Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Convergence of Pitch Accents in a Shadowing Task

Iona Gessinger[1,2], *Antje Schweitzer*[3], *Bistra Andreeva*[1], *Eran Raveh*[1,2],
Bernd Möbius[1], *Ingmar Steiner*[1,2,4]

[1]Language Science and Technology, Saarland University, Germany
[2]Multimodal Computing and Interaction, Saarland University, Germany
[3]Institute for Natural Language Processing, University of Stuttgart, Germany
[4]German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
gessinger@coli.uni-saarland.de

Abstract

In the present study, a corpus of short German sentences collected in a shadowing task was examined with respect to pitch accent realization. The pitch accents were parameterized with the PaIntE model, which describes the f_0 contour of intonation events concerning their height, slope, and temporal alignment. Convergence was quantified as decrease in Euclidean distance, and hence increase in similarity, between the PaIntE parameter vectors. This was assessed for three stimulus types: natural speech, diphone based speech synthesis, or hidden Markov model (HMM) based speech synthesis. The factors tested in the analysis were *experimental phase* – was the sentence uttered before or while shadowing the model, *accent type* – a distinction was made between prenuclear and nuclear pitch accents, and *sex of speaker & shadowed model*. For the natural and HMM stimuli, Euclidean distance decreased in the shadowing task. This convergence effect did not depend on the accent type. However, prenuclear pitch accents showed generally lower values in Euclidean distance than nuclear pitch accents. Whether the sex of the speaker and the shadowed model matched did not explain any variance in the data. For the diphone stimuli, no convergence of pitch accents was observed.

Index Terms: phonetic convergence, pitch accent, PaIntE model, speech synthesis, shadowing task

1. Phonetic convergence

Phonetic convergence describes the phenomenon that interlocutors become phonetically more similar to each other during spoken communication [1]. According to Pickering and Garrod [2], a reason for such alignment between speakers is an internal automatic priming mechanism. Others propose that alignment is externally, socially motivated [3]. It is difficult to model the complex phenomenon in its totality, as it affects a great variety of phonetic features, segmental as well as suprasegmental, and it occurs to different degrees in different speakers.

One line of the literature on convergence is concerned with prosodic alignment. Levitan and Hirschberg [4] found evidence for *entrainment* – a collective term that includes convergence – of features such as pitch, intensity, and speaking rate in spontaneous conversations. Michalsky and Schoormann [5] linked the convergence of several pitch measures, found in spontaneous dialogues, to the perceived mutual attractiveness and likability of the interlocutors. Schweitzer *et al.* [6] examined pitch accents as parameterized by the PaIntE model (cf. Section 2) and found that seeing one's interlocutor led to divergence in pitch accent realization, whereas convergence occurred when interlocutors did not see each other. Perceived mutual likability merely enhanced the respective effect.

The three studies cited above all worked with corpora of spontaneous dyadic conversations. Most frequently, and in the present study as well, phonetic convergence is examined in shadowing experiments, where speakers repeat utterances they heard from a model speaker [7]–[9]. As opposed to other shadowing experiments, which use mainly mono- or bisyllabic utterances, the corpus of the present study contains short sentences.

With the development of more sophisticated text-to-speech synthesis techniques for spoken dialogue systems, there is a growing interest to examine phonetic convergence in the context of human-computer interaction [10]–[12]. Apart from natural model speakers, the shadowing experiment of the present study uses two types of synthetic voices to assess whether speakers respond similarly to them.

2. PaIntE model

The PaIntE model [13], [14] ***pa****rametrizes* ***int****onation* ***e****vents* by approximating their f_0 contour with the sum of a rising and a falling sigmoid as shown in Figure 1. Each parametrization takes the syllable carrying the intonation event σ^*, as well as one preceding and one following syllable σ as the basis for the analysis. The length of each syllable is normalized to 1; the three syllables thus fit into the range of -1 to 2.

The model function is characterized by six parameters: *c1* and *a1* represent the height and slope of the rising sigmoid, respectively; *c2* and *a2* provide the same information for the falling sigmoid. The parameters *d* and *b* describe the absolute height and the relative syllable alignment of the event peak, respectively.

If the f_0 contour cannot be fitted with two sigmoids, only a single sigmoid is applied (either rising or falling, see the dashed lines in Figure 1), leaving one set of *c* and *a* parameters unspecified. If a single sigmoid is not a good fit either, PaIntE only provides the mean f_0 value as the *d* parameter, leaving all other parameters unspecified.

For extracting the PaIntE parameters, f_0 is tracked using *get_*f_0 from the Entropic Signal Processing System (ESPS) [15]. The resulting raw contour is smoothed by the *smooth_*f_0 algorithm provided by the Edinburgh Speech Tools [16].

3. Corpus

The speech material was elicited in a shadowing experiment. A total of 56 German native speakers (12 male and 44 female) read short German sentences (declaratives and questions) from a screen, shadowed the same sentences (i.e., repeated the sentences after hearing them) from a male and a female voice, and

 10.21437/SpeechProsody.2018-46

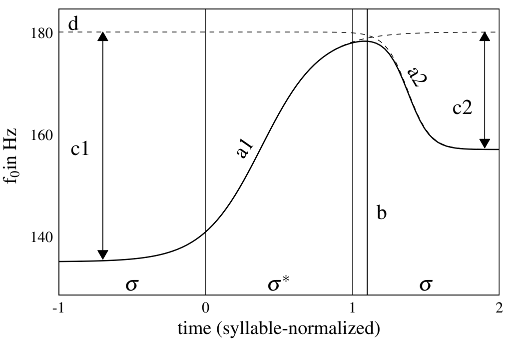


Figure 1: *Parametrization of an intonation event by the PaIntE model. The f_0 contour is approximated with the sum of a rising and a falling sigmoid function. The approximation is characterized by six parameters: a1, a2, b, c1, c2, and d. Figure adapted from Schweitzer* et al. *[6] and Möhler [14].*

finally read the sentences from a screen again. These three phases of the experiment are referred to as *baseline*, *shadowing*, and *post* production.

The corpus was originally collected to examine phonetic convergence at the segmental level. During the shadowing portion of the experiment, selected segmental features were manipulated to create room for convergence, i.e., participants were presented with the opposite of their preferred feature realization. Furthermore, different voice types were employed in the shadowing task. The participants heard either recordings of natural speakers, or one of two synthetic stimulus sets created using diphone and hidden Markov model (HMM) based synthesis, respectively. A detailed description of the procedure and analysis can be found in Gessinger *et al.* [17].

As it is likely that phonetic convergence occurs in the corpus with respect to other features than those manipulated explicitly, the present study looks for such convergence at the level of intonation. To that end, it focuses on the differences between baseline (1950 utterances) and shadowing (3360 utterances) production.

4. Selection of target syllables

To select the target syllables for the present study, the stimuli used in the shadowing task were manually annotated by a trained phonetician with respect to prenuclear (*P*) and nuclear (*N*) pitch accents. Nuclear pitch accents were further divided into those that occurred non-finally (*Nnf*) and those that coincided with the last syllable (the *ultima*) of a declarative (*Nud*) or a question (*Nuq*).

As f_0 contours and segment durations from the natural stimuli had been imposed onto the two types of synthetic stimuli during their generation, it was expected that the same pitch accent locations would be found in all three stimulus sets. Verification of this assumption revealed that this was true for the vast majority of the utterances. However, in a very few cases, additional pitch accents occurred in the synthetic stimuli. This is reflected in the numbers of identified pitch accents for each stimulus set in Table 1.

5. Hypotheses and operationalization

We assume that speakers converge, i.e., become more similar, to the stimuli they hear during the shadowing task, with respect to pitch accent realization. Based on the results of [17], we believe

Table 1: *Number of pitch accents identified in the three stimulus sets: prenuclear (P), non-final nuclear (Nnf), nuclear on ultima of declarative (Nud), nuclear on ultima of question (Nuq).*

Stimuli	P	Nnf	Nud	Nuq	Sum
Natural	129	60	12	18	219
Diphone	133	59	12	19	222
HMM	131	60	12	18	221

that this is the case for the group of speakers that shadowed natural recordings, as well as for the two groups of speakers that shadowed synthetic stimuli. The effect is expected to be stronger for nuclear than for prenuclear pitch accents, as the former are known to be perceptually more salient [18].

We do not expect the factor same-sex vs. mixed-sex pairing to have an effect on the degree of convergence, as such an effect was not observed in previous analyses of the same corpus. However, as it is often claimed in the literature that same-sex pairings show a higher degree of convergence [e.g., 19], the factor is taken into consideration during the data analysis.

Lastly, we include the two models per stimulus type as a factor in the analysis. The models differ with respect to their sex – speakers always shadowed a male and a female model – but of course exhibit a variety of other characteristics which could affect the degree of convergence to them. They might, for example, differ in the degree of attractiveness and likability, qualities which evidently have an effect on convergence between speakers [5], [6].

Similarity between a speaker and the shadowed model is assessed by calculating the Euclidean distance between the 6-dimensional PaIntE parameter vectors of the same syllable. The Euclidean distance is expected to decrease from baseline to shadowing production as an indication of increasing similarity, and hence, convergence.

6. Parameter handling

Parameters were extracted for every syllable of an utterance, target or non-target. Syllables for which PaIntE returned only the mean f_0 value as the *d* parameter, leaving all other parameters unspecified, were excluded from the analysis. This concerned about 6 % of the data.

From the remaining syllables, those cases were excluded for which one of the six (two sigmoids fitted) or four (one sigmoid fitted) parameter values fell in the 1st or 99th percentile for that parameter within the same speaker. About 10 % of the data were removed in this step. This procedure has proven to be good practice to remove potential measurement errors while keeping plausible yet atypical values in the data [6]. Such atypical values are expected to occur when a speaker converges to an interlocutor.

To subsequently calculate Euclidean distance between 6-dimensional PaIntE parameter vectors, the *c* (height) and *a* (slope) parameters were set to 0 wherever they were unspecified. Remember that this is the case when only a single sigmoid was fitted.

The values of the six parameters were standardized to speaker specific z-scores to eliminate differences linked to the speaker sex and give every parameter the same weight in the distance analysis.

Speaker syllables were matched with the corresponding stimulus syllables. During the shadowing condition, every sentence was shadowed once from a male and once from a female

Table 2: *Number of target syllable observations per stimulus type, condition, and accent type.*

Stimuli	Condition	P	Nnf	Nud	Nuq	Sum
Natural	Baseline	1578	484	222	164	2448
	Shadowing	1545	503	208	166	2422
Diphone	Baseline	1240	430	140	142	1952
	Shadowing	1220	438	137	141	1936
HMM	Baseline	1246	452	122	80	1900
	Shadowing	1205	426	121	92	1844

voice. To create a comparable data set for the baseline condition, every individual baseline production was matched twice: once with the male and once with the female voice.

Lastly, the Euclidean distance between the 6-dimensional PaIntE parameter vectors $\vec{s}$ of a speaker and $\vec{m}$ of the shadowed model was calculated for each syllable as shown in Equation (1).

$$d(\vec{s},\vec{m}) = \sqrt{\sum_{i=1}^{6} (s_{\mathrm{i}} - m_{\mathrm{i}})^2} \tag{1}$$

The target syllables as defined in Section 4 were selected for the final data set. It contains 12 502 observations that are distributed over the stimulus types, conditions, and accent types as given in Table 2.

7. Analysis and results

For the statistical analysis of the PaIntE data, linear mixed-effects models (LMMs) with Euclidean distance (eucDist) as the dependent variable were fitted for each stimulus type separately.[1] The model selection was carried out bottom-up, starting with a model which only included the random factor intercepts *speaker* and *sentence*. Then, theoretically relevant fixed factors were added to the model. Their influence on the model fit was assessed by means of the Akaike information criterion (AIC), which estimates the relative quality of a statistical model for a given data set by taking into account the likelihood function and the number of estimated parameters [22]. Similar to the method in Schweitzer *et al.* [23], the factor was kept in the model if the AIC value decreased by at least two points as compared to the model without the factor in question. After the individual factors, theoretically relevant interactions and eventually random slopes were tested and included or discarded in the same way. The models that were chosen to fit the data sets best are given in Equation (2), for the natural stimuli, in Equation (3), for the diphone stimuli, and in Equation (4), for the HMM stimuli.

$$\begin{aligned} \mathit{eucDist} &\sim \mathit{condition} * \mathit{accentType} + \mathit{model} + \\ &\quad (1 + \mathit{accentType} \mid \mathit{speaker}) + \\ &\quad (1 + \mathit{accentType} \mid \mathit{sentence}) \end{aligned} \tag{2}$$

$$\begin{aligned} \mathit{eucDist} &\sim \mathit{accentType} + \\ &\quad (1 + \mathit{accentType} \mid \mathit{speaker}) + \\ &\quad (1 + \mathit{accentType} \mid \mathit{sentence}) \end{aligned} \tag{3}$$

$$\begin{aligned} \mathit{eucDist} &\sim \mathit{condition} + \mathit{accentType} + \mathit{model} + \\ &\quad (1 + \mathit{condition} \mid \mathit{speaker}) + \\ &\quad (1 + \mathit{accentType} + \mathit{model} \mid \mathit{sentence}) \end{aligned} \tag{4}$$

[1] Models were fitted with the lme4 package (v1.1-13) [20] in R (v3.4.0) [21].

Table 3: *Parameter estimates, standard errors (SE), and t-values for the LMMs given in Equations* (2) *to* (4)*. The last column holds the p-values* $p(\chi^2)$ *calculated by likelihood ratio comparison of full model vs. null model.*

	Estimate	SE	t-value	$p(\chi^2)$
Natural				
(Intercept)	2.53	0.07	37.38	
cond.base	0.08	0.02	4.94	4×10^{-6}
acc.prenuc	−0.11	0.07	−1.74	0.04
mod.mod1	0.04	0.02	2.72	0.006
cond.1:acc.1	−0.03	0.02	−1.88	0.06
Diphone				
(Intercept)	2.73	0.09	31.56	
acc.prenuc	−0.11	0.07	−1.71	0.09
HMM				
(Intercept)	2.67	0.07	39.65	
cond.base	0.06	0.02	2.30	0.03
acc.prenuc	−0.14	0.07	−2.12	0.04
mod.mod1	0.07	0.06	1.19	0.2

The factor *condition* has two levels, *baseline* and *shadowing*; *accentType* has two levels as well, *prenuclear* and *nuclear*. The latter combines *Nnf*, *Nud*, and *Nuq*, since group sizes for each subgroup were too small for individual analysis. The factor *model* has the two levels *model1* and *model2*.

The factor *same-sex vs. mixed-sex pairing* did not account for variance in any of the data sets.

Table 3 shows the parameter estimates, and their respective standard errors and t-values. p-values ($p(\chi^2)$) were calculated by likelihood ratio comparison of the full model and the respective null model. The null model is constructed by removing only the factor in question from the fixed effects structure, while keeping the random effects structure unchanged. The significance level is at $\alpha = 0.05$.

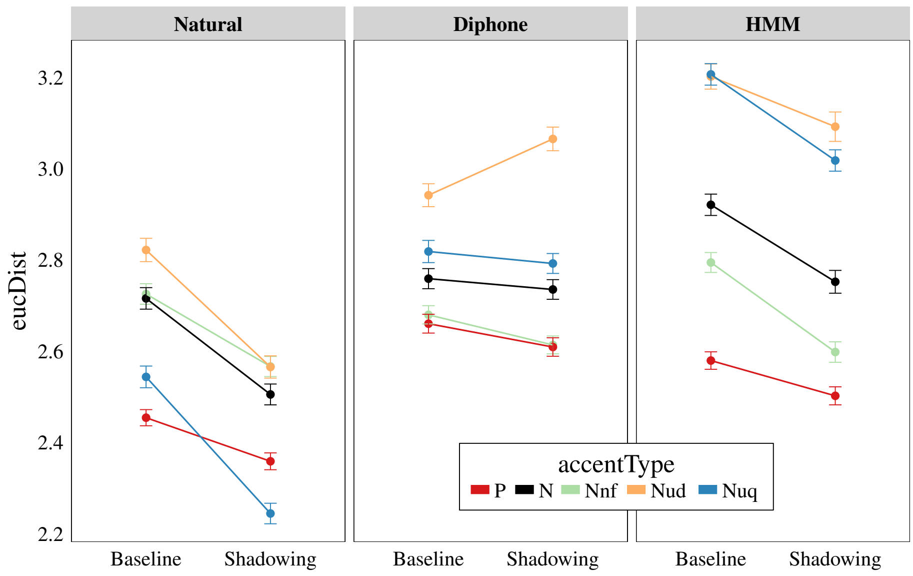
All model factors were sum coded and hence compare the first level of each factor to the grand mean. Remember that all three factors have two levels. The total effect size is therefore two times the estimate given in the table.

For the natural stimuli, the factor *condition* had an effect of size 0.16 on Euclidean distance; *accentType* (0.22) and *model* (0.08) influenced Euclidean distance as well. The interaction between *condition* and *accentType* was not significant ($p > 0.05$). For the diphone stimuli, none of the examined factors significantly influenced the dependent variable. For the HMM stimuli, *condition* (0.12) and *accentType* (0.28) had an effect on Euclidean distance. The effect of *model* was not significant ($p > 0.1$).

The direction of the effect needs to be interpreted while taking the reference level of each factor into account. This shows that, for the natural and the HMM stimuli, Euclidean distance is lower for prenuclear pitch accents than for nuclear pitch accents, and it decreases from baseline to shadowing condition. For the natural stimuli, Euclidean distance was also slightly higher when shadowing model1 (female) than when shadowing model2 (male).

Figure 2 shows a comparison of the mean Euclidean distance between the baseline and shadowing condition for each stimulus type. Results are given for the accent types prenuclear and nuclear, as well as for the three subgroups *Nnf*, *Nud*, and *Nuq* separately. Remember that for the nuclear pitch accent

Figure 2: *Comparison of mean Euclidean distance (eucDist) for the three stimulus types Natural, Diphone, and HMM, between the conditions Baseline and Shadowing. Results are given for the accent types prenuclear (P) and nuclear (N), as well as for the three subgroups, non-final nuclear (Nnf), nuclear on ultima of declarative (Nud), and nuclear on ultima of question (Nuq). A decrease in Euclidean distance indicates an increase in similarity and therefore convergence. The error bars illustrate the* 95 % *confidence interval, i.e.,* $2 \times SEM$ *(standard error of the means).*



subgroups, especially for *Nud* and *Nuq*, group sizes are fairly small when distributed over the three stimulus types.

The findings from the statistical analysis can be retraced in the graph and complemented with additional information about the nuclear pitch accent subgroups. For the natural and HMM stimuli, there is a clear drop in Euclidean distance from baseline to shadowing condition which is more pronounced in the natural stimuli. The effect appears to be stronger for the nuclear than for the prenuclear pitch accents. However, this was not confirmed by the statistical analysis. Although all nuclear accents combined show higher values in Euclidean distance for the natural and HMM stimuli, the *Nuq* accents fall below the level of the prenuclear accents in the natural condition.

The diphone condition clearly lacks the predicted increase in similarity. For the *Nud* accents, Euclidean distance even increases considerably. Overall, mean Euclidean distance is higher for the synthetic stimuli than for the natural ones.

8. Discussion

The results of the present study show that participants of a shadowing experiment converge to natural model speakers with respect to pitch accent realization as parameterized with the PaIntE model. A similar, yet slightly less pronounced effect was found for HMM based synthetic stimuli. Diphone based synthetic stimuli did not evoke such an effect in the participants. For nuclear accents on the ultima of a declarative, participants even diverged from the diphone stimuli. However, with regard to the low number of data points for the *Nud* subgroup, this result should be interpreted with caution.

Previous analysis of the same corpus at the segmental level showed that diphone stimuli are indeed capable of triggering convergence [17]. They were, for example, as successful as the natural and HMM stimuli in evoking convergence with respect to segmental pronunciation in the case of the phonemic allophone pair [ɪç] vs. [ɪk]. Therefore, the question arises whether diphone synthesis differs in a way from HMM based synthesis, which is particularly influential on the perception of intonation.

Remember that the f_0 contours and segment durations from the natural stimuli were imposed onto the two types of synthetic stimuli during their generation. The diphone stimuli were created with MBROLA [24]. The HMM based synthesis used HTS (v2.3) [25] with voices built from the BITS unit selection corpus [26]. Both techniques realize f_0 and segment duration according to the provided parameters. Deviations are minor in both cases. Therefore, deviation with respect to f_0 contour and segment duration does not seem to be a possible explanation for the difference in response.

One of the long-standing points of criticism toward diphone synthesis is the large number of concatenation points, which is detrimental to the perceived naturalness [27]. Although the diphone stimuli produced for the present study do not seem to be of inferior quality to the HMM stimuli, these underlying spectral discontinuities could play a role in their perception and prevent convergence.

Where convergence did occur, visual inspection of the data suggests a stronger effect for the nuclear pitch accents than for the prenuclear ones, as hypothesized. However, the statistical analysis did not support this observation. Remember that the analysis only distinguished between *P* and *N* pitch accents. As the nuclear pitch accent subgroups *Nnf*, *Nud*, and *Nuq* seem to behave rather differently, it would be informative to include them as separate factor levels. To that end, more data would have to be collected.

The significantly higher Euclidean distance for nuclear than for prenuclear pitch accents could be interpreted as an indication of overall greater variability in the realization of the former. Especially the *Nud* accents show the highest values in Euclidean distance for all stimulus types.

Whether the sex of the speaker and the shadowed model matched did not explain any variance in the given data. This goes against theories which suggest that same-sex pairings exhibit higher degrees of convergence than mixed-sex pairings and contributes to resolve the lack of shadowing studies which control for effects of the factor *same-sex* vs. *mixed-sex* pairing on convergence as recently diagnosed by Pardo *et al.* [9].

The results presented in this paper motivate further investigation of pitch accent convergence. Specifically, with respect to the PaIntE parametrization, a next step will be to give the parameters different weights, which reflect their importance in the perception of pitch accents.

9. References

[1] J. Pardo, “On phonetic convergence during conversational interaction”, *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[2] M. J. Pickering and S. Garrod, “Toward a mechanistic psychology of dialogue”, *Behavioral and Brain Sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[3] H. Giles, J. Coupland, and N. Coupland, “Accommodation theory: Communication, context and consequences”, *Contexts of Accommodation: Developments in applied sociolinguistics.*, pp. 1–68, 1991.

[4] R. Levitan and J. Hirschberg, “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions”, in *Interspeech*, 2011, pp. 3081–3084.

[5] J. Michalsky and H. Schoormann, “Pitch convergence as an effect of perceived attractiveness and likability”, in *Interspeech*, 2017, pp. 2253–2256.

[6] K. Schweitzer, M. Walsh, and A. Schweitzer, “To see or not to see: Interlocutor visibility and likeability influence convergence in intonation”, in *Interspeech*, 2017, pp. 919–923.

[7] A. Walker and K. Campbell-Kibler, “Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task”, *Frontiers in Psychology*, vol. 6, no. 546, 2015.

[8] J. W. Dias and L. D. Rosenblum, “Visibility of speech articulation enhances auditory phonetic convergence”, *Attention, Perception, & Psychophysics*, vol. 78, no. 1, pp. 317–333, 2016.

[9] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, “Phonetic convergence across multiple measures and model talkers”, *Attention, Perception, & Psychophysics*, vol. 79, no. 2, pp. 637–659, 2017.

[10] R. Levitan, *Acoustic-prosodic Entrainment in Human-human and Human-computer Dialogue*. Columbia University, 2014.

[11] L. Bell, J. Gustafson, and M. Heldner, “Prosodic adaptation in human-computer interaction”, in *15th International Congress of Phonetic Sciences (ICPhS)*, vol. 3, 2003, pp. 833–836.

[12] R. Coulston, S. Oviatt, and C. Darves, “Amplitude convergence in children’s conversational speech with animated personas”, in *Interspeech*, 2002, pp. 2689–2692.

[13] G. Möhler and A. Conkie, “Parametric modeling of intonation using vector quantization”, in *3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW)*, 1998, pp. 311–316.

[14] G. Möhler, “Improvements of the PaIntE model for F0 parametrization”, Institute for Natural Language Processing, University of Stuttgart, Tech. Rep., 2001.

[15] D. Talkin, “A robust algorithm for pitch tracking (RAPT)”, *Speech Coding and Synthesis*, pp. 497–518, 1995.

[16] S. King, A. W. Black, P. Taylor, R. Caley, and R. Clark. (1999). Edinburgh Speech Tools Library. version 1.2, [Online]. Available: http://www.cstr.ed.ac.uk/projects/speech_tools/.

[17] I. Gessinger, E. Raveh, S. Le Maguer, B. Möbius, and I. Steiner, “Shadowing synthesized speech – Segmental analysis of phonetic convergence”, in *Interspeech*, 2017, pp. 3797–3801.

[18] N. Jagdfeld and S. Baumann, “Order effects on the perception of relative prominence”, in *17th International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 958–961.

[19] G. Bailly and A. Lelong, “Speech dominoes and phonetic convergence”, in *Interspeech*, 2010, pp. 1153–1156.

[20] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4”, *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[21] R Core Team. (2017). R: A Language and Environment for Statistical Computing. version 3.4.0, [Online]. Available: https://www.r-project.org/.

[22] H. Akaike, “Information theory and an extension of the maximum likelihood principle”, in *2nd International Symposium on Information Theory*, 1973, pp. 267–281.

[23] K. Schweitzer, M. Walsh, S. Calhoun, H. Schütze, B. Möbius, A. Schweitzer, and G. Dogil, “Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation”, *Speech Communication*, vol. 66, pp. 65–81, 2015.

[24] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, “The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes”, in *4th International Conference on Spoken Language Processing (ICSLP)*, vol. 3, Oct. 1996, pp. 1393–1396.

[25] H. Zen and T. Toda, “An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005”, in *Interspeech*, 2005, pp. 93–96.

[26] T. Ellbogen, F. Schiel, and A. Steffen, “The BITS speech synthesis corpus for German”, in *International Conference on Language Resources and Evaluation (LREC)*, 2004, pp. 2091–2094.

[27] R. E. Donovan and P. C. Woodland, “A hidden Markov-model-based trainable speech synthesizer”, *Computer Speech & Language*, vol. 13, no. 3, pp. 223–241, 1999.