Social Factors in Convergence of F1 and F2 in Spontaneous Speech

Antje Schweitzer and Natalie Lewandowski

Institute for Natural Language Processing, Stuttgart University, Germany
{antje.schweitzer,natalie.lewandowski}@ims.uni-stuttgart.de

Abstract

We present results on phonetic convergence of normalized F1 and F2 values in German spontaneous speech (46 dialogs, 20.8 hrs of speech). We are interested in the influence of social factors, specifically of mutual likeability and competence ratings, on convergence. To this end we fitted linear mixed models with speakers' F1 and F2 values, using partners' averaged values as well as the mutual social ratings as predictors. Our results show significant general convergence effects as well as significant effects of the interaction between partners' F1 and F2 values and the social ratings on speakers' productions of F1 and F2. This indicates that vowel formants are subject to phonetic convergence in spontaneous speech, and that social factors have an effect on the degree of convergence.

Keywords: phonetic convergence, vowel formants, spontaneous speech

1. Introduction

Phonetic convergence is the process of adapting one's speech to an interlocutor. The opposite, i.e., the assumption of a speaking style that differs from that of the interlocutor, is called divergence. In both cases, the perception of the interlocutor's speech affects a speaker's current production targets. An issue related to convergence, sometimes even considered equivalent to convergence, is imitation, which occurs when speakers' production targets are influenced by properties of stimuli that they have been exposed to before.

According to Communication Accommodation Theory (CAT, e.g. Giles and Smith 1979; Giles, Coupland, and Coupland 1991; Shepard, Giles, and Le Poire 2001; Giles and Ogay 2006), the adaptation seen in convergence or divergence is a dynamic process and affects not only speech, but communicative behavior in general (i.e. linguistic and phonetic features, but also paralinguistic aspects). CAT proposes that convergence decreases social distance between interlocutors and thus reflects a speaker's (often unconscious) need for social integration or identification with the interlocutor's social group (Giles, Coupland, and Coupland 1991). In contrast, divergence is caused by the need to distance oneself from the interlocutor's group. Interlocutors may also converge to increase intelligibility and efficiency of communication (Triandis 1960; Natale 1975; Gallois et al. 1995). Thus, social factors and the communication setting are clearly important when investigating convergence.

However, most recent studies on phonetic convergence use rather controlled and limited speech material, often drawing on methodology that is typically used in imitation research, without real conversational interaction, or focus on only specific target words or phrases in conversations (Babel 2010; Abrego-Collier et al. 2011; Kim, Horton, and Bradlow 2011; Babel 2012; Pardo et al. 2012). Few recent studies on convergence use larger-scale fully annotated corpora such as the Columbia Games Corpus (Levitan and Hirschberg 2011). In our opinion, testing the reality of convergence "in the wild" by investigating convergence on such corpora is indeed overdue, but there is one possible drawback in using game task corpora to this end: Given that efficiency of communication is a prerequisite for successfully playing such a game task, the question arises whether convergence in a game corpus may be a consequence of the game concept instead of a natural phenomenon in conversation.

Furthermore, social factors are assumed to be central in convergence, but to our knowledge there are no corpora to date which take social aspects of the conversation into account while providing data from completely free, spontaneous conversations. To close this gap, we have created the German Conversations (GECO) database,[1] which provides data on speakers' mutual social assessment (in terms of likeability and competence), in addition to large-scale fully annotated recordings of high audio quality. In this paper, we investigate convergence of vowel formants in this corpus.

2. Speech data

GECO consists of spontaneous conversations between previously unacquainted female German speakers on topics of their choice. Most speakers were students between age 20 and 30. Each dialog lasted approx. 25 minutes. Participants wore AKG HSC271 head-sets with rubber foam windshields while talking to each other in a sound-attenuated booth. We recorded about one half of the dialogs in a unimodal (UM) condition, where speakers could not see each other, and the other half in a multimodal (MM) condition, where speakers could see each other through a transparent screen. There are 22 dialogs (approx. 10.3 hours of dialog) in the UM condition and 24 (approx. 10.5 hours) in the MM condition. Subjects were naïve to the research questions; in both conditions, they were told that the purpose of the study was to research how small talk between strangers works. They 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. In fact, participants rarely consulted the list. The recordings were automatically annotated on the segment, syllable, word, and prosodic levels. The resulting corpus amounts to 20.8 hrs. of dialog, with approx. 250,000 words, 360,000 syllables, and 870,000 phones.

2.1. Social factors

As elaborated above, it is well accepted that the degree of accommodation (and its direction, i.e., divergence or convergence) is related to social factors (e.g. Giles and Smith 1979;

[1] The GECO corpus is freely available for non-commercial use at http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.en.html

Street 1984; Pardo et al. 2012). To cater for such social factors in the present database, speakers rated their, after each conversation by filling in a questionnaire. We captured likeability by four items in the questionnaires: Participants were asked how likeable ("sympathisch"), friendly ("freundlich"), socially attractive ("sozial"), and relaxed ("locker") they found their partner on a 5-point Likert scale. Competence was assessed by asking how intelligent, competent, successful, and self-confident the partner was perceived. We transformed all values to integers from -2 to +2. For both aspects, likeability and competence, we added the values for the four corresponding items to obtain a composite score for overall likeability and competence, respectively. Even though negative scores were rare in this experiment, both composite scores exhibit reasonable variation (both range from -2 to 8). Following the usual procedure in linear regression, we centered these raw scores for the statistical analysis below.

3. Methodology

3.1. Data processing

We extracted F1 and F2 values for each non-reduced monophthong in our data, along with vowel identity, duration, word stress, word frequency, speaker ID, listener ID, F0 at vowel midpoint (as calculated by get_f0 from the ESPS software package). For calculating F1 and F2 we used Praat (Boersma and Weenink 2014) to extract the first two formants using the "Burg" method, allowing a maximum of five formants, an expected maximum of 5500 Hz, a window length of 25 ms, and a pre-emphasis from 50 Hz in time steps of 25% of the window length, i.e. approx. 6 ms. We then sampled F1 and F2 at vowel midpoint by linear interpolation. To remove outliers, we filtered the data in the following way: First we removed duration outliers separately for each vowel in the standard way, by discarding all instances where the duration was more than 1.5 times the interquartile range away from the upper and lower quartiles, respectively. Duration outliers are usually indicative of labeling errors, which do occur because the data were annotated automatically using forced alignment. Then, we eliminated F1 and F2 outliers analogously, this time separately for each speaker and each vowel. Finally, we still observed suspiciously low F1 values especially for /a:/ vowels, which were close to the estimated F0 values, thus F0 may have been mistaken for F1 in these cases. Therefore, 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 212,677 to 173,025. Visual inspection of F1 and F2 quantile-quantile plots revealed that after this step, both distributions were approximately normal.

3.2. Normalization

As the formant values of course are vowel-specific, we scaled and centered (i.e. z-scored) 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, because we need the same reference frame for normalization in order to assess whether speakers used similar values. The parameters resulting from our transformation will be referred to as F1' and F2', respectively. Thus a value of 0 for F1' for instance indicates that the respective vowel token was produced with an F1 that is exactly average across all speakers for this vowel, while a value of 2 indicates that the vowel token was higher than this average by two standard deviations. Bear in mind that for normally distributed data, only 2.5% of the values are more than 2 standard deviations higher than the mean. As our F1 and F2 values were approx. normally distributed, we can then interpret an F1' value of 2 as indicating that approx. 97.5% of all tokens of that vowel were produced with a lower F1 than this token, 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 speakers' tokens of the same vowel. Figure 1 illustrates the normalization technique: it depicts the vowel space of all speakers in terms of F1 and F2. The boxes around each vowel indicate the region of one standard deviation above and below the mean. They can thus be interpreted as reference frames for normalization: Values falling at the edges of these boxes yield normalized values of +1 or -1. For instance, the yellow box highlights the region for all /a:/ vowels, and the red arrows then indicate the normalized axes for /a:/ vowels. A hypothetical example token located at the point indicated by the asterisk in Fig. 1 would then have an F2' of 0, which indicates that its raw F2 is equal to the mean for all speakers, and an F1' of approx. 0.75, which indicates that its raw F1 is relatively high compared to that of all other /a:/ tokens.

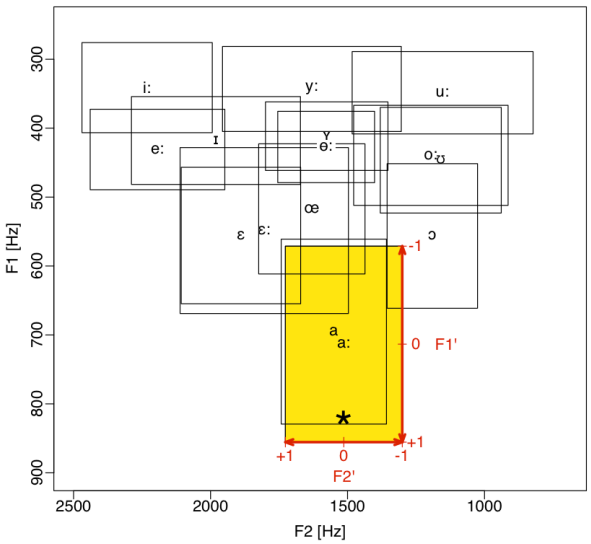

Figure 1: *Illustration of the formant normalization method. The boxes around each vowel indicate the range of values observed for all speakers: they indicate the region of one standard deviation above and below the mean. The yellow box highlights the region for /a:/ tokens, the arrows indicate the reference frame for their normalization, and the asterisk indicates a hypothetical example token.*

3.3. 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. A negative relationship on the other hand 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. 2014). 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'. 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 for every speaker and every dialog, we averaged partners' F1' and F2' values across that whole dialog. These averaged values were used as predictors. Thus F1' and F2' values of all vowel tokens of a speaker in a dialog were considered once individually as dependent variables, and then again indirectly when they contributed to the average F1' or F2', which then served as predictor variables for all vowel tokens of the other speaker in the same dialog. All F1' and F2' values as well as the averages were centered prior to fitting the models.

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. All random factors were justified, as confirmed by likelihood ratio tests for each factor, always comparing the model without the factor in question to the full model. This was done once at the beginning, including only partners' F1' or F2' averages as fixed factors. 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.[2] For the two winning models, we re-checked that all random effects were still justified for these richer models.

To assess the significance of the fixed effects in the winning models, we used the "Wald" method of the `confint` function provided by the `lme4` package (Bates et al. 2014). This function allows approximation of confidence intervals based on the estimated local curvature of the likelihood surface. 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 zero at this confidence level.

4. Results

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. Visual inspection of the residual plots of each winning model revealed no obvious deviations from normality and homoscedasticity.

Estimates for the coefficients in the two winning models are given in Table 1. They exhibit similar patterns. In both cases, we observed a general convergence effect, irrespective of the social ratings: we observed a positive coefficient for the main effect of partner's score (lines labeled *partner* in Table 1). The effect was more pronounced (i.e. with a higher coefficient) in case of F1' than in case of F2', but the effect was significant at a level of 0.975 in both cases. This means that the default behavior across all dialogs was convergence.

Table 1: *Coefficients of the winning linear mixed models. Estimates for the coefficients are in the second column, upper and lower bounds of the corresponding confidence intervals are listed in the third and fourth columns. The last columns indicates whether we consider the effect significant (*) or not significant (n.s.).*

F1' results				
Coefficient	estim.	upper	lower	sig.
(Intercept)	-0.439	-0.968	0.090	n.s.
partner	0.143	0.116	0.169	*
likeability	-0.005	-0.010	-0.001	*
competence	0.021	0.017	0.024	*
partner:likeability	0.043	0.029	0.056	*
partner:competence	-0.063	-0.077	-0.049	*
likeab.:competence	0.002	0.001	0.003	*
partner:likeab.:comp.	0.007	0.003	0.011	*

F2' results				
Coefficient	estim.	upper	lower	sig.
(Intercept)	0.082	-0.097	0.262	n.s.
partner	0.043	0.025	0.061	*
likeability	-0.006	-0.010	-0.002	*
competence	0.011	0.008	0.015	*
partner:likeability	0.014	0.003	0.025	*
partner:competence	-0.020	-0.031	-0.009	*
likeab.:competence	0.000	-0.001	0.001	n.s.
partner:likeab.:comp.	-0.003	-0.005	0.000	*

In addition, there are interactions of likeability and competence with partners' scores which are in opposition: there is a positive coefficient for the interaction between likeability and partners' scores (lines labeled *partner:likeability*), i.e., the more a speaker liked her partner, the more "influence" the partner's score had on the speaker's productions, i.e. the general convergence effect described above is strengthened with higher likeability scores. 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 (lines labeled *partner:competence*), 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', i.e. the general convergence effect is weakened for higher competence scores. This effect was also significant at a level of 0.975 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, while there was a negative effect in case of F2' (lines labeled *partner:likeab.:comp.*) .

It should be noted that we also observed main effects of the social ratings and their two-way interaction on speaker's F1' and F2', irrespective of the partner's score, which we found surprising. They indicate for instance that higher competence ratings for the partner raised speakers' F1' and F2' in general (lines labeled *competence*), and that higher likeability ratings for the partner lowered speakers' F1' and F2' in general (lines labeled *likeability*). We currently have no explanation for these findings.

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

5. Discussion

To summarize the results presented above, the main finding is that there is a general convergence effect both for F1' and F2', which is strengthened for increased likeability scores and weakened for increased competence scores. As mentioned above, our speakers' mutual ratings were mostly positive. Also, speakers usually indicated that they found the dialogs pleasant. Thus, the general convergence effect may be a consequence of the homogeneity of the participant group in terms of gender, age, and occupation. Irrespectively, our results confirm that convergence occurs naturally in fully spontaneous dialogs, and that it can be detected even using fully uncontrolled speech material.

On a more abstract level, the results clearly confirm the relevance of social factors in convergence: For both F1' and F2' including the social factors as fixed effects improved the fit of the models in all cases. In addition, the interactions between mutual social ratings and partner's F1' and F2' scores nicely demonstrate that social factors affect the degree of convergence. This suggests that accounts of phonetic convergence should acknowledge social factors and speaks against a purely biological account of convergence.

Concerning the asymmetry of likeability and competence effects, we can currently only speculate on possible causes. Even though the two variables were correlated (Pearson's r=0.70), there seem to be subtle differences between likeability and competence. We would argue that competence is a more competitive asset than likeability—people are more likely to compete with respect to competence than with respect to likeability. Some related evidence comes from investigating backchannel frequency in the GECO corpus using linear regression models (Schweitzer and Lewandowski 2012). We found that the more backchannels speakers produced, the more competent and likeable they found their partner, i.e. the effects of partners' likeability and competence on speakers' production of backchannels were symmetric. However, we also found that the more backchannels speakers produced, the less competent they tended to be rated themselves by their partners—there was a marginally significant (t(46)=-1.95,p=0.058,β=-0.37) negative relationship between how many backchannels speakers produced and how competent they were perceived by their partners. This effect was not present for likeability. Assuming that producing many backchannels is in a way similar to converging to the partner, both signaling appreciation in some way, it might be that the findings in the present paper are related to these earlier findings. It is possible that speakers are intuitively aware of the negative relationship between showing (maybe too much) appreciation and the impression of competence on the partner. Thus, when talking to a more competent partner, speakers might be inclined to be more subtle or careful in converging. As no adverse effect of showing appreciation on one's impression of likeability needs to be feared, and as we would not expect that conversation partners are competing with respect to likeability in the first place, there is no need for speakers to reduce convergence when talking to more likeable partners. We hope to shed more light on the asymmetric effects of likeability and competence in the future.

6. Acknowledgements

This study is part of the project *Phonetic Convergence in Spontaneous Speech* within the SFB 732 funded by the German Research Foundation (DFG).

7. References

Abrego-Collier, C., J. Grove, M. Sonderegger, and A. C. L. Yu (2011). "Effects of speaker evaluation on phonetic convergence". In: *Proceedings of the ICPhS XVII*, pp. 192–195.

Babel, M. (2010). "Dialect divergence and convergence in New Zealand English". In: *Language in Society* 39.4, pp. 437–456.

— (2012). "Evidence for phonetic and social selectivity in spontaneous phonetic imitation". In: *J. of Phonetics* 40.1, pp. 177–189.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-6. URL: `http://CRAN.R-project.org/package=lme4`.

Boersma, P. and D. Weenink (2014). *Praat: doing phonetics by computer (Version 5.3.64) [Computer program]*. Retrieved from `http://www.praat.org`.

Gallois, C., H. Giles, E. Jones, A. C. Cargile, and H. Ota (1995). "Accommodating intercultural encounters: Elaborations and extensions". In: *Intercultural communication theory*. Ed. by R. Wiseman. Thousand Oaks, CA: Sage, pp. 115 –147.

Giles, H., N. Coupland, and J. Coupland (1991). "Accommodation theory: Communication, context and consequence". In: *Contexts of Accommodation*. Ed. by H. Giles, N. Coupland, and J. Coupland. Cambridge University Press, pp. 1–68.

Giles, H. and T. Ogay (2006). "Communication Accommodation Theory". In: *Explaining communication: Contemporary theories and exemplars*. Ed. by B. Whaley and W. Samter. Mahwah, NJ: Lawrence Erlbaum, pp. 293–310.

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

Kim, M., W. S. Horton, and A. R. Bradlow (2011). "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance". In: *Journal of Laboratory Phonology* 2, pp. 125–156.

Levitan, R. and J. Hirschberg (2011). "Measuring Acoustic-Prosodic Entrainment with Respect to Multiple Levels and Dimensions". In: *Proceedings of Interspeech 2011*, pp. 3081–3084.

Lobanov, B. M. (1971). "Classification of Russian Vowels Spoken by Different Speakers". In: *The Journal of the Acoustical Society of America* 49.2B, pp. 606–608.

Natale, M. (1975). "Social desirability as related to convergence of temporal speech patterns". In: *Perceptual and Motor Skills* 40, pp. 827–830.

Pardo, J. S., R. Gibbons, A. Suppes, and R. M. Krauss (2012). "Phonetic convergence in college roommates". In: *Journal of Phonetics* 40.1, pp. 190–197.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Schweitzer, A. and N. Lewandowski (2012). "Accommodation of backchannels in spontaneous speech". In: *Booklet of the International Symposium on Imitation and Convergence in Speech*. Aix-en-Provence, pp. 65–66.

Shepard, C. A., H. Giles, and B. A. Le Poire (2001). "Communication Accommodation Theory". In: *The New Handbook of Language and Social Psychology*. Ed. by W. P. Robinson and H. Giles. John Wiley & Sons, pp. 33–78.

Street, R. (1984). "Speech Convergence and Speech Evaluation in Fact-Finding Interviews". In: *Human Communication Research* 11.2, pp. 139–169.

Triandis, H. C. (1960). "Cognitive similarity and communication in a dyad". In: *Human Relations* 13, pp. 175 –183.