Probing Theories of Speech Timing using Optimization Modeling

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

[1]Faculty for Linguistics and Literary Studies, Bielefeld University, Germany
[2]Institute of Behavioural Sciences, University of Helsinki, Finland
[1]firstname.lastname@uni-bielefeld.de, [2]juraj.simko@helsinki.fi

Abstract

We implement two theories about the temporal organization of speech in an optimization-based model of speech timing and conduct simulation experiments in order to test whether both theories can account for the phenomenon of foot-level shortening (FLS) observed in English speech corpora. Results suggest that a model that induces compensatory timing relations between syllables and feet predicts empirical results very accurately. However, we also observe that the FLS effect can equally well be explained under the assumption that suprasegmental timing is confined to localized lengthening effects at the heads and edges of prosodic domains. Implications for theories of speech timing are discussed.

Index Terms: Speech timing, computational modeling

1. Introduction

In this paper, we shall test predictions made by two theories of suprasegmental speech timing. To this end, we will employ a computational optimization model [1], by implementing both theories in the model and evaluating modeling results against attested speech timing patterns of English. Results allow for comparing the empirical adequacy of different predictions and demonstrate the potential of our model as a test bed for different theories of suprasegmental speech timing.

The phenomenon under study is polysyllabic shortening, i.e. the property of syllables to shorten as a function of the number of syllables in some larger prosodic unit. We will concentrate on foot-level shortening (FLS), an alleged shortening effect triggered by the interval from a lexically stressed syllable onset to the next, possibly spanning word boundaries [2]. While not leading to true periodicity of stressed syllable onsets, FLS in English does seem to be well-supported by both experimental studies [3, 4, 5] and corpus analyses [6, 7, 8, 9]. Figure 1 shows two examples of FLS patterns found in English speech corpora, as evident from vowel rather than syllable durations. There is a marked shortening effect on stressed vowel durations which is, however, not linear but seems to be attenuated as syllable count in the foot increases. Similar patterns have been observed in experimental studies and could be interpreted as an effect of durations moving towards a compressibility threshold [12]. Figure 1 does not indicate FLS in unstressed syllables, which has been linked to incompressibility as well [13]. However, [10] do report consistent shortening effects of various prosodic constituents also on unstressed vowel durations.

Findings on FLS seem to support theories in which prosodic timing effects are distributed over larger prosodic domains, leading to inverse relationships between the number of syllables assembled in these domains and the durations of those syllables [14, 2]. This has been explicitly formalized in a class of computational models which assume that the temporal organization of speech is governed by oscillatory mechanisms at different levels of the prosodic hierarchy [15, 16, 17, 18, 9]. On this account, surface speech timing patterns emerge as a result of an interaction, or *coupling*, between the different oscillators, as a result of which they entrain to a stable frequency pattern. FLS would be explained to arise from hierarchical coupling between a dominant oscillator at the inter-stress interval or foot level and a more compliant syllabic oscillator, thus generating a tendency for stressed syllables to reoccur at regular intervals.

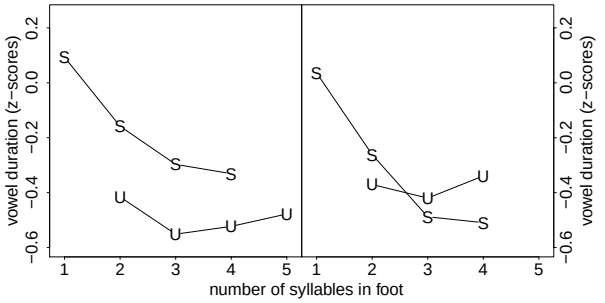

Figure 1: *Foot-level shortening effect on mean stressed (S) and unstressed (U) vowel/syllable durations in English. Graphs are reproduced from numerical results for two speakers from [8].*

In contrast to this, [19] and [20] propose a theory in which suprasegmental timing mechanisms are confined to localized lengthening at the heads and edges of prosodic domains. In this theory, there is little or no role for "domain-span effects", a term used by these authors to denote precisely the kind of inverse timing relationships that oscillatory approaches would predict. [19] and [20] base their claims on experimental findings suggesting that purported shortening effects at the word level in English can be largely accounted for by combinations of localized lengthening effects, such as accentual lengthening and word-final lengthening. However, [20] concede that their model may not necessarily account for FLS effects, which makes it a promising prospect to test for foot-level effects explicitly.

In this paper, we shall investigate whether both theories can account for FLS effects observed in English speech corpora. This will be done by implementing the mechanisms proposed by both theories, informally referred to as the "distributed" vs. the "localized" timing account, in our optimization-based model of speech timing. Predictions made by both theories will then be tested on input data derived from an authentic speech corpus and compared to published results. The paper is structured as follows: in section 2, we describe the general architecture of our model and the additions implemented in order to

accommodate the distributed and the localized timing account. Results of simulation experiments are presented in section 3 and discussed in section 4. Section 5 concludes the paper and addresses perspectives for further work.

2. Model Architecture

Our model is inspired by Hypo & Hyper-articulation (H&H) theory [21], implementing the assumption that speech patterns emerge from the resolution of conflicting demands related to minimization of effort and maximization of communicative success. It derives from an embodied optimization model of articulatory timing [22, 23]. The current model operates on specifications of sequences of lexically stressed and unstressed syllables, representing speech utterances. Given an input sequence, an optimization algorithm computes the vector S of syllable durations that minimizes the composite cost function C. C is a weighted sum of component functions that represent production and perception-related influences on constituent durations.

The basic architecture of the model includes three such components, D_S, T and P_S. D_S 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. This component provides a control mechanism for overall speech tempo. D_S is proportional to individual syllable durations, based on the assumption that the syllable is a basic unit of information which speakers strive to transmit in an efficient manner. Motivation for having both T and D_S in the model comes from evidence that different mechanisms may be responsible for changes of local durations and overall speaking rate, cf. [24] and references therein.

Conversely, component P_S represents an impetus to maximize perceptual clarity, by imposing costs on the reciprocal of syllable durations. P_S thus decreases with syllable duration in a convex decaying fashion, assuming that very short durations impede perception while facilitation induced by durational lengthening will eventually reach a ceiling. Independent evidence for this modeling decision comes from gating studies, where subjects have to identify phonemes from acoustic syllable fragments of varying duration [25, 26].

Weighting factors allow for globally imposing premiums on the individual components in order to simulate requirements regarding efficiency (α_D), perceptual clarity (α_P) or overall speaking rate (α_T). The vectors δ_S and ψ_S, assigning weights to individual syllables, can be used to boost their relative perceptual clarity and simultaneously lower the premium on efficient information transmission. This mechanism is used to account for prosodic prominence in the model, assuming that speakers prioritize clarity over efficiency in prominent syllables [27]. We usually set δ_S to the reciprocal of ψ_S, in order to reduce the number of free parameters. Formally, the basic model is thus defined as

$$C = \alpha_D \sum_S \delta_S D_S + \alpha_P \sum_S \psi_S P_S + \alpha_T T \qquad (1)$$

In order to accommodate the distributed timing account, we implemented a version of the basic model with two additional component costs, D_F and P_F, together with respective weighting factors α_{DF} and α_{PF}. D_F and P_F are basically copies of D_S and P_S operating at the stress foot rather than the syllabic level. Thus, the two new components in combination impose a tendency to produce stress feet with a certain optimal duration, while D_S and P_S tend to converge to optimal syllable durations. This leads to an obvious trade-off and will trigger compensatory relationships for feet with different syllable counts. By setting the respective weighting factors, precedence can be given to either the foot or the syllabic level, simulating purported tendencies towards "stress timing" and "syllable timing", respectively. This architecture is conceptually very similar to the coupled oscillator models mentioned above. Independent motivation for such a design might be derived from findings on convergence of syllable and foot durations with certain temporal windows of cognitive processing [28]. Figure 2 visualizes the architecture of the distributed timing model.

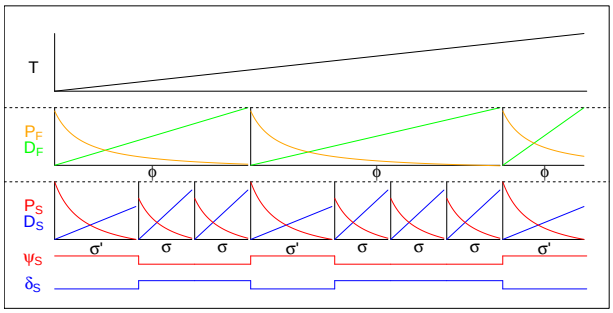

Figure 2: *Distributed timing model. Cost functions T (utterance level), D_F/P_F (stress foot level; ϕ) and D_S/P_S (syllabic level; σ; ' denotes lexical stress) as well as parameters δ_S and ψ_S are plotted as a function of respective constituent durations for a hypothetical utterance consisting of a trisyllabic, a tetrasyllabic and a monosyllabic foot. α_{DF} and α_{PF} are not shown.*

For the localized timing account, no additional cost functions had to be supplied. Its predictions were implemented by additionally enhancing ψ_S and thus boosting the perception-oriented component P for word-final syllables, leaving the model definition in (1) unchanged otherwise. This approach is in keeping with [19]'s reasoning that localized lengthening is utilized to increase the perceptual salience at important points in the speech signal, in this case word boundaries. There are thus no "domain-span" mechanisms in this version of the model (note that T does not induce compression as a function of utterance length if it is linear); only localized lengthening effects at the heads (stressed syllables) and edges (word-final syllables) of words are included. No attempt was made in either of the two versions of the model to incorporate utterance-final lengthening, since utterance-final syllables have usually been excluded in investigations of FLS. Effects of syllable structure and pitch accent were also ignored for the present purpose. Figure 3 visualizes the architecture of the localized timing model.

3. Simulations

Input data for the simulation experiments were derived from the Aix-MARSEC database, a corpus of English broadcast speech [29, 30, 31]. That is, we prepared input "utterances" for the model that were based on actual utterances from the corpus in terms of number of syllables and locations of lexical stress and word boundaries. The Aix-MARSEC database is ideally suited for this purpose because FLS and shortening effects in other domains have been documented in this corpus [6, 10].

Both versions of the model were implemented in R using the built-in optimization function *optim*. In order to keep computing time within reasonable limits, input data were restricted to 2000 utterances from the corpus, amounting to 7512 sylla-

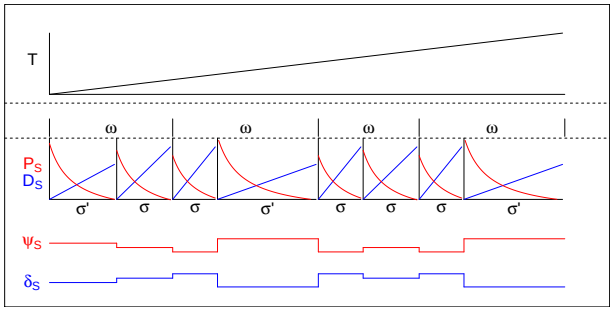

Figure 3: *Localized timing model. The utterance is the same as in Figure 2. The middle panel shows word boundaries (ω). Note the differentiation for word-final vs. non-final syllables in ψ_S and δ_S (in addition to stressed/unstressed differentiation).*

bles. Utterance-initial exametrical syllables, i.e., unstressed syllables not contained in a foot, were excluded. For the simulation with the distributed timing model, we set α_D and α_P to 0.5 and α_{DF} and α_{PF} to 1 in order to simulate the hypothesized dominant timing influence of the foot. ψ_S was set to 2 for lexically stressed and 1 for unstressed syllables, and δ_S was set to $1/\psi_S$, as explained above. All other parameters were set to 1. Simulation results from the distributed timing model are shown in Figure 4.

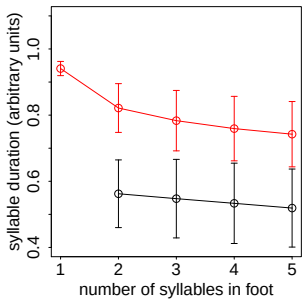

Figure 4: *Results from distributed timing model simulation for stressed (red) and unstressed (black) syllables.*

The distributed timing model reproduces the general pattern shown in Figure 1 remarkably well, particularly for the stressed syllables, which exhibit a marked asymptotic shortening tendency. The model does predict a consistent shortening effect in unstressed syllables as well, but it is quite weak compared to the effect in stressed syllables. Given that many other sources of durational variation are not taken into account in this version of the model, it may be assumed that the shortening effect on unstressed syllable durations is likely to vanish if simulations are carried out with a more full-blown model architecture. The fact that syllables shorten as a function of foot length is in itself of course rather trivial, given that the parameter settings impose temporal compression at the foot level. What is non-trivial about this result, however, is that the model (1) captures the asymptotic nature of the shortening effect in stressed syllables and (2) reproduces the finding of a weaker effect in unstressed syllables. We believe that both outcomes are indeed a consequence of incompressibility, which has been shown to emerge automatically from the architecture of our model [1].

The localized timing model was run on the same input data as the distributed timing model. In this simulation, ψ_S was increased by 0.5 for word-final syllables in order to simulate word-final lengthening. All other parameter settings were the same as in the distributed timing model simulation. Monosyllables were counted as final. Results are shown in Figure 5. Surprisingly, the localized timing model also captures the pattern of results shown in Figure 1 quite well, with asymptotic shortening in stressed and a weaker effect in unstressed syllables. The magnitude of the effect is somewhat smaller than in the distributed timing model simulation, but this of course depends on the exact numerical setting of the parameter values. The important result is that the localized timing model can account for the overall pattern of results. This is a striking finding, given that no explicit timing mechanism at the foot level is included in this version of the model.

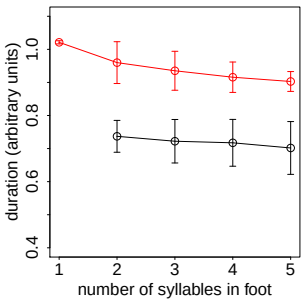

Figure 5: *Results from localized timing model simulation for stressed (red) and unstressed (black) syllables.*

On this account, the explanation for the FLS effect would be an entirely different one: rather than a genuine tendency towards temporal compression at the foot level, the phenomenon would be a mere statistical artifact, arising from an apparent tendency for word-final syllables to occur in shorter feet. An analysis of the whole MARSEC corpus was conducted in order to substantiate this correlation. We computed the probability of a syllable being in word-final position as a function of syllable count in the respective foot, by dividing the number of word-final syllables occurring in feet of a given length by the total syllable count for the respective foot length in the corpus. This was done separately for stressed and unstressed syllables. Results are shown in Figure 6. As can be seen, the probability of a syllable occurring in word-final position indeed decreases as a function of syllable count in the foot, in a fashion which is strikingly similar to the durational effect.

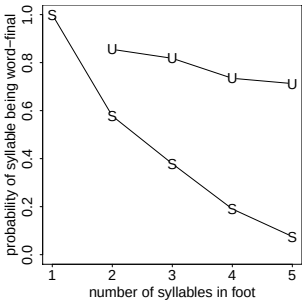

Figure 6: *Probability of syllables occurring in word-final position as a function of foot length in the MARSEC corpus.*

Upon closer inspection, this pattern of results is actually hardly surprising: if a stressed syllable is directly followed by

another stressed syllable and thus constitutes a monosyllabic foot, there is necessarily a word boundary intervening, since by definition, a word cannot contain more than one primary stressed syllable. The probability of being word-final therefore has to be one for monosyllabic stress feet. For bisyllabic feet, the probability of the stressed syllable being word-final is much lower – any bisyllabic word with initial stress followed by any word with initial stress will make for a non-word-final observation here – but there are still two frequent patterns that will generate final observations, 1) words with final stress followed by bisyllabic words with final stress, as in "[STAY a]WAKE", and 2) sequences of a word with final stress, a weak function word and a word with initial stress, as in "[JOHN the] BAPtist" (brackets mark target foot boundaries). For successively longer feet, the frequency of patterns that allow for word-final stressed syllables decreases – it is hard to imagine a pattern with a word-final stressed syllable followed by four unstressed syllables.

For unstressed syllables, the probability of occurring in word-final position decreases with foot length as well, but the effect is much weaker than in stressed syllables. A possible explanation is that longer feet are likely to involve polysyllabic words. In this case, some of the unstressed syllables in a long foot will be word-initial or medial, resulting in a weaker correlation between foot length and the probability to occur word-finally for unstressed syllables. This would explain the weaker or absent effect of foot length on unstressed syllables in the durational domain under a localized timing account.

4. Discussion

Our results show that both the distributed and, interestingly, also the localized account of speech timing can reproduce FLS patterns observed in English speech corpora. In the distributed timing account, the effect would be explained as a tendency towards periodicity of stressed syllable onsets, resulting from a trade-off between realizing optimal syllable and foot durations. Under the localized timing account, FLS would be classified as a mere by-product of language structure, i.e., the stronger tendency for syllables in shorter feet to occur word-finally, where they are subject to a localized lengthening effect.

If the distributed timing theory is correct, our implementation provides a promising explanatory platform for the FLS effect. Crucially, the model not only produces longer syllables in shorter feet, but also captures the precise nature of the shortening effect intriguingly well. We would therefore argue that our distributed timing model offers a more satisfactory account of the phenomenon than the oscillatory approach described in [17], who have to introduce ad-hoc assumptions in order to replicate the difference between stressed and unstressed syllables reported by [7]. [17] effectively "switch off" FLS in unstressed syllables, but it may well be that the effect was just masked by noise from other durational processes in [7], in keeping with our model's predictions. Indeed, [8] reports on the same data that unstressed durations do exhibit a significant shortening effect once the number of *phones*, rather than syllables per foot is used as the independent variable. Since [17] only compare bi- and trisyllabic feet, it is also not clear if their oscillatory model captures the non-linear nature of the shortening in stressed syllables. In our distributed timing model, both the weaker effect in unstressed syllables and the attenuation of the shortening effect in stressed syllables emerge automatically from the independently motivated property of durational incompressibility.

Results of the localized timing model simulations, however, show that it may be entirely unnecessary to invoke a timing mechanism at the foot level in order to reproduce the empirical results. This would be in keeping with [19]'s claim that "domain-span effects" have been falsely attributed to English due to ignorance of confounding influences such as final or accentual lengthening. Word-final lengthening in particular seems to be a well-attested effect in English [32, 33], although it is not entirely uncontroversial [34]. Of course, it cannot be decided based on our simulation results whether word-final lengthening is the trigger of a spurious FLS effect, or if, on the contrary, word-final syllables are longer than non-final ones *because* they tend to occur in shorter feet.

Thus, while our simulation results show that both accounts can generate the observed pattern, they do not allow for falsifying either theory. The predictions of the localized timing model converge with results by [19] and [20], as well as [33], who reports that in his large-scale corpus study, an apparent FLS effect on vowel durations disappears once a vowel's distance to the right word boundary is controlled. On the other hand, it seems that the localized timing model cannot fully account for *experimental* findings on FLS. For example, [5] report that a word-final stressed syllable is longer in a monosyllabic than in a bisyllabic foot, which is also acknowledged by [19] and [20]. There may be alternative analyses, such as a kind of stress clash effect here, however.

We have not tested for shortening effects in domains other than the stress foot. Results from [19, 20] and [33] are compatible with a domain-span effect at the *word rhyme* level, the unit that stretches from the onset of a stressed syllable to the next word boundary, alternatively referred to as *narrow rhythm unit* [35, 10]. However, [20] raise the possibility that this may in fact be a progressive word-final lengthening effect. Further empirical study is necessary in order to decide on these issues.

5. Conclusions

Using optimization modeling, we have shown that a model architecture that imposes distributed timing mechanisms can well account for patterns of foot-level shortening observed in English speech corpora. However, the effect may equally well be explained in a model of speech timing that only includes localized lengthening effects due to a tendency for shorter feet to contain mostly word-final syllables.

From a methodological point of view, results of our simulation experiments confirm that our optimization-based model provides a promising test bed for different theories of speech timing. The model itself is of course not theory-neutral, but it appears that the H&H assumptions it is based on are sufficiently general to accommodate other theories. Detailed studies on empirical data, including proper control for possible sources of durational variation, are required in order to determine which model architecture correctly captures the suprasegmental organization of English speech timing.

6. Acknowledgements

We would like to thank the present and former staff at the Laboratoire Parole et Langage at the University of Aix-Marseille, in particular Cyril Auran, Caroline Bouzon and Daniel Hirst, for making the MARSEC corpus publicly available. Further thanks go to three anonymous reviewers for helpful comments on an earlier draft of this paper. The first author gratefully acknowledges the Bielefeld graduate school of linguistics and literary studies for financial support.

7. References

[1] A. Windmann, J. Šimko, B. Wrede, and P. Wagner, "Modeling durational incompressibility," in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 1375–1379.

[2] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.

[3] T. P. Barnwell, "An algorithm for segment durations in a reading machine context." DTIC Document, Tech. Rep., 1971.

[4] M. Fourakis and C. Monahan, "Effects of metrical foot structure on syllable timing," *Language and Speech*, vol. 31, no. 3, pp. 283–306, 1988.

[5] B. Rakerd, W. Sennett, and C. Fowler, "Domain-final lengthening and foot-level shortening in spoken English," *Phonetica*, vol. 44, no. 3, p. 147, 1987.

[6] N. Campbell, "Foot-level shortening in the Spoken English Corpus," in *Proceedings of the 7th FASE Symposium*, Edinburgh, 1988, pp. 489–494.

[7] H. Kim and J. Cole, "The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase," in *Proceedings of Interspeech 2005*, Lisbon, 2005, pp. 2365–2368.

[8] H. Kim, "Speech rhythm in American English: A corpus study," Ph.D. dissertation, University of Illinois, 2006.

[9] J. Krivokapić, "Rhythm and convergence between speakers of American and Indian English," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013.

[10] C. Bouzon and D. Hirst, "Isochrony and prosodic structure in British English," in *Speech Prosody 2004, International Conference*, 2004.

[11] S. Shattuck-Hufnagel and A. Turk, "Durational evidence for word-based vs. prominence-based constituent structure in limerick speech," in *Proceedings of ICPhS 2011*, Hong Kong, 2011.

[12] D. 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.

[13] C. Hoequist, "Durational correlates of linguistic rhythm categories," *Phonetica*, vol. 40, no. 1, pp. 19–31, 1983.

[14] K. Pike, *The Intonation of American English.* Ann Arbor: University of Michigan Press, 1945.

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

[16] P. Barbosa, "From syntax to acoustic duration: A dynamical model of speech rhythm production," *Speech Communication*, vol. 49, no. 9, pp. 725–742, 2007.

[17] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008, pp. 175–184.

[18] S. Tilsen, "Multitimescale dynamical interactions between speech rhythm and gesture," *Cognitive Science*, vol. 33, no. 5, pp. 839–879, 2009.

[19] L. White, "English speech timing: a domain and locus approach," Ph.D. dissertation, University of Edinburgh, 2002.

[20] L. White and A. E. Turk, "English words on the procrustean bed: Polysyllabic shortening reconsidered," *Journal of Phonetics*, vol. 38, no. 3, pp. 459–471, 2010.

[21] 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.

[22] J. Šimko and F. Cummins, "Embodied task dynamics." *Psychological review*, vol. 117, no. 4, pp. 1229–1246, 2010.

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

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

[25] 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.

[26] 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.

[27] K. J. De Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *The journal of the acoustical society of America*, vol. 97, no. 1, pp. 491–504, 1995.

[28] P. Wagner, *The rhythm of language and speech: Constraints, models, metrics and applications.*, Habilitation thesis, University of Bonn, 2008.

[29] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "Marsec: A machine-readable spoken English corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47–53, 1993.

[30] G. Knowles, B. B. J. Williams, and L. Taylor, *A corpus of formal British English speech.* Longman, 1996.

[31] C. Auran, C. Bouzon, and D. Hirst, "The aix-marsec project: an evolutive database of spoken British English," in *Speech Prosody 2004, International Conference*, 2004.

[32] D. H. Klatt, "Vowel lengthening is syntactically determined in a connected discourse." *Journal of phonetics*, vol. 3, no. 3, pp. 129–140, 1975.

[33] J. P. Van Santen, "Contextual effects on vowel duration," *Speech Communication*, vol. 11, no. 6, pp. 513–546, 1992.

[34] A. E. Turk and S. Shattuck-Hufnagel, "Word-boundary-related duration patterns in English," *Journal of Phonetics*, vol. 28, no. 4, pp. 397–440, 2000.

[35] W. Jassem, *Intonation of Conversational English (educated Southern British)*. Nakl. Wroclawskiego Tow. Naukowego; skl. gl.: Dom Ksiazki, 1952, no. 45.