SPEAKER IDIOSYNCRATIC INTENSITY AND MOUTH OPENING-CLOSING VARIATIONS: THE CASE OF ENGLISH

Yu Zhang, Lei He, Volker Dellwo

Department of Computational Linguistics, University of Zurich, Switzerland
{yu.zhang | lei.he | volker.dellwo}@uzh.ch

ABSTRACT

This study investigated speaker idiosyncrasy in intensity and mouth opening-closing variations using an English corpus containing both acoustic and articulatory data (19 speakers × 59 read sentences). The speeds of intensity as well as mouth opening-closing movements were calculated and summarized in terms of the mean, standard deviation, and pairwise variability index per sentence. Multinomial logistic regressions were used to test the speaker effect and evaluate the amount of between-speaker variability explained by each measure. It was found that all measures showed significant speaker effect. Moreover, the measures pertaining to the speeds of intensity and mouth opening-closing movements explained more between-speaker variability in English.

Keywords: speaker idiosyncrasy, intensity variability, mouth opening-closing variability, English.

1. INTRODUCTION

People differ in how they speak and what they sound like. The major processes in speech production – phonatory and articulatory activities – all contain speaker idiosyncratic information; such speaker idiosyncrasies leave traces in the speech signal and are thus measurable acoustically (see [1] for a general review). Salient between-speaker variabilities in various temporal acoustic features have been explained in terms of speaker idiosyncratic articulatory movements (e.g., [2, 3] for duration variabilities; [4, 5] for formant variabilities; [6, 7] for intensity variabilities,). This paper focuses on between-speaker variability in intensity variability and corroborates it from the articulatory movements of the mouth.

We followed the method for examining between-speaker intensity variability developed by He and Dellwo [7], who partitioned the intensity curve into rising and falling segments between peaks and troughs (peaks were identified between syllable boundaries; troughs were identified between peaks). The speeds of intensity increases and decreases for all rising and falling segments[1] were calculated using the Zurich German TEVOID corpus [2, 3]. They found that the speeds of intensity decreases explained more between-speaker variability than the speeds of intensity increases across an utterance [7]. Similar results have been obtained from a Thai corpus [8]. The authors suggested that speakers might differ more in mouth-closing movements [7, 8], as a high degree of covariation between intensity and mouth aperture size was observed [9]. Similar results were obtained from the same Zurich German corpus by calculating the first formant variability, another acoustic feature that covaries with mouth opening and closing movements [5].

For the current study, we aim to (1) investigate whether the speeds of intensity decreases also explain more between-speaker variability in English, and (2) whether the findings from intensity variations can be supported by the speeds of mouth opening-closing movements.

2. METHOD

2.1. Corpus

The sentence-reading part of the EMA-MAE corpus[2] [10] was used. Twenty native speakers of American English were enrolled; each participant read 59 sentences. Both articulatory and acoustic data were recorded simultaneously (The NDI Wave electromagnetic articulograph; $Fs_{\_NDI}$ = 400 Hz; $Fs_{\_acoust}$ = 22.05 kHz). After data cleaning, 19 speakers (9F, 10M) were kept with 1108 analyzable sentences in total.

2.2. Data Preprocessing

The intensity curve of each sentence was extracted using the Praat [11] function "Sound: To Intensity…" with default settings[3]. The intensity curve was then normalized such that the average intensity equated to 65 dB (SPL).

The mouth opening-closing movements were characterized using three NDI sensors at the upper lip (UL), lower lip (LL) and lip corner (LC), as indicated in Fig. 1. The area of the triangle formed

by these three sensors was used to estimate the magnitude of mouth opening (abbreviated as MO henceforth) at each instant in time based on the x, y and z coordinates of these sensors [15][4]. The resultant MO curve was smoothed by low-pass filtering at 10 Hz.

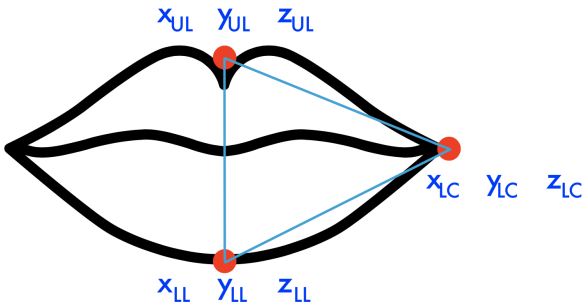


Figure 1: An illustration of estimating the mouth opening-closing movements from three NDI sensors at the upper lip, lower lip, and the lip corner. Endnote 4 shows the formula.

Syllable boundaries were placed using the BAS Web Services[5]. First, the WebMAUS Basic [12, 13] was used for phonemic transcriptions based on the acoustic signals and orthographic transcriptions. The annotations were saved in the BAS Partitur Format (*.par). The *.par files were then submitted to Pho2Syl [14] for the syllabification processing; syllable boundaries were saved in Praat TextGrids. The boundaries were cross-checked by the authors.

2.3. Speeds of change in intensity and mouth movements

For each sentence, the speeds of increase and decrease in intensity and mouth movement magnitude were calculated (i.e., the steepness of the arrows in red, purple and caramel in Fig. 2). For intensity changes, the peaks were pinpointed between syllable boundaries and the troughs were identified between consecutive peak points (see Fig. 2(a), the peak points were indicated by blue down arrows, and the trough point was indicated by a blue up arrow). The speeds of intensity increases from a trough point to its consecutive peak point (i.e., the steepness of the upward arrows in Fig. 2(a)) is notated as **v_I[+]**; the speeds of intensity decreases from a peak point to its neighboring trough point (i.e., the steepness of downward arrows in Fig. 2(a)) is notated as **v_I[–]**.

For the speeds in MO increases and decreases, two methods were followed: (1) the syllable-based method and (2) the derivative-based method.

Re (1) the syllable-based method: Similar to **v_I[+]** and **v_I[–]**, MO peaks were identified between syllable boundaries, and MO troughs were detected between peaks (see Fig. 2(b), the MO peaks were indicated by blue down arrows, and the trough point was indicated by an blue up arrow). The speeds of increases from an MO trough to its following peak (i.e., the steepness of the upward arrows in Fig. 2(b)) is notated as $\mathbf{v\_MO_{syl}[+]}$; the speed of MO decrease from a peak to the next trough (i.e., the steepness of the downward arrows in Fig. 2(b)) is notated as $\mathbf{v\_MO_{syl}[–]}$.

Re (2) the derivative-based method: The first derivative of the MO curve (abbreviated as derMO hereafter) was approximated in terms of the difference between the $(n+1)^{th}$ and the n^{th} samples. An MO peak was pinpointed where the derMO crossed zero with a downward trend; an MO trough was pinpointed where the derMO hit zero with an upward trend [16] (Fig. 2(d) indicates the peaks by blue down arrows, and a trough by an blue up arrow). This captured all local flexions in MO. The speed of local trough-to-peak increase (i.e., the steepness of the upward arrows in Fig. 2(d)) is notated as $\mathbf{v\_MO_{der}[+]}$; The speed of local peak-to-trough decrease (i.e., the steepness of the downward arrows in Fig. 2(d)) is notated as $\mathbf{v\_MO_{der}[–]}$.

MO speeds calculated using the first method focus on macroscopic MO movements within the frame of syllables, while MO speeds calculated using the second method focus on a finer scale of MO movements.

2.4. Variables calculated from the speeds of change in intensity and mouth movements

(a) Variables based on **v_I[+]** *and* **v_I[–]** Each sentence contains a number of **v_I[+]**s and **v_I[–]**s. Their central tendencies, dispersions and sequential variabilities were calculated in terms of the means, standard deviations, and pairwise variability indices[6] ad modum He and Dellwo [7]: mean_vI[+], stdev_vI[+] and pvi_vI[+], as well as mean_vI[–], stdev_vI[–] and pvi_vI[–].

(b) Variables based on $\mathbf{v\_MO_{syl}[+]}$ *and* $\mathbf{v\_MO_{syl}[–]}$ Similarly, each sentence contains a number of $\mathbf{v_MO_{syl}[+]}$s and $\mathbf{v_MO_{syl}[–]}$s. Their means, standard deviations, and pairwise variability indices were calculated: mean_vMO_{syl}[+], stdev_vMO_{syl}[+] and pvi_vMO_{syl}[+], as well as mean_vMO_{syl}[–], stdev_vMO_{syl}[–] and pvi_vMO_{syl}[–].

(c) Variables based on $\mathbf{v\_MO_{der}[+]}$ *and* $\mathbf{v\_MO_{der}[–]}$ In the same manner, the means, standard deviations and pairwise variability indices were calculated based on $\mathbf{v\_MO_{der}[+]}$ and $\mathbf{v\_MO_{der}[–]}$ for each sentence: mean_vMO_{der}[+], stdev_vMO_{der}[+] and pvi_vMO_{der}[+], as well as

mean_vMO$_{der}$[–], stdev_vMO$_{der}$[–] and pvi_vMO$_{der}$[–].

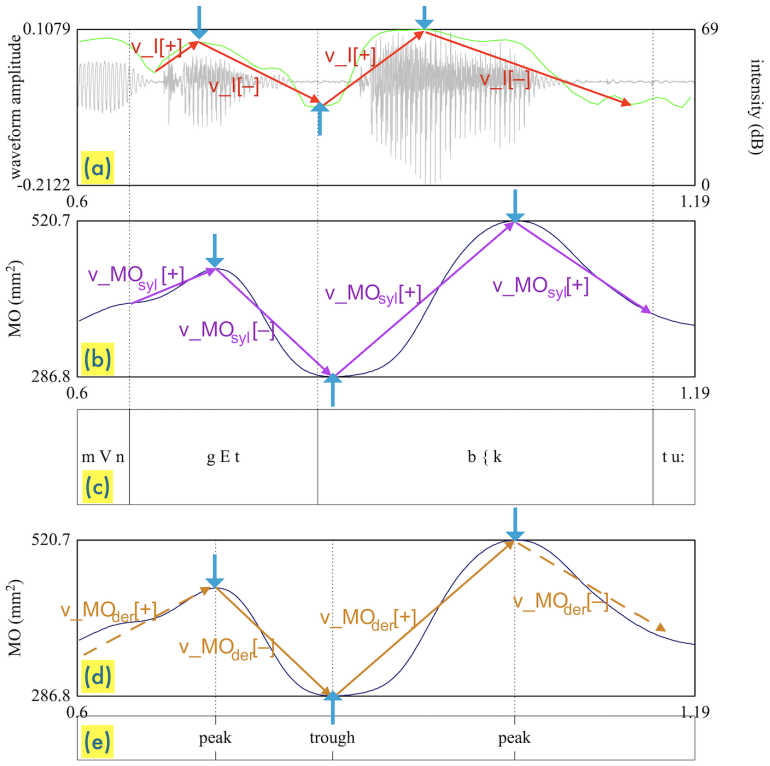


Figure 2: (a) An illustration of speeds of intensity increases and decreases; (b) an illustration of speeds of MO increases and decreases using the syllable-based method; (c) syllable boundaries with the SAMPA transcription ("get back"); (d) an illustration of speeds of MO increases and decreases using the derivative-based method; (e) MO peaks and troughs pinpointed from derMO.

2.5. Statistical analysis

JMP was used for statistical analysis. To quantify how between-speaker variability was differentially explained by the positive and negative variables in §2.4, the multinomial logistic regressions (MLRs) were fitted with *speaker* as the nominal response variable and the variables in §2.4 as the numeric predictor variables. *Sentence* effect was mitigated by z-score normalization. To bypass the collinearity issues, nine MLR models were separately fitted (M1 – M9 below).

For variables based on speeds of intensity increases and decreases: (M1) *speaker* ~ mean_vI[+] + mean_vI[–], (M2) *speaker* ~ stdev_vI[+] + stdev_vI[–], (M3) *speaker* ~ pvi_vI[+] + pvi_vI[–].

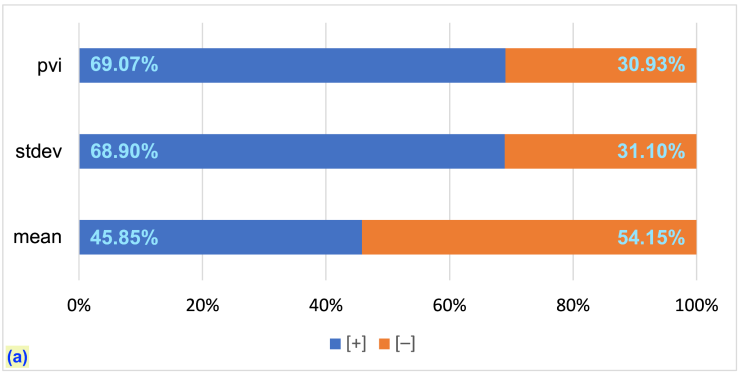
For variables based on speeds of MO increases and decreases calculated across syllable boundaries: (M4) *speaker* ~ mean_vMO$_{syl}$[+] + mean_vMO$_{syl}$[–], (M5) *speaker* ~ stdev_vMO$_{syl}$[+] + stdev_vMO$_{syl}$[–], (M6) *speaker* ~ pvi_vMO$_{syl}$[+] + pvi_vMO$_{syl}$[–].

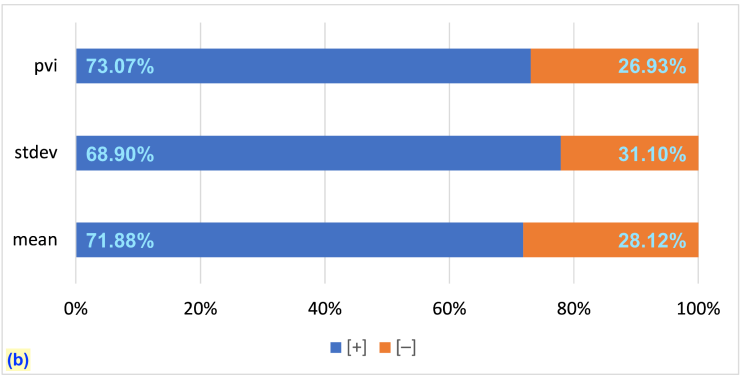
For variables based on speeds of MO increases and decreases calculated based on derivatives: (M7) *speaker* ~ mean_vMO$_{der}$[+] + mean_vMO$_{der}$[–], (M8) *speaker* ~ stdev_vMO$_{der}$[+] + stdev_vMO$_{der}$[–], (M9) *speaker* ~ pvi_vMO$_{der}$[+] + pvi_vMO$_{der}$[–].

For each model, the amount of between-speaker variability was calculated as $(\chi^2/\Sigma\chi^2)\times100\%$, where χ^2 refers to the likelihood ratio χ^2 of a particular variable, and $\Sigma\chi^2$ refers to the sum of likelihood ratio χ^2s of both variables in a model.

3. RESULTS

Due to the page constraint, the model fitting details and statistical results of the nine models are not tabulated in this paper; they are documented as a supplementary material accessible via https://osf.io/n7jcu. For each model, the variable based on the positive speed explained more between-speaker variability than its negative counterpart (Fig. 3 indicates the percentage of between-speaker variability explained by each variable per model).





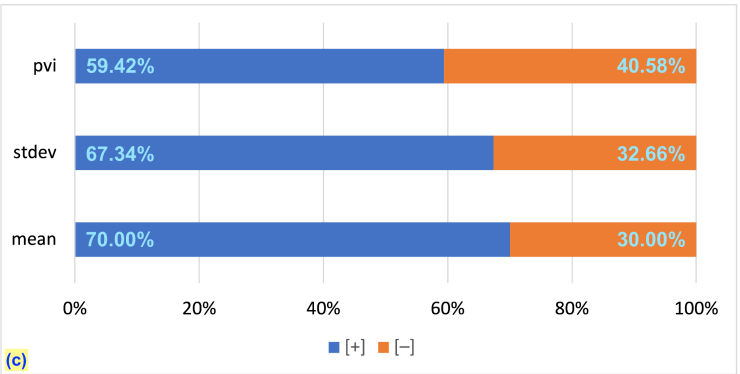


Figure 3: Between-speaker variability calculated from nine MLR models. (a) variables based on vI[+] and vI[−]; (b) variables based on vMO_{syl}[+] and vMO_{syl}[−]; (c) variables based on vMO_{der}[+] and vMO_{der}[−].

4. DISCUSSION

For native English speakers, measures based on speeds of intensity increases explained more between-speaker variability in general (except for the mean measure, although the difference is small-scale). The speeds of mouth opening and closing also differentially explain between-speaker variability in a similar manner: the measures based on the speeds of MO increases are more powerful in explaining speaker variability than the ones based on the speeds of MO decreases.

Congruency in the results obtained from both acoustic and articulatory data supports the attempt to investigate speaker idiosyncrasy in mouth opening-closing articulation via variabilities in intensity increases and decreases [7, 8]. This suggests that speaker idiosyncratic features can be shared across interconnected systems. In linguistics and speech communications, features shared between domains (visuo-articulatory, acoustic, neurological and gestual) are helpful, and sometimes crucial for comprehension [20, 21, 22].

In terms of the disparities in the speeds of increases and decreases for intensity and mouth opening-closing movements in encoding speaker-specific information, it is likely that during speech articulation, English speakers exhibited more freedom in the mouth opening phases, while in the mouth closing phases they coordinate their articulatory movements in a more controlled way. In a study on the reproduction of articulatory trajectories using a model-based method, it has already been demonstrated that the motor programs may be different in the opening and closing gestures [23].

It is also interesting to see that the English speakers exhibited the opposite pattern compared to Zurich German and Thai speakers [7, 8]. It is likely that the role of mouth opening-closing cycles is not universal across languages, and the way speaker-specificity is encoded in this dynamic process and its acoustic outcomes are also different in different languages. Using spontaneous speech, a number of studies have shown that the distribution of speaker-specific information in speeds related to the increases and decreases of intensity and F1 are more or less balanced in Dutch, English and German [24, 25, 26]. The differences might be attributed to the use of spontaneous speech, which is believed to be less controlled and thus displaying larger variations in articulatory patterns in general [27, 28].

Although different studies involving different languages and speaking styles exhibited different results in terms of the amount of speaker specificity explained by measures that are attributed by the speeds of mouth opening and closing movements, *speaker* effect remained significant regardless of the speed directions (see the model fitting results in the supplementary material https://osf.io/n7jcu) and the languages involved [5, 7, 8]. These findings have implications for forensic phonetics, ASR and other research fields that require the characterization of individual speakers. In forensic voice comparison closer attention should be paid to parts of the speech signals which are more informative about speaker identity in a particular language. In terms of ASR, it may be possible to achieve higher system performance if different parts of the acoustic signal are modeled differently. Of course more work needs to be done to gain better insights in our general understanding of the acoustic-articulatory dynamics.

5. ACKNOWLEDGEMENTS

We acknowledge support from the University of Zurich (Forschungskredit, Grant Nos. FK-19-069 (Y.Z.) and FK-20-078 (L.H.)), and the Swiss National Science Foundation (Grant No. PZ00P1_193328 (L.H.)).

6. REFERENCES

[1] Dellwo, V., French, P., He, L. 2018. Voice biometrics for speaker recognition applications. In: Frühholz, S., Belin, P. (eds), *The Oxford Handbook of Voice Perception*. Oxford University Press, 777–795.

[2] Leemann, A., Kolly, M.-J., Dellwo, V. 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Sci. Int'l* 238, 59–67.

[3] Dellwo, V., Leemann, A., Kolly, M.-J. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *J. Acoust. Soc. Am.* 137, 1513–1528.

[4] McDougall, K. 2006. Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies. *Int'l J. Speech, Lang. Law* 13, 89–126.

[5] He, L., Zhang, Y., Dellwo, V. 2019. Between-speaker variability and temporal organization of the first formant. *J. Acoust. Soc. Am.* 145, EL209–EL214.

[6] He, L., Dellwo, V. 2016. The role of syllable intensity in between-speaker rhythmic variability. *Int'l J. Speech, Lang. Law* 23, 243–273. .

[7] He, L., Dellwo, V. 2017. Between-speaker variability in temporal organizations of intensity contours. *J. Acoust. Soc. Am.* 141, EL488–EL494.
[8] Zhang, Y., He, L., Kerdpol, K., Dellwo, V. 2021. Between-speaker variability in intensity slopes: The case of Thai. Abstract presented at the XVII[th] AISV Conference (Zürich, 4–5 February 2021).
[9] Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A. A. 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, e1000436.
[10] Ji, A., Berry, J. J., Johnson, M. T. 2014. The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In: *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing* (Florence, 4–9 May, 2014), 7769–7773.
[11] Boersma, P., Weenink, D. 2020. Praat: Doing phonetics by computer (version 6.1.16) www.praat.org
[12] Schiel, F. 1999. Automatic phonetic transcription of non-prompted Speech. In *Proc. of the 14th ICPhS* (San Francisco, 1–7 August 1999), 607–610.
[13] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Comput. Speech Lang.* 45, 326–347.
[14] Reichel, U. D., Kisler, T. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In: Hoffmann, R. (ed), *Elektronische Sprachverarbeitung* (Studientexte zur Sprachkommunikation 71). TUDpress, 42–49.
[15] He, L. 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am.* 152, 567–579.
[16] Hass, J., Heil, C., Weir, M. 2019. *Thomas' Calculus* 14th edition, (SI units, international edition) Pearson Education.
[17] Low, E. L., Grabe, E., Nolan, F. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech* 43, 377–401.
[18] MacNeilage, P. F. 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499--546.
[19] Morrill, R. J., Paukner, A., Ferrari, P. F., Ghazanfar, A. A. 2012. Monkey lipsmacking develops like the human speech rhythm, *Develop. Sci.* 15, 557--568
[20] Strauss, A., and Schwartz, J.-L. 2017. The syllable in the light of motor skills and neural oscillations, *Lang. Cogn. Neurosci*. 32, 562--569.
[21] Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J, Kello, C., Schaefer, R.S., Wiggins, G.A. 2021. Multilevel rhythms in multimodal communication, *Phil. Trans. R. Soc*.B376: 20200334.
[22] Park, H., Kayser, C., Thu, G., Gross, J. 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility, *eLife* 5, e1452.
[23] Birkholz P., Kroger, B. J., and Neuschaefer-Rube, C. 2011. Model-based reproduction of articulatory trajectories for consonant-vowel sequences, *IEEE Trans. Audio, Speech Lang. Process*. 19, 1422--1433.
[24] Lins Machado, C. 2021. A cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech. In *Proc. of the* XVII AISV Conference (Zurich, 4-5 February 2021), 79-80
[25] He, L., Heeren, W. 2021. Between-speaker variability in dynamic formant characteristics in spontaneous speech. In Proc. of the XVII AISV Conference (Zurich, 4-5 February 2021), 71-72.
[26] Heeren, W., He, L. 202., Between-speaker variability in segmental F1 dynamics in spontaneous speech. In: Proc. of the 30th Annual Conference of the International Association for Forensic Phonetics and Acoustics (Marburg, 10-13 July 2021).
[27] De Nil, L.F., Abbs, J.H. 1991. Influence of speaking rate on the upper lip, lower lip and jaw peak velocity sequencing during bilabial closing movements. *J. Acoust. Soc. Am.* 89(2), 845-849.
[28] Illa, A., Ghosh, P.K. 2020. The impact of speaking rate on acoustic-to-articulatory inversion. In *Computer Speech & Language*, 59, 75-90.

---

[1] He and Dellwo [7] referred to the speeds of intensity increases and decreases as "positive and negative dynamics" of intensity. In the context of [7], "dynamic(s)" was used in its most general sense to mean "non-static" properties of the intensity curve. Here, we refrain from using the word "dynamics" to avoid confusion with its more technical sense in mechanics, where "dynamics" entails elucidating the forces underpinning movements. We did not characterize these forces.

[2] The EMA-MAE corpus is based on work supported by the National Science Foundation of the United States under Grant #IIS-1142826 to Marquette University, which support does not constitute an endorsement.

[3] The underlying signal processing details are described in He and Dellwo [7].

[4] The formula for calculating the triangle area based on 3D coordinates is $\frac{1}{2}\sqrt{\begin{vmatrix} y_{UL} & z_{UL} & 1 \\ y_{LL} & z_{LL} & 1 \\ y_{LC} & z_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} z_{UL} & x_{UL} & 1 \\ z_{LL} & x_{LL} & 1 \\ z_{LC} & x_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} x_{UL} & y_{UL} & 1 \\ x_{LL} & y_{LL} & 1 \\ x_{LC} & y_{LC} & 1 \end{vmatrix}^2}$. The MATLAB script (mouth_opening_area.mlx) for the computation is available via https://osf.io/qjh6p [15].

[5] URL: clarin.phonetik.uni-muenchen.de/BASWebServices

[6] For a series of numbers (x_1, x_2, x_3, …, x_n), the pairwise variability index (pvi) was calculated as $(|x_2 - x_1| + |x_3 - x_2| + \cdots + |x_n - x_{n-1}|) \div (n - 1)$. The pvi was initially developed to measure speech rhythm based on the durations of vocalic or intervocalic intervals [17].