Articulation During Voice Disguise: a Pilot Study

Lauri Tavi[1,2], Tomi Kinnunen[2], Einar Meister[3], Rosa González-Hautamäki[2,5], and Anton Malmi[4]

[1] School of Humanities, University of Eastern Finland, Joensuu, Finland
[2] School of Computing, University of Eastern Finland, Joensuu, Finland
[3] School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia
[4] Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia
[5] Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
lauri.tavi@uef.fi

Abstract. Speakers can conceal their identity by deliberately changing their speech characteristics, or disguising their voices. During voice disguise, speakers alter their normal movements of the articulators, such as tongue positions, according to a predetermined strategy. Even though technology for accurate articulatory measurements has existed for years, few studies have investigated articulation during voice disguise. In this pilot study, we recorded articulation of four speakers during regular and disguised speech using electromagnetic articulography. We analyzed imitation of foreign accents as a voice disguise strategy and utilized functional t-tests as a novel method for revealing articulatory differences between regular and disguised speech. In addition, we evaluated discovered articulatory differences in the light of the performance of an x-vector-based automatic speaker verification system.

Keywords: electromagnetic articulography · functional data analysis · foreign accent · voice disguise · automatic speaker verification.

1 Introduction

The human voice is extremely variable and flexible. Besides *inter-speaker* variation due to organic differences in vocal production systems, speakers can modify the content and the style of speaking flexibly, leading to *intra-speaker* speech variation of a given individual. Some of this variation is intentional and controllable by the speaker (*e.g.* whispering to enable private communication) while others are either automatic (*e.g.* Lombard reflex) or only weakly controllable by the speaker (*e.g.* accent of one's mother tongue) [7]. Depending on whether the focus is on speech *science* or speech *technology*, these variations are either the main object of interest or an unwanted nuisance. For automatic speaker verification (ASV), speech variation cause problems as it lowers even the state-of-the-art ASV system's accuracy [6]. Consequently, some speakers might try to

exploit this deficiency by deliberately varying their speech features, or disguising their voices.

One highly common form of variation is caused by *foreign accent*, which yields various phonetic changes in speech. Second language (L2) learners also tend to use phonemic substitution rules of their first language (L1), leading to *foreign-accented* speech that shares properties of both L1 and L2 [5]. Consequently, characteristics of L2 speech can be imitated as a form of voice disguise [13].

Human speech is essentially multimodal – it can be analyzed and represented in terms of articulatory, acoustic and perceptual attributes. While research in speech technology and acoustic phonetics has benefited from large acoustic datasets available in many languages, acoustic–articulatory speech data is much scarcer [4]. The study of speech production involves tracking the displacement, timing, and coordination of *articulators* (such as the tongue, the jaw and the lips) inside and outside of the vocal tract. Traditional *imaging* techniques such as x-ray, ultrasound and magnetic resonance imagining have been used to visualize articulators. These methods can be restrictive since speech is characterized by fast, complex and small 3-dimensional movements of the articulators. Our focus, *electromagnetic articulography* (EMA), is a 3D measurement technique designed to track and record articulatory movements during speech production. EMA is an invasive method but it allows precise tracking of the position and orientation of miniaturized sensor coils attached to various places on the articulators. An articulograph records articulator trajectories directly without the need for additional image processing techniques. However, disadvantages of EMA are rather heavy post-processing of raw data and time-consuming data collection.

In this study, we recorded L1 and imitated L2 speech from Finnish and Russian speakers using EMA and explored 1) articulatory differences between L1 and imitated L2 speech and 2) the effect of imitated L2 accent on a modern deep speaker embedding ASV. The latter was tested using an x-vector-based ASV system. In former, we utilized *functional data analysis* (FDA), particularly *functional t-tests*. Although there are existing EMA corpora in English (e.g. `MOCHA-TIMIT` [24], `mngu0` [19], `USC-TIMIT` [12]), Mandarin-accented English [8], German [2] and Italian [4], the authors are unaware of previous EMA corpora available for the combination of Finnish and Russian. To our knowledge, this study is also the first data collection that addresses voice disguise through EMA measurements.

2 Phonology of Finnish and Russian

Finnish and Russian have numerous differences in their phonological systems [21], which can affect production and imitation of Russian and Finnish accents. For instance, Finnish has eight vowels, which can occur short or long, while Russian has six. Finnish has also 18 diphthongs that are absent in Russian.

Russian sibilants and affricates can be particularly problematic for Finns. Additionally, in Russian most of the consonants can be *palatalized* in various

positions. In palatalization, the place of articulation is higher and more anterior [10]. Palatalized consonants acquire a secondary place of articulation on the *palatal region* of the mouth while preserving the primary constriction. Unlike Finnish, Russian has phonological oppositions in palatalized and non-palatalized consonants, where the usage of palatalization changes the meaning (e.g. мать, 'mother' and мат, 'check mate'). However, most Finnish consonants can occur either short or long. Furthermore, typical structure of a Finnish syllable is CV(C/V), but Russian syllables can contain one vowel and up to five consonants, e.g. a word взгляд, 'a look'.

Another essential difference between Finnish and Russian phonology is the position of word stress. In Russian, word stress can occur in any syllable resulting in stress-based minimal pairs, e.g. му'ка, 'flour' and 'мука, 'torment'. In Finnish, the first syllable of a word is always stressed.

3 EMA Data Collection

3.1 Corpus Design

We collected simultaneous recordings of read-aloud speech and articulatory movements tracked by EMA (EMA Wave by Northern Digital) for Finnish and Russian. The collected speech included three different speaking styles from each speaker: speech in native (L1) and non-native (L2) Finnish or Russian, and while imitating Finnish or Russian foreign accent (IL2) in the L1. In the IL2, participants were asked to imitate the foreign accent without further instructions or practise. In this study, we focused on the differences between the L1 and the IL2.

Overall, we collected data from six speakers including one Finnish female, one Finnish male, one Russian female, and three Russian male speakers. To have a balanced set of native languages and sexes for this pilot study, we focused on four speakers summarized in Table 1. Even though four participants can be considered as a limited number of speakers, EMA studies commonly have few participants: for example, [4], [12], [19] and [24] collected articulatory data from four or less speakers. Yet, we do *not* claim generality of the presented findings beyond the collected material — rather, the point is to demonstrate specific elicited speech variations in terms of articulatory changes, and to address the potential ramification of such changes on ASV performance.

To this end, read-aloud speech is elicited by Finnish and Russian versions of Aesop's Fable 'The north wind and the sun'. Additionally, we collected prompt sentences (ca 70 sentences in both languages) involving most frequent vowels, diphthongs, and consonants in different segmental and prosodic contexts, and text material contained spontaneous speech, elicited by a story telling task based on a cartoon. The prompted sentences and spontaneous speech based on a cartoon were excluded from analyses, but the prompted sentences were included in ASV evaluation (see 4.2).

Table 1. Speaker information. L2 and IL2 (i.e. "imitated L2") levels are self-evaluated by each speaker using categories of low–middle–high and 1–5 scale, respectively. For the IL2 level, the higher the number, the better the imitation.

speaker	sex	age	L1	L2	L2 level	IL2 level
FIN_M_001	male	28	Finnish	Russian	high	5
FIN_F_001	female	22	Finnish	Russian	high	2
RUS_M_003	male	30	Russian	Finnish	middle	3
RUS_F_001	female	18	Russian	Finnish	middle	3

3.2 Recording Procedure

Recording setup included AKG C444 close-talking microphone (headset), the EMA system and a Windows desktop computer running the EMA recording software. Sensor positions are presented in Table 2.

Table 2. Sensor positions. Left and right are in relation to the experimenter.

sensor number	position
10–12	biteplate (in a triangular shape)
9	nose
7 and 8	left and right mastoid
6	jaw (behind the lower lip, on the gum)
3 and 5	left and right lateral
4	laminal
1 and 2	tongue dorsum and anteo-dorsum

Articulatory movements and the audio signal were recorded simultaneously at sampling frequencies of 200 Hz and 22.05 kHz, respectively. The recording procedure included the following steps: 1) gluing reference sensors 9,8 and 7, 2) a *bite-plate* recording, 3) gluing sensors 1-6, 4) doing a palate trace recording, where participant produced syllables *ta*, *ti*, *tu*, *ka*, *ki* and *ku*, and 5) performing a rehearsal run and 6) finally recording itself. The prompt items were displayed on sheets of paper. Fig. 1 shows the EMA system and sensors 1–5 glued on the tongue.

Since EMA recordings involve physically attaching sensors on participants, our study underwent detailed ethical and safety evaluation by the Ethics Committee of the University of Eastern Finland. In January 2020, the Committee gave a supporting statement for the proposed research.

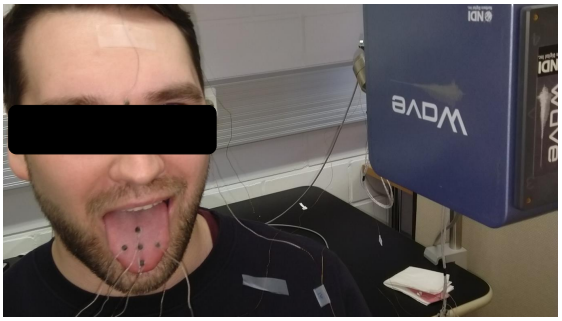

Fig. 1. Example of attached sensors on the tongue.

4 Methods

4.1 Functional Data Analysis

FDA involves a set of statistical methods, which are extended from their traditional counterparts to function of time [18]. First step in FDA is to transform discrete data points, such as EMA sensor movements, to continuous curves using basis functions. We used B-splines, a common choice with non-periodic signals. This step also smooths data trajectories, which helps to avoid overfitting [20].

The occurrence of palatalization, which requires tongue dorsum raising and fronting, is one of the major differences between Finnish and Russian accents (see Section 2). Therefore, we focused on *sagittal plane* of *tongue dorsum* (TD) sensor which tracks down–up and front–back movement of TD. TD sensor curves were constructed from each word of Finnish and Russian versions of 'The north wind and the sun'. However, words that contained less than 40 measured samples were excluded. This resulted in a total of 531 curves for both the sensor directions.

The second step in FDA is to perform statistical analysis on continuous curves. For comparing the mean TD sensor curves between the L1 and the IL2, we used *functional* t-tests, extensions of classical t-tests, where the t-statistic is defined as

$$T(t) = \frac{|\overline{x}_1(t) - \overline{x}_2(t)|}{\sqrt{\frac{1}{n_1}\mathrm{Var}[x_1(t)] + \frac{1}{n_2}\mathrm{Var}[x_2(t)]}}. \quad (1)$$

In Equation 1, $\bar{x}$s are the curve means and Var(xs) are the curve variances for the L1 and the IL2. In functional t-tests, the maximum value of *T(t)* is used as a test statistic, which critical value is found using permutation test. The labels of the curves are randomly shuffled and the maximum value of *T(t)* is recalculated with new labels. As a result, a null distribution is constructed. In this study, FDAs were carried out using `R` [16] package `fda` [17]. `WebMAUS` [9] and `Praat` [3] were used to time-align the word segments. Following list will summarize the FDA procedure used in this study:

1. First, TD sensor trajectories were recorded from four participants, who read 'The north wind and the sun' once with their native language and once while imitating a foreign accent.
2. Then, the sensor trajectories were segmented into individual word trajectories using word-level annotations and converted to smoothed functions (i.e. curves) using B-splines.
3. Finally, the means of the word curves of the two speaking styles were compared using functional t-tests; four functional t-tests were performed separately on each participants' speaking styles.

4.2 Automatic Speaker Verification

We considered the effects of accent imitation in an ASV experiment. The L1 read speech from the "The north wind and the sun" was used to train the speaker model (enrollment). In the evaluation phase, we considered gender-dependent trials, same speaker's (*target*) trials correspond to the speaker's audio samples from the prompted sentences, and for different speaker's (*nontarget*) trials the prompted sentences read by the other speaker. The test audio samples average duration is 3.5 sec. The trials with L1 are considered as the baseline case, and the effect of including IL2 trials as the disguise case. The number of trials is shown in Table 3.

Table 3. The number of ASV trials per speaking style.

speaking style	**sex**	**target**	**non-target**
L1	male	294	294
IL2	male	302	302
L1	female	296	296
IL2	female	296	296

In the experiments, we used an x-vector based ASV system [23]. The system is based on a speaker-discriminative training using deep neural network architecture [22]. The ASV system correspond to the implementation in Kaldi-toolkit [14] with the pre-trained model recipe [1] from augmented VoxCeleb 1 and 2 data [11]. The speech samples were turned from stereo to mono samples by selecting the left channel, and then downsampled to 16 kHz. The feature extraction configuration consists of 23-dimensional Mel-frequency cepstral coefficients (MFCCs) extracted from 25 ms long frames every 10 ms. A cepstral mean subtraction was applied over a 3-second sliding window and energy-based speech activity detection was used to filter out the non-speech frames. A probabilistic linear discriminant analysis (PLDA) [15] was used for scoring the extracted 512-dimensional x-vectors representation of the trial's samples. The x-vectors were centred, whitened, and unit length normalized.

5 Results

5.1 Functional T-tests

Functional t-tests were applied to check for significant differences between the mean values of the TD sensor curves in the L1 and the IL2. Although the curves contained variation caused by producing different words, the same speech material (i.e. 'The north wind and the sun') was used in comparison of the L1 and the IL2. Consequently, the mean word curves of each speaker's speaking styles show average TD movements revealing plausible articulatory changes related to the IL2.

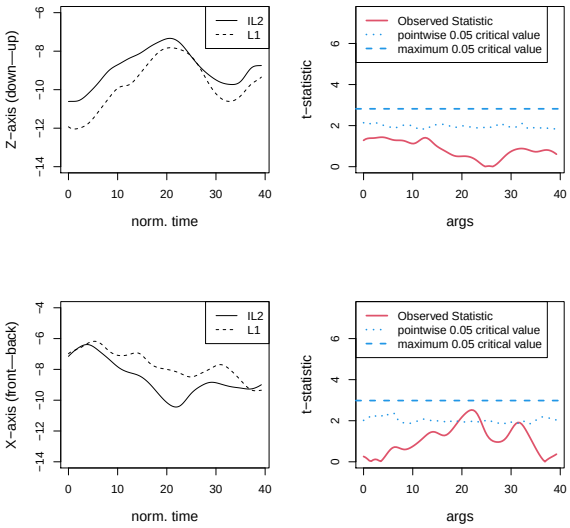

Fig. 2. FIN_F_001's mean TD movements in the IL2 and the L1 (left panels) and functional t-tests (right panels). The panels on the left show the down–up (above) and the front–back (below) movements.

Figures 2–5 show the mean curves in the L1 and the IL2 and the results of the functional t-tests for all four speakers. Comparing T-statistic to conservative reference line, *maximum critical value*, no statistically significant differences between the TD positions in the L1 and the IL2 were found for female speakers (see Figs. 2 and 4). However, male speakers' front–back TD curves differed statistically significantly between the speaking styles. These differences are at strongest approximately on the left and on the right side of the curve, while in the middle they are at weakest. This indicates that especially the beginning and the ending of the words are pronounced differently, i.e. the position of TD in the IL2 is more

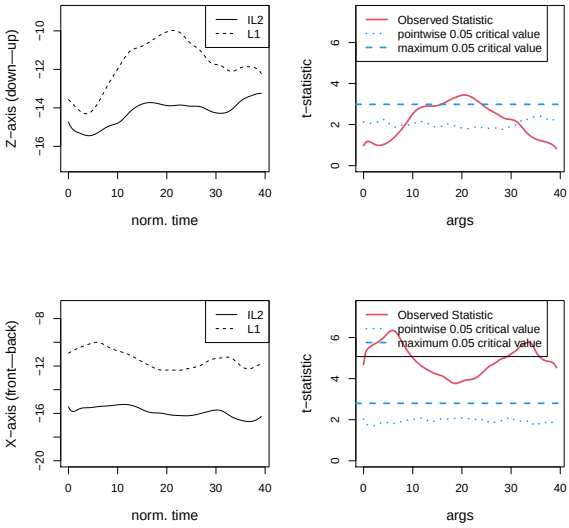

Fig. 3. FIN_M_001's mean TD movements in the IL2 and the L1 (left panels) and functional t-tests (right panels). The panels on the left show the down–up (above) and the front–back (below) movements.

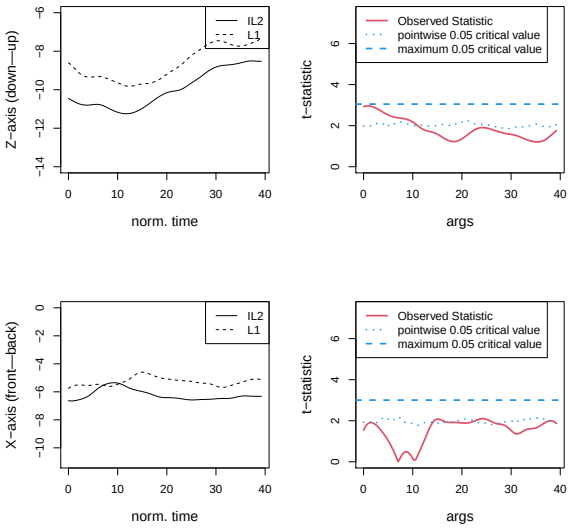

Fig. 4. RUS_F_001's mean TD movements in the IL2 and the L1 (left panels) and functional t-tests (right panels). The panels on the left show the down–up (above) and the front–back (below) movements.

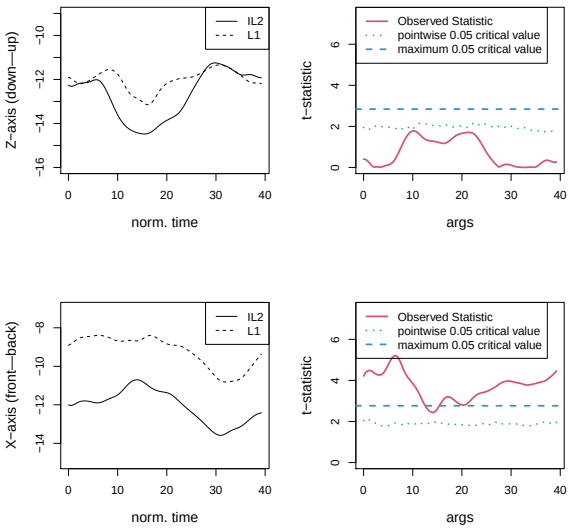

Fig. 5. RUS_M_003's mean TD movements in the IL2 and the L1 (left panels) and functional t-tests (right panels). The panels on the left show the down–up (above) and the front–back (below) movements.

anterior compared to the TD position in the L1. The down–up sensor movements showed no differences expect briefly at the middle of FIN_M_001's mean curves. Compared to other speakers, FIN_M_001's TD positions between the L1 and the IL2 differed the most, which most likely related to his self-evaluated imitation performance (i.e. 5 out of 5; see Table 1), which was the highest number of all four speakers.

5.2 Speaker Verification Results

We tested whether the articulatory differences found between the L1 and the IL2 (see Section 5.1) relate to the x-vector system (see Section 4.2) performance. Table 4 shows the effect of the IL2 on the ASV system as percentage of *equal error rates* (EERs). The EER is an error rate, which equates false acceptance rate and false rejection rate by adjusting a detection threshold. The higher the EER value, the lower the accuracy of the system.

Female speakers' IL2 had no effect on ASV accuracy since the EERs (%) were the same (0.34) for the L1 and the IL2. On the contrary, male speakers' IL2 caused a strong negative effect on ASV accuracy, increasing the EER (%) from 3.06 to 11.59. The same conclusions can be drawn from Fig. 6, where density distributions of genuine trials ASV scores for each speaker are presented. Because male speakers' IL2 yields lower scores compared to their L1 especially

Table 4. Gender-dependent equal error rates for the L1 and the IL2.

sex	speakers	L1–EER(%))	IL2–EER(%)
male	FIN_M_001 & RUS_M_03	3.06	11.59
female	FIN_F_001 & RUS_F_01	0.34	0.34

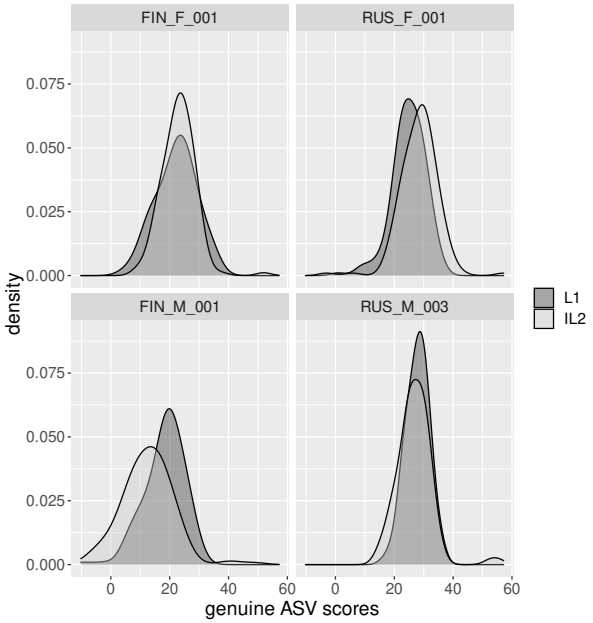

Fig. 6. Density distributions of ASV target scores for each speaker. The lower the ASV scores, the less confident the system is that the speakers are the same.

for FIN_M_001, the results from the ASV tests support the articulatory data indicating that tongue fronting was used as an effective voice disguise technique.

6 Conclusions

In this study, we investigated articulation of Finnish and Russian speakers during imitation of a foreign accent. The imitation of a foreign accent served as a method of voice disguise. We recorded tongue movements during regular and disguised speech using EMA and performed functional t-tests on the trajectories of TD movements. Additionally, we recorded the audio signal and tested the effect of voice disguise on an x-vector based ASV system. Using these two approaches, it was possible to investigate actual articulatory changes during different speaking styles and the effectiveness of the changes against an x-vector-based ASV system.

Functional t-tests revealed significant differences in the front–back TD movements between male speakers' L1 and IL2; for female speakers, there were no

significant differences in the TD positions. Although male speakers had different L2 and IL2, they both fronted their tongues during voice disguise. Fronting the tongue can occur during palatalization, which can be an expected when imitating (palatalized) Russian accent as a Finnish L1 speaker. However, also the Russian speaker's average TD position was more at front while imitating Finnish accent. In this case, the type of imitated Finnish feature was less clear.

The articulatory differences were also evaluated in the respect of ASV performance. The EERs(%), which were calculated using the x-vector-based ASV-system, supported the articulatory findings: while female speakers' IL2 had no effect on ASV performance, male speakers' IL2 increased the EERs(%) from 3.06 to 11.59. The declined ASV performance can be possibly explained by male speakers' tongue fronting.

This pilot study has shown that articulatory movements during voice disguise can be revealed using EMA measurements and functional data analysis. When the Covid-19 restrictions ease, we aim to continue collecting EMA data and reveal more articulatory mechanisms of voice disguise.

7 Acknowledgements

This project was partly funded by Academy of Finland (proj. 309629). Einar Meister's work was supported by the European Regional Development Foundation (the project "Centre of Excellence in Estonian Studies"). We thank Fabian Tomaschek for providing a set of R scripts for post processing of raw EMA data.

References

1. VoxCeleb Xvector models system 1a. https://kaldi-asr.org/models/m7, accessed: 2021-04-10
2. Arnold, D., Tomaschek, F.: The karl eberhards corpus of spontaneously spoken southern german in dialogues-audio and articulatory recordings. In: Kleber, C.D..F. (ed.) Tagungsband der 12. tagung phonetik und phonologie im deutschsprachigen raum. pp. 9–11. Ludwig-Maximilians-Universitat Munchen. Retriev (2016)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program] (2020), https://http:/praat.org
4. Canevari, C., Badino, L., Fadiga, L.: A new italian dataset of parallel acoustic and articulatory data. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
5. Fan, J., Yongbing, L.: The impact of l1 negative phonological transfer on l2 word identification and production. International Journal of Linguistics **6**(5), 37–50 (2014)
6. González Hautamäki, R., Hautamäki, V., Kinnunen, T.: On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. The Journal of the Acoustical Society of America **146**(1), 693–704 (2019)
7. Hansen, J.H., Bořil, H.: On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. Speech Communication **101**, 94–108 (2018)

8. Ji, A., Berry, J.J., Johnson, M.T.: The electromagnetic articulography mandarin accented english (ema-mae) corpus of acoustic and 3d articulatory kinematic data. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7719–7723. IEEE (2014)
9. Kisler, T., Reichel, U., Schiel, F.: Multilingual processing of speech via web services. Computer Speech & Language **45**, 326–347 (2017)
10. Malmi, A., Lippus, P.: Keele asend eesti palatalisatsioonis. Journal of Estonian and Finno-Ugric Linguistics **10**(1), 105–128 (2019)
11. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Science and Language p. 101027 (2019)
12. Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.C., Zhu, Y., et al.: Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). The Journal of the Acoustical Society of America **136**(3), 1307–1311 (2014)
13. Neuhauser, S.: Voice disguise using a foreign accent: phonetic and linguistic variation. International Journal of Speech, Language & the Law **15**(2), 131–159 (2008)
14. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding (ASRU). IEEE Signal Processing Society, Hawaii, US (2011)
15. Prince, S.J.D., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: Proc. of ICCV. pp. 1–8. Rio de Janeiro, Brazil (2007). https://doi.org/10.1109/ICCV.2007.4409052
16. R Core Team: R: A language and environment for statistical computing (2020), https://www.R-project.org/
17. Ramsay, J., G.S., Hooker, G.: fda: Functional data analysis. R package version 5.1.5.1. (2020), https://CRAN.R-project.org/package=fda
18. Ramsay, J.O., Silverman, B.W.: Functional data analysis (2nd edition). NY:Springer Verlag (2005)
19. Richmond, K., Hoole, P., King, S.: Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
20. Schötz, S., Frid, J., Gustafsson, L., Löfqvist, A.: Functional data analysis of tongue articulation in palatal vowels: Gothenburg and malmöhus swedish/i:, y:, 0ff. In: Proceedings of Interspeech. vol. 2013 (2013)
21. de Silva, V., Ullakonoja, R.: Introduction: Russian and finnish in contact. In: de Silva, V., Ullakonoja, R. (eds.) Phonetic of Russian and Finnish: General Description of Phonetic Systems: Experimental Studies on Spontaneous and Read-aloud Speech. pp. 15–20. Frankfurt am Main: Peter Lang (2009)
22. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proceedings of INTERSPEECH. pp. 999–1003. Stockholm, Sweden (2017)
23. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333. IEEE, Calgary, AB, Canada (2018)
24. Wrench, A.: The mocha-timit articulatory database (1999), www.cstr.ed.ac.uk/research/projects/artic/mocha.html