INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Assessing acoustic and articulatory dimensions of speech motor adaptation with random forests

Eugen Klein[1], *Jana Brunner*[1], *Phil Hoole*[2]

[1]Humboldt-Universität zu Berlin, Germany
[2]Ludwig-Maximilians-Universität München, Germany
{eugen.klein, jana.brunner}@hu-berlin.de, hoole@phonetik.uni-muenchen.de

Abstract

Although most modern theories of speech production assume that representations of speech sounds are multidimensional encompassing acoustic and articulatory information, speech motor learning studies which assess the degree of adaptation in both dimensions are few and far between. In the current paper, we present an auditory perturbation study of German sibilant [s] in which speakers' audio and articulatory movements were recorded by means of electromagnetic articulography. Random Forest, a supervised learning algorithm, was employed to classify speakers' responses produced under unaltered or perturbed feedback based either on acoustic or articulatory parameters. Preliminary results demonstrate that while classification accuracy increases in the acoustic dimension as the perturbation session goes on, the classification accuracy in the articulatory dimension, although overall higher, remains approximately at the same level. This suggests that the adaptation process is characterized by active exploration of the articulatory space which is guided by speakers' auditory feedback.

Index Terms: auditory perturbation, sibilants, electromagnetic articulography, random forest

1. Introduction

Substantial amount of behavioral evidence from auditory and oral-articulatory perturbation studies suggests that speakers employ auditory and somatosensory feedback signals to adjust for errors in their own speech production. For example, altering the formant frequencies in real-time results in automatic and largely unconscious compensation in articulation [1]. On the other hand, speakers also modify their articulatory movements when these are mechanically impeded without any consequences in the auditory dimension [2]. These findings are supported by more recent neuroimaging results which suggest that sensory representations of speech sounds are defined in both auditory and somatosensory cortex and are characterized by neural auditory-somatosensory mappings [3, 4].

Both types of evidence, behavioral and imaging, inform modern theories of speech production which all in all agree that speech sound representations are multidimensional in their nature. Subtle controversy, however, arises around the issues whether there is a hierarchical relation between auditory and somatosensory signals and how they are integrated during speech production. For instance, some authors assume that targets of speech are rather auditory than articulatory while an auditory-motor-somatosensory mapping is learned by a speaker during speech acquisition [5]. This position is similar to the idea that speech sounds are perceptuo-motor units comprising of articulatory movements which are shaped by perceptual properties and selected for their functional value for communication [6]. Other authors emphasize the role of the somatosensory signal which is assumed to be employed by speakers to fine-tune their articulatory movements [7] or to be in a constant trade-off relation with the auditory signal [8].

Set against this theoretical discussion, the findings of previous cross-modal perturbation studies, in which auditory and somatosensory feedback signals were perturbed alone or in combination during speech production, suggest that speakers display individual preferences regarding one of the feedback signals [9] or give systematic preference for the auditory feedback when adaptation to somatosensory perturbation increases the auditory error [10]. By design, these studies were set up to investigate specific trade-off scenarios between auditory and somatosensory feedback signals. However, a neutral account of speech motor learning which investigates acoustic and articulatory dimensions of adaptation in a more balanced fashion is missing in literature. There are reasons to assume that the degree of adaptation effects in acoustic and articulatory dimensions might indeed differ as it is known that certain acoustic outputs can be produced by employing different articulatory configurations [11].

In the current experiment, we perturbed acoustic spectra of the voiceless sibilant [s] in near real-time while participants' speech and articulatory movements were recorded by means of electromagnetic articulography (EMA). Since the applied perturbation decreased the spectral center of gravity of the target sound, we expected speakers to react to this perturbation by changing the position of their articulators. We chose the sound [s] for our investigation as sibilants are characterized by prominent features in both acoustic and articulatory dimensions [12, p. 388]. Acoustically, voiceless sibilants exhibit broad noise spectra with contrastive differences regarding their spectral means [13, p. 1176 ff.]. Articulatorily, accurate production of sibilants requires precise control of the constriction location, tongue grooving, and the jaw height [14, p. 706]. During the analysis, we employed the Random Forest (RF) algorithm [15] to classify speakers' responses produced under perturbed or unaltered auditory feedback on the basis of acoustic measurements and spatial displacements of their articulators. This methodology allowed us to investigate the degree of potential adaptation effects in both dimensions of participants' speech and to examine the amount of their congruence.

2. Methods

2.1. Participants

16 female, native monolingual speakers of German without reported speech, language, or hearing disorders participated in

http://dx.doi.org/10.21437/Interspeech.2019-1812

the study. The mean age of the group was 29.5 years. The study was approved by the ethics committee of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS) and all participants gave their written consent to participate in the experiment.

2.2. Electromagnetic articulography

To capture articulatory measurements, the AG501 3D electromagnetic articulograph (Carstens Medizinelektronik GmbH, Germany) was used. Articulatory sensors were placed midsagittaly on the tongue tip (TT), tongue mid (TM), tongue back (TB), and the lower incisors (JAW) with static head reference sensors placed on the gum above the upper incisors, the nasal bone and behind each ear on the mastoid process.

To identify the neutral position of reference sensors, recordings from a static pose were made while participants remained immobile for a few seconds. During post-processing this reference position was used to correct for any head movements that occurred during the experiment and to translate the articulatory data into a coordinate system centered on the upper incisors, thus allowing comparison of sensor displacement across all participants.

Sensor movements were recorded at 1250 Hz and downsampled to 250 Hz during post-processing. The speech audio was recorded with a Sennheiser ME 62 omni-directional microphone and digitalized at a sampling rate of 44.1 kHz.

2.3. Experimental procedure

Each experimental session was recorded in a sound attenuated booth and lasted for about 25 minutes. Participants were seated in front of a computer monitor which served to display the stimuli. Before the actual experiment began, participants completed a familiarization block in order to accustom themselves to speaking with articulatory sensors attached. Focus was made on the production of the sibilant [s] which later served as the target segment for the auditory perturbation. Overall, each participant produced 96 [s]-tokens embedded in five word sentences during the familiarization.

The following experiment included 160 repetitions of the German sentence [lasə (ʔ)ɛɐ̯hiːlt aɪ̯nə tasə] (*Lasse erhielt eine Tasse;* engl. *Lasse got a mug*) and was divided into five blocks. During a baseline block, no auditory perturbation was applied to participants' speech. After the baseline, three shift blocks followed during which the spectral properties of the [s]-sound contained in [lasə] were perturbed in near real-time. The sibilant contained in [tasə] remained unperturbed during the whole experiment to serve as a control condition. From previous research it is known that participants are able to adapt simultaneously to multiple auditory perturbations [16]. In order to test for potential after-effects of the perturbation, each experimental session ended with a post-shift block incorporating unperturbed feedback.

Participants were asked to produce the experimental stimuli with a neutral intonation and in a moderate speech tempo in order to improve the online tracking of the sibilant. To keep the speech amplitude equal across all blocks, participants were provided with a real-time graphic display of the microphone gain. Participants' speech was fed back via EARTONE 3A insert earphones. The earphones attenuated the air-conducted sound by 25–30 dB while the feedback level was set approximately at 75 dB SPL.

2.4. Auditory perturbation

For auditory perturbation we employed AUDAPTER, a C++ real-time signal processing application executable within MATLAB [17]. In order to perturb the target segment [s], we employed AUDAPTER's pitch shifting facilities. We applied a negative pitch shift of -6.5 semitones to the sibilant such that its spectrum was shifted in its entirety towards lower frequencies resulting in an average decrease of its center of gravity by 5 kHz (Figure 1). The pitch shift was applied exclusively to the sibilant contained in [lasə] and did not affect any of the other segments of the experimental stimulus.

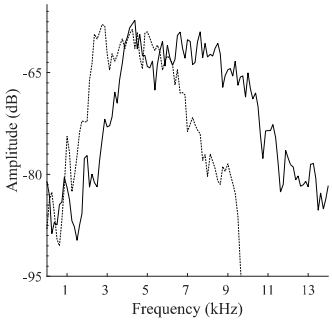

Figure 1: *Original (solid line) and perturbed (dashed line) power spectrum of a sibilant.*

To identify the onset and the offset of the sibilant in real-time, AUDAPTER performed an analysis of the speech signal's short-time root-mean-square (RMS) and RMS ratio curves (Figure 2a). While RMS is an acoustic intensity measure, RMS ratio is an indicator of high frequency intensity present in the signal and is computed by dividing a smoothed RMS curve by a high-pass filtered RMS curve. Operating with a set of heuristic rules, AUDAPTER enabled pitch shifting when a pre-defined high frequency energy threshold was crossed, which coincided with the onset of the sibilant, and turned it off immediately when the high frequency energy curve fell again under the threshold (Figure 2b). The audio signal was digitized and fed back to participants with a sampling rate of 32 kHz.[1]

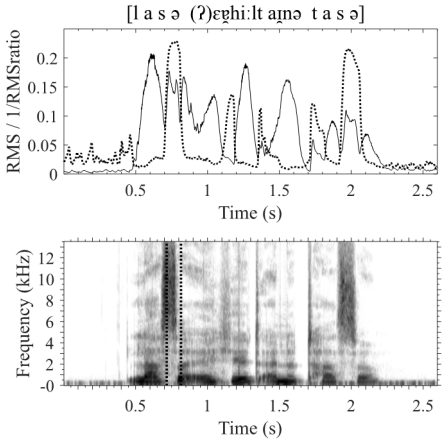

Figure 2: *Single experimental trial: (a) RMS (solid line) and RMS ratio curve (dashed line) of the speech signal. (b) Sibilant onset and offset (dashed lines) tracked by AUDAPTER overlaid over a spectrogram of the speech signal.*

[1]For this we had to modify the original AUDAPTER software to operate at a higher sampling frequency.

2.5. Experimental measurements

We chose several acoustic and articulatory measurements to investigate adaptation in participants' speech. To examine the acoustic dimension, we first divided sibilants' spectra into three frequency bands, 600–5500 Hz (low band), 5500–11000 Hz (mid band), and 11000–16000 Hz (high band). Then, we computed several absolute and relative measures to quantify the amount of acoustic energy present in different parts of the spectrum (Table 1). For a thorough overview of this method including empirical justifications, see [13]. In the articulatory dimension, we examined the vertical (z-axis) and anterior-posterior (y-axis) displacements of the sensors glued to the lower incisors (JAW) and the tongue tip (TT).

Table 1: *Definition of acoustic parameters.*

Parameter	Definition
$AmpMin_{Low}$*	Minimum amplitude over the low band
$AmpPeak_{Low}$*	Peak amplitude within the low band
$AmpPeak_{Mid}$*	Peak amplitude within the mid band
$AmpPeak_{High}$*	Peak amplitude within the high band
$AmpD_{Mid-MinLow}$	$AmpPeak_{Mid}$–$AmpMin_{Low}$
$AmpD_{High-Mid}$	$AmpPeak_{High}$–$AmpPeak_{Mid}$
$Level_{Low}$	Sound level (RMS) over the low band
$Level_{Mid}$	Sound level (RMS) over the mid band
$Level_{High}$	Sound level (RMS) over the high band
$LevelD_{High-Mid}$	$Level_{High}$-$Level_{Mid}$
$LevelD_{Mid-Low}$	$Level_{Mid}$-$Level_{Low}$

2.6. Data extraction

For each trial, the acoustic centers of the sibilant segments contained in [lasə] and [tasə] were labeled automatically employing the signal's RMS ratio curve in combination with a set of heuristic rules. The quality of the automatic labelling procedure was examined visually and mislabeled trials were discarded from further analysis. Subsequently, based on the labeled land marks, the experimental measurements were extracted from the acoustic and articulatory signals of each response. To this end, the acoustic signal was first high pass filtered at 600 Hz to exclude any potential influence of accidentally occurring voicing. Then, power spectral densities (PSD) were computed with MATLAB's pwelch() function. The final acoustic data set consisted of 2394 unperturbed and 2394 perturbed [s]-tokens; the articulatory data set consisted of 1774 pairs of perturbed and unperturbed [s]-tokens since additional trials had to be discarded due to sensor detachment.

2.7. Data analysis

All analyses were performed in R (version 3.4.1) [18]. To understand the involvement of acoustic and articulatory dimensions in the adaptation process, we applied the RF algorithm to responses produced by participants under perturbed vs. unperturbed feedback using the implementation provided in the randomForest package (version 4.6-14) [15].

RF is a supervised ensemble classification technique with the objective to predict the values of a particular response variable, in our case the perturbation condition (perturbed vs. unperturbed), from a set of predictive variables, in our case the acoustic and articulatory measurements. In addition to classification, RF models provide an importance measure for each predictive variable defined as a loss of prediction accuracy of classification which occurs when a given variable is excluded from the model. By applying a feature selection procedure implemented in the Boruta package (version 5.3.0) [19], we intend to better understand their relevance for the adaptation process.

All modelling procedures were applied to normalized parameter values which were obtained by subtracting each participants' mean of each acoustic and articulatory parameter produced during the baseline phase in the respective perturbation condition. This was done to exclude participant-specific differences regarding their absolute frequency magnitudes and to improve the performance (i.e., computation time) of the classification algorithm.

3. Results

3.1. Feature selection

As expected, during the baseline phase, none of the seven acoustic nor the four articulatory parameters was deemed important by the Boruta procedure to discriminate between the [s]-tokens produced in [lasə] and [tasə]. The number of important features, however, increased in the course of the three shift phases, suggesting that speakers adjusted an increasing number of production parameters as the trials with perturbed [s]-sounds went on. Here, we present the results of the variable selection procedure computed for the last shift phase for acoustic and articulatory parameters.

Table 2: *Variable importance computation performed on the acoustic data from the last shift block.*

Parameter	Importance score	Relevance
$LevelD_{Mid-Low}$	10.57	important
$Level_{Mid}$	7.43	important
$Level_{Low}$	6.28	important
$Level_{High}$	5.07	important
$AmpD_{Mid-MinLow}$	5.01	important
$AmpD_{High-Mid}$	4.01	important
$LevelD_{High-Mid}$	3.21	important
shadowMax*	-0.09	

alpha = 0.01

*shadowMax is a "dummy" parameter computed by the Boruta procedure to determine the importance decision boundary.

As can be seen from Table 2, the acoustic parameter that was deemed most important to categorize perturbed and unperturbed [s]-tokens was the difference between the sound levels of the mid and low frequency bands. The change in this parameter suggests that participants tried to adapt to the applied perturbation by balancing the acoustic energy in both corresponding frequency regions.

Table 3: *Variable importance computation performed on the articulatory data from the last shift block.*

Parameter	Importance score	Relevance
TT_Z	15.82	important
JAW_Z	13.74	important
TT_Y	12.80	important
JAW_Y	12.39	important
shadowMax*	-0.1	

alpha = 0.01

*shadowMax is a "dummy" parameter computed by the Boruta procedure to determine the importance decision boundary.

In the articulatory dimension, parameters describing vertical tongue tip and jaw displacements (z-axis) were deemed more important compared to anterior-posterior displacements (y-axis) to classify [s]-tokens with regard to their perturbation condition (Table 3). The vertical displacements are expected as a consequence of a constriction location change between alveolar and postalveolar regions.

3.2. RF models

Based on variable importance scores computed separately for acoustic and articulatory data for every experimental phase, we ran corresponding RF models. To establish an initial prediction accuracy score, we ran models on the baseline data knowing that no perturbation was applied in this phase. In these cases all investigated parameters were used to compute the corresponding RF model since no particular acoustic or articulatory parameter was deemed relevant to predict the perturbation condition. The prediction accuracy scores for acoustic and articulatory RF models are summarized in Figure 3 and Figure 4, respectively.

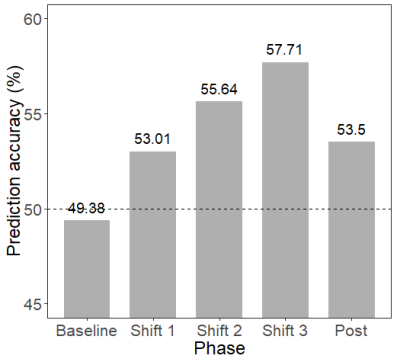

Figure 3: *Prediction accuracy scores of acoustic RF models across all experimental phases.*

During the baseline phase, the prediction accuracy stayed approximately at chance level for both models operating with acoustic (49.38%) and articulatory data (50.83%). This finding fits the idea that speakers' production of [s]-tokens contained in [lasə] were initially not different from the corresponding productions in [tasə].

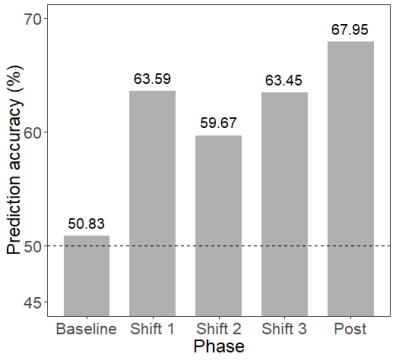

Figure 4: *Prediction accuracy scores of articulatory RF models across all experimental phases.*

In the course of the three shift phases, the prediction accuracy based on acoustic parameters increased incrementally to 57.71% hinting at progression of the adaptation process in the acoustic dimension. In other words, as the shift trials went on, speakers were increasingly adjusting the acoustic make-up of the perturbed [s]-tokens to match the unperturbed ones.

In the articulatory dimension on the other hand, the prediction accuracy increased to over 60% immediately after the baseline phase and remained approximately at the same level throughout all of the following three shift phases. This suggests that applied perturbation immediately impacted speakers' articulatory movements resulting in different production strategies for perturbed and unperturbed [s]-tokens. However, since there were no incremental increases of the prediction accuracy scores for the articulatory RF models, it seems that speakers were rather exploring the articulatory space than adapting to a certain production strategy. Correspondingly, the accuracy scores in the acoustic and the articulatory dimension do not match for the shift phases.

During the post phase, when auditory perturbation was no longer applied, the prediction accuracy in the acoustic dimension decreased to 53.5%, suggesting that speakers ceased to adapt to a certain acoustic goal. On the other hand, in the articulatory dimension the prediction accuracy increased to almost 68% maybe as a result of some kind of articulatory overshoot during the production of adapted [s]-tokens.

4. Conclusion

Preliminary results suggest that the applied spectral perturbation of the sibilant [s] had an immediate impact on speakers' articulation causing measureable articulator displacements. However, it took a certain amount of experimental exposure to the perturbation before speakers were able to adapt their speech in the acoustic dimension by balancing the acoustic energy in low and mid frequency regions of the perturbed sound. Interestingly, there was no congruency between adaptation effects observed in acoustic and articulatory dimensions which suggests that while speakers were adjusting the acoustic make-up of their speech towards a certain goal, they were rather exploring the articulatory space. In other words, it was not possible to infer the degree of adaptation in the acoustic dimension from the adaptation effects observed in the articulatory dimension. At this point, it remains open whether additional or other articulatory parameters can be identified which can serve as better indicators of the adaptation degree in the acoustic dimension. Overall, our current findings are consistent with the idea that speech sounds are perceptuo-motor units comprising of articulatory movements which are shaped by perceptual properties [6].

5. Acknowledgements

We gratefully acknowledge support by DFG grants 220199 and 247501188 to JB. We thank Megumi Terada for her support during data acquisition. We also thank all participants who took part in the study.

6. References

[1] J.F. Houde and M.I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[2] S. Tremblay, D.M. Shiller, and D.J. Ostry, "Somatosensory basis of speech production," *Nature,* vol. 423, no. 6942, p. 866–869, 2003.

[3] G. Hickok, J.F. Houde, and F. Rong, "Sensorimotor integration in speech processing: computational basis and neural organization," *Neuron,* vol. 69, no. 3, pp. 407–422, 2011.

[4] J.A. Tourville and F.H. Guenther, "The DIVA model: A neural theory of speech acquisition and production," *Language and Cognitive Processes*, vol. 26, no. 7, pp. 952-981, 2011.
[5] F.H. Guenther and G. Hickok, "Role of the auditory system in speech production," In *Handbook of Clinical Neurology*, vol. 129, pp. 161–175, Amsterdam: Elsevier B.V., 2015.
[6] J.L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.
[7] G. Hickok, "Computational neuroanatomy of speech production," *Nature Reviews Neuroscience*, vol. 13, no. 2, pp. 135–145, 2012.
[8] J.F. Houde and S.S. Nagarajan, "Speech production as state feedback control," *Frontiers in Human Neuroscience*, vol. 5, article 82, 2011.
[9] D.R. Lametti, S.M. Nasir, and D.J. Ostry, "Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback," *Journal of Neuroscience*, vol. 32, no. 27, pp. 9351–9358, 2012.
[10] Y. Feng, V.L. Gracco, and L. Max, "Integration of auditory and somatosensory error signals in the neural control of speech movements," *Journal of Neurophysiology*, vol. 106, no. 2, pp. 667–79, 2011.
[11] P. Perrier and S. Fuchs, "Motor Equivalence in Speech Production," *The Handbook of Speech Production*, pp. 225–247, Malden, MA: Wiley-Blackwell, 2015.
[12] J.S. Perkell, "Movement goals and feedback and feedforward control mechanisms in speech production," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 382–407, 2012.
[13] L.L. Koenig, C.H. Shadle, J.L. Preston, and C.R. Mooshammer, "Toward improved spectral measures of /s/: Results from adolescents," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 4, pp. 1175–1189, 2013.
[14] J. Brunner, P. Hoole, and P. Perrier, "Adaptation strategies in perturbed /s/," *Clinical Linguistics & Phonetics*, vol. 25, no. 8, pp. 705–724, 2011.
[15] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[16] A. Rochet-Capellan and D.J. Ostry, "Simultaneous acquisition of multiple auditory–motor transformations in speech," *Journal of Neuroscience*, vol. 31, no. 7, pp. 2657–2662, 2011.
[17] S. Cai, S.S. Ghosh, F.H. Guenther, and J.S. Perkell, "Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2033–2048, 2010.
[18] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2017.
[19] M.B. Kursa and W.R. Rudnicki, "Feature selection with the Boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.