Modeling Durational Incompressibility

Andreas Windmann[1], *Juraj Šimko*[1], *Britta Wrede*[2], *Petra Wagner*[1]

[1]Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[2]Faculty of Technology, Bielefeld University, Germany
{awindmann2,juraj.simko,petra.wagner}@uni-bielefeld.de, bwrede@techfak.uni-bielefeld.de

Abstract

We show how incompressibility, a well-described property of some prosodic timing effects, can be accounted for in an optimization-based model of speech timing. Preliminary results of a corpus study are presented, replicating and generalizing previous findings on incompressibility as a function of increasing speaking rate. We then introduce the architecture of our model and present results of simulation experiments that reproduce the results of the corpus analysis. Results suggest that incompressibility can be interpreted as a consequence of trade-offs between competing requirements of production efficiency and communicative efficacy.

Index Terms: speech timing, computational modeling

1. Introduction

It is a long-established finding that acoustic durations of speech segments exhibit *incompressibility*, i.e., the property of approaching a threshold under the influence of some shortening processes [1][2][3]. In the descriptive model by Klatt [2][3], this effect was captured by positing equations of the form

$$D_j = k(D_i - D_{min}) + D_{min} \quad (1)$$

where D_{min} is a constant that implements the compressibility threshold. Klatt hypothesized that this reflects a minimum duration required for executing articulatory movements. Incompressibility seems to be a property of most, though not necessarily all [4] shortening processes, whereas prosodic lengthening effects show less evidence of it [5][6].

A generalized prediction that follows from (1) is that under identical conditions, inherently longer segments should shorten more strongly than inherently shorter segments, in absolute as well as proportional terms. To see this, consider Figure 1: if two segments with the same minimum duration $D_{min} > 0$ are shortened by a factor k, the ratio between inherent duration D_i and output duration D_j will be greater for the inherently longer segment, as exemplified by D_{i1} and D_{i2} in the figure. Without incompressibility, the D_i/D_j ratio will be the same, regardless of differences in D_i (cf. D_{i3} and D_{i4}). Indeed, this is what prompted Klatt [2] to introduce incompressibility, after finding that an earlier model which assumed constant percentage changes over-predicted the combined effects of postvocalic voicing and polysyllabic shortening on vowel durations.

In this paper, we shall demonstrate how incompressibility can be accounted for in our optimization-based model of speech timing [7]. This model is aimed at providing explanations for speech timing phenomena, implementing the assumption that speech patterns emerge from the resolution of conflicting demands related to minimization of effort and maximization of communicative success, as posited by Hypo & Hyper-articulation (H&H) theory [8]. In the following section, we present results from a corpus study, supporting and generalizing previous findings on incompressibility at increasing speaking rates. We then describe the model architecture and present simulation results that replicate observations from the corpus analysis. The paper is concluded by a general discussion of the account of incompressibility proposed by our modeling paradigm.

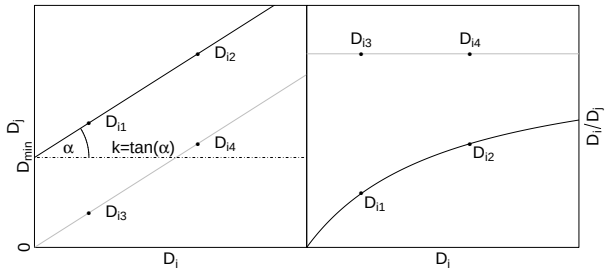

Figure 1: *Left panel: schematic illustration of timing processes with (black line) and without (gray line) compressibility threshold. Right panel: ratios between inherent duration D_i and output duration D_j for both processes. See text for more details.*

2. Corpus Study

Previous studies have attested incompressibility by crossing the effects of speech rate and categorical phonological distinctions such as vowel tensity and postvocalic voicing [4][9]. In order to test the generalized incompressibility hypothesis laid out above, we wanted to examine the effect in the continuous phonetic domain, and in higher-level prosodic constituents. This was done by computing linear regressions on syllable durations in fast productions as a function of corresponding syllable durations produced at a "normal" conversational rate in text readings recorded at different tempos. In this analysis, the regression intercept would correspond to D_{min} in Figure 1, interpreting syllable durations at normal rate as "inherent" durations. A significantly positive intercept would thus be evidence for incompressibility, indicating that the proportional magnitude of shortening varies as a function of inherent duration.

Analyses were conducted on data from the BonnTempo Corpus [10]. The corpus comprises readings of a short paragraph in German, English, French, and Italian, produced at five different rates (very slow – slow – normal – fast – fastest pos-

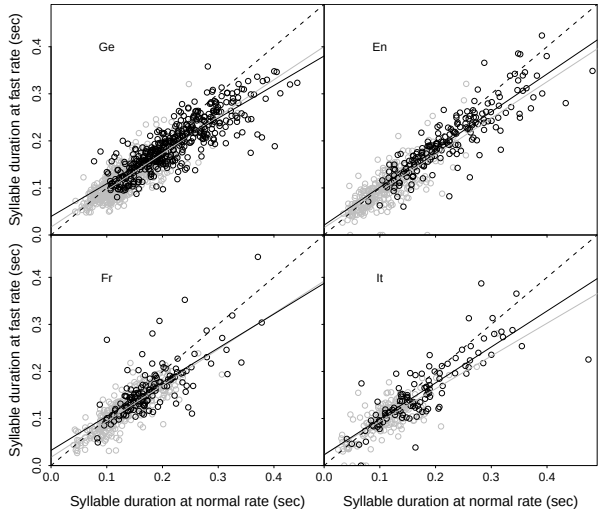

Figure 2: *Regression models of fast as a function of normal duration, fitted to stressed (black) and unstressed (gray) syllables separately.*

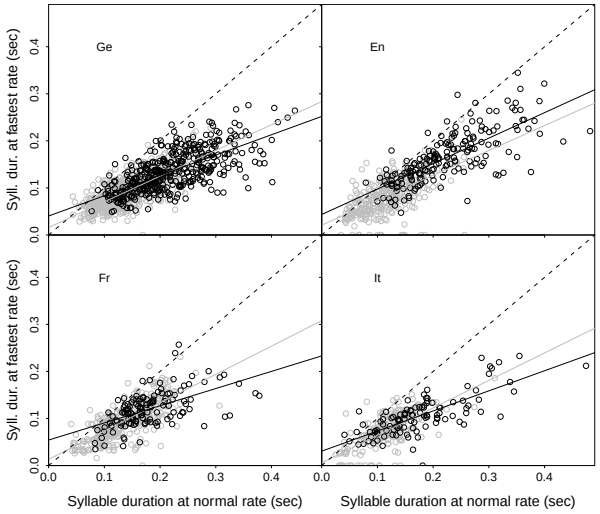

Figure 3: *Regression models of fastest possible as a function of normal duration, fitted to stressed (black) and unstressed (gray) syllables separately.*

sible). For this paper, we examined the fast and fastest possible data, using the normal condition as a reference point for comparison. Analyses were carried out on the existing segmentation of the corpus data. Phrase-final syllables were excluded from the analysis, in order to prevent a possible confounding influence of final lengthening. Elisions in the fast conditions were treated as cases of 100% shortening. This strategy was adopted as a conservative approach towards handling elisions, as it should bias intercept estimates towards zero and thus favor the null hypothesis. Table 1 summarizes the corpus data.

Table 1: *Summary of corpus data. Unbracketed: numbers of syllables in normal rate productions. In brackets: numbers of elided syllables in fast/fastest-possible productions.*

lang.	spkrs.	stressed	unstressed	total
Ge	15	408 (0/1)	609 (0/0)	1017 (0/1)
En	7	182 (0/0)	304 (4/14)	486 (4/14)
Fr	6	134 (0/0)	311 (0/1)	445 (0/1)
It	3	107 (1/1)	180 (4/18)	287 (5/19)

Separate models were fitted to stressed and unstressed syllables in the four languages. Figures 2 and 3 show plots of syllable durations at the fast rates as a function of duration at normal rate with regression lines fitted to the data. Table 2 summarizes the regression models, showing estimates (in ms) and significance levels for intercepts (Int) as well as the amount of variance (R^2) explained by the models (*: $p < .05$; **: $p < .01$; ***: $p < .001$; all slopes are significant at $p < 0.001$).

As can be seen in Table 2, all models have significantly positive intercepts, with the exception of the model fitted to the fastest possible unstressed Italian data. Inspection of these data suggests that this is indeed due to a substantial proportion of complete elisions in this condition. Our results thus generally support the hypothesis that the magnitude of rate-induced shortening of syllables varies as a function of their "inherent" duration at normal rate, replicating and generalizing findings from previous investigations [4][9][11].

Table 2: *Summary of regression models for stressed and unstressed syllables in German, English, French, and Italian.*

	fast		fastest possible	
language	Int	R^2	Int	R^2
Ge str.	39^{***}	0.71	40^{***}	0.46
Ge unstr.	17^{***}	0.72	16^{***}	0.55
En str.	21^{**}	0.77	43^{***}	0.54
En unstr.	16^{***}	0.60	21^{***}	0.35
Fr str.	32^{**}	0.52	54^{***}	0.31
Fr unstr.	18^{***}	0.71	13^{**}	0.52
It str.	23^{*}	0.69	31^{***}	0.57
It unstr.	24^{***}	0.58	6 n.s.	0.45

Without further statistical inquiry, two preliminary observations can be made on Table 2: first, the models fitted to stressed syllables have higher intercepts than their unstressed counterparts, with the exception of Italian. This supports the assumption that unstressed segments are more compressible than stressed ones [12]. Second, intercept estimates also seem to increase from the fast to the fastest-possible condition, at least for stressed syllables. This phenomenon is at odds with the traditional descriptive account captured by equation (1), in particular with the assumption of a fixed minimal duration D_{min} predicting tempo-independent value of the intercept (see Figure 1). If proved significant (a dedicated study of a large corpus will be required for this purpose), this discrepancy would call for a revision of Klatt's incompressibility theory, or, rather, for its possible application for syllable durations.

3. Model Architecture

Our model stands in the tradition of [13] and other H&H–inspired approaches [14][15][16][17][18], which employ optimization algorithms to simulate trade-offs between production and perception demands. It operates on specifications of sequences of stressed and unstressed syllables, represent-

ing speech utterances. Given an input sequence, an optimization algorithm computes the vector of syllable durations S that minimizes the composite cost function C. This function is a weighted sum of component functions that represent production and perception-related influences on constituent durations.

For the present purpose, we have included three such components, D, T and P. D and T implement constraints associated with efficiency of information transmission. The durational cost component T captures the overall duration of the utterance, i.e., the time used for conveying the message encoded in the sentence of a part thereof. The component D represents an approximation of the articulatory effort parsimoniously assumed to be proportional to the average segment duration. Cost component P capturing the parsing effort imposed on the listener, on the other hand, favors longer durations facilitating perception. This model architecture implements the assumption that speech patterns emerge from the resolution between conflicting requirements of production efficiency and communicative efficacy. Formally, this can be written as

$$C = \alpha_D \sum_S \delta_S D_S + \alpha_P \sum_S \psi_S P_S + \alpha_T T \tag{2}$$

where

$$P = \frac{n}{S} \tag{3}$$

$$D = \frac{S}{n} \tag{4}$$

$$T = \sum S \tag{5}$$

$$[\alpha_D, \alpha_P, \alpha_T, \delta_S, \psi_S] \in \mathbb{R}^+ \tag{6}$$

While both D and T are essentially linear functions of duration, the difference lies in their scope: T acts "globally", at utterance level, while D operates at the syllabic level, being modified locally by the number of segments per syllable, n, and by the parameter δ_S, which we describe below. Motivation for having both in the model comes from evidence that different mechanisms may be responsible for changes of local durations and overall speaking rate, cf. [19] and references therein.

Component P, in contrast, is non-linear, being modeled by a convex decaying function such as $1/S$. This technique has an intuitive appeal if one interprets P as the *inverse of the probability of recognition* of a syllable, which should grow as a function of its duration up to the point where recognition is 100%, staying constant afterwards. Direct evidence for this modeling decision comes from gating studies, where subjects have to identify syllables from acoustic fragments of varying duration [20][21]. P is also modified by syllabic structure, which can be interpreted as an affordance to produce a syllable with sufficient time for each segment to be perceived.

The objects in (6) are scalar weighting factors, which allow for locally or globally imposing premiums on individual component cost functions in order to model different prosodic conditions. Increasing the value of α_T, for example, increases the premium placed on requirement of more efficient, faster transmission; this parameter is used in this work to elicit speaking rate variations. Similarly, weights α_D and α_P conceptualize the premiums placed on requirements of production and perception efficiency. Of particular interest in the context of current modeling work are δ_S and ψ_S, which locally modify D and P. This is how stress is accounted for in the model: assuming that stressed syllables are particularly critical for decoding linguistic structure, they are modeled by locally increasing ψ_S, based on the intuition that it is precisely the greater degree of *perceptual* salience that differentiates stressed from unstressed syllables. At the same time, D is "released" for stressed syllables, implementing the assumption of "extra energy" expenditure [19].

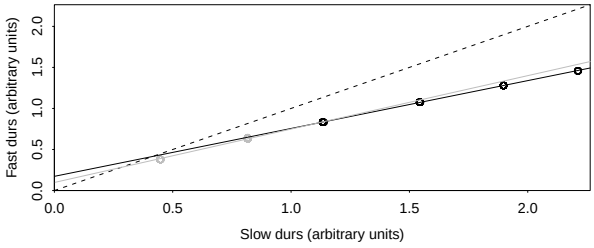

Figure 4: *Simulated durations of stressed (black) and unstressed (gray) syllables at a faster ($\alpha_T = 1.5$) rate as a function of corresponding durations at a slower ($\alpha_T = 0.5$) rate.*

4. Simulations

The model was set up using the implementation of the Nelder-Mead algorithm in the R function *optim* [22]. Input data were generated by compiling random "corpora" of roughly 900–1000 syllables, with utterance lengths and distributions of stresses roughly modeled on our corpus data. Since the BonnTempo Corpus is not annotated for syllabic structure, we derived n from the distributional statistics on syllable types in English given in [23], preliminarily restricting the investigation to English. While the simulations thus do not directly reflect the input speech corpora, they should provide a reasonable approximation of English language structure. In the simulations reported, α_D and α_P were set to 1, ψ_S to 0.75 for stressed and 0.5 for unstressed syllables, and δ_S was set to $1/\psi_S$, so as to reduce the number of free parameters. α_T was varied in order to simulate increasing speaking rate. With such a simple model definition, it would actually be possible to solve the optimization problem analytically, but we decided to run simulations, with reference to future investigations with a more elaborate architecture.

Figure 4 shows simulated syllable durations at a faster ($\alpha_T = 1.5$) as a function of durations at a slower ($\alpha_T = 0.5$) rate. The substantial positive intercepts of the stressed and unstressed fits show that the increase in speaking rate in the model simulations is characterized by incompressibility. While the structure of the input data is not identical to that of the real corpora, the general pattern of results of the corpus study, which seems to be independent of the language under consideration, is thus reproduced by the model. It also replicates the greater compressibility of unstressed compared to stressed syllables.

We ran simulations across a range of values for α_T in order to investigate whether further rate increases would result in higher regression intercepts, as observed in the corpus study. Figure 5 shows that the model indeed predicts increasing intercepts as a function of further rate increases, but does so only up to a certain point, where the trend is reversed. Since α_T values cannot be straightforwardly related to actual speaking rates, it is not clear whether this prediction is of relevance for speech.

5. Discussion

Preliminary results indicate that our optimization model can account for incompressibility under increasing speaking rate. The model also replicates observed differences between stressed and

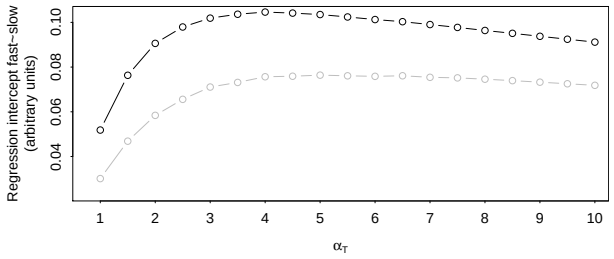

Figure 5: *Regression intercepts for simulated stressed (black) and unstressed (gray) syllables at faster rates as a function of corresponding durations at slow ($\alpha_T = 0.5$) rate.*

unstressed syllables. Crucially, the effect emerges from the interaction of independently motivated components, representing in a very general manner production and perception-related requirements. No ad-hoc mechanisms, such as explicit minimum duration constraints, have to be posited. One problem is that the model does not necessarily account for deceleration, which may not exhibit incompressibility [5]. However, slowing down one's speech is arguably a very different process from speeding up: Investigations of decelerating speech report a variety of possible strategies, including effects such as spontaneously emerging breaks, which are not straightforwardly captured by either incompressibility or constant-ratio models [24][25][26]. We have not yet investigated the possibility of elisions in the model, which might be an interesting perspective, given that it seems to be influenced by rhythmic factors [27].

In order to understand how the model incorporates incompressibility, it is helpful to visualize the cost functions for different conditions, as in Figure 6. In the left panel, C is plotted for a "stressed" value of $\delta_S = 0.5$ in black and an "unstressed" $\delta_S = 1$ in gray at three different rates, $\alpha_T = 1$ (solid), $\alpha_T = 2$ (dashed) and $\alpha_T = 100$ (dotted). The filled circles mark the minima of C, corresponding to the optimal durations for the respective settings. Increasing values of α_T shorten durations, pushing the minimum of C further to the left. The plots for $\alpha_T = 100$ show that even for unrealistically high durational weights, the optimal duration is still considerably greater than 0. Although the optimal duration will asymptotically converge to 0 as α_T increases, the perceptual consequences of such shortening are too high to get compensated by gains in transmission efficiency in a linear fashion. The underlying trade-offs captured by the model architecture are at the core of incompressibility phenomena as accounted for in the present work.

The right panel of Figure 6 shows time-normalized plots of C for $\delta_S = 0.5$ (black) and $\delta_S = 1$ (gray) at $\alpha_T = 1$ (solid) and $\alpha_T = 2$ (dashed), setting the minima of C at $\alpha_T = 1$ to 1. This plot shows that the inherently longer syllable undergoes proportionally stronger shortening, as indicated by the larger leftward displacement of the black compared to the gray minimum in the faster condition. Given the way we model stress, this would lead to stronger shortening in stressed than in unstressed syllables, which is at odds with some empirical findings, e.g. for Dutch [28]. It is conceivable that more fine-grained modeling of sub-syllabic constituents would be required to accommodate both this result and durational incompressibility.

Inspection of Figure 6 suggests that it is the interaction between various aspects of production and perception efficiency requirements introducing the surface phenomena of incompressibility. Rather than introducing a threshold for syllable duration, compressing speech constituents beyond a certain point becomes prohibitively expensive in terms of perception. Longer elements, on the other hand, can undergo substantial compression without marked perceptual loss.

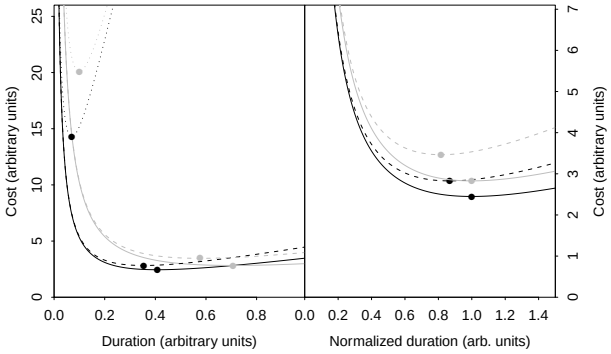

Figure 6: *Left panel: Plot of C for $\delta_S = 0.5$ (gray) and $\delta_S = 1$ (black) at $\alpha_T = 1$ (solid), $\alpha_T = 2$ (dashed) and $\alpha_T = 100$ (dotted). Right panel: Time-normalized plots of C for $\delta_S = 0.5$ (gray) and $\delta_S = 1$ (black) at $\alpha_T = 1$ (solid) and $\alpha_T = 2$ (dashed).*

Interestingly, our cost functions can be re-interpreted in terms of dynamical systems theory: the minima can be thought of an *attractors*, stable states the system tends to converge towards. Rather than being the consequence of executing explicit timing rules, the result represents the optimal solution, given the constraints that act upon the system. On this view, the interplay between P, D and T can be envisioned as giving rise to oscillatory behavior, settling on a "natural frequency" fully determined by syllabic structure when unperturbed. T exerts linear compression on syllable durations, to which inherently longer syllables offer less resistance, being more elastic than inherently shorter ones. Our proposal thus bears resemblance to the technique introduced in [29], where an oscillatory approach was applied to inter-stress interval durations expressed by linear regression [30]. We have recently replicated [29]'s modeling result within the optimization paradigm [7], using additional components that impose tendencies towards periodicity of syllable and foot onsets. Our current results complement this finding, demonstrating the potential of optimization modeling as a unified account of various speech timing phenomena.

6. Conclusion

We have shown that our optimization-based model of speech timing replicates incompressibility as a property of syllable durations under increasing speaking rate. The model provides a principled explanation for this effect, grounding it in a cognitively plausible model architecture that exploits production-perception trade-offs to account for speech phenomena. Our results thus add to previous findings [13][7], suggesting that timing effects observed at various levels of speech description (gestural, segmental, suprasegmental) can be given a unified explanation within an optimization-based modeling paradigm

7. Acknowledgements

This work was partially supported by a Bielefeld University LiLi-Kolleg grant to the first and a von Humboldt Fellowship grant to the second author. We thank Michael O'Dell and three anonymous reviewers for helpful suggestions.

8. References

[1] I. Lehiste, "The timing of utterances and linguistic boundaries," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2018–2024, 1972.

[2] D. H. Klatt, "Interaction between two factors that influence vowel duration," *The Journal of the Acoustical Society of America*, vol. 54, no. 4, pp. 1102–1104, 1973.

[3] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, p. 1208, 1976.

[4] R. Port, "Linguistic timing factors in combination," *The Journal of the Acoustical Society of America*, vol. 69, no. 1, pp. 262–274, 1981.

[5] R. Port, "The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words," Ph.D. dissertation, University of Connecticut, 1976.

[6] F. Cummins, "Some lengthening factors in English speech combine additively at most rates," *The Journal of the Acoustical Society of America*, vol. 105, no. 1, pp. 476–480, 1999.

[7] A. Windmann, J. Šimko, B. Wrede, and P. Wagner, "Optimization-based model of speech timing and rhythm," in *The 13th conference on Laboratory Phonology*, Stuttgart, Germany, 2012.

[8] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech production and speech modeling*, W. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer, 1990, pp. 403–439.

[9] H. Gopal and A. K. Syrdal, "Interaction of speaking rate and postvocalic consonantal voicing on vowel duration in American English," *The Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S16–S16, 1987.

[10] V. Dellwo, I. Steiner, B. Aschenberner, J. Dankovicova, and P. Wagner, "BonnTempo-Corpus & BonnTempo-Tools: A database for the study of speech rhythm and rate,," in *Proceedings of Interspeech 2004*, Jeju Island, Korea, 2004, pp. 777–780.

[11] J. P. van Santen and J. P. Olive, "The analysis of contextual effects on segmental duration," *Computer Speech & Language*, vol. 4, no. 4, pp. 359–390, 1990.

[12] D. H. Klatt, "Synthesis by rule of segmental durations in English sentences," *Frontiers of Speech Communication Research*, vol. 1, pp. 287–300, 1979.

[13] J. Šimko and F. Cummins, "Sequencing and optimization within an embodied task dynamic model," *Cognitive Science*, vol. 35, no. 3, pp. 527–562, 2011.

[14] P. Boersma, *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Holland Academic Graphics/IFOTT, 1998.

[15] R. Kirchner, *An effort based approach to consonant lenition*. Routledge, 2001.

[16] E. Flemming, "Phonetic optimization: Compromise in speech production," *University of Maryland Working Papers in Linguistics*, vol. 5, pp. 72–91, 1997.

[17] E. Flemming, "Scalar and categorical phenomena in a unified model of phonetics and phonology," *Phonology*, vol. 18, no. 1, pp. 7–44, 2001.

[18] J. Katz, "Compression effects, perceptual asymmetries, and the grammar of timing," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[19] K. S. Harris, "Vowel duration change and its underlying physiological mechanisms," *Language and Speech*, vol. 21, no. 4, pp. 354–361, 1978.

[20] W. Grimm, "Perception of segments of English-spoken consonant-vowel syllables," *The Journal of the Acoustical Society of America*, vol. 40, no. 6, pp. 1454–1461, 1966.

[21] M. Tekieli and W. Cullinan, "The perception of temporally segmented vowels and consonant-vowel syllables," *Journal of Speech, Language and Hearing Research*, vol. 22, no. 1, p. 103, 1979.

[22] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

[23] R. Dauer, "Stress-timing and syllable-timing reanalyzed." *Journal of Phonetics*, vol. 11, no. 1, pp. 51–62, 1983.

[24] J. Trouvain and M. Grice, "The effect of tempo on prosodic structure," in *Proceedings of ICPhS 1999*, San Francisco, 1999, pp. 1067–1070.

[25] J. Trouvain, "Tempo variation in speech production: Implications for speech synthesis," Ph.D. dissertation, Saarland University, 2003.

[26] J. Šimko and S. Beňuš, "Emergence of prosodic boundaries: continuous effects of temporal affordance on inter-gestural timing," *Journal of Phonetics*, submitted.

[27] S. Tilsen, "Relations between speech rhythm and segmental deletion," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, vol. 44, no. 1. Chic Ling Society, 2008, pp. 211–223.

[28] E. Janse, S. Nooteboom, and H. Quené, "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," *Speech Communication*, vol. 41, no. 2, pp. 287–301, 2003.

[29] M. O'Dell and T. Nieminen, "Coupled oscillator model of speech rhythm," in *Proceedings of ICPhS 1999*, San Francisco, 1999, pp. 1075–1078.

[30] A. Eriksson, "Aspects of Swedish speech rhythm," Ph.D. dissertation, University of Gothenburg, 1991.