Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Identifying a speaker's regional origin: the role of temporal information

Adrian Leemann[1], *Marie-José Kolly*[2], *Francis Nolan*[1]

[1]Phonetics Laboratory, DTAL, University of Cambridge
[2]Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich
{al764|fjn1}@cam.ac.uk, marie-jose.kolly@uzh.ch

Abstract

Previous studies have revealed that, depending on the language, listeners can identify speakers' dialects quite well. The role of segments and prosody in this task is largely unknown, however. In a between-subjects design, we tested a total of 30 listeners in two conditions: in the unmorphed condition, listeners heard original sentences from two Swiss German dialects; in the duration morphed condition, listeners heard the same material, but with syllable durations exchanged between the two dialects. In a two-alternative forced choice design, subjects judged the regional origin of the stimuli heard. Results revealed near perfect identification performance for both conditions, thus underlining the overriding dominance of segmental cues in dialect identification tasks. The findings reported are pertinent to forensic phonetics, enhancing the diagnostic power of naïve and expert listeners' claims about suspect speakers' voices.

Index Terms: temporal cues, dialect, cue weighing, forensic speaker comparison, forensic phonetics, Swiss German

1. Introduction

The speech stream consists of segmental and suprasegmental (i.e. prosodic) features, both of which can be successfully exploited for dialect identification. [1, 2] reported that naïve Swiss German (SwG) listeners, for example, can identify a speaker's dialect at rates of 86% and 74% respectively. [3, 4] showed that Dutch listeners correctly identified Dutch dialects at a rate of 60%. In other languages naïve listeners may perform somewhat less well: [5] report identification rates of 30–50% for American and British English dialects, and [6, 7] report similarly poor recognition rates for German dialects.

What these accounts did not address, however, is the role of segmental and suprasegmental cues in this identification process. Existing research in this field is mostly exploratory: in a pilot study, [8] created stimuli with multi-dimensional combinations of Indian and British English segmental and suprasegmental cues. Results showed that segmental cues seemed most diagnostic for the recognition of a variety, followed by f0 and temporal information. [2] examined the perceptual saliency of different linguistic cues: in an eight-alternative forced-choice experiment using SwG stimuli, listeners were asked to note down which features they judged to be most salient for dialect recognition. Most listeners reported dialect-specific lexical items as being most diagnostic. Segmental features such as differences in /r/ realization were also considered important diagnostics. Suprasegmental features were classified as less indicative by naïve listeners. [3] examined the degree to which the amount of segmental information in the signal affects dialect identification for Dutch and English, using monotonization and low-pass filtering. They report a tendency for dialect identification scores to decrease the less segmental information there is in the signal. It remains unclear, however, exactly which parameters of prosody – duration, f0, or intensity – contribute to dialect identification. Speech manipulation in the sense of morphing prosodic parameters of one variety onto the segmental string of another – as currently applied in research on L2 accentedness and intelligibility [9, 10, 11, 12] – would allow these parameters to be disentangled and their respective contribution to dialect identification to be evaluated.

In this proof-of-concept study, we examine the role of segmental and suprasegmental cues in the identification of a speaker's dialect. In particular, we examine the role of temporal cues in the identification process. This question is approached by means of an experiment in which segmental and temporal cues are played off against each other. This can be achieved by manipulating the speech signal so that the durational features of dialect X will be morphed onto the segments of dialect Y and vice versa [cf. 13]. These manipulated stimuli are then presented to naïve listeners who are asked to judge whether the stimulus they hear is from a speaker of dialect X or Y. On the basis of the survey of the scarce existing literature, it is conceivable that segmental features will override temporal cues as more important cues in the identification of a dialect. The findings are pertinent to the reliability and diagnostic power of naïve and expert listeners' claims about a suspect speaker's dialect in forensic speaker comparison.

2. Methods

In a between-subject design we tested 30 listeners: 15 in the unmorphed condition and 15 in the duration morphed condition. The between-subject design is preferred given a potential learning curve on the part of the listeners when exposed to morphed stimuli: [14] report that distorted speech becomes more intelligible with experience.

2.1. Dialects

Bern (BE) and Valais (VS) SwG were chosen. Previous research has shown distinct differences in suprasegmentals [1, 15, 16] and segmentals [17, 18]. As for differences in the time domain, [15] reported faster speaking rates for VS SwG. [16] reported significant rhythmical differences (%V [19] and VarcoV [20]).

2.2. Speakers

12 speakers provided the sentence material for this study (6 VS SwG speakers (3f/3m); 6 BE SwG speakers (3f/3m)).

10.21437/SpeechProsody.2016-222

Speakers claimed to speak the respective dialects on a daily basis and were aged between 18 and 33.

2.3. Material

To circumvent the question of how to create sentences that balance salient dialectal features, we randomly selected 10 from the 336 sentences of the Bamford-Kowal-Bench (BKB) corpus [21]. We made BE and VS SwG transcripts of these 10 sentences. 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). Sentences were manually labeled in Praat [22]. We performed a syllable-based segmentation, following sonority hierarchy principles [23, 24]. Before durations were morphed, we calculated the means of the males' and females' f0s (males: 127.5 Hz, females: 216.1 Hz). We then normalized both males' and females' sentences to these f0 values, using Praat's 'change gender' function (which adjusts f0 medians, leaving f0 variability intact). On the one hand, this prevents listeners from identifying speakers based on their f0s; on the other, it renders even baseline – unmorphed – stimuli synthetic. This is important as it means both groups of subjects heard artificially manipulated speech. For duration morphing, the syllable durations of sentence 01 of speaker BE01, for example, were morphed onto the syllables of sentence 01 of speaker VS01. Analogously, the syllable durations of sentence 01 of speaker VS01 were morphed onto the syllables of sentence 01 of speaker BE01, see Figure 1.

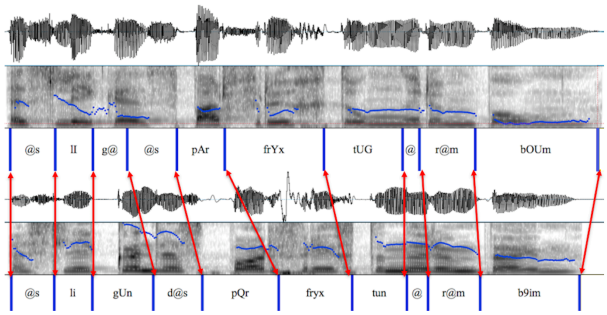

Figure 1: *Duration morphing on the syllable level.*

The bottom panel of Figure 1 shows a sentence from a VS speaker, segmented on the syllable level. The top panel shows the same sentence from a BE speaker. Both sentences feature the same syllable counts. We used a duration morphing script from [25]. Speakers were paired in such a way that they exhibited similar speaking rates (as established by the number of syllables / sec). This, too, counteracts potential artefacts in the morphed stimuli – the less stretching between the donor and the recipient, the more natural the stimulus sounds. Each of the 12 speakers provided 10 sentences, which amounted to 120 sentences per condition.

2.4. Subjects

30 listeners (15 / condition) were recruited at the University of Zurich (unmorphed: 11f, 4m; duration morphed: 14f, 1m). Listeners were all speakers of Zurich German and students at the University of Zurich. Mean age for the unmorphed speech condition was 26 (SD=6.2), mean age for the duration morphed condition was 23 (SD=5). It can be assumed that all listeners were relatively equally exposed to VS and BE SwG either through personal contacts or through the media.

2.5. Procedure

Listeners were familiarized with the experiment interface and the types of stimuli used. In the main part of the experiment they heard 120 stimuli of roughly 5–8 seconds per stimulus; each stimulus was played once and heard through high-quality headphones. The order of the stimuli was randomized separately for each subject. Following stimulus presentation, listeners decided whether this sentence was BE or VS SwG by clicking on the corresponding button on a laptop screen, using the experiment interface shown in Figure 2.

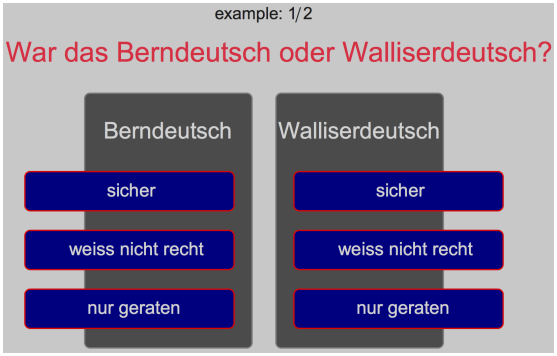

Figure 2: *Experiment interface used. To give the response, listeners clicked on the respective button.*

Listeners were able to indicate their confidence level on a 3-point scale (ranging from 1='certain' to 3='only guessed').

2.6. Data analysis

To examine production differences in the time domain, we applied three temporal metrics on the dialects: speaking rate, the rate-normalized standard deviation of syllable interval durations *VarcoSyl* [26], and the rate-normalized average differences between consecutive syllable interval durations *nPVISyl* [26]. We used a signal detection theory framework to assess identification performance, calculating A' (with BE SwG as the signal and VS SwG as the noise). A' is a non-parametric measure of sensitivity used to eliminate response bias [27]. It ranges from 0 to 1 (chance level at A'=0.5), 1 indicating perfect sensitivity (i.e. perfect identification of both dialects) and 0 indicating no sensitivity. As a measure of listeners' response bias we report the non-parametric B''_D, ranging from -1 (bias towards responding BE SwG) to +1 (bias towards responding VS SwG), where 0 indicates no bias. We further report the percentage of correct responses when investigating the effect of *dialect*. All data were analyzed using R [28] and the R package lme4 [29]. If not indicated otherwise, we analyzed data using linear mixed effect models. *Dialect* and *gender* were treated as fixed effects, *item* and *speaker* as random effects. We included *by-speaker* random slopes on the effect of *dialect*.

3. Results

3.1. Production

Differences between the two dialects in speaking rate – as captured by the number of syllables per second – were not

significant. Descriptively, however, BE SwG (M=4.3, SD=.75) speakers articulated more slowly than VS speakers (M=4.7, SD=.78). As for the rhythm metrics applied, we found differences between the two dialects in *VarcoSyl*, as shown in Figure 3.

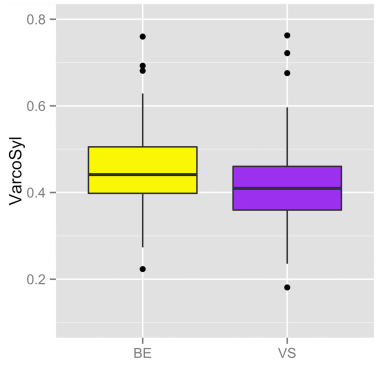

Figure 3: *Boxplots of VarcoSyl by dialect.*

VS speakers (M=.42, SD=.11) showed less syllable duration interval variation than BE speakers (M=.48, SD=.12). This difference was significant (p=.0015, AIC=-260).

3.2. Perception

3.2.1. Effect of condition

Figure 4 shows the boxplots of the dialects' *A′*. The left-hand panel shows the full *A′* scale from 0 to 1; on the right-hand side the scale is truncated to zoom in on the effect.

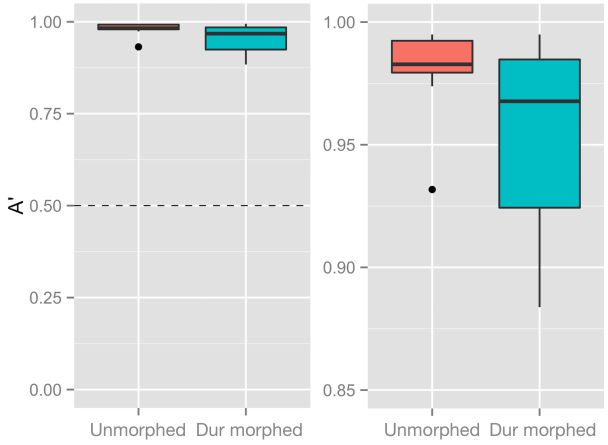

Figure 4: *Boxplots of A′ by condition. Left panel: full representation; right panel: y-axis truncated from .85 to1.0.*

In unmorphed speech, mean sensitivity was at .98 (SD=.02); when durations were morphed, mean sensitivity dropped to .95 (SD=.04). A one-sample t-test showed that *A′* was significantly above chance (.05) for unmorphed speech (t(14)=118.4, p<.0001*) and duration morphed speech (t(14)=48.9, p<.0001*). We obtained a distinct ceiling effect. There further was a significant effect of *condition*: listeners' sensitivity was significantly higher in unmorphed speech (linear model F(1, 29)=7.4, p=.011). Listeners' confidence did not differ between the unmorphed (M=1.13, SD=.08) and the duration morphed condition, however (M=1.20, SD=.24).

3.2.2. Between-dialect differences

We further computed the percentage of correct responses for the BE and VS SwG stimuli separately. Figure 5 shows boxplots of *%correct* crossed against *dialect* and *condition*.

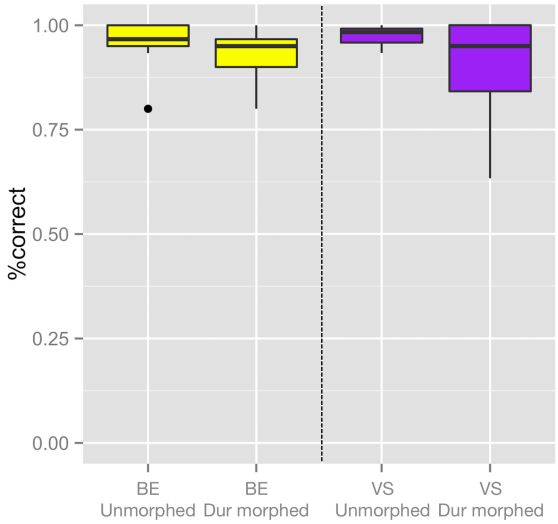

Figure 5: *Boxplots of %correct crossed against dialect and condition.*

For both dialects, unmorphed speech showed a higher *%correct* (BE: M=.96, SD=.05; VS: M=.97, SD=.02) than duration morphed speech (BE: M=.96, SD=.06; VS: M=.92, SD=.09). The variability in responses for the VS duration morphed condition was much wider than in VS unmorphed condition. These differences were not significant, however (Bonferroni-adj. α). Further, listeners' confidence scores did not differ between the VS (M=1.16, SD=0.21) and BE samples (M=1.18, SD=.17). For VS SwG, confidence scores slightly decreased from the unmorphed speech (M=1.11, SD=.10) to the duration morphed condition (M=1.21, SD=.27).

3.2.3. Bias

To examine whether listeners had a significant preference for responding to either of the dialects, we calculated B''_D for every condition (cf. 2.6). There was no clear preference in both conditions (unmorphed: M=.07, SD=.42; morphed: M=.05, SD=.6), as shown in Figure 6.

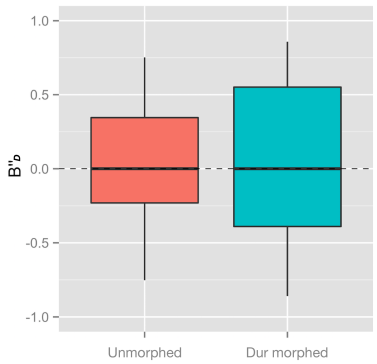

Figure 6: *Boxplots of* B''_D.

The spread of the responses is wider in duration morphed speech. This indicates a greater variability in the listeners' biases towards either dialect.

4. Discussion & Conclusion

Results of the production data revealed between-dialect differences for BE and VS SwG in the time domain, which legitimizes the choice of the two dialects for our research question. VS SwG speakers articulated faster, a trend that has has been previously reported [15]. The dialects further differed in the rhythm metric *VarcoSyl*. VS SwG's lower *VarcoSyl* may be explained by the fact that this dialect tends to exhibit non-reduced syllables, i.e. retaining full vowels, in unstressed environments [18]. This feature – not found in BE SwG – may lower temporal variability in VS SwG.

Results of the perception test indicate that (i) listeners are very sensitive to discriminating between the two dialects, (ii) durational cues seem to occupy a marginal – albeit statistically significant – role when it comes to identifying a speaker's dialect, and (iii) listeners' responses were more variable in the duration condition.

(i) is consistent with data reported in [1, 2], which showed that SwG listeners perform above chance level at identifying BE and VS SwG. Naïve SwG listeners are well aware of dialectal variation; this knowledge allows them to make relatively accurate identification judgments.

(ii) seems to match the observations reported in [2, 3, 8]: all three studies underlined the critical role of segmental information in dialect identification. Even if syllable durations are swapped between two dialects, listeners are able to identify the dialect almost perfectly. However, despite the very high level of identification performance, dialect identification deteriorates significantly when durations are morphed.

(iii) Figure 4 suggests more variation in listener response for the morphed stimuli. The rightmost boxplot in Figure 5 reveals that this variation particularly stems from the responses given to the VS samples in the duration morphed condition. Listeners' confidence scores may help explain this variability: descriptively, there was a decrease in confidence from the unmorphed to the duration morphed condition, particularly for the VS samples. Moreover, listener bias was more variable in the duration morphed condition than in the unmorphed condition (cf. Figure 6). These findings suggest that response behavior tends to become more variable in duration morphed speech, which seems to be caused by the mismatch in segmental and temporal information.

These findings, while preliminary, help us understand the role of segmental and durational cues in the identification of speakers' dialects. The task of dialect identification is exploited in forensic phonetic research. In forensic speaker comparison (FSC), experts compare speech in criminal and suspect recordings to evaluate the evidence under competing prosecution and defense hypotheses. A speaker's dialect significantly contributes to the uniqueness of a speaker's voice, which is why dialectal information is typically incorporated in the analysis of criminal and suspect recordings [30, 31]. In FSC, not only experts evaluate speech material but also witnesses or victims who might have heard the criminal's voice. In the course of the investigation, these so-called '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 the speaker s/he believes was heard at the crime scene [32]. When earwitnesses are asked to describe the voice they heard, dialect forms a central part in that description [33]. The results presented in this study contribute to the increasing knowledge base of dialect identification. In general, therefore, these findings may enhance the reliability of claims (both expert and naïve) about a suspect speaker's dialect, potentially making these claims diagnostically more conclusive.

To develop a full picture of the role of segmental and suprasegmental features in the identification of a speaker's dialect, the current experiment needs to be expanded:

(a) f0 information will also be morphed and added as a separate, third, condition. This will allow for an assessment of the role of intonation in dialect identification tasks.

(b) There is abundant room for progress in determining exactly which segmental cues in the material listeners perceived as diagnostic for BE or VS SwG. More in-depth analyses of the sentences used will provide these insights.

(c) Forensic crime material frequently includes background noise (e.g. bar noise, traffic noise, reverberation, hotel lobby noise etc.). Noise has detrimental effects on speech perception. This raises the question to what degree naïve earwitnesses and experts are able to assess a suspect's dialect in noisy conditions: speech in noise conditions need thus to be added to the experiment design. Moreover, will the contribution of segmental and suprasegmental information in the identification of a speaker's dialect be the same in adverse listening conditions? Recent research reports that background noise can trigger a re-ranking of acoustic cues to linguistic categories, where "secondary" cues in laboratory settings become the only accessible, and thus "primary", cues in noisy environments [34]. The kinds of coping strategies that listeners use when exposed to speech in adverse listening conditions can reveal mechanisms that may not come to light when speech is observed in laboratory conditions.

An expansion of the experiment incorporating the points mentioned here will further heighten the reliability and diagnostic power of naïve and expert listeners' claims about a suspect speaker's dialect.

5. Acknowledgements

The authors thank Volker Dellwo for helpful comments and technical assistance in the experiment design. This research is funded by the *Swiss National Science Foundation*, grant Nr. P300P1_151210, http://p3.snf.ch/project-151210.

6. References

[1] A. Leemann and B. Siebenhaar, "Perception of dialectal prosody," Proceedings of Interspeech, Brisbane 22.-26.09.2008: 524-527.

[2] M. Guntern, "Erkennen von Dialekten anhand von gesprochenem Schweizerhochdeutsch," Zeitschrift für Dialektologie und Linguistik, vol. 78, no. 2, pp. 155-187, 2011.

[3] R. Van Bezooijen and C. Gooskens, "Identification of Language Varieties: The Contribution of Different Linguistic Levels", Journal of Language and Social Psychology, vol. 18, no. 31, pp. 31-48, 1999.

[4] R. Van Bezooijen and J. Ytsma, "Accents of Dutch: Personality impression, divergence, and identifiability," Belgian Journal of Linguistics, vol. 25, pp. 105-129, 1999.

[5] C. G. Clopper and D. Pisoni, "Perception of dialect variation," in: D. Pisoni and R. E. Remez (Eds.), The Handbook of Speech Perception. Oxford: Blackwell, pp. 313–337, 2005.

[6] R. Kehrein, A. Lameli, C. Purschke, "Stimuluseffekte und Sprachraumkonzepte," in: C. Anders, M. Hundt, and A. Lasch (Eds.), "Perceptual dialectology". Neue Wege der Dialektologie. Berlin/New York: de Gruyter, pp. 351–384, 2010.

[7] R. Kehrein, "Wen man nicht alles für einen Sachsen halt?! Oder: Zur Aktivierung von Sprachraumkonzepten durch Vorleseaussprache," in: K. Jakob, R. Hünecke (Eds.), Die obersächsische Sprachlandschaft in Geschichte und Gegenwart. Heidelberg: Winter, pp. 223-263, 2012.

[8] R. Fuchs. "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, S. Herment (Eds.), Prosody and language in contact: L2 acquisition, attrition and languages in multilingual situations. Frankfurt: Springer, 121-146, 2015.

[9] M. J. Munro and T. M. Derwing, "Second language accent and pronunciation teaching: A research based approach," TESOL Quarterly, vol. 39, pp. 379–398, 2005.

[10] B. Vieru-Dimulescu and P. Boula de Mareüil, "Contribution of prosody to the perception of a foreign accent: a study based on Spanish/Italian modified speech," Proceedings of ISCA Workshop on Plasticity in Speech Perception, London 15.-17.06.2005.

[11] S. Winters, M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of F0 and duration," Speech Communication, vol. 55, pp. 486-507, 2013.

[12] C. Ulbrich, "German pitches in English: Production and perception of cross-varietal differences in L2," Bilingualism: Language and Cognition, vol. 16, no. 2, pp. 397-419, 2013.

[13] J. Vaissière and P. Boula de Mareüil, "Identifying a language or an accent: from segments to prosody," Paper read at MIDL, Paris 29.-30.11.2004.

[14] M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan, "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," Journal of Experimental Psychology, vol. 134, no. 22, pp. 222–241, 2005.

[15] A. Leemann, Swiss German Intonation Patterns. Amsterdam: John Benjamins, 2012 (=Studies in Language Variation, vol. 10).

[16] A. Leemann, V. Dellwo, M.J. Kolly, and S. Schmid, "Rhythmic variability in Swiss German dialects," Proceedings of Speech Prosody, Shanghai 22.-25.05.2012.

[17] W. Marti, Berndeutsch-Grammatik. Bern: Francke, 1985.

[18] E. Wipf, Die Mundart von Visperterminen im Valais. Frauenfeld: Huber, 1910.

[19] R. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," Cognition, vol. 73, pp. 265-292, 1999.

[20] V. Dellwo, "Rhythm and Speech Rate: A Variation Coefficient for deltaC", in: P. Karnowski and I. Szigeti (Eds.), Language and language-processing. Frankfurt am Main: Peter Lang, pp. 231-241, 2006.

[21] J. Bench, A. Kowal, and J. Bamford,. "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," British Journal of Audiology, vol. 13, pp. 108-112, 1979.

[22] P. Boersma and D. Weenink, Praat: Doing phonetics by computer. http://www.praat.org/, 2013.

[23] E. Sievers, (1881). Grundzüge der Phonetik. Leipzig: Breitkopf und Hartel, 1881.

[24] B. Siebenhaar, "Phonological and phonetic considerations for a classification of Swiss German dialect as a word language or syllable language," in: R. Javier Caro and R. Szczepaniak (Eds.), Syllable and Word Languages. Berlin: de Gruyter: pp. 327–345, 2014 (= Linguae et litterae, vol. 40).

[25] P. Boula de Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," Phonetica, vol. 63, pp. 247–267, 2006.

[26] C. Lai, K. Evanini, and K. Zechner, "Applying Rhythm Metrics to Non-native Spontaneous Speech," in Proceedings of SLaTE 2013. Interspeech 2013 Satellite Workshop on Speech and Language Technology in Education, Grenoble 30-31.08.2013, pp. 159-163.

[27] D. M. Green and J. A. Swets, Signal detection theory and psychophysics. New York: Wiley, 1966.

[28] 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.

[29] D. M. Bates and M. Maechler, lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-32, 2009.

[30] M. Jessen, "Speaker Classification in Forensic Phonetics and Acoustics," in: C. Müller (Ed.), Speaker Classification, vol. 1, pp. 180-204, 2007

[31] O. Köster, R. Kehrein, K. Masthoff, and Y. H. Boubaker, "The tell-tale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians," Journal of Speech, Language and the Law, vol. 19, no. 1, pp. 51–71, 2012.

[32] F. Nolan and E. Grabe, "Preparing a voice lineup," Forensic Linguistics, vol. 3, no. 1, pp. 74-94, 1996.

[33] H. Hollien, Forensic Voice Identification. San Diego, CA: Academic Press, 2002.

[34] G. Parikh and P. C. Loizou, "The influence of noise on vowel and consonant cues," Journal of the Acoustical Society of America, vol. 118, no. 6, pp. 3874-3888, 2005.