S., and Wodak, R. (1989). 'Ricerche sociolinguistiche sullalingua urbana di Vienna – Sociolinguistic research on the urban language of Vienna', in Parlare in citta: studi di sociolinguistica urbana – Talk in the city: Urban sociolinguistic studies, edited by G. Klein, 93–110 (Congedo, Galatina).

Dressler, W. U. and Wodak, R. (1982). 'Sociophonological methods in the study of sociolinguistic variation in Viennese German', Language in Society 11, 339–370.

Hollenstein, J., Pucher, M., and Schabus, D. (2013). 'Visual control of hidden-semi-Markov-model-based acoustic speech synthesis', in Proc. of the 12th International Conference on Auditory-Visual Speech Processing (AVSP 2013), 31–35 (Annency, France).

Hornung, M. (1964). Mundartkunde Osttirols. Eine dialektgeographische Darstellung mit volkskundlichen Einblicken in die altbäuerliche Lebenswelt – Dialectology of Eastern Tyrol. A geographical account of the dialects with folkloristic insights into farm life (Böhlau, Wien), 182 pages.

Hornung, M. and Roitinger, F. (2000). Die österreichischen Mundarten. Eine Einführung – Austrian dialects. An introduction (öbv&hpt, Wien), 160 pages.

Imai, S. (1983). 'Cepstral analysis synthesis on the Mel frequency scale', in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-83), 93–96 (Boston, USA).

Isogai, J., Yamagishi, J., and Kobayashi, T. (2005). 'Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis', in Proc. of the 9th European Conference on Speech Communication and Technology (EUROSPEECH 2005), 2597–2600 (Lisbon, Portugal).

Karhila, R. and Wester, M. (2011). 'Rapid Adaptation of Foreign-Accented HMM-based speech synthesis', in Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), 2801–2804 (ISCA, Florence, Italy).

King, S., Tokuda, K., Zen, H., and Yamagishi, J. (2008). 'Unsupervised adaptation for HMM-based speech synthesis', in Proc. of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008),1869–1872 (Brisbane, Australia).

Lei, M., Yamagishi, J., Richmond, K., Ling, Z.-H., King, S., and Dai, L.-R. (2011). 'Formant-controlled HMM-based speech synthesis', in Proc. INTERSPEECH, 2777–2780 (Florence, Italy).

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2008). 'Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge', in Proc. INTERSPEECH, 573–576 (Brisbane, Australia).

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2009). 'Integrating articulatory features into HMM-based parametric speech synthesis', Trans. Audio, Speech, and Language Processing 17, 1171–1185.

Moosmüller, S. (1991). Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck – Standard and dialect in Austria – A sociophonological study on teasing apart standard and dialect in Vienna, Graz, Salzburg, and Innsbruck (Böhlau, Wien), 212 pages.

Moosmüller, S. and Scheutz, H. (2013). 'Chain shifts revisited: The case of Monophthongisation and E-confusion in the city dialects of Salzburg and Vienna', in Language variation – European Perspectives IV, edited by P. Auer, J. Caro Reina, and G. Kaufmann, 173–186 (Benjamins, Amsterdam).

Moosmüller, S., Schmid, C., and Brandstätter, J. (2015). 'Standard Austrian German', Journal of the International Phonetic Association 45, 339–348.

Noll, A., White, J., Balazs, P., and Deutsch, W. A. (2007). STX – Intelligent Sound Processing, Programmer's Reference, Acoustics Research Institute, Austrian Academy of Science, http://www.kfs.oeaw.ac.at. OEAW (2015).

Oppenheim, A. V., Schafer, R. W. (1999). Zeitdiskrete Signalverarbeitung, R. Oldenbourg Verlag, München Wien.

Picart, B., Drugman, T., and Dutoit, T. (2014). 'HMM-based speech synthesis with various degrees of articulation:▪A perceptual study', Neurocomputing 132, 142–147.

Pucher, M., Schabus, D., and Yamagishi, J. (2010a). 'Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners', in INTERSPEECH 2010, 2186–2189 (Makuhari, Japan).

Pucher, M., Schabus, D., Yamagishi, Y., Neubarth, F., and Strom, V. (2010b). 'Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis', Speech Communication 52, 164–179.

Qin, L., Ling, Z., Wu, Y., Zhang, B., and Wang, R. (2006). 'HMM-based emotional speech synthesis using average emotion model', in Proc. of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP-2006), 233–240 (KentRidge, Singapore).

Schatz, J. (1903). 'Die tirolische Mundart – The dialect of Tyrol', Zeitschrift des Ferdinandeums für Tirol und Vorarlberg 47(3), 1–94.

Scheutz, H. (2016). 'Deutsche Dialekte in Südtirol. Erste Ergebnisse eines Dialektatlas-Projektes – German dialects in South Tyrol. First results of a project on a dialect atlas', in Bayerisch-österreichische Varietäten zu Beginn des 21. Jahrhunderts – Dynamik, Struktur, Funktion – Bavarian-Austrian varieties at the beginning of the 21st century – Dynamics, structure, function, edited by A. Lenz, L. M. Breuer, P. Ernst, M. Glauninger, T. Kallenborn, and F. Patocka, 407–432 (Steiner, Stuttgart).

Soukup, B. (2009). Dialect use as interaction strategy. A sociolinguistic study of contextualization, speech perception, and language attitudes in Austria (Braumüller,Wien), 253 pages.

SPTK (2015). Speech Signal Processing Toolkit (SPTK) Version 3.9, http://sp-tk.sourceforge.net/.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). 'Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing', IEICE Transactions on Information and SystemsE88-D, 2484–2491.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (1998). 'Speaker adaptation for HMM-based speech synthesis system using MLLR', in the Third ESCA/COCOSDA Workshop on Speech Synthesis, 273–276 (NSW, Australia).

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). 'Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR', in Proc. of the Internation-

al Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP 2001), 805–808 (Salt Lake City, USA).

Toman, M., Pucher, M., Moosmüller, S., and Schabus, D. (2015). 'Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis', Speech Communication 72, 176–193.

Veaux, C., Yamagishi, J., and King, S. (2012). 'Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders', in Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012), 967–970 (Portland, USA).

Wester, M. and Karhila, R. (2011). 'Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation', in Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP2011), 5372–5375 (Prague, Czech Republic).

Wiesinger, P. (1995). Schreibung und Aussprache im älteren Frühneuhochdeutschen – Spelling and Pronunciation in Early High German (de Gruyter, Berlin), 265 pages.

Yamagishi, J. and Kobayashi, T. (2007). 'Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training', IEICE Transactions on Information and Systems E90-D, 533–543.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). 'Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm', IEEE Audio, Speech, & Language Processing 17, 66–83.

Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). 'MLLR adaptation for hidden semi-Markov model-based speech synthesis', in Proc. of the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004), 1213–1216 (Jeju Island, Korea).

Yamagishi, J., Ogata, K., Nakano, Y., Isogai, J., and Kobayashi, T. (2006). 'HSMM-based model adaptation algorithms for average-voice-based speech synthesis', in Proc. of the International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP 2006), 77–80 (Toulouse, France).

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., and Kitamura, T. (1997). 'Speaker interpolation in HMM-based speech synthesis system', in 5th European Conference on Speech Communication and Technology (EUROSPEECH 1997), 2523–2526 (Rhodes, Greece).

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007). 'Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005', IEICE Trans. Inf. & Syst. E90-D, 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). 'Hidden semi-Markov model-based speech synthesis', in Proc. of the 8th International Conference on Spoken Language Processing (INTERSPEECH 2004), 1393–1396 (Jeju Island, Korea).