Modeling of Between-Speaker and Within-Speaker Variation in Spontaneous Speech Tempo

Hugo Quené

Utrecht institute of Linguistics OTS
Utrecht University, Utrecht, The Netherlands
hugo.quene@let.uu.nl

10.21437/Interspeech.2005-754

Abstract

Speech tempo (speaking rate) varies both between and within speakers. Previous research suggests several relevant factors and predictors. The present study investigates all these factors combined, both between and within speakers, in a large corpus of spoken Dutch interviews. This is done by means of multi-level modeling of sex, age, and dialect region (all between speakers) and phrase length and sequential position of phrase within session (both within speakers). Results show that speech tempo depends mainly on phrase length, and not on between-speaker factors sex, age, or dialect region. Within-speaker tempo variations exceed the JND. Separate modeling of phrase length itself reveals significant negative effects of age and of sequential position, but not of region or sex. When taken together, these results underline the phonetic and communicative importance of within-speaker variations in speech tempo.

1. Introduction

Human speech is produced by moving the vocal organs and articulators. These movements result in an articulated speech signal, in which phonetic events occur at particular moments in time. The rate at which these speech events occur constitutes the tempo or speed or rate of speech. This tempo is often expressed in syllables per second [1] or in average syllable duration (ASD) [2, 3].

Many textbooks in phonetics state that speakers vary their speaking rate, in anticipation of the time listeners will need to process their words. Hence, important or unpredictable portions are spoken at a relatively slower rate, e.g. [4]. Obviously, the segmental content also affects speech tempo; if syllables contain fewer speech sounds, then a speaker can produce more of such syllables per time unit.

Recent studies, using the same Dutch corpus material, have reported on several other factors that may affect speech tempo, viz. age, sex, and region. Speakers' average tempo is reported to be faster for younger speakers than for older speakers, and faster for men than for women [5, 6]. For speakers of Dutch from the Netherlands, four dialect regions were investigated [5]. The West region (Zuid-Holland) is considered the linguistic center of the Netherlands. The Mid region (Utrecht, Gelderland) is a transition zone. The North (Groningen, Drenthe) and South (Limburg) regions have distinct regional dialects, although the "western" variety of Standard Dutch is widely used. Speakers' average tempo is found to be related to their distance from the "linguistic center": average tempo was fastest in the West region, intermediate in the Mid region, and slowest in the North and South regions [5].

These analyses of the corpus material were limited to between-speaker variation in tempo. For communicative purposes, however, within-speaker changes in tempo are at least as important, if not even more. The present research extends previous attempts at modeling speech tempo in Dutch, by including within-speaker predictors in the model.

First, longer phrases, containing more syllables, tend to be spoken at a faster rate, with shorter average syllable durations; this is known as 'anticipatory shortening' [7, 8, 9, 10]. A plausible model of speech tempo should therefore include phrase length as a factor.

The speech corpus under analysis consists of interviews with high-school teachers of Dutch language and literature. Each interview lasts about 15 minutes. During the interview, speakers may gradually speed up (due to arousal, etc) or slow down (due to fatigue, etc). Hence, the sequential position of a phrase may constitute a second within-speaker predictor of the speech tempo within that phrase.

The first aim of this study is to adequately model speakers' variation in tempo, both between and within speakers, by means of a large corpus of spontaneous speech. In addition, the second aim is to estimate the within-speaker variation in tempo. If tempo changes are indeed relevant for speech communication, as stated above, then these changes (and hence within-speaker variance) should exceed the just noticeable difference for speech tempo. This JND is reported to be about 5% change in tempo [6]. The second aim is therefore to confirm that a speaker's tempo changes are indeed noticeable.

2. Method

The Corpus of Spoken Dutch [11] was used to investigate which factors contribute to variation in speaking rate. For this purpose, we concentrated on the sub-corpus containing interviews with $N = 80$ high-school teachers of Dutch in the Netherlands [12]. Interviewed speakers ('interviewees') were stratified by dialect region (four regions within the Netherlands), sex, and age group (below 35 vs. over 45 years of age), with $n = 5$ speakers in each cell. At present, this paper reports data from 58 out of 80 speakers, distributed over regions as follows: West 16, Mid 13, North 13, South 16. All speakers are assumed to speak a variety of Standard Dutch as used in the Netherlands. All interviews were conducted by the same interviewer (male, age 26), and similar topics were discussed across interviews. Hence, language variety, conversation partner, and conversation topic were eliminated as confounding factors, and the speech samples were highly comparable among speakers.

For each interview, the orthographic transcripts were extracted from the annotations provided, broken down by inter-pause chunks. The speaking time of each chunk or phrase was

determined from the time marks in the transcript. The number of orthographic syllables in each phrase was determined by dictionary look-up of the orthographic words (with manual correction where necessary). Speaking rate is expressed here as average syllable duration (ASD) [2, 3]. Inter-pause chunks or phrases, as determined by the original transcribers, constitute the units of observation. Most of the short phrases (of 1 or 2 orthographic syllables) consisted of hesitation sounds, filled pauses, backchannel sounds, etc. These were excluded from the data set [3]. The proportion of excluded phrases per speaker ranged from .01 to .23 (median .10, inter-quartile range .08). In total, $N = 20886$ phrases from 58 speakers were analyzed.

Response variables (e.g. syllable durations) were modeled by means of multi-level analysis [13, 14], with speakers and phrases as two nested random factors. This type of analysis has several important advantages over more conventional techniques such as repeated measures ANOVA, or linear regression (see [15] for a longer review and tutorial). First, it allows for multiple nested random effects, such as phrases within speakers. Second, multi-level modeling does not require homogeneity of variance, nor sphericity. Between-speaker s_u^2 and within-speaker variances s_e^2 variances are modeled explicitly, instead of assumed to be constant everywhere. Third, multi-level modeling allows for incomplete designs, and for varying numbers of observations per cell.

Before multi-level modeling, the 4 levels of the region factor were converted to 4 binary (dummy) factors. Sex was included as a binary dummy factor (0 female, 1 male). Age was not included as a classification factor discriminating two speaker groups (like [5, 6]), but as a linear predictor, centralized to the mean age of 44 before modeling [13, 14].

In any type of statistical modeling, the aim is to obtain a model that contains the least number of predictors but explains the highest variance of the dependent variable. In multi-level modeling, this is complicated somewhat because the optimal model can be different for the fixed part, and for the random parts at each level. Therefore one may find different predictors in the various parts of the models. For each model reported below, the fixed part contains regression coefficients (β), and the random parts contain amounts of variance. Only final, optimal models are reported, for the sake of clarity.

3. Results

3.1. Speech tempo

First a model was fitted that was somewhat similar to the between-speaker ANOVA reported by [5], including only region, age, and sex. Phrase length and sequential position are not included. Contrary to ANOVA models, however, within-speaker variances need not be homogeneous in the present analysis. Instead, variances are modeled explicitly, which allows us to investigate the effects of sex, age and region on these variance components [14, 15]. Results for this first model (1) are listed in the lefthand part of Table 1.

Results for this model (1) confirm previous analyses of speakers' average tempo in this corpus [5, 6]. First, comparisons of the four regional means show that speakers from the West region (the linguistic center of the Netherlands) produced significantly shorter syllables (i.e. faster tempo) than did those from the other regions ($\chi^2 = 14.6, df = 3, p = .002$). Second, male speakers produced significantly shorter syllables (i.e. faster tempo) than female speakers [5]. Thirdly, the results show that older speakers produce significantly longer syllables (i.e. slower tempo) than younger speakers. For each additional year of age, ASD increases with 0.5 ms. With a grand mean ASD of 223 ms, the tempo difference between speakers aged 25 and 65 is $40 \times 0.5/223$ or about 9%.

Table 1: *Estimated parameters (with standard error of estimate in parentheses) of multi-level modeling of syllable durations (in ms). Significant parameters are printed in boldface.*

	Model 1		Model 2	
fixed				
reg.West	221.4	(4.7)	225.9	(10.9)
reg.Mid	237.6	(5.2)	247.0	(11.9)
reg.North	242.9	(5.1)	255.0	(11.6)
reg.South	236.4	(4.7)	231.8	(10.8)
sex	**-17.5**	(4.3)	-7.2	(9.9)
age	**0.5**	(0.2)	-0.3	(0.5)
log phr length			**-139.3**	(6.3)
seq position			**-0.012**	(-0.003)
random				
$\sigma^2_{u_{0j}}$	**232.**	(48.0)	**1311.**	(246.8)
$\sigma^2_{u_{\text{length }0j}}$			**2256.**	(427.8)
$\sigma^2_{e_{\text{W }ij}}$	7979.	(177.2)	3763.	(81.5)
$\sigma^2_{e_{\text{M }ij}}$	8404.	(187.5)	3630.	(80.7)
$\sigma^2_{e_{\text{N }ij}}$	11270.	(287.2)	4308.	(112.8)
$\sigma^2_{e_{\text{S }ij}}$	7911.	(191.4)	3766.	(87.9)
$\sigma^2_{e_{\text{age }ij}}$	**5.32**	(1.17)	**1.62**	(0.51)
deviance	249810		232898	

The multi-level analysis also allows us to inspect the between-speaker and within-speaker variances in syllable durations, in the random part of Table 1. between-speaker variances σ_u^2 are remarkably low, which indicates that speakers' averages are relatively similar across regions, ages and sexes.

The average within-speaker variance in this model is $s_e^2 = 8601$ ($s = 93$ ms), yielding a coefficient of variation of .42, which is indeed far larger than the JND of about 5%, Q.E.D.

But the within-speaker variances also show interesting effects of region and age. This indicates that within-speaker (error) variances are in fact not homogeneous. Comparisons of the four regional within-speaker variances show that speakers from the North region (Groningen–Drenthe) varied their tempo significantly more than did speakers from other regions ($\chi^2 = 143.2, df = 3, p < .001$). In addition, the within-speaker variance increases significantly with older age: older speakers vary their speech tempo more than younger speakers do.

This first model was extended by including as predictors both *phrase length* (in orthographic syllables, converted to log units, and centralized to the mean log length) and *sequential position* of each phrase within its interview. Results for this model (2) are listed in the righthand part of Table 1.

First, results confirm that phrase length indeed has a large and highly significant effect on speaking rate, as known from previous research [7, 8, 9, 10]. Speakers produce longer phrases with shorter average syllable duration, hence with faster speech tempo. Secondly, the regional differences in speech tempo disappear, if phrase length is included as a predictor in the model. Although speakers from the West region still produce the shortest syllable durations, the main effect of region is no longer significant ($\chi^2 = 5.8, df = 3, p = .12$). Thirdly, the effects of the sex and age factors are no longer significant, if phrase length is

included in the model. Fourthly, the effect of sequential position is extremely small, although significant. Speakers tend to speed up by a small amount (yielding shorter syllable durations) during the 15-minute interview. However, because this position effect is extremely small, the major result for model (2) is that differences in speech tempo are adequately modeled by phrase length as the only predictor.

Again, the multi-level analysis allows us to model between-speaker and within-speaker variances explicitly, rather than assuming that these are homogeneous. In fact, they are not. First, between-speaker variance significantly increases with phrase length. This means that the tempo differences between speakers increase for longer phrases: speakers' tempi differ more in longer phrases than in shorter phrases.

The within-speaker variances in model (2) are about half of those in model (1), due to the inclusion of extra predictors in the model. Phrase length now explains a lot of within-speaker variance, so that considerably less error variance remains. The remaining within-speaker variance in this model (2) is $s_e^2 = 3829$ on average, corresponding to coefficient of variation of .28. This is still far larger than the JND of about 5%, Q.E.D.

As in model (1), speakers from the North region (Groningen–Drenthe) varied their tempo significantly more than did speakers from other regions ($\chi^2 = 30.9, df = 3, p < .001$). Likewise, within-speaker variance again increases significantly with older age: older speakers vary their speech tempo more than younger speakers do. Older speakers are somewhat slower (although not significantly so), but they vary their tempo more than younger speakers do. Other variance components in the random parts were not significantly different from zero, and have been excluded.

In summary, these results suggest that average speech tempo is similar for speakers from various regions within the Netherlands, if tempo measurements are corrected for phrase length. In turn, this raises the question whether previously reported regional differences between speakers in speaking rate [5] may have been due to regional differences in phrase length. This was investigated by multi-level modeling of phrase length as the dependent variable.

3.2. Phrase length

Phrase length was also modeled by means of multi-level analysis, similar to the analysis of speech tempo above. For each phrase, the observed length in orthographic syllables was converted to log units; resulting data were centralized to the grand mean of 2.22 log units. Speakers and phrases were two nested random factors. The optimal model for the phrase length data is summarized in Table 2.

The fixed part of this model suggests that the average phrase length is similar for speakers from different regions within the Netherlands, yielding a nonsignificant main effect of region ($\chi^2 = 1.74, df = 3, n.s.$). The effect of age is marginally significant (Wald $Z = 1.94, p = .052$), which indicates a tendency for older speakers to produce somewhat shorter phrases (containing fewer syllables) than younger speakers do. The log of phrase length decreases by 0.0049 for each one-year increment of age. Third, male and female speakers produce phrases of similar length, as indicated by the insignificant coefficient for the sex factor. Fourth, the effect of sequential phrase position is also significant. As a speaker progresses in the interview, he or she produces somewhat shorter phrases.

In the random part, it may be noticed that the between-speaker variation σ_u^2 is far smaller than the within-speaker variation σ_e^2. Speakers do not differ very much in their *average* phrase length. This explains why regional (between-speaker) differences in phrase length were not significant in this model: only a small amount of the total variance is due to differences between speakers (intra-class correlation $\rho_I = .099$). Only part of that small amount can be attributed to regional differences between speakers; this results in an insignificant main effect.

Table 2: *Estimated parameters (with standard error of estimate in parentheses) of multi-level modeling of log of phrase length (in orthographic syllables). Significant parameters are printed in boldface.*

fixed		
reg.West	-0.0209	(0.0536)
reg.Mid	-0.0102	(0.0588)
reg.North	0.0409	(0.0570)
reg.South	-0.0923	(0.0536)
sex	0.0839	(0.0487)
age	-0.0049	(0.0025)
seq position	-4.5×10^{-5}	2.3×10^{-5}
random		
$\sigma^2_{u_{0j}}$	**0.0317**	(0.0060)
$\sigma^2_{e_{\mathrm{W}\,ij}}$	0.2663	(0.0067)
$\sigma^2_{e_{\mathrm{M}\,ij}}$	0.2263	(0.0059)
$\sigma^2_{e_{\mathrm{N}\,ij}}$	0.2303	(0.0071)
$\sigma^2_{e_{\mathrm{S}\,ij}}$	0.2649	(0.0070)
$\sigma^2_{e_{\mathrm{age}\,ij}}$	17.0×10^{-5}	3.6×10^{-5}
$\sigma^2_{e_{\mathrm{seq.pos}\,ij}}$	2.0×10^{-7}	0.3×10^{-7}
deviance	33239	

Regional differences in within-speaker variation in phrase length turn out to be unstable across several analyses (not reported here), and they may reflect artefacts in the present data set rather than true regional differences in variance. The effect of age on within-speaker variance, however, is very stable and highly significant. Older speakers produce significantly more within-speaker variation in phrase length. Likewise, the effect of sequential position is also highly significant. Speakers vary the length of their phrases more towards the end of their 15-minute interview, than in the beginning. Other variance components were not significantly different from zero; these insignificant components have been discarded from the multi-level model summarized in Table 2.

4. Discussion and Conclusion

The results for the full model (2) in Table 1 show no main effects for any of the three between-speaker factors in this study. Neither dialect region, nor sex, nor age, influences a speaker's average tempo, if phrase length is taken into account. In other words, speech tempo in a phrase is a function mainly of the length of that phrase (in orthographic syllables). This finding supports previous findings for American English read speech [3]. In the latter study, ASD was predicted in single-level fashion by 7 phrase-internal predictors (e.g. proportion of stressed syllables, number of phones, etc.). This yielded an $R^2 = .579$. Since all 6 speakers read the same text, ASD's could be correlated between phrases spoken by pairs of speakers, yielding correlations between .66 and .88. Although this may suggest that ASD is largely controlled by the contents of a phrase ([3], p.111), there is also considerable between-speaker variation in speech tempo.

Previous reports on speakers' average tempi have suggested that dialect region, sex and age [5, 6] are significant descriptors of this between-speaker variation. However, the present study shows that speakers' mean ASD is similar across these between-speaker factors. The apparent discrepancy between similar analyses of the same material needs to remain unexplained, until more data become available. At this time, data from more speakers and from more predictors are still being processed.

One possible explanation for the above discrepancy is that the previously reported effect of dialect region may be an artefact of regional variation in phrase length. If speakers from the West produce relatively shorter phrases on average, and if those from the North and South produce longer phrases on average, then this difference in phrase length would explain the absence of a region effect on speech tempo — if phrase length does indeed affect tempo, as in model (2). This tentative explanation appears to be false, however, because phrase length appears to be equal across dialect regions (see Table 2).

A similar line of reasoning, however, does hold for the absence of an age effect in model (2). Age has a (marginally) significant effect on a speaker's average phrase length. On average, older speakers produce shorter phrases than older speakers do. Hence, the independent effect of age in model (1) disappears if phrase length is included as an independent predictor, as in model (2).

The results on phrase length also show that older speakers produce significantly more *variation* in the length of their phrases than younger speakers do. If we return to the ASD model (2), we see a similar age effect: older speakers produce significantly more variation in speech tempo than younger speakers do, even if the ASD data are corrected for the logarithmic effect of phrase length. These age grading effects could be due to two effects. These older speakers may succesfully vary tempo and phrase length for communicative purposes, after decades of experience in expressing themselves as teachers. On the other hand, older teachers may increasingly suffer from cognitive constraints (e.g. in retrieving words from their mental lexicon), which forces them to produce shorter phrases occasionally, yielding more variation in phrase length. Again, these possible explanations need to be further investigated.

The resulting multi-level model of speech tempo ascribes most within-speaker variance to the speaker's dialect region. However, this apparent effect of dialect region could be artefactual, as explained above. In any case, most of the within-speaker variance remains unexplained in the final model (2). This suggests that other unknown factors, *not* related to phrase length or any other predictor, control a speaker's tempo variations. The present data do not allow us to determine whether these unknown factors are phonetic (segmental content) or communicative (emphasis, attitude) in nature.

Although these analyses have raised many new questions, they also provide us with positive answer to some questions. The coefficient of within-speaker variation in tempo is about 28%, which by far exceeds the JND for speech tempo of about 5%. This confirms similar results on speech tempo in interviews [16]. Apart from indicating the speaker's communicative intent, tempo also works as a scaling factor for other phonetic distinctions, e.g. voicing of consonants [17], or quantity of vowels [18]. Such changes in spontaneous speech tempo do exceed the just-noticeable difference for speech tempo in Dutch, which underlines their relevance for speech communication.

5. References

[1] R.H. Stetson, *Motor Phonetics* (retrospective ed.), J.A.S. Kelso and K.G. Munhall, Eds. Boston: Little, Brown and Company, 1988.

[2] F. Goldman-Eisler, *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic, 1968.

[3] T.H. Crystal and A.S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *J. Acoust. Soc. Am.*, vol. 88, pp. 101–112, 1990.

[4] S.G. Nooteboom and W. Eefting, "Evidence for the adaptive nature of speech on the phrase level and below," *Phonetica*, vol. 51, pp. 92–98, 1994.

[5] J. Verhoeven, G. De Pauw and H. Kloots, "Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands", *Language and Speech*, vol. 47, pp. 297–308 (2004).

[6] H. Quené, "On the just noticeable difference for tempo in speech". Manuscript under review.

[7] S.G. Nooteboom, "Production and perception of vowel duration", PhD thesis, Rijksuniversiteit Utrecht, 1972.

[8] B. Lindblom and K. Rapp, "Some temporal regularities of spoken Swedish," *Papers in Linguistics, University of Stockholm*, vol. 21, pp. 1–59, 1973.

[9] J.J. de Rooij, "Speech Punctuation: An acoustic and perceptual study of some aspects of speech prosody in Dutch," PhD thesis, Rijksuniversiteit Utrecht, 1979.

[10] L.H. Nakatani, K.D. O'Connor, and C.H. Aston, "Prosodic aspects of American English speech rhythm," *Phonetica*, vol. 38, pp. 84–105, 1981.

[11] N. Oostdijk, "The Spoken Dutch Corpus: Overview and first evaluation," in *Proc. Second Int. Conf. Language Resources and Evaluation*, vol. 2, M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, Eds., 2000, pp. 887–894.

[12] R. Van Hout, G. De Schutter, E. De Crom, W. Huinck, H. Kloots and H. Van de Velde, "De uitspraak van het Standaard-Nederlands: Variatie en varianten in Vlaanderen en Nederland," in *Artikelen van de Derde Sociolinguïstische Conferentie*, E. Huls and B. Weltens, Eds. Delft: Eburon, 1999, pp. 183–196.

[13] T.A.B. Snijders and R.J. Bosker, *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. London: Sage, 1999.

[14] D.A. Luke, *Multilevel Modeling*. Thousand Oaks, CA: Sage, 2004.

[15] H. Quené and H. van den Bergh, "On Multi-Level Modeling of data from repeated measures designs: A tutorial", *Speech Communication*, vol. 43, pp. 103–121 (2004).

[16] J.L. Miller, F. Grosjean, and C. Lomanto, "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," *Phonetica*, vol. 41, pp. 215–225, 1984.

[17] L.E. Volaitis and J.L. Miller, "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," *J. Acoust. Soc. Am.*, vol. 92, pp. 723–735, 1992.

[18] H. Traunmüller and D. Krull, "The effect of local speaking rate on the perception of quantity in Estonian," *Phonetica*, vol. 60, pp. 187–207, 2003.