# Speech events are recoverable from unlabeled articulatory data: Using an unsupervised clustering approach on data obtained from Electromagnetic Midsaggital Articulography (EMA)

*Daniel Duran* [1], *Jagoda Bruni* [1]
*Grzegorz Dogil* [1], *Hinrich Schütze* [2]

Institute for Natural Language Processing, University of Stuttgart, Germany
[1] <firstname.lastname>@ims.uni-stuttgart.de, [2] hs999@ifnlp.org

## Abstract

Some models of speech perception/production and language acquisition make use of a quasi-continuous representation of the acoustic speech signal. We investigate whether such models could potentially profit from incorporating articulatory information in an analogous fashion. In particular, we investigate how articulatory information represented by EMA measurements can influence unsupervised phonetic speech categorization. By incorporation of the acoustic signal and non-synthetic, raw articulatory data, we present first results of a clustering procedure, which is similarly applied in numerous language acquisition and speech perception models. It is observed that non-labeled articulatory data, i.e. without previously assumed landmarks, perform fine clustering results. A more effective clustering outcome for plosives than for vowels seems to support the motor view of speech perception.

**Index Terms**: Speech production/perception, modeling, clustering, EMA.

## 1. Introduction

Speech is continuous and highly ambiguous and variable, within and across speakers. In first language acquisition, we face the task of recognizing patterns within the speech stream and partitioning it into (linguistically relevant) units. One of the first steps towards the acquisition of linguistic categories and units is thus segmentation of the continuous speech stream and identification of individual speech events or items. However, no two utterances are identical. Therefore, some method is needed to group speech events according to their similarity into distinct groups. Speech events, in that sense, may be any linguistic units like phrases, words, syllables, phones etc. While lots of computational models of language acquisition and speech segmentation are based on a symbolic speech representation consisting of a sequence of discrete units, e.g. phonetic or syllabic segments [1], there are some models which are based directly on quasi-continuous representations of the acoustic signal, e.g. spectral or mel-frequency cepstral coefficient (MFCC) vectors. These models address the question of how linguistic categories can be acquired without using 'top-down' information (like pragmatic, semantic, lexical or phonemic knowledge generating expectations or guiding the perception on ambiguous and highly variable input). Learning linguistic categories like phonemes or syllables in an essentially unsupervised fashion based exclusively on the raw speech signal is a hard, if not impossible to solve, problem. Despite large acoustic inter- and intra-speaker variability, more or less similar sounds have to be identified and extracted from the continuous speech stream. As linguistic knowledge is usually assumed to be categorical [2, 3] a clustering technique is applied in some computational experiments, to group similar acoustic samples or patterns together [4, 5, 6]. The goal is to group percepts such that similarity within a cluster is maximized while similarity between different clusters is minimized. These clusters could then form a first step towards categorization of short stretches of speech. Clustering, in general, is considered a fundamental cognitive strategy for speech segmentation in language acquisition [7] and is also applied in symbolic approaches [8]. However, while there may be no purely acoustic cues that would group a sequence of, for example, three events like: *silence, plosion, release* into one phone, the fact that all three are caused by one continuous articulatory gesture will bias any learning algorithm to view them as a single phone. Some computational speech perception/production models incorporate articulatory information such that they use an articulatory speech synthesizer in order to simulate somatosensory feedback information [6, 9, 10].

From the perspective of Motor Theory of Speech Perception [11], units of speech are stored as articulatory patterns and motor commands, providing a basis for using different phonetic categories during speech production. Thus, speech perception is a process of conversion from acoustic signal to gestural comprehension. It is claimed [11] that differences between phonetic categories during stop perception (as in [ba] and [pa]) can be perceived without certain cues, for example voicing. Instead, phonetic category definition comprises significance of articulatory gestures and is more than just acoustic encoding.

In his speech perception model, [3] introduced a notion of *landmarks*, i.e. regions in a signal which contain acoustic evidence for particular phonetic features. It is posited that configuration of the articulatory movements exhibits categorical acoustic effects. The relation between the acoustic properties and the articulatory displacement shows either minimum/maximum values or is discontinuous. Thus, the occurrence of spectral discontinuity, particular frequency amplitude or measurement of its abruptness provides a landmark for phonetic features. Similarly, [12] 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 also said to enclose parameter values extorted from the speech signal. Thus, according to [12], speech perception relies on the activation of the perceived landmarks along with the successful context-matching.

In the view of articulatory phonology [13, 14, 15] gestures are basic units of speech production. These dynamic articula-

tory actions contain specified parameters correlating with vocal tract settings (including lips, tongue, glottis, velum etc.), which occur sequentially or undergo overlapping during the course of speech. Analysis of articulatory aspects of speech deals also with its temporal organization within syllable structure. [16] proposed an intrinsic model of syllable coordination, where the 'in-phase mode' generates 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). The authors demonstrated competitive articulatory patterns of complex CCV English onsets, the so called C-Center Effect. This correlation is obtained by measuring the interval between the mean value of the onset consonantal targets and the vowel, where the consonants maintain a stable distance with regards to the vowel target. Furthermore, it is claimed that in the VCC syllables the first consonant gesture is related to the gesture of a vowel target, exhibiting local organization of coordination. In analogue studies conducted on Italian [17] and Polish [18], C-Center like coordination has been demonstrated in CV and CCV clusters, with no such correlation in their Polish coda counterparts.

Recording articulatory movements with an Electromagnetic Midsagittal Articulograph (EMA) is a labor-intensive task and, so far, EMA data has been used primarily to extract a discrete set of features or cues at specific landmark positions (for example C-Centers [19]). These features are obtained from manual annotations of the signal, which is not only an additional very labor-intensive task, but it is also inherently subjective and depends on a more or less arbitrarily fixed annotation scheme.

In the experiments presented in this paper, we address the question whether raw EMA signals could be used analogously to the usage of acoustic data in computational models of speech perception, without having to manually label articulatory gestures or landmarks. We investigate whether computational models of speech perception and production which are based on auditory representations could benefit from the incorporation of *real* articulatory information. For this purpose, we combine the acoustic speech signal with raw EMA signals and apply a clustering procedure on the speech data.

## 2. Method

We use data from a corpus that was constructed to investigate the C-Center effect in Polish [18]. The acoustic and articulatory signals were obtained using a 2D Electromagnetic Articulograph, Carstens AG100, 10 channels. Three Polish native speakers (2 female, 1 male) participated in the study. Sensors were placed on the vermillion border of the upper and lower lip and on the tongue (3 sensors: 1cm, 3cm and 4cm behind the tongue tip). Coronal sounds, vowel articulation and velar consonants were analyzed with the sensor placed on the tongue tip and two sensors attached to the dorsum. 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 stored in Simple Signal File Format (SSFF), and manually labeled in EMU Speech Database System[1]. 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: (1) onset position: *"Ona mówi pranie aktualnie"* (*She is saying laundry currently*), (2) coda position: *"Ona powiedziała Cypr aktualnie"* (*She said Cyprus currently*). The underlined target word was recorded with an emphasis mode of articulation. For the list of words, see table 1.

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 |

The EMA signals were combined to form a sequence of 4-dimensional, length normalized vectors (3 signals from the tongue and one for lip aperture). Following [20], the acoustic data in our experiments was converted to an 8-dimensional representation with amplitude envelopes from 8 logarithmically spaced frequency bands, sampled at 250 Hz for computational efficiency. This corpus allowed us to compare 3 different data types for the same speech material: (1) using only the 4-dimensional EMA data ('EMA'), (2) using only the 8-dimensional envelope representation of the acoustic data ('ENV') and (3) using a combined 12-dimensional representation of EMA and acoustic signal ('EMA+ENV').

We selected a total of 336 ($3 \times 112$) utterances for our experiment. We do not use the manually created annotation. However, since we evaluate against the given reference annotation, we can use merely the labeled parts of the corpus. Only the consonants and consonant clusters in the target words along with their corresponding vowel were labeled at the phone level. Because of this selective annotation, the different phone classes' proportions are very unbalanced in our speech corpus. For speaker 1, for example, 21% of all data frames are labeled [a], while only 2% of the data are [u] frames. In order to avoid any negative effects caused by this unequal distribution, we applied a random sampling procedure to create corpus samples with equally-sized phone classes per speaker. To further account for effects caused by the random selection of frames from the original data, we repeated this sampling procedure 10 times for every speaker and every data type (EMA, ENV and EMA+ENV). Thus, we created a total of 90 data sets or corpus samples (3 speakers $\times$ 3 data types $\times$ 10 random samples). We apply a clustering procedure separately for all corpus samples using bisecting k-means [21].

## 3. Evaluation and results

We use *purity* (P) [22] to evaluate the quality of the clusters. To compute P, each cluster $\omega_l$ is assigned to the reference phone label class $c_i$ whose frames are most frequent in $\omega_l$, and then the accuracy of this assignment is measured by counting the number of correctly assigned frames and dividing by the total number of frames $N$:

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_l \max_i |\omega_l \cap c_i| \qquad (1)$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $C = \{c_1, c_2, \ldots, c_J\}$ is the set of reference phones labels or classes. We interpret $\omega_k$ as the set of frames assigned to $\omega_k$ and $c_j$ as

[1] Available online at: `emu.sourceforge.net`

Table 2: Purity of clustering per speaker

| speaker | ENV | EMA | EMA+ENV |
|---|---|---|---|
| 1 | 0.44 | 0.42 | 0.44 |
| 2 | 0.49 | 0.54 | 0.48 |
| 3 | 0.48 | 0.54 | 0.48 |

Table 3: Unmatched phone classes per data type (speakers 1-3)

| | k | p | r | l | u | y | i | a |
|---|---|---|---|---|---|---|---|---|
| EMA | 3 | 3 | 0 | 6 | 0 | 2 | 3 | 17 |
| ENV | 3 | 7 | 6 | 7 | 10 | 5 | 0 | 0 |
| EMA+ENV | 4 | 3 | 8 | 6 | 11 | 6 | 0 | 0 |

the set of frames labeled with label $c_j$ in equation 1. A perfect clustering has purity 1, while bad clusterings have purity values close to 0. Purity is an intuitive measure of cluster quality, but high purity is easy to achieve when the number of clusters is large – in particular, purity is 1 if each frame gets its own cluster. The number of classes in our experiments corresponds to the number of reference labels and depends on the decision, which labels should be used for evaluation. Note also, that annotation of continuous data like a speech signal with a discrete set of labels always contains some more or less arbitrary decisions, like, for example, labeling a plosive with a single label or labeling the silence and release parts of the phone with separate labels etc. We evaluate P for the case where $C = \{$p, r, a, l, k, i, y[2], u$\}$ corresponds to the set of phone labels and $|\Omega| = |C| = 8$. Table 2 shows P averaged over the 10-fold sampling procedure with equally-sized phone classes per speaker.

Tables 4, 5 and 6 show confusion count matrices for ENV, EMA and EMA+ENV data, respectively (i.e. they show how often a particular data frame of a cluster was actually labeled with a phone label). The results were obtained by summing the confusion counts for all 3 speakers and all 10 repeated clusterings. The table rows correspond to the clusters and the columns to the phones. They are labeled according to the same criterion used for the computation of P: each cluster is assigned to the phone class which is most frequent in that cluster. We will refer to this as a phone class being 'matched by a cluster majority'. An ideal clustering would assign all frames of one class to one separate cluster resulting in a diagonal confusion matrix. Figure 1 shows a visualization of the confusion matrices of tables 4, 5 and 6. The horizontal barplots show the relative distribution of phone classes for a given cluster. The three data type conditions are shown side-by-side, with dark grey corresponding to EMA, medium grey corresponding to EMA+ENV and light grey corresponding to ENV data. Note, that while on average all phone classes are matched by such a cluster majority (i.e. the highest numbers per row are on the diagonal and there are no completely empty rows in the tables), this is not always the case for each individual run on a corpus sample. Table 3 shows a summary of how often each phone was not matched by a cluster majority. The most often unmatched phone classes were: [a] on EMA, [u] on ENV and EMA+ENV data.

The phones can be divided into 3 groups: (1) plosives (phones [p] and [k]), (2) sonorants ([r] and [l]), and (3) vowels ([u], [ɨ], [i] and [a]). A direct comparison of the different data types in figure 1 reveals that clustering EMA data is more successful for plosives than for vowels (as indicated by the higher peaks). Clustering acoustic data, on the other hand, is more successful for vowels (with some confusion between [u] and [ɨ]). The flatter distributions of ENV and EMA+ENV indicate that clustering generally performs better on EMA data than on amplitude envelopes (if evaluated against the phonetic segments as reference classes). This tendency can also be seen by the relatively higher peaks in the diagonal for EMA data. However, only the extremes of the confusion matrix show very clear results: EMA does clearly provide enough information to identify [p] and [k], while ENV provides comparably more information to identify vowels, especially [a]. The results in table 3 show that while [a] is frequently not matched by a cluster majority, in total there is slightly more confusion on ENV data (as indicated by the higher total number of unmatched phone classes).

[2]The (orthographically motivated) label 'y' is used for the Polish central high vowel [ɨ].

Table 4: Cluster×phone confusions; speakers 1-3: ENV

| | k | p | r | l | u | y | i | a |
|---|---|---|---|---|---|---|---|---|
| c:k | 1153 | 898 | 489 | 162 | 165 | 112 | 171 | 33 |
| c:p | 781 | 974 | 365 | 114 | 84 | 121 | 118 | 9 |
| c:r | 492 | 319 | 1103 | 324 | 85 | 435 | 132 | 91 |
| c:l | 158 | 223 | 372 | 1269 | 322 | 240 | 401 | 77 |
| c:u | 70 | 145 | 283 | 581 | 1332 | 650 | 297 | 123 |
| c:y | 122 | 239 | 306 | 310 | 881 | 1726 | 34 | 257 |
| c:i | 490 | 479 | 311 | 509 | 431 | 16 | 2146 | 31 |
| c:a | 34 | 23 | 71 | 31 | 0 | 0 | 1 | 2679 |

Table 5: Cluster×phone confusions; speakers 1-3: EMA

| | k | p | r | l | u | y | i | a |
|---|---|---|---|---|---|---|---|---|
| c:k | 1750 | 24 | 47 | 15 | 0 | 280 | 469 | 189 |
| c:p | 4 | 1436 | 16 | 0 | 399 | 85 | 19 | 248 |
| c:r | 81 | 211 | 1890 | 1191 | 1 | 125 | 76 | 736 |
| c:l | 187 | 12 | 436 | 1314 | 0 | 9 | 96 | 336 |
| c:u | 48 | 1234 | 219 | 241 | 2596 | 591 | 116 | 363 |
| c:y | 222 | 196 | 415 | 423 | 21 | 1768 | 458 | 492 |
| c:i | 917 | 16 | 147 | 69 | 0 | 347 | 1952 | 389 |
| c:a | 91 | 171 | 130 | 47 | 283 | 95 | 114 | 547 |

## 4. Conclusions

We found indications that with respect to purity, using EMA signals can improve the results of a clustering procedure. Clustering experiments show better results for EMA data in case of plosives rather than vowels. This might serve as evidence for a motor speech perception/production approach [11] according

Table 6: Cluster×phone confusions; speakers 1-3: EMA+ENV

| | k | p | r | l | u | y | i | a |
|---|---|---|---|---|---|---|---|---|
| c:k | 828 | 629 | 384 | 130 | 121 | 115 | 119 | 29 |
| c:p | 1128 | 1364 | 547 | 185 | 173 | 160 | 153 | 6 |
| c:r | 462 | 244 | 993 | 312 | 48 | 398 | 112 | 82 |
| c:l | 148 | 175 | 351 | 1313 | 335 | 215 | 396 | 68 |
| c:u | 81 | 168 | 323 | 532 | 1310 | 725 | 291 | 134 |
| c:y | 142 | 241 | 312 | 297 | 927 | 1664 | 48 | 296 |
| c:i | 479 | 456 | 289 | 499 | 386 | 23 | 2180 | 20 |
| c:a | 32 | 23 | 101 | 32 | 0 | 0 | 1 | 2665 |

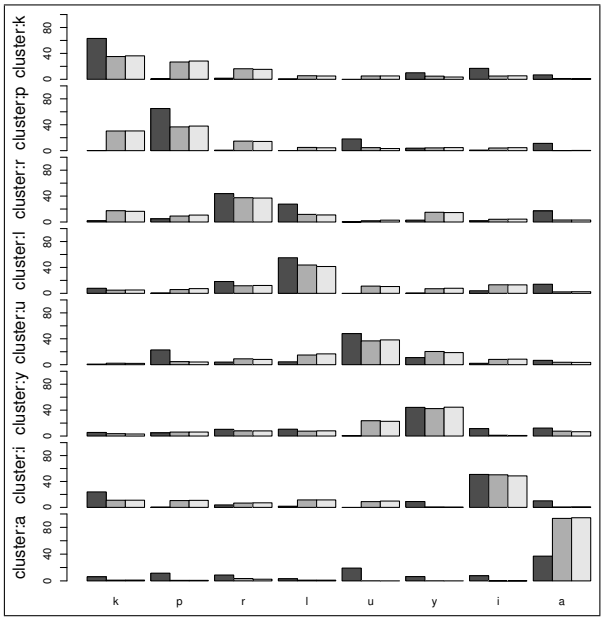


Figure 1: *Relative cluster×phone confusions. Dark grey: EMA; medium grey: EMA+ENV; light grey: ENV.*

to which plosives, involving usage of a variety of articulators, are more distinct than vowels, which production is more gradual in terms of articulatory settings. Overall, it seems possible to model speech perception without taking all phonetically distinct cues into account. Our study shows that acoustic landmarks embedded in spectral information need not be always accounted for in speech modeling.

Articulatory information as represented by raw EMA signals provides a useful resource for unsupervised categorization of speech samples. Excluding discrete landmarks supports computational models of speech perception using a rich memory representation where every bit counts [20]. It also allows incorporation of articulatory information into an acquisition model without having to justify the employment of a particular set of discrete gestural features or landmarks (and, without having to manually extract these data from EMA measurements). That way, EMA data can be incorporated analogously to acoustic signals and processed using the same methods.

Although we implemented an unsupervised clustering of the corpus data, our experiments depended on manually created annotations in order to evaluate them. This reduced the amount of available data drastically, even though the full acoustic and EMA signals for all utterances were available. We think that this fact additionally supports our claim – that it might be advantageous to use unlabeled articulatory data.

## 5. Acknowledgments

This study is part of the research project SFB 732 A2, funded by the German National Science Foundation (DFG).

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

## 6. References

[1] M. R. Brent, "Speech segmentation and word discovery: a computational perspective," *Trends in Cognitive Sciences*, vol. 3, no. 8, pp. 294–301, 1999.

[2] F. Cooper, A. Liberman, J. Borst, and L. Gerstman, "Some experiments on the perception of synthetic speech sounds," *JASA*, vol. 24, pp. 597–606, 1952.

[3] K. N. Stevens, "Features in speech perception and lexical access," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds., 2005, pp. 125–155.

[4] K. Gold and B. Scassellati, "Audio speech segmentation without language-specific knowledge," in *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 2006.

[5] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1084–1095, 2010.

[6] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, pp. 85–117, 2011.

[7] J. V. Goodsitt, J. L. Morgan, and P. K. Kuhl, "Perceptual strategies in prelingual speech segmentation," *Journal of Child Language*, vol. 20, pp. 229–252, 1993.

[8] D. Swingley, "Statistical clustering and the contents of the infant vocabulary," *Cognitive Psychology*, vol. 50, pp. 86–132, 2005.

[9] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain and Language*, vol. 96, pp. 208–301, 2006.

[10] B. J. Kröger, P. Birkholz, and A. Lowit, "Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production (ACT)," in *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2010, ch. 2, pp. 23–36.

[11] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[12] T. Wade and B. Möbius, "Speaking rate effects in a landmark-based phonetic exemplar model," in *Interspeech*, Aug 2007, pp. 402–405.

[13] C. Browman and L. Goldstein, "Towards an articulatory phonology," in *Phonology Yearbook*. Cambridge University Press, 1986, vol. 3, pp. 219–252.

[14] ——, "Articulatory gestures as phonological units," *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

[15] ——, "Articulatory phonology: An overview," Haskins Laboratories, Status Report on Speech Research SR-111/112, 1992.

[16] H. Nam, L. Golstein, and E. Saltzman, "Self organization of syllable structure: a coupled oscillator model," in *Approaches to phonological complexity*, F. Pellegrino, E. Marisco, and I. Chiotran, Eds., 2009, pp. 299–328.

[17] A. Hermes, M. Grice, D. Mücke, and H. Niemann, "Articulatory indicators of syllable affiliation in word initial consonant clusters in Italian," in *8th International Seminar on Speech Production (ISSP)*, R. Sock, S. Fuchs, and Y. Laprie, Eds., 2008, december 8th to 12th 2008, Strasbourg, France.

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

[19] C. Browman and L. Goldstein, "Some notes on syllable structure in articulatory phonology," *Phonetica*, vol. 45, pp. 140–155, 1988.

[20] T. Wade, G. Dogil, H. Schütze, M. Walsh, and B. Möbius, "Syllable frequency effects in a context-sensitive segment production model," *Journal of Phonetics*, vol. 38, no. 2, pp. 227–239, 2010.

[21] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000.

[22] A. Strehl, "Relationship-based clustering and cluster ensembles for high-dimensional data mining," Ph.D. dissertation, UTexas Austin, 2002.