Speaker idiosyncratic variability of intensity across syllables

He, Lei ; Dellwo, Volker

Speaker Idiosyncratic Variability of Intensity across Syllables

Lei He, Volker Dellwo

Phonetics Laboratory, University of Zurich, Switzerland

{lei.he, volker.dellwo}@uzh.ch

Abstract

This study explored speaker idiosyncrasy by measuring the syllabic intensity variability in the speech signal. Sixteen speakers of the TEVOID corpus, each producing 256 read sentences, were analyzed. Characteristics of intensity variability (average or peak) between syllables were measured either holistically (standard deviation of intensity changes between syllables) or locally (pairwise variability indices of intensity changes between syllables). The results indicated significant effects of the speakers in all the metrics, suggesting a potential application of the methods for speaker recognition, and in particular for forensic speaker comparison.

Index Terms: intensity variability, speaker idiosyncrasy

1. Introduction

Speech production is a complicated process that involves much neuromuscular programming to control the movements of articulators [1]. The motor control in speech, similar to other modes of human movements like human gain [2, 3], is highly individual, and it seems conceivable that such individual characteristics are reflected in the physical properties of the speech signal. Enlightened by the idiosyncratic temporal characteristics in human gait, the research team in our laboratory adopts a time-domain approach to voice identification. The widely used speech rhythm metrics [4, 5, 6, 7] were employed to find speaker individualities in the speech signal. [8] and [9] discovered that the percentages over which speech is vocalic (%V) and the percentage over which speech is voiced (%VO) showed fair success in detecting speaker idiosyncrasy with spontaneous speech. %VO also turned out to show speaker specific characteristics independent of the language in bilingual speakers [10]. Moreover, newly developed metrics (ΔPeak) also succeeded in finding speaker individualities [8, 9]. ΔPeak is calculated by taking the standard deviations of the intervals between syllabic amplitude peaks. Such measures are motivated by the idea that the combined movements of the articulators result in a temporal organization of amplitude envelope characteristics like syllabic peak points. This idea also motivates the present study in which we studied amplitude peak and syllabic intensity variability between speakers. Similar to temporal measures we previously found that such measures show language specific effects between English and Mandarin or L2 English by Mandarin natives [11]. In the present study we tested to what degree such measures reveal within-language variability as a function of speakers, if speaker specific controls of the articulators are responsible for the individual timing organization of speech.

2. Methods

2.1. The TEVOID corpus

The TEVOID (Temporal Voice Idiosyncrasy) corpus [8, 9] was constructed in the Phonetics Laboratory of the University of Zurich to study temporal variability in the speech signal. The speakers were all native speakers of Zurich German. This German variety shows little if any socio-economic variability, which could be a potential artifact in between-speaker variability of temporal characteristics [8]. Recordings of both read and spontaneous speech of 16 speakers are in the current corpus. All the recordings were digitized in a sound attenuated booth with the sampling rate of 44.1 kHz and a quantization depth of 16 bit. The read speech (256 sentences * 16 speakers = 4'096 sentences) was analyzed for the present study.

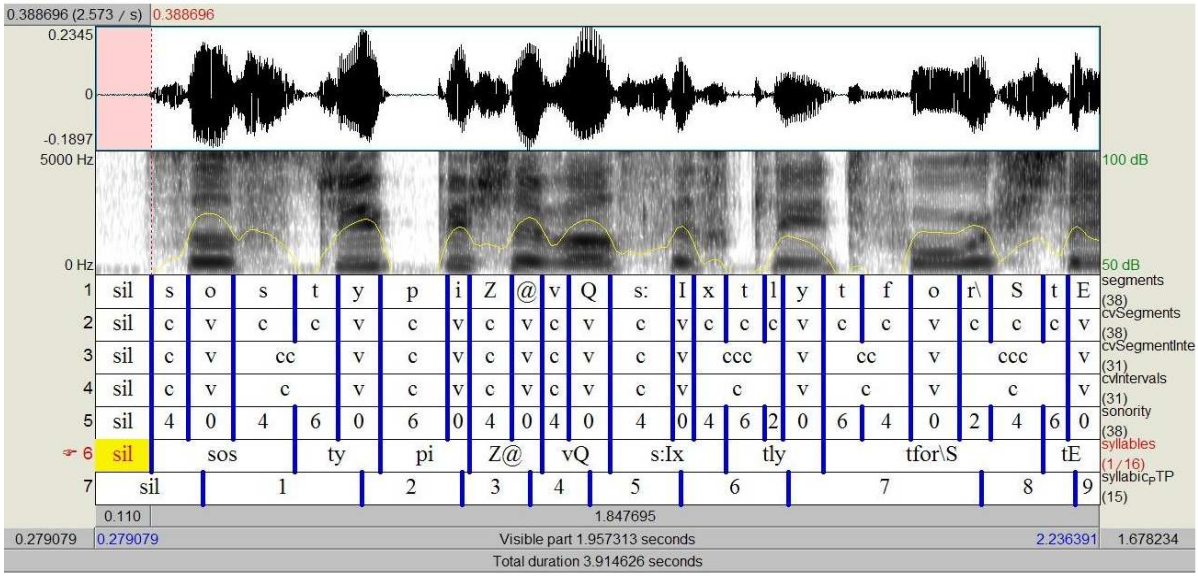

Figure 1. *A screenshot of the TEVOID annotations. Tier 6 and tier 7 mark the syllables and the corresponding syllable peaks in the utterance.*

Segments of the sound files were annotated manually, on the basis of which tiers of different interval details were created. However, for the present study, only the tier containing syllable on- and offset information (Tier 6, Figure 1) and the tier containing the times of syllabic amplitude peak points (Tier 7, Figure 1) were relevant for the measurements described below (§2.2). For more about the TEVOID corpus, please refer to [8, 9].

2.2. Techniques of measurement

Two sets (mean metrics and peak metrics) of intensity variability measures were devised based on the durational metrics of speech rhythm (ΔC/V, varcoC, rPVI, nPVI) [4, 5, 6]. The basic calculation unit of the mean metrics is the average intensity across each syllable in the utterance, while the calculation unit of the peak metrics is the intensity of each automatically detected syllable peak. For both sets, global intensity variations are quantified by taking the standard deviations of both mean and peak intensity of the syllables in a sentence utterance, hence the metrics stdevM and stdevP. Local intensity variations are quantified by taking the cumulative intensity differences between adjacent syllables, either in the mean tier or peak tier. Formulas (1) and (2) express the idea more explicitly:

$$\text{rPVIm} = \sum_{j=1}^{n-2} | L_{\text{M}j} - L_{\text{M}j+1} | / (n-2) \quad (1)$$

$$\text{rPVIp} = \sum_{j=1}^{n-2} | L_{\text{P}j} - L_{\text{P}j+1} | / (n-2) \quad (2)$$

where rPVI means the raw pairwise variability index; $L_{\text{M}j}$ and $L_{\text{P}j}$ refer to the mean intensity level and the peak intensity level of the jth syllable in the utterance; and n refers to the total number of intervals in the utterance.

However, some speaker may be intrinsically "louder" than others, and the distance between the mouth and the microphone may not be precisely controlled either. Hence the global metrics are normalized by taking the ratios of the original scores and the average intensity levels in the mean and peak tiers: varcoM = 100 * stdevM / $\bar{L}_{\text{M}}$, varcoP = 100 * stdevP / $\bar{L}_{\text{P}}$, where varco is short for variation coefficient; $\bar{L}_{\text{M}}$ and $\bar{L}_{\text{P}}$ refer to the average syllabic intensity levels in the mean and peak tiers. The local measures are normalized by dividing the absolute difference of each neighboring pair by their own average value prior to the final summation:

$$\text{nPVIm} = \sum_{j=1}^{n-2} \left| \frac{L_{\text{M}j} - L_{\text{M}j+1}}{[L_{\text{M}j} + L_{\text{M}j+1}]/2} \right| \times \frac{100}{n-2} \quad (3)$$

$$\text{nPVIp} = \sum_{j=1}^{n-2} \left| \frac{L_{\text{P}j} - L_{\text{P}j+1}}{[L_{\text{P}j} + L_{\text{P}j+1}]/2} \right| \times \frac{100}{n-2} \quad (4)$$

where nPVI is the normalized pairwise variability index, and the denotations of the other symbols are the same as those in formulas (1) and (2). The scalar 100 in both varcoM, varcoP as well as in (3) and (4) makes the integer parts of the scores greater than zero.

The calculations were automated in Praat [12] using a script (available from the first author). The parameters for extracting and querying the intensity objects were set default (minimum pitch = 100 Hz, time steps = 0.0, subtract mean = True, averaging method = dB, interpolation method = Cubic). The initial and final syllables were excluded from analysis because the duration was sometimes too short for the intensity values to be measured reliably.

3. Data analysis and results

3.1. Data normality and transformations

Data distributions were assessed by constructing Q-Q plots for all the metrics. The data deviate from the normal Q-Q lines as the top left panel of Figure 2 (only rPVIp is displayed due to limited space, but the patterns are similar across metrics) shows, not meeting the distribution assumption of parametric statistics. Therefore, natural logarithmic transformations ($X_{\text{Trans}} = \ln X$), square root transformations ($X_{\text{Trans}} = X^{1/2}$) and arcsine transformations ($X_{\text{Trans}} = (2/\pi) * \sin^{-1}(X / 100)^{1/2}$) [13] were performed to see which methods optimally transform the data into normally distributed ones. As the Q-Q plots in Figure 2 indicate, the natural log transformations had the least success, whereas the square root and arcsine transformations showed similar success in transforming the data into normally distributed sets. Given similar effects of both square root and arcsine transformations, we will only analyze the square root transformed data in the coming sections, because the calculations are more straightforward compared with the arcsine transformations.

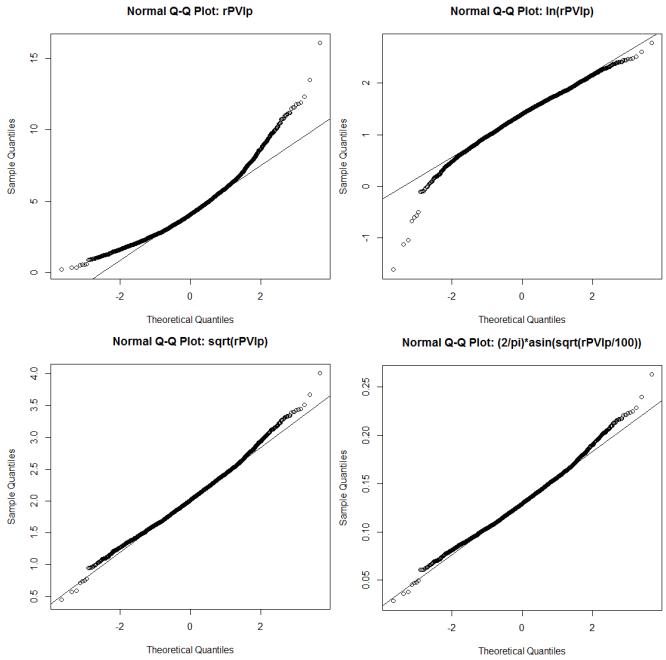

Figure 2. *Normal Q-Q plots of the original rPVIp (top left), natural log transformed rPVIp (top right), square root transformed rPVIp (bottom left), and arcsine transformed rPVIp (bottom right).*

3.2. Correlations between the metrics scores

First of all, the correlations between the raw and normalized metrics (stdevM/P vs. varcoM/P; rPVIm/p vs. nPVIm/p) were evaluated to see how well one metric can predict another. Figure 3 demonstrates the scatter plot matrices of both mean and peak metrics. It is evident that the raw scores are highly correlated with their normalized counterparts. The Pearson's correlation coefficients confirm that very high correlations exist: r = 0.974 for stdevM and varcoM, r = 0.981 for rPVIm

and nPVIm, $r = 0.991$ for stdevP and varcoP, and $r = 0.993$ for rPVIp and nPVIp (all p values < 0.0001).

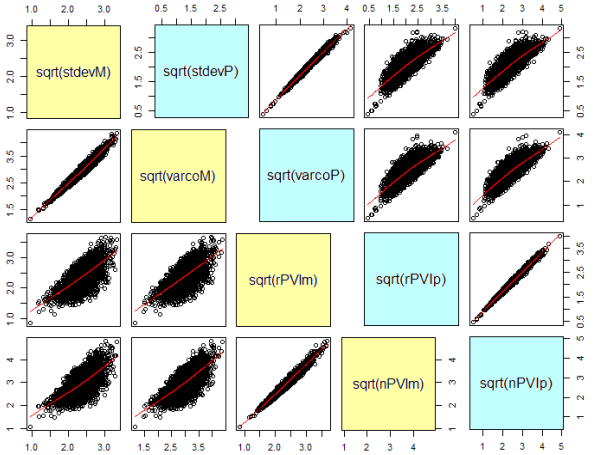

Figure 3. *The scatter plot matrices of both mean metrics (lower panel) and peak metrics (upper panel). The red LOWESS lines indicate that the relationships are linear to a large extent, and the raw scores and their normalized counterparts are highly correlated.*

In addition, correlations between the holistic and local metrics within the mean and peak measures are also significantly high: $r = 0.666$ for stdevM and rPVIm, $r = 0.662$ for varcoM and rPVIm, $r = 0.670$ for stdevM and nPVIm, and $r = 0.707$ for varcoM and nPVIm (all p values < 0.0001); $r = 0.794$ for stdevP and rPVIp, $r = 0.784$ for varcoP and rPVIp, $r = 0.804$ for stdevP and nPVIp, and $r = 0.809$ for varcoP and nPVIp (all p values < 0.0001).

Finally, we also examined the correlations between the mean metrics and the peak metrics. As the scatter plots (Figure 4) show, the correlation of both holistic metrics and local metrics between the mean and peak measures are rather poor (the highest correlation coefficient being merely 0.322 for nPVIm and nPVIp, $p < 0.0001$). This means that mean and peak measures cannot be reliably predicted from each other, suggesting that both measures contain different information about the amplitude contour.

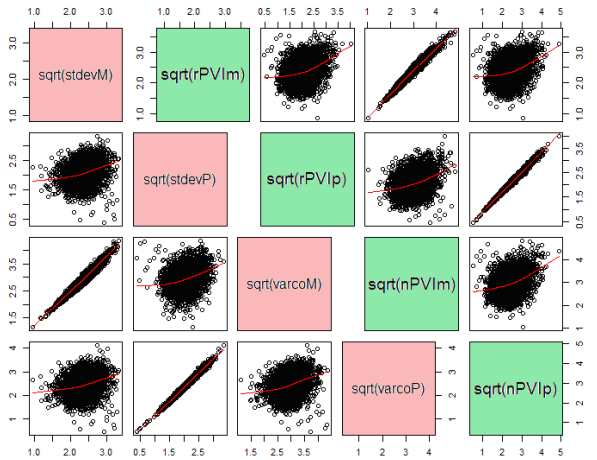

Figure 4. *The scatter plot matrices of both holistic measures (lower panel) and local measures (upper panel) between the mean and peak measures. The red LOWESS lines also indicate that the poorly correlated scores deviate from linearity to some extent.*

3.3. Significant effects of speakers

Figures 5 to 8 contain box plots for the measures of varcoM, nPVIm, varcoP, and nPVIp as a function of the 16 speakers. It is visible that there are strong differences between the speakers. For some of the measures the between speaker variability seems to be rather similar (nPVIp and varcoP), but not for the other measures. Univariate ANOVAs with speakers as the independent variable were performed on the square root transformed data in R [14]. For each metric, the Bartlett test of variances homogeneity was run, so as to adjust the "var.equal" argument in the ANOVA commands. If the equality of variances is violated, an approximate method of Welch [15] is applied in the computation. As Table 1 shows, significant effects of the speakers were found on all the metrics, both mean measures and peak measures.

Table 1. *Statistical outputs of the square root transformed data (var.equal=FALSE for all metrics as indicated by K^2).*

	Bartlett's K^2 (*df*)	*F*	*df* (num, denom)
stdevM	82.21⋆ (15)	47.43⋆	15, 1540.69
varcoM	109.30⋆ (15)	66.53⋆	15, 1540.17
rPVIm	85.46⋆ (15)	32.08⋆	15, 1540.74
nPVIm	108.93⋆ (15)	43.37⋆	15, 1540.65
stdevP	54.87⋆ (15)	88.25⋆	15, 1540.74
varcoP	60.14⋆ (15)	98.52⋆	15, 1540.72
rPVIp	64.15⋆ (15)	90.23⋆	15, 1540.76
nPVIp	64.94⋆ (15)	95.30⋆	15, 1540.76

⋆ $p < 0.0001$

The speaker individualities are also visualized by box plots (Figures 5 – 8). As Figure 3 shows, the raw metrics scores and the normalized ones are highly correlated, thus we only plot the normalized metrics (varcoM, nPVIm, varcoP and nPVIp), because their raw counterparts should have similar patterns.

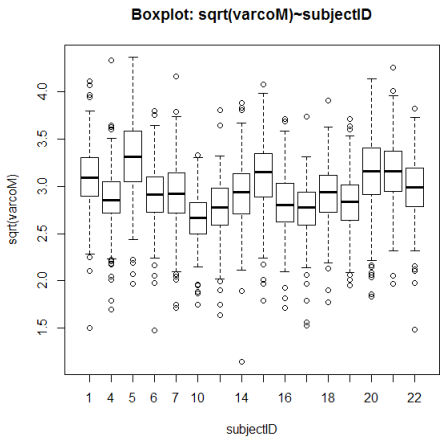

Figure 5. *Box plot of square root transformed varcoM.*

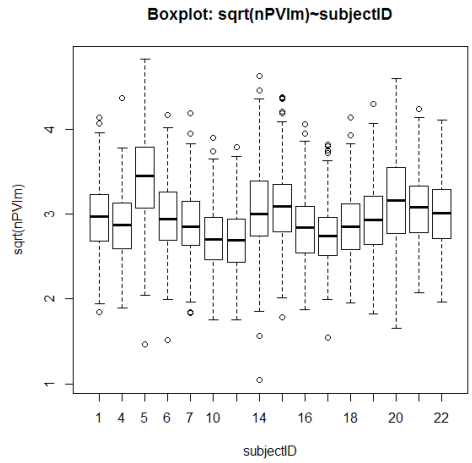

Figure 6. *Box plot of square root transformed nPVIm.*

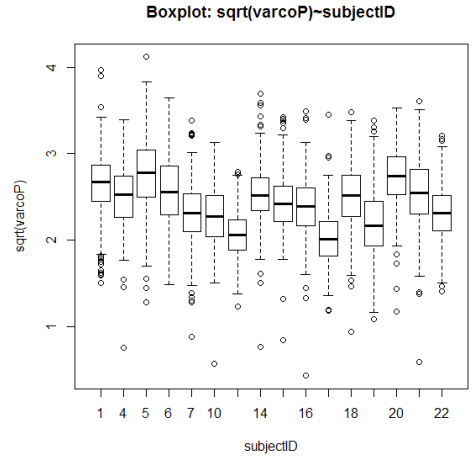

Figure 7. *Box plot of square root transformed varcoP.*

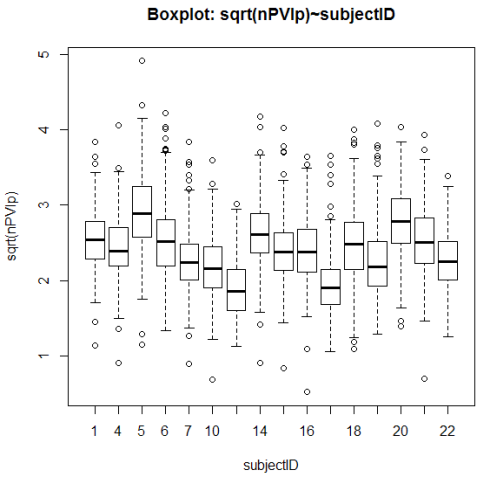

Figure 8. *Box plot of square root transformed nPVIp.*

4. Discussion

This study investigated speaker individualities in the speech signal through variations of intensity levels. Both holistic and local measures of intensity fluctuations (either average syllabic intensities, or syllable peak intensities) were employed. The results indicated that a significant speaker effect exists in all the calculation methods, suggesting a potential application of the methods in speaker recognition.

A closer inspection of the box plots shows that some speakers may not be differentiated from each other on one metric, but this is sometimes not the case on other metrics. For example, speakers 6 and 7 have similar scores on varcoM (Figure 5), but have different ones on varcoP or nPVIp. In a similar vein, speakers 15 and 16 may not be distinguished on varcoP and nPVIp, but they are very likely to be differentiated on varcoM and nPVIm. As Figure 4 suggests, the mean metrics and peak metrics are independent of each, and a combination of the two sets of measures should increase the probability of speaker recognition. Given the fair speaker discriminative strength of durational metrics (%V, %VO, ΔPeak, and varcoPeak) [8], we envisage that speakers can be differentiated even better by a combination of a variety of measures, including the ones presented here.

Moreover, idiosyncratic intensity variability is potentially important for forensic phonetic applications. [16] applied a 1-bit requantization to the speech signal, by setting all positive amplitude values to 1 and all negative ones to 0, thus getting rid of the information of intensity variability contained in the undulating amplitude envelope. However, the 1-bit requantized speech is highly intelligible, suggesting that intensity variability is something in the signal that the speakers may not be aware of. In such situations speakers also have less control over these variables. Therefore, intentional disguise of intensity variability might be more difficult since there is a lack of possible auditory feedback.

For our further research, we would like to see if such idiosyncratic characteristics in intensity levels could also be found in spontaneous speech. In most forensic speaker comparisons, the speech signal is deteriorated to a greater or lesser degree. [17] listed a number of sources of speech degradation, such as reduction of frequency bandwidth, presence of noise, reduction of energy level, spectral distortion, inadequacy of transmission links, and inadequate pickup transducers. Standard audio signal processing techniques like "compressor-limiter" normalization of amplitude levels are non-linear and might introduce a significant amount of noise to between-speakers intensity or amplitude variability. We would like to see if the metrics could, or to what extend, survive these adversities, and how they could be optimized to be useful in actual forensic case works.

5. Conclusion

This study investigated speaker idiosyncrasy via syllabic intensity fluctuations. The results showed that significant effects of the speakers existed in all the intensity metrics, and therefore, are potentially useful in speaker recognition tasks, especially in forensic settings. Future speaker recognition experiments will show whether this hypothesis holds.

6. Acknowledgements

This study is supported by the Gebert-Rüf Stiftung (Grant No. GRS-027/13) and the Swiss National Science Foundation (Grant No. 100015_135287). Adrian Leemann and Marie-José Kolly played a significant role in building the TEVOID corpus. Stephan Schmid made a significant contribution to the conceptualization.

7. References

[1] Dellwo, V., Huckvale, M., and Ashby, M., "How is individuality expressed in voice? an introduction to speech production and description for speaker classification", in C. Müller [Ed], Speaker Classification I, 1-20, Springer Verlag, 2007.

[2] Loula, F., Frasad, S., Kent, H., and Shiffrar, M., "Recognizing people from their movement", J. Exp. Psychol. Hum. Percept. Perform., 31: 210-220, 2005.

[3] Matovski, D., Nixon, M., Mahmoodi, S., and Carter, J., "The effect of time on the performance of gait biometrics", IEEE 4th Conference on Biometrics, Washington, DC, USA, 2010.

[4] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73: 265-292, 1999.

[5] Grabe, E. and Low, E. L., "Durational variability in speech and rhythm class hypothesis", in N. Warner and C. Gussenhoven [Eds], Papers in Laboratory Phonology 7, 515-543, Mouton de Gruyter, 2002.

[6] Dellwo, V., "Rhythm and speech rate: A variation coefficient for deltaC", in P. Karnowski and I. Szigeti [Eds], Language and Language Processing, 231-241, Peter Lang, 2006.

[7] White, L., and Mattys, L. S., "Calibrating rhythm: first language and second language studies", J. Phonet., 35: 501-522, 2007.

[8] Dellwo, V., Leemann, A., and Kolly, M-J., "Speaker idiosyncratic rhythmic features in the speech signal", in Interspeech, Portland, USA, 2012.

[9] Leemann, A., Kolly, M-J., and Dellwo, V., "Speech-individuality in suprasegmental temporal features: implications for forensic voice comparion", Forensic Sci. Int., 238: 59-67, 2014.

[10] Dellwo, V., Schmid, S., Leemann, A., Kolly, M-J., and Müller, M., "Speaker identification based on speech rhythm: the case of bilinguals", Abstract presented at PoRT2012, Glasgow, Scotland, 2012.

[11] He, L., "Syllabic intensity variations as quantification of speech rhythm: evidence from both L1 and L2", in Speech Prosody 6, Shanghai, China, 2012.

[12] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", version5365, http://www.praat.org/, 2014.

[13] Johnson, K., Quantitative methods in linguistics, Wiley-Blackwell, 2008.

[14] R Core Team, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/, 2014.

[15] Welch, B. L., "On the comparison of several mean values: an alternative approach", Biometrika, 38: 330-336, 1951.

[16] Kolly, M-J., and Dellwo, V., "Cues to linguistic origin: the contribution of speech temporal information to foreign accent recognition", J. Phonet., 42: 12-23, 2014.

[17] Hollien, H., "About forensic phonetics", Linguistica, 52: 27-53, 2012.