Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2aSC: Linking Perception and Production (Poster Session)

2aSC47. Acoustic and articulatory information as joint factors coexisting in the context sequence model of speech production

Daniel Duran*, Jagoda Bruni and Grzegorz Dogil

***Corresponding author's address: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, Stuttgart, 70569, BW, Germany, Daniel.Duran@ims.uni-stuttgart.de**

This simulation study presents the integration of an articulatory factor into the Context Sequence Model (CSM) (Wade et al., 2010) of speech production using Polish sonorant data measured with the Electromagnetic Articulograph technology (EMA) (Mücke et al., 2010). Based on exemplar-theoretic assumptions (Pierrehumbert 2001), the CSM models the speech production-perception loop operating on a sequential, detail-rich memory of previously processed speech utterance exemplars. Selection of an item for production is based on context matching, comparing the context of the currently produced utterance with the contexts of stored candidate items in memory. As demonstrated by Wade et al. (2010), the underlying exemplar weighing for speech production is based on about 0.5s of preceding acoustic context and following linguistic match of the exemplars. We extended the CSM by incorporating articulatory information in parallel to the acoustic representation of the speech exemplars. Our study demonstrates that memorized raw articulatory information--movement habits of the speaker--can also be utilized during speech production. Successful incorporation of this factor shows that not only acoustic but also articulatory information can be made directly available in a speech production model.

Published by the Acoustical Society of America through the American Institute of Physics

©2013 Acoustical Society of America [DOI: 10.1121/1.4799009]
Received 21 Jan 2013; published 2 Jun 2013

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

INTRODUCTION

We present results from a computer simulation study on the integration of an articulatory factor into the *Context Sequence Model* (CSM) of speech production (Wade et al., 2010) using Polish speech data. We enrich the model's original auditory memory with articulatory information, using continuous EMA signals directly in a speech production model.

In the view of articulatory phonology (Browman & Goldstein, 1989) gestures, i.e. dynamic actions containing specified parameters correlating with the vocal tract settings (including lips, tongue, glottis, velum etc.), occur sequentially or undergo overlapping during the course of speech production and perception. In the current simulation, articulatory gestures are investigated on exemplar-theoretic grounds (Pierrehumbert, 2001), and are depicted with the help of EMA recordings as articulatory habits of speakers.

Temporal organization of gestural movements has received broad attention in recent articulatory studies (Browman & Goldstein, 2000; Hermes et al., 2008). For example Nam et al. (2009) describe intrinsic model of syllable coordination based on 'coupled oscillators'. In this model CV structures (where C is a syllable onset) are described to exhibit the in-phase type of coordination, whereas VC structures are said to be organized by the anti-phase mode (where C is a syllable coda). Additionally, the authors (Nam et al., 2009) demonstrated a phenomenon described as C-Center Effect, which illustrates the stability of an articulatory distance maintained between the consonant and the vowel target in the onset CCV constructions for English. On the other hand, it has been shown that VCC constructions exhibit local organization of coordination, in which the first consonant gesture is related to the gesture of a vowel target. Moreover, analogous studies conducted on Italian (Hermes et al., 2008) and Polish (Mücke et al., 2010) seem to strengthen the observations on the C-Center Effect, showing presence of this type of coordination in the CV and CCV clusters, with no such bounding in the Polish coda VCC sequences.

Wade and Möbius (2007) proposed a model of speech perception which operates on a set of acoustic cues extracted from a rich memory representation at landmark positions. These landmarks are said to contain parameter values (like amplitude, speech rate and other information) extorted from the speech signal. Newly perceived sounds are identified by a comparison between stored speech items in context, and immediately encountered auditory instances. Thus, speech perception relies on activation of the perceived landmarks and robustness of the context undergoing the matching process. One of the central assumptions of this exemplar model is that the representations of speech, that are to be stored, have to be immediately available to the auditory cortex. The less abstraction that takes place at the front-end, the higher the plausibility granted to the speech representation.

The CSM models the speech *production—perception loop* operating on a sequential, detail-rich memory of previously processed speech utterance exemplars, grounding its assumptions in Exemplar Theory (Wade et al., 2010). In this model, selection of an item for production is based on context matching, comparing the context of the currently produced utterance with the contexts of stored candidate items in memory. According to Wade et al. (2010), context matching involves two types of information: *left* acoustic context and *right* linguistic context. Their simulations on a large speech corpus involved counting context similarities between the current and previously produced contexts. The authors conclude that the amount of context relevant for exemplar weighting during speech production is around 0.5 s, preceding and following the exemplar. Moreover, it is claimed that the context-level speech production is highly correlated with frequency effects previously assumed to be associated only with higher levels of speech organization.

Our study is an extension of the Context Sequence Model by enriching it with articulatory information in parallel to the acoustic representation of the speech exemplars. Successful incorporation of this factor shows that raw articulatory information, i.e. memorized movement habits of the speaker, can also be made directly available and utilized during speech production.

SPEECH MATERIAL

The speech material for the present simulation experiment is taken from a Polish speech data base containing acoustic and articulatory recordings from three native speakers. The data was collected originally for a study on sonorant voicing in word onset or coda position (Mücke et al., 2010; Bruni, 2011).

The corpus contains audio recordings and time-aligned articulatory measurements obtained through Electromagnetic Midsagittal Articulography (EMA) using a Carstens AG100, Electromagnetic Articulograph with 10 channels. Signals from four sensors were used for the simulation experiments: two for the tongue body movements (recorded traces of sensors placed 3 cm and 4 cm behind the tongue tip), one for the tongue tip and one

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

for the lip distance (from two sensors placed on the vermillion border of the upper and lower lip). Three adult, native speakers (one male, two female) were recorded producing a set of carrier phrases with embedded, systematically varied target words. The target words were produced in two different conditions: with and without emphasis. Utterances that could not be processed automatically (due to inconsistent labeling or some missing or incomplete signal files) have been omitted for this study. The resulting two splits of the corpus comprise in total 336 utterances in the emphasis part and 337 utterances in the non-emphasis part. These two parts of the corpus are labeled "emph" and "noemph" in the remainder of the text. Manual annotation at the phonetic level covers single consonants and consonant clusters in onset and coda positions of the target words along with the syllables' vowel. Only these labeled phone segments are used in this present study, along with stretches of the signals preceding the first segment to provide a left context for it–see below.

The EMA measurements are originally sampled at 250 Hz. The four EMA signals are combined such that each frame is represented by an eight-dimensional vector with each dimension corresponding to one EMA measurement in the horizontal or in the vertical planes. The acoustic data has been converted to provide a structurally similar representation. Amplitude envelopes with a sampling rate of 250 Hz were computed for eight logarithmically spaced frequency bands. This representation was chosen according to earlier work on the CSM by Wade et al. (2010). The choice of using such a representation is particularly motivated by the idea to reduce the amount of signal processing to a level which seems plausible from an auditory or cognitive point of view. In addition to the amplitude envelopes, we convert the audio signals to a mel-frequency cepstral coefficient (MFCC) representation. The 13-dimensional MFCCs were computed using the *mfcc* function of the Auditory Toolbox (Slaney, 1998) for Matlab. The parameter *frameRate* of the *mfcc* function was set to 250 (corresponding to a 2 ms window shift). All remaining parameters were not changed from their respective default values. The corresponding velocity and acceleration data is added for both acoustic and articulatory signals and was computed with Matlab's *diff* function.

METHOD

The present simulation experiment is based primarily on exemplar-theoretic assumptions formulated in the Context Sequence Model. In particular, the setup is designed as an extension to the experiments presented by Wade et al. (2010) in order to investigate the incorporation of articulatory data in the representations of speech items. The simulation experiment is implemented and carried out in Matlab.

The production targets of the CSM are defined according to the labeled phone segments from the Polish EMA corpus. The simulation is carried out on each speaker sub-corpus separately, not mixing data from different speakers in the memory. In order to avoid selection of segments from their original utterances, each utterance in the corpus is in turn excluded from the model's memory and treated as a new target utterance to be produced by the model. The remaining corpus data is treated as the memory sequence of speech exemplars. This approach forces the model to select a segment from a different utterance stored in memory. Note, that this also means that there will never be a perfect match as there are no identical acoustic/articulatory contexts in the memory at the signal level. All segments from the current target utterance are produced sequentially by the model. Candidate selection from the memory sequence is based on context matching.

The original proposal of the CSM is extended by incorporating articulatory data. We compare the performance of the model on three different data types: acoustic speech data, articulatory speech data and a combined representation of both acoustic and articulatory data. All data types are processed in the same way. The algorithms do not treat the articulatory signals different from the acoustic signals. According to Wade et al.'s (2010) study we first set the size of the left context to 0.5 s as our baseline. As this value is based on experiments considering an acoustic speech signal representation, we additionally investigate the influence of the context length. Let w_a and w_b denote the length of the context in seconds and let n_a and n_b denote the length of the context in frames (or samples) for the articulatory and the acoustic domains, respectively. The context lengths are varied systematically from a maximum of $w_a = 1.0$ s to a minimum length of $w_a = 0.004$ s, which is the minimum length at a sampling rate of 250 Hz corresponding to one frame of the discretized signal. The parameter w_b is varied accordingly between $w_b = 1.0$ s and $w_b = 0.004$ s for both the amplitude envelope representation and the MFCC representation of the acoustic signals.

The output sequence is initialized by copying w_a and/or w_b seconds of the original acoustic and/or the articulatory signals which immediately precede the first target segment (note, that there are no utterance-initial target segments in this study such that there is always a non-empty left context for each segment). This copied stretch of speech is interpreted as the original left context of the first segment that is to be produced by the model. Then, for each target segment, a stretch of w_a and/or w_b seconds from the output sequence provides the left context for the current production target. We follow Wade et al. (2010) and define the left-context similarity, or the *context-match*, according to the following formula:

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

$$\mathrm{cmatch}_{\mathrm{left}}(t_0, t_e, n_a, n_b) = \exp\{\sum_{d=1}^{D} A_{d,t_e-n_a:t_e-1} \cdot A_{d,t_0-n_a:t_0-1} + \sum_{d=1}^{D} B_{d,t_e-n_b:t_e-1} \cdot B_{d,t_0-n_b:t_0-1}\} \ ,$$

where $A_{d,m:n} = (A_{d,m}, \ldots, A_{d,n})^{\mathsf{T}}$ and $B_{d,m:n} = (B_{d,m}, \ldots, B_{d,n})^{\mathsf{T}}$ are the articulatory and acoustic sequences in dimension d from index m to n, respectively, D is the number of dimensions, t_0 is the start index of the current target segment, t_e is the start index of the candidate segment, and, n_a and n_b are the lengths of the left context in the articulatory and the acoustic domain, respectively.

The similarity is computed for the entire cloud of candidate segments, comparing the context of each candidate segment from the memory with the context of the current production target. The one exemplar with the highest match score wins and is selected for production. An important modification to the original CSM in this simulation is the exclusion of the right, i. e. the "linguistic" context from the context matching procedure. This is done due to the relatively small size of the corpus and its regular and, therefore, highly predictable structure. In order to avoid an unwanted selection bias, the right context is thus not considered. Exemplar selection in this scenario is more difficult as it has to rely solely on the raw acoustic and/or articulatory signal information of the left context.

Despite the underlying exemplar-theoretic assumption that all feedback during speech production is stored in memory and immediately available for future productions, the produced utterances in this simulation are not added to the corpus. For the sake of simplicity and in order to avoid artifacts, the underlying memory representation is not changed. Thus, the simulation has to be interpreted as a static simulation for each produced utterance which does not take into account processes such as memory decay or interference effects or any other kind of individual language change over time.

Evaluation Method

The manual annotation of the corpus is taken as the reference against which the results produced by the simulation experiments are evaluated. A *context accuracy* measure is defined for the evaluation at the segment-label level. It is defined as the proportion of produced segments for which their original context in the memory sequence from which they were selected matches the production context. The *context*, in this sense, is defined as the labels preceding and following a given segment. If, for example, a [p] segment was selected from a […upr…] context in the memory sequence for the production of that segment in a […ipr…] context, its right context would be counted as correct, while its left context would be counted as wrong. The baseline for this measure is defined as a random selection of a segment from the set of available candidates for each target item. The corresponding baseline values are estimated for each speaker sub-corpus based on the proportion of available segments with correct contexts.

RESULTS

Due to space limitations, we report only the total context accuracy which considers both the left and the right context of each produced segment. Tables 1 and 2 show the baseline values and the context accuracy results for all three speakers and all data types for the "emph" and the "noemph" parts of the corpus, respectively. The tables show that the context accuracy is consistently higher for articulatory data (column "EMA") than it is for acoustic data alone (columns "ENV" and "MFCC") or the combined representations (columns "ENV+EMA" and "MFCC+EMA"). For all data types, the performance of the production model is clearly above the baseline.

Tables 3, 4 and 5 show the context accuracy as a function of the two experimental parameters w_a and w_b for the combined data type from the "emph" part of the corpus for speakers 1, 2 and 3, respectively. Due to space limitations, not all results are shown for every tested context window length combination. Table columns correspond to specific lengths w_a of the articulatory context window and rows correspond to settings of the acoustic context window length w_b. The corresponding context accuracy results for the "noemph" part of the corpus are shown in tables 6, 7 and 8 for speakers 1, 2 and 3, respectively.

The results of the context length variations for the combined data show that performance is improved in general by decreasing the acoustic context size in comparison to the initially assumed optimal length of half a second. A comparison between tables 1 and 2 on the one hand the results shown in tables 3-8 indicate that combined data representations with asymmetric context sizes for articulatory and acoustic data yield the best results in terms of context accuracy. A direct comparison of the model's performance on the uni-modal data shows mostly better results on the amplitude envelopes than on the MFCCs. However, in combination with the EMA data, MFCCs yield higher accuracies, especially with asymmetric context sizes, as shown in tables 3-8.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

TABLE 1. Context accuracy on the "emph" part of the Polish corpus for both audio representations using amplitude envelopes (ENV) and MFCCs and the articulatory EMA data with $w_a = w_b = 0.5$ s.

	baseline	**acoustic**		**articulatory**	**combined**	
		ENV	**MFCC**	**EMA**	**ENV+EMA**	**MFCC+EMA**
Speaker 1	0,221	0,705	0,698	0,772	0,712	0,747
Speaker 2	0,220	0,754	0,744	0,840	0,754	0,765
Speaker 3	0,219	0,782	0,811	0,875	0,786	0,814

TABLE 2. Context accuracy on the "noemph" part of the Polish corpus for both audio representations using amplitude envelopes (ENV) and MFCCs and the articulatory EMA data with $w_a = w_b = 0.5$ s.

	baseline	**acoustic**		**articulatory**	**combined**	
		ENV	**MFCC**	**EMA**	**ENV+EMA**	**MFCC+EMA**
Speaker 1	0,219	0,781	0,749	0,795	0,781	0,763
Speaker 2	0,217	0,832	0,775	0,857	0,832	0,782
Speaker 3	0,217	0,768	0,757	0,871	0,768	0,779

TABLE 3. Context accuracy for speaker 1 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "emph" part of the corpus. Maxima are printed in bold face, and minima in italics.

ENV+EMA

w_b	w_a				
	1.0	**0.6**	**0.5**	**0.4**	**0.004**
1.0	*0.566*	0.569	0.569	0.569	0.569
0.5	0.612	0.712	0.712	0.712	0.705
0.2	0.609	0.772	0.772	0.772	0.779
0.05	0.605	0.733	0.737	0.751	0.733
0.01	0.612	0.794	0.786	0.783	0.676
0.004	0.619	0.769	**0.815**	0.794	0.690

MFCC+EMA

w_b	w_a				
	1.0	**0.6**	**0.5**	**0.4**	**0.004**
1.0	*0.573*	0.587	0.591	0.591	0.584
0.5	0.609	0.744	0.747	0.747	0.694
0.2	0.644	0.808	**0.829**	0.804	0.783
0.05	0.655	0.772	0.779	0.754	0.694
0.01	0.580	0.740	0.783	0.790	0.641
0.004	0.577	0.719	0.747	0.779	0.644

TABLE 4. Context accuracy for speaker 2 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "emph" part of the corpus. Maxima are printed in bold face, and minima in italics.

ENV+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.665	0.665	*0.662*	*0.662*	*0.662*
0.5	0.765	0.754	0.754	0.754	0.754
0.2	0.811	0.815	**0.815**	0.815	0.811
0.05	0.762	0.790	0.783	0.808	0.783
0.01	0.737	0.783	0.779	0.754	0.690
0.004	0.754	0.783	0.790	0.772	0.665

MFCC+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.673	0.676	0.673	0.673	0.658
0.5	0.751	0.754	0.765	0.758	0.744
0.2	0.815	0.833	0.815	0.819	0.797
0.05	0.801	0.847	0.843	0.829	0.726
0.01	0.829	0.854	**0.861**	0.847	*0.601*
0.004	0.815	0.840	0.847	**0.861**	0.612

TABLE 5. Context accuracy for speaker 3 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "emph" part of the corpus. Maxima are printed in bold face, and minima in italics

ENV+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.632	0.632	0.632	0.632	0.632
0.5	0.654	0.782	0.786	0.786	0.782
0.2	0.596	0.779	0.782	0.782	0.782
0.05	*0.593*	0.721	0.696	0.700	0.707
0.01	0.604	0.754	0.725	0.736	0.693
0.004	0.646	**0.796**	0.764	0.754	0.643

MFCC+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.639	0.636	0.636	0.636	0.632
0.5	0.675	0.814	0.814	0.814	0.811
0.2	0.629	0.779	0.796	0.814	0.796
0.05	0.682	0.825	0.829	0.836	0.721
0.01	0.686	0.886	0.864	0.861	0.689
0.004	0.664	0.882	**0.889**	0.879	*0.604*

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

TABLE 6. Context accuracy for speaker 1 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "noemph" part of the corpus. Maxima are printed in bold face, and minima in italics.

ENV+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	*0.451*	0.455	0.459	0.459	0.459
0.5	0.485	0.760	0.781	0.781	0.781
0.2	0.485	0.770	0.777	0.774	0.770
0.05	0.451	0.777	0.753	0.774	0.728
0.01	0.478	0.774	0.802	0.799	0.742
0.004	0.470	0.802	**0.809**	0.784	0.763

MFCC+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.429	*0.425*	0.437	0.437	0.437
0.5	0.470	0.763	0.763	0.760	0.753
0.2	0.485	0.802	0.792	0.788	0.763
0.05	0.478	0.802	**0.809**	0.784	0.686
0.01	0.451	0.784	0.806	0.792	0.622
0.004	0.448	0.756	0.799	0.763	0.625

TABLE 7. Context accuracy for speaker 2 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "noemph" part of the corpus. Maxima are printed in bold face, and minima in italics.

ENV+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	*0.718*	*0.718*	*0.718*	*0.718*	*0.718*
0.5	0.807	0.836	0.832	0.832	0.832
0.2	0.818	0.829	0.825	0.825	0.825
0.05	0.782	0.804	0.800	0.800	0.754
0.01	0.804	0.829	0.825	0.804	0.743
0.004	0.804	**0.861**	0.843	0.839	0.725

MFCC+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.732	0.732	0.736	0.739	0.718
0.5	0.768	0.782	0.782	0.782	0.775
0.2	0.764	0.832	0.843	0.836	0.807
0.05	0.775	0.839	0.839	0.832	0.725
0.01	0.814	0.857	0.854	0.843	0.618
0.004	0.814	0.846	0.854	**0.861**	*0.593*

TABLE 8. Context accuracy for speaker 3 as a function of context window length based on ENV+EMA data (left) and MFCC+EMA data (right) of the "noemph" part of the corpus. Maxima are printed in bold face, and minima in italics.

ENV+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.573	0.573	0.573	0.573	0.573
0.5	0.562	0.764	0.768	0.768	0.768
0.2	0.551	0.796	0.793	0.793	0.793
0.05	*0.507*	0.739	0.743	0.743	0.743
0.01	0.522	0.796	0.779	0.779	0.732
0.004	0.540	**0.836**	0.829	0.832	0.779

MFCC+EMA

w_b	w_a				
	1.0	0.6	0.5	0.4	0.004
1.0	0.584	0.584	0.584	0.584	0.577
0.5	0.573	0.775	0.779	0.771	0.757
0.2	0.584	0.793	0.789	0.789	0.754
0.05	*0.540*	0.829	0.829	0.821	0.768
0.01	0.599	0.839	0.839	0.850	0.650
0.004	0.591	0.875	**0.900**	0.896	0.689

CONCLUSION

We presented an extension to the Context Sequence Model which integrates articulatory information into its exemplar based, context-sensitive production process. Candidate exemplars are specified in context based on a similarity score which takes into account acoustic and articulatory information.

It has been documented, that Polish sonorants preceded by voiceless obstruents in word-final positions are desyllabified, i.e. they are not licensed for [voice] (Gussmann, 1997). Moreover, articulatory investigation of Polish CCV and VCC clusters (Mücke et al., 2010), demonstrated no coupling relations like C-Center Effect in the coda positions contrary to the strong bonding in onsets. The Polish EMA corpus contains precisely such clusters. Thus, the fact that the model selects segments from the memory which are appropriate in the given contexts indicates the presence of contextual information. This observation holds for both the acoustic as well as the articulatory domains.

This present computer simulation study demonstrates that memorized raw articulatory information—movement habits of the speaker—can be available during speech production. Both modalities can be represented in memory and processed in parallel. Successful incorporation of this factor shows that not only acoustic but also articulatory information can be made directly available during speech production.

It is hypothesized that without involving any complex front-end transformations (like acoustic/articulatory conversion and match), the amplitude envelope representation is robust enough and immediately available to the auditory cortex. Such a representation appears to be ideally suited for memory representations for exemplar based speech perception and production.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36

ACKNOWLEDGMENTS

This research was funded by the German Research Foundation (DFG), grant SFB 732, A2, "Incremental Specification in Context". EMA recordings were conducted thanks to the courtesy of Martine Grice and Doris Mücke from the Institute of Linguistics at the University of Cologne.

REFERENCES

Bruni, J. (**2011**). *Sonorant voicing specification in phonetic, phonological and articulatory context*. Dissertation, Universität Stuttgart. http://elib.uni-stuttgart.de/opus/volltexte/2011/6311

Browman, C.P., Goldstein, L. **(1989)**. Articulatory gestures as phonological units. Phonology, 6, 20-251.

Browman, C.P., Goldstein, L. **(2000)**. Competing constraints on intergestural coordination and self-organization of phonological structures. Bulletin de la Communication Parlee, 5, 25-34.

Gussmann, E. **(1992)**. Resyllabification and Delinking: The Case of Polish Voicing. Linguistic Inquiry 23, 29-56.

Hermes, A., Grice, M., Mücke, D., Niemann, H. **(2008)**. "Articulatory indicators of syllable affiliation in word initial consonant clusters in Italian." Proceedings of the 8th International Seminar on Speech Production, Strasbourgh, France, 433-436.

Mücke, D., Sieczkowska, J., Niemann, H., Grice, M., and Dogil, G. **(2010)**. "Sonority Profiles, Gestural Coordination and Phonological Licensing: Obstruent-Sonorant Clusters in Polish". Presented at the 12th Conference on Laboratory Phonology (LabPhon), Albuquerque, New Mexico.

Nam, H., Golstein, L., Saltzman, E. **(2009)**. "Self organization of syllable structure: a coupled oscillator model." In F. Pellegrino, E. Marisco, & I. Chiotran (Eds.). Approaches to phonological complexity, 299-328.

Pierrehumbert, J. **(2001)**. "Exemplar dynamics: Word frequency, lenition, and contrast." In J. Bybee, & P. Hopper (Eds.), Frequency and the emergence of linguistic structure, 137-157. Amsterdam: Benjamins.

Slaney, M. **(1998)**. *Auditory Toolbox*. https://engineering.purdue.edu/~malcolm/interval/1998-010/. (accessed 2009-06)

Wade, T., and Möbius, B. **(2007)**. "Speaking rate effects in a landmark-based phonetic exemplar model," 8th Annual Conference of the International Speech Communication Association — Interspeech, pp. 402–405.

Wade, T., Dogil, G., Schütze, H., Walsh, M., and Möbius, B. **(2010)**. "Syllable Frequency Effects in a Context-sensitive Segment Production Model." Journal of Phonetics **38** (2): 227–239.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 141.58.160.176 On: Thu, 06 Mar 2014 10:16:36