CONVERGENCE OF HARMONIC VOICE QUALITY PARAMETERS IN SPONTANEOUS DIALOGUES

Antje Schweitzer[1], Wolfgang Wokurek[1], Manfred Pützer[2]

University of Stuttgart[1], Saarland University[2]
schweitzer@ims.uni-stuttgart.de[1], wokurek@ims.uni-stuttgart.de[1], puetzer@coli.uni-saarland.de[2]

ABSTRACT

This study investigates the adaptation of voice quality between dialogue partners in natural spontaneous conversations using data from a German database containing approximately 20 hours of speech from 46 dialogues. The voice quality parameters are spectral decay ratios and relative bandwidth of the first formant. Spectral decay rates are estimated using amplitudes and frequencies of the harmonic peaks of the first and second harmonics and the harmonics near the first four formants. Results of linear mixed effects models predicting speakers' current parameters from partners' preceding ones indicate that speakers converge, i.e. consistently adapt several of these parameters to those of their partner. In some cases, the adaptation depends on mutual likeability and competence ratings, and can be negative. In addition, voice quality parameters vary with these ratings in general, indicating that perception of the partner has an effect on speakers' voice quality irrespective of partners' voice quality.

Keywords: convergence, voice quality, mutual social perception, spontaneous speech

1. INTRODUCTION

For a number of years now there has been growing interest in the phenomenon of phonetic convergence. This phenomenon, sometimes also referred to as accommodation, alignment, or entrainment, with possibly slightly different conceptualizations related to the different terms, refers to the fact that speakers may adapt their style of speech to become more similar to that of an interlocutor. According to Communication Accommodation Theory (CAT), e.g. [9, 8], convergence decreases social distance between conversation partners and can signal identification with the conversation partner's social group [7].

There is also a considerable number of studies by now which have investigated convergence of various phonetic parameters in conversation, for instance formants [20, 21, 22], voice onset time [26], articulation rate [21, 13, 24], keyword duration [22], pitch [13], or spectral amplitude envelopes [14]. Some have looked at perceptual similarity as assessed by AXB tests rather than acoustic parameters (e.g., [19, 10]). In addition to these studies on conversational speech, a number of studies have confirmed speakers' ability to adapt these parameters in non-conversational settings such as naming or shadowing tasks (e.g., [17, 5, 3, 1, 2, 18, 6, 16]).

However convergence of voice quality (VQ) has received much less attention so far. To our knowledge, only [13] have looked at such parameters in a conversation setting. Specifically, they looked at jitter, shimmer, and harmonics-to-noise ratio (HNR) as well as other phonetic parameters unrelated to VQ. All in all they found some evidence of convergence of the VQ parameters, but less than for other phonetic parameters.

Hence the question whether VQ parameters are subject to phonetic convergence effects is not yet exhaustively answered. Voice quality is often considered a speaker-inherent parameter and thus the question arises if speakers can be expected to be influenced at all by a dialogue partner's VQ characteristics. On the other hand, [12] posits that some aspects of VQ, for instance harshness, can be imitated (and are used to index social information).

The present study contributes to this question by investigating convergence of VQ. We use two kinds of VQ parameters suggested by [29], viz. spectral decay rates and the relative bandwidth of the first formant. The spectral decay rates aim at the shape of the excitation spectrum, i.e. they mirror the glottal cycle. A rapid closing of the vocal folds e.g. causes a sharper voice and a whiter (flatter) spectrum than slower closing.

Such VQ parameters based on the harmonic spectrum were introduced as amplitude differences (in dB) by [27] and named after temporal phenomena such as open quotient (OQ), glottal opening (GO), rates of closure (RC) and skewness (SK). To reduce the influence of changes in fundamental frequency (F0) they are modified here to spectral decay gradients as suggested by [29]. An advantage of this set of VQ parameters is their noise robustness [15].

In addition to these parameters, the relative bandwidth of the first formant is intended to capture the damping of the vocal tract resonances, in particular of the first formant [28], caused by the open glottis.

2. DATA AND METHOD

2.1. Speech corpus

This study uses recordings of spontaneous conversations from the GECO database [24, 25]. This database comprises 46 conversations between German females, adding up to just over 20 hours of speech. They were recorded over headsets in an anechoic chamber. Speakers in that database were free to choose and switch topics in their conversations, there was no joint task to be carried out. The database provides annotations of all conversations on the phone, syllable, and word levels. Speakers rated each other after every conversation regarding various aspects of social attractiveness on 5-point Likert scales from +2 to -2. [24, 25] aggregated the ratings of friendliness, likeability, relaxedness, and social attractiveness to form a composite score for a broader concept of *likeability* with scores ranging from +8 to -8; likewise *competence* was aggregated from ratings of self-confidence, successfulness, intelligence, and competence. We adopt this procedure here.

2.2. Quantifying convergence

The present study assesses convergence by calculating linear mixed models [4] that predict a speaker's VQ parameters in each turn by the partner's ones from the preceding turn, including social ratings as additional factors. If speakers converge, i.e. if they adapt their VQ to become more similar to that of the partner, then the partner's VQ parameters should be significant predictors of the speaker's parameters, with a positive coefficient. A negative coefficient on the other hand would indicate divergence.

We expect that convergence is related to social factors, as posited by CAT, so the social scores should be included in interactions with the partner's preceding parameters. Positive coefficients for the interactions again indicate a positive effect: more similarity for higher than for lower social scores. In turn, negative coefficients indicate that there is less similarity for higher scores, i.e. divergence in case of higher social scores. For ease of interpretability we exclude the interaction between all three factors. Potentially the social scores could also affect a speaker's VQ in general, and they are also included as main factors. Hence the models we will use for assessing convergence below look as indicated in Eq. 1, where VQP is a place holder for the respective VQ parameter, $prec.VQP$ refers to the partner's parameter from the preceding turn, and *like* and *comp* refer to the likeability and competence score that the speaker has given to the partner.

$$\begin{aligned} \text{VQP} \sim\ & \text{prec.VQP} + \text{like} + \text{comp} \\ & + \text{like:prec.VQP} + \text{comp:prec.VQP} \\ & + (1|\text{speaker}) + (1|\text{partner}) \end{aligned} \tag{1}$$

The last two terms in Eq. 1 are random factors (intercepts) to account for speaker-dependent effects on VQ both for the speaker and the partner. Please note that both fixed factors and random terms were not determined by a model selection process but were chosen because they are theoretically motivated.

2.3. Voice quality parameters

The spectral decay rate parameters are based on amplitude and frequency measurements of several harmonic peaks, an estimate of F0, and formant parameter estimates [29]. Concretely, the first two harmonics (H1, H2) and the harmonics near the first four formants (A1P through A4P) are employed. The harmonic peaks are sought in a short term spectrum with a 25ms (Hamming) window. This window is long enough to show the spectrum of two or more fundamental periods in order to reveal the speech signal's harmonic structure. The analysis is repeated every 10ms.

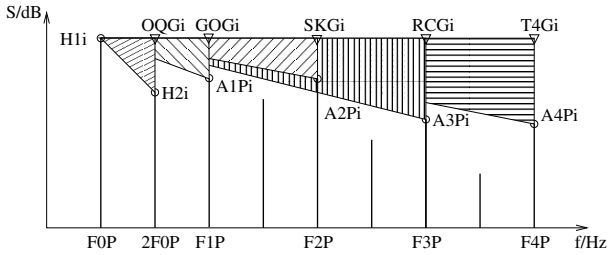
In a next step, the vocal tract resonance contribution is subtracted from these peak amplitudes. The compensation is indicated by symbols ending in i (for *inversely* filtered, as in H1i) in Fig. 1 and in the parameter names in the following.

Spectral decay gradients were introduced by [29] instead of amplitude differences to compensate changes in both F0 and formant frequencies at unchanged VQ: An increase of F0 shifts all harmonics to higher frequencies and also increases their distance. Assuming an unchanged slope of the spectrum of the voice source, the increased distance between the harmonics also increases their amplitude difference (e.g. if in Figure 1 the slope of the line connecting the peaks of $H1i$ and $H2i$ is unchanged, and the peak of $H2i$ occurs at a higher frequency, then its amplitude will necessarily be lower). By this mechanism the parameters, for instance $OQi = H1i - H2i$, increase just by increasing F0 without changing the phonation. The corresponding spectral gradient however, e.g.

$$OQGi = \frac{OQi}{\log_2 2F0P - \log_2 F0P}$$

stays constant: in case of OQGi, it is the decay slope of the leftmost triangle in Figure 1. The same argument applies to changes of each VQ parameter defined by harmonic amplitude differences. Figure 1 shows the appropriate spectral decay gradient triangles. Parameter names (OQGi through T4Gi) are indicated at the top right corner of each triangle.

Figure 1: Spectral decay gradient triangles. Vertical lines correspond to harmonics. Triangles visualize the spectral decay gradients, see text.



In calculating these parameters, we discard frames where the probability of voicing is below 50%, or where the harmonic structure is insufficient. Further we eliminate cases where the second harmonic coincides with the harmonic near the first formant, because OQGi and GOGi would be the same.

In addition to the spectral shape parameters, we employ the IC parameter instead of the CC parameter suggested by [27]. IC ("incompleteness of closure") is the relative bandwidth of the first formant and increases with the amount of glottal opening introduced in [15].

2.4. Data preprocessing

We estimated all parameters frame by frame (i.e. every 10 ms) within tense long vowels where the partner was silent. This yielded almost 280,000 frames.

All further steps were carried out in R [23]: Next we averaged each parameter over all frames pertaining to the same phone realization, keeping only these averaged values for each realization. These values were then scaled and centered separately for each phoneme category. The resulting values thus indicate whether a specific realization exhibited higher or lower values of that parameter than all other realizations pertaining to the same phoneme category. They allow for comparing values for different phones to one another irrespective of differences that may be a consequence of the two instances belonging to different phoneme categories. After this aggregation and normalization step, we were left with approx. 41,000 data points, each for one long tense vowel. Finally, we removed outliers for each parameter on a by-vowel basis, following the widely used strategy to remove data points that were more than 1.5 times the interquartile range higher (lower) than the upper (lower) quartile. This applied to no more than 50–200 data points for each parameter and thus did not significantly reduce the data further.

We then identified 6761 "turns" by detecting switches between speakers in the remaining data and calculated VQ averages for each parameter for each turn. In order to exclude turns that consist of only a backchannel, which in our experience are often more murmured than what is observed in fluent speech, we excluded turns where no more than 2 vowels had contributed in calculating the turn averages, leaving 3842 turns for statistical analysis.

3. RESULTS

For each VQ parameter, we fit a linear mixed model [4] to predict the values observed in the current turn by the VQ parameter of the partner from the preceding turn, as explained in Section 2.2, in Eq. 1. We used the implementation in *lmerTest* [11] to estimate Satterthwaite degrees of freedom and associated p-values. The VQ parameters are already scaled and centered after preprocessing as described above; the social scores were also scaled and centered for statistical analysis. Table 1 lists the results for all six VQ parameters. First cells in the gray rows indicate the name of the VQ parameter. In each subtable, *prec* refers to partners' values of the respective parameter in the preceding turn.

As discussed above, positive coefficients (column "Est."') for *prec* (if significant) indicate a general effect of convergence, and negative ones indicate divergence. Table 1 thus attests general convergence effects for OQGi, RCGi, T4Gi, and IC. For GOGi and SKGi, there is no significant effect. In no case do we observe a general effect of divergence. In addition to these general convergence effects, we find main effects of the mutual social scores: for all parameters, partners' perceived competence negatively significantly affects speakers' VQ parameters, as evidenced by the negative coefficients for *comp* in all cases, i.e. speakers "raise their voice" (exhibit a flatter spectrum) in conversations with partners that are perceived as more competent. We observe the opposite effect for perceived likeability (*like*) for 3 out of 5 spectral shape parameters (and no effect for OQGi and T4Gi). Interestingly, for *IC* the effect is the same as for competence, with lower *IC* values when talking to more likeable partners. Since *IC* is intended to capture glottal opening, this could suggest that speakers produce slightly less breathiness with both competent and likeable partners.

Table 1: Estimated (slope) coefficients resulting from fitting linear mixed models to predict the VQ parameters, along with Satterthwaite degrees of freedom (df), t-values, and p-values, as estimated by *lmerTest*. Coefficients and t-values were rounded to 2 digits; df to 4 significant digits. Intercepts are not indicated. Levels of significance are *** ($p<0.001$), ** ($p<0.01$), and * ($p<0.05$).

OQGi	Est.	df	t	p
prec	0.04	3648	2.45	0.01*
comp	-0.05	1930	-4.39	0.00***
like	0.02	809.5	1.66	0.10 n.s.
prec:comp	0.01	3817	0.44	0.66 n.s.
prec:like	-0.02	3813	-0.76	0.45 n.s.
GOGi	Est.	df	t	p
prec	0.02	3534	1.44	0.15 n.s.
comp	-0.05	1784	-3.98	0.00***
like	0.04	734.8	2.82	0.00**
prec:comp	0.04	3823	2.08	0.04*
prec:like	-0.04	3821	-2.14	0.03*
SKGi	Est.	df	t	p
prec	0.02	3681	1.14	0.26 n.s.
comp	-0.08	3000	-6.14	0.00***
like	0.03	1704	2.29	0.02*
prec:comp	0.06	3814	2.68	0.01**
prec:like	-0.05	3825	-2.2	0.03*
RCGi	Est.	df	t	p
prec	0.06	3511	3.75	0.00***
comp	-0.1	2845	-7.27	0.00***
like	0.05	1589	3.42	0.00***
prec:comp	0.02	3810	0.91	0.36 n.s.
prec:like	-0.01	3825	-0.29	0.77 n.s.
T4Gi	Est.	df	t	p
prec	0.07	3415	4.87	0.00***
comp	-0.06	3124	-4.2	0.00***
like	0.02	1845	1.5	0.13 n.s.
prec:comp	0.04	3815	2.24	0.03*
prec:like	-0.04	3822	-2.04	0.04*
IC	Est.	df	t	p
prec	0.06	3748	4.25	0.00***
comp	-0.05	3202	-4.12	0.00***
like	-0.03	1968	-2.28	0.02*
prec:comp	-0.04	3818	-1.81	0.07 n.s.
prec:like	0.03	3827	1.31	0.19 n.s.

As for the interactions of *prec* with *comp* and *like*, these are significant for 3 spectral parameters (GOGi, SKGi, T4Gi). In the case of *prec:comp*, the slope is positive indicating that the higher the competence of the partner, the higher the contribution of the preceding VQ parameter, i.e. the stronger the convergence effect. In contrast, the slope for the interaction with *prec:like* is negative but in most cases smaller than that for *prec:comp*, indicating that if the likeability scores are higher than the competence scores, divergence may be observed.

It should be noted that the observed effects are significant but small: the parameters have been standardized before statistical analysis, thus a coefficient of, say, 0.05 indicates that when a conversation partner produces a VQ parameter that is 1 standard deviation above (or below) the mean, the speaker is predicted to raise (or to lower) their parameter by 5% of the standard deviation. We claim that this is because there are many other factors that affect VQ that we do not consider here, such as the segmental context, positional prosodic factors, stress, and maybe also paralinguistic factors beyond speaker identity (which we cater for through the random effects).

4. DISCUSSION & CONCLUSION

In contrast to [13], we observe small but consistent turn-based convergence effects for most VQ parameters, while [13] found that only convergence of shimmer was (marginally) significant when averaging over the whole session, and only convergence of jitter was significant at turn-level. At turn-level, they observed synchrony, but not convergence, for the three VQ parameters. This difference may be due to the fact that we investigated VQ via parameters that depend on the spectral shape, while they investigated jitter, shimmer, and HNR. Possibly these parameters are less easily changed dynamically than the spectral parameters we use. Our parameters reflect phonation details of the glottal cycle such as the extent of modal voice quality and might be adapted more easily. It should be noted however that [13] used both another method and other data than the present study, so a direct comparison is difficult. For instance the earlier study assessed convergence in mixed as well as same-gender conversations, while the present study investigates only female-female dialogues (but a study on mixed-gender dialogues is planned for the future). The earlier study used correlations to assess convergence, while we use linear mixed models.

In any case this study presents new findings that indicate that the social ratings (i) affect VQ directly, irrespective of convergence or divergence effects, and (ii) that they affect the degree of convergence, as would be expected for instance from the perspective of CAT. It also confirms and strengthens the hypothesis that VQ parameters are in general subject to convergence effects, similar to other phonetic parameters that have been observed to be adapted in conversation.

5. REFERENCES

[1] Babel, M. 2010. Dialect divergence and convergence in New Zealand English. *Language in Society* 39(4), 437–456.

[2] Babel, M., Bulatov, D. 2011. The role role of fundamental frequency in phonetic accommodation. *Language and Speech* 55(2), 231–248.

[3] Babel, M. E. 2009. *Phonetic and Social Selectivity in Speech Accommodation*. PhD thesis University of California Berkeley.

[4] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[5] Delvaux, V., Soquet, A. 2007. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica* 145–173.

[6] Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., Steiner, I. 2018. Convergence of pitch accents in a shadowing task. *Proceedings of Speech Prosody 2018, Poznań* 225–229.

[7] Giles, H., Coupland, N., Coupland, J. 1991. Accommodation theory: Communication, context and consequence. In: Giles, H., Coupland, N., Coupland, J., (eds), *Contexts of Accommodation*. Cambridge University Press 1–68.

[8] Giles, H., Ogay, T. 2006. Communication accommodation theory. In: Whaley, B., Samter, W., (eds), *Explaining communication: Contemporary theories and exemplars*. Mahwah, NJ: Lawrence Erlbaum 293–310.

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

[10] Kim, M., Horton, W. S., Bradlow, A. R. 2011. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Journal of Laboratory Phonology* 2, 125–156.

[11] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13), 1–26.

[12] Laver, J. D. M. 1968. Voice quality and indexical information. *British Journal of Disorders of Communication* 3(1), 43–54.

[13] Levitan, R., Hirschberg, J. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* 3081–3084.

[14] Lewandowski, N. 2011. *Talent in nonnative phonetic convergence*. Doctoral dissertation, Universität Stuttgart.

[15] Lugger, M., Yang, B., Wokurek, W. May 2006. Robust estimation of voice quality parameters under realworld disturbances. *2006 IEEE International Conference on Acoustics Speech and Signal Proces sing Proceedings* volume 1 1097–1100.

[16] Michalsky, J., Schoormann, H. 2018. Opposites attract! Pitch divergence at turn breaks as cause and effect of perceived attractiveness. *Proceedings of Speech Prosody 2018, Poznań* 265–268.

[17] Namy, L., Nygaard, L., Sauerteig, D. 2002. Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology* 21(4), 422–432.

[18] Nielsen, K. 2011. Specificity and abstractness of vot imitation. *Journal of Phonetics* 39(2), 132–142.

[19] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4), 2382–2393.

[20] Pardo, J. S. 2010. Expressing oneself in conversational interaction. In: Morsella, E., (ed), *Expressing oneself/expressing one's self: Communication, cognition, language, and identity*. London: Taylor & Francis 183–196.

[21] Pardo, J. S., Cajori Jay, I., Krauss, R. M. 2010. Conversational role influences speech imitation. *Attention, Perception, & Psychophysics* 72(8), 2254–2264.

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

[23] R Core Team, 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[24] Schweitzer, A., Lewandowski, N. 2013. Convergence of articulation rate in spontaneous speech. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)* 525–529.

[25] Schweitzer, A., Lewandowski, N. 2014. Social factors in convergence of F1 and F2 in spontaneous speech. *Proceedings of the 10th International Seminar on Speech Production, Cologne* 391–394.

[26] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Perception & Psychophysics* 66(3), 422–429.

[27] Stevens, K. M., Hanson, H. M. 1998. Classification of glottal vibration from acoustic measurements. In: Fujimura, O., Hirano, M., (eds), *Vocal Fold Physiology*. Cambridge MA: Hiltop University Press 147–170.

[28] Stevens, K. N. 1998. *Acoustic Phonetics*. Current Studies in Linguistics. Cambridge, Massachusets: The MIT Press.

[29] Wokurek, W., Pützer, M. 2003. Automated corpus based spectral measurement of voice quality parameters. *Proc. 15th ICPhS (Barcelona)* 2173–2176.