Social Factors in Convergence of F1 and F2 in Spontaneous Speech

Antje Schweitzer and Natalie Lewandowski
Institute for Natural Language Processing, University of Stuttgart

Introduction. We present results on the influence of social factors on phonetic convergence of F1 and F2 values in spontaneous speech in German. Our data consists of 46 spontaneous conversations (approx. 20.8 hrs) between female German speakers on topics of their choice (approx. 25 minutes/dialog). Participants wore head-set microphones while talking to each other in a sound-attenuated booth. Speakers were provided with a list of potential topics to ease conversation, but were explicitly told that they were completely free to choose other topics as well. The recordings were automatically annotated on the segment and word level. As it is well accepted that convergence is influenced by social factors (e.g. Giles and Smith, 1979; Street, 1984; Pardo et al., 2012), speakers rated their conversational partner after each conversation in terms of likeability and competence.

Data processing. We extracted F1 and F2 values at vowel midpoint using Praat (Boersma and Weenink, 2013), and F0 by `get_f0` from the ESPS software package. We then removed outliers in terms of duration, F1, and F2. Duration outliers are usually indicative of problems in the automatic alignment. Finally, we excluded all cases where F1 and F0 were less than 100 Hz apart, except for the high vowels "i:", "y:", and "u:", for which F1 values in the range of female F0 could be expected. Outlier removal reduced the number of vowels in the analysis from 221,656 to 197,034. Visual inspection of F1 and F2 quantile-quantile plots revealed that after this step, both distributions were approximately normal.

Normalization. As the formant values of course are vowel-specific, we scaled and centered all formant values using vowel-specific means and standard deviations. Note that while this may sound reminiscent of Lobanov's (1971) speaker normalization procedure, our normalization technique is actually different: The aim in applying Lobanov's technique is to express the formant values in terms of their location in a specific speaker's vowel space. The aim of our technique is to express the formant values in terms of their location in the region that all recorded speakers used for this specific vowel. The parameters resulting from our transformation will be referred to as F1' and F2', respectively. A value of 0 for F1' for instance indicates that the respective vowel token was produced with an F1 that is exactly average for this vowel, while a value of 2 indicates that the vowel token was higher than average by two standard deviations. Thus the token is located at the upper edge of the distribution in terms of F1. In this way, F1' and F2' indicate each vowel token's position relative to all other tokens of the same vowel.

Statistical analysis. Our aim is to find out whether speakers' F1 and F2 values are influenced by their partners' F1 and F2 values. Specifically, if there was a positive relationship (i.e. if speakers produce higher values when confronted with higher partners' values), this would indicate convergence. Vice versa, a negative relationship would indicate divergence. To assess the relationship between partners' and speakers' F1' and F2' values, we performed two sets of linear mixed effects analyses using R (R Core Team, 2013) and the `lme4` package (Bates et al., 2013). The dependent variables were F1' and F2', respectively. If speakers converge to their partners, we would expect that partners' F1' and F2' productions are significant predictors of speakers' F1' and F2', and that the corresponding coefficients are positive. As it is not yet clear how much context is needed for speakers to converge, i.e., how many vowels must have been perceived before speakers' productions are affected, we do not want to make any assumptions as for exactly which of the partner's preceding tokens affect each produced vowel. Therefore, while we predicted F1' and F2' for each vowel token of a speaker, we averaged partners' F1' and F2' values across the whole dialog. These averaged values were used as predictors. To control

for random factors (for instance reduction effects due to stress, vowel duration, and word frequency, but also speaker-specific effects on vowel formants) we included intercepts for speaker, as well as by-vowel slopes for duration, stress, and word frequency. We then iteratively added the social factors and their interactions as fixed effects to both models, always confirming that including the factor was justified by way of likelihood ratio tests of the model with the factor in question compared to the model without the factor in question.[1] For the two winning models, we checked that all random effects were justified analogously, always removing the factor in question from the model. Visual inspection of the residual plots of these models revealed no obvious deviations from normality or homoscedasticity. To assess the significance of the fixed effects in the winning models, we used the `confint` function provided with the `lme4` package (using the "Wald" method) to approximate confidence intervals for the fixed effects. We chose a confidence level of 0.975 (Bonferroni correction for two tests, one for F1', one for F2'). We regard effects as significant if the estimated confidence interval does not contain the zero at this confidence level.

Results. Our results clearly confirm the relevance of social factors in convergence: The best model both for F1' and F2' was the model which included as predictors (i) partners' average F1' (or F2') scores, (ii) the likeability score for the partner (iii) the competence score for the partner, and (iv) their interactions. The coefficients for the two winning models exhibit similar patterns: in both cases, the effects of likeability and competence are in opposition: there is a positive coefficient for the interaction between likeability and partners' F1' (or F2'), i.e., the more a speaker liked her partner, the more influence has the partner's F1'/F2' on the speaker's productions. However in case of F1', the effect does not quite reach significance,[2] while in case of F2', the effect is significant at a level of 0.975. We find the opposite for the competence scores: for both F1' and F2' we observe negative coefficients for the interaction between competence and partners' scores, i.e., the more competent a speaker rated her partner, the lower the contribution of the partner's score in predicting the speaker's F1' or F2'. This effect was significant in both cases. In case of F1', there was also a small but significant positive effect of the three-way interaction between partners' average F1' and the competence and likeability scores. In addition, we observed a strong general convergence effect in case of F1', irrespective of the social ratings: we observed a high positive coefficient for the main effect of partner's score in case of F1'. This means that the default behavior across all dialogs was convergence of F1'. There was no significant effect like this in the F2' model, which indicates that the default was maintenance of F2'.

References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-4.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer (version 5.3.44) [computer program]. Retrieved from http://www.praat.org.

Giles, H. and Smith, P. M. (1979). Accommodation theory: Optimal levels of convergence. In Giles, H. and St. Clair, R., editors, *Language and Social Psychology*, pages 45–65. Blackwell, Oxford.

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608.

Pardo, J. S., Gibbons, R., Suppes, A., and Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40(1):190–197.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Street, R. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

[1] In all cases the better fit was also corroborated by lower AIC scores of the winning models.

[2] the estimated confidence interval is positive only at a confidence level of 0.95, but not at 0.975