ALIGNMENT OF PITCH AND ARTICULATION RATE

Lotte Eijk[1], Mirjam Ernestus[1] & Herbert Schriefers[2]

[1]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
l.eijk@let.ru.nl; m.ernestus@let.ru.nl; h.schriefers@donders.ru.nl

ABSTRACT

Previous studies have shown that speakers align their speech to each other at multiple linguistic levels. This study investigates whether alignment is mostly the result of priming from the immediately preceding speech materials, focussing on pitch and articulation rate (AR). Native Dutch speakers completed sentences, first by themselves (pre-test), then in alternation with Confederate 1 (Round 1), with Confederate 2 (Round 2), with Confederate 1 again (Round 3), and lastly by themselves again (post-test). Results indicate that participants aligned to the confederates and that this alignment lasted during the post-test. The confederates' directly preceding sentences were not good predictors for the participants' pitch and AR. Overall, the results indicate that alignment is more of a global effect than a local priming effect.

Keywords: alignment; pitch; articulation rate.

1. INTRODUCTION

Alignment (also often referred to as entrainment, convergence or accommodation) refers to the phenomenon that speakers adapt their speech to an interlocutor's speech on multiple levels (e.g. prosodic, phonetic, syntactic). Although alignment has been thoroughly investigated in the (recent) past, e.g. [3, 5, 7], many empirical questions are still open.

This study investigates whether alignment is mostly due to priming from the immediately preceding speech materials by addressing three questions. (RQ1) How long does alignment persist when the interlocutor is no longer present? If alignment exclusively results from adaptation to recent input, it should disappear rapidly. (RQ2) Do speakers align more rapidly to a speaker they have been talking to before? If alignment is exclusively driven by the immediately preceding input, this should not be the case. (RQ3) Do the features of the immediately preceding utterance predict how speakers adapt their speech in a given sentence?

We investigated these questions for both pitch and articulation rate, henceforth AR. By investigating two prosodic features, we can see in how far the results are feature specific, that is, whether and to what extent different prosodic features converge or differ in their alignment patterns.

Previous research has shown that both pitch, and AR are susceptive to alignment [3, 5, 7], although conflicting results have been reported for both features. For instance, research on pitch alignment by Gijssels et al. [5] has shown that speakers align their pitch to a confederate's pitch on a turn-by-turn basis (see also [7]), that the degree of alignment does not increase over time, and that alignment disappears immediately when the confederate is no longer present. In contrast, Bonin et al. [3] reported that pitch alignment fluctuates over time and that speakers do not always align in every turn. Research on AR alignment also shows conflicting results. For instance, whereas Levitan and Hirschberg [7] found alignment, Schweitzer and Lewandowski [10] found divergence in AR between speaker and interlocutor, though this effect was modulated by how much the participant liked the interlocutor.

We addressed our research questions in a sentence completion task consisting of five parts, which was originally designed to investigate other forms of alignment (phonological and syntactic). Participants first completed sentence beginnings by themselves (pre-test). Then, they alternated between sentence completion and listening to sentences completions from a confederate's pre-recorded speech. They did so, first with Confederate 1 (in Round 1), then with Confederate 2 (Round 2), and then with Confederate 1 again (Round 3). After these parts, they completed sentences by themselves again (post-test).

Our first question can be answered by comparing (the speed of change in) pitch and AR in the post-test with the other parts of the experiment. The second question can be addressed by comparing (the speed of change in) pitch and AR between Rounds 1 and 3 (the rounds with the same confederate). The third question can be addressed by testing whether the pitch or AR of a given sentence is predicted by the confederate's pitch or AR in the directly preceding utterance.

2. METHOD

2.1. Participants

Twenty-five female native Dutch speakers, aged 18 to 26 years (M = 22.4, SD = 2.1) participated in the

experiment. Participants received course credits or gift vouchers.

2.2. Materials

Two sets of materials were designed. The first set contained 268 Dutch sentence beginnings that had to be completed by the participants. These sentence beginnings were designed to elicit as much speech as possible. An example of a stimulus is shown in (1).

(1) Otto is een stuk vrolijker sinds…
'Otto has been a lot happier since…'

The second set of materials consisted of 198 complete Dutch sentences, which were uttered by the confederates and functioned as auditory primes. During the experiment, participants saw the beginnings of the confederates' full sentences on the computer screen. These beginnings were similar in length and grammatical structures to the sentence beginnings the participants had to complete. The two sets of stimuli included 205 stimuli that were adapted from Hartsuiker and Westenberg [6].

The complete sentences were recorded by the confederates in a sound-attenuated booth with a table-mounted Sennheiser K6/ME 64 microphone connected to a pre-amplifier and a Roland R-05 recorder. Speech was digitised at a sampling rate of 44.1 kHz, a 16-bit quantisation. Confederate 1 (23-year-old female) had an average median pitch of 224 Hz (ranging from 189 to 256) and an average AR of 5.0 syllables per second (ranging from 3.4 to 6.0), while Confederate 2 (24-year-old female) had averages of 215 Hz (ranging from 193 to 241) and 4.7 syllables per second (ranging from 3.4 to 6.5), see §2.4 for the measurement method.

Six pseudo-randomised stimuli lists were generated to make sure that, across participants, a given sentence (beginning) appeared in different parts of the experiment.

2.3. Procedure

Participants were tested in a sound-attenuated booth. The participants' speech was recorded using the same equipment as mentioned above. The confederates' speech was presented over Sennheiser HD 215 MKII DJ headphones.

Participants were presented with a sentence beginning via the Presentation software (Version 20.2, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com) in Times New Roman, font size 34, centered on the screen. They were instructed to read aloud the sentence beginning and to complete the beginning with whatever came to mind. In the pre- and post-test (both 35 trials), the participants completed the sentences by themselves. In Rounds 1 (60 trials), 2 (60 trials) and 3 (78 trials), the participants alternated with the pre-recorded speech from Confederate 1, Confederate 2, and Confederate 1, respectively. During these rounds, they saw the picture of the respective confederate on the screen.

Participants were asked to indicate for each sentence produced by the confederates, on a 7-point Likert scale, whether they would finish the sentence in the same way. This way we ensured that they paid attention to the confederates' speech. Instructions ('I would finish the sentence in the same way' plus the scale) were shown on the computer screen during confederates' trials. Participants were told that the confederates would rate their sentences as well. The experiment took less than one hour in total.

2.4. Measurements

Median pitch and articulation rate were calculated per sentence in Praat [2]. Median pitch was calculated with a script [8] which measured F0 values every 10 ms by using the *To Pitch...* command in Praat with a pitch range of 75 to 500 Hz. The script cleaned the raw values from errors resulting in pitch doubling and halving and from values based on speech produced with creaky voice by removing F0 values that were more than a factor of 1.5 bigger or smaller than the second to last F0 value. Then, the median F0 value per sentence was calculated. We removed all sentences with a minimum F0 lower than 110 Hz or a maximum F0 higher than 400 Hz. After deletion of these outliers, outliers more than 2.5 SD from the mean were deleted, resulting in 6230 data points for analyses (93.22% of the total).

The AR per sentence was calculated with a script [4] using the following parameters: a silence threshold of -25 dB (default), a minimum dip between peaks of 3 dB and a minimum pause duration of 0.3 seconds (default). The script divides the number of syllables (based on a number of syllable-related acoustic properties) of a sentence by the vocalisation time (the total time minus pauses). Outliers more than 2.5 SD from the mean were excluded, which resulted in 6588 data points for analyses (98.58% of the total).

2.5. Statistical analysis

Linear Mixed Effects models were performed in R [9] using the lme4 package [1]. Unless otherwise mentioned, our dependent variable was either the participant's median F0 or the AR per sentence. Fixed effects were ExperimentPart (EP) (pre-test, Round 1, Round 2, Round 3 and post-test) and EPtrialnr, which codes the sequential position of sentences within a given part of the experiment. We also tested for a potential quadratic trend of EPtrialnr, but adding the

quadratic predictor did not improve the models. We further tested for an interaction of the two fixed effects. Random effects were added for participant and sentence. For the final models, we removed data points deviating more than 2.5 SD from the predicted values. No random slopes were added for participant and sentence, because this caused non-convergence.

3. RESULTS

Figures 1 and 2 show the participants' median pitch and AR as a function of the trial number in the experiment. Different parts of the experiment are indicated by lines in different shades of grey. The figures also show the confederates' average pitch and AR, which were generally higher than the participants' pitch and AR.

Figure 1: Participants' median F0 over pre-test, Rounds 1, 2 and 3 and post-test; lines were fitted using lm. Points represent Confederates' means.

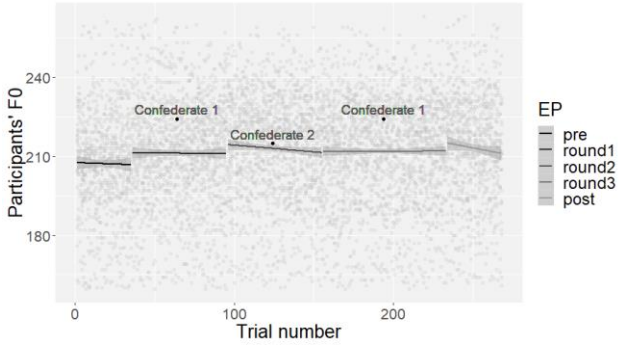
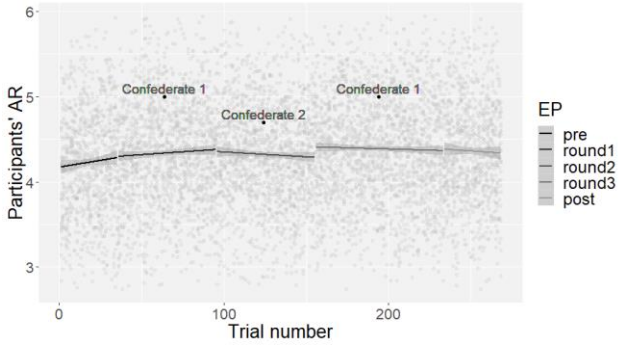


Figure 2: Participants' AR over pre-test, Rounds 1, 2 and 3 and post-test; lines were fitted using lm. Points represent Confederates' means.



3.1. RQ1: Difference between post-test and other parts

To see whether alignment lasts when the confederate is no longer present, we compared the post-test to the other parts of the experiment. If alignment lasts in the absence of the interlocutor, we would expect a significant difference between the pre-test and the post-test, reflecting that the participant's pitch and AR do not immediately return to the level of the pre-test. We would further expect no difference between Round 3 and the post-test if the alignment of Round 3 lasts in the post-test. Table 1 shows the results of the pitch model and Table 2 of the AR model, both with the post-test as the reference level.

Table 1: Pitch model with post-test as a reference.

Parameter	Estimate	SE	T value
Intercept	212.618	3.704	57.40
EPpre	-5.398	0.869	-6.21
EPround1	-2.784	0.795	-3.50
EPround2	1.261	0.789	1.60
EPround3	-0.650	0.754	-0.86
EPtrialnr	-0.061	0.030	-2.01
EPpre:EPtrialnr	0.031	0.042	0.75
EPround1:EPtrialnr	0.084	0.033	2.52
EPround2:EPtrialnr	0.022	0.033	0.67
EPround3:EPtrialnr	0.062	0.032	1.95

Table 2: AR model with post-test as a reference.

Parameter	Estimate	SE	T value
Intercept	4.414	0.063	70.30
EPpre	-0.247	0.052	-4.74
EPround1	-0.128	0.047	-2.72
EPround2	-0.056	0.047	-1.18
EPround3	-0.012	0.045	-0.26
EPtrialnr	-0.003	0.002	-1.77
EPpre:EPtrialnr	0.006	0.003	2.53
EPround1:EPtrialnr	0.005	0.002	2.37
EPround2:EPtrialnr	0.002	0.002	1.16
EPround3:EPtrialnr	0.003	0.002	1.46

Tables 1 and 2 show that participants did not immediately return to their habitual median pitch and AR in the post-test, as there are statistically significant differences between the pre-test and post-test. This is further supported by the lack of significant differences between the post-test and Round 3. Furthermore, participants gradually returned to their habitual pitch in the post-test as reflected in a significant effect of EPtrialnr within the post-test. This is not the case for AR.

3.2. RQ2: Difference between Round 1 and Round 3

To see whether speakers aligned more rapidly to Confederate 1 in Round 3 than in Round 1, we focussed on the differences between Round 1 and Round 3. If participants aligned more rapidly, i.e. within the first few trials, in Round 3 than in Round 1, this should result in an overall positive significant difference in median pitch and AR between Rounds 1 and 3. More rapid alignment could also be reflected in a positive statistically significant difference in the effect of EPtrialnr, i.e. an interaction between EPtrialnr and Round. Tables 3 and 4 show the models of Tables 1 and 2, with Round 1 as the reference.

Table 3: Pitch model with Round 1 as a reference.

Parameter	Estimate	SE	T value
Intercept	209.834	3.681	57.00
EPpost	2.784	0.795	3.50
EPpre	-2.615	0.778	-3.36
EPround2	4.045	0.661	6.12
EPround3	2.134	0.636	3.35
EPtrialnr	0.023	0.014	1.68
EPpost:EPtrialnr	-0.084	0.033	-2.52
EPpre:EPtrialnr	-0.053	0.033	-1.60
EPround2:EPtrialnr	-0.062	0.019	-3.30
EPround3:EPtrialnr	-0.022	0.016	-1.33

Table 4: AR model with Round 1 as a reference.

Parameter	Estimate	SE	T value
Intercept	4.286	0.058	74.18
EPpost	0.128	0.047	2.72
EPpre	-0.119	0.047	-2.55
EPround2	0.072	0.040	1.83
EPround3	0.116	0.038	3.06
EPtrialnr	0.001	0.001	1.82
EPpost:EPtrialnr	-0.005	0.002	-2.37
EPpre:EPtrialnr	0.005	0.002	0.86
EPround2:EPtrialnr	-0.002	0.001	-2.12
EPround3:EPtrialnr	-0.002	0.001	-1.97

Tables 3 and 4 show statistically significant differences between Rounds 1 and 3 for both pitch and AR. This could mean that speakers aligned very rapidly in Round 3, but see §4. We do not see positive values for the interaction between Round 3 and EPtrialnr. This means that participants did not align more rapidly throughout Round 3 than in Round 1.

There is one potential caveat to this pattern of results. Because Rounds 1 and 3 do not consist of the same number of trials (see §2.3 above), the differences between the rounds could simply be due to this length difference. To control for this possibility, we checked whether the results change when we only analyse the first 60 trials of Round 3 (so it contains the same number of trials as Round 1). This analysis did not show any important changes in the pattern of results.

3.3. RQ3: Locality of Pitch and AR alignment

We finally investigated whether participants aligned to the immediately preceding utterance produced by the confederate, i.e. whether they aligned on a turn-by-turn basis. We therefore added the median F0 or AR of the immediately preceding sentence produced by the confederate as a fixed predictor to the models discussed above. Furthermore, we analysed the data from only Rounds 1, 2, and 3, excluding trials with outlier values from the confederates. In these models, turn-by-turn alignment should be reflected as an effect of the pitch or AR of the preceding sentence produced by the confederate on the following participant's sentence. The models showed that the preceding median pitch and AR did not have a significant effect on the participants' pitch ($\beta = 0.012$, $t = 0.91$) and AR ($\beta = 0.005$, $t = 0.37$), indicating that alignment was not a local turn-by-turn effect.

We also studied locality of the alignment effects by analysing the difference between the participant's median F0 and AR and the confederate's median F0 and AR in the directly preceding prime. We tested the same models as in §3.1 and §3.2, but replaced the participants' F0 and AR by the absolute values of the difference scores. Results showed that there were no statistically significant effects of EPtrialnr for any of the three rounds. This suggests alignment on a turn-by-turn basis did not increase within any round.

4. DISCUSSION

We investigated alignment of two prosodic features. The main results are as follows. First, speakers do not immediately go back to their habitual pitch and AR when they no longer hear the interlocutor. This differs from the findings by Gijssels et al. [5], who found that participants' pitch immediately returns to a speaker's base value in the interlocutor's absence. Our results thus suggest that alignment has more long-lasting effects than suggested before.

Second, we saw a difference in overall pitch and AR between Rounds 1 and 3, with the same confederate. This could mean that participants aligned very rapidly, within the first few trials of Round 3, when they heard Confederate 1 again. Alternatively, it could be a spill-over effect from Round 2 (with a different confederate). This alternative could be tested, for example, by having participants finish sentences by themselves again in Round 2 instead of alternating with Confederate 2.

Lastly, unlike Gijssels et al. [5] and Levitan and Hirschberg [7], we did not find effects from the immediately preceding utterance. Taken together, these results indicate that alignment is not the exclusive result of immediate local priming from an interlocutor's preceding utterance, but rather a more global effect.

Although participants globally aligned to the confederates in both median pitch and AR, our data also show differences between median pitch and AR alignment (e.g. the effect of EPtrialnr in the post-test). Alignment of different prosodic features does thus not behave the same in all aspects in this experiment.

In conclusion, the present study suggests that prosodic alignment of pitch and AR is more than a local reaction to the acoustic characteristics of the immediately preceding utterance.

5. ACKNOWLEDGEMENTS

This work was supported by the Netherlands Organisation for Scientific research, through a gravitation grant 024.001.006 to the Language in Interaction Consortium.

6. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.<doi:10.18637/jss.v067.i01>.

[2] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/.

[3] Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., Campbell, N. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proc. INTERSPEECH* Lyon, 539–543.

[4] De Jong, N. H., Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2), 385-390.

[5] Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., Casasanto, D. 2016. Speech accommodation without priming: The case of pitch. *Discourse Processes*, 53(4), 233-251.

[6] Hartsuiker, R. J., Westenberg, C. 2000. Word order priming in written and spoken sentence production. *Cognition*, 75, B27-B39.

[7] Levitan, R., Hirschberg, J. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proc. INTERSPEECH* Florence. 3081-3084.

[8] Marcoux, K., Ernestus, M. 2019. Pitch in native and non-native Lombard speech. *Proc. 19th ICPhS* Melbourne.

[9] R Core Team 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[10] Schweitzer, A., Lewandowski, N. 2013. Convergence of articulation rate in spontaneous speech. In *INTERSPEECH* Lyon. 525-529.