1. Introduction

Dialect entails regional variation on various linguistic levels including phonetics, phonology, morphology, syntax, and the lexicon (Petyt 1980). It is an indexical property that is encoded in everyday language situations. Upon making a new acquaintance in German-speaking Switzerland, for instance, it is not unusual that the first topic of small talk involves dialects: 'Judging by your dialect, you come from Bern – right?'. The speech stream consists of segmental and prosodic features, both of which carry diagnostic information for dialect identification. If, for example, Hans pronounces the word 'milk' as [mˈɪʊɣ̊] instead of [mˈɪlɣ̊], he will likely be placed in Western Switzerland, given that the vocalization of /l/ is a typical feature of this area. Swiss German (SwG) listeners perform well at dialect identification: Leemann & Siebenhaar (2008) as well as Guntern (2011) have shown that naïve listeners can accurately identify a speaker's dialect much above chance. Identification performance varies by language: Williams et al. (1999) report successful dialect identification of varieties of English in Wales. For dialects of Dutch and dialects of British English, Van Bezoojien & Gooskens (1999) found identification rates much above chance level – the same is valid for dialects of American English, as shown by Clopper & Pisoni (2004). In recent research, Bent et al. (2016) conducted a large-scale study using regional dialects and non-native accents of English to examine how they are perceptually organized on the part of the listeners. In a free classification task, listeners exhibited sensitivity to distinguishing between the 24 native and non-native accents, providing insight into listeners' representations of varieties of English. Overall, these findings reveal that naïve listeners are aware of dialectal variation and can perform an identification task with different degrees of accuracy, depending on the language.

Recent research has revealed that naïve listeners can successfully identify dialects using prosodic features alone. Leemann & Siebenhaar (2008) report that naïve SwG listeners are able to identify dialects in delexicalized speech (low-pass filtered <250Hz) above chance level. Dialect identification based on prosodic features alone has also been documented for German (Schaeffler & Summer 1999, Gilles et al. 2001, Peters et al. 2002) as well as for American English dialects (Vicenik 2011, Vicenik & Sundara 2013). These studies used delexicalized speech with different types of signal manipulations. Vicenik & Sundara (2013), for example, who studied the role of temporal and f0 information in the identification of two varieties of English (American and Australian) created three conditions: low-pass filtered speech, *sasasa* speech (Ramus and Mehler 1999), i.e. rhythm only, and a resynthesis of the f0 contour onto a steady /a/ sound, i.e. intonation only. In a 2 AFC design, they found that American English listeners were able to identify the dialects in the first two conditions, but not in the intonation-only condition. Working on dialect identification of Dutch and British English, Van Bezooijen & Gooskens (1999) monotonized f0 in one condition – retaining segmental information – and applied a low-pass filter (350 Hz) – removing segmental information – in the other condition for eight dialects of Dutch and six dialects of British English. As a control, listeners judged unmanipulated speech. The authors report that dialect identification scores decreased the less segmental information there was in the signal. It is uncertain, however, whether a low-pass filter of 350 Hz effectively strips segmental information from the speech signal, when vowels such as /i/ and /u/ produced by males, for example, have first formants smaller than 350 Hz (cf. Peterson & Barney 1952). Further, low-pass filtered speech contains multiple layers of acoustic information (rhythm, intonation, and loudness) which makes the actual role of segments, rhythm, intonation, and intensity difficult to assess. Fuchs (2015) examined the cues listeners of Indian and British English use when asked to identify the two varieties. To disentangle prosodic and

segmental information, Fuchs created stimuli with multi-dimensional combinations of Indian and British English segmental and prosodic information (108 unique types of combinations consisting of different permutations of monotone f0, low-pass filtering, and swapping consonantal and vocalic interval durations). Preliminary results indicated that segmental cues were most diagnostic for the identification of the variety, followed by f0 and rhythmic information. The role of segmental and prosodic cues in dialect identification is ultimately one of saliency, i.e. the diagnostic accessibility of linguistic features (phonetic, morphological, syntactic, or lexical). Lenz (2010) conceives of saliency as the cognitive conspicuousness of a linguistic feature: a linguistic element stands out from a given context and is thus cognitively more quickly accessible than non-salient features. Guntern (2011) provides a qualitative approach to examining the saliency of different linguistic cues for identifying SwG dialects. While taking part in an 8 AFC dialect identification experiment, listeners were asked to write down features that they perceive as most salient for the individual SwG dialects. What was striking in the notes of the participants was that most of them mentioned segmental features as the most salient diagnostic cue to identifying a speaker's dialect, such as dialect-specific realizations of /r/ or the presence or absence of /l/-vocalization. As containing equally important diagnostic cues, the subjects mentioned dialect-specific lexical items. Prosodic features such as intonation, rhythm, or speaking rate were not considered as carrying much diagnostic power. By implication, Guntern's (2011) study suggests that listeners particularly need diagnostic segmental information in the sentence material to make judgments about the origins of a speaker, e.g. if there is material where /r/s and /l/s are lacking, it may be more difficult for a listener of Swiss German to identify the speaker's regional origin.

To examine the individual role of segments and prosody in the identification of a speaker's dialect even further, segmental and prosodic information can be disentangled, which can be achieved by swapping the two levels (cf. Vaissière & Boula de Mareüil 2004). This so-called prosody transplantation or prosody morphing paradigm has gained much attention in second language research: a number of studies have attempted to show which features – segmental or prosodic – are more important for accentedness and intelligibility in second language speakers (Derwing & Munro 2005, Vieru-Dimulescu & Boula de Mareüil 2005, Boula de Mareüil & Vieru-Dimulescu 2006, Holm 2008, Winters & O'Brien 2013, Ulbrich 2013, Ulbrich & Mennen 2015). Typically, prosodic features of one language variety are morphed onto the segments of another variety and vice versa, involving a form of intelligibility, accentedness, or accent rating task conducted by native listeners.

In the present study, we employ this methodology on a dialect identification task through a set of perception experiments where we separate segmental and prosodic features from one another and play them off against each other: we manipulate the speech signal in a way that prosodic features of dialect X are morphed onto the segments of dialect Y and vice versa. These manipulated stimuli are then played to naïve listeners of dialect Z (familiar with both dialects X and Y) who are then asked to indicate whether the stimulus heard is from a speaker of dialect X or dialect Y. We will do so by pursuing the following specific research questions:

RQ1: What is the role of segments and prosody (rhythm alone and rhythm combined with intonation) in the identification of a speaker's dialect?
RQ2: How is dialect identification contingent on the sentence material used?

These research questions will be studied in the context of the above-mentioned prosody morphing paradigm: to answer RQ1 we will use material from two SwG dialects: in the first condition, listeners judge unmorphed speech, in the second condition different listeners judged rhythm morphed speech, and in the third condition, different listeners judged speech that was morphed in rhythm combined with intonation. To answer RQ2, we will look at sentence material individually and study how identification performance varies as a function of the different segmental and prosodic make-up of the sentences.

Given the literature survey presented above, for RQ1 we predict that segmental features will carry more diagnostic weight than prosodic features. While this may be the case in this study, this expectation – ultimately – will depend on which dialects are studied and, more specifically, on the (dis)similarities in the segmental and prosodic domain. It is likely that there are dialects that are very different in the prosodic domain and possibly more similar in the segmental domain; and there are dialects where this relationship is diametrically opposite. The predictions we make for RQ1 are thus only valid for the dialects examined in the current study. As for RQ2, we expect listeners' dialect identification performance to vary substantially depending on the diagnostic cues that are available to them in the sentence material. A first pilot study conducted on unmorphed and rhythm morphed speech only – using a subset of the listeners of this study – revealed that segmental information appeared to dominate dialect identification (Leemann, Kolly, Nolan 2016). In the present study, we expanded the listener set, included a third experiment (i.e. a third condition in which we swapped rhythm and intonation combined), and tested for the effect of sentence material to explain the trends observed. In the discussion section of this contribution we will argue how the results of this study may have repercussions for speaker identification by victims and witnesses in criminal contexts, for automatic speech recognition and for cognitive bases for storing and accessing indexical information such as a speaker's dialect.

2. Methods

For this study we conducted a dialect identification experiment with three conditions using material from two SwG dialects (Bern and Valais SwG): in one condition, 21 listeners (Zurich SwG) judged unmorphed speech (Bern and VS SwG), in the second condition 20 different listeners judged rhythm morphed speech (Bern SwG segments with Valais SwG rhythm and vice versa), and in the third condition, 21 different listeners judged speech that was morphed in rhythm combined with intonation (i.e. Bern segments with Valais rhythm combined with intonation and vice versa). We preferred a between-subject design over a within-subject design as the latter would require different sentence material for each of the three conditions. This is disfavored because a difference in speech material would mean a difference in diagnostic segmental and prosodic cues that listeners may use in the dialect identification process.

2.1 Dialects

We chose Bern SwG and Valais SwG for the purposes of this study. Both dialects have been shown to be substantially different in segmental and prosodic make-up. Some of the most salient segmental differences are (a) /l/-vocalization in some phonological contexts in BE SwG, (b) /nd̥/-velarization in BE SwG, (c) hiatus diphthongization in BE SwG, (d) full final vowel realizations in word-final position in VS SwG, (e) the application of *Staub's Law* in most VS SwG dialects (i.e. the deletion of nasals before homorganic fricatives or affricates, resulting in the lengthening or diphthongization of the preceding stem vowel; cf. Werlen 1977), (f) 'palatalized' /s/ in VS

SwG, and (g) vocalic differences such as rounding of low-back vowels in VS SwG. See examples below:

(a) Vocalized /l/:	'milk'	BE SwG [mɪɻ̊ɣ̊]	VS SwG [milɣ̊]
(b) Velarized /nd̰/:	'child'	BE SwG [ɣ̊ɪŋ]	VS SwG [ɣ̊ind̰]
(c) Hiatus diphthongization:	'to snow'	BE SwG [ʃnˈei.ə]	VS SwG [ʃnˈiː.ə]
(d) Full final vowels:	'cut' (p.p.)	BE SwG [kʃnˈɪtə]	VS SwG [kʃnˈitʊ]
(e) Staub's Law:	'to drink'	BE SwG [trˈɪŋkxə]	VS SwG [trˈiːxɐ]
(f) Palatalized /s/:	'she'	BE SwG [sɪ]	VS SwG [ʃi]
(g) Vocalic differences:	'street'	BE SwG [ʃtrɑːsː]	VS SwG [ʃtrɒːsː]

In terms of prosodic differences, Leemann et al. (2012) report more syllable-timed rhythm in VS SwG, which shows less vocalic interval variability (*VarcoV*) than BE SwG. Alpine dialects, which includes VS SwG, have a tendency of realizing word-final syllables in non-reduced forms, i.e. they retain full vowels in otherwise unstressed environments; a relic from Old German. Using a corpus of spontaneous speech consisting of ten speakers per dialect, Leemann (2012) further reports substantial differences in articulation rate, with the Bernese articulating nearly 1 syllable / sec. slower than the Valais speakers (BE SwG 4.98 syll/sec; VS SwG 5.8 syll/sec, excluding filled pauses). In terms of intonation, Leemann (2012) – who applied Fujisaki's intonation model (Fujisaki & Hirose 1984) on the same corpus – found rises in prenuclear position in declarative statements for both dialects (contrary to Northern German varieties' H*+L accents in prenuclear and nuclear position, cf. Fitzpatrick-Cole 1999). The two dialects behaved differently in the phonetic realization of the rises, however: contrary to Bern Swiss German, Valais speakers were reported to show a displacement of a trailing tone to the right, reaching f0 maxima in subsequent unstressed syllables (cf. Grabe et al. 2001, Leemann 2012). It has been shown previously that associations of tones with the segmental string can undergo restructuring and can thus cause modifications of the location of the tones in the time domain (Gussenhoven 1990).

2.2 Speakers

Twelve speakers provided the sentence material for this study (6 VS SwG speakers (3f/3m); 6 BE SwG speakers (3f/3m)). The speakers, aged between 18 and 34, claimed to speak the respective dialects on a daily basis. The BE SwG speakers said to speak the dialect of Bern city. Speakers from the Valais dialect came from different localities in the canton – most of them from the city of Brig, others from Zermatt or adjacent valleys. Because of this, it is likely that intra-dialect variability is greater for the Valais SwG speakers. The speakers were screened for voice quality before the recordings: it was imperative that they had sonorant modal voices with little to no laryngealization, as recordings with laryngealization would have been highly problematic for the f0 moprhing algorithm described in 2.3. Speakers were given the sentence material a few days before the recordings; they had ample time for sentence reading rehearsals. Speakers were recorded in a sound-treated booth using an omnidirectional Earthworks QTC40 high definition condenser microphone (sampling rate of 44.1kHz; 16-bit quantization).

2.3 Materials

Given that we do not know the degree of salience for the features outlined in 2.1 – we do not know, for example, whether /l/-vocalization is a more salient feature for BE SwG than /nd̥/ velarization – we tried to bypass this issue altogether: to circumvent the question of how to create sentences that balance salient dialect-specific features, we randomly selected ten from the 336 sentences of the Bamford-Kowal-Bench (BKB) corpus (Bench et al. 1979), see Appendix. All of the salient features except for (e) (cf. 2.1.) are present in the material. Sentences were manually labeled in Praat (Boersma and Weenink 2016). We performed a syllable-based segmentation on our sentence material. Syllabification in Swiss German does not strictly proceed according to morpheme boundaries (Siebenhaar 2014); rather, it follows the onset-maximization principle, where syllable onsets are maximized before syllabification proceeds to the coda and nucleus of the preceding syllable.

We ran acoustic analyses to test for qualitative between-dialect differences in the speech material used. To examine these differences in the temporal domain – speech rhythm and articulation rate – we applied various rhythm metrics commonly used in the field. The acoustic parameter that is most frequently associated with speech rhythm is duration, although studies have shown that changes in f0, for example, can influence the percept of speech rhythm (Barry et al. 2009). A number of indices have been developed to capture variation in duration; these have often been referred to as 'rhythm metrics'. Rhythm metrics standardly use differently defined intervals as a basis – some of them use consonantal and vocalic variability, others suggest syllables as the basis of speech rhythm (Loukina et al. 2011). Throughout the rest of this paper, when we mention speech rhythm, we are referring to the syllabic duration information captured by various rhythm metrics. For the analysis of intonation, we performed an autosegmental metrical-based analysis (cf. Ladd 2008) for all the 120 sentences (two dialects, six speakers per dialect, ten sentences per speaker). The intonation transcription system used is based on IViE (Grabe et al. 2001) and – at this stage – is largely provisional, given the lack of systematic work on Swiss German intonation using the autosegmental-metrical framework.

The intonation analysis was conducted by the first and third author via auditory inspection and visual inspection (pitch tracking). We began by deciding on the location of rhythmically strong syllables in each sentence. Typically, rhythmically strong syllables occurred where the dialects' dictionary entries would mark lexically stressed syllables; – yet, the determination of strong syllables in speech perception has been debated and can be less than straightforward (cf. Kochanski et al. 2005). In cases where we were not fully certain if a syllable was rhythmically strong, we examined (auditorily and visually) duration, loudness, and pitch cues on or in the vicinity of the respective syllables and then made a decision. We subsequently examined, by ear and eye, the pitch movement surrounding these prominent syllables, with a view to attributing the shape to categorically distinct intonation patterns. This task is not trivial, given that (i) there is no previous systematic research on the intonational phonology of Swiss German that we could have referred to for a definitive inventory of phonologically contrastive pitch accents and (ii), more generally, there is a lack of accepted tests for phonological category membership (cf. Grabe et al. 2001). Given the first and third authors' familiarity with IViE (ibid.), we used IViE-based labels for transcription purposes. Although the IViE inventory is based on existing analyses of English, it is rich enough to capture most potential contrasts in languages with strong stress such as Swiss German, yet sufficiently constrained by its restriction to left-headed pitch accents that it removes some potential ambiguities of analysis. The contour was analyzed, with both auditory impressions and the f0 taken into account, but priority given to

auditory impressions, using the following IViE categories: (a) offset boundary tones 0%, L%, H%, which capture pitch targets at the IP-final boundary, or in the case of 0% the absence of such a targets; (b) onset IP boundary tones %L, %H, which capture non-default pitch targets on unstressed syllables before the first pitch accent, i.e. where the speaker starts high or low rather than in the middle of their normal pitch range; (c) prenuclear pitch accents – H*, H*+L, L*, L*+H, i.e. pitch accents which occur before the nucleus; and (d) nuclear pitch accents – H*, H*+L, L*, L*+H, i.e. the last pitch accent of the phrase. We used the symbol '!' to mark a downstepped accent, i.e. a compression in the pitch range which lowers the f0 targets for a H tone relative to a previous H. We further used '_' to mark the displacement of trailing tones to the right, in instances where the f0 target is not reached immediately after the accented syllable but as far to the right as is compatible with the subsequent pitch target. We did not, to any substantial degree, encounter problems such as laryngealized voice quality – which may result in gaps or errors of pitch measurements – in our data, as we selected speakers and sentences with favorable modal voice quality (see 2.4(3)). Neither did we encounter substantial between-speaker pitch variation due to different emotional involvement on the part of the speakers. To assess rhythm differences between the dialects, we applied measures based on interval variability of syllable durations (Lai et al. 2013) and based on voiced and unvoiced interval durations (Dellwo et al. 2007).

Results revealed between-dialect differences for *VarcoSyl* – the rate-normalized standard deviation of syllable interval durations: VS SwG speakers (M=.42, SD=.11) showed less syllable duration interval variability than BE SwG speakers (M=.48, SD=.12). Results further showed a trend that the two dialects differ in articulation rate – as captured by the number of syllables per second. The difference is minor, however: BE SwG speakers (M=4.3, SD=.75) articulated more slowly than VS speakers (M=4.7, SD=.78) – albeit not statistically significant. In terms of intonation, the two dialects differ in the proportions of %H and %L boundary tones. VS SwG prefers %L boundaries (63%), BE SwG %H boundaries (90%). In prenuclear position, both dialects feature L*+H: BE SwG 85%; VS SwG 44%. The rise in BE is often completed within the accented syllable, labelled L*+H, whereas VS features an additional rise not generally found in the BE inventory, one where the peak is delayed beyond the accented syllable, labelled L*+_H. This pitch accent itself makes up 30% of the VS prenuclear accent inventory (cf. Leemann 2012). L*+H and L*+_H combined, make up 74% of the VS SwG prenuclear accent inventory. For BE, this results in a glide up pattern, while the VS SwG pattern appears to be more step-up than gliding. Table 1 shows the distributions of pitch accent types for the two dialects.

		!H*	!H*+L	%H	%L	0%	H*	H*+L	L*	L*+H
	onset boundary	0	0	90.48	9.52	0	0	0	0	0
	offset boundary	0	0	0	0	100	0	0	0	0
	prenuclear	5.56	0	0	0	0	3.7	5.56	0	85.19
	nuclear	5	16.67	0	0	0	0	0	66.67	11.67

	!H*	!H*+L	%H	%L	0%	H*	H*+L	L*	L*+_H	L*+H
onset boundary	0	0	36.84	63.16	0	0	0	0	0	0
offset boundary	0	0	0	0	100	0	0	0	0	0
prenuclear	10.19	0	0	0	0	12.96	3.7	0	29.63	43.52
nuclear	8.33	53.33	0	0	0	0	0	30	0	8.33

Table 1: Distribution of pitch accent types (in %) for the two dialects: Bern Swiss German (top panel) and Valais Swiss German (bottom panel).

Figure 1 juxtaposes BE SwG L*+H with the VS SwG-typical L*+_H pitch accent. The phrase reads *Das Kind nimmt ein Spielzeug in die Hand*, 'The child is taking a toy in its hand' – both speakers have pitch accents on *Kind*, *Spiel-*, and *Hand*.

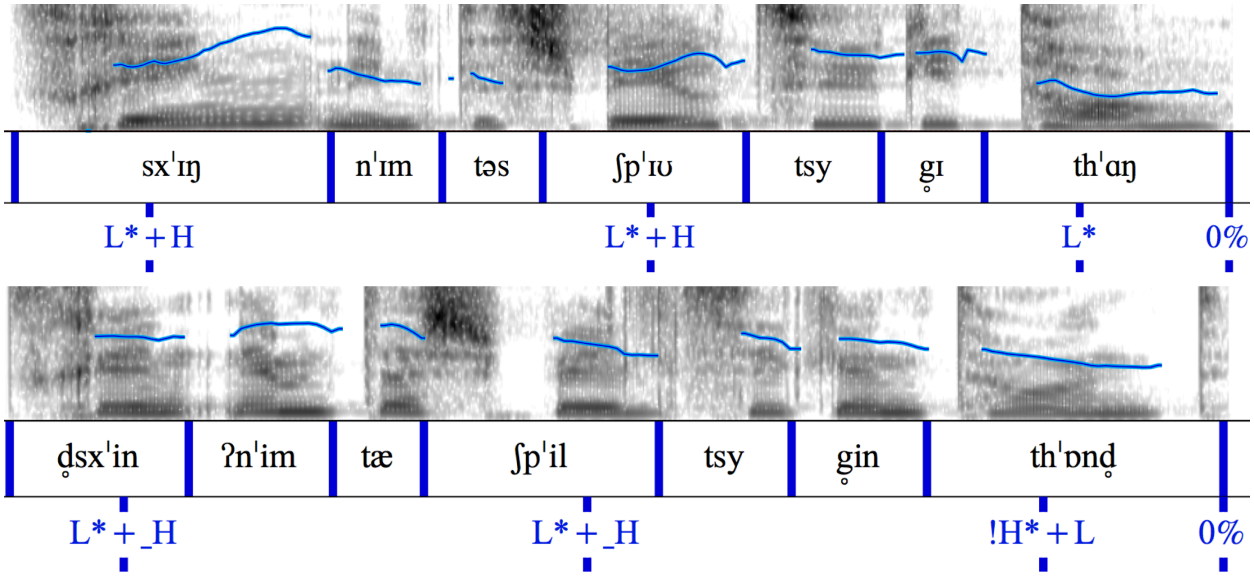

Figure 1: BE and VS SwG-typical prenuclear pitch accents: L*+H and L*+_H, where '_' indicates a delayed peak.

In prenuclear accents, the dialects further differ in the relative proportions of H*, with VS SwG revealing a higher proportion (13%) than BE SwG (4%). In terms of nuclear accents, VS SwG stands out with more !H*+L falls (53%) while BE features proportionately more L* nuclei (67%).

2.4 Signal processing

To create the stimuli for the three conditions, we proceeded as follows:

(1) Condition I: unmorphed speech

The first step in the stimulus creation process was to normalize all recordings to 68 dB. Following this normalization in intensity we further normalized the male speakers' f0s and the female speakers' f0s. This would prevent listeners from identifying the dialect based on speaker-specific f0: if, hypothetically, one of the male speakers had a high-pitched voice in all the ten sentences, it may have been possible that listeners rate the dialect in the stimulus based on the speaker's f0 rather than rating the dialect in question. We therefore adjusted the medians to 127.5 Hz for males and 216 Hz for females (these values represent the grand f0 means for male and female speakers in this corpus). We used *Praat*'s (Boersma & Weenink 2016) 'change gender' function to adjust f0 medians, which leaves f0 variability intact.

(2) Condition II: rhythm morphed speech

Using the stimuli created in *Condition I*, we chose to transplant syllable durations rather than segment durations for two reasons (see also Winters & O'Brien 2013): (a) speech rhythm, at least in the dialects under scrutiny, appears to operate on the syllable rather than the segmental level (cf. Barry et al. 2003); (b) to morph rhythm, the sentences of both dialects needed to have the identical syllable or segment count. Chances are greater that syllable counts are the same between the dialects, given that dialect-specific phonological processes can easily change the number and type of segments in a sentence (e.g. /nd̥/ velarization in *Kind*, 'child', BE SwG [ɣɪŋ], VS SwG [ɣɪnd̥]). It is likely, of course, that segment durations do not match between the dialects. Leemann & Siebenhaar (2010) have shown, for example, that BE SwG exhibits more pronounced phrase-final lengthening (which essentially amounts to stretching the final vowel of the final syllable) than VS SwG. Such segmental level effects and how they may affect within the prosody transplantation paradigm would be worth exploring in future studies.

Material was morphed using a script from Boula de Mareüil & Vieru-Dimulescu (2006) which was adapted for the present purposes. The script selects a rectangular window to transfer durations syllable-by-syllable, either stretching or compressing the syllable. Finally, it creates a synthetic version of the signal using PSOLA. Speakers were paired in such a way that they were maximally similar in articulation rate, which would counteract potential artefacts in the morphed stimuli – the less stretching between the source and the recipient signal, the more natural sounding the stimulus (Quené and van Delft 2010). We calculated global articulation rates (i.e. the global mean of ten sentences) for each speaker. The fastest VS SwG female was paired with the fastest BE SwG female (5.3 sylls/sec and 4.7 sylls/sec); the slowest VS SwG female was paired with the slowest BE SwG female (4.3 sylls/sec and 3.7 sylls/sec) – the two medium fast female speakers were then paired (VS SwG 4.4 sylls/sec and BE SwG 4.2 sylls/sec). The same procedure was applied to the male subjects: fastest speakers (VS SwG 5.2 sylls/sec and BE SwG 4.9 sylls/sec); medium fast speakers (VS SwG 4.7 sylls/sec and BE SwG 4.6 sylls/sec) slowest speakers (VS SwG 4.2 sylls/sec and BE SwG 4 sylls/sec). For rhythm morphing, the syllable durations of sentence 01 of the fastest BE speaker, for example, were morphed onto the syllable durations of sentence 01 of the fastest VS speaker. Analogously, the syllable durations of sentence 01 of the fastest VS speaker were morphed onto the syllable durations of sentence 01 of fastest BE speaker. Figure 2 illustrates the process of rhythm swapping.

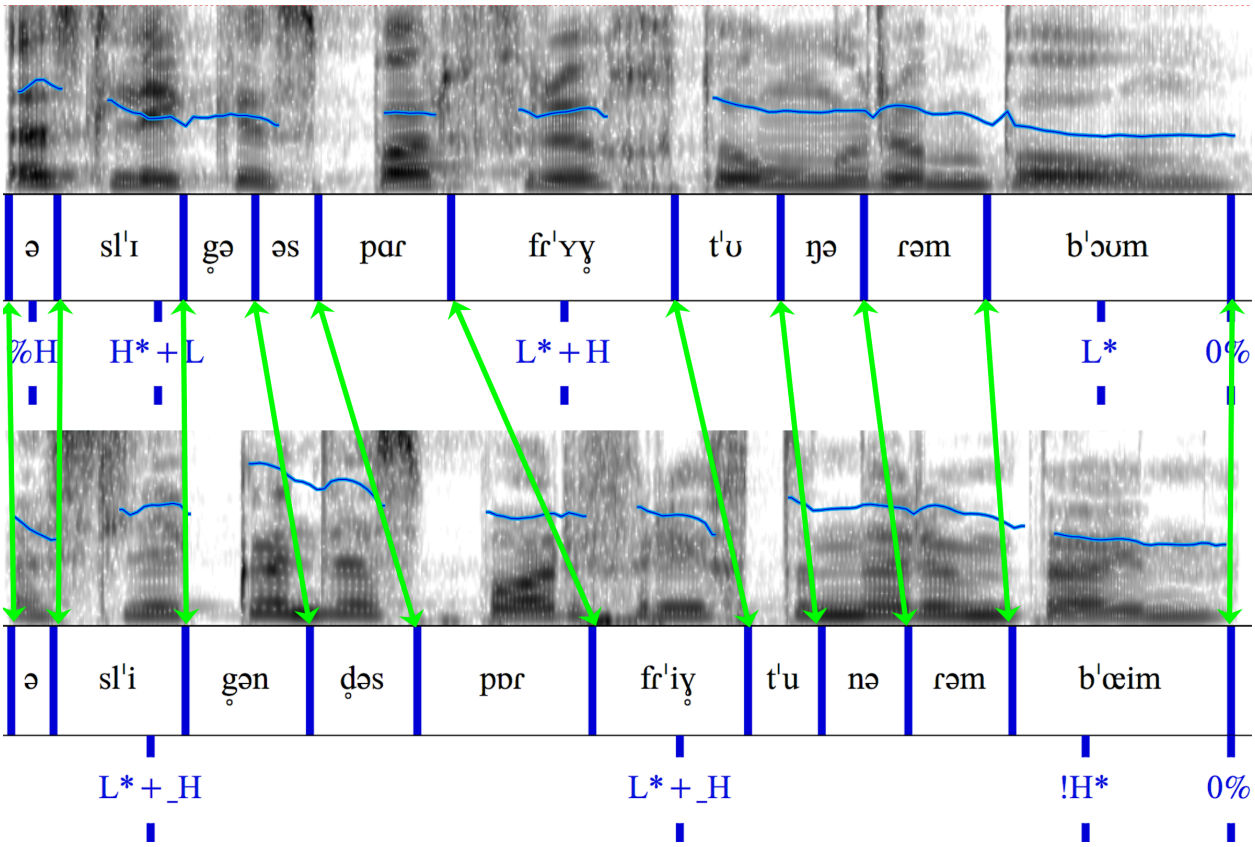

Figure 2: Rhythm swapping of a BE SwG (top) and a VS SwG sentence (bottom).

The top panel of Figure 2 shows a sentence from a BE speaker, segmented on the syllable level; the bottom panel shows the same sentence for a VS SwG speaker. The sentence reads *Es liegen ein paar Früchte unter dem Baum*, 'Some fruit is lying under the tree'. Both sentences feature the same syllable counts. In the example shown here, BE SwG [frˈʏɣ̯] (320 ms) is compressed in the morphing process as the VS SwG [frˈiɣ̯] is much shorter (232 ms); conversely, the VS SwG variant would be stretched. Note that in some instances – such as shown in Figure 1 – syllable structures differed between the dialects: *liegen* ('to lie') is realized as [lˈɪgə̯], i.e. CV.CV in BE SwG, but as [lˈigə̯nd̥], CV.CVCC in VS SwG. As a result, morphing VS SwG [gən] onto BE SwG [gə̯] causes a stretching of the syllable (a compression vice versa), which is due to differences in syllable structure. A check of the material revealed that in eight syllables of the total of 77 syllables of the sentence material (i.e. all syllables of the ten sentences counted) there are differences in syllable structure between BE and VS SwG, i.e. c. 10%.

(3) Condition III: rhythm and intonation morphed speech

For intonation morphing, we used stimuli of *Condition II* as a basis. Intonation cannot be morphed on its own because (a) there may be a potential mismatch in the length of the source and recipient sentence and, more importantly, (b) the temporal synchronization of the f0 contour with the segmental string, in particular with stressed syllables, plays a critical role for the type of pitch accents observed. Our approach, as far as possible, respects the character of intonational events by matching the pitch fluctuations. If we had replaced the f0 curve and aligned the start and end by normalizing only for overall duration, one would obtain intonation events of a

changed character (e.g. what was a peak aligned with a stressed vowel nucleus may now be lagging into a coda consonant or even a following unstressed syllable). An even more sophisticated way of transplanting intonation would be to morph the f0 contour on a syllable-by-syllable basis (in the same way how the stimuli for Condition II were created). This would avoid misaligning the tones and syllables altogether. In our approach, where we morph f0 after rhythm has been morphed, only the larger scale f0 fluctuations – the overall intonation pattern – survives the manipulation process.

The PSOLA-based f0 transplantation algorithm of Boula de Mareüil & Vieru-Dimulescu (2006) morphed voiced stretches from the source signal to voiced stretches in the recipient signal. Morphing can create artefacts in the f0 contour, which is why the f0 contour was manually checked and edited in *Praat* after the manipulation procedure. Because there can be gaps in the f0 contour due to laryngealized voice quality in either the source or the recipient sentence for example, we selected speakers who featured as little laryngealized voice as possible, see 2.2. Figure 3 illustrates the process of rhythm and intonation swapping on the sentence shown in Figure 2.

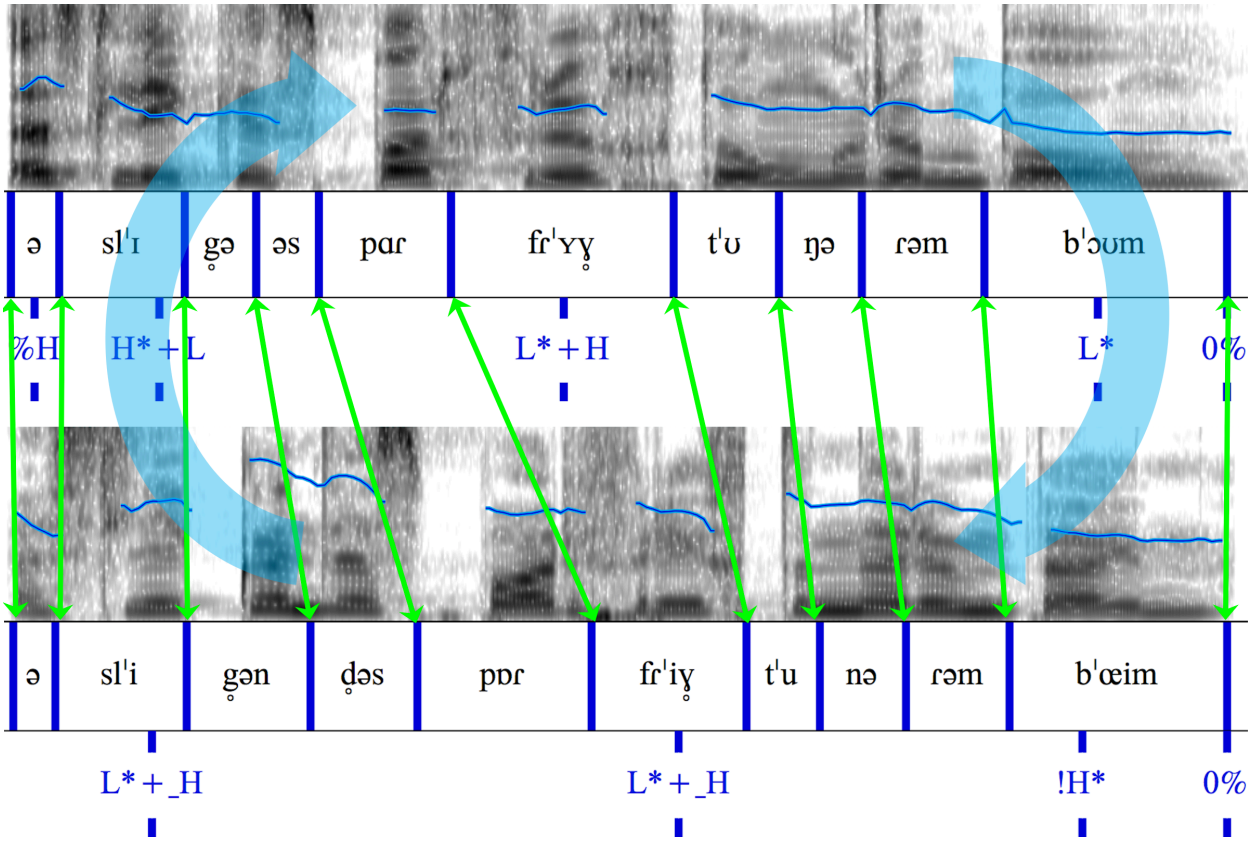

Figure 3: Rhythm and intonation swapping of a BE SwG (top) and a VS SwG sentence (bottom).

Figure 4 shows the sentence with swapped prosody – BE SwG segments with VS SwG prosody on top, and VS SwG segments with BE SwG prosody on the bottom.

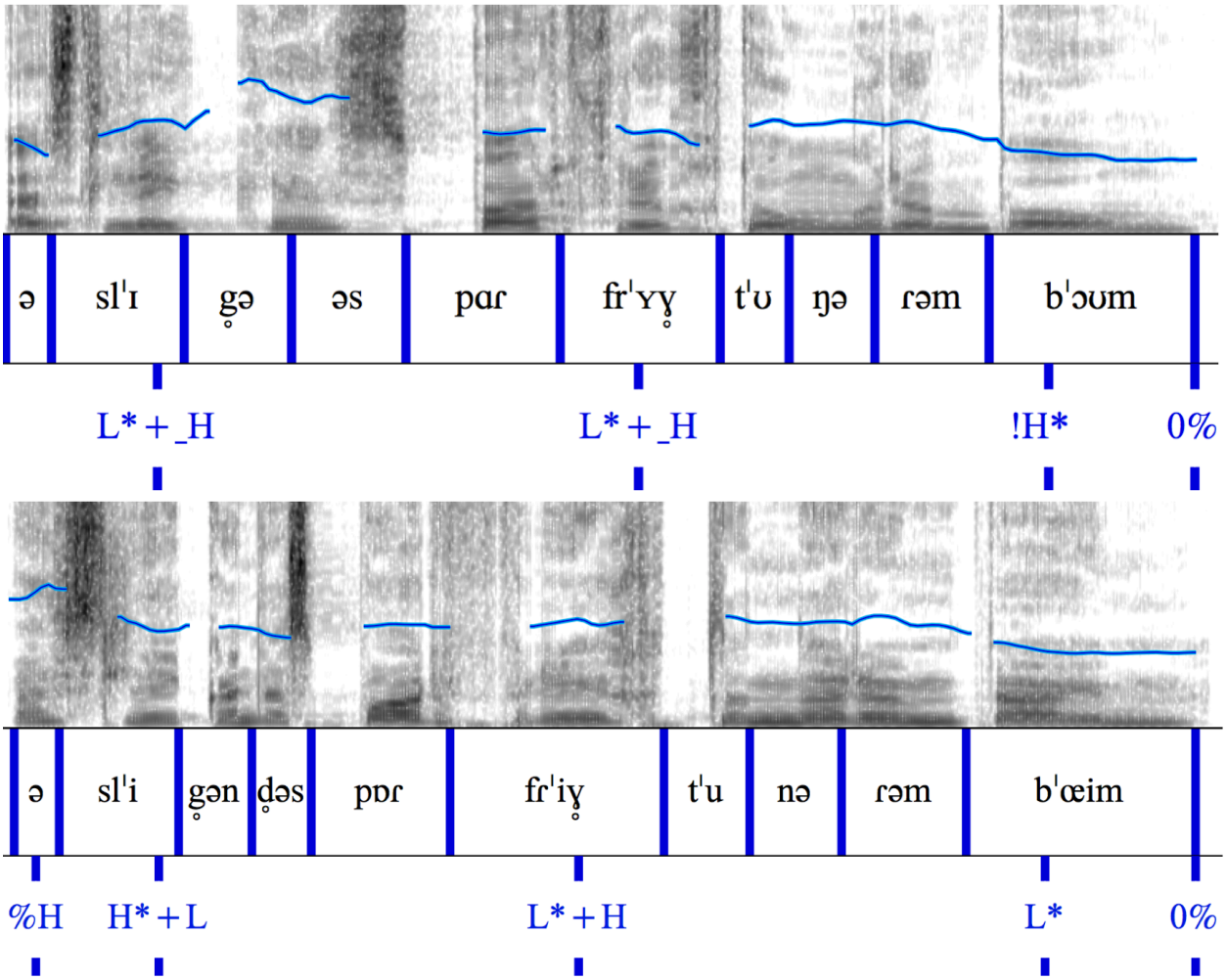

Figure 4: Example of where rhythm and intonation have been swapped: BE SwG segments with VS prosody (top panel) and VS SwG segments with BE SwG prosody (bottom panel).

In the example shown here, after rhythm and intonation transplantation, the BE SwG sentence (top) no longer has an %H boundary, but features L*+_H, L*+_H, and a downstepped !H* nucleus instead; likewise, after transplantation, VS SwG (bottom) now carries a %H, a prenuclear H*+L, L*+H, and ends with a low (L*) nucleus.

Each of the 12 speakers provided ten sentences, which amounted to 120 sentences per condition. To make the stimuli equally artificial in all conditions, the manipulations of rhythm and intonation described for *Condition III* were equally applied on the unmorphed stimuli presented in *Condition I*. That is, rhythm and intonation of BE SwG speaker 01 sentence 01, for example, were morphed onto the same token. This rendered these baseline stimuli in *Condition I* equally synthetic as the ones created in the rhythm morphing and rhythm and intonation morphing conditions. In doing so we ensured that we are not testing for an effect of artificiality of the stimuli but rather of the conditions themselves.

2.5 Subjects

62 listeners (21 for the unmorphed and rhythm combined with intonation morphed conditions each; 20 for the rhythm morphed condition) were recruited at the University of Zurich. For *Condition I* (unmorphed) mean age was 24.7 (SD=5.5), 71% f, 29% m; for *Condition II* (rhythm morphed) mean age was 23 (SD=4.3), 80% f, 20% m; for *Condition III* (rhythm and intonation morphed), mean age was 24 (SD=4.5), 71% f, 29% m. Listeners were students at the University of Zurich and fully competent in Zurich German dialect. We assume similar exposure to both dialects, VS and BE SwG, on the part of the listeners either through personal contacts or through the media. Had we had participants who, because of job requirements or other, would frequently travel to the cantons of Bern or the Valais, for example, exposure would have been different. The majority (an estimated 60%) of Swiss National Television programs are broadcast in dialect (Siebenhaar & Wyler 1997) – both, BE and VS SwG, are represented dominantly in TV broadcasts. The rest of the programs are broadcast in Swiss High German, where TV and Radio presenters loosely follow pronunciation guidelines established at Swiss National Radio and TV (Buri 1993).

2.6 Procedure

Listeners were tested in a quiet room at the University of Zurich using laptop computers. They heard stimuli over high-quality closed Beyerdynamics DT 770 PRO headphones. Stimulus order was randomized for each listener. Listeners were instructed as follows: they would hear a sentence; after sentence presentation they had to decide whether it was Bern or Valais SwG, and indicate how confident they were in their response. They were encouraged to respond intuitively. Listeners responded using a binary forced choice experiment interface presented over the *Praat* demo window. The interface is shown in Figure 5.

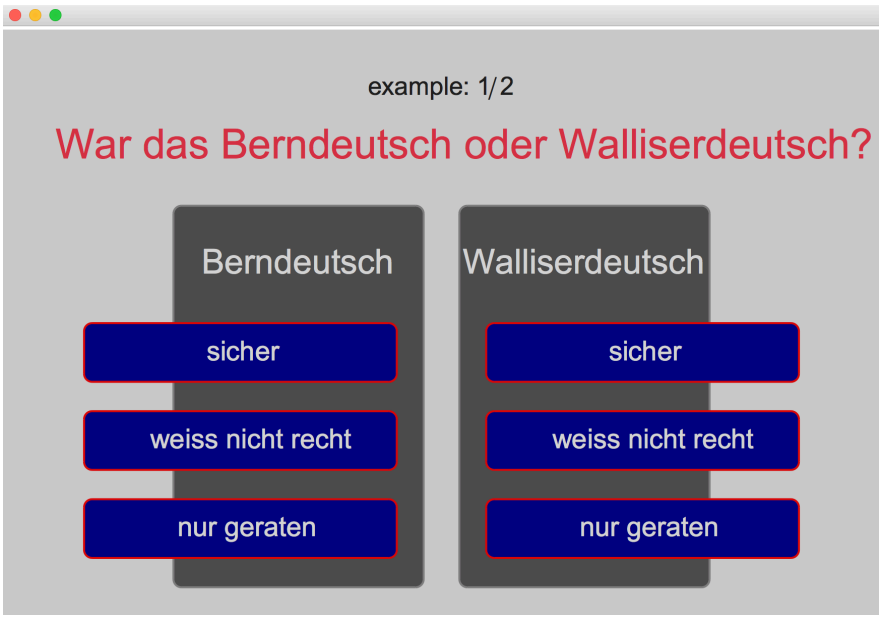

Figure 5: Experiment interface. To give the response, listeners clicked on the respective button. In red 'Was this Bern German or Valais German?'; on dark grey background 'Bern German', 'Valais German'; on blue background 'sure', 'not certain', 'only guessing'.

The red text reads ‘Was this Bern German or Valais German?’. The blue boxes reads ‘sure’, ‘not certain’, ‘only guessing’) – from top left to bottom left. Listeners clicked on one of the blue rectangles shown in Figure 3, indicating whether they judged the stimulus as being BE or VS SwG German. At the same time, they indicated their confidence level for each stimulus on a three-point scale (1=sure, 2=not certain, 3=only guessing). We included these graded responses as a means of introducing some form of freedom for the subjects while still operating within the 2 AFC design. Before the beginning of the experiment, listeners were familiarized with the experiment interface and with manipulated speech through the presentation of two randomly selected stimuli. The experiment, including instructions, lasted about 15 minutes and listeners were paid 10 Swiss Francs for participation.

2.7 Data analysis

We used R (R Core Team, 2012) and lme4 (Bates et al. 2017, default dummy-coding) to construct a mixed effects logistic regression model to test for the relationship between listeners' response (coded as 1=correct and 0=incorrect) and signal condition. We arbitrarily defined ‘correct’ to mean ‘listener perceives dialect based on segments’, i.e. 1=segmental response, 0=prosodic response. When we speak of percent correct below we thus mean ‘percent response based on segmental information’ – for ease of understanding and phrasing we maintain ‘percent correct’ throughout the paper, however. For the full specification of the model we adopted, please see the Appendix.

3. Results

3.1 Effect of condition

Figure 6 shows violin plots of percent correct for the three conditions: unmorphed (red), rhythm morphed (green), and rhythm and intonation morphed (blue); chance level = 50%.

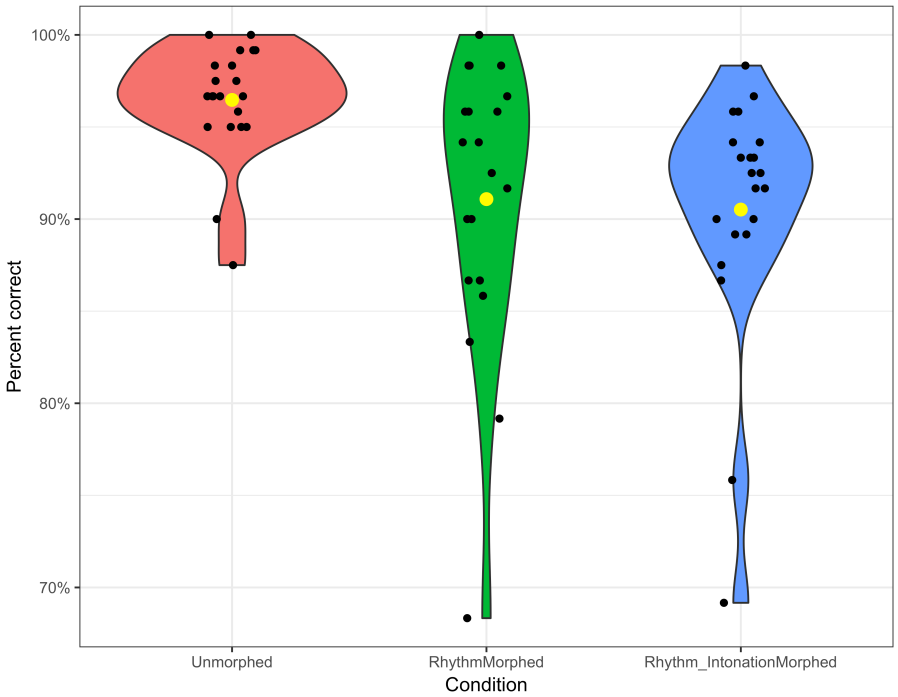

Figure 6: Violin plot of percent correct by condition: unmorphed condition (red), rhythm morphed (green), and rhythm combined with intonation morphed (blue). Points in the graph are

horizontally jittered, and the shapes of the boxes convey the overall distribution of data within the conditions. The yellow point indicates the median correct percentage in each condition.

The yellow dot represents the mean, the scattered dots show percent correct for individual listeners. The kernel density estimation indicates the distribution of the shape of the data – where wider sections illustrate a higher probability that members of a population will take on a given value. More narrow sections represent lower probability. Figure 6 reveals a near-ceiling effect: identification scores are high in all three conditions. They are highest in unmorphed speech (red, M=96.47%) followed by rhythm morphed speech (green, M=91.08%) and drop to their lowest in rhythm combined with intonation morphed speech (blue, M=90.52%). Table 2 shows the estimates of the logistic regression model.

Table 2: Estimates of the logistic regression model, showing the effects of conditions, speaker dialects, and their interaction. The baseline (intercept) contains unmorphed utterances by Bern (BE) Swiss German speakers.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.9537	0.4132	9.569	< 2e-16***
rhythmMorphed	-1.0273	0.4002	-2.567	0.01025*
rhythm&intonationMorphed	-1.2155	0.3849	-3.158	0.00159**
dialectSpeakerVS	0.2777	0.5622	0.494	0.62139
rhythmMorphed:dialectSpeakerVS	-0.11	0.5688	-0.193	0.84666
rhythm&IntonationMorphed:dialectSpeakerVS	-0.1156	0.5441	-0.212	0.83179

All three conditions are above chance level: the intercept is significantly different from 0 in the current model specification (logodds = 1 to 1 odds = 50%). When the other two conditions are specified as a reference level both are also significantly above 0. Multiple comparisons (Tukey method, using the R-package *multcomp*) revealed significant differences between the unmorphed (red) and rhythm morphed conditions (green) (p=.027*) and the unmorphed and the rhythm combined with intonation morphed conditions (blue) (p<.0045*). We used the same mixed effects logistic regression model as specified in the Appendix, except the response variable here is listener certainty (coded as 1 = 'sure', 2 = 'not certain', 3 = 'only guessing') instead of identification performance. Listener certainty did not vary in a statistically significant way between the three conditions, with 35.79% of the unmorphed stimuli having been rated with the listener claiming to be 'sure', while in the rhythm morphed condition this drops to 32.2% and in the rhythm and intonation morphed condition this is at 32.02%.

3.2 Effect of condition by dialect

We were interested in examining whether listener responses varied depending on which dialect they judged and which condition they judged the dialect in. Table 1 showed that there was no effect of dialect and no interaction between dialect and condition in the model. To probe possible causes for this non-significance, we look at detailed distributions of listener identification performance within the two dialects. Figure 7 shows the violin plot of percent correct crossed against *dialect* and *condition*.

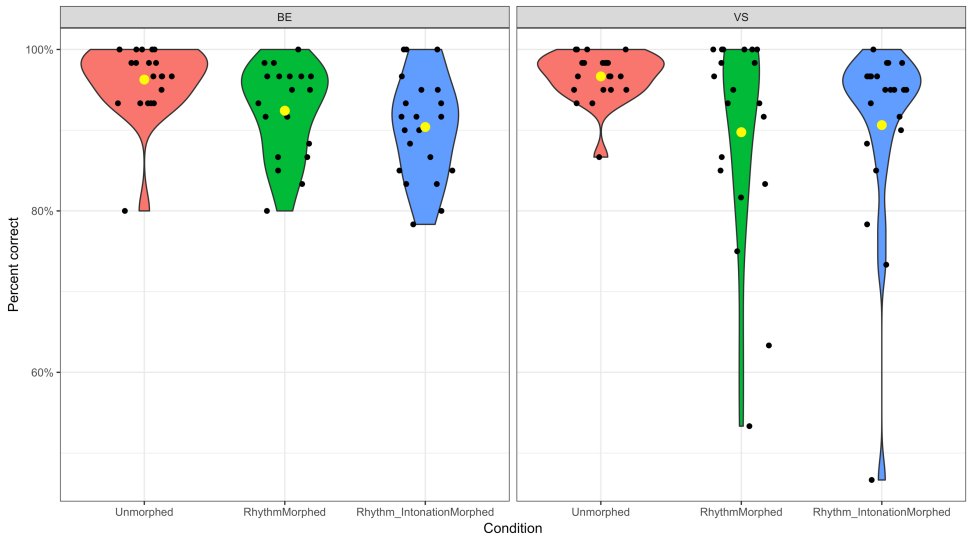

Figure 7: Violin plot of percent correct crossed against *dialect* and *condition*. Unmorphed condition (red), rhythm morphed (green), and rhythm combined with intonation morphed (blue); Bern Swiss German (left), Valais Swiss German (right).

Figure 7 shows that identification scores are high in both dialects in all three conditions. Variability across listeners, however, appears to be somewhat more pronounced for VS SwG in the rhythm (green) and rhythm combined with intonation morphed conditions (blue). When BE SwG rhythm and rhythm combined with intonation are morphed onto VS SwG segments, we obtain more variability across some listeners. This appears to be consistent with the estimated standard errors in the model output (see Table 2) (BE SwG=.41 and VS SwG=.56).

3.3 Effect of sentence material by condition

To explore the influence of segmental material on dialect identification performance, we calculated percent correct using responses to each sentence (twelve responses per listener for each sentence). Figure 8 presents percent correct for the ten sentences (see appendix) by the three conditions: unmorphed (red), rhythm morphed (green), rhythm combined with intonation morphed (blue); Table 3 shows the model coefficients, given that we modeled random intercepts for each sentence as well as by-sentence random slopes on condition.

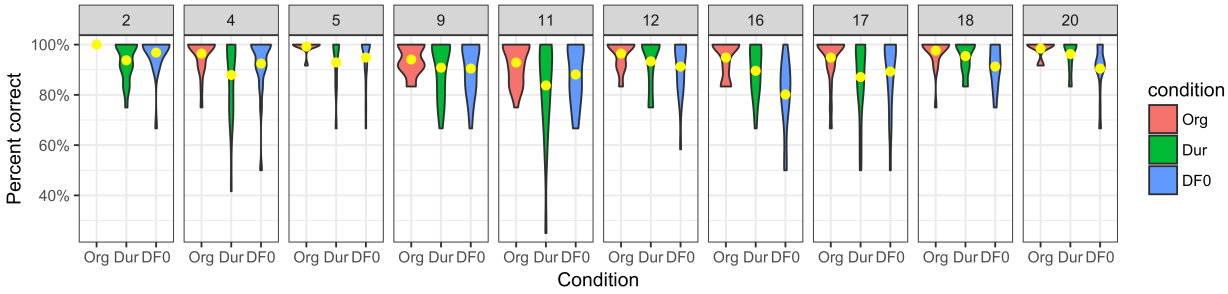

Figure 8: Violin plot of percent correct for each sentence by the three conditions. Unmorphed condition (red), rhythm morphed (green), and rhythm combined with intonation morphed (blue).

Table 3: Estimates of the random effects coefficients of the logistic regression model, showing the effects of conditions on each of the 10 sentences. The baseline (intercepts) correspond to unmorphed utterances.

sentence	intercept	rhythmMorphed	rhythm&intonationMorphed
2	5.066156	-1.8961316	-1.651005
4	4.382231	-0.9358807	-0.5295886
5	4.548806	-1.4807179	-1.4022713
9	3.89534	-0.5030658	-0.6322497
11	2.673042	-0.8624793	-0.7766129
12	3.862637	-0.7576124	-1.1085922
16	2.57597	-0.4539705	-1.2866299
17	3.250109	-0.7395579	-0.7075012
18	3.937334	-1.2029486	-1.8677595
20	4.147719	-0.8230777	-1.5658028

Figure 8 suggests that dialect identification appears to be influenced by the sentence material used. Sentence 2, for example, is identified well in all three conditions (unmorphed: 100%, rhythm morphed: 93.75%, and rhythm combined with intonation morphed: 96.83%). While sentence 16 was identified comparatively poorly, particularly as morphing progressed (unmorphed: 94.84%, rhythm morphed: 89.58%, and rhythm combined with intonation morphed: 80.16%). We are mindful that random effects might contain noise and that the random coefficients undergo shrinkage (i.e. the difference in estimates is smaller than the difference in actual values) – but the tendency that dialect identification appears to be influenced by the sentence material is reflected in the coefficients in Table 2. Sentence 2, for example, shows an intercept of 5.06 (unmorphed), the coefficient is lower for the rhythm morphed condition (-1.9) and the rhythm combined with intonation morphed condition (-1.65). Here, too, zero intercept means chance level, i.e. 50%. It might be the case that using sentence 2, for example, enabled listeners to more readily identify the dialect as either Bern or Valais SwG, particularly so in the unmorphed condition; using sentence 16, however, listeners had more difficulty in the identification process, especially so in the rhythm morphed condition, as reflected in the intercept shown in Table 3 (2.57), the coefficient for the rhythm morphed condition (-.45) and the unmorphed condition (-1.28).

4. Discussion

In this study we examined the role of segments and prosody in the identification of a speaker's dialect. We used two dialects which have been known to differ in segments and prosody. The qualitative acoustic analyses of the production data largely corroborate these between-dialect production differences. Based on the data from 12 speakers (6 VS SwG, 6 BE SwG, ten sentences each), VS SwG had a tendency of being spoken more quickly, and had more regular speech rhythm. The differences found in speech rhythm are consistent with previous studies by

Leemann et al. (2012), who report less vocalic interval variability for VS SwG – which may be said to reflect a more regular, syllable-timed rhythm. In articulation rate, too, previous studies have reported slower rates for BE SwG (e.g. Leemann 2012) – although our data only shows a trend in this direction. It is possible that the speaking style elicited, read speech, did not enable full permeation of naturally occurring articulation rates for the two dialects. Had we elicited spontaneous speech – in particular longer sentences – we speculate that dialectal differences in articulation rate would have been more pronounced. Spontaneously produced material would have hindered the application of the prosody transplantation paradigm, however, which requires the same material for both dialects. Findings from the intonation analysis revealed that the dialects feature different types of onset boundaries as well as different distributions of prenuclear and nuclear pitch accents. Both dialects demonstrated a preference for L*+H in prenuclear position, which has been reported by Leemann (2012) and Fitzpatrick-Cole (1999). The delayed peaks L*+_H for VS SwG have previously been reported in Leemann (2012) as well. Aside from prosodic differences, segmentally, too, our sentence material contains dialect-specific diagnostic features, such as vocalized /l/ in BE SwG (e.g. sentence 4 Appendix *geschmolzen*, 'melted', BE SwG [kʃmˈɔʊtsə], VS SwG [kʃmˈʊltsə]), velarized /nd̥/ in BE SwG (e.g. sentence 9, *Kind* 'child', BE SwG [ɣ̊ɪŋ], VS SwG [ɣ̊ind̥]), palatalized /s/ in VS SwG (e.g. sentence 10 *sie* 'they', BE SwG [siː] VS SwG [ʃiː]), and vocalic differences (e.g. sentence 11, *Strasse* 'street', BE SwG [ʃtrɑːsː], VS SwG [ʃtrɒːsː]), amongst others.

Using this material, 62 listeners who were familiar with the two varieties participated in a dialect identification task where we tested whether they pay more attention to segments, rhythm, or rhythm and intonation combined when identifying a dialect. Results of the perception experiments revealed the following findings pertinent to RQ1: What is the role of segments and prosody (rhythm alone and rhythm combined with intonation) in the identification of a speaker's dialect?:

(a) Listener performance was generally high at identifying the two dialects
(b) Speech rhythm and rhythm and intonation combined seem to occupy only a marginal – albeit statistically significant role – when it comes to identifying a speaker's dialect
(c) Morphing BE SwG prosody onto VS SwG segments seems to cause more variability in the listeners' responses than the other way around

Results further revealed the following finding pertinent to RQ2: How is dialect identification contingent on the sentence material used?

(d) Dialect identification performance depends on the sentence material use.

(a) The finding that SwG listeners perform well at identifying SwG dialects is consistent with the results reported in Leemann & Siebenhaar (2008) as well as Guntern (2011) (4 AFC, 8 AFC paradigms). The very high performance scores in this study (96.47%) can partly be explained by the fact that it is only a 2 AFC paradigm (chance level was 50%), but also because SwG dialects are the prestige varieties in Switzerland and are met with an exceptionally high approval in society (cf. Sieber & Sitta 1986, Christen 2010). Contrary to the situation in most parts of Germany and Austria, where Standard German represent the prestige variety, dialects represent the prestige varieties in Switzerland. Because of this, SwG listeners are exposed to different

dialects on a daily basis, which perhaps contributes to them having high dialect identification scores – as reported here and in previous studies.

(b) Our findings revealed that morphing rhythm and morphing rhythm combined with intonation caused a significant drop in dialect identification performance. For all three conditions, however, identification performance remained exceptionally high. In other words, morphing a dialect's rhythm and intonation onto the segments of another dialect does have a slight influence on listeners, yet, listeners are still very much able to identify the underlying dialect, i.e. they appear to identify dialects particularly based on segmental information. Fuchs' (2015) exploratory study found a similar trend: exposing Indian and British English listeners to speech for which segmental and prosodic information was morphed, he reports segmental information to be dominant over temporal and f0 information in the process of dialect identification. In a more qualitative approach, Guntern (2011), too, found that SwG listeners particularly report segmental cues (e.g. velar or uvular /r/ realizations for Eastern SwG dialects, /l/-vocalization for BE SwG etc.) rather than prosodic cues when asked about which features they pay attention to the most when identifying a dialect. One unanticipated finding was that whether one morphs rhythm or rhythm and intonation combined does not seem to affect listener judgment (the two morphing conditions were not significantly different from each other). A possible explanation for this may be that speech rhythm differences between the two dialects may be particularly salient in perception, which is why, in performing the task, listeners' identification performance drops significantly when rhythm is morphed. Perhaps the two dialects are not perceived as being very different from each other in intonation, which may be why identification performance does not drop further when intonation is morphed along with speech rhythm. For reasons outlined in section 2.4(3), intonation alone cannot be morphed sensibly in the prosody swapping approach we used.

(c) Results in section 3.2 showed that the morphing conditions elicited somewhat different listener responses depending on the dialect examined: when morphing BE SwG rhythm and rhythm combined with intonation onto VS SwG, some listeners performed more poorly at identifying VS SwG. Presumably, BE SwG is marked by high vocalic variability in speech rhythm (Leemann et al. 2012) and L*+H accents in prenuclear and nuclear position without a displacement of trailing tones. Superimposing these prosodic features onto the segmental string of VS SwG caused more variability in the identification performance across some listeners than superimposing VS SwG prosody onto BE SwG segments. This may suggest that, to some Zurich German listeners, BE SwG prosodic cues may carry substantial saliency – causing the confusion in listeners. And, arguing along the same lines, VS SwG segments may not be very strong indicators of that dialect when occurring in conjunction with BE SwG prosody. More generally, this raises the question of what would happen if other dialects had been investigated in this study. As pointed out in the introduction, results of a study of this kind heavily rely on the segmental and prosodic similarities and differences of the dialects selected. Had we selected dialects that are particularly different in the prosodic domain but less so in the segmental domain, it is likely that applying prosody transplantation would have caused greater listener confusion. In the context of Swiss German, however, substantial differences in prosody typically go hand in hand with differences in segmental make-up. Expanding this paradigm to varieties of the same language that particularly differ in prosody and much less so in segments would be worthwhile exploring in future research.

(d) Effect of sentence material: In the literature review and in the results presented earlier we pointed out how the two dialects differ segmentally and prosodically. These differences may be reflected in different ways in the sentence material used: some sentences may carry more diagnostic information about a dialect, other sentences less. These differences in diagnostic cues in the sentence material can affect listener judgment in a dialect identification task. For this reason, we examined identification performance for each sentence and conditions, results of which we presented in section 3.3. We feel that it is worth exploring these triggers further as this may provide additional insights regarding which cues govern dialect identification, albeit in a qualitative fashion. Looking at one case we found, for instance, that listeners revealed the highest identification performance on the basis of sentence 2 (100%) and somewhat lower scores for sentence 16 in unmorphed speech (94.84%). Which means that, in this condition, when listeners judged sentence 2 they were readily able to tell whether it was BE or VS SwG; dialect identification appeared to be just a bit more difficult when listeners heard sentence 16. However, when listeners judged sentence 2 in the full prosody morphed condition, identification rates were still at 96.8%; whereas when they judged sentence 16, rates dropped to 80.16%. Let us examine the segmental and prosodic make-up of these two sentences in a bit more detail.

Sentence 2
BE: Es lige es paar Frücht ungerem Boum [əs ˈlɪgə əs pɑːɾ frʏxt ˈʊŋərəm bɔʊm]
VS: Es liggunt äs par Fricht unnärum Böim [æs ˈl̥igunt æs pɒːɾ frɪxt ˈʊnærʊm bœim]
German: Es liegen ein paar Früchte unter dem Baum̊
English: Some fruit is lying under the tree

In sentence 2, the dialects differ in terms of vowel quality (tense high front vowels and rounded low back vowels in VS SwG), velarization of /nd̥/ (featured in BE SwG, e.g. [ˈʊŋərəm]), and tense marking (3rd p. sing. [-unt] in VS SwG [ˈligunt]). In terms of the prosodic make-up of this sentence, the two dialects showed virtually the same speech rates (BE SwG 5.1 syll / sec, VS SwG 5 syll / sec) but quite different VarcoSyll scores (BE SwG M=.52, VS SwG M=.39) – i.e. less vocalic variability in VS SwG. As for intonation, four of the six BE SwG speakers realized onset boundaries (%H); only one VS SwG speaker did so; in prenuclear position, BE SwG featured predominantly H*+L (50%) and L*+H (41%); VS SwG L*+H (58%) and L*+_H (33%). In nuclear position, BE SwG speakers showed L* only, VS SwG speakers featured L* (50%) and L*+H (33%).

Sentence 16
BE: Er het sech i ne Finger gschnitte [ɛɾ hɛt səx ɪ nə ˈfɪŋər ˈkʃnɪtə]
VS: Är hät schich in ä Fingär gschnittu [æɾ hæt ʃix i næ ˈfiŋær ˈkʃnitʊ]
German: Er hat sich in den Finger geschnitten
English: He cut his finger
In terms of segments, BE and VS SwG differ particularly in terms of vowel quality (tense high front vowels in VS SwG), vowels in open syllables (full vocalic articulation in VS SwG, e.g. [kʃnˈitu]), and in the quality of /s/ (palatalized in VS SwG, e.g. in [ʃiɣ̊]). Prosodically, BE SwG has an articulation rate of 5.2 syll/sec, VS SwG 5.9 syll/sec. *VarcoSyl* is also lower in VS SwG (M=.37) than it is in BE SwG (M=.41) in this sentence – suggesting more regular, syllable-timed rhythm. In terms of intonation, the two dialects differed substantially: virtually all of the 12

speakers realized an onset boundary tone, %L in VS SwG (N=6) and %H in BE SwG (N=5). As for prenuclear pitch accents, BE SwG featured L*+H only, while VS SwG had L*+_H (33%), H* (33%), L*+H (17%), and H*+L (17%). In nuclear position, BE SwG demonstrated 67% !H*+L and 33% L*; VS SwG featured !H*+L pitch accents only.

Prosodically and segmentally, the dialect-specific features observed for these two sentences reflect some general between-dialect differences found in the sentence material used (see 2.3). One question that remains is why listener performance drops more for sentence 16, the more the signal was morphed while this decrease in identification performance is much less pronounced for sentence 2. Segmentally, above analysis showed that both sentences – 2 and 16 – differ quite substantially. In the domain of prosody, however, the two dialects perhaps differed more in sentence 16 than they did in sentence 2: while articulation rates were virtually identical for sentence 2, they differed substantially in sentence 16. In terms of intonation, too, there may be more pronounced differences between the dialects in sentence 16 – particularly the difference in onset boundary tones (%L in VS SwG vs. %H in BE SwG) as well as in the types of prenuclear accents. It is possible that, because prosodic differences are more pronounced in sentence 16, the identification performance dropped for sentence 16 when prosody was morphed.

More generally, results of this study may have implications for speaker identification by victims and witnesses, the cognitive bases for storing and accessing indexical information such as a speaker's dialect, and for automatic speech recognition.

The phenomenon that a speaker's dialect constitutes an integral component of a speaker's identity is exploited in forensic casework. In forensic casework, experts evaluate not only speech material but also the evidence of witnesses and victims who might have heard the criminal's voice. In these instances, it may be appropriate to test whether the voice of a suspect matches the memory of the one heard during the crime. In the course of the investigation, the 'earwitnesses' may attend a voice parade where a recording of the suspect's voice is presented along with a number of recordings of similar voices. The earwitness is then asked to pick out, if it is present, the speaker s/he believes to have heard in the crime scene (Nolan & Grabe 1996, Nolan 2003). When earwitnesses are asked to describe the voice they heard, dialect forms a central part in that description (cf. Hollien 2002). Knowing more about which cues are diagnostic for dialect recognition may affect the reliability of claims (both expert and naïve) about a suspect speaker's dialect, making these claims diagnostically more conclusive. It may also assist the process of selecting 'foil' voices, ensuring a fair voice parade. In future studies it will be worthwhile exploring the current experiment design with conditions that include background noise. Introducing noise will pose a forensically more realistic scenario, as disputed samples often include some form of background noise (e.g. multi-talker babble, bar noise, traffic noise, reverberation etc. – personal communication Dr. Olaf Köster, German Federal Criminal Police Office BKA). Noise, obviously, distorts the shape of the spectrum and can make speech highly unintelligible. Multi-talker babble with spectral features between 100 and 6000 Hz has the most detrimental effects on speech perception, given that speech energy is concentrated in this region (cf. Assmann & Summerfield 2004). The introduction of noise will provide additional insights how indexical information about dialects is ordered: using background noise with the stimuli at hand raises the question to what degree naïve witnesses, for example, can still assess a suspect speaker's dialect in noise conditions. Will the contribution of segmental and prosodic information for the identification of a speaker's dialect be the same in adverse and ideal listening

conditions? In the current study, we reported that segmental information overrides prosodic information in the identification of a dialect in ideal listening conditions. Studies have shown, however, that background noise can trigger a re-ranking of acoustic cues to linguistic categories; 'secondary' cues in laboratory settings, i.e. prosodic information, may become the only accessible cues when background noise is present (cf. Summerfield & Haggard 1977, Parikh & Loizou 2005). Van Zyl & Hanekom (2011), for example, report that word recognition strongly deteriorates as SNRs decrease, while the recognition of prosodic patterns remains robust. Cognitive coping strategies that listeners use when exposed to speech in adverse listening will expose mechanisms that may not come to light in laboratory conditions.

Remaining with the cognitive bases of the present study, our results have shown that listeners perform well at extracting the information necessary to make the right judgments about the dialectal origin of a speaker, despite there being conflicting information of segments and prosody in the signal. It is currently largely unknown how exactly dialectal information and other social information affects linguistic processing (cf. Foulkes 2010). We know that dialectal information is stored in memory alongside linguistic knowledge (Clopper & Pisoni 2005) and that processing speeds of dialects differ depending on frequency of encounter with the dialects and social saliency of the dialects (Clopper et al. 2016). Awareness of dialectal variation has been reported from an early age: children aged four are able to perform dialect classification tasks, with performance increasing incrementally with age, reaching adult-like classification scores in late adolescence (Jones et al. 2016). Results from the current study suggest that it is particularly a dialect's segmental information that assists the perceptual identification of the dialects, which in turn implies that regional variation in prosody serves as a redundant, supplementary function that may not be so relevant for dialect identification in ideal listening. Whether segmental information plays such a central role in dialect identification in children, or whether, perhaps, prosodic cues bear more diagnostic value, remains to be studied. Newborns have been shown to be able to identify mother's voices from alien ones given their exposure to the mother's voice, particularly her prosodic features, when still inside the womb (Smith et al. 2003); arguable this would contribute to the children exploiting prosodic information more for dialect identification in early years. Studying children at an earlier age than four is questionable, however, as they are unlikely to have the necessary sociolinguistic competence and social perception skills needed to be able to conceptualize of dialect variation (cf. Labov 1964).

The present study also has implications for automatic speech recognition (ASR). ASR systems are still highly limited with respect to variation in the input signal. The more we know about dialectal variability in speech and the more we know about how variation in speech is perceived, the more we will be able to apply these insights to building robust ASR systems. Biadsy & Hirschberg (2009), for example, show that prosodic information significantly improves ASR-based dialect identification. Here, too, future studies that include noise in the signal will be of relevance: today, small departures from quiet listening conditions can lead to a rapid drop in the recognition accuracy of ASRs, as taking noise into account is perhaps the greatest obstacle in today's ASR technology (cf. Barker et al. 2005). Results of future studies using the same paradigm as in the present study but including background noise will reveal which features – segmental or prosodic – are more critical for dialect identification in noise.

5. Conclusion

This study set out to determine the relative importance of segments, rhythm alone, and rhythm combined with intonation in the identification of a speaker's dialect. One of the more significant findings to emerge from this study was that listeners seem to pay attention particularly to segmental information; rhythm alone and rhythm combined with intonation do play a role in the identification process, as performance significantly decreases when this information is swapped between the dialects, yet, the original target dialects are still identified much above chance. The investigation has further shown that morphing BE SwG prosody onto VS SwG segments tends to cause more variation in listeners' responses than morphing VS SwG prosody onto BE SwG segments. The experiments have also shown that sentence material used seems to occupy a critical role in this type of experiment, as listener judgment appears to vary depending on the sentence material used. This research extends our knowledge of the functional load of segments and prosody in the process of dialect identification. The main weakness of this study is that the contribution of intonation itself could not be examined, as it was coupled with speech rhythm in the morphing process. More research is required to determine the role of intonation relative to segments in the identification of a dialect. Furthermore, we were not able to isolate in detail exactly what segmental differences were used by the listeners to enable successful dialect identification. What we presented in the discussion (cf. Section 4 (d)) are analyses that remain on the sentence level. Future studies will have to explore in detail the actual saliency of specific segments in the process of dialect identification. Notwithstanding these limitations, findings of the study may have implications for forensic phonetics, the cognitive bases of accessing indexical information about a speaker's dialect, and automatic speech recognition.

6. Acknowledgements

The authors thank Volker Dellwo for useful comments and technical assistance in the experiment design. We thank Timo Röttger (Northwestern), the Editor, and two other anonymous reviewers for their very helpful comments during the review process. This research is funded by the Swiss National Science Foundation, grant Nr. P300P1_151210, http://p3.snf.ch/project-151210.

7. References

Assmann, P.F., A.Q. Summerfield (2004). The perception of speech under adverse conditions. In S. Greenberg, W.A. Ainsworth, A.N. Popper, R.R. Fay (Eds.), *Speech Processing in the Auditory System 14, Springer Handbook of Auditory Research (pp. 123-148).* New York: Springer.

Barker, J.P., Cooke, M.P., & Ellis, D.P.W. (2005). Decoding speech in the presence of other sources. *Speech Communication* 45, 5-25.

Barr, D. J., Levy, R., Scheepers, C. & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language,* 68, 255–278.

Barry, W., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, 66(1-2), 78-94.

Barry, W., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? *Proceedings of the 15th ICPhS*, 2693-2696.

Bates, D., Maechler, M., Bolker, B., Walker, S., Haubo Bojesen Christensen, R., Singmann, H., Dai, B., Grothendieck, G., Green, P. (2005). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. R package version 1.1-13.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research*

Foundation 3.
Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British journal of audiology*, 13(3), 108-112.
Bent, T., Atagi, E., Akbik, A., & Bonifield, E. (2016). Classification of regional dialects, international dialects, and nonnative accents. *Journal of Phonetics*, 58, 104-117.
Biadsy, F., Hirschberg, J. (2009). Using Prosody and Phonotactics in Arabic Dialect Identification, *Proceedings of Interspeech 2009*, 208-211.
Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer*. http://www.praat.org.
Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, 63(4), 247-267.
Buri, R. M., et al. 1993. *Deutsch Sprechen am Schweizer Radio DRS*. Bern: Schweizer Radio DRS.
Christen, H. (2010). Was Dialektbezeichnungen und Dialektattribuierungen über alltagsweltliche Konzeptualisierungen sprachlicher Heterogenität verraten. In C. Anders, M. Hundt & A. Lasch (Eds.), *"Perceptual dialectology". Neue Wege der Dialektologie* (pp. 269–290). Berlin/New York: de Gruyter.
Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32(1), 111-140.
Clopper, C. G., & Pisoni, D. B. (2005). Perception of Dialect Variation. In D. Pisoni & R. Remez (Eds.), The Handbook of Speech Perception (pp. 312-337). New York: Wiley & Sons.
Clopper, C. G., Tamati, T. N., & Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. Journal of phonetics, 58, 87-103.
Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1129-1132.
Dellwo, V., Leemann, A., & Kolly, M. J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. *Proceedings of Interspeech 2012*, 1584-1587.
Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *Tesol Quarterly*, 39(3), 379-397.
Fitzpatrick-Cole, J. (1999). The alpine intonation of Bern Swiss German. *Proceedings of the 14th International Congress of Phonetic Sciences*, 941-944.
Fuchs, R. (2015). You're not from around here, are you? – A dialect discrimination experiment with speakers of British and Indian English. In E. Delais-Roussarie & M. Avanzi (Eds.), *Prosody and Languages in Contact: L2 Acquisition, Attrition, Languages in Multilingual Situations* (pp. 123–148). Heidelberg: Springer.
Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5(4), 233-242.
Foulkes, P. (2010). Exploring social-indexical knowledge: A long past but a short history. *Laboratory Phonology*, 1(1), 5-39.
Gilles, P., Peters, J., Auer, P., & Selting, M. (2001). Perzeptuelle Identifikation Regional Markierter Tonhöhenverläufe. Ergebnisse einer Pilotstudie zum Hamburgischen. *Zeitschrift für Dialektologie und Linguistik*, 68(2), 155-172.
Grabe, E., Post, B., & Nolan W. [sic] (2001). Modelling intonational variation in English: the IViE system. In S. Puppel & G. Demenko (Eds.), *Proceedings of Prosody 2000*.
Guntern, M. (2011). Erkennen von Dialekten anhand von gesprochenem Schweizerhochdeutsch.

Zeitschrift für Dialektologie und Linguistik, 78(2),155-187.
Gussenhoven, C. (1990) Tonal association domains and the prosodic hierarchy in English. In S. Ramsaran (Ed.), *Studies in the pronunciation of English* (pp. 27–37). London: Routledge.
Hollien, H. (2002). *Forensic Voice Identification*. San Diego, CA: Academic Press.
Holm, S. (2008). *Intonational and durational contributions to the perception of foreign-accented Norwegian: an experimental phonetic investigation*. Unpublished doctoral dissertation. Trondheim: Norwegian University of Science and Technology.
Jessen, M. (2007). Speaker Classification in Forensic Phonetics and Acoustics. In C. Müller (Ed.), *Speaker Classification* (1) (pp. 180-204). New York: Springer.
Jones, Z., Yan, Q., Wagner, L., & Clopper, C. G. (2017). The development of dialect classification across the lifespan. *Journal of Phonetics*, 60, 20-37.
Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2), 1038-1054.
Köster, O., Kehrein, R., Masthoff, K., & Boubaker, Y. H. (2012). The tell-tale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians. *International Journal of Speech, Language & the Law*, 19(1).
Labov, W. (1964). Stages in the acquisition of standard English. In R. Shuy (Ed.), Social dialects and language learning (pp. 77–104). Champaign, IL: National Council of Teachers of English.
Lai, C., Evanini, K., & Zechner, K. (2013). Applying rhythm metrics to non-native spontaneous speech. *Proceedings of SLaTE*, 159-163.
Ladd, D R (2008). *Intonation Phonology* (2nd ed.). Cambridge: CUP.
Leemann, A. (2012). *Swiss German Intonation Patterns*. Studies in Language Variation 10, Amsterdam: John Benjamins.
Leemann, A., Dellwo, V., Kolly, M. J., & Schmid, S. (2012). Rhythmic variability in Swiss German dialects. *Proceedings of Speech Prosody 2012*, 607-610.
Leemann, A., Kolly, M. J., & Nolan, F. (2016). Identifying a speaker's regional origin: the role of temporal information. *Proceedings of Speech Prosody 2016*.
Leemann, A., & Siebenhaar, B. (2008). Perception of dialectal prosody. *Proceedings of Interspeech 2008*, 524-527.
Leemann, A., & Siebenhaar, B. (2010). Statistical modeling of F0 and timing of Swiss German dialects. *Proceedings of Speech Prosody 2010*, 1-4.
Lenz, A. (2010). Zum Salienzbegriff und zum Nachweis salienter Merkmale. In C.M. Anders, M. Hundt & A. Lasch (Eds.), "*Perceptual dialectology". Neue Wege der Dialektologie* (pp. 89–110). Berlin/New York: de Gruyter.
Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, 129(5), 3258-3270.
Nolan, F. (2003). A recent voice parade. *International Journal of Speech Language and the Law*, 10(2), 277-291.
Nolan, F., & Grabe, E. (1996). Preparing a voice lineup. *International Journal of Speech Language and the Law*, 3(1), 74-94.
Parikh, G., Loizou, P. C. (2005). The influence of noise on vowel and consonant cues. *Journal of the Acoustical Society of America* 118(6), 3874-3888.
Peters, J., Gilles, P., Auer, P., & Selting, M. (2002). Identification of Regional Varieties by

Intonational Cues An Experimental Study on Hamburg and Berlin German. *Language and Speech*, 45(2), 115-138.
Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2), 175-184.
Petyt, K. M. (1980). *The study of dialect: An introduction to dialectology*. Westview press.
Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, 52(11), 911-918.
R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Version 3.0.0., 2013. http://www.R-project.org.
Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1), 512-521.
Schaeffler, F., & Summers, R. (1999, August). Recognizing German dialects by prosodic features alone. *Proceedings of the 14th International Congress of Phonetic Sciences,* 2311-2314).
Sieber, P., & Sitta, H. (1986). *Mundart und Standardsprache als Problem der Schule*. Salzburg: Sauerländer.
Siebenhaar, B., & Wyler, A. (1997). *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. Zürich: Pro Helvetia.
Smith, S. L., Gerhardt, K. J., Griffiths, S. K., Huang, X. & Abrams, R. M. (2003). Intelligibility of sentences recorded from the uterus of a pregnant ewe and from the fetal inner ear. *Audiology and Neuro Otology*, 8, 347 - 353.
Summerfield, Q. & M. Haggard (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America* 62, 435-448.
Ulbrich, C. (2013). German pitches in English: Production and perception of cross-varietal differences in L2. Bilingualism. *Language and Cognition*, 16(02), 397-419.
Ulbrich, C., & Mennen, I. (2015). When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accent. *International Journal of Bilingualism*.
Vaissière, J., & Boula de Mareüil, P. (2004). Identifying a language or an accent: from segments to prosody. *Colloque MIDL*, 1-6.
Van Bezooijen, R., & Gooskens, C. (1999). Identification of Language Varieties: The Contribution of Different Linguistic Levels. *Journal of language and social psychology*, 18(1), 31-48.
Vicenik, C. (2011). The Role of Intonation in Language Discrimination by Infants and Adults. PhD dissertation, UCLA.
Vicenik, C., & Sundara, M. (2013). The role of intonation in language and dialect discrimination by adults. *Journal of Phonetics*, 41(5), 297-306.
Vieru-Dimulescu, B., & Mareüil, P. B. D. (2005). Contribution of prosody to the perception of a foreign accent: a study based on Spanish/Italian modified speech. *ISCA Workshop on Plasticity in Speech Perception*.
Werlen, I. (1977). Das" Staubsche Gesetz" im Schweizerdeutschen. Zeitschrift für Dialektologie und Linguistik, 44(3), 257-281.
Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic*

applications. arXiv:1308.5499.
Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, 55(3), 486-507.

Appendix

The sentences below were presented to the subjects. The material was transliterated in the respective SwG dialects (BE or VS) with the help of two native writers (as there is no formal writing system for SwG) – this would enable fluent reading on the part of the speakers. Had we used Standard German writing, we would have expected potential interference effects when the subjects try to articulate the sentences in dialect. For ease of reading, we also provide IPA transcriptions of the individual dialect realizations.

Sentence 1 (2)
BE: Es lige es paar Früecht ungerem Boum [əs ˈlɪgə əs pɑːɾ fɾʏxt ˈʊŋərəm bɔʊm]
VS: Es liggunt äs par Fricht unnärum Böim [əs ˈlɪ̊gunt æs pɒːɾ fɾixt ˈʊnæɾʊm bœim]
German: Es liegen ein paar Früchte unter dem Baum̊
English: Some fruit is lying under the tree

Sentence 2 (4)
BE: D Sunnä hät der Schnee gschmouze [d̥ ˈsʊnːə hət d̥əɾ ʃnɛː ˈkʃməʊt͡sə]
VS: D Sunnu hät där Schnee gschmolzu [d̥ ˈsʊnːu hæt d̥əɾ ʃneː ˈkʃmɔlt͡sʊ]
German: Die Sonne hat den Schnee geschmolzen
English: The sun melted the snow

Sentence 3 (5)
BE: Si hei mit chautem Wassr gwäsche [si hei mɪt ˈxɑʊtəm ˈɣɑsəɾ ˈkɣæʃə]
VS: Schi heint mit chaaltum Wassär gwäschu [ʃi heint mit ˈxɑltʊm ˈɣɒsəɾ ˈkɣæʃʊ]
German: Sie haben mit kaltem Wasser gewaschen
English: They washed with cold water

Sentence 4 (9)
BE: Ds Ching nimt es Schpiiuzüüg id Hang [d̥s xɪŋ nɪmt əs ˈʃpɪʊt͡sɣːg̊pɪt hɑŋ]
VS: Ds Chind nimmt es Schpilzig in d Hand [d̥s xind̥ nimt æs ˈʃpiːlt͡siːg̊piːlhɒnd̥ɒ
German: Das Kind nimmt ein Spielzeug in die Hand
English: The child is taking a toy in his hand

Sentence 5 (11)
BE: D Straass isch uf der Chaarte [d̥ ʃtɾɑːs ɪʃ uf d̥əɾ ˈxɑːɾtə]
VS: D Straass isch uf där Chaartu [d̥ ʃtɾɒːs iʃ uf d̥æɾ ˈxɒːɾtʊ]
German: Die Strasse ist auf der Karte
English: The street is on the map

Sentence 6 (12)
BE: Dr aut Maa macht sech Soorgä [d̥ɾ ɑʊt mɑː mɑxt səx ˈsɔːɾg̊ə]
VS: Där aalt Maa macht schich Soorgä [d̥æɾ ɒlt mɒː mɒxt ʃix ˈsɔːɾg̊æ]
German: Der alte Mann macht sich Sorgen

English: The old man is worried

Sentence 7 (16)
BE: Er het sech i ne Finger gschnitte [ɛɾ hɛt səx ɪ nə ˈfɪŋər ˈkʃnɪtə]
VS: Är hät schich in ä Fingär gschnittu [æɾ hæt ʃix i næ ˈfiŋær ˈkʃnitʊ]
German: Er hat sich in den Finger geschnitten
English: He cut his finger

Sentence 8 (17)
BE: E Chatz isch vom Wage gschtigä [ə xɑts ɪʃ fɔm ˈɣɑgə ˈkʃtɪgə]
VS: Ä Chatz isch vom Wagu gschtigu [æ xɒ͡ts iʃ fom ˈɣɑ̝gʊ ˈkʃtɪ̝gʊ]
German: Eine Katze ist vom Wagen gestiegen
English: A cat climbed down a car

Sentence 9 (18)
BE: Ds Outo het e Wang gschtreift [d̥s ˈɔʊtɔ hɛt ə ɣɑŋ kʃtɾeift]
VS: Ds Auto hät en Wand gschtriift [d̥s ˈɒʊtɔ hæt ən ɣɑnd̥ kʃtɾiːft]
German: Das Auto hat eine Wand gestreift
English: The car touched a wall

Sentence 10 (20)
BE: Si hei zwöi lääri Fläsche ghaa [sɪ hei t͡sɣœi ˈlæːɾi ˈflæʃə khɑː]
VS: Schi heint zwei leeri Fläsche gha [ʃi heint t͡sɣei ˈleːɾi ˈflæʃə khɒː]
German: Sie hatte zwei leere Flaschen
English: She had two empty bottles

Full Model
correctness ~ condition + dialectSpeaker + condition * dialectSpeaker + (condition | speaker) + (dialectSpeaker + condition | sentence) + (dialectSpeaker | listener)

We included by-listener random slopes on speaker dialect to account for listener-specific bias, as listeners typically vary in leaning towards a particular response. We further included by-speaker random slopes on condition and by-sentence random slopes on speaker dialect and condition to make the model as maximal as the data allow for (following Barr et al. 2013 and Winter 2013).

Random effects

Groups	Name	Variance	Std.Dev.	Corr		
listener	(Intercept)	0.63382	0.7961			
	dialectSpeakerVS	1.79683	1.3405	-0.42		
speaker	(Intercept)	0.19358	0.44			
	rhythmMorphed	0.1222	0.3496	-0.69		
	rhythm&intonationMorphed	0.07854	0.2802	-0.62	1	
sentence	(Intercept)	0.67523	0.8217			
	dialectSpeakerVS	0.89805	0.9477	-0.35		
	rhythmMorphed	0.24228	0.4922	-0.68	-0.2	
	rhythm&intonationMorphed	0.24052	0.4904	-0.32	-0.72	0.46