Crowdsourcing regional variation in speaking rate through the iOS app 'Dialäkt Äpp'

Adrian Leemann[1], Marie-José Kolly[1], Volker Dellwo[1]

[1]Phonetics Laboratory, Department of General Linguistics, University of Zurich
{adrian.leemann, marie-jose.kolly}@pholab.uzh.ch, volker.dellwo@uzh.ch

Abstract

It is a common stereotype in Switzerland that speakers from Bern speak slowly and speakers from Zurich speak quickly. Are these differences in perception at all mirrored in production? We present a new method of crowdsourcing speaking rate through a free of charge iOS application. Astonishingly, results indicate that the temporal structure of a few words alone – as spoken by a few hundred speakers – are sufficient to tell apart the two dialects in speaking rate. In line with previous literature, females articulate more slowly than males. Further potential fields of application of the introduced method are discussed.

Index Terms: Speaking rate, crowdsourcing, dialectology, Swiss German, iOS application

1. Introduction

Swiss German dialects are spoken by roughly 4.5 million people [1] and enjoy high prestige in Swiss society [2, 3, 4]. Speakers of Swiss German (SwG) are well aware of regional variation and many dialects are stereotyped: Zurich Swiss German (ZH SwG), for example, is perceived as fast. Bern Swiss German (BE SwG), which enjoys the status of being Switzerland's most popular regional variety [5], is perceived as very slow [6, 7, 8]. Whether these differences in perception are reflected in production has been examined in passing by [9, 10]. Based on a corpus of spontaneous speech for ten speakers per dialect, [9, 10] reported that ZH SwG speakers articulate nearly one syllable more per second than BE SwG speakers (5.8 syll./sec. vs. 5.0 syll./sec – excluding pauses), thereby corroborating the previously mentioned stereotypes. Possible reasons for these differences in speaking rate were given in [10], who showed that BE SwG speakers produced distinctly longer mean durations of vowels and, in particular, exhibited more distinct phrase-final lengthening.

The studies mentioned present one major weakness: while clearly highlighting trends in regional variation in speaking rate, these tendencies have yet to be validated on a large set of speakers and on controlled material. We aim to alleviate this issue: based on crowdsourced data from an iOS application, we provide a more precise estimate of regional variation in speaking rate on the basis of nearly 250 speakers who articulated a controlled set of words. The term 'crowdsourcing' refers to "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" [11].

The use of crowdsourcing applications for studying linguistic phenomena has, until recently, received relatively little attention. This is somewhat surprising given that iPhone microphones, for example, feature wide frequency ranges of 50Hz-20kHz that enable high-quality audio recordings [12]. Previous research showed that a first generation iPhone from 2007 provides very useful for speech analysis and allows for reliable acoustic measurements – particularly for F1 and F2 [13]. Currently, a number of smartphone applications are in use or in development for crowdsourcing linguistic data. [14, 15] developed Android applications as a means to collect speech for the training of acoustic models. [16, 17] are applications currently in development for the purpose of documenting endangered languages, putting language documentation in the hands of the speakers. The mentioned apps are primarily used for acoustic modeling, dictionary building, text collection, translation, as well as dialect mapping. We present a novel method for crowdsourcing data to conduct research on prosodic features of dialects.

2. Data and methods

2.1. iOS application: 'Dialäkt Äpp'

'Dialäkt Äpp' [18] capitalizes on the Swiss' public interest in dialectology. It provides functionality that allows users (1) to localize their own Swiss German dialect by indicating – i.e. listening to pre-canned recordings and then tapping on the screen – their dialectal pronunciation of 16 tokens, i.e. words, and (2) to articulate and anonymously record these 16 tokens in their dialect. Data used in the current study stem from this second function.

The user interface (UI) prompts the users to indicate their dialect (possible localities are those used in [19]), age, and gender (Figure 1, left panel) before they proceed to the recording instructions screen (Figure 1, right panel), see Figure 1.

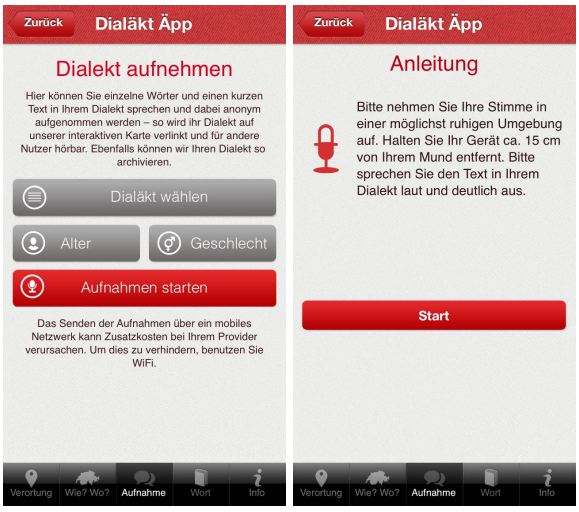

Figure 1: *UI for dialect, age, and gender selection (left panel) and recording instructions (right panel).*

10.21437/SpeechProsody.2014-31

The right panel in Figure 1 reads: "Please record your voice in as quiet an environment as possible. Keep an approximate distance of about 15 cm between your device and your lips. Please articulate the text loudly and clearly in your own dialectal pronunciation". Next, the user articulates and records the token shown on the screen (see Figure 2, left panel). The 16 tokens in this recording function are the same as those used for the localization function. Once the recordings are finished they are anonymously uploaded on our servers where each audio file is given a unique ID. Following the upload, users can navigate to an interactive map of Switzerland where they can listen to their own recordings and those of other users (Figure 2, right panel, green and purple pins).

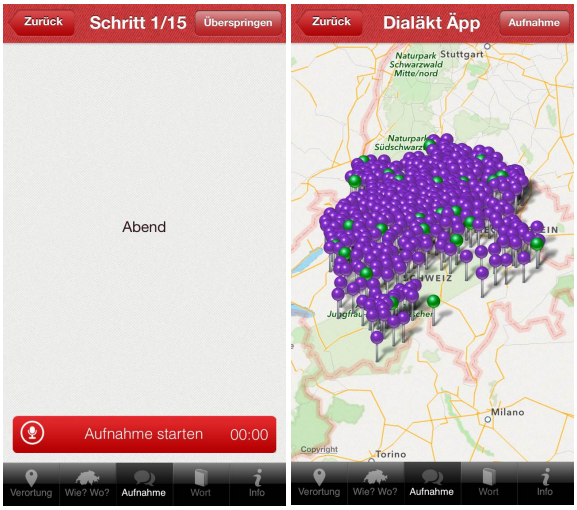

Figure 2: *UI for token recording (left panel) and audio playback map of one's own and other users' recordings (right panel).*

In Switzerland, 'Dialäkt Äpp' was the number one downloaded free app for iPhones after its release on March 22, 2013 [20]. It received major media attention and, so far, has >58,000 downloads. More than 2300 users from all over German-speaking Switzerland have uploaded voice recordings.

2.2. Subjects

Users who declared BE SwG (i.e. Bern city) and ZH SwG (i.e. Zurich city) as their local dialect served as subjects. In total there were 115 unique BE SwG speakers and 205 unique ZH SwG speakers. Not all speakers read all of the presented words, which is why the number of observations varies from token to token (cf. 2.3). On average speakers were 32-years-old, ranging between 4 years of age and 75 years of age, with 60% males and 40% females.

2.3. Material

We selected six out of a total of 16 'Dialäkt Äpp' tokens (cf. 2.1) for analysis of speaking rate. Selection criteria were that each token consisted of two syllables, given that we measured the temporal distance between adjacent vowel onsets (cf. 2.4). Half of our selected words further featured phonologically Middle High German long vowels or diphthongs while the other half featured underlying short vowels. The selected words with underlying long vowels were *Abend* 'evening', *Augen* 'eyes', and *fragen* 'to ask'; those with underlying short vowels *Donnerstag* 'Thursday', *heben* 'to lift', and *trinken* 'to drink'. Typically, these words are articulated as follows:

Long vowels/diphthongs:
Abend: BE SwG: [ˈɑːb̥ə], ZH SwG: [ˈɒːb̥ig̥]
Augen: BE SwG: [ˈɔug̥ə], ZH SwG: [ˈæug̥ə]
fragen: BE SwG: [ˈfrɑːg̥ə], ZH SwG: [ˈfrœːg̥ə]

Short vowels:
Donnerstag: BE SwG: [ˈd̥ɔnʃt̥i], ZH SwG: [ˈd̥unʃt̥ig]
heben: BE SwG: [ˈlʏp͡fə], ZH SwG: [ˈlup͡fə]
trinken: BE SwG: [ˈtrɪŋk͡xə], ZH SwG: [ˈtriŋk͡xə]

Table 1 presents the total number of observations, i.e. speakers, with figures on gender.

	ZH SwG	BE SwG
Abend 'evening'	n=188 (114m, 74f)	n=100 (58m, 42f)
fragen 'to ask'	n=193 (118m, 75f)	n=103 (64m, 39f)
Augen 'eyes'	n=186 (113m, 73f)	n=96 (60m, 36f)
Donnerstag 'Thursday'	n=186 (114m, 72f)	n=100 (63m, 37f)
heben 'to lift'	n=194 (118m, 76f)	n=104 (64m, 40f)
trinken 'to drink'	n=199 (120m, 79f)	n=105 (66m, 39f)

Table 1: *Summary of the number of total observations, i.e. speakers.*

The majority of recordings were usable, i.e. demonstrated little background noise interference nor were the speakers goofing off. Instances of unfavorable audio quality or otherwise unusable material were disregarded from the analyses (percentage of discarded tokens: approximately 20%).

2.4. Procedure

There are various approaches to measuring speaking rate. Most commonly one measures a linguistic unit per second (words, syllables, segments, consonantal intervals, vocalic intervals; cf. [21]). Since our corpus contains words that exhibit cross-dialectal differences in syllable structure (e.g. *Abend*: BE SwG V.CV [ˈɑːb̥ə] vs. ZH SwG V.CVC [ˈɒːb̥ig̥] or *Donnerstag*: BE SwG CVC.CCV [ˈd̥ɔnʃt̥i], vs. ZH SwG CVC.CCVC [ˈd̥unʃt̥ig]), we refrained from applying conventional speaking rate measures such as number of syllables per second, and instead measured the temporal duration between the two vowel onsets. In theory, this is motivated by [22]'s findings that vowel onsets represent perceptually prominent centers of a syllable. We call this measure of vowel-onset-to-vowel-onset duration *durVonVon*. Figure 3 schematically shows the measurement technique applied in the present study

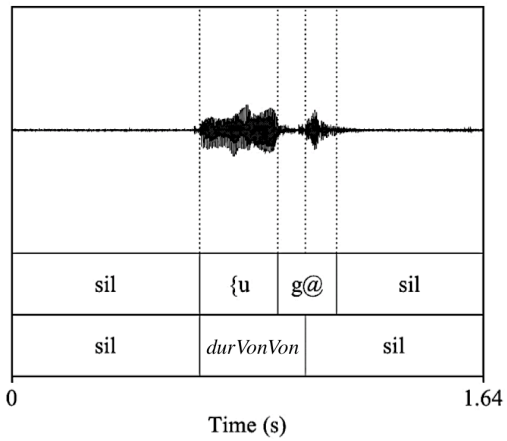

Figure 3: *Schematic of vowel-onset-to-vowel-onset measurement (2nd tier).*

Figure 3 shows the oscillogram of a ZH SwG speaker articulating the token *Augen* as [ˈæugə] (cf. 1st tier). The 2nd tier shows the boundaries placed at the vowel onsets. 'sil' indicates silence. Altogether there were 2920 measurement points (1460 intervals). Temporal duration between these two vowel onsets was measured in Praat [23].

3. Results

3.1. Statistical analyses

All data were analyzed using R [24] and the R packages *lme4* [25] and *languageR* [26, 27]. If not indicated otherwise, we analyzed data using linear mixed effect models (LMEs). Normality was checked by visual inspection of quantile plots. *Dialect*, *gender*, and *vowel type* were treated as fixed effects, *token* and *age* as random effects. Effects were tested by model comparison between a full model, in which the factor in question is entered as either a fixed or a random effect, and a reduced model without this effect. p-Values were obtained by comparing the results from the two models using ANOVAs. For the assessment of the relative goodness of fit, we report *AIC* (Akaike Information Criterion) values that decrease with goodness of fit. Only p-values that are considered significant at the α= 0.05 level are reported.

3.2. Overall regional differences

Figure 4 shows the boxplots representing the two dialects' *durVonVon* data.

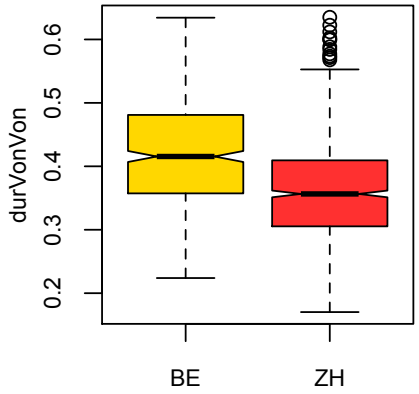

Figure 4: *Boxplots of the dialects' durVonVon.*

The yellow boxplot indicates the values for BE SwG, the red boxplot those of ZH SwG. Visually, the two boxes' notches do not overlap, which can be taken as strong evidence that their medians differ. The comparison between the full and reduced models showed a significant difference for dialect, with the full model exhibiting an increased goodness of fit (BE SwG M=.42, SD=.08; ZH SwG M=.36, SD=.08; p<.0001; *AIC*=-3806). There was thus a significant difference in *durVonVon* between the two dialects. BE SwG speakers showed longer *durVonVon* intervals than ZH SwG speakers.

3.3. Cross-gender differences by dialect

Figure 5 shows the distribution of *durVonVon* across *dialect* and *gender*.

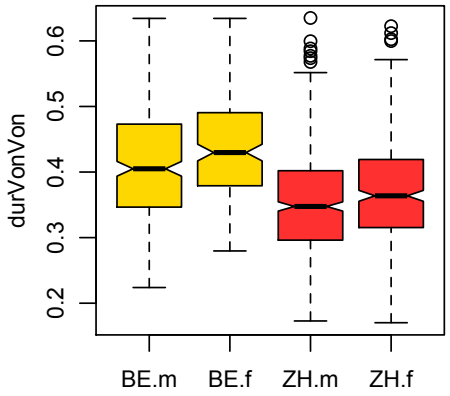

Figure 5: *Boxplots of durVonVon across dialect and gender.*

The differences between the genders were significant in both dialects (Bonferroni adjusted for *gender*, α=0.025; both p<.0001; BE SwG: *AIC*=-1244, ZH SwG: *AIC*=-2598). The boxplots in Figure 5 indicate that for both BE SwG and ZH SwG, females demonstrated significantly longer vowel-onset-to-vowel-onset durations. There was no significant interaction of *dialect***gender* (p=.20; *AIC*=-3799).

4. Discussion

Based on a controlled set of words spoken by a large number of speakers, the current study found that BE and ZH SwG strongly differ in terms of speaking rate. For the sake of

illustration, let us extrapolate these results to a more realistic scenario. Say a BE SwG and a ZH SwG speaker read Aesop's fable "The North Wind and the Sun". In Zurich German, the fable consists of 129 syllables, i.e. approximately 128 vowel-onset-to-vowel-onset intervals (cf. [28]). Based on raw findings of the current study – disregarding contextual factors such as phrase-final lengthening – the BE SwG speaker would need 54 seconds to read the text while the ZH SwG speaker would only take 46 seconds.

This finding is intriguing in a number of ways. [10] notes that BE SwG speakers speak more slowly particularly because they exhibit more distinct phrase-final lengthening. Results of the current study show, however, that the two dialects strongly vary from one another in speaking rate irrespective of such contextual factors. Our findings are unique in just this sense: the temporal information contained in a few words alone is already sufficient to tell apart the two dialects (cf. Figure 4) – regardless of contextual factors such as phrase-final lengthening. In future studies it would be interesting to test whether vowel length differences are the major influential factor. Moreover, it will be interesting to examine if we find these between-dialect differences in each of the 6 tokens individually.

Concerning gender differences in speaking rate. The result that female speakers articulate more slowly than males is in line with previous studies on cross-gender differences in speaking rate on British English dialects [29] and on American English dialects [30, 31]. It is further in line with [32] who shows that in German, durations of female vowels are systematically longer than durations of male vowels.

Several questions remain unanswered at present. Is it conceivable that listeners can tell apart the two dialects based solely on time domain information in perception experiments? To answer this question one would require tokens that are identical in segmental, syllabic, and prosodic structure, which (deliberately) does not apply to the tokens used in 'Dialäkt Äpp'. One could, however, use 'Dialäkt Äpp' tokens that are identical in syllabic structure and delexicalize them (e.g. *sasasa*-delexicalization, where every consonantal interval is replaced with a pre-recorded [s] and every vocalic interval with a pre-recorded [a] (cf. [33]).

Speaking rate represents only one of countless areas of application in speech prosody where crowdsourcing is useful. In the present region-wide 'Dialäkt Äpp' corpus, more than 2300 speakers have uploaded voice recordings, and in most cases speakers recorded all 16 words. This amounts to approximately 36,000 voice recordings. In future studies we will further examine fundamental frequency distributions, temporal, stress and intonational patterns, and generate vowel plots. These phenomena can be explored in multiple dimensions, enabling us to test for effects of speaker, age, gender, locality or region. Crowdsourcing applications for American English, German regional varieties, and British English are currently being developed, inspired by the 'Dialäkt Äpp' framework. The acoustic data crowdsourced through the 'Dialäkt Äpp' is further used to train an automatic speech recognition system. This system will be part of a follow-up smartphone application [34] that will perform dialect localization based on spoken language input.

5. Conclusion

Based on crowdsourced utterances from a large number of speakers, results of the current study corroborate previous impressionistic and empirical observations that ZH SwG is fast and BE SwG slowly spoken [6, 7, 8, 9, 10]. Results of the current study revealed that regional differences in speaking rate are prevalent on the basis of a few words alone. We further showed that female speakers articulate more slowly than male speakers, which is in line with other research findings.

6. Acknowledgments

We thank Daniel Wanitsch for server-side technical assistance and audio data extraction and Ingrid Hove for database maintenance. We are indebted to 65 backers who made 'Dialäkt Äpp' possible through crowdfunding. Thank you!

7. References

[1] BFS = Bundesamt für Statistik, Statistisches Lexikon der Schweiz. Personenverkehr: Entwicklung der Tagesmobilität, 2005, http://www.bfs.admin.ch/.

[2] Werlen, I., "Zur Sprachsituation in der Schweiz mit besonderer Berücksichtigung der Diglossie in der Deutschschweiz", Bulletin VALS-ASLA (Vereinigung für angewandte Linguistik in der Schweiz), 79:1-30, 2004.

[3] Christen, H., "Was Dialektbezeichnungen und Dialektattribuierungen über alltagsweltliche Konzeptualisierungen sprachlicher Heterogenität verraten", in C. Anders, M. Hundt and A. Lasch [Eds], "Perceptual dialectology". Neue Wege der Dialektologie, 269-290, Berlin/New York: de Gruyter, 2010.

[4] Hotzernköcherle, R., Die Sprachlandschaften der deutschen Schweiz. Ed. by N. Bigler, R. Schläpfer, Aarau: Sauerländer, 1984.

[5] Schwarzenbach, R., Die Stellung der Mundart in der deutschsprachigen Schweiz. Studien zum Sprachgebrauch der Gegenwart (= Beiträge zur schweizerdeutschen Mundartforschung XVII), Frauenfeld: Huber, 1969.

[6] Ris, R., "Innerethik der deutschen Schweiz", in P. Hugger [Ed], Handbuch der schweizerischen Volkskultur, vol. II, 749–766, Zürich: Offizin, 1992.

[7] Berthele, R., "Wie sieht das Berndeutsche so ungefähr aus? Über den Nutzen von Visualisierungen für die kognitive Laienlinguistik", in H. Klausmann [Ed], Raumstrukturen im Alemannischen. Beiträge der 15. Arbeitstagung zur alemannischen Dialektologie, Schloss Hofen (Vorarlberg) vom 19.-21.9.2005 (= Schriften der VLB 15), 163–176, Graz-Feldkirch: Neugebauer, 2006.

[8] Werlen, I., "Zur Einschätzung von schweizerdeutschen Dialekten", in I. Werlen [Ed], Probleme der schweizerdeutschen Dialektologie. 2. Kolloquium der Schweizerischen Geisteswissenschaftlichen Gesellschaft 1978, 195–257, Fribourg, 1985.

[9] Leemann, A., Swiss German Intonation Patterns, Amsterdam/Philadelphia: Benjamins, 2012.

[10] Leemann, A. and Siebenhaar, B., "Statistical Modeling of F0 and Timing of Swiss German Dialects", Proceedings of Speech Prosody, 2010.

[11] crowdsourcing, Merriam-Webster.com, Retrieved December 14, 2013, http://www.merriam-webster.com/dictionary/crowdsourcing

[12] http://blog.faberacoustical.com/2009/ios/iphone/iphone-microphone-frequency-response-comparison/.

[13] De Decker, P. and Nycz, J., "For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?," University of Pennsylvania Working Papers in Linguistics 17(2):51-59, 2011.

[14] Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. and LeBeau, M., "Building transcribed speech corpora quickly and cheaply for many languages", Proceedings of Interspeech 26.-30.10.2010, Makuhari, Chiba, Japan: 1914-1917.

[15] de Vries, N., Davel, M. H., Badenhorst, J., Basson, W. D., de Wet, F., Barnard, E. and de Waal, A., "A smartphone-based ASR data collection tool for under-resourced languages", Speech Communication, 56: 119-131, 2014.
[16] Ma! Iwaidja, https://itunes.apple.com/au/app/ma-iwaidja/id557824618?mt=8.
[17] Hanke, F. R., Byrd, S., "Large-scale text collection for unwritten languages", International Joint Conference on Natural Language Processing, 1134-1138; http://lp20.org/aikuma/, 2013.
[18] Leemann, A., and Kolly, M.-J., Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8, 2013.
[19] SDS = Sprachatlas der deutschen Schweiz, Bern (I-VI), Basel: Francke (VII-VIII), 1962-2003.
[20] http://www.appannie.com/,
[21] Roach P., "Myth 18: Some languages are spoken more quickly than others", in L. Bauer and P. Trudgill [Ed], Language Myths, 150–158, London: Penguin 1998.
[22] Allen, G. D., "The location of rhythmic stress beats in English: An experimental study I.", Language and Speech, 15:72–100, 1972.
[23] Boersma, P. and Weenink D., Praat: doing phonetics by computer, www.praat.org, 2013.
[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Version 3.0.0. http://www.R-project.org, 2013.
[25] Bates, D. M. and Maechler, M., lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-32, 2009.
[26] Baayen, R. H., Analyzing Linguistic Data: A Practical introduction to statistics using R, CUP, Cambridge, 2008.
[27] Baayen, R. H., languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics using R", R package version 0.955, 2009.
[28] Fleischer, J. and Schmid, S., "Zurich German", Journal of the International Phonetic Association, 36(2):243-253, 2006.
[29] Whiteside, S. P., "Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker gender differences", Journal of the International Phonetic Association, 26:23-40, 1996.
[30] Jacewicz, E., Fox, R. A., O'Neill, C. and Salmons, J., "Articulation rate across dialect, age, and gender", Language Variation and Change, 21:233-256, 2009
[31] Byrd, D., "Preliminary results on speaker-dependent variation in the TIMIT database", Journal of the Acoustical Society of America, 92(1):593-596, 1992.
[32] Simpson, A. P., „Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung", Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 33, 1998.
[33] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: A study based on speech resynthesis", Journal of the Acoustical Society of America, 105(1):512–521, 1999.
[34] Swiss Voice App, www.voiceapp.ch