Deriving the onset and offset times of planning units from acoustic and articulatory measurements

Joe Rodd,[1, 2, a)] **Hans Rutger Bosker,**[1, 3] **Louis ten Bosch,**[2, 1, 3] **and Mirjam Ernestus**[2, 1]

[1)] *Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH, Nijmegen, The Netherlands*

[2)] *Radboud University, Centre for Language Studies, P.O. Box 9103, 6500 HD, Nijmegen, The Netherlands*

[3)] *Radboud University, Donders Institute for Brain, Cognition and Behaviour, P.O. Box 9104, 6500 HE, Nijmegen, The Netherlands*

joe.rodd@mpi.nl,
hansrutger.bosker@mpi.nl,
l.tenbosch@let.ru.nl,
m.ernestus@let.ru.nl

Abstract: Many psycholinguistic models of speech sequence planning make claims about the onset and offset times of planning units, such as words, syllables, and phonemes. These predictions typically go untested, however, since psycholinguists have assumed that the temporal dynamics of the speech signal is a poor index of the temporal dynamics of the underlying speech planning process. This article argues that this problem is tractable, and presents and validates two simple metrics that derive planning unit onset and offset times from the acoustic signal and articulatographic data.

[a] Author to whom correspondence should be addressed.

1. Background

Typically, the inverse mapping between the acoustic signal and the articulator configuration is characterized as highly non-linear and one-to-many, in that many speech sounds can be produced by multiple configurations of the vocal tract (e.g. Lindblom, 1983). This assumed intractability complicated the evaluation of psycholinguistic models of speech planning, specifically claims about the implementation of abstract linguistic planning units by speech motor programs.

While it is the case that speakers can make use of alternative vocal tract configurations to achieve speech sounds when articulatory freedom is constrained (Lindblom, 1979), or to reduce the required movement from the previous configuration (e.g. Boyce and Espy-Wilson, 1997), the opacity of the correspondences between acoustics, articulation, and the dynamics of higher planning processes may be overestimated (Hogden *et al.*, 1996). This paper posits that the problem is tractable, and proposes methods to characterize the dynamics of higher planning processes from the acoustic signal or from tracked articulator movements. Thus, the testing of previously untestable predictions of psycholinguistic models is facilitated.

1.1 Acoustic change largely reflects articulatory change

Despite assumptions to the contrary, in practice, the inverse mapping from the acoustic signal to articulatory configurations can be defined in a appropriate way to predict articulatory configurations from the acoustic signal, within a certain tolerance for deviations in the articulatory domain. For speech sounds that intrinsically consist of multiple acoustic events (such as diphthongs, plosives), the mapping results in an estimated trajectory in articulatory space. For a subset of stable speech sounds, ‘codebooks’ of articulatory configurations associated with acoustic outcomes can be compiled (e.g. Hogden *et al.*, 1996). Moreover, machine learning approaches that can make use of contextual information and sufficiently large corpora of training data have proven successful in predicting articulatory configuration from the acoustic signal with no constraints on speech materials (e.g. Illa and Ghosh, 2018; Richmond, 2006; Uria *et al.*, 2011).

Relatedly, it holds that when the vocal tract is in a stable configuration, the acoustic output is also stable, and that when the acoustic output is changing, the vocal tract configuration must also be changing. This observation has been exploited in blind speech segmentation, where frame-by-frame changes in the acoustic spectrum are tracked, and peaks in spectral change are detected. These peaks correspond to perceptually relevant phone boundaries (e.g. Dusan and Rabiner, 2006; Hoang and Wang, 2015; ten Bosch and Cranen, 2007). These approaches are intended to automate the preparation of corpora to test speech recognition systems, and assume that segments are concatenated without overlap, making these algorithms unsuited for the retrieval of onset and offset times of overlapping planning units predicted by psycholinguistic models. They can, however, serve as inspiration for the development of new techniques to retrieve planning unit dynamics.

Note that although changes in the acoustic signal must reflect changes in the articulatory configuration, it does not follow that when the vocal tract configuration is changing, the acoustic signal always changes with it, since for many speech sounds, the precise positioning of non-critical articulators is unimportant (such as tongue position during the realization of /m/).

1.2 The mapping between planning units and acoustics and articulation

A class of psycholinguistic speech production models (which we will term phoneme-based models) characterize the units that mediate between formulation (lexical access and phonological encoding) and execution (speech motor programming and articulation itself) as phonemes, or sequences of phonemes, such as syllables, demi-syllables, or whole words (e.g. Dell and O'Seaghdha, 1992; Levelt, 1989; Levelt *et al.*, 1999; Tourville and Guenther, 2011). Phoneme-based models also conceptualize the execution process as an obedient servant of formulation (e.g. Dell and O'Seaghdha, 1992; Levelt, 1989; Levelt *et al.*, 1999; Tourville and Guenther, 2011), which entails that the observable movements of the articulators and the resulting speech acoustics are inherently a consequence of planning units in formulation becoming active and subsequently being deactivated. That the dynamics of the activation of planning units directly influences the articulatory configuration and thereby the acoustic output seems plausible in the light of findings that competing representations in the formulation phase exert some influence on fine detail in articulation (e.g. Goldrick and Blumstein, 2006).

The DIVA model (Tourville and Guenther, 2011) operationalizes the planning units by defining them in terms of upper and lower bounds for articulator positions, and upper and lower bounds of the expected auditory outcome in terms of fundamental frequency and formants. Planning units typically overlap in time, and all simultaneously active planning units exert influence on both the articulatory configuration and speech acoustics directly via the feedforward route. They also influence articulation and acoustics indirectly by shaping the expected acoustic and somatosensory outcomes, which in turn lead to corrective feedback.

The temporal overlap of adjacent planning units (at the output stage of phoneme-based psycholinguistic speech planning models) results in local coarticulation in the overt speech. Equivalently, low level pre-activation (priming) of upcoming planning units and incomplete deactivation of preceding planning units result in longer-range coarticulation in the overt speech.

The retrieval of planning units from articulatory measurements has previously been attempted by Steiner and Richmond (2009), who developed an analysis-by-resynthesis approach that reconstructs a gestural score from electromagnetic articulography (EMA) data in terms of vocalic and consonantal gestures for the VocalTractLab (VTL) synthesizer (Birkholz *et al.*, 2007). This representation differs somewhat from that inherent to phoneme-based models, in that vowel and consonants are treated as fundamentally distinct units of representation on distinct tiers of the gestural score, while phoneme-based models instead predict a chain of potentially overlapping planning units of the same class, on the same tier.

Vaz *et al.* (2016) described an algorithm to retrieve underlying structure from multivariate time series data, and tested it on vocal tract constriction distances measured from real-time MRI vocal tract data. The algorithm was able to construct an inventory of gestures from the data, and an activation time series for each of these gestures, which are collectively analogous to a gestural score in the articulatory phonology (AP) framework. AP diverges from phoneme-based production models in that the planning units it supposes are not phonemes or sequences of phonemes, but rather articulatory gestures defining articulatory events, such as opening of the glottal aperture, or the creation of a labial closure (Browman and Goldstein, 1992), which cannot easily be translated into phonemes.

The direct retrieval of the timings of planning units from the acoustic signal has been attempted by Nam *et al.* (2012), again with an analysis-by-synthesis approach, and similarly rooted in the articulatory phonology (AP) framework. Their procedure involves constructing a task dynamic gestural score (encoding the speech to be produced in terms of degrees of constriction at different positions in the vocal tract) from an orthographic transcription of the speech. Then, the TADA model (Nam *et al.*, 2004; Saltzman and Munhall, 1989) is used to predict time-varying vocal tract dimensions from the gestural score, which is then synthesized to produce a speech signal. Next, dynamic time warping (DTW) is applied between the synthesized and natural speech signals. This involves stretching and compressing the synthesized speech signal in the temporal dimension, to improve the temporal alignment with the natural speech signal. The result of the DTW is a warping scale, which can then be applied to the gestural score, yielding a warped gestural score from which activation and deactivation times of individual gestures can be established.

Aside from requiring potentially difficult to acquire articulatory measurements (EMA in the case of Nam *et al.* (2012), real time MRI in the case of Vaz *et al.* (2016)), these procedures that construct multivariate gestural scores cannot readily be applied to phoneme-based models of speech production, since the gestures are not consistent with or easily mapped to the planning units hypothesized by phoneme-based models of lexical access and multi-word processes of speech production (e.g. Bohland *et al.*, 2010; Dell and O'Seaghdha, 1992; Levelt, 1989; Levelt *et al.*, 1999). An additional concern is that the process leaves the researcher relatively unconstrained in the construction of the gestural score for a given utterance, either directly or through their parameterization of the linguistic model.

2. Study aims

This study aims to provide a means to estimate the onset and offset times of phoneme-based planning units (such as words, syllables or phonemes) from recorded speech materials. The tight temporal locking between formulation and execution processes in speech production (e.g. Goldrick and Blumstein, 2006) suggests that reconstructing the activation dynamics of planning units from measurements of articulator movement is feasible. That the inverse mapping between acoustics and articulation is transparent enough to construct codebooks describing the mapping implies that reconstructing the activation dynamics of planning units from the acoustic signal should also be feasible for a constrained repertoire of speech sounds.

We propose two approaches to retrieve planning unit onset and offset times from speech materials; from the acoustic signal, and from EMA data. We compare the outcomes of the two techniques, to establish that recovering planning unit onset and offset times from the acoustic signal is broadly equivalent to recovering planning unit timing from articulatographic data.

The first metric uses fleshpoint position data gathered by electromagnetic articulography, and begins by deriving upper and lower bounds for each fleshpoint position for each segment from corpus data. Subsequently, a multi-dimensional, time-varying target for a multi-segmental speech sequence is constructed, the temporal parameters of which are adjusted to achieve a good fit to the observed data.

The second is a metric that exploits the acoustic signal directly with no need to record articulator motion, but constrains the speech sounds that can be evaluated. This metric depends on the claim that acoustic instability mirrors articulatory instability, which in turn reflects simultaneous activation of multiple planning units.

Neither metric is predicated on any specific theoretical treatment of speech production, aside from the assumption that planning units are phonemes or sequences of phonemes, and the parameterization of both metrics is wholly data-driven. For the experimental psycholinguist, a metric that can be collected from the acoustic signal alone is clearly preferable, since that reduces the burden of data collection on both researcher and participant, and makes recording of electrophysiological or other measures during speech production possible because no articulatographic data needs to be collected.

The two metrics were tested on acoustic and articulatory data for the same vowel-consonant sequences, taken from the electromagnetic articulography subset of the mngu0 corpus (Richmond *et al.*, 2011), where monophthongs transitioned into continuant consonants. The choice of this limited subset was driven by the need to use segments that were acoustically stable during realization, for the acoustic metric. Comparing the performance of the metrics against a ‘gold standard’ baseline annotation of the onsets and offsets of speech planning units is clearly impossible, given that any hand annotation of speech planning unit onsets and offsets would inherently be largely arbitrary and noisy.

3. Speech materials

The EMA subset of the mngu0 corpus (Richmond *et al.*, 2011) was used, which consists of TIMIT sentences read by a single male speaker of British English. EMA sensors were placed on the lower and upper lips, at the tongue tip, blade and dorsum and on the lower incisors (to track jaw motion). A further sensor was placed on the upper incisors to serve as a reference for the others. For technical details relating to the data collection and preparation see Richmond *et al.* (2011).

3.1 Post processing and annotation

From the 1263 sentences of the mngu0 corpus, vowel - consonant (VC) sequences of interest were identified, where a monophthong transitioned into a continuant consonant. The sequences of interest were all one of the following: /am/, /aʃ/, /av/, /ɪʃ/, /ɪv/, /im/, /iv/, /ʌm/, /is/, /ʌs/, /ɒn/. n/. Note that in the context of phoneme-based speech planning models, where no distinction is made between planning units for different classes of phonemes, there is no reason to suppose that sequences of a different composition (CVs, or CCs, for instance) would behave any differently from the VCs tested here. This means that the predictions of phoneme-based speech planning models can effectively be tested by this reduced set of sequences. This yielded 775 sequences of interest, which were identified based on the forced aligned transcriptions available in the corpus. Analysis intervals from the temporal center of the forced aligned vowel to the temporal center of the forced-aligned consonant were defined (see Figure 1(a)). The analysis interval served as a landmark to identify the planning unit transitions found; so the precision of the start and end points of the interval was not critical, as long as the transition between the planning units was included.

In the EMA data, lateral movement was discarded, yielding articulator positions on the mid-sagittal plane only. To facilitate annotation, the remaining two dimensional data was rotated independently for each sensor by means of principal component analysis, so that PC1 captured the most informative direction of movement for that sensor, which in all cases was the open-close dimension. Since PC2 is orthogonal to PC1, it captured forward-backward movement of each sensor. Then, manual annotation was undertaken (by the first author) to identify articulatory stable periods of each segment for use in the preparation of the targets used in the articulatory metric. In the manual annotation procedure, movement tracks in PC1-PC2 dimensions were displayed on a graphical interface, in which the periods of stability associated with the vowel and continuant consonant could be highlighted. The articulatory configuration was considered stable if there was little to no change (assessed visually) in several sensors. Since the targets were defined in terms of 95% highest density intervals (see section 4.1), some noise in this annotation procedure was acceptable.

4. Planning unit timing from articulatory measurements

The articulatory metric approaches the identification of planning unit onset and offset times from EMA data by essentially inverting the motor control process: reconstructing a multidimensional articulatory target that could have lead to the recorded movements during

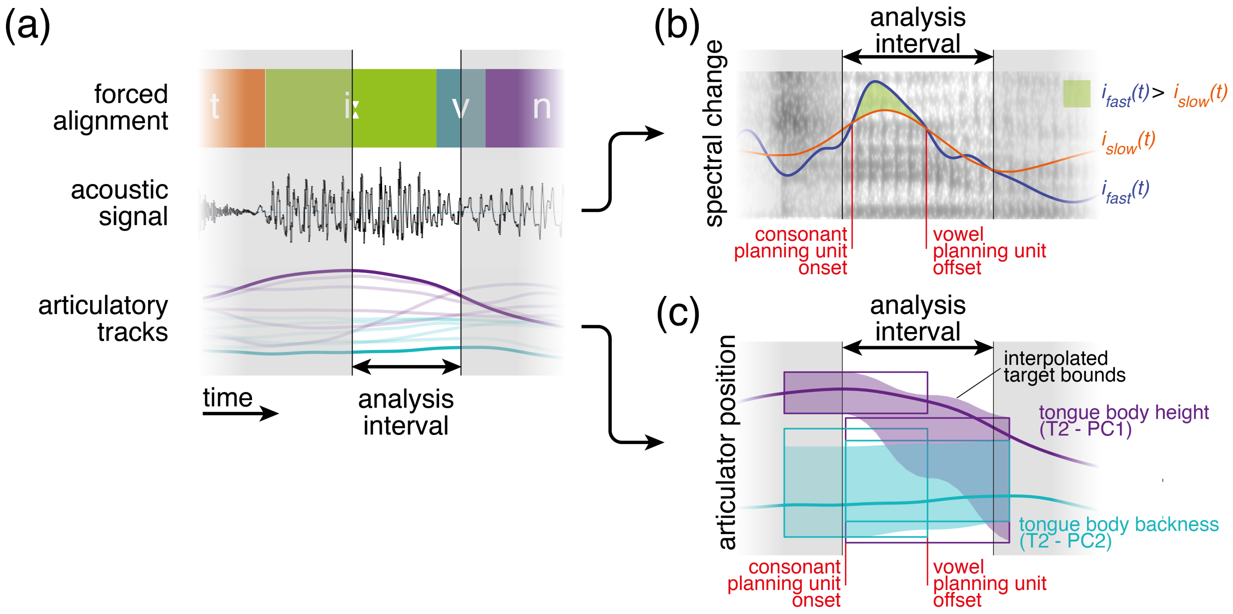
(a)
forced alignment
t
iː
v
n
acoustic signal
articulatory tracks
time
analysis interval
(b)
spectral change
analysis interval
ifast(t)> islow(t)
islow(t)
ifast(t)
consonant planning unit onset
vowel planning unit offset
(c)
articulator position
analysis interval
interpolated target bounds
tongue body height (T2 - PC1)
tongue body backness (T2 - PC2)
consonant planning unit onset
vowel planning unit offset

Fig. 1. (color online) An example analysis. (a) An analysis interval is defined that stretches from the temporal center of the forced aligned vowel to the center of the forced aligned consonant. (b) The acoustic metric. Lines show $I_{slow}(t)$ and $I_{fast}(t)$, the Gaussian smoothed, interpolated spectral distance functions used to identify the acoustically evident planning unit overlap during the analysis interval. $I_{slow}(t)$ has a 90 ms kernel, $I_{fast}(t)$ has a 30 ms kernel. Shading identifies periods of atypically fast acoustic change (where $I_{fast}(t) > I_{slow}(t)$), from which the onset of the consonant planning unit and the offset of the vowel planning unit are derived. (c) The articulatory metric. The heavy lines indicate the recorded movement of the tongue body sensor, in the open-close dimension (PC1) and the forward-backward dimension (PC2). The outlined boxes indicate the segmental targets, the shading indicates the interpolated sequence level target.

a vowel-consonant sequence. This was done separately for each vowel-consonant transition token, using a parameter optimization routine which adjusted the onset and offset times of the segment targets to construct a target that fitted the recorded movements well.

4.1 Establishing segment targets

First, separate segmental targets are established for the vowels and for the consonants, defined in terms of upper and lower bounds for the positions for each fleshpoint (lower jaw, upper and lower lips, tongue tip, blade and dorsum) on the two dimensions (principal components) of the mid-sagittal plane. These maxima and minima are derived from the distribution of sensor positions during the hand-annotated stable periods of those segments in the corpus, irrespective of context, by extracting the 95% highest density interval(s). When the positioning of a fleshpoint is of crucial importance to the identity of the segment, the positioning of that fleshpoint varies little between realizations, and the target is therefore narrow (e.g. the positioning of the tongue tip in /s/). When the positioning of a fleshpoint is only marginally relevant for the identity of the segment, the target is broad (e.g. the positioning of the tongue back in /v/), since there is lots of variability in the source data.

4.2 Combining segmental targets to form a sequence target

The sequence targets were constructed by temporally-overlapping the vowel and consonant targets. Figure 1(c) depicts an example of the construction of the targets, for the sequence /iːv/, showing the target bounds for each segment as boxes (purple for PC1, blue for PC2), for the tongue body sensor. The segmental targets are fixed at the outer edges, such that the vowel target begins at the hand-annotated onset of vowel stability, and the consonant target ends at the hand-annotated offset of consonant stability. The other two temporal parameters, the offset of the vowel target and the onset of the consonant target are free parameters that can be optimized.

The upper bound of the sequence target is calculated as an exponential moving average (with a window of 20 ms) of the upper bounds of the segmental targets over time. This means that for time points when only the vowel target is engaged, the upper bound is equal to the upper bound of the vowel target. When both segmental targets are engaged, however, the upper bound switches smoothly from following the upper bound of the vowel target to reflecting the average upper bound of both targets. Once the vowel target is disengaged, the upper bound again smoothly shifts to reflect the upper bound of the consonant target. The lower bound of the target is derived in the same way.

4.3 Parameter optimization

For each analysis interval, an independent parameter optimization routine is conducted. Two parameters, the onset time of the consonant target and the offset time of the vowel target, are optimized with the BOBYQA algorithm (Powell, 2009; Ypma *et al.*, 2018).

To evaluate how well a sequence target defined by a pair of consonant target onset and vowel target offset times fitted the observed movements, the proportion of time points where the recorded sensor positions are outside the bounds of the multidimensional target is counted. This proportion is used as a score to be minimized during the parameter optimization process.

For each realization, 200 starting points for these parameters are tried, sampled from normal distributions ($SD = 25$ ms) centered around the annotated end of vowel stability (this is the center-point of the starting distributions for the consonant onset parameter) and the annotated beginning of consonant stability (this is the center-point of the starting distributions for the vowel offset parameter). A search space constraint ensures that the algorithm only considers solutions where the overlap between the segment targets is greater than 0. Having multiple starting points allows us to assess how consistently the algorithm selects the best performing parameter sets, and offers more protection from premature convergence to local minima. To select a single vowel offset time and a single consonant onset time from the distributions that resulted from the 200 initializations, a two-dimensional distribution is estimated from the resulting parameters, where the dimensions are the vowel offset time parameter and consonant onset time parameter. The distribution is weighted by one minus the score achieved in each attempt, so as to weight the best performing solutions most heavily, and the peak is identified. The coordinates of this peak define the planning unit onset and offset times.

5. Planning unit timing from the acoustic signal

The acoustic metric quantifies the rate of change in the acoustic signal (the spectral change). Local peaks in this signal identify periods where the speech acoustics, and therefore the underlying vocal tract configuration, are changing. At the transition between two planning units, this change is due to the interaction of the two overlapping planning units, and the duration of the instability is equated with the duration of the overlap. We term this overlap 'acoustically evident planning unit overlap'. To be able to establish the onset and offset of instability, a method is required to transform a continuous signal into a categorical one: to distinguish acoustic stability from instability. This is done by overlaying two different smoothings of this signal; a 'fast' smooth that captures local changes in the signal, and a 'slow' smooth that captures longer trends. We identify periods when the 'fast' smooth exceeds the 'slow' smooth as unstable, and other periods as stable. The onset of the second planning unit is equated with the start of such a period of instability. The offset of the first planning unit is equated with the end of that same period of instability. This is illustrated in Figure 1(b).

5.1 Step 1: quantifying acoustic change

To identify the period of overlap, the MFCC vectors (mel frequency cepstral coefficient; 25 ms analysis frame length, samples every 10 ms) for the analysis intervals (with a margin of 40 ms before and after) are extracted using the HTK front end (Young *et al.*, 2006). MFCC vectors may be seen as a numeric representation of the spectral content of the speech signal during a short (25 ms) window, and are one of the best spectro-temporal representations of speech acoustics. From each frame to the next, the Euclidean distance in MFCC space was calculated as follows, where j is the index of the MFCC coefficient and t is the index of the frame:

$$D_{spec} = \sqrt{\sum_{j=0}^{12} (\mathrm{MFCC}_{j_t} - \mathrm{MFCC}_{j_{t+1}})^2} \qquad (1)$$

This gives $D_{spec}(t)$, a spectral distance function quantifying the degree of spectral change evident in the acoustic signal, sampled every 10 ms.

5.2 Step 2: identifying periods of fast acoustic change

This spectral distance function is smoothed twice, once with a 30 ms wide Gaussian kernel, yielding $D_{fast}(t)$, which captures relatively fast changes in the spectral distance function; and once with a 90 ms wide Gaussian kernel, yielding $D_{slow}(t)$, which captures longer term trends in the function.

Spline interpolation (every 0.1 ms) is then applied to these functions in order to improve temporal resolution, yielding $I_{fast}(t)$ and $I_{slow}(t)$. The two interpolated functions are overlaid, and parts of the signal in each analysis interval where $I_{fast}(t)$ is larger than $I_{slow}(t)$ are identified as candidate overlaps (in Figure 1(b) shown as green shading). Where $I_{fast}(t)$ exceeds $I_{slow}(t)$, atypically fast acoustic change is occurring: acoustically evident planning unit overlap. It is possible that there are multiple periods where $I_{fast}(t)$ exceeds $I_{slow}(t)$, however, typically one period is longer and the associated peak is larger. Therefore, a heuristic is engaged to select precisely one period per analysis window: the duration of each of these periods is calculated. Periods that cross the boundaries of the analysis interval (into the margins) are discarded. When an analysis interval still contains multiple periods, all but the longest candidate are discarded. This yields precisely one period of acoustically evident planning unit overlap per analysis interval. The onset of the remaining period of overlap (where $I_{fast}(t)$ becomes larger than $I_{slow}(t)$) yields the onset of the consonant planning unit.

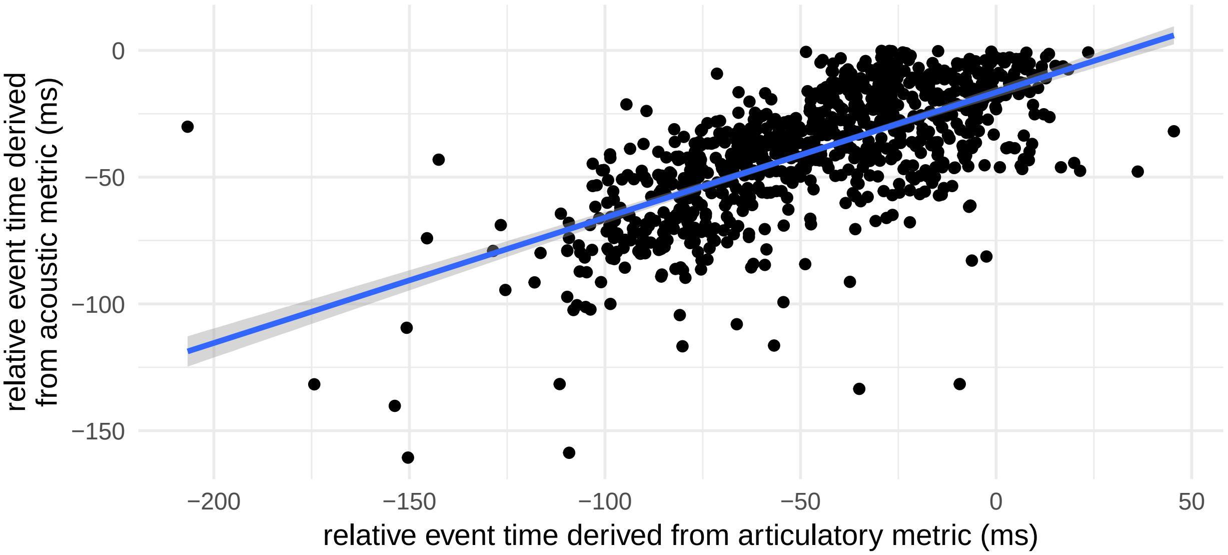

Fig. 2. (color online) The correlation between planning unit onset and offset times, derived from the articulatory (x-axis) and acoustic metrics (y-axis). All event times are relative to the forced-aligned offset of the consonant segment, meaning that times less than 0 are to be expected.

The offset of the overlap (where $I_{fast}(t)$ becomes smaller than $I_{slow}(t)$) yields the offset of the vowel planning unit.

This procedure was refined by testing various kernel widths and interpolations via a grid search, in which the the parameters that resulted in the highest spectral change peak were selected.

R scripts implementing the two metrics and the data preprocessing method are available from https://git.io/fh8EM.

6. Results and discussion

6.1 Validity of the metrics

Figure 2 shows the onsets and offsets of planning units (event times) as predicted by the articulatory (x-axis) and acoustic metrics (y-axis). All event times are relative to the forced-aligned offset of the consonant segment, meaning that times less than 0 are to be expected. An r^2 of 0.447 was calculated between the event times derived by the two metrics. This moderately high correlation between the predictions of the two metrics indicates that they both capture the same underlying dynamic process of planning unit activation.

The intercept of -10.64 indicates that the acoustic metric systematically predicts earlier event times than the articulatory metric does. This is approximately half the width of the 25 ms analysis window employed in the acoustic metric, which suggests that this anticipation may be an artifact of the spectral analysis inherent to the acoustic metric.

6.2 Reliability

The metrics were evaluated by comparing the planning unit onset and offset times predicted by each metric. Because the two metrics are so divergent in the modality of the data used and the approach used to derive event times from the data, we interpreted the finding that the two metrics predicted comparable event times as evidence that they are both indexing the onset and offset times of planning units. This is of course weaker evidence in support of the validity of a metric than comparison against data capturing the ground truth, but the ground truth is clearly unobtainable for psychological processes such as the activation dynamics of planning units. Comparison against the results obtained by Nam *et al.* (2012) is also problematic given the AP theoretical framing inherent to their procedure.

6.3 Applicability and ecological validity

The articulatory metric is in principle equally suited to examining transitions between any pair of segments where there is at least a short period of articulatory stability in each segment, including stops. Of course, given the metric-comparison approach we took to evaluate the performance of the two metrics, the articulatory metric was only tested on materials also suitable for the acoustic metric.

The acoustic metric is inherently limited to identifying planning unit onset and offset times at transitions between a subset of segment types involving at least a short period of articulatory stability and incomplete obstruction of the airflow: monophthong vowels, nasals and continuant fricatives. Nevertheless, for the experimental psycholinguist, the convenience of the acoustic-only recording may well outweigh the disadvantage of constrained material selection.

Both metrics share the inherent assumption that the onsets of all the movements or gestures involved in the production of a phoneme are synchronized. This assumption is inherent to the class of phoneme-based models, which form the mainstream in psycholinguistic models of higher speech planning. Adhering to it was necessary to achieve this paper's goal of making it possible to test and refine phoneme-based models by relating activation dynamics to the speech signal. Models based on a multivariate gestural score may achieve better fits to the data given that they are not constrained by this synchronicity assumption.

The metrics were developed and tested using the mngu0 corpus (Richmond *et al.*, 2011), which contains a large quantity of English data from a single speaker, rather than smaller quantities of data from multiple speakers available in other corpora (e.g. the Wisconsin x-ray microbeam database, Westbury *et al.*, 1990). The mngu0 corpus was selected because we sought to have a large number of realizations of each segment to reliably compute the static segment targets for the articulatory metric. It remains to be seen how the articulatory metric would perform given a smaller dataset from which to derive target boundaries. A requirement for a large speaker-specific dataset would be disadvantageous in the context of experimental psycholinguistics, where it is typically desirable to test multiple speakers on a small set of materials, though recent success in using a generalized background model and a speaker-specific adaptive model in acoustic-to-articulatory inversion (Illa and Ghosh, 2018) offers hope that a comparable approach could work for this metric too.

7. Conclusion

This paper presented two techniques to identify planning unit onsets and offsets from articulographic and acoustic data in the context of phoneme-based models of speech production. The first metric requires articulographic recording, but imposes less constraint on speech material selection. The second metric exploits the acoustic signal directly, with no need to record articulator motion, but constrains the speech sounds that can be evaluated. This metric depends on the claim that acoustic instability mirrors articulatory instability, which in turn reflects simultaneous activation of multiple planning units. The two metrics are agnostic to the duration of planning units (syllables, demi-syllables, phonemes, entire words), and make minimal assumptions about precisely what is encoded by the planning unit, other than that upper and lower bounds for articulatory positions are encoded. A moderately high correlation between the event times predicted by the two metrics indicates that they capture the same underlying dynamic process of planning unit activation. This correlation means in turn that temporal predictions arising from phoneme-based psycholinguistic models of speech planning can be tested using the acoustic signal without the need to collect articulographic data.

Acknowledgments

This research was supported by Netherlands Organization for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium. We are grateful to Antje S. Meyer for helpful discussions and useful comments on this manuscript.

References and links

Birkholz, P., Steiner, I., and Breuer, S. (**2007**). “Control concepts for articulatory speech synthesis,” in *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis*, edited by P. Wagner, J. Abresch, S. Breuer, and W. Hess, Bonn.

Bohland, J. W., Bullock, D., and Guenther, F. H. (**2010**). “Neural representations and mechanisms for the performance of simple speech sequences,” Journal of cognitive neuroscience **22**(7), 1504–1529.

Boyce, S., and Espy-Wilson, C. Y. (**1997**). “Coarticulatory stability in American English /r/,” The Journal of the Acoustical Society of America **101**(6), 3741–3753, https://asa.

scitation.org/doi/abs/10.1121/1.418333, doi: 10.1121/1.418333.

Browman, C. P., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**(3-4), 155–180.

Dell, G. S., and O'Seaghdha, P. G. (**1992**). "Stages of lexical access in language production," Cognition **42**(1–3), 287–314, http://www.sciencedirect.com/science/article/pii/001002779290046K, doi: 10.1016/0010-0277(92)90046-K.

Dusan, S., and Rabiner, L. (**2006**). "On the relation between maximum spectral transition positions and phone boundaries," in *Ninth International Conference on Spoken Language Processing*.

Goldrick, M., and Blumstein, S. E. (**2006**). "Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters," Language and Cognitive Processes **21**(6), 649–683, doi: 10.1080/01690960500181332.

Hoang, D.-T., and Wang, H.-C. (**2015**). "Blind phone segmentation based on spectral change detection using Legendre polynomial approximation," The Journal of the Acoustical Society of America **137**(2), 797–805.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., and Saltzman, E. (**1996**). "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," The Journal of the Acoustical Society of America **100**(3), 1819–1834.

Illa, A., and Ghosh, P. K. (**2018**). "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," Proc. Interspeech 2018 3122–3126.

Levelt, W. J. M. (**1989**). *Speaking: From intention to articulation* (MIT press, Cambridge, MA).

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (**1999**). “A theory of lexical access in speech production,” Behavioral and brain sciences **22**(01), 1–38.

Lindblom, B. (**1979**). “Formant frequencies of some fixed-mandible vowel and a model of speech motor programming by predictive simulation,” Journal of Phonetics **7**, 147–162.

Lindblom, B. (**1983**). “Economy of Speech Gestures,” in *The Production of Speech*, edited by P. F. MacNeilage (Springer New York), pp. 217–245, http://link.springer.com/chapter/10.1007/978-1-4613-8202-7_10, doi: 10.1007/978-1-4613-8202-7_10.

Nam, H., Goldstein, L., Saltzman, E., and Byrd, D. (**2004**). “TADA: An enhanced, portable Task Dynamics model in MATLAB,” The Journal of the Acoustical Society of America **115**(5), 2430–2430.

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., and Goldstein, L. (**2012**). “A procedure for estimating gestural scores from speech acoustics,” The Journal of the Acoustical Society of America **132**(6), 3980–3989.

Powell, M. J. (**2009**). “The BOBYQA algorithm for bound constrained optimization without derivatives,” Cambridge NA Report NA2009/06, University of Cambridge, Cambridge 26–46.

Richmond, K. (**2006**). “A trajectory mixture density network for the acoustic-articulatory inversion mapping,” in *Ninth International Conference on Spoken Language Processing*.

Richmond, K., Hoole, P., and King, S. (**2011**). "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Twelfth Annual Conference of the International Speech Communication Association*.

Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological psychology **1**(4), 333–382.

Steiner, I., and Richmond, K. (**2009**). "Towards unsupervised articulatory resynthesis of German utterances using EMA data," in *Tenth Annual Conference of the International Speech Communication Association*.

ten Bosch, L. F. M., and Cranen, B. (**2007**). "A computational model for unsupervised word discovery," in *Proc. Interspeech 2007*, pp. 1481–1884.

Tourville, J. A., and Guenther, F. H. (**2011**). "The DIVA model: A neural theory of speech acquisition and production," Language and Cognitive Processes **26**(7), 952–981, http://dx.doi.org/10.1080/01690960903498424, doi: 10.1080/01690960903498424.

Uria, B., Renals, S., and Richmond, K. (**2011**). "A deep neural network for acoustic-articulatory speech inversion," in *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, Citeseer.

Vaz, C., Toutios, A., and Narayanan, S. S. (**2016**). "Convex Hull Convolutive Non-Negative Matrix Factorization for Uncovering Temporal Patterns in Multivariate Time-Series Data.," in *INTERSPEECH*, pp. 963–967.

Westbury, J., Milenkovic, P., Weismer, G., and Kent, R. (**1990**). "X?ray microbeam speech production database," The Journal of the Acoustical Society of America **88**(S1), S56–S56, https://asa.scitation.org/doi/abs/10.1121/1.2029064, doi: 10.1121/1.2029064.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D. *et al.* (**2006**). “The HTK book (for HTK version 3.4),” Cambridge University Engineering Department **2**(2).

Ypma, J., Borchers, H. W., and Eddelbuettel, D. (**2018**). “Package ‘nloptr”’ R package.