MODELING MULTI-MODAL FACTORS IN SPEECH PRODUCTION WITH THE CONTEXT SEQUENCE MODEL

Daniel Duran, Jagoda Bruni, Grzegorz Dogil

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Abstract. This article describes modeling speech production with multi-modal factors integrated into the Context Sequence Model (Wade et al. 2010). It is posited that articulatory information can be successfully incorporated and stored in parallel to the acoustic information in a speech production model. Results demonstrate that a memory sensitive to rich context and enlarged by the additional inputs facilitates exemplar weighing and selection during speech production.

1 Introduction

This study describes the integration of articulatory and acoustic factors in the exemplar-based Context Sequence Model (CSM) of speech production (Wade et al. 2010). Based on exemplar-theoretic assumptions (Pierrehumbert 2001), the CSM models the speech *production-perception loop* as operating on a flat, sequential, detail-rich memory of previously processed speech utterance exemplars.

Wade et al. (2010) describe speech production as taking place at the segmental level, where each segment of an utterance is represented by an exemplar cloud taken from the context-preserving memory of previously stored speech items. It is posited that in speech production an exemplar cloud is created for each target segment and each token undergoes weighting through a match between the current production context and the originally produced one. Thus, production of whole utterances containing segments is modeled based on more than the unit specification, where acoustic information is completed step by step and rooted in the developing production context. Characterization of the segments is made through the analysis of both the preceding and the following contexts of currently produced utterances. The preceding (*left*) context is composed of acoustic information from recently produced segments, while the following (*right*) context is the linguistic information about what will be produced in the following step. Disambiguation among all available exemplars is modeled as a process of token weighting by matching the current production context with the memorized one in which the token originally occurred. As a result, Wade and colleagues demonstrated that the amount of context relevant for exemplar weighting during speech production oscillates around 0.5 s, preceding and following the exemplar.

Numerous, recent articulatory studies are concerned with the temporal organization of gestural movements (e.g. Browman & Goldstein, 2000; Hermes et al., 2008). Nam et al. (2009) describe an intrinsic model of syllable coordination based on 'coupled oscillators', where CV structures (where C is a syllable onset) are said 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). Moreover, the authors (Nam et. al, 2009) describe a so-called C-Center Effect, which demonstrates stability of articulatory distance maintained between the consonant and vowel targets in CCV English clusters. It has been observed that their coda VCC counterparts exhibit local type of organization, in which only the first consonant gesture is related to the gesture of a vowel target. Similar studies conducted on Italian (Hermes et al., 2008) and Polish (Mücke et al., 2010) demonstrate similar C-Center Effect patterning, showing this type of coordination in the CV and CCV clusters, with no such bounding in the Polish coda VCC sequences (see figure 1).

Studies conducted on Polish (Mücke et al. 2010) demonstrated competitive articulatory patterns of complex CCV onsets, 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. VCC constructions, on the other hand, showed local organization of coordination, in which the first consonant gesture is related to the gesture of a vowel target.

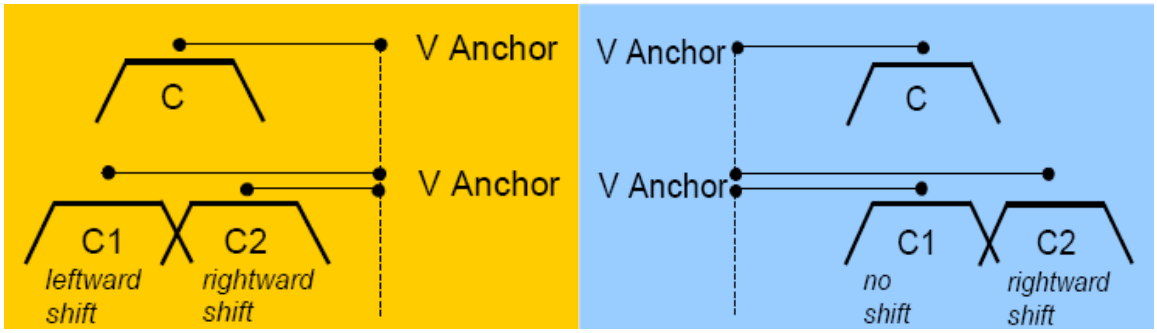

Figure 1. C1 leftward and C2 rightward shifts of the Polish onset (orange graph) and coda (blue graph) consonant clusters.

1.1 Extension of the Context Sequence Model

In this present study, articulatory gestures are investigated in the framework of Exemplar Theory, and are depicted with the help of EMA recordings as articulatory habits of individual speakers. In an exemplar theoretic model of the speech production-perception loop, all feedback during production, including articulatory habits of speakers, is assumed to be stored in detail in the memory providing a basis for future productions. We extended the CSM such that it exploits both articulatory and acoustic information.

In addition to the incorporation of articulatory speech representations we also incorporate multiple speakers in the model's memory. Alluding to Johnson's (1997) simulation study, we assume that speech items are not normalized but stored as detailed exemplars in memory. Therefore disambiguation of candidate exemplars necessarily involves the distinction between stored ego-exemplars and exemplars originating from other speakers. In a full-fledged exemplar-theoretic model of speech production and perception it could be assumed that each exemplar is associated with a multitude of labels, including the speaker identity or gender. In our current implementation of the CSM, however, we assume a flat structure omitting all higher levels of abstraction and semantic associations in the memory representations.

2 Method

We use data from the investigation on Polish consonant clusters (Mücke et al. 2010; Bruni 2011). Three adult native speakers were recorded (two female one male) with a Carstens AG100 2D Electromagnetic Articulograph, 10 channels. The corpus comprises a set of target words which are embedded in identical carrier phrases spoken with and without emphasis on the target word (this splits the corpus into two parts which are labeled "emph" and "noemph" below). We refer to each such carrier phrase as an "utterance" in the remainder of the text. In total, we use 336 utterances from the "emph" part and 337 utterance from the "noemph" part of the corpus. Only the target onset and coda syllables from the target words are labeled and only these are treated as production targets.

The production experiment was implemented such that it uses the continuous acoustic and articulatory data. Apart from the segmentation, we do not employ any discrete information (such as articulatory features or cues extracted at specific landmark positions from the signal—cf. Wade & Möbius, 2007). Instead, the continuous acoustic and articulatory signals are both processed in the same way. We simulate the production of one target utterance, by taking that target out of the corpus and using the remaining corpus data as the model's memory of remembered exemplars. The simulation iterates over all utterances of one speaker

and takes a sequence of phonetic segments $T = [t_1 ... t_n]$ 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 the output for each target utterance by copying 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, and compared against the left contexts of the candidate exemplars available in the memory. This procedure is repeated for each speaker, once on the "emph" part of the corpus and once on the "noemph" part. For each speaker, three memory structures are generated: first, using only the current speaker's data from the corpus. Additionally, data from each of the other two speakers is added in turn to the memory. Such that the memory of one simulation run consist of either the single target speaker's data or a combination of the current target speaker's data and the data of another speaker. We do not mix data from the "emph" and the "noemph" parts. As candidate segments then, we allow all segments from the memory irrespective of their corresponding speaker or their segment label. Only the current production target is excluded to avoid selection of the original segment from memory for production.

2.1 Evaluation

The simulation's performance is evaluated by computing the "segment accuracy" as well as the "segment context accuracy". We take the manually created annotation of the corpus as the reference against which the model's outputs are evaluated. Note that the labels are not used in the production process. The context accuracy is defined as the proportion of segments which have been selected for production from an identical context in the memory. The context, in this sense, is defined as the labels of the segments 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…] target context, its right context is counted as correct, while its left context is counted as wrong. The segment accuracy, on the other hand, considers only the segments' labels and compares the original label of the selected exemplar with the target label.

The performance of the implemented model is compared in total on three data type conditions: (1) using only acoustic data according to the original implementation of the CSM, (2) using articulatory data from the EMA recordings and (3) using a combination of acoustic and articulatory data.

3 Results

The results are shown in tables 1—6 for the two female speakers F1 and F2, the male speaker M and the combined memory representations. Tables 1 and 2 show the segment accuracy on the "emph" part of the corpus and tables 3 and 4 show the accuracy for the "noemph" part. Tables 5 and 6 show how many segments have been selected for production from the non-target speaker's speech exemplars.

The results allow the interpretation that using a combined corpus of EMA and the acoustic signals can 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' contexts can be seen by comparing the acoustic-based results with the results for the combined data. Moreover, for two of our three speakers, the results based on EMA signals are better as compared to the acoustic data.

The implementation using continuous acoustic and articulatory signals is based on the assumption that data for both modalities is stored in parallel and processed in a similar way. We conclude that both articulatory and acoustic information facilitates exemplar candidate selection in exemplar theoretic speech production models with context-sensitive representations.

Speaker	articulatory	acoustic	combined
F1	0.648	0.491	0.498
F1 × M	0.648	0.488	0.495
F1 × F2	0.644	0.491	0.498
M	0.819	0.633	0.633
M × F1	0.794	0.623	0.626
M × F2	0.819	0.616	0.616
F2	0.868	0.604	0.604
F2 × F1	0.864	0.604	0.604
F2 × M	0.868	0.604	0.604

Table 1. Segment accuracy on the "emph" part of the corpus. The first speaker label designates the target speaker, and the second label the added speaker data in the mixed memory cases.

Speaker	articulatory	acoustic	combined
F1	0.537	0.381	0.391
F1 × M	0.537	0.370	0.381
F1 × F2	0.537	0.381	0.388
M	0.694	0.520	0.520
M × F1	0.680	0.509	0.512
M × F2	0.694	0.505	0.505
F2	0.779	0.496	0.496
F2 × F1	0.779	0.496	0.496
F2 × M	0.779	0.496	0.496

Table 2. Segment context accuracy on the "emph" part of the corpus.

Speaker	articulatory	acoustic	combined
F1	0.580	0.580	0.587
F1 × M	0.580	0.583	0.590
F1 × F2	0.580	0.565	0.572
M	0.796	0.629	0.629
M × F1	0.786	0.625	0.625
M × F2	0.807	0.607	0.607
F2	0.839	0.596	0.596
F2 × F1	0.839	0.596	0.596
F2 × M	0.839	0.596	0.596

Table 3. Segment accuracy on the "noemph" part of the corpus.

Speaker	articulatory	acoustic	combined
F1	0.527	0.491	0.498
F1 × M	0.527	0.495	0.502
F1 × F2	0.527	0.470	0.477
M	0.682	0.518	0.518
M × F1	0.675	0.507	0.507
M × F2	0.689	0.507	0.507
F2	0.771	0.500	0.500
F2 × F1	0.771	0.500	0.500
F2 × M	0.771	0.500	0.500

Table 4. Segment context accuracy on the "noemph" part of the corpus.

Speaker	articulatory	acoustic	combined
F1 × M	1	9	8
F1 × F2	4	3	3
M × F1	9	13	12
M × F2	10	9	9
F2 × F1	1	0	0
F2 × M	0	0	0

Table 5. Number of segments selected from the non-target speaker speech data in the memory, on the "emph" part of the corpus.

Speaker	articulatory	acoustic	combined
F1 × M	2	4	4
F1 × F2	0	14	14
M × F1	1	16	16
M × F2	5	6	6
F2 × F1	0	1	1
F2 × M	0	0	0

Table 6. Number of segments selected from the non-target speaker speech data in the memory, on the "noemph" part of the corpus.

The above accuracy values need to be compared against the baseline which we define as a random selection of one exemplar from the set of available candidate exemplars in the corpus. The values are computed for each configuration separately, but the numbers are very similar due to the nearly identical structures of the corpora. The baseline segment accuracy is around 0.145. For the context accuracy the baseline is around 0.066.

4 Conclusion

The results show that a speech production model that operates on a quasi-continuous speech representation can exploit articulatory information and process it in the same way that acoustic information is analyzed. As indicated, for example, by Johnson's (1997) exemplar theoretic simulation studies, explicit speaker normalization might not be necessary. Stored speech items can retain full details including speaker specific properties. To our knowledge,

this simulation study is the first to investigate such detailed speech representations on a multi-modal memory representation which incorporates both acoustic and articulatory information. The model successfully selects exemplars from the target speaker without explicit labels. Candidate exemplars are specified in context based only on a similarity score which takes into account acoustic and articulatory information. Moreover, the fact that the segment accuracy is well above the baseline shows that segments can be specified by their context.

The above described simulation provides an empirical proof of Context Specification during speech production (Dogil 2010) where exemplar-cloud formation stems from complex processes of signal analysis – starting from the peripheral auditory system analysis of important acoustic landmarks, going through lexicon syllabary formation and finishing by internal analysis-by-synthesis.

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.

References

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.

Bruni, J. (2011). Sonorant voicing specification in phonetic, phonological and articulatory context. Dissertation. Universität Stuttgart.

http://elib.uni-stuttgart.de/opus/volltexte/2011/6311

Dogil, G. (2010). Hard-wired phonology: limits and latitude of phonological variation in pathological speech. In: C. Fougeron, B. Kühnert, M. D'Imperio, N. Valleè (eds.): Laboratory Phonology, Vol. 10, p. 343-380. Mouton de Gruyer. Berlin, NY.

Hermes, A., Grice, M., Mücke, D., and Niemann, H. (2008). Articulatory indicators of syllable affiliation in word initial consonant clusters in Italian. In: R. Sock, S. Fuchs, and Y. Laprie (Eds.), 8th International Seminar on Speech Production (ISSP).

Johnson, K. (1997). Speech Perception without Speaker Normalization: An Exemplar Model. In: Johnson, K., and Mullennix, J., Eds.: Talker Variability in Speech Processing. Academic Press, pp. 145–165.

Mücke, D., Sieczkowska, J., Niemann, H., Grice, M., Dogil, G. (2010). Sonority profiles, gestural coordination and phonological licensing: obstruent-sonorant clusters in Polish. Poster Presentation during LabPhon Conference 2010, Albuquerque, New Mexico.

Nam, H., Goldstein, L., and Saltzman, E. (2009). Self organization of syllable structure: a coupled oscillator model. In: F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé (Eds.), Approaches to phonological complexity. De Gruyter Mouton, pp. 299–328.

Pierrehumbert, J. (2001): Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J., Hopper, P. (eds.), Frequency and the emergence of Linguistic Structure. Benjamins, Amsterdam, p. 137-157.

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

Wade, T., Dogil, G., Schütze, H., Walsh, M., Möbius, B. (2010): Syllable frequency effects in a context-sensitive segment production model. Journal of Phonetics, Vol. 38, p. 227-239.