USING UNLABELED EMA DATA IN A SPEECH PRODUCTION MODEL WITH A RICH MEMORY

Daniel Duran, Jagoda Bruni, Hinrich Schütze and Grzegorz Dogil
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
<firstname.lastname>@ims.uni-stuttgart.de

ABSTRACT

We present a pilot study which integrates articulatory information into the Context Sequence Model (CSM) of speech production [1]. The CSM is an exemplar-theoretic model which builds on the concept of the speech perception—production loop and incorporates a rich acoustic memory of past speech items which are stored sequentially in their original context. In the present study, we enrich the original acoustic memory of the CSM with articulatory information by using continuous Electromagnetic Midsaggital Articulography (EMA) measurements. To our knowledge, there are no existing speech production models which use the full continuous EMA signals directly and in the same way as acoustic speech signals. In a first series of experiments, we used data from a Polish corpus [2] designed to investigate the coordination between articulatory gestures within syllables in onset and coda positions (particularly the so-called C-Center effect — a distance of the consonants in a cluster with regards to a vowel [3]). The corpus is composed of a set of repeated target words with simple onsets and codas containing single sonorants, as well as onset and coda clusters containing a voiceless stop and a sonorant, embedded into carrier phrases which guarantee identical contexts of tongue movements for all target consonants and clusters. Their structure is as follows (target words are underlined): onset: "*Ona mówi pranie aktualnie*" ("She is saying laundry currently"); coda: "*Ona powiedziała Cypr aktualnie*" ("She is saying Cyprus currently"). The database contained speech recordings and EMA measurements from one male and two female native speakers, recorded with a 2D Electromagnetic Articulograph, Carstens AG100. The EMA data was sampled at 400 Hz, post-processed, and manually annotated with phone segments and articulatory landmarks using the EMU Speech Database System [http://emu.sourceforge.net]. The target words were recorded with an emphasis and a non-emphasis articulation mode. We selected the phonetically labeled (C)CV and VC(C) sequences from approximately 670 target words for our production simulation. Our implementation of the simulation reproduces the original CSM (see [1] for a detailed description) with two important modifications: First, we run the simulations on three different conditions: (i) using only the acoustic speech recordings according to the original CSM, (ii) using the continuous EMA signals instead of the acoustic data, and (iii) using a multidimensional combined representation of both acoustic and EMA data. Second, we consider only the left context for the context matching procedure. When selecting an item for production from a set (or "cloud") of candidate exemplars, the CSM uses a left and a right context, comparing the candidates' original contexts with the context of the currently produced utterance. The left context stretches into the past and contains the acoustic signal (in our case also the EMA signals) of what was produced preceding the current target segment, whereas the right context estimates what is going to be produced in the future and contains the linguistic information (the phone labels) of what should be produced next (or what was originally produced after the respective candidate exemplars).

Our results indicate that it might be possible to incorporate articulatory information into speech perception—production models using raw EMA data (without having to manually label specific articulatory landmarks). This also allows using unlabeled EMA traces in acquisition models without having to justify the employment of an a priori defined set of discrete gestural features or landmarks. We are currently in the process of repeating the initial simulations using a database of English speech and EMA recordings (MOCHA-TIMIT corpus).

REFERENCES

[1] 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.

[2] 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 at LabPhon Conference 2010, Albuquerque, New Mexico.

[3] Browman, C., Goldstein, L. (1988). Some Notes on Syllable Structure in Articulatory Phonology. *Phonetica* 45, p.140-155.