CONTEXT SEQUENCE MODEL OF SPEECH PRODUCTION ENRICHED WITH ARTICULATORY FEATURES

Daniel Duran, Jagoda Bruni, Hinrich Schütze & Grzegorz Dogil

Institute for Natural Language Processing, University of Stuttgart, Germany
daniel.duran@ims.uni-stuttgart.de; jagoda.bruni@ims.uni-stuttgart.de
hs999@ifnlp.org; grzegorz.dogil@ims.uni-stuttgart.de

ABSTRACT

This study describes integration of an articulatory factor into the exemplar-based Context Sequence Model of speech production, CSM [10], which builds on the concept of a speech perception-production loop. It has been demonstrated that selection of new exemplars for speech production is based on about 0.5 s of preceding acoustic context and following linguistic match of the exemplars. This investigation presents the role of the articulatory features integrated in the exemplar weighing processes.

Keywords: speech production, EMA, articulatory phonology, Context Sequence Model

1. INTRODUCTION

In the view of articulatory phonology [1] 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 [4]. In this study, articulatory gestures are investigated in the framework of Exemplar Theory [10], and are depicted with the help of EMA as articulatory habits of individual speakers.

Since temporal organization of the articulators has become an important factor of analysis, it is the organization of gestural settings within the syllable structure that has received particularly broad attention in the literature ([2, 6] and the others). In their work on 'coupled oscillators', [8] propose an intrinsic mode of syllable coordination, where the in-phase mode produces the coordination of CV structures (where C is a syllable onset), whereas VC structures are coordinated by the anti-phase mode (where C is a syllable coda). Moreover, the authors demonstrated competitive articulatory patterns of complex CCV onsets for English, observing characteristics described as C-Center Effect–a stable distance of the consonants with regards to the vowel target, measured as the interval between the mean value of the onset consonantal targets and the vowel. On the other hand, VCC constructions are said to show local organization of coordination, in which the first consonant gesture is related to the gesture of a vowel target. Similar studies conducted on Italian [6] and Polish [7] demonstrated onset C-Center coordination in the CV and CCV clusters with no such bounding in the Polish coda VCC sequences. While the studies described above operate on a unit-level speech analysis, [10] and [11] proposed an exemplar-based model where representations of speech are considered as unstructured stretches of continuous speech.

The model of speech perception described in [11] operates on a set of acoustic cues extracted from the 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 items in context, and immediately encountered auditory instances. Thus, speech perception relies on the activation of the perceived landmarks and robustness of the context undergoing matching process. Moreover, 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 auditory cortex (the less abstraction that takes place at the front-end, the better for the model).

The study described in [10] demonstrates that speech production takes place at the segmental level, where each segment of an utterance is represented by an exemplar cloud taken from the memory of previously stored speech items. Production of speech is thus a process of token weighing by matching the currently produced context with the one in which the token occurred originally. According to [10], context matching involves two types of information: *left* acoustic context and *right* linguistic context. The simulations on a large speech corpus, involved counting

context similarities between the current and previously produced contexts. As a result, the authors demonstrated that the amount of context relevant for exemplar weighting during speech production oscillates around 0.5 s, preceding and following the exemplar. Moreover, it is claimed that the 'context-level speech production' [10] is highly correlated with frequency effects previously assumed to be associated only with higher levels of speech organization.

The simulation experiments presented here are based on the CSM; see [10] for a detailed technical description. We enrich the solely auditory memory of the original CSM with articulatory information, using continuous EMA signals directly in a speech production model.

2. METHOD

2.1. Database

We use data from a corpus that was originally created to investigate the C-Center effect in Polish [7]. Three native speakers were recorded (two female, one male) with a 2D Electromagnetic Articulograph, Carstens AG100, 10 channels. Sensors were placed on the vermillion border of the upper and lower lip and on the tongue (3 sensors: 1 cm, 3 cm and 4 cm behind the tongue tip). The sensor on the tongue tip and two sensors attached to the dorsum were used for analyzing coronal sounds, vowel articulation and velar consonants. Two additional reference sensors were attached to the nose and the upper gums to correct head movements. The data was sampled at 400 Hz, down sampled to 250 Hz, smoothed with a low-pass filter at 40 Hz. All data was converted to Simple Signal File Format (SSFF), and manually labeled in EMU Speech Database System [3]. Target words with simple onsets and codas, as well as onset and coda clusters containing a voiceless stop and a sonorant were recorded in the following carrier phrases, which guarantee identical contexts of tongue movements for all target consonants and clusters: 1. onset position: '*Ona mówi pranie aktualnie*' ('*She is saying laundry currently*'), 2. coda position: '*Ona powiedziała Cypr aktualnie*' ('*She is saying Cyprus currently*'). The underlined target word was recorded with an emphasis articulation mode. See Table 1 for a word list. A total of 336 (3 × 112) utterances was selected for our experiment. In each utterance, only the consonant or consonant cluster of the target word along with the corresponding following or preceding vowel was labeled at the phone level. We did not use the gestural landmark labels for our experiments but only the continuous EMA signals which depict the articulatory habits of the speakers.

Following [10], the acoustic data was converted to an 8-dimensional envelope representation, sampled at 250 Hz for computational efficiency.

Table 1: Structure of target words.

	Onset	Coda
/p/	padnij	typ
/k/	kadisz	tik
/l/	labrys	gil
/r/	rabin	tir
/p/+/l/	plamić	ZUPL
/p/+/r/	pranie	Cypr
/k/+/l/	klawisz	cykl
/k/+/r/	krasić	WIKR

2.2. Integration into the CSM

We implemented our production experiment such that it uses unprocessed acoustic and articulatory data. In an exemplar theoretic model of speech production all feedback, including articulatory habits of speakers, is stored in detail in the memory providing a basis for future productions. We simulate the production of one target utterance, by taking that target out of the corpus and using the remaining corpus data as the memory. Accordingly, the next target is than taken aside and the remaining corpus is used as the memory, and so on. As production targets for each utterance we take the labeled phonetic segments (842 in total; 281 + 281 + 280) and production proceeds as in [10] on the segmental level. The underlying motor commands and articulatory gestures are considered only indirectly through their resulting vocal tract shapes reflected in the EMA signals. Therefore, there is no motor-planning involved in the sense that a stored speech item (a phone, a syllable etc.) has associated motor-commands and articulatory gestures from which articulatory movements and specific vocal tract shapes have to be generated. The stored gestures are real, ecologically observed recordings of speech organ configurations.

The simulation iterates over all utterances and takes a sequence of phonetic labels $T = [t_1 \ldots t_n]$, with $n \leq 3$, from the currently produced utterance (e.g. T = [p r a] for a carrier phrase with the word 'pranie'). This is the production target for which an output sequence is then produced. First, we initialize each iteration by copying 0.5 s of the original acoustic and/or EMA signal preceding the first segment t_1 to the output sequence, treating it

as if it were the target's initial left context. Then, for each segment $t_i \in T$ its left context is taken from the output sequence, i.e. a stretch of 0.5 s from the speech that has been produced immediately before the current segment t_i.

Another modification to the original CSM is that we did not use the right context to match the candidate exemplars' contexts with the production targets' right context. This was done because of the small size of our corpus and the regular, highly predictive structure of the carrier phrases. We thus wanted to exclude this potential selection bias and test whether the model still operates under the harder conditions of relying only on the left context (i.e. the raw acoustic and/or articulatory signal) without the linguistic information of the right context.

The entire corpus, excluding the utterance in which *T* originally occurs, is treated as the memory sequence of stored speech items. For the sake of simplicity we do not add the produced speech to the memory for the following production iterations. The underlying memory representation is not changed. Particularly, memory decay or interference effects are not considered, i.e. we treat the corpus data as a snapshot of the memory at one instance in time. Therefore, the actual order of the sentences in the corpus and that of their respective production is arbitrary.

3. EXPERIMENTAL RESULTS

Tables 2-4 show summaries of the model's confusion in consonant segment selection in terms of syllable position and type. 'Wrong type' means that the model's actual production target was a segment in a CC (or C) syllable context but it selected a candidate from a C (or CC) context instead. 'Wrong pos.' indicates errors of the model with respect to syllable position, e.g. selecting a candidate from a syllable coda instead of a syllable onset. The columns 'Ema' and 'Env.' show the results where the model used only the EMA signals and the amplitude envelopes of the acoustic signal, respectively. The third column 'Ema+Env.' shows the results for the combined signals. Note that the total number of consonant targets was 169 for speakers 1 and 2 and 168 for speaker 3.

Assuming a random selection of segments from the set of consonants gives a baseline of 50% error rate since all consonants appear in both onset and coda position and CC and C syllable contexts. The results are clearly better than this baseline and thus allow the interpretation that using a combined corpus of EMA and the acoustic signals might improve the model's performance as compared to its original implementation based only on the acoustic signal. A slight improvement with respect to the production segments' syllable contexts can be seen by comparing the 'Env.' with the combined 'Ema+Env.' results. Moreover, for two of our three speakers, the results based on EMA signals are better as compared to the acoustic data. However, the differences are so small that more research is needed to further investigate this observation.

Table 2: Confusion summary for Speaker 1.

	Ema	Env.	Ema+Env.
wrong type	25	24	24
wrong pos.	3	2	2
correct type	144	145	145
correct pos.	166	167	167
all correct	144	145	145
all wrong	3	2	2

Table 3: Confusion summary for Speaker 2.

	Ema	Env.	Ema+Env.
wrong type	20	21	21
wrong pos.	0	0	0
correct type	149	148	148
correct pos.	169	169	169
all correct	149	148	148
all wrong	0	0	0

Table 4: Confusion summary for Speaker 3.

	Ema	Env.	Ema+Env.
wrong type	15	22	21
wrong pos.	0	0	0
correct type	153	146	147
correct pos.	168	168	168
all correct	153	146	147
all wrong	0	0	0

The model makes only few mistakes in choosing between onset and coda contexts. Only for speaker 1, the model selects some segments from the wrong syllable position (making 7 such errors in total). Choosing between CC and C syllable types, on the other hand, causes more problems. This is shown in more detail in tables 5-7. The rows 'false C' show how often the model has erroneously produced a consonant segment from a simple syllable onset or coda instead of a complex one, and, accordingly, 'false CC' designates the selection of candidates from complex consonant clusters for simple C targets.

Table 5: Syllable type errors for Speaker 1.

	Ema	Env.	Ema+Env.
false C	9	8	8
false CC	16	16	16

Table 6: Syllable type errors for Speaker 2.

	Ema	Env.	Ema+Env.
false C	8	11	11
false CC	12	10	10

Table 7: Syllable type errors for Speaker 3.

	Ema	Env.	Ema+Env.
false C	7	10	9
false CC	8	12	12

4. DISCUSSION AND CONCLUSIONS

We expanded a context-sensitive segment production model, the CSM [10], adding continuous, raw EMA signals (i.e. without any gestural landmark annotations) to its memory. We have shown that a speech production model originally based on a rich representation of an acoustic speech signal can also be applied to articulatory data as represented by raw EMA signals. These results seem to indicate that it might be possible to base a speech production model with a rich memory representation on a combination of the acoustic signal and EMA traces, or even on EMA data only.

The fact that the model produces simple codas from existing complex codas or even onsets might lie in the irregularity and variability of the articulatory and, as a consequence, also acoustic characteristics of Polish codas. It has been documented, that sonorants preceded by voiceless obstruents in word final positions are desyllabified, i.e. they are not licensed for [voice] [5]. Moreover, articulatory investigation of Polish CCV and VCC clusters [7], demonstrated no coupling relations like C-Center Effect in the coda positions contrary to the strong bonding in onsets.

It has been observed, that the speech envelope representation is robust enough and immediately available to the auditory cortex, without involving any complex front end transformations (like acoustic/articulatory conversion and match). Such a representation appears to be ideally suited for memory representations for exemplar based speech perception and production. Similarly, articulatory habits of speakers, stored as 'raw' movement trajectories can be made directly available for the context sequence model speech production.

5. FUTURE WORK

As the size of the corpus used is small, more research is needed to investigate to what degree raw EMA data can be used on its own, and to what degree the addition of EMA data can improve acoustic-based models of speech production.
Moreover, it seems to be worth investigating the importance of acoustic and articulatory landmarks in correlation with a rich and context-dependent memory representation of speech (as assumed in the CSM) and EMA signals.

6. ACKNOWLEDGMENTS

This research was funded by the German Research Foundation (DFG), grant SFB 732, A2.

EMA recordings were conducted thanks to the courtesy of Martine Grice and Doris Mücke from the Institute of Linguistics at the University of Cologne.

7. REFERENCES

[1] Browman, C.P., Goldstein, L. 1989. Articulatory gestures as phonological units. *Phonology* 6, 20-251.

[2] 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.

[3] EMU Speech Database System. Available online: *http://emu.sourceforge.net*

[4] Fowler, C.A. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14, 3-28.

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

[6] 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.

[7] Mücke, D., Sieczkowska, J., Niemann, H., Grice, M., Dogil, G. 2010. *Sonority Profiles, Gestural Coordination and Phonological Licensing: Obstruent-Sonorant Clusters in Polish.* Presented at the *LabPhon Conference 2010* Albuquerque, New Mexico.

[8] Nam, H., Golstein, L., Saltzman, E. 2009. Self organization of syllable structure: a coupled oscillator model. In Pellegrino, F., Marisco, E., Chiotran, I. (eds.), *Approaches to Phonological Complexity*, 299-328.

[9] Pierrhumbert, J. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic structure*. Amsterdam: Benjamins, 137-157.

[10] Wade, T., Dogil, G., Schütze, H., Walsh, M., Möbius, B. 2010. Syllable frequency effects in a context-sensitive segment production model. *Journal of Phonetics* 38, 227-239.

[11] Wade, T., Möbius, B. 2007. Speaking rate effects in a landmark-based phonetic exemplar model. *Proceedings of Interspeech* Antwerp, Belgium, 402-405.